Российская академия наук
Российский фонд фундаментальных исследований
Карельский научный центр РАН
Институт прикладных математических исследований
Петрозаводский государственный университет
Институт проблем информатики РАН
Московская секция ACM SIGMOD

# Электронные библиотеки: перспективные методы и технологии, электронные коллекции

**XI Всероссийская научная конференция RCDL'2009**

Петрозаводск, 17–21 сентября 2009 г.

*Труды конференции*

---

Russian Academy of Sciences
Russian Foundation for Basic Research
Karelian Research Center of the Russian Academy of Sciences
Institute of Applied Mathematical Research
Petrozavodsk State University
Institute of Informatics Problems of the Russian Academy of Sciences
Moscow ACM SIGMOD Chapter

# Digital Libraries: Advanced Methods and Technologies, Digital Collections

**XI All-Russian Research Conference RCDL'2009**

Petrozavodsk, September 17–21, 2009

*Proceedings of the Conference*

Петрозаводск • 2009

УДК 002:004.9 (063)
ББК 78.3
Э 45

Электронные библиотеки: перспективные методы и технологии, электронные Э 45 коллекции: Труды XI Всероссийской научной конференции RCDL'2009 (Петрозаводск, Россия, 17–21 сентября 2009 г.). – Петрозаводск: КарНЦ РАН, 2009. – 487 [1] с.: ил.

ISBN 978-5-9274-0374-5

Электронные библиотеки – область исследований и разработок, направленных на развитие теории и практики обработки, распространения, хранения, анализа и поиска цифровых данных различной природы. Основная цель серии конференций RCDL (http://rcdl.ru/) заключается в формировании сообщества специалистов России, ведущих исследования и разработки в области электронных библиотек и близких областях. Всероссийская научная конференция 2009 г. (RCDL'2009) является одиннадцатой конференцией по данной тематике (1999 г. – Санкт-Петербург, 2000 г. – Протвино, 2001 г. – Петрозаводск, 2002 г. – Дубна, 2003 г. – Санкт-Петербург, 2004 г. – Пущино, 2005 г. – Ярославль, 2006 г. – Суздаль, 2007 г. – Переславль-Залесский, 2008 г. – Дубна).

Настоящий сборник включает тексты докладов, коротких сообщений и стендовых докладов, отобранных Программным комитетом RCDL'2009 в результате проведенного рецензирования.

Конференция организована при поддержке Российского фонда фундаментальных исследований и Российской академии наук.

Digital Libraries: Advanced Methods and Technologies, Digital Collections: Proceedings of the XI All-Russian Research Conference RCDL'2009 (Petrozavodsk, September 17–21, 2009). – Petrozavodsk: KRC RAS, 2009. – 487 [1] p.: il.

ISBN 978-5-9274-0374-5

Digital Libraries is a field of research and development aiming to promote the theory and practice of processing, dissemination, storage, search and analysis of various digital data. The purpose of the series of All-Russian Research Conferences on Digital Libraries (RCDL, http://rcdl.ru) is to stimulate consolidation of the Russian digital libraries community and encourage research in this field. All-Russian Research Conference RCDL'2009 is the Eleventh Conference on this subject (1999 – St. Petersburg, 2000 – Protvino, 2001 – Petrozavodsk, 2002 – Dubna, 2003 – St. Petersburg, 2004 – Puschino, 2005 – Yaroslavl, 2006 – Suzdal, 2007 – Pereslavl-Zalessky, 2008 – Dubna). The RCDL'2009 Proceedings include the texts of reports, short papers and posters selected by Program Committee RCDL'2009.

The conference was organized with the support of the Russian Foundation for Basic Research and the Russian Academy of Sciences.

ISBN 978-5-9274-0374-5 

# XI Всероссийская научная конференция RCDL'2009

# «Электронные библиотеки: перспективные методы и технологии, электронные коллекции»

**Петрозаводск, 17 – 21 сентября 2009 года**

http://rcdl2009.krc.karelia.ru/

## Руководящий комитет конференции RCDL

**Председатель:** Л.А. Калиниченко (Москва)

**Члены комитета:**

А.Е. Авраменко (Пущино)
А.Б. Антопольский (Москва)
П.И. Браславский (Екатеринбург)
В.Т. Вдовицын (Петрозаводск)
В.Н. Захаров (Москва)
М.Р. Когаловский (Москва)
А.Г. Марчук (Новосибирск)
И.С. Некрестьянов (Санкт-Петербург)
Ю.Г. Сметанин (Москва)
В.Н. Смирнов (Ярославль)
В.П. Шириков (Дубна)

## Координация ECDL/RCDL: Andreas Rauber, Леонид Андреевич Калиниченко

## Организаторы конференции

Российская академия наук
Российский фонд фундаментальных исследований
Карельский научный центр РАН
Институт прикладных математических исследований
Петрозаводский государственный университет
Институт проблем информатики РАН
Московская секция ACM SIGMOD

## Организационный комитет

**Председатель:** Мазалов Владимир Викторович – директор Института прикладных математических исследований (ИПМИ) Карельского научного центра (КарНЦ) РАН

**Члены оргкомитета:**

Вдовицын Владимир Трофимович – Карельский научный центр РАН
Захаров Виктор Николаевич – Институт проблем информатики РАН
Калиниченко Леонид Андреевич – Институт проблем информатики РАН
Луговая Наталья Борисовна – ИПМИ КарНЦ РАН, ученый секретарь конференции

Печников Андрей Анатольевич – ИПМИ КарНЦ РАН
Рогов Александр Александрович – Петрозаводский государственный университет
Рогова Ксения Александровна – Петрозаводский государственный университет
Рузанова Наталья Сократовна – Петрозаводский государственный университет
Сметанин Юрий Геннадиевич – Российский фонд фундаментальных исследований
Сорокин Анатолий Дмитриевич – ИПМИ КарНЦ РАН

## Программный комитет

**Сопредседатели:**

Вдовицын Владимир Трофимович – Карельский научный центр РАН
Когаловский Михаил Рувимович – Институт проблем рынка РАН

**Члены программного комитета:**

Авраменко Аркадий Ефимович – Пущинская радиоастрономическая обсерватория
Антонов Александр Викторович – ЗАО "Галактика Софт", Москва
Богоявленский Юрий Анатольевич – Петрозаводский государственный университет
Borbinha Jose – IST / INESC-ID
Браславский Павел Исаакович – Институт машиноведения УрО РАН
Вольфенгаген Вячеслав Эрнстович – Московский инженерно-физический институт
Губин Максим Вадимович – Ask.Com IAC Search&Media
Добров Борис Викторович – НИВЦ МГУ
Елизаров Александр Михайлович – НИИ математики и механики им. Н.Г. Чеботарева
Жижимов Олег Львович – Институт вычислительных технологий СО РАН
Захаров Виктор Николаевич – Институт проблем информатики РАН
Знаменский Сергей Витальевич – Институт программных систем РАН
Калиниченко Леонид Андреевич – Институт проблем информатики РАН
Kulovits Hannes – Vienna University of Technology
Лавренова Ольга Александровна – ФГУ Российская государственная библиотека
Litvine Vladimir – California Institute of Technology
Малков Олег Юрьевич – Институт астрономии РАН
Марчук Александр Гурьевич – Институт систем информатики им. А.П. Ершова СО РАН
Мусульманбеков Женис Жумкенович – Объединенный институт ядерных исследований
Некрестьянов Игорь Сергеевич – Санкт-Петербургский государственный университет
Палей Дмитрий Эзрович – Ярославский государственный университет
Плешко Владимир Владимирович – RCO
Проскудина Галина Юрьевна – Институт программных систем НАН Украины
Rauber Andreas – Vienna University of Technology
Рогов Александр Александрович – Петрозаводский государственный университет
Sandkuhl Kurt – Jönköping University, Sweden
Сегалович Илья Валентинович – ООО "Яндекс"
Серебряков Владимир Алексеевич – Вычислительный центр им. А.А.Дородницына РАН
Смирнов Владимир Николаевич – Ярославский государственный университет
Solvberg Ingeborg Torvik – Norwegian University of Science and Technology
Ушаков Алексей Семенович – Калифорнийский университет, Санта-Барбара, США
Федотов Анатолий Михайлович – Новосибирский государственный университет
Шириков Владислав Павлович – Объединенный институт ядерных исследований

# XI All-Russian Research Conference RCDL'2009

# «Digital Libraries: Advanced Methods and Technologies, Digital Collections»

**Petrozavodsk, September 17 – 21, 2009**

http://rcdl2009.krc.karelia.ru/

## Steering committee of the RCDL conferences

**Chair:** Leonid Kalinichenko RAS (Moscow)

**Members:**

Arkady Avramenko (Pushchino)
Alexander Antopolsky (Moscow)
Pavel Braslavsky (Ekaterinburg)
Mikhail Kogalovsky (Moscow)
Alexander Marchuk (Novosibirsk)
Igor Nekrestjanov (Saint Petersburg)
Vladislav Shirikov (Dubna)
Yury Smetanin (Moscow)
Vladimir Smirnov (Yaroslavl)
Vladimir Vdovitsyn (Petrozavodsk)
Victor Zakharov (Moscow)

## Coordination of ECDL/RCDL Andreas Rauber, Leonid Kalinichenko

## Conference organizers

Russian Academy of Sciences
Russian Foundation for Basic Research
Karelian Research Centre of the Russian Academy of Sciences
Institute of Applied Mathematical Research
Petrozavodsk State University
Institute of Informatics Problems of the Russian Academy of Sciences
Moscow ACM SIGMOD Chapter

## Organizing committee

**Chairman:** Vladimir Mazalov – Director of the Institute of Applied Mathematical Research (IAMR) of Karelian Research Centre (KarRC) of the Russian Academy of Sciences

**Members:**

Leonid Kalinichenko – Institute of Informatics Problems, RAS
Natalia Lugovaya – IAMR KarRC RAS, scientific secretary of the conference

Andrey Pechnikov – IAMR KarRC RAS
Alexander Rogov – Petrozavodsk State University
Kseniya Rogova – Petrozavodsk State University
Natalia Ruzanova – Petrozavodsk State University
Yury Smetanin – Russian Foundation for Basic Research
Anatoly Sorokin – IAMR KarRC RAS
Vladimir Vdovitsyn – Karelian Research Centre, RAS
Victor Zakharov – Institute of Informatics Problems, RAS

## Program Committee

**Co-chairs:**

Vladimir Vdovitsyn – Karelian Research Centre, RAS, Russia
Mikhail Kogalovsky – Market Economy Institute, RAS, Russia

**Members:**

Arkady Avramenko – Pushchino Radio Astronomy Observatory, RAS, Russia
Alexander Antonov – Galaktika Soft, Moscow, Russia
Yury Bogoyavlensky – Petrozavodsk State University
Jose Borbinha – IST / INESC-ID
Pavel Braslavsky – Institute of Engineering Science, Ural Branch of RAS, Russia
Boris Dobrov – Research Computing Center of Moscow State University, Russia
Alexander Elizarov – Institute of Mathematics & Mechanics, Russia
Anatoly Fedotov – Novosibirsk State University, Russia
Maxim Gubin – Ask.Com IAC Search&Media, USA
Leonid Kalinichenko – Institute of Informatics Problems, RAS, Russia
Hannes Kulovits – Vienna University of Technology
Olga Lavrenova – Russian State Library
Vladimir Litvine – California Institute of Technology, USA
Oleg Malkov – Institute of Astronomy, RAS, Russia
Alexander Marchuk – A.P. Ershov Institute of Informatics Systems, RAS, Russia
Genis Musulmanbekov – Joint Institute for Nuclear Research, Russia
Igor Nekrestjanov – Saint-Petersburg State University, Russia
Dmitry Paley – P.G.Demidov Yaroslavl State University, Russia
Vladimir Pleshko – OOO "ErSiO"
Galina Proskudina – Institute of Software Systems, NAS Ukraine
Andreas Rauber – Vienna University of Technology, Austria
Alexander Rogov – Petrozavodsk State University
Kurt Sandkuhl – Jönköping University, Sweden
Ilya Segalovich – Yandex Ltd, Russia
Vladimir Serebryakov – A.A Dorodnitsyn Computing Center, RAS, Russia
Vladislav Shirikov – Joint Institute for Nuclear Research, Russia
Vladimir Smirnov – Yaroslavl State University, Russia
Ingeborg Torvik Solvberg – Norwegian University of Science and Technology
Alexey Ushakov – Santa Barbara University, California, USA
Vyacheslav Wolfengagen – Moscow Engineering Physical Institute, Russia
Victor Zakharov – Institute of Informatics Problems, RAS, Russia
Oleg Zhizhimov – Institute of Computational Technologies, Siberian Branch of RAS, Russia
Sergey Znamensky – Program System Institute, RAS, Russia

# СОДЕРЖАНИЕ / CONTENTS

## Приглашенный доклад / Invited Paper



## Борьба с плагиатом, электронные библиотеки в аттестации научных кадров / Antiplagiarism Means, Digital Libraries Use for Scientists Attestation



## Онтологическое моделирование-1 / Ontological Modeling-1

## Человеческий фактор / Human Factor



## Инструментальные средства / Digital Libraries Tools



## Диссертационный семинар-2 / PhD Workshop-2

## Приглашенный доклад / Invited Paper



## Извлечение информации из текстов / Extraction of Information from Texts



## Лексика, перевод, синтаксическая разметка текстов / Lexicology, Translation, Syntactic Markup of Texts

## Семантический анализ текстовых коллекций / Semantic Analysis of Text Collections



## Веб-технологии / Web Technologies



## Интеграция информационных ресурсов / Integration of Information Resources

## Онтологическое моделирование-2 / Ontological Modeling-2



## Коллекции научных данных / Collections of Scientific Data

## Мультимедийные коллекции, информационная безопасность / Multimedia Collections, Information Security



## Цифровые архивы / Digital Archives

## Стендовые доклады / Poster Papers

# Предисловие

RCDL'2009 – одиннадцатая в цикле ежегодных российских научных конференций, посвященных проблематике электронных библиотек. На конференциях RCDL обсуждаются различные аспекты разработки, организации и использования электронных библиотек – методологические вопросы, новые технологии и стандарты, инфраструктура таких систем, их контент, предоставляемые ими сервисы, приложения электронных библиотек. Значительное место в тематике конференций занимает обсуждение проблем цифрового сохранения и формирования электронных коллекций информационных ресурсов в различных областях науки, образовании, культуре и других сферах деятельности.

Основная цель цикла конференций RCDL – формирование и поддержка деятельности отечественного профессионального сообщества специалистов России, ведущих исследования и разработки в области электронных библиотек, привлечение молодых ученых и студентов к работе над современными проблемами в этой области, создание форума для неформального общения представителей науки, высшей школы и индустрии. Конференции RCDL открыты для участия в них как российских, так и зарубежных ведущих специалистов, что дает возможность широкого обмена опытом, идеями и полученными результатами, а также установления контактов для более тесного научного сотрудничества. За прошедшие годы в работе конференций приняло участие несколько сотен российских и зарубежных специалистов из Австрии, Германии, Греции, Италии, Новой Зеландии, США, Украины и других стран. Традиционно особое внимание уделяется исследованиям и разработкам в области электронных библиотек, выполняемых в рамках проектов, поддержанных Российским фондом фундаментальных исследований.

Программа RCDL'2009 включает 32 полных, 19 кратких и 3 стендовых доклада, отобранных Программным комитетом в результате рецензирования из 68 поступивших работ. В программу включены два приглашенных доклада: «IT Research Challenges in Digital Preservation» (А. Раубер, Австрия) и «Развитие электронных библиотек – путь к открытой науке» (С.И. Паринов, Россия), а также лекторий (Tutorial) «Demand-oriented Information Supply of Digital Content» (К. Сенкюль, Швеция). В рамках конференции RCDL'2009 впервые организован семинар молодых ученых «Диссертационные исследования по тематике информационных технологий, связанных с электронными библиотеками». Цель семинара – предоставить диссертантам возможность обсудить с более опытными коллегами текущие результаты и идеи их исследований; их слабые и сильные стороны, а также направления дальнейшего развития проводимых исследований, установить контакты для научного сотрудничества, развить навыки выступлений на научных конференциях. Как и в прошлые годы, совместно с RCDL'2009 проводится Всероссийский научный семинар по оценке методов текстового поиска РОМИП. В 2009 году к конференции примыкает также Третья Российская летняя школа по информационному поиску RuSSIR'2009.

Основное внимание в программе конференции RCDL'2009 уделено таким проблемам, как моделирование электронных библиотек, инструментальные средства разработки таких систем, онтологическое моделирование и его использование в электронных библиотеках, семантическая интеграция неоднородных информационных ресурсов, социальные сети и электронные библиотеки, человеческий фактор в электронных библиотеках, различные аспекты обработки текстовых информационных ресурсов и обеспечения доступа к ним, веб-технологии и электронные библиотеки, проблемы создания мультимедийных коллекций информационных ресурсов, реализация конкретных проектов создания цифровых архивов и электронных коллекций.

Руководители Оргкомитета и Программного комитета RCDL'2009 выражают благодарность всем авторам, представившим свои работы для обсуждения на конференции, членам Программного комитета за большую работу по рецензированию и отбору докладов, а также РФФИ и Отделению нанотехнологий и информационных технологий РАН за оказанную ими поддержку конференции. Руководящий комитет конференций RCDL выражает благодарность сотрудникам Института прикладных математических исследований КарНЦ РАН и Петрозаводского государственного университета за организацию и проведение конференции и особо отмечает большой вклад Н.Б. Луговой в разработку и поддержку сайта конференции, а также эффективное выполнение ею обязанностей ученого секретаря Программного комитета.

Председатель Оргкомитета RCDL'2009 – В.В. Мазалов
Сопредседатели Программного комитета – В.Т. Вдовицын, М.Р. Когаловский
Председатель Руководящего комитета конференций RCDL – Л.А. Калиниченко

# Preface

RCDL'2009 is the eleventh conference in the series of annual national research conferences held in Russia as a forum on digital libraries. RCDL participants discuss various aspects of digital library development, organization and utilization – methodologies, new technologies and standards, infrastructure of such systems, their content, services provided, and DL applications. An important issue on the conference agenda is digital preservation and developing digital collections of information resources in different areas of science, education, culture and other spheres.

The main objective of the RCDL cycle is organization and support of activities of the national professional community of specialists involved in DL research and development; encouragement of young scientists and students to work on topical problems in the area; establishment of a forum for informal communication between the science, higher education and industry representatives. RCDL events are open to both Russian and foreign specialists, promoting wide exchange of experience, ideas and results, as well as closer contacts for scientific cooperation. Over the past years, several hundreds specialists from Russia, Austria, Germany, Greece, Italy, New Zealand, Ukraine, the USA, and other countries took part in the conferences. Traditionally, special focus is on DL research and development performed within projects supported by the Russian Foundation for Basic Research.

RCDL'2009 program includes 32 full-scope, 19 brief and 3 poster presentations selected by the Program Committee upon reviewing of 68 papers submitted. The program includes two invited papers: "IT Research Challenges in Digital Preservation" (A. Rauber, Austria) and "Digital Libraries Development as a Way to Open Science" (S. Parinov, Russia), as well as the tutorial "Demand-oriented Information Supply of Digital Content" (K. Sandkuhl, Sweden). For the first time, RCDL'2009 comprises the young scientists' workshop "Doctoral Research on Information Technologies Related with Digital Libraries". The aim of the workshop is to give candidate for a degree the possibility to discuss current results and ideas of their researches, their strengths and weaknesses with more experienced colleagues; to consider the ways to move on; to establish contacts for scientific cooperation; to improve presentation skills. Like in previous years, Russian Information Retrieval Evaluation Seminar – ROMIP is held together with RCDL'2009. Also, adjoining the conference in 2009 is the Third Russian Summer School on Information Retrieval RuSSIR'2009.

Key issues on RCDL'2009 agenda are digital library modeling, tools for development of such systems, ontological modeling and its applications in digital libraries, semantic integration of heterogeneous information resources, social networks and digital libraries, human factor in digital libraries, various aspects of treatment of textual information resources and access to them, web technologies and digital libraries, problems of generation of multimedia collections, implementation of specific projects for creation of digital archives and digital collections.

Heads of RCDL'2009 Organizing and Program Committees thank all authors who have submitted their papers for discussion at the conference, Program Committee members for their great effort in reviewing and selecting papers, as well as Russian Foundation for Basic Research and The Nanotechnologies and IT Division of Russian Academy of Sciences for the support provided to the conference. RCDL Steering Committee highly appreciates the work of the staff of the Institute of Applied Mathematical Research of RAS Karelian Research Centre and Petrozavodsk State University for the conference organization, and especially the great contribution of Natalia Lugovaya to development and maintenance of the conference website, as well as her efficiency as the Program Committee Secretary.

Chair of RCDL'2009 Organizing Committee – V.V. Mazalov
Program Committee Co-chairs – V.T. Vdovitsyn, M.R. Kogalovsky
Chair of RCDL Steering Committee – L.A. Kalinichenko

**ЛЕКТОРИЙ**

**TUTORIAL**

# Demand-oriented Information Supply of Digital Content

© Kurt Sandkuhl

School of Engineering at Jönköping University, Sweden

Information overload has been a phenomenon observed and discussed in the literature since many decades. One of the pioneers of computer-supported collaborative work, Vannevar Bush, foresaw already in 1945 that it would not be possible to manage all information we collect in our "bewildering store of knowledge". However, the attention in the scientific community in this field increased significantly during the last 10 years with the intensification of Internet, e-mail, and information systems use. The problem no longer seems to be that information does not exist electronically, but that it is difficult to find in the huge amount of available data.

The research field information logistics addresses the problem of information overload by developing concepts and technologies for improving information flow in organizations. The core idea is to use principles from material logistics, like just-in-time delivery, in the area of information supply for improved information supply. This is based on demands with respect to content, time of delivery, location, presentation, and quality of information. The scope can be a single person, a target group, a machine/facility, or any kind of networked organization. The aim is to explore, develop, and implement concepts, methods, technologies, and solutions for the above mentioned purpose.

A core subject of demand oriented information supply is how to capture the needs and preferences of a user in order to get a fairly complete picture of the demand in question. Among the different approaches for this purpose are user profiles, situation-based and context-based demand models. *User profiles* are usually created for functionality provided by a specific application. They are based on a predefined structured set of personalization attributes and assigned default values at creation time. The basic idea of the *situation-based* approach is to divide the daily schedule of a person into situations and to determine the optimal situation for transferring a specific message. This approach defines a situation as an activity in a specific time interval including topics and location relevant for the activity. Information value is a relation between a message and a situation, which is based on relevance of the topics of a message for the situation, utility of the message in specific situations and acceptance by the user. The *context-based* is based on the idea that information demand of a person in an enterprise to a large extent depends on the work processes this person is involved in, on the co-workers or superiors and on the products, services or machines the person is responsible for. This led to the proposal to capture the context of information demand, i.e. a formalized representation of the setting in which information demand exists.

Digital content has been subject of research since several decades. An established since the 1980's is the differentiation between logical structure, layout structure and meta-data for describing content, presentation or both. A prominent example for meta-data is the Dublin Core Standard. This principal structure is valid even for contemporary document management applications and is complemented by navigation structure or time-related information. For storing meta-data two principal ways can be distinguished:

- Embedding meta-data in the document, i.e. meta-data are part of the document model,
- Managing meta-data separated from the document in the management systems.

Most contemporary content management solutions use a hybrid approach, as file formats contain a core set of meta-data for specific purposes, which are complemented with additional information on system side.

When trying to identify digital content that fits to a given demand model, matching between demand model and meta-data is considered an important element. Matching is often characterized as "searching with imprecise specification of information needs". In this context we have to discuss the aspect of relevance of information, e.g. algorithmic, topical and cognitive relevance. The underlying concepts for algorithmic relevance, i.e. the relation between the query features and the search result, and for topical relevance, i.e. relation between aboutness of content objects and query, have to be observed when implementing matching systems. Semantic matching based on ontologies in combination with information retrieval techniques for string matching are promising approaches for information logistics in order to match information demand and content.

.

**Proceedings of the 11th All-Russian Research Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» - RCDL'2009, Petrozavodsk, Russia, 2009.**

# МОДЕЛИ ЭЛЕКТРОННЫХ БИБЛИОТЕК

# MODELS OF DIGITAL LIBRARIES

# Концептуальная модель электронной библиотеки

© Резниченко В.А., Проскудина Г.Ю., Кудим К.А.
Институт программных систем НАН Украины, г. Киев 03187, пр. Академика Глушкова, 40
reznich@isofts.kiev.ua, gupros@isofts.kiev.ua, kuzma@isofts.kiev.ua

**Аннотация**

Работа посвящена задаче создания концептуальной модели электронной библиотеки. Обсуждаются некоторые известные связанные проекты – CIDOC CRM, FRBR, DELOS DLRM. Предложен оригинальный вариант информационной составляющей концептуальной модели.

"Пора подумать, – Морж сказал, –
О множестве вещей:"

Л.Кэрролл "Алиса в Зазеркалье"

## 1 Введение

Появление новых электронных библиотек (ЭБ), увеличение числа хранимых в них документов и повышение качества предоставляемых ими услуг способствует развитию науки, облегчая, а иногда и просто открывая единственно возможный доступ к источникам информации для ученого, предоставляя ему замечательное средство донести плоды своей деятельности до широчайшей аудитории. В последние несколько лет при нашем непосредственном участии научное сообщество Украины продвинулось в этом направлении. В частности, в прошлом году создан портал периодических изданий НАН Украины NASPLIB[1]. Два года назад создана ЭБ Института программных систем НАН Украины ISS EPrints[2]. В первом случае использовалось программное обеспечение DSpace, во втором – EPrints. Обе системы были полностью украинизированы. Были отработаны основные сценарии использования, создан ряд методик и рекомендаций по созданию и использованию электронных библиотек на основе данных программных систем. Были также изучены программные продукты Greenstone[3] и Fedora[4]. Этот опыт оказался очень ценным для понимания современного состояния дел в мире программных систем ЭБ.

В настоящее время нет какой-либо универсальной ЭБ, которая отвечала бы всем требованиям и ожиданиям пользователей. Анализ существующих систем ЭБ [1-3] показывает их разнородность на нескольких уровнях:

- на уровне информационной модели, которую они обеспечивают;
- на уровне поддержки пользователей и групп пользователей;
- на уровне функциональных возможностей.

Из-за этой гетерогенности ЭБ и игнорирования нужд их пользователей возникает ряд проблем:

- интеграция информации из различных ЭБ;
- сравнение ЭБ по предоставляемой функциональности;
- оценка и сравнение производительности различных систем ЭБ;
- добавление новых типов хранимых объектов;
- добавление новых функциональных возможностей;
- резервное копирование.

Решить эти и другие возникающие проблемы на первом этапе поможет аккуратное и полное рассмотрение области ЭБ. Именно для этого создаются концептуальные модели, обобщающие накопленный опыт в сфере создания и использования ЭБ.

В последнее время в мире предпринимаются усилия по полному и всестороннему описанию сферы ЭБ. Во втором разделе мы обсуждаем следующие известные модели и стандарты, которые могут применяться для описания ЭБ в целом и ее частей: CIDOC CRM [4], FRBR [5], DELOS DLRM [6].

Третий раздел посвящен описанию информационной составляющей разрабатываемой нами концептуальной модели ЭБ. Мы рассматриваем ЭБ как информационную систему, поэтому в дальнейшем планируется дополнить описание модели пользовательской и функциональной составляющими, которые не затронуты в данной работе.

## 2 Обзор известных проектов

Нацелившись на описание информационного пространства библиотечной системы, мы естественно не могли не рассмотреть что-то подобное, что уже сделано или делается сегодня в мире.

### 2.1 CIDOC CRM

Концептуальная эталонная модель (Conceptual Reference Model, CRM) CIDOC, разработанная Международным комитетом по документации Международного совета музеев (The International Committee

**Труды 11[ой] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

for Documentation of the International Council of Museums, ICOM-CIDOC), предназначена для интеграции, посредничества и обмена информацией в области мирового культурного наследия и связанных областей.

Представляя определения и формальную структуру описания неявных (implicit) и явных (explicit) сущностей и отношений, модель претендует на общий язык для экспертов и разработчиков, формулирующих требования к информационным системам, способствует общему пониманию информации, обеспечивает интеграцию, посредничество и обмен информацией между музеями, библиотеками, архивами [7, 8].

До 1994 года разрабатывалась ER-модель для музейной информации, начиная с 1996 года подход разработки модели сместился к методологиям объектно-ориентированного моделирования и привел в 1999 году к появлению первой Концептуальной эталонной модели CIDOC. С 2000 года начался процесс стандартизации, который успешно завершился принятием стандарта ISO 21127:2006 – "Эталонная онтология для обмена информацией культурного наследия" (A reference ontology for the interchange of cultural heritage information).

Разработчики CIDOC CRM поставили своей целью написать стандарт модели, пригодной как для машинной обработки, так и для легкого понимания человеком. Модель совместима с формализмом RDF.

Версия 4.2.4 модели CIDOC CRM [4] состоит из 87 классов и 148 свойств, описывающих предметы, понятия, людей, события, место, время и их отношения. CIDOC CRM предлагает только высокоуровневые понятия, описывающие сущности и их связи, и никак не связана с документированием либо реализацией таких систем.

На рис. 1 представлена часть иерархии классов CIDOC CRM. Все классы, за исключением класса *Простое значение* (*Primitive Value*) и его подклассов, прямо или опосредовано являются подклассами класса *E1 Сущность CRM*, охватывающего все сущности, которые могут быть описаны в CIDOC CRM.

Модель может быть расширена добавлением необходимых для конкретной задачи сущностей в иерархию классов.

Резюмируя рассмотрение данной модели, отметим, что для наших целей это – нужный и полезный стандарт. Важным преимуществом стандарта является его формальный подход. Он вполне может служить основой для информационной составляющей концептуальной модели ЭБ. Конечно, стандарт нуждается в расширении более конкретными сущностями, которые часто используются во многих ЭБ. Кроме того, CIDOC CRM не охватывает пользовательского и функционального аспектов ЭБ.

### 2.2 FRBR и FRBRoo

Независимо от CIDOC CRM в 1991-1997 годах Международной федерацией библиотечных ассоциаций и учреждений (International Federation of Library Associations and Institutions, IFLA) была разработана ER-модель "Функциональные требования к библиографическим записям" (Functional Requirements for Bibliographic Records, FRBR) как обобщенное представление библиографического универсума, независимого от какого-либо кода каталогизации или реализации.

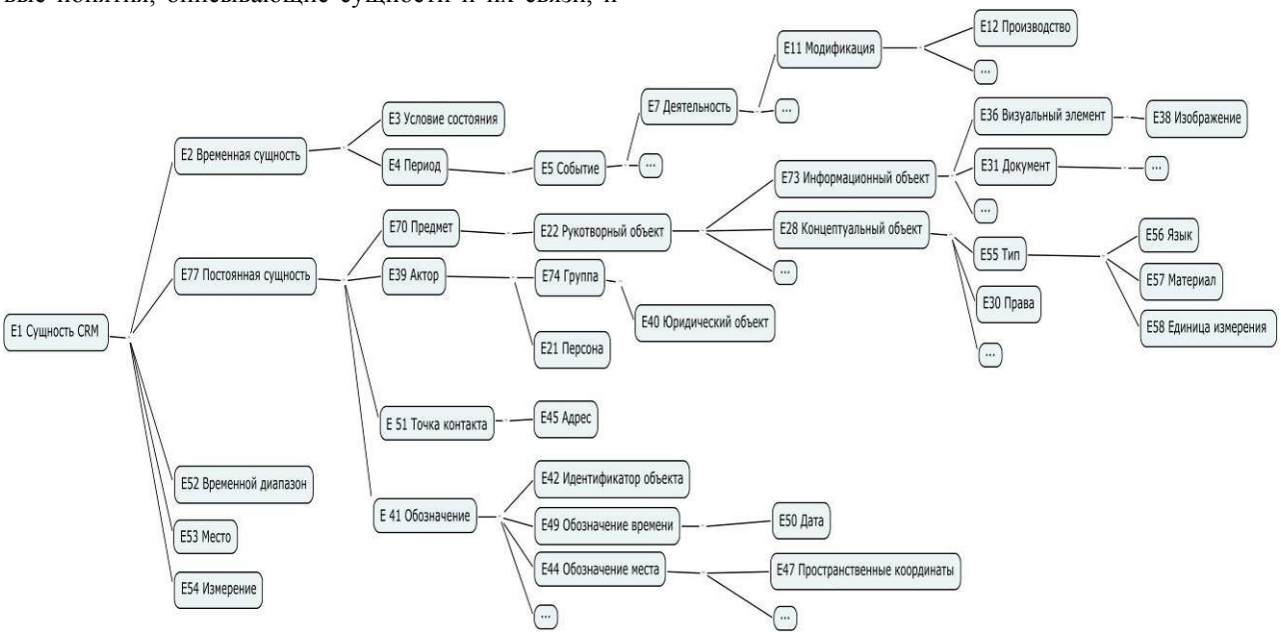



Рис. 1. Часть иерархии классов в модели CIDOC CRM

В 1998 году модель была опубликована [5]. В настоящее время IFLA продолжает контролировать приложения модели FRBR и поддерживает ее использование и развитие.

FRBR включает описание концептуальной модели (сущности, их отношения и атрибуты), предлагает универсальные библиографические записи для всех типов материалов и пользовательских задач, связанных с библиографическими ресурсами, описанными в каталогах, библиографиях и других библиографических инструментах [9, 10].

Модель FRBR различает три группы сущностей (рис. 2):

- для описываемых объектов: *произведение (work)*, *выражение (expression)*, *воплощение (manifestation)*, *экземпляр (item)*;
- для описателей-субъектов: *человек (person)* и *организация (corporate body)*;
- для описателей-объектов: *концепт, объект, событие* и *место (concept, object, event, place)*.

Ниже приведен пример [11] экземпляров сущностей *произведения* (w1) и его *выражений* (e1-e2):

w1 *Tennis--bis zum Turnierspieler* Элльвангера

- e1 оригинальный текст на немецком языке
- e2 перевод на английский язык Венди Джилл
- ...

Большое внимание в модели уделено отношениям между сущностями.

Отношения могут быть отражены в библиографических записях многими способами. Те, что изображены на ER-диаграмме FRBR (рис. 2), описывают логические связи между сущностями и часто реализуются простой конкатенацией одной сущности с атрибутами связанной сущности в одной записи.

Помимо логических связей в модели выделена группа так называемых *контентных* связей (для первой группы сущностей). Они идентифицируют основные типы отношений, которые существуют между экземплярами сущности одного типа (например, сущности *произведения*) или между экземплярами разных типов сущностей (например, экземпляров сущностей *произведение* и *воплощение*). Например, в группе отношений *произведение-произведение* выделены такие типы отношений: *имеет адаптацию* (свободный перевод); *имеет приложение* (сходство, соответствие), *имеет продолжение*; *имеет резюме* (обзор, аннотацию); *имеет преобразование (*стихотворную форму*)*; *имеет имитацию* (пародию). В группе отношений *выражение-выражение* перечислены следующие типы отношений: *имеет сокращение* (корректировку, уплотнение); *имеет пересмо*тр (исправленную редакцию, расширенную редакцию); *имеет перевод* (буквальный перевод) и некоторые другие типы отношений, касающиеся музыкальных произведений.

И наконец, отношения *часть/целое* и *часть в части* также представлены в модели FRBR.

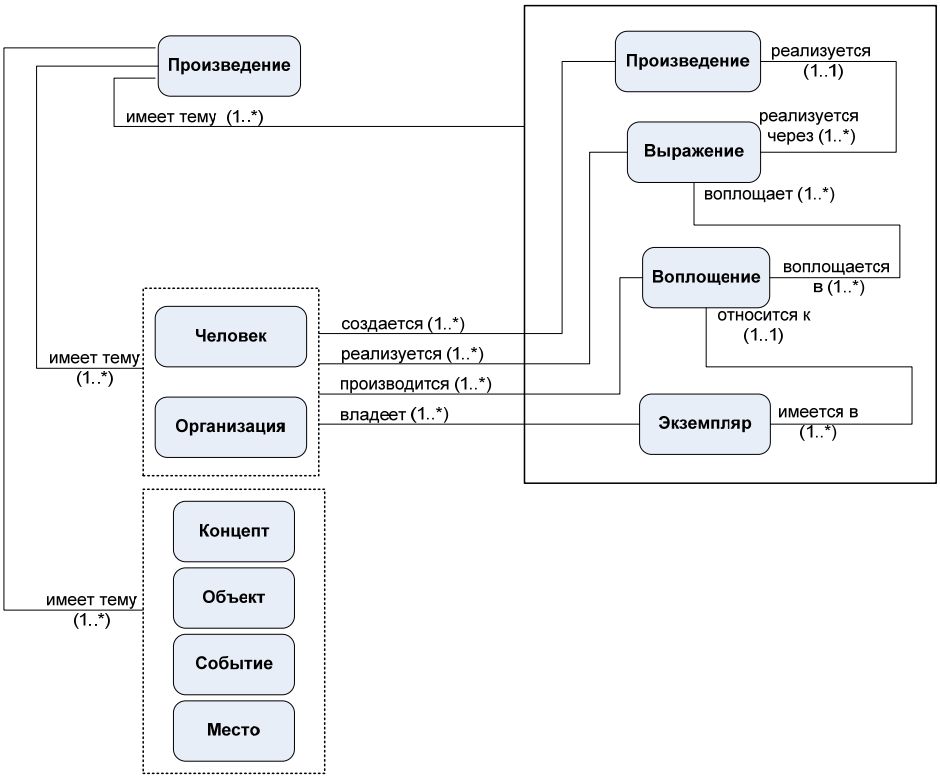



Рис. 2. Модель FRBR

В 2006 году был опубликован первый полный проект модели FRBRoo, т.е. объектно-ориентированной версии FRBR, согласованной с CIDOC CRM. Эта формальная модель предназначена для представления основной семантики библиографической информации, интеграции и обмена библиографической и музейной информацией.

Главное новшество FRBRoo – реалистичная, явная модель процесса интеллектуального творчества, которая еще должна получить свое дальнейшее развитие для библиотекарей и ученых [12].

Следует отметить, что в FRBR границы между различными типами основных сущностей (произведение и выражение) размыты и окончательное решение по тому, к какому типу отнести тот или иной объект, отдается на откуп каталогизатору. Кроме того, сущностей этих совсем немного и явно недостаточно для большинства конкретных библиотечных приложений. Для нас основной интерес представляет очень богатый набор атрибутов и отношений в этой модели. Как и в случае CIDOC CRM, в соответствии с решаемыми задачами, FRBR применима только для описания информационной составляющей концептуальной модели.

### 2.3 DELOS DLRM

Группа специалистов ассоциации в сфере ЭБ DELOS в 2006-7 гг., основываясь на анализе имеющихся библиотечных систем [3], где большое внимание было уделено функциональным возможностям современных ЭБ, начали разработку эталонной модели ЭБ (Digital Library Reference Model, DLRM) [6]. Цель проекта – разобраться с фундаментальными понятиями, существенными объектами и их отношениями, стандартными функциональными и структурными блоками и процессами, из которых состоит универсум ЭБ. Эталонная модель предназначена для разработки более узких моделей с конкретной архитектурой для последующей реализации программных систем.

Прежде всего, в модели было выделено три понятия для разграничения того, что обычно называется ЭБ:

- *ЭБ* – конкретная ЭБ с ее пользователями, правилами, содержимым, интернет-сайтом и ведущей организацией. Например: библиотека института программных систем ISS EPrints http://eprints.isofts.kiev.ua;
- *система ЭБ* – программное обеспечение, на основе которого создаются ЭБ. Например: EPrints 3.0.
- *система управления ЭБ* – программное обеспечение для создания и управления системами ЭБ. Например: система OpenDLib[5].

Далее модель DELOS DLRM рассматривается в ролевом аспекте, т.е. с точки зрения разных категорий пользователей:

- конечный пользователь ЭБ;
- разработчик ЭБ;
- системный администратор ЭБ;
- разработчик приложений для ЭБ.

Соответственно DELOS DLRM имеет четыре уровня пользовательских представлений.

Весь универсум ЭБ разбит на шесть высокоуровневых ключевых областей (рис. 3):

- контент;
- пользователь;
- функциональные возможности;
- качество;
- политики;
- архитектура;

и несколько дополнительных. Эти шесть областей объединены в одну область ресурса. В каждой из них вводятся и определяются свои сущности и их свойства.

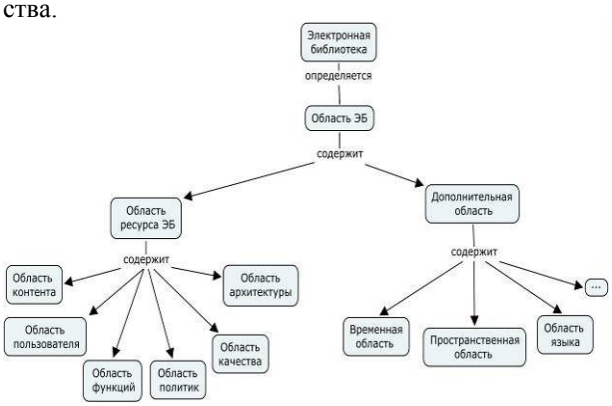



Рис. 3. Иерархия областей ЭБ в модели DELOS DLRM

Теперь вкратце рассмотрим наиболее важные области ЭБ и их структуру.

**Область ресурса ЭБ** – наиболее общая область в данной модели, представляет все сущности и связи, населяющие универсум ЭБ. *Ресурс* – наиболее общее понятие, включающее любую сущность ЭБ. По аналогии с ресурсом в Веб, ресурс – это все то, что может быть идентифицировано, названо или адресовано. Представленная здесь модель ресурса исходит из веб-архитектуры, но дополнена некоторыми аспектами, специфичными для предметной области ЭБ.

Ресурс – абстрактное понятие, в том смысле, что непосредсвенно не имеет экземпляров, он только выражен экземплярами одной из своих специализаций. В частности, экземплярами понятия ресурс в универсуме ЭБ являются экземпляры *информационного объекта* любого типа (например, документы, изображения, видео, мультимедийные объекты, наборы метаданных и аннотаций, потоки, базы даных, коллекции, запросы и результаты запросов), *акторы* (как одушевленные так и неодушевленные сущности), *функции*, *политики*, *параметры качества* ЭБ и *архитектурные компоненты*. Каждый из этих экземпляров представляет главное понятие в своей области, т.о. в представленной модели ЭБ каждая область состоит из ресурсов, а ресурсы – строительные блоки всех областей ЭБ. Каждый ресурс:

- имеет идентификатор;

– организован в соответствии с *форматом ресурса*. Формат здесь выражен *онтологией*. Ресурс может быть сложным и структурированным, поскольку он, в свою очередь, может состоять из меньших ресурсов и иметь связи с другими ресурсами;
– может характеризоваться параметрами качества;
– может регулироваться политиками, управляющими его жизненным циклом;
– выражается через информационный объект;
– может быть описан или дополнен информационным объектом, обычно – метаданными и аннотациями.

С организационной точки зрения, ресурсы могут группироваться в наборы ресурсов, которые рассматриваются как единая сущность. Например, *коллекции* в области контента или *группы* в области пользователя.

**Область контента** (рис. 4) представляет все объекты, связанные с информацией, которой управляет ЭБ. *Информационный объект* – наиболее общее понятие в этой области, представляет произвольную единицу информации, управляемую в универсуме ЭБ. В DELOS DLRM различают *информационный объект по уровню абстракции,* где заимствуются типы объектов из модели FRBR (произведение, выражение, воплощение) и *информационный объект по связи* – "абстрактный концептуальный контейнер для классов, которые порождают эти объекты", а именно:

– *первичный информационный объект* – информационный объект, который используется самостоятельно, например, текстовые документы, изображения;
– объект *метаданные* – информационный объект, главная цель которого состоит в том, чтобы дать информацию о целевом ресурсе (как правило о первичном информационном объекте);
– объект *аннотация* – информационный объект, главная цель которого состоит в том, чтобы аннотировать целевой ресурс или его часть (на рис. 4 части ресурса соответствует сущность *регион*). Примеры таких аннотаций включают примечания, структурированные комментарии и связи. Объекты аннотации помогают интерпретировать целевой ресурс, содержат либо поддержку, либо возражения, либо более детальные объяснения.

Поскольку информационный объект является ресурсом, то он наследует все вышеперечисленные свойства ресурса.

Информационные объекты также могут быть сложными объектами и могут быть сгруппированы в *коллекции* информационных объектов. Коллекции, в свою очередь, тоже являются информационными объектами, они наследуют все аспекты моделирования информационных объектов и средства их обслуживания, например они могут аннотироваться. Кроме того, коллекция – специализация понятия *набора ресурсов*. Коллекции определяются *критерием отбора* (hasIntension) либо *перечислением элементов* (hasExtension). Другая специализация понятия набор ресурсов в данной модели – *результирующий набор*. В традиционных ЭБ он представляет собой набор документов, которые извлекаются в ответ на *запрос*.

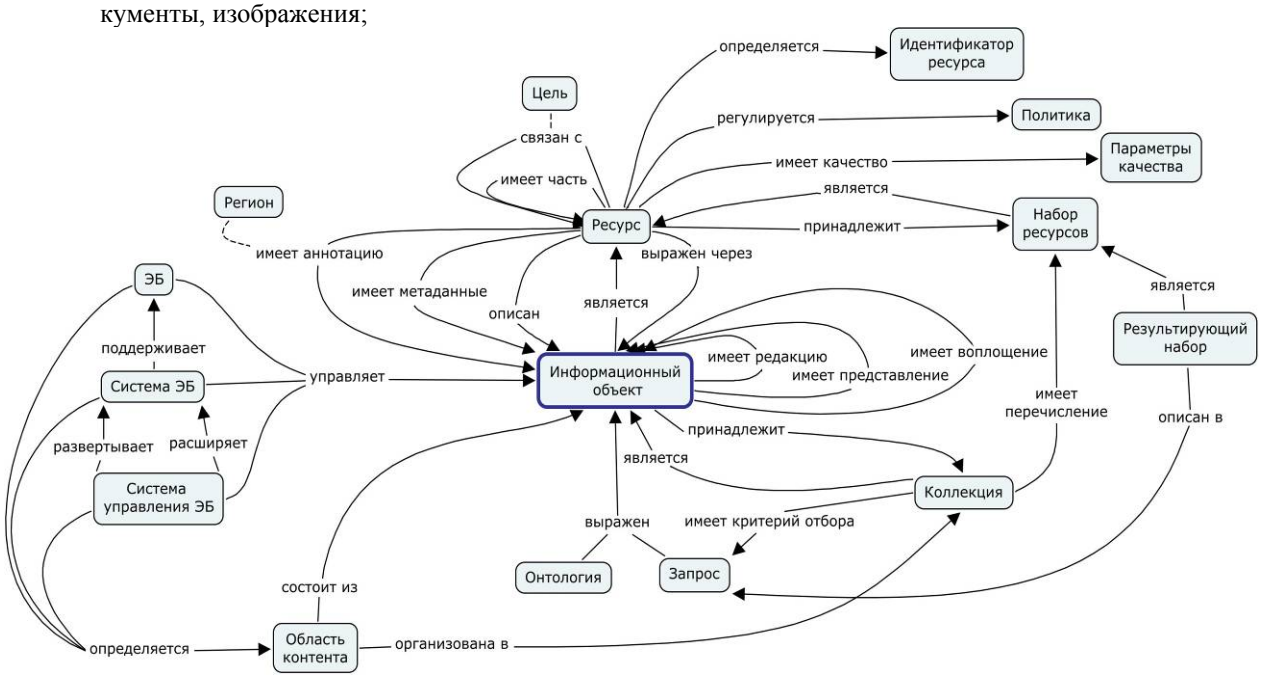



Рис. 4. Область контента ЭБ в модели DELOS DLRM

**Область пользователя** в DELOS DLRM содержит все объекты, которые являются "внешними по отношению к системе ЭБ и с ней взаимодействуют:

люди и неодушевленные объекты, например, программы или физические инструменты… или даже другая ЭБ может быть среди пользователей ЭБ".

Поскольку главная сущность в этой области – *актор* является ресурсом и следовательно, наследует все его свойства, а именно:

- имеет уникальный идентификатор (идентификатор пользователя);
- организован в соответствие с форматом (модель пользователя);
- благодаря свойствам ресурса композиции и соединения может быть составлен в различные сложные и структурированные группы например, сотрудничество пользователей или соавторов;
- описан или дополнен метаданными и аннотациями.

**Область функций** представляет наиболее объемную и наиболее открытую часть модели DELOS DLRM, поскольку охватывает всю обработку ресурсов, а также действия пользователей в ЭБ. Здесь наиболее общим понятием является сущность *функция*.

Функция – специфическая задача обработки, которая может быть реализована на наборе ресурсов или одном ресурсе в результате действий отдельного пользователя. Описание функций основано на пользовательском аспекте и ресурсе, представляющем все объекты, вовлеченные в ЭБ. Хотя функции в традиционных моделях ЭБ обычно связываются с контентом в ЭБ и выполняются людьми, здесь, в данной модели, функции могут выполняться неодушевленными пользователями на любом типе ресурсов.

В данной модели ЭБ каждая функция также является ресурсом и потому наследует все его характеристики.

Функции разделены на пять классов:

- доступа к ресурсам;
- управления ресурсами;
- совместной работы;
- управления ЭБ;
- настройки ЭБ.

Подводя итог, нужно признать, что именно модель DELOS DLRM вдохновила изначально нашу работу. Внимательное изучение данной модели помогло не только обозреть всю сферу ЭБ, но и найти некоторые пробелы в самой модели. Вот некоторые из них:

- недостаточно формализованные определения, оставляющие размытыми границы многих сущностей (например, сущности, заимствованные из FRBR, или граница между *метаданными* и *аннотацией*);
- в некоторых местах остаются не ясными критерии выделения сущностей (в частности, область качества наименее убедительна в этом отношении);
- неоднородность описания различных областей ЭБ, скрытая за внешне однообразным описанием (достаточно сравнить простую иерархию области функций со сложной структурой области контента).

К преимуществам DELOS DLRM следует отнести наибольшую полноту охвата среди существующих концептуальных моделей ЭБ.

## 3 Информационная модель ЭБ

Концептуальная модель должна описывать то, какие сущности могут существовать в данной предметной области (для нас – области электронных библиотек, а точнее – электронных научных библиотек), т.е. существуют в данный момент, существовали ранее или когда-либо смогут существовать. А также она должна фиксировать их правила, связи, что в частности предполагает классификацию сущностей, абстрагирование, обобщение.

Основываясь на рассмотренных выше моделях, учитывая их достоинства и недостатки, мы попытались построить свою модель ЭБ, начав описание с ее информационной составляющей.

### 3.1 Сущности

На рис. 5 изображена иерархия сущностей или объектов, представленных в электронной научной библиотеке.

*Физический объект* – корневой объект в представляемой модели, он охватывает все объекты, информация о которых хранится в электронной библиотеке.

Физический объект, как и все другие объекты, обладает *атрибутами*. Набор атрибутов объекта зависит от его типа. Так физический объект имеет следующие атрибуты:

- идентификатор физического объекта;
- название;
- тема;
- ключевые слова;
- версия;
- аннотация.

Эти атрибуты наследуются всеми другими объектами представленной иерархии (рис. 5).

Как правило, в системах ЭБ предусматривается хранение *рукотворных объектов* – основного типа объектов информационного контента, а также некоторых других объектов, имеющих к ним отношение:

- *организации, отделы* организаций и *издательства*, где создавались или публиковались рукотворные объекты;
- люди (на схеме это сущность *человек*), работающие в этих организациях (отделах) – авторы рукотворных объектов;
- *проекты* в рамках которых создаются рукотворные объекты;
- научные *журналы* (периодические издания) и *конференции* их публикующие.

Объект *коллекция* может быть применим к любой совокупности (группировке, агрегации) физических объектов. Физические объекты здесь могут быть лю-

бого типа, т.е. коллекциями могут быть как совокупности физических объектов, так и рукотворных объектов, совокупности организаций, журналов и т.д. Критерии для таких совокупностей могут определяться, например, общностью местоположения, общностью авторов, хронологией, тематикой, происхождением или принадлежностью и т.д. [1]. Коллекции могут содержать любое число объектов и критерии отбора этих объектов со временем могут изменяться.

*Организации,* как правило, представляют научно-исследовательские институты или образовательные организации. Помимо наследуемых атрибутов этот класс имеет также следующие атрибуты и связи:

- тип организации;
- дата основания;
- местонахождение (страна, город);
- вышестоящая организация;
- руководитель;
- подразделение;
- адрес (почтовый, юридический, сайт, e-mail, телефон).

Класс *человек,* наряду с наследуемыми, имеет также и собственные свойства:

- место работы (организация, отдел);
- пол;
- дата рождения;
- место рождения (страна, регион, город);
- адрес (почтовый домашний, личного сайта, e-mail, телефон);
- ученое звание, ученая степень;
- специальность ВАК (Высшая аттестационная комиссия);
- соавторство.

Приведем также перечень атрибутов для класса *проект*:

- название программы;
- название конкурса;
- период выполнения;
- организация (где выполняется проект);
- руководитель;
- спонсор;
- бюджет.

*Журнал* и *конференция* – объекты, связанные с публикацией (а значит и с производством) главного вида научной продукции – статьи, одного из представителей класса рукотворных объектов.

Перечень возможных атрибутов для класса *журнал*:

- ISSN (Международный стандартный серийный номер);
- издатель,

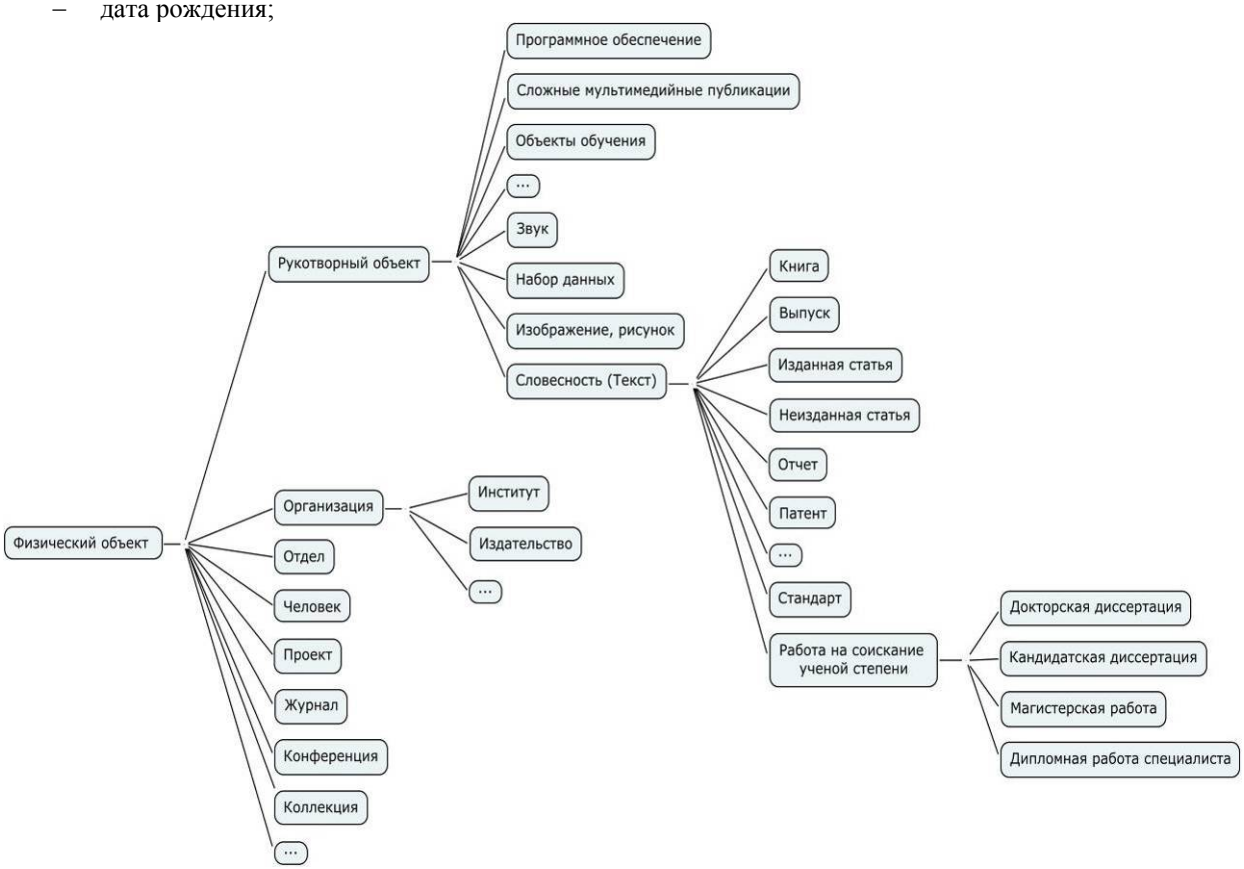



Рис. 5. Иерархия классов в информационной модели научной электронной библиотеки

а также для класса *конференция*:

- дата и место проведения;
- программный комитет;
- ответственная организация.

*Рукотворный объект* – класс существующих в библиотеке объектов, созданных в процессе научной деятельности людей. К свойствам физического объекта здесь добавляется новое отношение *имеет автора*.

Рукотворные объекты, в свою очередь, подразделяются на:

- текстовые объекты или *словесность;*
- графические объекты (*изображение, рисунок*);
- аудио объекты или *звук;*
- *сложные мультимедийные публикации*;
- *объекты обучения*;
- *наборы данны;*
- *программное обеспечение*.

Объекты *словесность* (текстовые объекты) в представляемой модели используется для обозначения любого электронного текстового контента различных типов – *книги; выпуски* научных периодических журналов; опубликованные в них статьи (*изданная статья*); еще *неизданные* публикации; различные научные *отчеты*; документы, вошедшие в разряд принятых *стандартов*; *патенты*, а также работы, представленные на соискание ученой степени (докторская, кандидатская диссертация, магистерская и дипломная работа). На схеме (рис. 5) для целью простоты и наглядности не показаны такие типы объектов словесности как инструкции, методические материалы, тексты и презентации докладов и выступлений научных конференций, симпозиумов, семинаров, школ[5].

Всем текстовым объектам, помимо атрибутов, что наследуются из вышестоящих объектов, присущи свойства:

- язык контента;
- количество страниц (или диапазон) в опубликованной версии.

Эти объекты могут также иметь такие атрибуты как:

- содержание;
- набор файлов, когда объекты данного типа располагают полным текстом, и он хранится в файлах,

а также связи с аналогичными объектами:

- является переводом;
- имеет перевод;
- является версией;
- имеет версию;
- цитируется;
- цитирует.

Этот перечень связей может быть существенно дополнен связями, рассмотренными в модели FRBR.

Каждый объект, имеющий тип *словесность*, обладает своими присущими только ему атрибутами или связями. Например, *книга*, помимо перечисленных атрибутов для вышестоящих в иерархии физического, рукотворного объекта и объекта словесность также имеет:

- ISBN (Международный стандартный номер книги);
- издатель;
- издание;
- место и дата публикации;
- автор предисловия (послесловия);
- редактор;
- автор перевода (если она переводная).

Изданная статья помимо общих атрибутов имеет: обязательный атрибут-связь *выпуск*, связывающий экземпляры класса *изданная статья* с соответствующим экземпляром класса *выпуск*. Объект *выпуск* (имеется ввиду журнала), который мы также отнесли к классу *словесность,* дополнительно имеет атрибуты *дата*, *номер*, может иметь *том* и *тему выпуска*, а также связующие свойства *журнал* и *изданная статья*.

Обсуждая иерархию объектов, нужно также перечислить классификаторы, используемые при задании некоторых их атрибутов. В модели CIDOC CRM эта категория сущностей выделена в *концептуальный объект* (рис. 1). Так, например, атрибуты *тема* и *ключевые слова* как правило задаются с помощью распространенных тематических или предметных классификаторов: УДК, ББК, тематического классификатора ВАК и некоторых других: DDC, LCC, LCSH, MESH. Атрибут *язык* желательно определять в соответствии со стандартами RFC 1766 (ISO 639-2, ISO 3166); *географическое положение* – в стандарте GEO; форматы файлов задавать контролируемым словарем MIME; при описании сущности *человек* использовать набирающий все большую популярность FOAF.

### 3.2 Связи

Выше в описании иерархии объектов и их свойств уже упоминались некоторые связи, важной составляющей концептуальной модели и любой ЭБ. Объекты информационного пространства ЭБ связаны между собой бинарными ориентированными обязательными связями вида "1..1" (один к одному) или "1..*" (один ко многим) и необязательными связями вида "0..1" или "0..*". Примеры таких связей показаны на рис. 6.

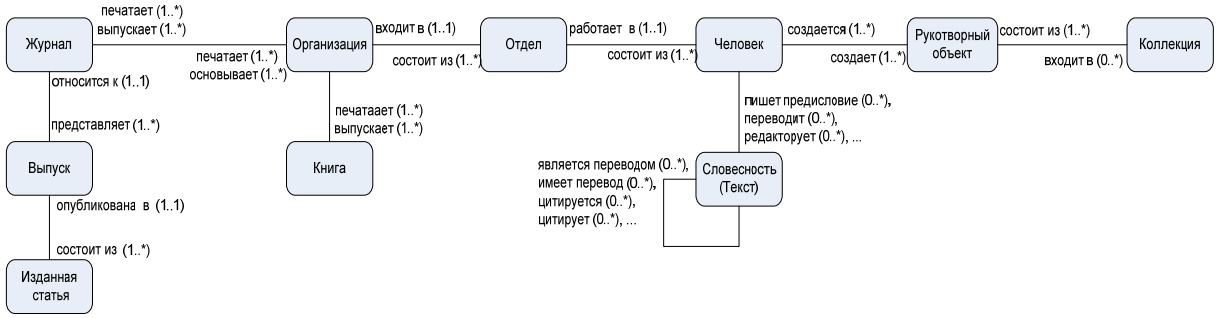



Рис. 6. Связи между сущностями в информационной модели научной электронной библиотеки

## Заключение

В мире электронных библиотек существуют задачи, которые могут быть решены с помощью качественного концептуального описания таких систем. На сегодняшний день не существует всеохватывающей модели ЭБ, которую можно было бы с полным правом назвать эталонной. Построение хорошей модели, учитывающей мировой опыт подобных разработок, является нашей целью.

В данной работе дан краткий обзор трех известных проектов из этой области: CIDOC CRM, FRBR, DELOS DLRM, а также рассмотрена информационная составляющая разрабатываемой нами концептуальной модели ЭБ по состоянию на текущий момент. Пользовательская и функциональная составляющие модели ЭБ находятся в зачаточной стадии разработки и пока что не представлены. Авторы надеются на интересное и полезное обсуждение.

## Литература


1. *Когаловский М.Р., Паринов С.И.* Информационные ресурсы, наукометрические показатели и показатели качества метаданных системы Соционет // Труды Девятой Всероссийской конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" — RCDL'2007, г. Переславль-Залесский, Россия, 15-18 октября 2007 г. — С.45-54.
2. *A Guide* to Institutional Repository Software. 3rd Edition. Open Society Institute. 2004. http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_v3.pdf
3. *Candela L., Castelli D., Fuhr N., Ioannidis Y., Klas C.-P., Pagano P., Ross S., Saidis C., Schek H.-J., Schuldt H., Springmann M.* Current Digital Library Systems: User Requirements vs Provided Functionality. IST-2002-2.3.1.12. Technology-enhanced Learning and Access to Cultural Heritage. March 2006.
4. *Crofts N., Doerr M., Gill T., Stead S., Stiff M. (editors),* Definition of the CIDOC Conceptual Reference Model, January 2008. Version 4.2.4
5. *Functional Requirements for Bibliographic Records, Final Report* / IFLA Study Group on the Functional Requirements for Bibliographic Records. – München: K.G. Saur, 1998. (UBCIM Publications, New Series; v. 19) http://archive.ifla.org/VII/s13/frbr/frbr.htm
6. *Candela L., Castelli D., Dobreva M., Ferro N., Ioannidis Y., Katifori H., Koutrika G., Meghini C., Pagano P., Ross S., Agosti M., Schuldt H., Soergel D.* The DELOS Digital Library Reference Model Foundations for Digital Libraries. IST-2002-2.3.1.12. Technology-enhanced Learning and Access to Cultural Heritage. Version 0.98, December 2007. http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf
7. *The CIDOC Conceptual Reference Model. http://cidoc.ics.forth.gr*
8. *Doerr M., Iorizzo D.* The Dream of a Global Knowledge Network — A New Approach // ACM Journal on Computing and Cultural Heritage, Vol. 1, No. 1, Article 5, Publication date: June 2008.
9. *Tillett*, *Barbara B.* "Bibliographic Relationships." In: *Relationships in the Organization of Knowledge*, edited by Carol A. Bean and Rebecca Green. — Dordrecht : Kluwer Academic Publishers, 2001, p. 19-35.
10. *Barbara Tillett* What is FRBR? A Conceptual Model for the Bibliographic Universe. http://www.loc.gov/cds/downloads/FRBR.PDF
11. *Функциональные* требования к библиографическим записям : окончат. отчет / Рос. библ. ассоц., Рос. гос. б-ка ; пер. с англ. [В.В. Арефьев ; науч. ред. пер.: Т.А. Бахтурина, Н.Н. Каспарова, Н.Ю. Кулыгина]. — Москва : РГБ, 2006. — [150] с.
12. *Doerr M., Leboeuf P.* Modelling intellectual processes: The FRBR—CRM harmonization // In Conference Proceedings of ICOM-CIDOC Annual Meeting. Gothenburg, Sweden. — 2006. — pp. 10–14.


## Conceptual Model of Digital Library

Reznichenko V.A., Proskudina G.Yu, Kudim K. A.

The paper is concerned with creation of scientific digital library conceptual model. Several well-known connected projects are surveyed, those are CIDOC CRM, FRBR, DELOS DLRM. Informational part of a conceptual model being developed is presented.

[1] http://dspace.nbuv.gov.ua:8080/dspace
[2] http://eprints.isofts.kiev.ua/
[3] http://www.greenstone.org/
[4] http://www.fedora-commons.org/
[5] http://www.opendlib.com
[6] Подробный перечень типов документов представлен, например, в одной из вспомогательных таблиц «Формы документов» классификатора УДК.

# Логическая модель цифровых библиотек в онтологии ЕНИП

© А.А. Захаров В.И. Филиппов

Вычислительный Центр РАН
vicoff@yandex.ru, andreya@sufler.ru

**Аннотация**

В работе представлены концепции Единого Научного Информационного Пространства РАН (ЕНИП РАН), обеспечивающие поддержку реализации системы управления электронной библиотекой (СУЭБ). Описываются ключевые концепции, такие как документоподобные объекты, медиа-представления, коллекции и расширенные атрибуты.

## 1 Введение

В настоящее время научно-исследовательский процесс неотделим от использования Интернета. Значительную часть своего времени научные сотрудники проводят за компьютерами в поиске и анализе информации, в электронной переписке с коллегами во всем мире. В частности, все большую роль в этом процессе начинает играть использование электронных библиотек.

Электронная (цифровая) библиотека — структурированная коллекция разнородных электронных документов (в отличие от печатных изданий, микрофильмов и других носителей), снабженных средствами навигации и поиска и доступных через компьютеры. Как правило, это Web-сайт, где накапливаются различные тексты (чаще литературные, научные и технические, но также и любые другие, вплоть до компьютерных программ) и медиа-файлы, каждый из которых самодостаточен и может быть востребован пользователем.

Электронные библиотеки следует отличать от смежных структурных типов сайта, в частности литературного и свободных публикаций. В отличие от литературного журнала, электронная библиотека не подразделяется на выпуски и обновляется перманентно по мере появления новых материалов. В отличие от сайта со свободной публикацией, электронная библиотека, как правило, подбирается координатором проекта по определенным правилам и не всегда предусматривает создания вокруг публикуемых текстов коммуникативной среды.

Не следует представлять себе, что электронная библиотека появляется простым выставлением в Интернете внутрибиблиотечной информационно-поисковой системы, осуществляющей поиск в каталоге и поддержку работы с единицей хранения. Электронная библиотека способна не только обеспечить многосторонний поиск в каталоге, но и предоставить пользователю непосредственно найденный текст (или другой ресурс), а также дополнительные сведения о его контексте: авторах, библиографии, издательстве и т.п. В связи с этим и специалисты в области библиотечного дела видят в электронных библиотеках новые возможности для совершенствования автоматизированных библиотечных систем, превращения их в публичные электронные библиотеки нового поколения с развитыми средствами представления разнообразных цифровых информационных ресурсов и доступа к ним, создаваемые с учетом необходимости интеграции издательских и библиотечных технологий.

## 2 Электронные библиотеки и ЕНИП

На протяжении ряда последних лет в РАН ведутся работы по разработке концепции и реализации Единого Научного Информационного Пространства РАН (ЕНИП РАН) [1], призванного удовлетворить потребность научных сотрудников в необходимости, как поиска качественной информации, так и в выставлении собственной информации в сети Интернет [2].

Основу ЕНИП РАН составляют, прежде всего, стандарты на метаданные информации, циркулирующей в ЕНИП. Эти стандарты должны отвечать следующим требованиям:

Включать в себя основные типы информации, требующейся для поддержки работы научного сотрудника.

Быть открытыми, т.е. обеспечивать доступ к соответствующей информации по этим описаниям.

Быть расширяемыми, т.е. обеспечивать возможность детализации описаний.

Обеспечивать возможности интеграции информации.

Обеспечивать возможности уникальной идентификации информации.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

Обеспечивать возможности размещения и поиска информации в распределенной среде.

Быть ориентированными на современные и перспективные технологии описания и использования информации (в нашем понимании – ориентироваться на семантический Веб (Semantic Web)).

Обеспечивать возможности интероперабельности с внешней средой.

## 3 Международные стандарты электронных библиотек

В связи с этим возникает целый ряд серьезных проблем, связанных с интегрированностью информации (под интегрированностью понимается обеспечение полноты и связанности информации, предоставляемой пользователю). Как и во многих других прикладных областях, обеспечение интегрированности неотделимо от разработки стандартов представления.

Международные стандарты электронных библиотек

С точки зрения потребностей научных сотрудников существенным недостатком многих схем метаданных электронных библиотек является то, что они работают лишь с так называемыми документоподобными объектами, определяют метаданные, описывающие только такие ресурсы, не выделяют другие виды важных объектов, например, персоналии, организации, коллекции и т.п. В итоге, например, встретив упоминание персоны в одном месте, невозможно точно установить соответствие с ее упоминанием в другом месте. Это обусловлено тем, что метаданные рассматриваются как нечто, связанное только с документом, их используют как средства идентификации ресурсов только для документов и только для целей их извлечения.

Набор элементов в специализированном профиле метаданных ЕНИП для электронных библиотек основан на предложениях наиболее влиятельных сообществ и организаций, выдвигающих или поддерживающих проекты стандартов (хотя значительное влияние на выбор решений оказал и анализ специфики работы научных сотрудников).

Это, прежде всего такие организации, как DELOS (четыре года спонсируется EU-ICT), представившая программный документ DELOS Digital Library Reference Model [3], где приводится современная концептуальная модель данной области, с определениями важнейших представлений об архитектуре, ресурсах и функциональности DL.

Существенно исследовался тематический сетевой проект Europeana [4], представивший метаданные и функциональную структуру прототипа портала DL, в плане метаданных уточняющий широко используемый стандарт метаданных Dublin Core (в ЕНИП применяются три словаря DC, в том числе Словарь элементов DC-Library). Кроме того, Europeana предоставляет прототип типового европейского портала цифровой библиотеки.

Для представления коллекций в ЕНИП используется Словарь элементов описания коллекций UKOLN, организации, спонсируемой Museums, Libraries and Archives Council (MLA) Великобритании.

Также изучались проекты Библиотеки конгресса США, прежде всего стандарт METS представления описательных, административных и структурных метаданных цифровых библиотек - стандартизованный XML формат передачи DL-объектов между системами, аналогичный Reference Model OAIS.

Была частично использована структура описания концепций и связей в документах культурного наследия, предложенная CIDOC, представленная в формальной онтологии Definition of the CIDOC Conceptual Reference Model, Dec.2008 [5].

При организации взаимодействия DL целесообразно использовать протоколы сбора метаданных и взаимообмена описаниями OAI - Open Archives Initiative – организации, занимающаяся разработкой интероперабельных стандартов доступа к электронным архивам.

При выборе набора библиографических элементов учитывался стандарт Publishing Requirements for Industry Standard Metadata (PRISM), разработанный издательскими организациями (Рабочей Группой) для обмена метаданными о публикациях (документах, журналах, книгах и пр.). Основан на DCMI, но в большей степени ориентирован на библиографические ресурсы. PRISM предлагает среду обмена и сохранения контента и метаданных, набор элементов описания контента и ряд контролируемых словарей значений этих элементов. В ЕНИП используются схема контролируемых словарей PRISM и основной набор элементов PRISM.

## 4 Системы управления электронными библиотеками

Схемы метаданных играют в ЕНИП двоякую роль. С одной стороны, они служат «обменными схемами», с разными уровнями детализации, для обмена данными между системами, входящими в Единое Научное Информационное Пространство. С другой стороны, в рамках ЕНИП стоит задача не только предложить обменные схемы, но и разработать конкретные типовые информационные системы для научных институтов, библиотек, издательских отделов и пр., которые дали бы стимул к информационному наполнению ЕНИП.

Поскольку цифровые библиотеки во многом похоже друг на друга, и, зачастую различаются лишь такими параметрами как визуальное оформление, набор дополнительных метаданных,

содержимое, виды каталогизации содержимого, то логично провести аналогию с базами данных, как это описывается в DELOS и в качестве типового решения разрабатывать не конкретную цифровую библиотеку, а систему управления цифровыми библиотеками. Такая система позволяет после установки и минимальной настройки администратором системы работать в цифровой библиотеку практически любой направленности и сложности.

## 5 Метаданные электронной библиотеки

В профиле метаданных ЕНИП для электронных библиотек активно используются ресурсы, представленные в основном профиле и некоторых его расширениях, такие как Организации, Персоны, Коллекции и т.д. Тем не менее, центральным остается библиографическое описание публикации, отвечающее за представление метаданных об официально зарегистрированных печатных изданиях.

Использование публикаций в научно-исследовательском процессе выдвигает необходимость быстрого ознакомления с содержимым публикации, и аннотация здесь оказывается часто недостаточной. В связи с этим в инструментарии ЕНИП разработаны средства полуавтоматического выделения оглавления с обеспечением ссылок на соответствующие разделы документа, а также средства работы с библиографическими ссылками.

В интерфейсе администратора системы имеется возможность отменить представление в интерфейсах пользователя каких-либо из перечисленных выше (необязательных) свойств.

С точки зрения потребностей научных сотрудников существенным недостатком многих схем метаданных электронных библиотек является то, что они работают лишь с так называемыми документо-подобными объектами (ДПО), определяют метаданные, описывающие только такие ресурсы, не выделяют другие виды важных объектов, например, персоналии, организации, конференции и т.п. В итоге, например, встретив упоминание персоны в одном месте, невозможно точно установить соответствие с ее упоминанием в другом месте. Даже идентифицировав каким-то образом персону, зачастую нет возможности получить документы, связанные только с ней. Это обусловлено тем, что метаданные рассматриваются как нечто, связанное только с документом, как качественные данные для “полнотекстовой” индексации значений атрибутов. Они не выделяют типы ресурсов, используют средства идентификации ресурсов только для документов и только для целей их извлечения.

В связи с этим в профиле метаданных ЕНИП для электронных библиотек активно используются ресурсы, представленные в основном профиле и некоторых его расширениях, такие как Организации, Персоны, Мероприятия и т.д. Тем не менее, центральным остается библиографическое описание публикации, отвечающее за представление метаданных об официально зарегистрированных печатных изданиях.

В целях обеспечения поддержки различных уровней детализации информации о публикациях, необходимых различным приложениям, библиографическая специализация разделена на базовую и расширенную подсхемы, а также выделяется академическая подсхема, отражающая специфику научных публикаций. Уже на базовом уровне требуется структурировать информацию обо всех вышестоящих библиографических уровнях для каждой публикации. Например, для описания ряда статей в журнале, необходимо описать сам журнал как издание сводного уровня, далее описать интересующие выпуски этого журнала как издания монографического уровня, и, наконец, сами статьи как издания аналитического уровня. И статья, и выпуск, и журнал как таковой являются полноценными структурированными ресурсами, описываемыми лишь единожды, и связываемыми с помощью URI-ссылок.

Такой структурированный подход требует некоторого усилия со стороны систем с «планарным» описанием публикаций. Однако, структуризация информации обо всех библиографических уровнях необходима и крайне важна для схем ЕНИП. Она позволяет избежать дублирования информации, эффектов наличия опечаток в названиях группирующих выпусков, серий и пр., позволяет представить пользователю информацию в целостном и непротиворечивом виде.

Базовый уровень Публикации включает следующие свойства:

- Название - Имя, сопоставленное ресурсу, обычно, под которым он официально известен.
- Альтернативный заголовок* - Любая форма заголовка, используемая как замена или альтернатива официального заголовка ресурса.
- Аннотация - Краткое описание или содержание источника.
- Ключевые слова - Классификация с помощью списка слов с разделителями (например, через запятую).
- Источник - Описание источника информации о данном ресурсе, например, наименование организации, ФИО и пр.
- Авторские права - Авторские права («копирайт») на ресурс.
- Web-адрес* - URL, в частности, HTTP-адрес контактной web-страницы, либо адрес FTP.
- Язык (элемент словаря: Язык) - Язык интеллектуального содержания ресурса.
- Выпущен - Дата формального выхода издания в свет.
- Идентификатор* (подструктура: Идентификатор, рекомендуемые значения: ISBN) - Указание идентификатора ресурса с

помощью рекомендуемых стандартных систем идентификации (см. класс "Идентификатор").
- Авторы* (ссылка: Персона) - Автор(ы) данной публикации.
- Издатель (ссылка: Организационная единица) - Организация, ответственная за публикацию данного издания.
- Редактор* (ссылка: Персона) - Редактор издания.
- Входит в состав (ссылка: Публикация) - Данный ресурс является физически или логически частью указанного ресурса.
- Включает* (ссылка: Публикация) - Данный ресурс физически или логически включает указанный ресурс.
- Кол-во страниц - Количество страниц в публикации.
- Реферат* (подструктура: Файл данных) - Реферат(ы) по данной публикации.
- Библиографическое описание - Библиографическое описание публикации по ГОСТ целиком, строкой. Может быть указано помимо отдельных элементов биб.описания, указываемых полями «название», «номер тома/выпуска» и пр.
- Полный код УДК - Тематическая классификация с помощью полного кода УДК (Универсального Десятичного Классификатора).
- Примечания - Произвольные примечания к публикации.
- ББК* (элемент классификатора: Рубрика ББК) - Ссылка на рубрику Библиотечно-Библиографической Классификации, либо вложенное описание рубрики с указанием кода и, возможно, словесной расшифровки.
- Основной код УДК* (элемент классификатора: Основной код УДК) - Тематическая классификация с помощью ссылки на рубрику основной таблицы УДК (Универсального Десятичного Классификатора).

Использование публикаций в научно-исследовательском процессе выдвигает необходимость быстрого ознакомления с содержимым публикации и аннотация здесь оказывается часто недостаточной. В связи с этим в инстументарии ЕНИП разработаны средства полуавтоматического выделения оглавления с обеспечением ссылок на соответствеющие разделы документа, а также средства работы с библиографическими ссылками.

Приведем описание фрагмента профиля электронных библиотек, отражающего решение этих задач. Расширенная схема описания библиографической информации:

- Список литературы (текстом) (подструктура: Файл данных) - Список библиографических ссылок в текстовом виде, если не может быть разобран по отдельным подструктурам поля "список литературы (структурированный)".
- Оглавление (подструктура: Файл данных) - Оглавление данной публикации в виде отдельного файла, либо текстового или XHTML-фрагмента.
- Список литературы (структурированный)* (подструктура: Библиографическая ссылка) - Список библиографических ссылок, указанных в тексте данной публикации, в виде списка структур «Библиографическая ссылка». Поля подструктуры:
  - Приоритет - Число, определяющее порядок вывода элементов. Чем меньше число, тем выше в списке находится данный элемент. При этом не накладывается требования нумеровать элементы сплошной последовательностью (1,2,3..), допустимо указывать приоритеты с пропуском (10,20,30...).
  - Идентификатор ссылки - Идентификатор библиографической ссылки, например «DC», или «12».
  - Текст ссылки - Исходный текст библиографической ссылки, желательно отформатированный как биб. описание по ГОСТ. Как правило, указывается в случае, когда цитируемая работа не может быть указана ссылкой на публикацию как ресурс.
  - Цитируемая публикация (ссылка: Публикация) - Публикация, на которую ссылается данная библиографическая ссылка (цитируемая работа).
- Сведения об издании - Сведения, относящиеся к изданию: в какой редакции, данные об оригинале для переводной литературы, место(а)/город(а) издания.
- Составитель* (ссылка: Персона) - Составитель(и) данной публикации (сборника).
- Коллективный автор публикаций* (ссылка: Организационная единица) - Организация или подразделние, выступающие как коллективный автор данной публикации.
- Переводчик публикаций* (ссылка: Персона) - Переводчик(и) данной публикации.
- Редколлегия* (ссылка: Должность) - Члены редколлегии издания, с указанием должностей и исполняющих лиц.
- Входит в состав* (ссылка: Коллекция) – Колекции, членом которых является данный ресурс..

В приведенных описаниях элементов профиля электронной библиотеки можно видеть использование элементов основного профиля ЕНИП: Персона, Организационная единица, Файл данных, Должность, Коллекция и др. Приведем состав наиболее частно используего класса – Персоны:

- Домашняя страница* - URL-адрес домашней страницы.
- Дата рождения - Дата рождения лица.
- Адрес* - Полный почтовый адрес.

- Имя (подструктура: Имя персоны) - ФИО персоны. Поля подструктуры:
- Фамилия - Фамилия персоны.
- Имя - Личное имя персоны.
- Отчество - Отчество или дополнительные имена персоны.
- Значение - Полное (не разобранное) значение описываемой подструктуры.
- Пол (элемент словаря: Пол) - Пол субъекта.
- Ученая степень* (подструктура: Ученая степень) - Ученая степень персоны (доктор физ.-мат. наук, кандидат технич. наук и т.д.). Поля подструктуры:
- Обладатель (ссылка: Персона) - Обратная связь с лицом-обладателем ученой степени (заполняется автоматически).
- Дата присуждения - Дата присуждения ученой степени/звания.
- Ученая степень (элемент словаря: Ученая степень) - Наименование ученой степени как ссылка на элемент справочника (доктор физ.-мат. наук, кандидат технич. наук и т.д.).
- Специальность ВАК (элемент классификатора: Специальность ВАК) - Рубрика классификатора специальностей ВАК.
- Ученое звание* (подструктура: Ученое звание) - Академическое или ученое звание (типа профессор, академик, доцент, ...). Поля подструктуры:
- Дата присуждения - Дата присуждения ученой степени/звания.
- Присудившая организация - Название организации, присудившей ученое звание (если организация не может быть указана ссылкой).
- Ученое звание (элемент словаря: Ученое звание) - Собственно само ученое звание как ссылка на элемент справочника (профессор, академик, доцент, с.н.с. и пр.).
- Присудившая организация (ссылка: Организационная единица) - Организация, присудившая ученое звание.
- Дата смерти - Дата смерти, в случае описании информации об исторической личности. По наличию данной даты историческая информация отличается от актуальной.
- Место рождения - Место рождения данной личности, указывается в произвольной форме. Ввиду сложности поддержки исторической информации об административно-территориальном делении, классификатор регионов не используется для указания места рождения (поскольку на момент рождения административно-территориальное деление могло быть другим).
- Место смерти - Место смерти данной исторической личности, указывается в произвольной форме, как и Место рождения.
- Электронная почта* - Контактный адрес электронной почты.
- Телефон* - Контактный телефон.
- Факс* - Факс (код/номер).
- WWW-страница* - HTTP-адрес контактной web-страницы.
- FTP-адрес – URL адрес FTP.

Сближение задач электронных библиотек, архивов и музеев в представлении научного наследия выдвигает требование стандартизации метаданных физических музейных предметов и мультимедийных (фото, видео, аудио) ресурсов. В связи с этим в ЕНИП разработан дополнительный прикладной профиль Поддержки музейной деятельности, в котором для новой сущности **Музейный предмет** определены следующие базовые свойства и связи:

- Состояние – Состояние или сохранность предмета.
- Автор описания (ссылка: Персона) – составитель описания предмета.
- Автор сбора (ссылка: Персона) – персона, выполнившая сбор предмета.
- Год сбора – дата сбора предмета.
- Дата поступления – дата поступления предмета в музей.
- Способ поступления (элемент словаря) – способ поступления предмета в музей.

Соответствующие дополнения и изменения внесены в представления участвующих ресурсов основного профиля, такие как Персоны, Результат деятельности, Коллекции.

В отличие от публикаций, описания музейных объектов могут значительно отличаться в различных музеях и здесь невозможно обеспечить всеобъемлющий набор необходимых свойств. В связи с этим для данных объектов реализуется возможность определения дополнительных свойств в виде связей с двумя вспомогательными объектами: Дополнительные свойства и Значения дополнительных свойств. Соответственно, в интерфейсе администратора системы предоставляется возможность определять дополнительные свойства предмета, при этом в интерфейсах ввода и вывода данных создаются представления соответствующих полей. Введенные значения дополнительных полей выдаются в полных сведениях о предмете, но поиск по ним не производится. Таким образом, администратор может добавить такие свойства, как Количество предметов, Автор описания, География, Размеры, Возраст, Способ поступления, Препараты и т.п.

## 6 Медиа-представления

Для обеспечения хранения цифровых представлений ресурсов и абстрагирования от конкретных методов хранения данных, в ЕНИП разработан дополнительный прикладной профиль Расширенной поддержки хранения данных, в котором вводится ряд новых сущностей.

### 6.1 Класс Медиа-объект (MediaObject)

Класс предназначен для описания медиа-объекта как единого целого, состоящего из частей данных с различной функциональной нагрузкой. Медиа-объект включает в себя следующие свойства:

- Части* - Собственно сами части целого медиа-объекта.

### 6.2 Класс Часть медиа-объекта (MediaObjectPart)

Класс позволяет в пределах одного целого медиа-объекта, например, публикации, иметь несколько частей с различной функциональной нагрузкой, такие как содержание, образы страниц в виде изображений, текст публикации в чисто текстовом формате, отформатированный текст публикации и тому подобное. Свойствами части медиа-объекта являются:

- Тип данных – Формат представления данных, хранимых в данной части, например «Документ Microsoft Word» или «Изображение в формате JPEG».
- Функциональный тип – Функциональный тип части медиа-объекта, показывающий какую функциональную нагрузку несёт часть, например «содержание», «страница книги».
- Потоки данных* - Потоки двоичных данных, связанные с частью медиа-объекта в формате соответствующем типу данных.
- Порядок в медиа-объекте – Порядок отображения части в списке частей медиа-объекта.
- Название части – отображаемое название части медиа-объекта.

### 6.3 Класс Тип данных (MediaType)

Класс тип данных представляет собой элемент классификатора форматов представления двоичных данных. Помимо стандартного для ЕНИП описания классификатора включается дополнительно свойство MIME-тип, связывающее данный классификатор со словарём IMT базового профиля ЕНИП.

### 6.4 Класс Функциональный тип (PartType)

Класс функциональный тип представляет собой элемент словаря функциональных типов частей медиа-объектов.

### 6.5 Класс Единица хранения (StorageItem)

Класс Единица хранения, представляющий единый и неделимый поток двоичных данных, позволяет абстрагироваться от конкретных методов хранения данных и позволяет собирать медиа-объекты состоящие из частей расположенных в разных местах и хранимых различными способами.

### 6.6 Класс Файл в файловой системе (FSFile)

Класс, представляющий собой поток двоичных данных, хранимый в файле на файловой системе локального компьютера. Единственным свойством данного класса является Путь к файлу.

### 6.7 Класс Ссылка (Reference)

Класс, представляющий собой поток двоичных данных, хранимый в виде URL-ссылки на внешний источник. Свойствами класса ссылка являются части URL, такие как:

- Схема – вид ресурса (URL scheme).
- Имя компьютера – FQDN имя компьютера или его IP адрес.
- Номер порта – номер порта для протокола TCP или UDP.
- Путь (path) – дополнительная часть URL, назначение которой зависит от схемы.

### 6.8 Класс BLOB запись в БД (DBBLOB)

Класс, представляющий собой поток двоичных данных, хранимый в BLOB поле записи в базе данных. Содержит единственное свойство Данные.

Принцип использования представленного выше класса Медиа-объект в ЕНИП несколько отличается от общепринятого в электронных библиотеках. Для обеспечения цифровых представлений не только публикаций, но и музейных объектов, а также мультимедийных изображений коллекций, фотографий персон, коллективова, зданий организации и т.п., в класс Ресурс, являющийся суперклассом для всех основных объектов онтологии, вводится свойство:

- Медиа-представление* - Медиа-объект (MediaObject).

Таким образом одно или несколько мультимедийных представлений может сопровождать любой объект информационной Web-системы, наследуемый от Ресурс.

## 7 Коллекции

В базовых метаданных ЕНИП предусмотрена поддержка коллекций, однако требования цифровых библиотек, а в особенности с поддержкой хранения музейных предметов, не позволяют их использовать. В связи с эти базовый профиль дополняется коллекциями со следующими атрибутами:

- Название – Название коллекции
- Тип коллекции (элемент словаря)
- Ключевые слова
- Описание
- Администратор (ссылка: Персона)
- Количество элементов в коллекции
- Место хранения
- Примечание
- Элементы коллекции* (ссылка: Ресурс)

Коллекции такого рода позволяют хранить классические коллекций (архивные, музейные) и иметь любые вложенные наборы объектов (выставочные, выездные, по хранению…).

## Заключение

В процессе разработки возник ряд вопросов, которые, по-видимому, можно будет решить в процессе первых установок и использования системы, а именно: делать ли единую ЭБ-систему с публикациями и предметами? Или делать административно настраиваемую на публикации или предметы? Или две отдельных (но единых инструментально) системы? При единой системе делать ли совместный поиск и выдачу списка публикаций и предметов?

Оглавление публикации – отдельный объект или медиа-представление (или группа элементов)? Изображения страниц и распознанный текст – два медиа-представления или постараться объединить? Пытаться ли реализовать распределенность хранения (и вывода) медиа-представлений?

Изучается также вопрос о включении в систему средств реализации распределённого поиска и каталогизации (по OAI-PMH, Z39.50). Однако подобные автономные средства предоставляются в настоящее время многими организациями, и в нашей стране пока еще более актуальны создание и наполнение электронных библиотек, нежели их интеграция.

## Литература


[1] Бездушный А.Н, Бездушный А.А., Серебряков В.А., Филиппов В.И. «Интеграция метаданных Единого Научного Информационного Пространства РАН». М.:ВЦ РАН. 2006.

[2] Бездушный А.Н, Бездушный А.А., Нестеренко А.К., Серебряков В.А.,Сысоев Т.М., Теймуразов К.Б., Филиппов В.И. «Информационная Web-система «Научный институт на платформе ЕНИП». М.:ВЦ РАН. 2007.

[3] Рабочая группа DELOS http://www.delos.info

[4] Europeana Digital Library http://www.europeana.eu/portal/

[5] Definition of the CIDOC Conceptual Reference Model, Dec.2008 http://cidoc.ics.forth.gr/docs/cidoc_crm_version_5.0_Dec08.doc


## Digital library logical model in ENIP ontology


A.A. Zakharov V.I. Filippov



This paper covers key aspects of digital library management system (DLMS) in RAS United Science Informational Space (ENIP). Some important concepts of document-like objects, media objects, collections and extended attributes are described.

# Метод динамического создания связей между информационными объектами базы знаний

© Обухова О.Л., Бирюкова Т.К., Гершкович М.М., Соловьев И.В., Чочиа А.П.
Институт проблем информатики РАН
support@intergallery.ru

**Аннотация**

В статье рассматривается формальная модель организации информационного поиска в базе знаний с использованием возможностей разработанной авторами адаптивной фасетной навигации. Поисковый образ объекта строится в форме фасетной формулы объекта. Построение поискового запроса представляет собой итерационной диалоговый процесс уточнения поискового запроса с учетом прямых и ассоциативных связей информационных объектов. Формальная модель положена в основу реализационной модели, в которой визуальный интерфейс представляет обзор содержания базы знаний и служит удобным механизмом построения поисковых запросов.

## 1 Введение

Интеллектуальный ресурс научно-исследовательского института находит своё отображение в электронной коллекции научных материалов по основным направлениям исследований, ведущихся научными работниками, и отраслям знаний, находящимся в сфере интересов сотрудников. Коллекция научных материалов может служить информационным источником для получения необходимых сведений о поставленных задачах и методах их решения в интересующей исследователя области знаний. Основную часть знаний аналитики получают в результате сравнения, анализа и синтеза информации, размещенной в текстах документов. Для доступа к требуемым материалам целесообразно использовать не только поиск по формальным признакам документов, таким как название, автор, год издания, но и поиск по смысловому содержанию документов. Для выполнения данной задачи авторы поставили целью разработать программную систему размещения научных материалов в базе знаний с возможностью информационного поиска.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

«Когда удачно завершился поиск и нужная информация найдена - она становится знанием, необходимым для принятия решения или совершения действий.»[1] Своевременно предоставленные знания служат повышению эффективности деятельности научного сообщества. «Особое значение имеет повышение продуктивности интеллектуальной деятельности.»[2]

Задача разработки программной системы поддержки базы знаний является актуальной. Вопросам организации информационного поиска в базе знаний с использованием возможностей фасетной навигации в электронных коллекциях[3] посвящена данная статья.

## 2 Концептуальная модель

В основе нашего подхода к разработке информационного поиска лежит тот факт, что информация о научно-исследовательской работе научного института хранится в базе знаний в виде информационных объектов (ИО), сопровождаемых *поисковым образом документа (ПОД)* [4], в котором кратко выражается основное смысловое содержание документа. Информационными объектами являются темы НИР, публикации, зарегистрированные продукты и патенты, диссертации, доклады на конференциях. Применение гибкого механизма образования логически связанных цепочек информационных объектов позволяет получать те знания по выбранному направлению, которые интересуют конкретного пользователя. Для этого авторами предлагаются методы поиска, основанные на технологии адаптивной фасетной навигации.

Используя принципы фасетной классификации [5] и разработанную авторами технологию адаптивной фасетной навигации [6], мы предлагаем концептуальную модель создания ПОД в форме фасетной формулы объекта, представленной в виде совокупности: фасетный признак – список значений фасетного признака для данного объекта, и поискового запроса (строящегося итерационно в процессе диалога) в форме фасетной формулы запроса. Данный подход развивает идеи, изложенные авторами ранее [7], в том направлении, что для каждого информационного объекта по

каждому (в общем случае) фасетному признаку задается список значений фасетного признака, характеризующих объект. Набор фасетных признаков диктуется направленностью предметной области, к примеру, для коллекции электронных документов сайта научного института набор фасетных признаков определяется видами и характером научной деятельности. Количество фасетных признаков выбирается в соответствии с принципом, сформулированным американским психологом Миллером [8]: для того, чтобы выбор был эффективным, количество элементов в нем не должно быть больше семи-девяти.

Список значений для каждого фасетного признака формируется и обновляется в административном режиме в процессе работы программной системы поддержки базы знаний в момент занесения информационных объектов. Результирующий список доступных значений фасетного признака является объединением всех заданных значений данного фасетного признака для каждого ИО. Занесение нового или удаление существующего ИО приводит к обновлению списка доступных значений каждого фасетного признака. Анализ сочетаний фасетных признаков и их заданных значений для информационных объектов позволяет программным образом в процессе работы системы выстраивать логические связи между ними. Связи динамически создаются и модифицируются в зависимости от актуального состояния электронной коллекции научных материалов, размещенной в базе знаний.

Общий обзор всех фасетных признаков и полного списка их значений дает представление о сферах, характере и направлениях научной деятельности, которые представлены в работах, хранящихся в базе знаний.

## 3 Формирование фасетной формулы объекта

Введем определение понятия «Фасетная формула объекта».

$G = \{g_i \mid i = 1,...k\}$ - коллекция информационных объектов, где $g_i$ - информационный объект;

$\Phi(G) = \{\Phi_j(G) \mid j = 1,...l\}$ - множество фасетов на $G$.

При рассмотрении конкретной коллекции $G$ будем обозначать: $\Phi_j(G) = \Phi_j$.

$\varphi^{\Phi_j} = \{\varphi_{j1},...\varphi_{jm_j}\}$ - множество допустимых значений фасета $\Phi_j$ (здесь $m_j$ - количество допустимых значений фасета $\Phi_j$).

Если для объекта $g_i$ задано хотя бы одно значение фасета $\Phi_j$, то

$\varphi_j(g_i) = \{\varphi_{js}(g_i), s \in \{1,...m_j\}\} \subset \varphi^{\Phi_j}$ - набор (множество) заданных значений фасета $\Phi_j$ для объекта $g_i$.

Если для объекта $g_i$ значения фасета $\Phi_j$ не заданы, то $\varphi_j(g_i) = \varnothing$ ($\varnothing$ - пустое множество).

$\varphi(g_i) = \{\varphi_j(g_i)\ \forall j : j \in \{1,...l\}, \varphi_j(g_i) \neq \varnothing\}$ - множество наборов заданных значений фасетов для объекта $g_i$ .

**Фасетная формула** *FF(g$_i$)* **объекта** $g_i$ представляет собой множество совокупностей следующего вида:

{[фасет $\Phi_j$; набор заданных для объекта $g_i$ значений фасета $\Phi_j$]}, то есть:

$$FF(g_i) = \{[\Phi_j, \varphi_j(g_i)] \mid \Phi_j \in \Phi(G), \varphi_j(g_i) \in \varphi^{\Phi_j} \ \forall j : j \in \{1,...l\}, \varphi_j(g_i) \neq \varnothing\}$$

Фасетная таблица содержит фасетные формулы для всех объектов коллекции:

$$FT(G) = \{FF(g_i) \mid g_i \in G, i = 1,...k\} = \{(\Phi_j, \varphi^{\Phi_j}) \mid j = 1,...l\}$$ -

множество фасетов с их допустимыми значениями на объектах коллекции $G$. Фасетная таблица служит базисом для выявления взаимосвязей различных информационных объектов.

Для формирования результирующего запроса с целью получения искомой выборки объектов $RG \subset G$ необходимо в процессе пошагового алгоритма задать значения некоторого подмножества фасетов для построения фасетной формулы запроса.

## 4 Формирование фасетной формулы запроса

Поисковый запрос строится в форме фасетной формулы запроса в процессе выполнения пошагового алгоритма выбора фасетных признаков и их значений из списка допустимых значений. На каждом шаге пользователь уточняет фасетную формулу, выбирая одно или несколько значений для очередного фасетного признака из множества допустимых значений фасетного признака на данном шаге.

Определим фасетную формулу запроса RFF как множество совокупностей следующего вида:

{[фасет $\Phi_j$; набор значений фасета $\Phi_j$, формирующих запрос]}

Операция выборки (*retrieve*) формирует подмножество объектов $RG \subset G$ в соответствии с фасетной формулой запроса:

$$RG = \text{retrieve}(G, RFF)$$

Такой способ построения поискового запроса обеспечивает получение на каждом шаге непустого подмножества объектов. Список допустимых значений для каждого фасета обновляется в соответствии с состоянием полученного подмножества объектов. Новые списки значений фасетов позволяют пользователям продолжить уточнение поискового запроса.

Авторами разработана модель фасетной навигации, в которой на каждом шаге актуальное подмножество информационных объектов определяется полным или частичным совпадением фасетной формулы объекта с фасетной формулой запроса.

В настоящей статье предлагается развитие этого подхода в направлении формирования расширенного актуального подмножества ИО, включающего объекты, объединенные прямыми и ассоциативными связями.

Определение 1. Информационные объекты $g_1$ и $g_2$ объединены прямой связью, если $\exists$ хотя бы один фасетный признак $\Phi_j$, для которого $\varphi_j(g_1) \bigcap \varphi_j(g_2) \neq \varnothing$

Определение 2. Информационные объекты $g_1$ и $g_2$ объединены ассоциативной связью, если для всех фасетных признаков $\Phi_j$, j=1,…l, $\varphi_j(g_1) \bigcap \varphi_j(g_2) = \varnothing$ , но $\exists$ хотя бы один информационный объект $g_3$, у которого для хотя бы одного фасетного признака $\Phi_\kappa$ $\left(k \in \{1,..l\}\right)$ выполняется условие: $\varphi_k(g_3) \bigcap \varphi_k(g_1) \neq \varnothing$ и $\varphi_k(g_3) \bigcap \varphi_k(g_2) \neq \varnothing$.

Данные определения используются в предлагаемом авторе подходе к формированию расширенного подмножества ИО в соответствии со следующими правилами: расширенное подмножество формируется из ИО, для которых фасетная формула объекта полностью или частично совпадает с фасетной формулой запроса, и расширяется информационными объектами, которые объединены с ними прямой или ассоциативной связью.

Список допустимых значений каждого фасета перегружается в соответствии с состоянием расширенного множества ИО.

Авторы поставили перед собой задачу разработки макета поисковой системы с целью апробации изложенного метода.

## 5 Визуальный интерфейс диалоговой модели информационного поиска

Для визуализации пользовательского интерфейса разработана технология динамического создания интерфейсных окон на основе технологии Adobe Flash со встроенным языком ActionScript.

Задача пользовательского интерфейса заключается в том, чтобы на каждом шаге выполнения алгоритма:

1. представить для просмотра все фасеты со множеством доступных значений;

2. обеспечить возможность выбора списка значений в определенном фасете;

3. получить текущую выборку ИО, анализ которой позволит пользователю определить свои последующие шаги по формированию фасетной формулы запроса.

Элементы визуального интерфейса формируются программным образом и определяются текущим состоянием подмножества информационных объектов и предысторией формирования фасетной формулы запроса.

Список допустимых значений для каждого фасетного признака обновляется в соответствии с состоянием полученного подмножества объектов. Новые списки значений фасетов позволяют пользователям продолжить уточнение поискового запроса. Текущее состояние визуального интерфейса позволяет получить представление о той части базы знаний, границы которой определяются выбранными значениями фасетных признаков на предыдущих шагах. Дополнительно к этой информации пользователь может проанализировать текущую выборку информационных объектов и сделать очередной шаг при выборе объектов, продолжив уточнение фасетной формулы запроса для достижения совпадения со своей информационной потребностью.

Динамически появляющаяся информация на каждом шаге доступа к объекту предоставляет пользователю данные, анализ которых позволяет ему сделать следующий выбор.

На рисунке 1 продемонстрирована работа с фасетом «Авторы». Список всех фасетов размещен в левой панели. Активация одного из фасетов приводит к генерации интерфейсного окна, содержащего множество доступных значений данного фасета с поддержкой возможности выбора списка значений, представляющих интерес для пользователя.

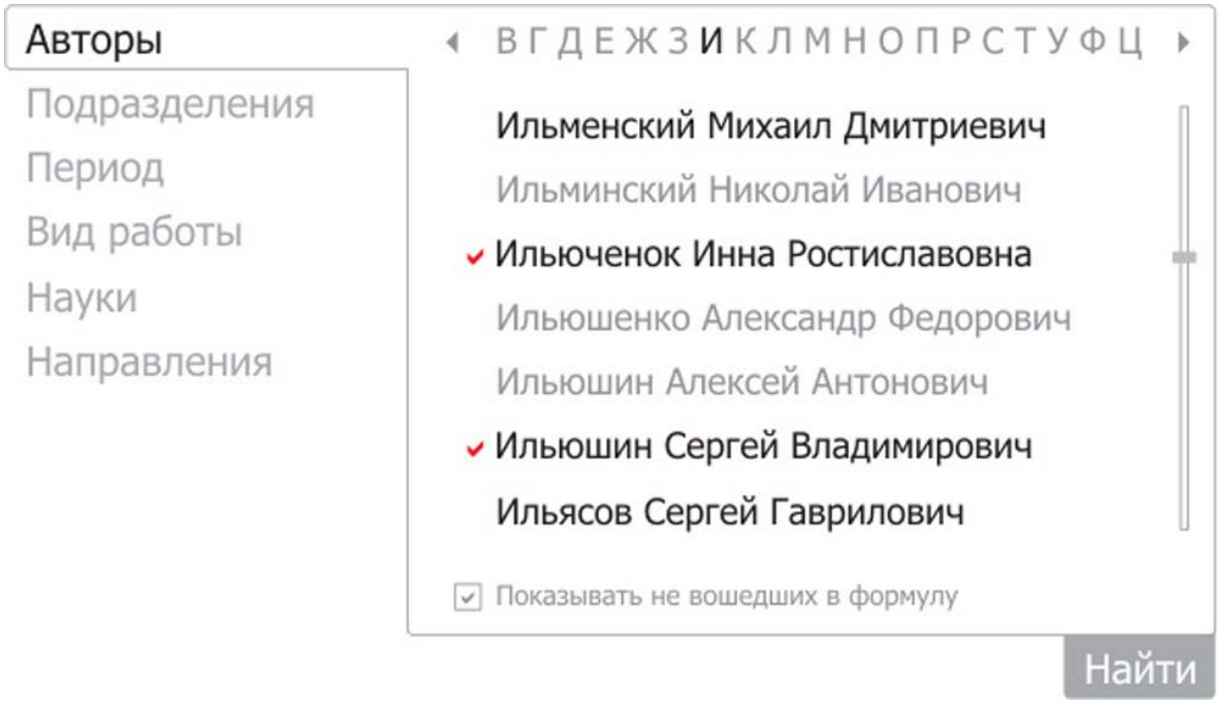



Рисунок 1

Выбранные значения формируют фасетную формулу запроса. Эти значения будут определять определенные границы подмножества ИО. Данное подмножество ИО на каждом шаге выполнения алгоритма будет обновляется с учетом прямых и ассоциативных связей объектов, соответственно, список доступных значений фасета на следующем шаге будет определяться состоянием расширенной выборки. Но, возможно, пользователю интересно будет в какой-то момент посмотреть и те значения фасета, которые он «проигнорировал» в силу логики своей работы. Для этого введена функция «Показывать не вошедших в формулу» (рисунок 2).

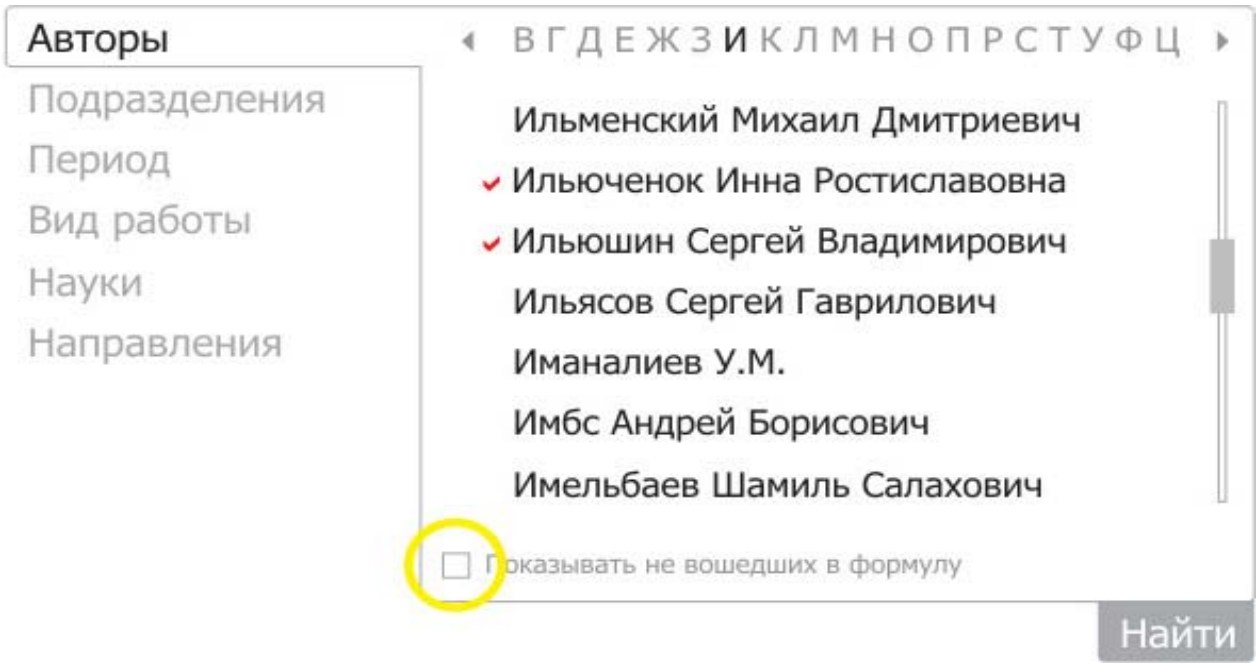



Рисунок 2

На рисунке 3 представлена работа с фасетом «Период».

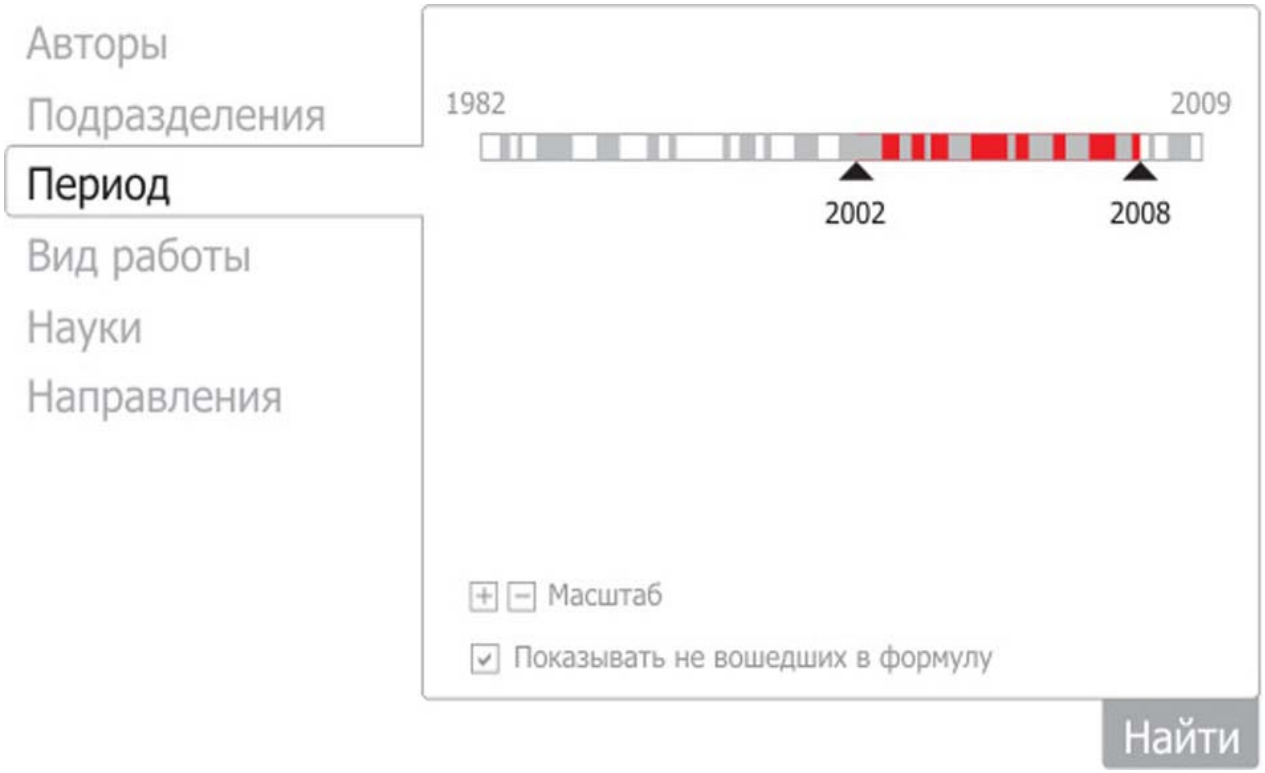



Рисунок 3

Для фасета «Период» введена дополнительная функция «Масштаб». Если пользователю необходимо более детальное уточнение периода появления той или иной публикации, то он может использовать данную возможность (рисунок 4).

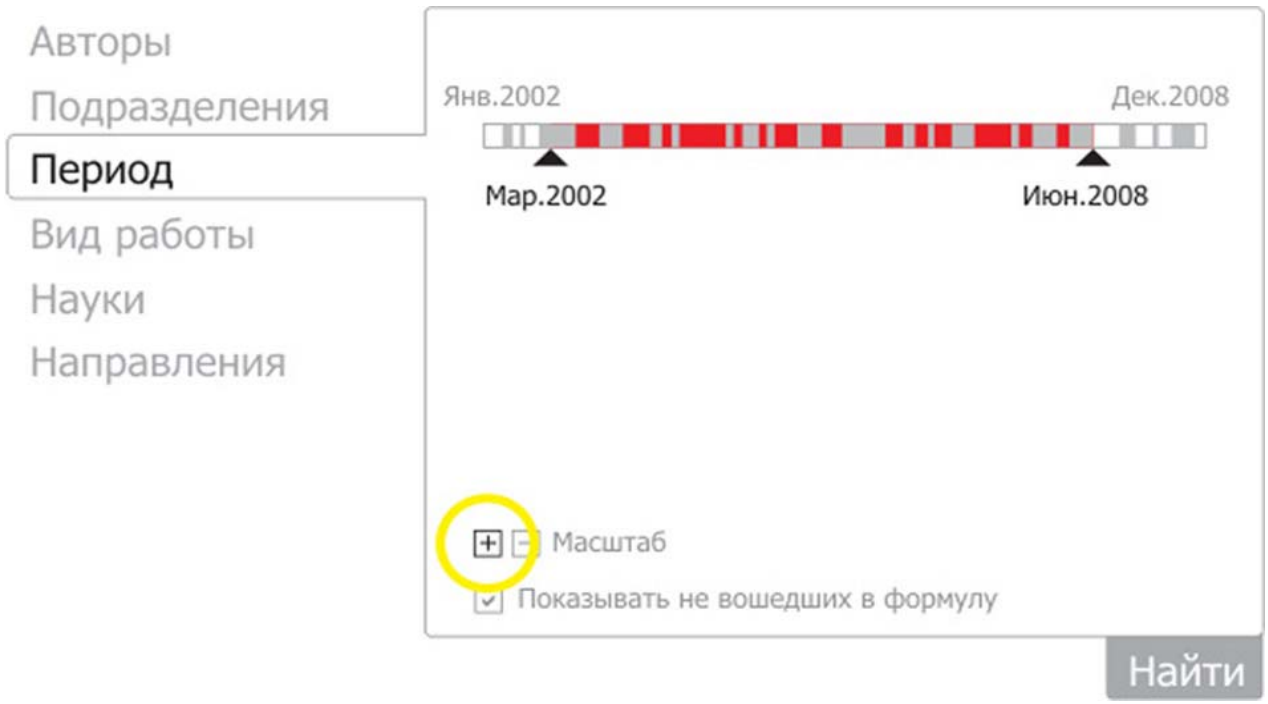



Рисунок 4

Особого подхода требуют фасеты, у которых список доступных значений представляет иерархию, к примеру, фасет «Науки» (рисунок 5).

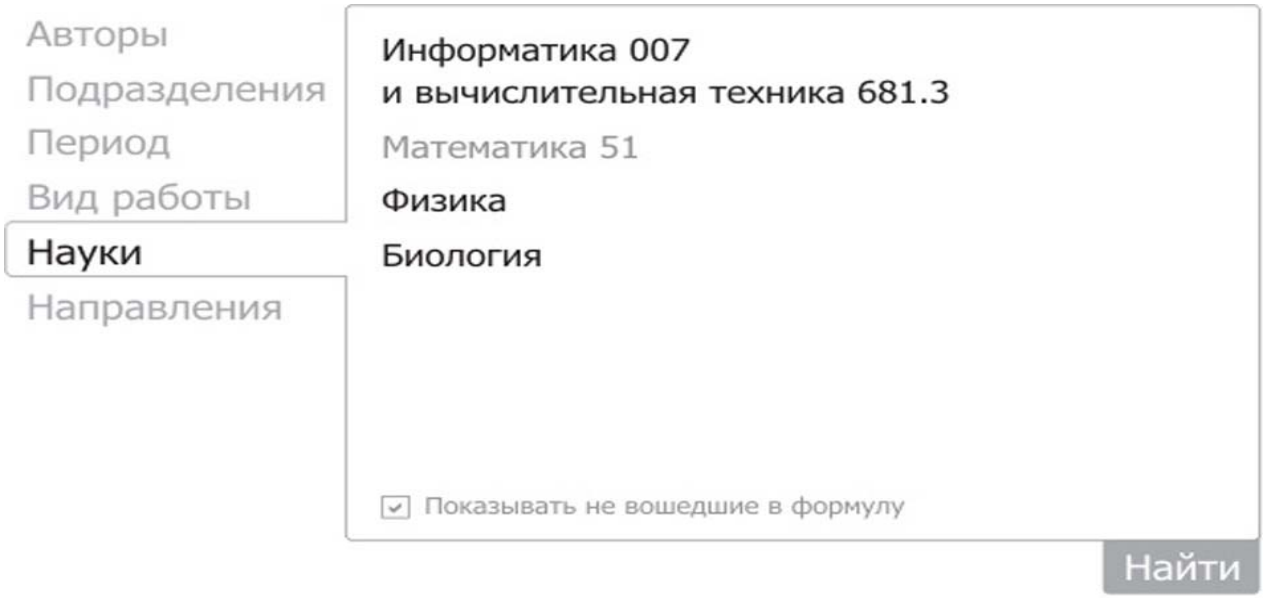



Рисунок 5

После фиксации значения фасетного признака первого уровня генерируется окно, в котором представлен второй уровень иерархии значений, что показано на рисунке 6.

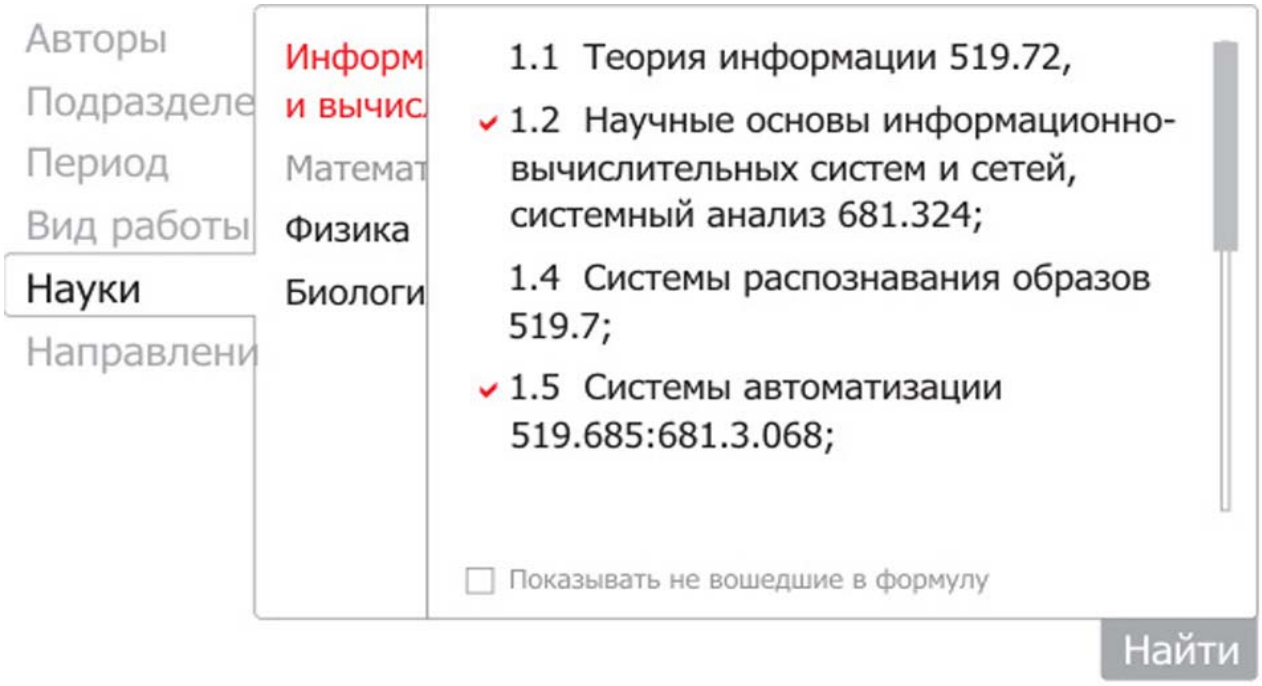



Рисунок 6

## 6 Задачи реализационной модели

Изложенные идеи авторами воплощаются в реализационной модели.

Используемые программные средства:

- реализация осуществляется в среде ОС: Microsoft Windows 2000/2003;
- веб-сервер: Microsoft Internet Information Server 5.0/6.0;
- контейнер серверного приложения: Microsoft ASP.NET 2.0;
- СУБД: Microsoft SQL Server 2005;
- в качестве средств разработки используется Microsoft Visual Studio 2005, ASP.NET 2.0;
- для построения пользовательского интерфейса используется Adobe Flash со встроенным языком ActionScript.

ASP.NET Web Services используют индустриальные стандарты:

-XML. Обмен данными между визуальным пользовательским интерфейсом и сервером

-SOAP. Протокол обмена сообщениями между Web службой и клиентом, основанный на XML .

-Web Services Description Language (WSDL). Описывает параметры сообщений Web-службы для взаимодействия с клиентами .

Опытная эксплуатация реализационной модели позволит вести файлы протокола и собирать статистическую информацию с тем, чтобы анализировать и обобщать возможные действия пользователей по формированию фасетной формулы запроса. В частности, только на практике можно проверить, интересен ли пользователям механизм учета прямых и ассоциативных связей информационных объектов. Авторы планируют сделать этот сервис опциональным, и сбор статистики об использовании этой опции позволит разработчикам определить, является ли востребованной эта возможность и стоит ли её поддержку перенести в действующую модель поисковой системы базы знаний.

## 7 Заключение

Авторы предложили свой взгляд на организацию информационного поиска в базе знаний научного института на базе адаптивной фасетной навигации. Реализационная модель позволит разработчикам исследовать вопрос - может ли предложенный подход повысить эффективность информационного поиска.

## Литература


[1]Андрей Гребенюк, Сергей Киселев. Информационные системы управления знаниями компании / Журнал "Коммерческий директор" Москва, 2 апреля 2007 года http://www.komdir.ru

[2] Ильин В. Д., Соколов И. А. Символьная модель системы знаний информатики в человеко-автоматной среде // Информатика и её применения. 2007 г. том 1. выпуск 1.

[3] Чочиа А.П., Соловьев И.В., Обухова О.Л., Бирюкова Т.К., Гершкович М.М. Модель адаптивной фасетной навигации в открытых электронных коллекциях // ”Системы и средства информатики”, выпуск 18, , Москва, Наука, 2008 .

[4] ГОСТ 7.73 — 96. МЕЖГОСУДАРСТВЕННЫЙ СТАНДАРТ. ПОИСК И РАСПРОСТРАНЕНИЕ ИНФОРМАЦИИ

[5] Ранганатан Ш.Р. Классификация двоеточием. Основная классификация: Пер. с англ. / Под ред. Т.С. Гомолицкой и др. - М., 1970. – 422 с.

[6] Обухова О.Л., Гершкович М.М., Бирюкова Т.К., Соловьев И.В., Чочиа А.П. Открытые электронные коллекции с адаптивным визуальным интерфейсом фасетной навигации // Труды 10-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2008, Дубна, Россия, 2008

[7] Н.А. Маркова, О.Л. Обухова, И.В. Соловьев, А.П. Чочиа. Эффективная фасетная навигация в электронных коллекциях // Системы и средства информатики, № 17, Москва, Наука, 2007, С. 214-222

[8] Миллер Дж. Магическое число семь, плюс или минус два. - В кн.: Инженерная психология. М. 1964


# The method for dynamic association of informational objects in knowledge base


Olga Obuhova, Maxim Gershkovich, Tatiana Biryukova, Ivan Soloviev, Anton Chochia



We propose the formal model for the process of informational search in the knowledge base utilizing features of adaptive facet navigation. The ‘search image’ of the object is build as object’s facet formula. Generation of the search request is conducted as a sequence of iterations, where search request being adjusted in dialogue mode using both direct and associative links between informational objects. The formal model is used to develop executive model, where visual interface shows content of the knowledge base and serves as a suitable tool for generation of the search requests.

# Метод выявления неявных связей объектов

© Снарский А.А.[1], Ландэ Д.В.[1,2], Женировский М. И.[3]

[1]НТУУ «Киевский политехнический институт»,
[2]Информационный центр «ЭЛВИСТИ»,
[3]Институт теоретической физики им. Н.Н. Боголюбова НАН Украины
asnarskii@gmail.com, dwl@visti.net

**Аннотация**

Описывается метод, позволяющий выявлять неявные связи в сложных сетях, представленных матрицами инцидентности. Описывается применение данного метода, базирующегося на теории электрических сетей, для выявления силы взаимосвязей понятий, извлекаемых из неструктурированных текстов, в частности, персон.

## 1 Матрицы инцидентности

В настоящее время в теории и практике аналитической деятельности получила большое развитие концепция сложных сетей (complex networks) [16], являющаяся с одной стороны, развитием теории графов, а с другой стороны, областью применения подходов, применяемых в физической науке, например, в теории электрических цепей или теории перколяции. Переход к физической парадигме объясняется, по-видимому, именно сложностью этих сетей, которые, на самом деле окружают нас повсюду – это и транспортные сети, и сети цитирования, и, безусловно, Интернет [8]. В частности, сети, образуемые персонами, совместно упоминаемыми в одних и тех же публикациях, позволяют аналитикам делать выводы об общих интересах отдельных групп персон во времени [15], выявлять неявные связи [10], пренебрегать несущественными и т.п.

Технологиям выявления понятий из неструктурированных текстов посвящено достаточно много публикаций [1,2,12], эта проблематика выходит за рамки нашего исследования. В данной работе предлагается метод исследования сети понятий, характеризующейся большим количеством узлов, ребер (связей) с различными весовыми значениями, высокой динамикой появления новых узлов и связей. Известно, что матрицы взаимосвязей понятий (МВП) [5,6] являются одной из форм представления сетевых структур, аналогичной по функциональности их графовому представлению. На практике эти матрицы чаще всего отражают близость отдельных понятий (совместную встречаемость в документах или близость по сопутствующему контексту в разных документах). При самых различных подходах к их построению - это, как правило, симметричные матрицы, элементы которых – коэффициенты взаимосвязей. Если отношения между понятиями не носят направленного характера, то их также можно рассматривать как неориентированные графы и применять к ним соответствующие методы. Чаще всего ребрам этих графов приписываются весовые коэффициенты, которые пропорциональны количеству документов из некоторого массива, одновременно соответствующие обоим узлам (понятиям), соединяемым этими ребрами. Существуют и другие многочисленные подходы к определению близости понятий в массивах неструктурированных текстов, среди таких можно назвать контекстные, вероятностные и энтропийные (Mutual Information) [5,9,13], но все они являются лишь предпосылками для построения матриц взаимосвязей, их перегруппировки и визуализации [11,14].

Рассмотрим одно из формальных определений матрицы взаимосвязей понятий $M$, соответствующее приведенным в работах [5] и [6]. Обозначим $p_i$ $(i = 1,...,K)$ – понятие, $d^j$ $(j = 1,...,N)$ – документ, $d^j \in D$ – массив документов, $e_i^j$ – признак соответствия понятия $p_i$ документу $d^j$:

$$e_i^j = \begin{cases} 1, & p_i \in d^j \\ 0, & p_i \notin d^j \end{cases}$$

Можно определить уровень связи понятий $p_i$ и $p_k$:

$$M_{ik} = \sum_{j=1}^{N} e_i^j e_k^j .$$

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

Введя обозначение: $E = \left\| e_i^j \right\|_{i=1,\ldots,K}^{j=1,\ldots,N}$, получаем:

$$M = EE^T = \left\| M_{ik} \right\|_{i,k=1,\ldots,K}.$$

Будем называть данную матрицу инцидентности $M$ матрицей взаимосвязей понятий. Недиагональный элемент $M_{ik}$ $(i \neq k)$ этой матрицы равен количеству одновременных упоминаний узлов (персон) $i$ и $k$ во всех статьях из базы данных. Диагональный элемент матрицы $M_{ii}$ - это количество упоминаний $i$ - того узла (персоны) во всех документах. На рис. 1 приведен пример трехмерного изображения матрицы взаимосвязи понятий, состоящей из 84 узлов (персон). Данная матрица была получена на основании анализа массива веб-публикаций, сосканированных системой контент-мониторинга InfoStream [3] в течение первого квартала 2009 года по тематике деятельности Киевской городской государственной администрации. Объем исходных данных составил свыше 10 тыс. документов.

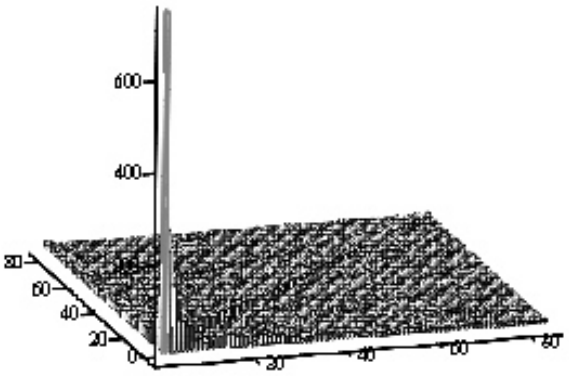



*Рис.1.Изображение модифицированной матрицей связей $M$. По горизонтальным осям отложены номера узлов (персон), по вертикальной – весовые значения связей*

## 2 Коэффициент сцепления

В рамках рассматриваемого метода предлагается новая характеристика сложных сетей - коэффициент сцепления (cohesion). Пусть мы имеем некоторую сеть, узлами которой являются люди (персоны), а ребрами - некоторые отношения между ними (как такие отношения, можно рассматривать общие интересы, упоминаемость в одних и тех же документах, и т.п.). При этом каждый из узлов связан с некоторым количеством других узлов этой сети. Актуальной является задача исследования такой сети - выяснение, какие узлы в ней играет ведущую роль и, главное, насколько эти главные узлы хорошо связаны (сцеплены), между собой. То есть на входе имеется стандартная матрица инцидентности $M$, соответствующая исходному графу связей, а на выходе хотим получить номера так называемых главных узлов и узнать, насколько хорошо эти узлы сцеплены между собой.

Будем трактовать значение элемента матрицы $M$ - $M_{ik}$, как числа, которое приписывается весу связи (ребра) между $i$ и $k$, в качестве проводимости этой связи, по аналогии с теорией электрических цепей (см., например, [4, 7]). Тогда по аналогии с этой теорией можно ввести так называемую матрицу инцидентности проводимости $A$ для матрицы $M$:

$$A_{ik} = -M_{ik},$$
$$A_{ii} = \sum_{j \neq i} \left| A_{ij} \right|$$

Здесь $A_{ii}$ - сумма проводимостей ребер, инцидентных данному узлу, а $A_{ik}$ - проводимость прямой связи между узлами (персонами) $i$ и $k$ взятая со знаком минус.

Зная матрицу $A$ можно найти матрицу кондактанса (полной проводимости) $G$, каждый элемент которой $G_{ik}$ соответствует полной проводимости с учетом всех прямых и непрямых связей между двумя узлами $i$ и $k$ ($i \neq k$). Будем называть величину $G_{ik}$ коэффициентом когезии (сцепления):

$$G_{ik} = \frac{\det(A)}{\det\left(A_{(i+k)(i+k)}\right)}.$$

Здесь $A_{(i+k)(i+k)}$ - это минор матрицы $A$, который вычисляется следующим образом: строка $i$ прибавляется к строке $k$ и затем вычеркивается, столбец $i$ прибавляется к столбцу $k$ и затем также вычеркивается. Если один из индексов равен нулю, то просто вычеркивается столбец и строка, соответствующие ненулевому индексу.

Для реальной базы данных персон, для которой построена матрица взаимосвязей, полученная матрица $G$ графически представлена на рис. 2

Характеристикой всей системы является средний коэффициент сцеплений (когезии) $G_{av}$, равный

$$G_{av} = \frac{1}{N(N-1)} \sum_{\substack{i,k \\ i \neq k}}^{N} G_{ik}.$$

Для пояснения смысла вводимых параметров рассмотрим также «игрушечную» небольшую базу данных связей (сеть) из четырех узлов (персон), для которой матрица $M$ имеет вид:

$$M^1 = \begin{pmatrix} 5 & 2 & 3 & 0 \\ 2 & 3 & 0 & 1 \\ 3 & 0 & 6 & 3 \\ 0 & 1 & 3 & 4 \end{pmatrix}$$

На рис. 3 эта же база данных изображена в виде графа, около каждой связи, которой проставлен значение проводимости (совместных упоминаний).

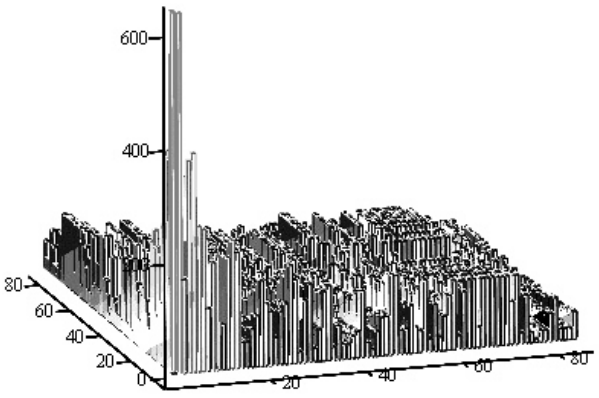



*Рис. 2. Матрица когезии. По горизонтальным осям отложены номера узлов (персон), по вертикальной – значения матрицы*

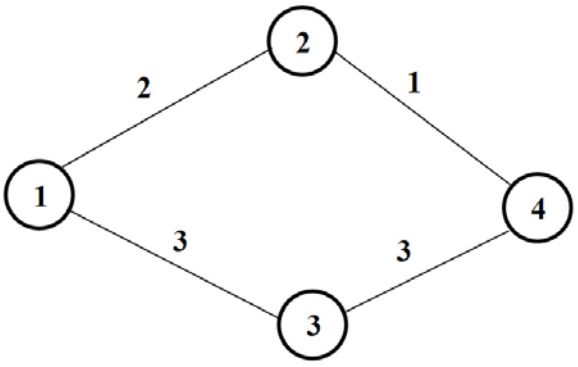



*Рис. 3. Граф модифицированной матрицы связей из четырех узлов (персон) $M^1$*

Матрица сцепления (когезии) $G^1$, соответствующая матрице $M^1$, равна:

$$G^1 = \begin{pmatrix} - & 2.6 & 3.6 & 2.2 \\ 2.6 & - & 2 & 1.9 \\ 3.6 & 2 & - & 3.5 \\ 2.2 & 1.9 & 3.5 & - \end{pmatrix},$$

а среднее значение $G^1_{av} = 2.2$.

Как видно из этого примера матрица когезии, в отличие от матриц $A$ и $M$ учитывает (с соответствующим весом) все, а не только прямые связи. В самом деле, элемент $M^1_{1,4}$ равен нулю, между узлами 1 и 4 нет прямой связи. В тоже время, между этими узлами есть опосредованные связи через узлы 2 и 3. Заметим, что коэффициент когезии $G^1_{1,4}$ узлов 1 и 4, непосредственно не связанных между собой больше чем этот же коэффициент $G^1_{2,4}$ для связанных между собой узлов 2 и 4.

## 3. Применение

Будем теперь исследовать только непрямые связи между узлами (персонами), условно назовем их скрытыми или неявными связями. Для этого обнулим все значения $G_{ik}$ для тех пар $i$ и $k$, которые связаны непосредственно (полученную матрицу обозначим как $K$). Нас будут интересовать те пары узлов (персон) между которыми нет прямых связей, а коэффициент когезии скрытых связей больше среднего коэффициента когезии всей базы. Для удобного представления последних введем матрицу скрытости $F$ (скрытость – furtive) :

$F = K - G_v$,

где нас будут интересовать только положительные элементы $F$.

На рис. 4 изображена матрица скрытости для реальной базы данных персон, для которой была построена матрица взаимосвязей.

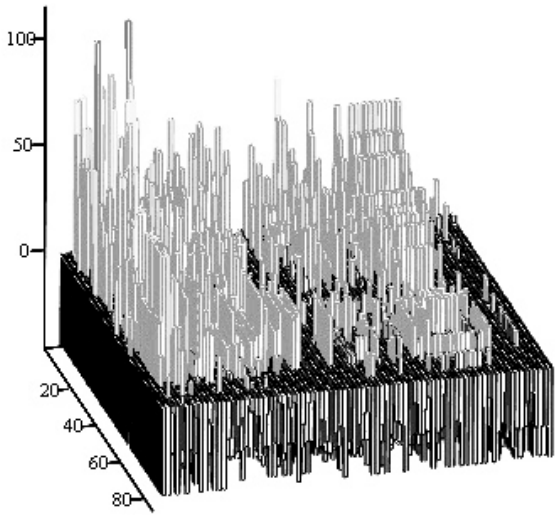



*Рис. 4. Матрица скрытости $F$. По вертикальной оси показан коэффициент когезии, из которого вычтено среднее значение когезии по всей базе, тех пар узлов, между которыми нет непосредственной связи*

Рассмотрим несколько конкретных узлов (персон) из МВП (см. Табл.). Для выбранных персон значения матрицы $A$ равны: $A_{3,3} = 170$, $A_{4,4} = 526$, $A_{9,9} = 234$, $A_{12,12} = 242$, $A_{27,27} = 20$, откуда в частности следует, что максимальное число связей у Ю. Тимошенко. А значения матрицы $M$ равны: $M_{3,4} = 0$, $M_{3,9} = 0$, $M_{3,12} = 0$, $M_{3,27} = 0$, $M_{4,9} = 18$, $M_{4,12} = 26$, $M_{4,27} = 0$, $M_{9,12} = 18$, $M_{9,27} = 0$, $M_{12,27} = 0$. Приведем также коэффициент сцепления для узлов 3 и 9 ($G_{3,9} = 89.4$). Анализируя полученные значения,

можно, в частности, заметить, что между узлами 3 и 9 нет прямой связи. В тоже время коэффициент сцепления равен 89.4, что более чем в два раза выше среднего коэффициента сцепления по МВП.

Табл. Несколько узлов сети персон

| Номер узла | Фамилия человека, соответствующая данному номеру |
|---|---|
| 3 | Басс Д. |
| 4 | Тимошенко Ю. |
| 9 | Яценюк А. |
| 12 | Янукович В. |
| 27 | Кильчицкая И. |

Приведенный метод во многом напоминает подходы, базирующиеся на комбинаторном кластерном анализе, однако его принципиальное отличие в том, что он основывается на законах Максвелла, из которого, в частности, следуют известные закономерности Кирхгофа о протекании электрического тока в разветвленных цепях. При этом не ставилась задача поиска прямых аналогий, а скорее целью было использование методов, уже наработанных в теории электрических сетей.

В отличие от существующих в настоящее время подходов к выявлению взаимосвязей понятий, предложенный метод позволяет выявлять, определять относительный вес и визуализировать неявные связи любых уровней. Следует отметить, что аналоги подобных методов из теории электрических цепей, до сих пор не находили широкого применения в практике аналитической обработки информации.

Вместе с тем рассмотренное направление анализа сложных сетей сегодня актуально в маркетинговых и социальных исследованиях, в конкурентной разведке, в задачах выявления и визуализации различных сообществ.

## Литература


[1] Гаврилова Т.А., Червинская К.Р. Извлечение и структурирование знаний для экспертных систем. - М.: Радио и связь, 1992.

[2] Гершензон Л. М., Ножов И.М. , Панкратов Д. В. Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности // Компьютерная лингвистика и интеллектуальные технологии: труды Международного семинара Диалог'2005. – М.: Наука, 2005.

[3] Григорьев А.Н., Ландэ Д.В. и др. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – Киев: ООО «Старт-98», 2007. – 40 с.

[4] Джексон Дж. Классическая электродинамика - М., Мир, 1965. - 694 с.

[5] Додонов А.Г., Ландэ Д.В. Выявление понятий и их взаимосвязей в рамках технологии контент-мониторинга // Регистрация, хранение и обработка данных, 2006, Т. 8, № 4.– С. 45 - 52.

[6] Калиткин Н.Н., Карпенко Н.В., Михайлов А.П. и др. Математические модели природы и общества –М.: Физматлит, 2005. -360 с.

[7] Попов В.П. Основы теории цепей – М.: Высшая школа, 1985. - 496 с.

[8] Albert R., Jeong H., Barabasi A. Attack and error tolerance of complex networks // Nature. - 2000. - Vol. 406. - pp. 378-382.

[9] Church K.W. , Hanks P. Word association norms, mutual information, and lexicography, Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 1989.

[10] Clauset A., Moore C., Newman M. E. G. Hierarchical structure and the prediction of missing links in networks // Nature. - 2000. - Vol 453. - pp. 98–101.

[11] Danon L. , Dнaz-Guilera A. , Duch J., Arenas A.. Comparing community structure identification // J. Stat. Mech. (2005) P09008. doi: 10.1088/1742-5468/2005/09/P09008 PII: S1742-5468 (05) 07477-7.

[12] Grishman R. Information extraction: Techniques and challenges. In Information Extraction (International Summer School SCIE-97). Springer-Verlag, 1997.

[13] Guiasu, S. Information Theory with Applications, McGraw-Hill, New York, 1977.

[14] Knepper M.M. , Killam R., Fox K.L., Frieder O. Information Retrieval and Visualization using SENTINEL / TREC 1998: 336-340.

[15] Lande D.V., Snarskii A.A. Dynamic network of concepts from web publications // ePrint Arxiv (0806.1439).

[16] Newman M.E.J. The structure and function of complex networks // SIAM Review. - 2003. - Vol. 45. - pp. 167-256.


## Discovering implicit relations of concepts


*A.A. Snarskii, D.V. Lande, M.I. Zhenirovsky, NTUU "KPI", ElVisti IC, Bogolyubov Institute for Theoretical Physics, Kyiv, Ukraine*



The method of discovering implicit relations in complex networks presented by incidence matrixes is described, along with implementation of this method based on electric circuit theory, for relevation of concepts correlations, e.g. persons, derived from unstructured texts.

# СОЦИАЛЬНЫЕ СЕТИ И ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ

# SOCIAL NETWORKS AND DIGITAL LIBRARIES

# Технология поддержки электронных научных публикаций как «живых» документов ♣

© С.И. Паринов

Центральный экономико-математический институт РАН
sparinov@gmail.com

© М.Р. Когаловский

Институт проблем рынка РАН
kogalov@cemi.rssi.ru

## Аннотация

Практика размещения самими учеными результатов своих исследований в форме научных статей и материалов в открытом доступе в сети Интернет постепенно получает организационную поддержку. Все большее распространение в научной среде получают идеи открытых архивов, свободного доступа к результатам исследований, самоархивирования в форме препринтов или постпринтов, а также требования к ученым от гранто- и работодателей по обязательному электронному депонированию всех законченных результатов исследований. Научные статьи и материалы, депонируемые их авторами в электронном репозитории своей организации, являются частью профессиональной информационной среды. Они цитируются наряду с "полноценными" публикациями в рецензируемых журналах. При этом онлайновые средства для электронного депонирования являются общедоступными и достаточно просты в использовании. Как следствие, авторы научных статей и материалов могут вносить в них изменения в общем случае в течение всей своей профессиональной жизни. При массовом использовании подобной практики электронные научные статьи и материалы получают статус "живого" документа (в зарубежной литературе “liquid publication”- «текучая (или неустойчивая) публикация»). Появление научных статей с подобным статусом имеет как положительные, так и отрицательные последствия. В работе сформулированы условия, при которых происходит ослабление негативных и усиление позитивных аспектов превращения научных статей в "живые" документы. Предложенные решения практически реализуются в системе Соционет. Описаны особенности реализации данных решений в виде инфраструктурных онлайновых сервисов, что открывает научному сообществу достаточно разнообразные возможности для их использования.

## 1 Введение

В последние годы у ученых появилась возможность самостоятельно выкладывать (депонировать) свои статьи и материалы (в виде пре- и постпринтов) в открытый доступ благодаря многочисленным электронным библиотекам, онлайновым репозиториям и открытым электронным архивам. Все более распространенной особенностью электронных репозиториев является неограниченная возможность авторов электронных статей вносить в них изменения на протяжении всей своей профессиональной жизни, что превращает эти ресурсы в "живые" документы.

Появление электронных научных документов с подобным статусом имеет как позитивные, так и негативные последствия.

Данная проблема становится актуальной, и ее исследование обладает достаточной новизной как в российской науке, так и в международной. Первая известная нам статья, в которой данная проблема рассматривается в терминах "текучей публикации" (“liquid publication”) была опубликована итальянскими учеными на английском языке в 2007 г. [1, 2]. На русском языке, по-видимому, первое обсуждение данной проблематики в терминах "живого" документа появилось также в 2007 г. в

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2009, Петрозаводск, Россия, 2009.**

журнальной публикации [3] одного из авторов данной статьи.

Отражением актуальности данной темы является выделение гранта 7-й Рамочной Программы ЕС проекту LiquidPub (http://project.liquidpub.org/), а также все более частое упоминание проблематики "living publication" в различных крупных международных проектах по открытому доступу к результатам исследований (что, например, было замечено авторами этой статьи в презентациях на OAI6 - CERN workshop on Innovations in Scholarly Communication, http://indico.cern.ch/conferenceDisplay.py?confId=48321).

В предлагаемой статье содержится анализ данной проблемы, сформулированы условия, при которых, по мнению авторов, происходит ослабление негативных и усиление позитивных аспектов превращения электронных научных материалов в "живые" документы. Предложены практические решения для программной реализации данных инноваций в системе Соционет (http://socionet.ru/). Описаны особенности реализации данных решений в виде инфраструктурных онлайновых сервисов, что открывает научному сообществу достаточно разнообразные возможности их использования.

## 2 "Живые" документы и связанные с ними проблемы

Прежде всего, уточним смысл понятия "живой" документ. В ряде случаев авторы представленных в цифровом виде научных статей, экспериментальных данных или других электронных научных документов, публикуемых в электронных библиотеках или в других свободно доступных онлайновых репозиториях, осуществляют обновление контента этих научных документов на протяжении их жизненного цикла в соответствующей информационной среде. Такие обновляемые электронные научные документы, поддерживаемые в онлайновой информационной среде, будем называть «живыми» документами (living document).

Конечно, не каждый из публикуемых в онлайновых информационных средах электронных научных документов может стать «живым» документом. Не станут такими документами, например, авторефераты или полные тексты уже защищенных диссертационных работ. Многие научные статьи также не попадут в эту категорию. Вместе с тем, имеется целый ряд типов научных документов, которые могут стать «живыми» документами. Приведем лишь несколько примеров:

- обзорная статья в некоторой области науки, которую автор может пожелать актуализировать по прошествии некоторого времени с момента ее публикации в связи с появлением новых значимых результатов в охватываемой данным обзором области
- пополняемая научная библиография по какой-либо тематике исследований
- описание пополняемой музейной коллекции в систематизирующих областях науки, например, ботанического гербария, коллекции насекомых, минералов и т.д.
- отчет, представляющий результаты многоэтапного научного проекта
- опись архивных документов
- математическая статья, автор которой нашел более изящное доказательство теорем по сравнению с приведенными в данной статье.

Поддержка «живых» научных документов, как уже отмечалось, порождает ряд проблем, которые подробно обсуждаются вместе с предлагаемыми подходами к их решению в дальнейших разделах статьи. Здесь мы лишь заметим, что сложность поддержки «живых» документов связана с динамичностью структуры информационной среды, элементами которой они являются. Эта структура образуется явно представленными ссылками между документами, обладающими различными свойствами. Однако семантическая адекватность этих ссылок зависит от состояния документов, на которые направлены эти ссылки. Обычно ссылка на какой-либо документ из другого документа относится к некоторому фрагменту целевого документа. Если этот фрагмент в «живом» целевом документе исчезает при его редактировании автором или его содержание изменяется, ссылка на него становится или может стать семантически неадекватной.

Таким образом, проблема поддержки «живых» документов связана именно со структурными аспектами информационной среды, в которой они существуют. Поддержка этой структуры требует наличия как специальных сервисов в соответствующем онлайновом информационном репозитории, так и в онлайновой научной инфраструктуре, создающей единую научную информационную среду из репозиториев отдельных организаций.

Функциональные требования к таким сервисам и другие технологические вопросы поддержки «живых» документов подробно рассматриваются далее.

## 3 Проблемы поддержки «живых» документов

Массовое использование информационных и интернет-технологий в научно-исследовательской деятельности, кроме прямых и очевидных выгод, иногда порождает ситуации, которые, на первый взгляд, являются проблемными и мешают нормальному функционированию созданных технологий. Подобные проблемы иногда сигнализируют, что данные технологии имеют скрытый потенциал, который также может быть использован, если будут созданы необходимые для этого условия. При этом важно определить когда

эффект от использования подобного потенциала информационных технологий превосходит затраты на создание необходимых дополнительных условий. Следовательно, усилия по созданию необходимых условий имеют смысл и будут оправданы.

Примером подобной ситуации является появившаяся в последние годы возможность для ученых самостоятельно выкладывать свои статьи и материалы (в виде пре- и постпринтов) в открытый доступ через многочисленные онлайновые репозитории и открытые электронные архивы. Явным позитивным следствием этого является ускорение кругооборота научных знаний. Статистически зафиксировано [4], что подобная практика увеличивает количество цитирований работ ученых примерно в два раза. Оборотная сторона - у ученых появляется неограниченная возможность вносить изменения в электронные версии своих статей на протяжении всей своей профессиональной жизни, что превращает такие научные статьи в «живые» документы.

Появление у научных статей и материалов статуса "живых" документов имеет для научного сообщества, как минимум, следующие важные последствия:

• ученый в ответ на критику или по собственной инициативе может вместо создания новой статьи доработать свой исходный текст и внести исправления в текущую версию сделанного им электронного депонента данной статьи; в результате данная статья, формально представляя собой информационный объект с тем же идентификатором (URI, URL или "handle" в Соционет) в электронном информационном пространстве, содержательно будет существенно отличаться от предыдущей версии;

• как результат, все связи цитирования, уже установленные с этой статьей из других научных материалов могут потерять свою актуальность и должны быть пересмотрены их авторами как содержательно (может измениться, или исчезнуть объект цитирования), так и технически (объект цитирования может исчезнуть, или переместиться в другую часть статьи).

Очевидное административное решение данной проблемы путем запрещения изменений в уже депонированных электронных научных документах (т.е. их новые версии заведомо получают статус нового электронного документа), фактически, воспроизводит в электронной среде традиционные принципы бумажных научных публикаций. Таким образом, остаются не исследованными и не использованными потенциальные выгоды от превращения научных статей в "живые" документы.

Сохранение в открытом доступе предыдущих версий научных статей и использование механизма поддержки многоверсионности документов, которым располагают некоторые широко распространенные инструментальные средства для разработки электронных библиотек, является, безусловно, полезным в данном контексте. Однако поддержка связей между версиями документов является, на наш взгляд, недостаточной, т.к. для использования потенциальных выгод от превращения научных статей в "живые" документы требуется более тонкий механизм, позволяющий создавать, отслеживать и поддерживать сети разнокачественных связей между фрагментами исходных и цитируемых документов. В первую очередь здесь представляют интерес сети связей между результатами исследований различных ученых, формирующиеся вследствие взаимного научного цитирования.

Предполагается, что создаваемый в среде системы Соционет механизм поддержки "живых" документов, позволит сделать более эффективным процесс использования результатов исследований научным сообществом и кругооборота научных знаний.

Использование научным сообществом положительных моментов, связанных с "живыми" документами, означает попытку усовершенствовать профессиональные взаимодействия между учеными, которые в данном конкретном случае облечены в форму написания и редактирования статей, а также их цитирования в других статьях, которые в свою очередь тоже могут изменяться.

Наиболее важные, на наш взгляд, потенциальные выгоды для научного сообщества от полного и эффективного использования феномена "живого" документа связаны с созданием качественно новых условий для функционирования сети профессиональных связей между учеными. Подобные связи возникают, динамически меняются или устойчиво существуют между учеными в процессе создания ими нового научного знания за счет специализации ученых, определенного разделения труда между ними и использования ими результатов друг друга. Переход в научной практике от традиционной статьи к "живому" документу означает в этом контексте существенное повышение эффективности сети профессиональных связей, выполняющей важную роль в процессе современного кругооборота научных знаний.

Поддержка статуса «живых» документов для научных статей требует определенных доработок и развития применяемых сейчас в исследовательских организациях электронных библиотек, онлайновых репозиториев и открытых архивов, включая встроенные в них алгоритмы электронного депонирования и цитирования статей и материалов. Кроме этого, требуется создание новых и развитие некоторых из уже существующих онлайновых служб и сервисов во внешней по отношению к исследовательским организациям среде, которые в данном проекте рассматриваются как элементы научной онлайновой инфраструктуры. Система Соционет [7] является хорошим объектом для реализации этого, т.к. содержит как средства электронного депонирования и создания институтских репозиториев, так и представляет

собой пример реализации онлайновой научной инфраструктуры.

Важная и сложная проблема в данном случае – определение возможного комплексного эффекта от превращения научных статей в «живые» документы и сопоставление его с затратами на создание программно-технических условий для поддержки «живых» документов в практической деятельности научного сообщества.

## 4 Условия эффективного использования «живых» документов

Одна из задач - спецификация набора информационных технологий, при которых научное сообщество может получить максимально возможную выгоду от превращения научных статей в «живые» документы. И в последующем - создание пилотных образцов данных технологий. Проведение их опытной эксплуатации авторы планируют в рамках программы «Открытый доступ к результатам исследований» институтов Отделения общественных наук РАН.

Ситуация превращения научной статьи в «живой» документ может быть представлена следующим образом:

- научная статья, обозначим ее «СТ1», выложена в электронном виде в открытый доступ ее автором «А1» (или другим лицом, являющимся представителем автора);
- научная статья «СТ2», выложенная в открытый доступ автором «А2», цитирует содержание «СТ1» (в общем случае цитат может быть несколько, и цитироваться может любое количество статей).

Превращение научных статей в «живые» документы не приведет к нарушению процесса научных исследований при наличии комплекса онлайновых сервисов, обеспечивающих следующие возможности:

- автор А1 может получить список всех статей, в которых цитируется его статья СТ1, а также может включить уведомления о появлении новых цитирований ее или исключении ранее созданных;
- автор А1 при изменении электронной версии статьи СТ1 получает автоматическое уведомление о списке существующих на текущий момент связей цитирования с его статьей СТ1, которые могут быть нарушены в результате его действий по изменению СТ1;
- автор А2 получает автоматическое уведомление о необходимости проверки в его статье СТ2 связи цитирования со статьей СТ1 ввиду того, что автор А1 внес изменения в СТ1 (информирование о возможно необходимом обновлении цитат);
- автор А2 может получить список текущих изменений в СТ1 по сравнению с предыдущей версией СТ1, для которой А2 создал в своей статье СТ2 связь цитирования со статьей СТ1;
- читатели статьи СТ1 должны видеть список всех связей цитирования, которые установлены из других статей со СТ1;
- читатели СТ1 должны видеть для каких связей цитирования с СТ1 не было выполнено обновление цитат после изменения СТ1;
- читатели статьи СТ2 должны видеть предупреждение, если автором А2 не было выполнено обновление цитат на статью СТ1 после изменения СТ1.

Подобные сервисы создадут условия для оперативной координации между авторами статей, связанных друг с другом ссылками цитирования, что обеспечит непрерывную цепную реакцию согласований и обновлений в содержании статей, представляющих собой «живые» документы.

Читатели статей, имеющих статус «живого» документа, также будут знать, какие ссылки цитирования в них гарантированно соответствуют их текущему содержанию, а какие возможно устарели и могут быть проигнорированы.

В системе Соционет существуют все необходимые предпосылки для реализации данного набора сервисов, включая средства для электронного депонирования (размещение статей в открытом доступе), а также инструменты электронного цитирования (создание связей цитирования между статьями, размещенными в электронных библиотеках, открытых архивах или электронных репозиториях).

## 5 Реализация поддержки "живых" документов в системе Соционет

Для практической проверки описанного подхода осуществляется разработка пилотной версии необходимого программного обеспечения, которое может быть свободно использовано научным сообществом для извлечения потенциальных выгод от превращения статей в "живые" документы.

Создаваемое программное обеспечение входит в состав системы Соционет (http://socionet.ru/) и представляет собой развитие уже существующих в Соционет средств электронного депонирования научных статей и материалов, а также совершенствование уже существующих процедур формирования и контроля связей электронного цитирования.

На данный момент система Соционет в целом представляет собой комплекс следующих подсистем:

1. *Информационный хаб*. Эта подсистема интегрирует в стандартизованную базу данных метаданные открытых архивов и тематических коллекций, принадлежащих различным научным организациям, а также предоставляет по нескольким популярным протоколам все собранные метаданные в стандартизованном виде для их внешнего использования в целях обеспечения возможности внешним разработчикам создавать на их основе новые тематические ресурсы и сервисы.

2. *Профессиональное информационное пространство*, которое предлагает пользователям средства навигации и поиска по содержанию всех метаданных, собранных информационным хабом Соционет.

3. *Онлайновое рабочее место ученого* (в терминах системы – "личная зона Соционет"), которое предоставляет пользователям разнообразные возможности для электронного депонирования результатов исследований, создания коллекций электронных материалов, создания связей между материалами, управления открытыми архивами организаций и т.п.

4. *Персональный информационный робот*, входящий в онлайновое рабочее место, который позволяет ученому в автоматическом режиме отслеживать появление в информационном пространстве Соционет интересных для него материалов.

5. *Профессиональная социальная сеть*, которая визуализирует различные типы взаимозависимостей между электронными материалами, заданные их авторами.

6. *Подсистема представления информационных ресурсов научных организаций* в виде Открытых Архивов (ОА) по протоколу OAI-PMH, который стал стандартом де-факто для распространения материалов в международном научном сообществе.

7. *Фрагмент наукометрической сигнальной системы*, аккумулирующей статистику просмотров метаданных материалов, загрузок их полных текстов и развития сети связей между материалами для расчета и визуализации показателей востребованности, использования и активности по отношению к авторам материалов, научным организациям и другим информационным объектам [6], [8].

Создаваемые в рамках перечисленных выше подсистем средства поддержки статуса «живых» документов должны информировать:

• автора изменяемой статьи, которая цитируется в других статьях, какие связи и цитаты установлены с его статьей и что из них он нарушает в момент ее изменения;

• авторов статей с цитатами из изменяемой статьи о сделанных изменениях в цитируемой статье, для принятия решений о корректировке соответствующих связей цитирования;

• читателей электронных статей о том, что определенные цитаты в них могут быть нарушены.

В связи с этим для создания пилотного образца предлагаемых онлайновых сервисов осуществляется следующая модернизация системы Соционет:

1. Расширение набора атрибутов связей цитирования, включая временные метки создания и обновлений связей цитирования. Развитие модели и алгоритмов электронного цитирования в целом, что включает в себя внедрение в научную практику использования учеными качественных характеристик связей цитирования, а также систему мониторинга связей цитирования и оповещений об их нарушении или об изменении свойств связей. Подробнее о развитии модели электронного цитирования см. в [5].

2. Модификация структуры базы данных, содержащей сведения о связях между информационными объектами системы.

3. Создание сервисов автоматического мониторинга событий, связанных с изменениями в научных статьях, включая изменение набора и атрибутов связей цитирования.

4. Создание системы генерации отчетов для читателей научных статей и автоматических уведомлений для авторов статей, связанных через цитирование.

Организационной поддержкой данных работ является комплекс следующих мероприятий, проводимых в Отделении общественных наук (ООН) РАН в рамках программы "Открытый доступ к результатам исследований":

• все исследовательские организации ООН РАН создают онлайновые открытые архивы, интегрированные в информационное пространство Соционет и совместимые с международными системами распространения научных материалов (75% институтов ООН РАН на апрель 2009 г. сформировали содержание своих ОА);

• исследовательские организации обязывают своих научных сотрудников депонировать в институтских ОА результаты всех открытых исследований (один из институтов – ЦЭМИ РАН – в апреле 2007 г. ввел в действие данное положение приказом директора);

• использование онлайновой наукометрии при определении персональных научных надбавок исследователей, первая версия наукометрической сигнальной системы функционирует с 01.01.2007, она собирает данные о востребованности статей и материалов из институтских ОА, рассчитывает для публичного просмотра ряд наукометрических показателей, характеризующих статьи, их авторов и научные организаций.

## 6 Заключение

Анализ рассматриваемой проблемы, на наш взгляд, показал реальность создания программных средств, способных ослабить (если не исключить полностью) негативные моменты от превращения научных статей в "живые" документы, а также усилить возможные позитивные последствия данного объективного процесса развития научно-исследовательской среды.

Превращение научных статей в "живые" документы даст заведомо положительный результат для научного сообщества при сочетании двух факторов: удачного программно-технического воплощения данной системы и ее массового применения исследователями. Затраты на реализацию этих факторов, на наш взгляд, являются оправданными, т.к. ожидаемая отдача от этих усилий достаточно весома.

Для исследователей это -

– получение оперативных сигналов о новых цитированиях материалов ученого, удалений уже существующих цитирований, а также об изменении цитируемых результатов исследований;

- возможность быть в курсе развития/изменения научных результатов, которые используются данным ученым в своей работе, а также автоматически оповещать ученых, которые используют результаты твоих исследований, о развитии/улучшении данных результатов;

- наличие комплексной картины, включая историю того, кто, когда и зачем (при внедрении модели цитирования с качественными атрибутами [5]) цитировал материалы ученого и т.п.

Для научного сообщества –

- более высокий уровень информированности ученых о появлении новых результатов исследований;

- новые стимулы и лучшие условия для развития собственных результатов исследований как следствие новых результатов у других ученых;

- улучшение среднего уровня использования результатов исследований в научном сообществе;

- повышение степени профессиональной связанности ученых, использующих результаты друг друга, и, как следствие, ускорение процессов создания нового научного знания.

При массовом использования технологий поддержки "живых" документов просматриваются перспективы превращения научной статьи в элемент профессиональной социальной сети ученого, в которой связи цитирования между статьями становятся поводом для устойчивых профессиональных взаимодействий между исследователями. Как следствие, возможно превращение корпуса научных результатов во множество взаимосвязанных "живых" документов, когда изменения в одном документе могут порождать цепную реакцию изменений в связях цитирования и изменение содержания других "живых" документов.

## Литература


[1] Fabio Casati, Fausto Giunchiglia, Maurizio Marchese. Publish and perish: why the current publication and review model is killing research and wasting your money, ACM Ubiquity 8 (3), Feb 2007. http://www.acm.org/ubiquity/views/v8i03_fabio.html

[2] Fabio Casati, Fausto Giunchiglia, Maurizio Marchese. Liquid Publications: Scientific Publications meet the Web, Version 2.3, October 1, 2007, http://liquidpub.org/attachment/wiki/WikiStart/LiquidPub%20paper-latest.pdf

[3] Паринов С.И. e-Science - онлайновое будущее науки. // Информационные технологии. - 2007. - № 9. Приложение.

[4] Стивен Харнад. Максимизация научного эффекта через институциональные и национальные обязательства самоархивирования для открытого доступа. Постпринт, 2006. http://socionet.ru/publication.xml?h=repec:rus:mqijxk:10

[5] Паринов С.И. Новый подход к оценке результатов научно-исследовательской деятельности. Соционет: электронный депонент, 2008, http://socionet.ru/publication.xml?h=repec:rus:mqijxk:20

[6] Когаловский М.Р., Паринов С.И. Информационные ресурсы, наукометрические показатели и показатели качества метаданных системы Соционет. Труды девятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции, Переславль, 2007». – Переславль-Залесский: Изд. «Университет города Переславля», 2007.

[7] Паринов С.И., Ляпунов В.М., Пузырев Р.Л. Система Соционет как платформа для разработки научных информационных ресурсов и онлайновых сервисов //Электронные библиотеки. – 2003. - Том 6. - Выпуск 1.

[8] Когаловский М.Р., Паринов С.И. Метрики онлайновых информационных пространств. Журнал Экономика и математические методы, 2008, т. 44, №2, с. 108-120


## An approach to support electronic research publications as «living» documents


Sergey Parinov and Mikhail Kogalovsky



Currently popular researchers' practice of research results self-archiving in institutional repositories results in researchers' ability to modify text of their publications during a long time. In this case a research publication becomes a living document (or "liquid publication" [1]). In the article we analyze possible benefits of it for a research community and necessary technical conditions to get the benefits. We discuss technical solutions as development of the Socionet system (http://socionet.ru/).



♣ Данная работа поддержана грантами РФФИ проект 09-07-00378 и РГНФ проект 09-02-12117-в

# Изучение профилей русскоязычных сообществ LiveJournal: региональный аспект

© А.В. Сычев

Воронежский государственный университет
sav@cs.vsu.ru

**Аннотация**

В докладе представлены результаты исследования на основе данных из профилей сообществ блог-хостинга *LiveJournal*. Рассматриваются такие задачи как реконструкция хронологии создания сообществ, вычисление корреляции между атрибутами профилей. Наибольшее внимание в работе было уделено интересам сообществ. Был изучен характер распределения интересов в профилях, как путем расчета простейших числовых характеристик, так и с помощью процедуры кластеризации. Для сравнения приведены результаты исследования регионального распределения читателей в профилях сообществ.

## 1 Введение

Социальные сети в WWW, привлекающие сегодня к себе всеобщее внимание пользователей Интернета, сформировались за очень короткий промежуток времени (буквально 5 - 8 лет). Они объединяют в себе блоги (сетевые дневники), сети медиа-ресурсов, сети персональной информации (MySpace, LinkeIn, Facebook, «В контакте.ru», «Одноклассники.ru» и другие), системы закладок (del.icio.us), wiki-энциклопедии и другие. Количество пользователей в этих сетях увеличивается с беспрецедентной скоростью, вызывая вполне обоснованный интерес к ним у представителей науки, бизнеса и ИТ-индустрии.

Социальные сети фактически представляют собой гигантское хранилище общедоступной информации, в первую очередь, персонального характера. Анализ этой информации может дать ценные сведения о структуре современного информационного общества и процессах, протекающих в нем.

За рубежом данная область является предметом интенсивных исследований, что подтверждается многочисленными публикациями по этой тематике. Направление «Социальные сети» было выделено отдельно в рамках такой авторитетной международной конференции как *World Wide Web* в 2008 и 2009 годах.

Хороший обзор актуального состояния исследований, подходов, инструментов и приложений, связанных с блогосферой, приведен в [1]. В этой работе рассмотрены следующие направления исследований: моделирование блогосферы, кластеризация, сбор данных, влияние и распространение информации, доверие и репутация, фильтрация спам-блогов.

Результаты исследований на примере конкретных коллекций данных приведены в работах [2-5]. В [5] представлены результаты широкомасштабного исследования на основе данных из 4 популярных социальных сетей *Flickr*, *YouTube*, *LiveJournal* и *Orkut*. Общий объем данных составил 11.3 млн. пользователей и 328 млн. связей между ними. Как показало исследование, социальные сети структурно отличаются от ранее исследованных сетей, в частности Веб. Для них характерна существенно большая доля симметричных связей и гораздо более высокий уровень локальной кластеризации. В [3] была исследована и проведена экспериментальная оценка (на основе данных из базы DBLP) предложенной авторами модели кластеризации, использующей помимо атрибутов объектов информацию о связях между ними. В [4] предложена методика FacetNet, основанная на байесовской оценке максимального правдоподобия, комбинирующая задачи выявления сообществ и их эволюции в рамках единого процесса. Для экспериментальной оценки была использована база DBLP. Феномен микроблоггинга на примере социальной сети *Twitter* рассмотрен в [2]. В этой работе приведены результаты исследования топологических и географических характеристик этой сети.

Русскоязычная блогосфера исследована в очень малой степени, при этом публикации, содержащие серьезные структурные исследования этой социальной сети, практически отсутствуют. Отдельные аспекты русскоязычной блогосферы

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

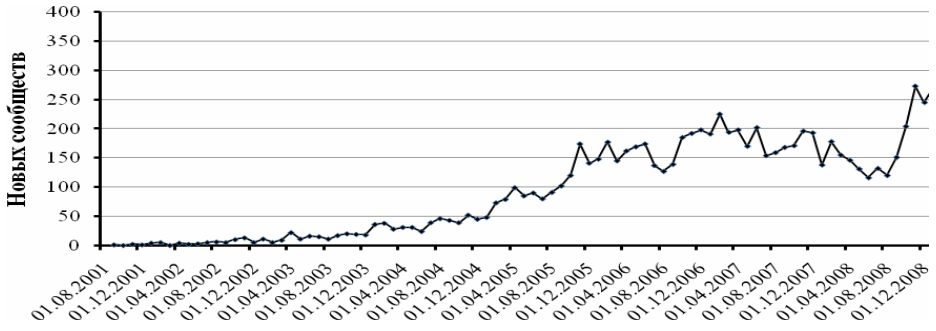



**Рисунок 1**. Хронология создания сообществ.

**Таблица 1**. Усредненные значения атрибутов профилей русскоязычных сообществ *LiveJournal*.

| **Атрибут** | **1** | **2** | **3** | **4** | **5** | **10** | **11** | **12** | **13** | **14** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Значение** | 22 | 1,5 | 0,4 | 131 | 126 | 219 | 10 | 69 | 1,0 | 1763 |

**Таблица 2**. Корреляция между атрибутами профилей сообществ.

| | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 0,09 | 0,12 | 0,12 | 0,12 | -0,06 | -0,07 | 0,24 | 0,04 | 0,06 | 0,21 | 0,23 | 0,21 | 0,04 |
| **2** | | 0,21 | 0,17 | 0,17 | -0,11 | -0,12 | 0,12 | 0,04 | 0,15 | 0,12 | 0,17 | 0,12 | 0,15 |
| **3** | | | 0,14 | 0,16 | 0,01 | 0,02 | 0,09 | 0,05 | 0,04 | 0,11 | 0,10 | 0,11 | 0,04 |
| **4** | | | | 0,95 | -0,13 | -0,29 | 0,13 | 0,10 | 0,72 | 0,18 | 0,13 | 0,09 | 0,63 |
| **5** | | | | | -0,13 | -0,27 | 0,13 | 0,09 | 0,59 | 0,20 | 0,15 | 0,10 | 0,51 |
| **6** | | | | | | 0,61 | 0,14 | -0,03 | -0,09 | -0,01 | -0,21 | 0,05 | -0,06 |
| **7** | | | | | | | 0,22 | -0,06 | -0,21 | -0,02 | -0,29 | 0,03 | -0,17 |
| **8** | | | | | | | | 0,03 | 0,08 | 0,14 | 0,26 | 0,19 | 0,05 |
| **9** | | | | | | | | | 0,13 | 0,02 | 0,02 | 0,04 | 0,15 |
| **10** | | | | | | | | | | 0,15 | 0,05 | 0,06 | 0,90 |
| **11** | | | | | | | | | | | 0,17 | 0,14 | 0,14 |
| **12** | | | | | | | | | | | | 0,16 | 0,02 |
| **13** | | | | | | | | | | | | | 0,05 |

регулярно освещаются в отчетах компании Яндекс (раздел портала “Исследования Яндекса”, http://company.yandex.ru/researches/), например в [6].

В работе [7] уже были представлены результаты расчета характеристик для сообществ блог-хостингов *LiveJournal* и *LiveInternet*, полученных на основе атрибутов профилей сообществ. Там же приведены графики и таблицы, полученные в результате кластеризации сообществ и их интересов.

В данной работе приведены новые результаты исследования профилей русскоязычных сообществ блог-хостинга *LiveJournal* (www.livejournal.com). Во-первых, было существенно увеличено количество сообществ - вместо прежних примерно трех тысяч русскоязычных сообществ *LiveJournal* - теперь для исследования были использованы 8870 профилей сообществ. Во-вторых, появился новый атрибут – *регион*. Так в 1898 профилях в качестве региона была указана Российская федерация, в 575 – бывший СССР, а в остальных – регионы РФ.

Ввиду ограничений поискового сервиса *LiveJournal*, удалось получить неполный список сообществ по РФ и Москве (только первые 2000). По всем остальным регионам списки сообществ были получены полностью. Данное ограничение, безусловно, вносит определенные коррективы при анализе распределений атрибутов профилей и их характеристик, но, как представляется, несущественно влияет на результаты анализа связей между всеми остальными регионами РФ.

Целями данного исследования являются определение того, насколько информативным может быть анализ только данных из профилей сообществ, исследование региональной специфики русскоязычных сообществ *LiveJournal* и изучение возможности использования данных из профилей сообществ для анализа связей между регионами.

**Таблица 3**. Рейтинг регионов РФ по числу сообществ (представлены первые 10 регионов)

| Регион | Сообществ | Среднее значение атрибута профиля сообщества | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **10** | **11** | **12** | **13** | **14** |
| 0 | 1898 | 31 | 1,8 | 0,6 | 345 | 331 | 0,5 | 18.06.2007 | 18.02.2009 | 593 | 24 | 119 | 1,4 | 5397 |
| 1 | 1733 | 20 | 1,5 | 0,3 | 79 | 76 | 0,4 | 27.09.2006 | 20.06.2008 | 108 | 6 | 67 | 0,9 | 442 |
| 2 | 1146 | 30 | 1,7 | 0,6 | 164 | 162 | 0,6 | 10.08.2007 | 15.01.2009 | 289 | 17 | 92 | 1,3 | 2321 |
| 3 | 575 | 26 | 1,6 | 0,3 | 139 | 133 | 0,2 | 06.04.2006 | 07.03.2008 | 159 | 7 | 73 | 0,8 | 1582 |
| 4 | 294 | 15 | 1,4 | 0,2 | 38 | 38 | 0,3 | 22.08.2006 | 15.01.2008 | 67 | 3 | 34 | 0,7 | 429 |
| 5 | 241 | 15 | 1,4 | 0,3 | 24 | 23 | 0,6 | 17.04.2007 | 16.05.2008 | 52 | 3 | 38 | 0,8 | 258 |
| 6 | 195 | 16 | 1,4 | 0,2 | 31 | 30 | 0,4 | 17.10.2006 | 09.03.2008 | 54 | 4 | 34 | 0,9 | 201 |
| 7 | 190 | 13 | 1,5 | 0,2 | 28 | 29 | 0,5 | 05.12.2006 | 11.04.2008 | 36 | 3 | 36 | 0,7 | 120 |
| 8 | 186 | 13 | 1,5 | 0,3 | 39 | 32 | 0,5 | 09.11.2006 | 15.01.2008 | 82 | 4 | 32 | 0,6 | 650 |
| 9 | 156 | 15 | 1,4 | 0,2 | 42 | 43 | 0,4 | 28.09.2006 | 21.04.2008 | 72 | 3 | 45 | 0,9 | 240 |
| 10 | 145 | 14 | 1,4 | 0,3 | 30 | 30 | 0,5 | 13.01.2007 | 30.04.2008 | 43 | 4 | 43 | 0,8 | 106 |

**Таблица 4**. Корреляция между количеством сообществ в регионе и средним значением атрибутов профиля сообщества. В первой строке учтены все регионы, включая “РФ”, во второй – исключен регион “РФ”, в третьей – исключены “РФ”, Москва, Санкт-Петербург и бывший СССР.

| **атрибут** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **10** | **11** | **12** | **13** | **14** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **корреляция** | 0,42 | 0,44 | 0,20 | 0,82 | 0,80 | -0,11 | -0,05 | 0,34 | 0,77 | 0,70 | 0,48 | 0,25 | 0,76 |
| | 0,32 | 0,34 | 0,11 | 0,67 | 0,63 | -0,16 | -0,12 | 0,19 | 0,55 | 0,49 | 0,34 | 0,17 | 0,53 |
| | 0,08 | 0,34 | -0,07 | 0,32 | 0,25 | -0,27 | -0,26 | -0,07 | 0,31 | 0,23 | 0,10 | 0,02 | 0,28 |

## 2 Исследование атрибутов профилей русскоязычных сообществ LiveJournal

### 2.1 Хронология создания сообществ

На графике, представленном на рисунке 1, отражена хронология формирования русскоязычных сообществ в блог-хостинге LiveJournal. На нем четко просматриваются сезонные колебания скорости появления новых сообществ с характерными весенними пиками и летними спадами

### 2.2 Атрибуты профилей сообществ

Также было проведено изучение характеристик атрибутов профилей сообществ. В таблице 1 приведены усредненные значения для атрибутов профилей, а в таблице 2 - корреляция между атрибутами профилей сообществ. Номерами в них обозначены атрибуты: 1 - количество интересов, 2 - количество смотрителей, 3 - количество модераторов, 4 - количество участников, 5 - количество читателей, 6 - тип аккаунта, 7 - дата создания, 8 - дата последнего обновления, 9 - количество вирт. подарков, 10 - количество записей, 11 - количество меток, 12 - количество избранных записей, 13 - количество картинок пользователя, 14 - количество полученных комментариев.

Анализ корреляции между атрибутами профилей сообществ (таблица 2) показал тесную связь между количеством участников (4) и количеством читателей (5), между количеством записей (10) и количеством комментариев на них (14).

Прослеживается также связь между количеством участников (читателей) и количеством записей (10) и комментариев (14) в сообществах.

### 2.3 Региональное распределение сообществ и характеристик их атрибутов

В таблице 3 приведены количество сообществ и средние значения атрибутов профилей для первой десятки в рейтинге регионов (по числу сообществ): 0 - РФ, 1 - Санкт-Петербург, 2 - Москва, 3 - бывший СССР, 4 - Новосибирская область, 5 - Московская область, 6 - Нижегородская область, 7 - Республика Татарстан, 8 - Самарская область, 9 - Свердловская область, 10 - Ростовская область.

Значения корреляции между количеством сообществ в регионе и средними значениями атрибутов профилей этих сообществ представлены в таблице 4. Обращает на себя внимание заметная зависимость средних значений числа читателей, участников, записей и комментариев от числа сообществ в регионе.

## 3. Кластеризация регионов по интересам

Как представляется, наиболее информативным является анализ распределения интересов и

пользователей (читателей), указанных в профилях сообществ, по регионам.

В исследовании не ставилась задача поиска наиболее эффективных для данной задачи алгоритмов кластеризации регионов по сообществам, скорее преследовалась цель получения группировок регионов в первом приближении, поэтому выбор конкретного алгоритма был достаточно произвольным.

В работе был использован метод аггломеративной кластеризации Ланса-Уильямса. Первичное расстояние между сообществами рассчитывалось по формуле:

$$\rho(c_1, c_2) = \frac{|c_1 \cap c_2|}{\sqrt{|c_1|} \cdot \sqrt{|c_2|}}$$

Регион $c_i$ рассматривался как множество интересов, указанных в профилях сообществ этого региона. При проведении процедуры кластеризации расстояние между кластерами рассчитывалось по формуле среднего расстояния.

В качестве исходных данных для процедуры кластеризации регионов была использована матрица "*регион-интерес*", из которой была сформирована матрица "*регион-регион*".

На основе матрицы "*регион-регион*" были получены кластеры регионов при различных значениях порога кластеризации $Th = 1/2^k$, $k = 0,1,2,3,4,5,6$.

**Таблица 5**. Невырожденные кластеры регионов (по интересам) при значении порога $Th = 0.125$

| Кластер | Регион | Сообществ | Интересов |
|---|---|---|---|
| 6 | Астраханская область | 20 | 297 |
| | Псковская область | 17 | 124 |
| 7 | Башкортостан Республика | 91 | 1066 |
| | Свердловская область | 156 | 1929 |
| 13 | Вологодская область | 26 | 408 |
| | Новосибирская область | 295 | 3794 |
| | Самарская область | 186 | 2010 |
| 35 | Москва | 1165 | 22571 |
| | Санкт-Петербург | 1737 | 22596 |
| 38 | Нижегородская область | 195 | 2530 |
| | Ростовская область | 145 | 1790 |
| 44 | Пермский край | 103 | 1104 |
| | Чувашская Республика | 16 | 173 |
| 45 | Приморский край | 88 | 1205 |
| | Хабаровский край | 49 | 783 |

**Таблица 6**. Невырожденные кластеры регионов (по интересам) при значении порога $Th = 0.0625$.

| Кластер | Регион | Сообществ | Интересов |
|---|---|---|---|
| 3 | Алтайский край | 22 | 242 |
| | Бурятия Республика | 6 | 29 |
| 5 | Архангельская область | 22 | 196 |
| | Астраханская область | 20 | 297 |
| | Башкортостан Республика | 91 | 1066 |
| | Белгородская область | 25 | 221 |
| | Владимирская область | 17 | 305 |
| | Волгоградская область | 53 | 483 |
| | Вологодская область | 26 | 408 |
| | Воронежская область | 96 | 1015 |
| | Ивановская область | 12 | 77 |
| | Иркутская область | 105 | 894 |
| | Калининградская область | 110 | 1679 |
| | Карелия Республика | 18 | 387 |
| | Кемеровская область | 40 | 386 |
| | Костромская область | 9 | 115 |
| | Краснодарский край | 112 | 1483 |
| | Курская область | 9 | 44 |
| | Липецкая область | 17 | 154 |
| | Москва | 1165 | 22571 |
| | Московская область | 241 | 3034 |
| | Нижегородская область | 195 | 2530 |
| | Новосибирская область | 295 | 3794 |
| | Омская область | 86 | 678 |
| | Приморский край | 88 | 1205 |
| | Псковская область | 17 | 124 |
| | Ростовская область | 145 | 1790 |
| | Рязанская область | 42 | 637 |
| | Самарская область | 186 | 2010 |
| | Санкт-Петербург | 1737 | 22596 |
| | Саратовская область | 66 | 814 |
| | Сахалинская область | 16 | 238 |
| | Свердловская область | 156 | 1929 |
| | Ставропольский край | 122 | 1072 |
| | Татарстан Республика | 190 | 2109 |
| | Томская область | 43 | 781 |
| | Тюменская область | 61 | 784 |
| | Удмуртская Республика | 34 | 455 |
| | Хабаровский край | 49 | 783 |
| | Челябинская область | 122 | 1443 |
| 13 | Курганская область | 7 | 79 |
| | Чукотский автономный округ | 1 | 5 |
| 16 | Марий Эл Республика | 8 | 164 |
| | Новгородская область | 29 | 351 |
| 22 | Пермский край | 103 | 1104 |
| | Чувашская Республика | 16 | 173 |

**Таблица 7**. Список интересов, указанных в профилях сообществ из всех регионов, входящих в кластер *№45* при значении порога *Th=0.125*.

| № | Интерес | № | Интерес | № | Интерес | № | Интерес |
|---|---|---|---|---|---|---|---|
| 1 | 35 мм | 38 | дизайн | 74 | мужчины | 111 | репортаж |
| 2 | art | 39 | дожди | 75 | музыка | 112 | ресторан |
| 3 | arthouse | 40 | дружба | 76 | музыканты | 113 | Россия |
| 4 | canon | 41 | женщины | 77 | набережная | 114 | самореализация |
| 5 | depth of field | 42 | живопись | 78 | наркотики | 115 | секс |
| 6 | foto | 43 | жизнь | 79 | Находка | 116 | семья |
| 7 | linux | 44 | загадки | 80 | негатив | 117 | слайд |
| 8 | minolta | 45 | здоровье | 81 | недвижимость | 118 | сон |
| 9 | nikon | 46 | зенит | 82 | новости | 119 | спорт |
| 10 | olympus | 47 | зеркалка | 83 | ночные клубы | 120 | старый город |
| 11 | pentax | 48 | игра | 84 | общение | 121 | стихи |
| 12 | photo | 49 | иллюстрации | 85 | огонь | 122 | счастье |
| 13 | photography | 50 | интересные люди | 86 | Оле Нидал | 123 | съемка |
| 14 | pinhole | 51 | интернет | 87 | оптика | 124 | творчество |
| 15 | slr | 52 | информация | 88 | отдых | 125 | театр |
| 16 | unix | 53 | искусство | 89 | панк | 126 | тигры |
| 17 | автомобили | 54 | история | 90 | пиво | 127 | трамваи |
| 18 | алкоголь | 55 | Карма Кагью | 91 | пленка | 128 | троллейбусы |
| 19 | архитектура | 56 | Кармапа | 92 | политика | 129 | учеба |
| 20 | барабаны | 57 | катастрофа | 93 | помощь | 130 | философия |
| 21 | бесплатно | 58 | кафе | 94 | портрет | 131 | флирт |
| 22 | брусчатка | 59 | кино | 95 | портфолио | 132 | фонтаны |
| 23 | буддизм | 60 | Китай | 96 | Поэзия | 133 | фото |
| 24 | Буддизм | 61 | классика | 97 | поэзия | 134 | фотоаппарат |
| 25 | бытовая техника | 62 | книги | 98 | Приморье | 135 | фотовыставки |
| 26 | вино | 63 | книги по фотографии | 99 | Природа | 136 | фотографии |
| 27 | Владивосток | 64 | компьютеры | 100 | природа | 137 | фотография |
| 28 | водка | 65 | кофе | 101 | провинция | 138 | Фотография |
| 29 | встречи | 66 | красота | 102 | программирование | 139 | фотожурналистика |
| 30 | выдержка | 67 | лето без воды | 103 | прогулки | 140 | фотопечать |
| 31 | гламур | 68 | литература | 104 | проза | 141 | художники |
| 32 | город | 69 | любовь | 105 | психология | 142 | цвет |
| 33 | горячая вода | 70 | люди | 106 | путешествия | 143 | цифровая фотография |
| 34 | графика | 71 | макро | 107 | работа | 144 | черно-белая фотография |
| 35 | гулять | 72 | медитация | 108 | радость | 145 | шашлыки |
| 36 | деньги | 73 | море | 109 | развлечение | 146 | Япония |
| 37 | диафрагма | | | 110 | реинкарнация | | |

В таблицах 5 и 6 приведены невырожденные кластеры (т.е. содержащие более одного элемента) регионов, в которых расстояние между регионами (отражающее общность интересов, указанных в профилях сообществ) оказалось не ниже порога ***Th***.

Первые невырожденные кластеры образуются при значении ***Th*** = 0.125. Далее по мере снижения порога происходит резкое расширение единственного суперкластера при уменьшении общего количества невырожденных кластеров, что и отражено в таблице 8.

**Таблица 8**. Характеристики невырожденных кластеров регионов (по интересам).

| **Порог *Th*** | 0.125 | 0.0625 | 0.03125 | 0.015625 |
|---|---|---|---|---|
| **Кол-во кластеров** | 7 | 5 | 4 | 3 |
| **Средний размер кластера** | 2,2 | 9,2 | 15 | 16,5 |
| **Медиана размера кластера** | 2 | 2 | 2,5 | 3 |

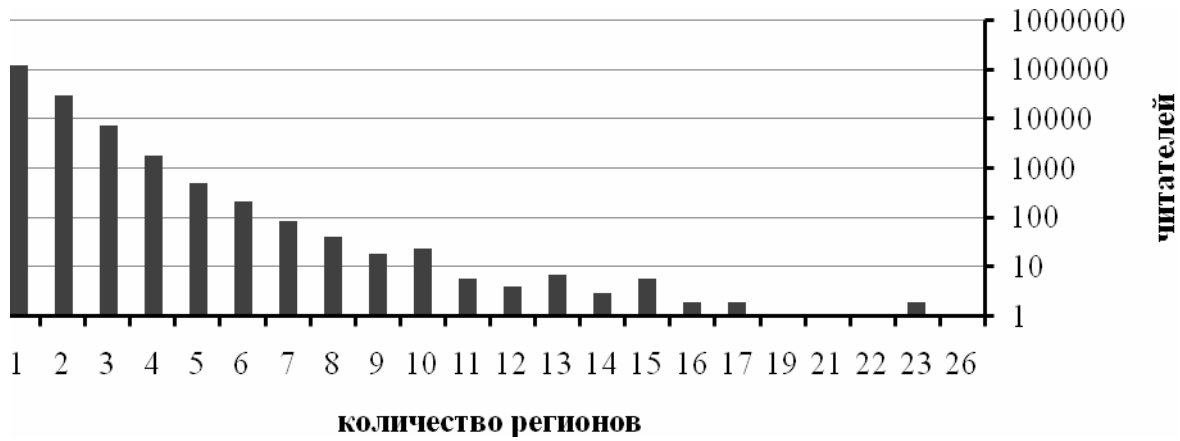



**Рисунок 2**. Распределение читателей по количеству регионов

**Таблица 9.** Характеристики невырожденных кластеров регионов (с выборкой интересов из средней части рангового распределения).

| **Порог *Th*** | 0.125 | 0.0625 | 0.03125 | 0.015625 |
|---|---|---|---|---|
| **Количество кластеров** | 4 | 9 | 15 | 10 |
| **Средний размер кластера** | 2,0 | 4,0 | 7,2 | 13,3 |
| **Медиана размера кластера** | 2 | 2 | 2 | 3 |

Список общих интересов для одного из кластеров, содержащего два региона (Приморский край и Хабаровский край), приведен в таблице 7.

Анализ частотного распределения интересов в профилях региональных сообществ показал, что в большинстве регионов доминирующим по частоте употребления в профилях сообществ является название областного (краевого, республиканского) центра, далее по рангу может встречаться название региона. Высокую частоту также имеют общераспространенные интересы, например: музыка, фото, компьютеры и т.п. В целом ранговое распределение интересов по частоте напоминает распределение Ципфа, что и объясняет характер кластеризации регионов по интересам. Однако согласно принципу Луна (Luhn), наиболее специфичными в распределении Ципфа являются элементы, находящиеся в средней части рангового распредления. Как представляется, исключение из анализа самых высокочастотных и самых редких (т.е. неспецифических) интересов может дать более содержательные результаты при кластеризации регионов по интересам. В таблице 9 приведены характеристики кластеров регионов, полученных для матрицы “регион-регион”, построенной для выборки интересов, имеющих частоту в интервале от 2 до 20.

## 4. Кластеризация регионов по читателям

Для исследования были использованы списки читателей сообществ, указанных в их профилях. В 19-ти самых крупных сообществах представленное в списке количество читателей отличалось от реального (до полутора раз).

Общее количество читателей (уникальных) составило порядка 160 тысяч.

На рисунке 2 представлено распределение читателей по количеству регионов, журналы сообществ из которых они читают. Например, всего лишь один пользователь является читателем блогов сообществ из 26 регионов, а число читателей блогов сообществ из не более чем одного региона составило порядка 100 тысяч.

В таблице 11 дается более детальное представление о распределении читателей по регионам. В ней указано, сколько пользователей являются читателями журналов сообществ только одного региона, двух и т.д Результаты кластеризации регионов по читателям существенно отличаются от кластеризации по интересам. В таблице 10 представлены характеристики полученных кластеров (невырожденных).

**Таблица 10**. Характеристики невырожденных кластеров регионов (по читателям).

| **Порог *Th*** | 0,125 | 0,0625 | 0,03125 | 0,015625 |
|---|---|---|---|---|
| **Кол-во кластеров** | 1 | 2 | 7 | 16 |
| **Средний размер кластера** | 2 | 2,5 | 2,4 | 2,7 |
| **Медиана размера кластера** | 2 | 2,5 | 2 | 2 |

При значении порога ***Th*** = 0.125 был выделен единственный невырожденный кластер, содержащий Москву и Санкт-Петербург. Кластеры для других значений порога ***Th*** приведены в таблице 12.

## Заключение

В данной работе была рассмотрена возможность использования ограниченного объема данных

**Таблица 11**. Распределение читателей по регионам.

| | Всего | Только 1 регион | Только 2 региона | | | | Только 3 региона | Только 4 региона | Только 5 регионов |
|---|---|---|---|---|---|---|---|---|---|
| | | | Регион+ Москва | Регион+ Спб | Регион+ другой | Всего | | | |
| **Читателей** | 219969 | 122618 | 25104 | 20762 | 16380 | 62246 | 21939 | 7424 | 2560 |
| **%** | 100 | 56 | 11 | 9 | 7 | 28 | 10 | 3 | 1 |

**Таблица 12**. Невырожденные кластеры регионов (по читателям) при различных значениях порога ***Th***.

**Th = 0.0625**

| Кл-р | Регион | Читателей |
|---|---|---|
| 20 | Калининградская область | 3760 |
| | Москва | 96520 |
| | Санкт-Петербург | 60616 |
| 41 | Томская область | 559 |
| | Новосибирская область | 5952 |

**Th = 0.03125**

| Кл-р | Регион | Читателей |
|---|---|---|
| 1 | Адыгея Республика | 62 |
| | Башкортостан Республика | 1143 |
| 7 | Белгородская область | 481 |
| | Владимирская область | 487 |
| 18 | Калининградская область | 3760 |
| | Москва | 96520 |
| | Санкт-Петербург | 60616 |
| | Свердловская область | 4023 |
| | Краснодарский край | 2088 |
| 33 | Мордовия Республика | 9 |
| | Тамбовская область | 62 |
| 38 | Новосибирская область | 5952 |
| | Томская область | 559 |
| 42 | Пензенская область | 166 |
| | Саратовская область | 724 |
| 44 | Приморский край | 3422 |
| | Хабаровский край | 646 |

**Th = 0.015625**

| Кл-р | Регион | Читателей |
|---|---|---|
| 1 | Адыгея Республика | 62 |
| | Башкортостан Республика | 1143 |
| 3 | Алтайский край | 373 |
| | Ростовская область | 2597 |
| | Тюменская область | 1212 |
| 5 | Архангельская область | 612 |
| | Вологодская область | 517 |
| | Мурманская область | 790 |
| 7 | Белгородская область | 481 |
| | Владимирская область | 487 |
| 8 | Брянская область | 462 |
| | Пермский край | 1456 |
| 9 | Курганская область | 106 |
| | Бурятия Республика | 94 |
| 11 | Воронежская область | 1463 |
| | Нижегородская область | 3504 |
| 15 | Томская область | 559 |
| | Челябинская область | 2348 |
| | Новосибирская область | 5952 |
| | Иркутская область | 1915 |
| 17 | Новгородская область | 444 |
| | Калининградская область | 3760 |
| | Москва | 96520 |
| | Санкт-Петербург | 60616 |
| | Свердловская область | 4023 |
| | Краснодарский край | 2088 |
| | Ленинградская область | 601 |
| | Удмуртская Республика | 569 |
| 25 | Костромская область | 318 |
| | Липецкая область | 236 |
| 29 | Мордовия Республика | 9 |
| | Тамбовская область | 62 |
| | Чувашская Республика | 256 |
| 30 | Татарстан Республика | 3131 |
| | Московская область | 3634 |
| 34 | Пензенская область | 166 |
| | Саратовская область | 724 |
| 35 | Приморский край | 3422 |
| | Хабаровский край | 646 |
| 38 | Самарская область | 2941 |
| | Тверская область | 378 |
| 40 | Сахалинская область | 463 |
| | Ханты-Мансийский автономный округ | 130 |

(только информации, извлекаемой из профилей сообществ пользователей) для изучения региональной структуры русскоязычной блогосферы. В рамках данной задачи наиболее информативными атрибутами из профилей сообществ являются списки интересов и читателей (участников).

Для анализа связей между регионами была использована процедура кластеризация.

Конечно, приведенные результаты выглядит скорее как набор статистических фактов, а использованные в работе общие подходы и их реализации в конкретных алгоритмах требуют более детального анализа и дальнейшей проработки. Тем не менее, исходя из этих результатов, можно сделать ряд интересных выводов.

Так, результаты кластеризации регионов по интересам и по читателям проявляют разные аспекты связи между регионами. Негеографические (экономические, культурные и подобные им) связимогут выявляться в большей степени с помощью кластеризации по интересам.

Кластеризация по читателям отражает в большей степени географическую близость регионов (включая и степень мобильности пользователей между регионами), причем, в большей степени это заметно для регионов, находящихся в неевропейской части РФ. Повышение уровня экономического и технологического развития региона приводит к конвергенции этих двух аспектов межрегиональных связей.

Можно провести аналогию с методами, которые применяются в поисковых системах для гипертекстовых документов. Для них используются 2 взаимно дополняющих друг друга подхода: содержательный (документ – как множество терминов) и топологический (документ – как узел гипертекстового графа). В случае регионов (а также отдельных сообществ и даже пользователей) эквивалентом *термина* является *интерес*, а топологическая связь между ними реализуется через отдельных читателей (участников) сообщества (или друзей пользователя). Как представляется, использование данной аналогии позволило бы применять хорошо проработанные методы анализа и поиска документов для поиска суперсообществ (регионов), сообществ и отдельных пользователей в социальных сетях.

Представленные в работе подходы могут быть использованы как для изучения связей между регионами, так и для построения поискового образа конкретного региона.

## Литература


[1] N. Agarwal, H. Liu. "Blogosphere: Research Issues, Tools, and Applications"//SIGKDD Explorations, 2008, Vol. 10, Issue 1. - pp. 18 - 31.

[2] A. Java, X. Song, T. Finin, B. Tseng "Why we twitter: understanding microblogging usage and communities" // WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD (2007), pp. 56-65.

[3] R. Ge, M. Ester, B. J. Gao, Z. Hu, B. Bhattacharya, B. Ben-Moshe "Joint Cluster Analysis of Attribute Data and Relationship Data: The Connected k-Center Problem, Algorithms and Applications". ACM Transactions on KDD, Vol. 2, No. 2. (2008), Article 7.

[4] Y.-R. Lin, Y.Chi, S.Zhu, H.Sundaram, B.L.Tseng "Analyzing Communities and Their Evolutions in Dynamic Social Networks" // ACM Trans. Knowl. Discov. Data, Vol. 3, No. 2. (2009), pp. 1-31.

[5] A. Mislove, M. Marcon, K. P.Gummadi, P. Druschel, B. Bhattacharjee "Measurement and Analysis of Online Social Networks"// Proceedings of the 7th ACM SIGCOMM conference on Internet measurement , 2007, San Diego. pp. 29-42.

[6] Состояние блогосферы российского интернета. По данным поиска по блогам Яндекса. Весна 2008 г. [Электрон. ресурс] – Режим доступа: (http://download.yandex.ru/company/yandex_on_blogosphere_spring_2008.pdf)

[7] Сычев А.В., Гадебский И.А. Изучение характеристик сообществ русскоязычной блогосферы //Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Десятой Всероссийской научной конференции «RCDL'2008». – Дубна: ОИЯИ, 2008. – С. 200-2009.


## The Study of The Profiles Features for Russian Blog Communities Hosted at LiveJournal: Regional Aspect


A.V. Sychev



On the base of data driven from profiles of russian communities, hosted at LiveJournal, some research tasks were carried out, like: new communities emerging chronology reconstructing, profiles attributes correlation estimating and Russian Federation regions clustering using both interests list and readers list from communities profiles.

# ДИССЕРТАЦИОННЫЙ СЕМИНАР–1

# PHD WORKSHOP–1

# Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска

© Рабчевский Е.А.

Пермский Государственный Университет
Кафедра компьютерных систем и телекоммуникаций
seus@rabchevsky.name

**Аннотация**

Обсуждается проблема автоматического построения онтологий на основе семантического анализа текстов на естественном языке. В качестве метода предлагается использование лексико-синтаксических шаблонов. Раскрывается синтаксис и семантика языка лексико-синтаксических шаблонов LSPL. Описывается программный комплекс, который позволяет:
- хранить шаблоны и корпус текстов на русском языке в базе данных
- редактировать и проводить валидацию шаблонов на корпусе русскоязычных текстов
- проводить семантический анализ текстов корпуса на основе шаблонов.
Для оценки предложенной методики семантического анализа предлагается оценивать результаты применения методики в приложении к информационному поиску. Предлагается модель информационного поиска на основе метрик TF*IDF, в которой понятие терма заменяется триплетом (атомарной единицей результатов семантического анализа). Обсуждаются результаты применения предложенной модели поиска к заданиям семинара РОМИП'2009.

## 1 Введение

Важнейшей проблемой в развитии Интернет является его интеллектуализация, и связанные с этим интеграция данных, качественный поиск, интеграция Веб служб и многое другое. В рамках подхода Semantic Web предлагаются эффективные средства для решения указанных задач.

Однако данные технологии предполагают наличие качественных источников семантических данных. Сегодня к таким можно отнести лишь источники, созданные при значительной поддержке корпораций или государственных структур, например база знаний Cyc или разработка языка DAML военным ведомством США. Также существуют достаточно качественные источники знаний, переведенные в форматы знаний Semantic Web из накопленных за многие годы источников. Например, онтологии генов, географических названий или RDF представление тезауруса Wordnet.

В общеупотребительных предметных областях не представлено открытых источников данных, которые бы реально повторно использовались Интернет сообществом. В этом отношении лидером можно считать RDF представление Wikipedia – DBPedia.

В целом, можно утверждать, что повторное использование и интеграция с указанными источниками данных находится на низком уровне. Это связано с тем, что источники данных не так совершенны, чтобы разработчики приложений могли с удобством их использовать или интегрировать в свои приложения. При этом потребность в приложениях подобного плана есть во всех сферах общества, отраженных в Интернет.

С русскоязычными источниками данных дело обстоит еще тяжелее.

В связи с этой проблемой проблема автоматического формирования онтологий на основе анализа текстов на естественном языке является весьма актуальной. Это подтверждается рядом современных исследований в данной области.

Методы автоматического построения используют средства компьютерной лингвистики, включающие все уровни анализа естественного языка, графематику, морфологию, синтаксис и семантику. Отличие между различными системами заключается в полноте комбинирования уровней анализа. Однако существенных результатов в данной области, особенно в применении к русскому языку, не представлено.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

Распространенный критерий качества онтологии основан на оценке работы приложения, использующего онтологию. Поэтому оценка автоматически построенных онтологий является еще отдельной сложной задачей, как и их построение. Ввиду наличия отработанных методик по оценке качества информационного поиска, последний можно рассматривать, как приложение, с помощью оценки которого, можно оценивать качество соответствующих онтологий.

Данная работа является попыткой решения проблемы автоматизации построения онтологий и оценки полученных результатов.

## 2 Лексико-синтаксические шаблоны как метод семантического анализа текста

Лексико-синтаксические шаблоны представляют собой характерные выражения (словосочетания и обороты), конструкции из определенных элементов языка. Такие шаблоны позволяют построить семантическую модель, соответствующую тексту, к которому они применяются.

Как метод семантического анализа, лексико-синтаксические шаблоны используются в компьютерной лингвистике более 20-ти лет. В своих исследованиях мы использовали работы таких авторов, как Марти Хэрст, Е.И. Большакова , Christopher Brewster, Fabio Ciravegna и Yorick Wilks; Ермаков А.Е., Плешко В.В., Митюнин В.А., Плешко В.В. (компания RCO).

Марти Хэрст предположила, что лексические отношения можно описать с помощью метода интерпретации образцов (шаблонов). Такой метод использует иерархию шаблонов, которые состоят главным образом из индикаторов части речи и групповых символов.

Хэрст выявила существенное количество шаблонов для идентификации отношения гипонимии [21]. Её исследования показали, что при использовании шаблонов на большом корпусе текстов одной тематики, можно построить «достаточно адекватную» таксономию понятий соответствующей предметной области. В её шаблонах в качестве элементов используются, например, понятие именной группы (NP), знаки препинания, конкретные слова.

Таким образом, шаблон «NP {, NP}* {,} and other NP», где NP – условное обозначение именной группы, определяет отношение гипонимии, которое продемонстрировано на части предложения «... temples, treasuries, and other important civic buildings …». С помощью указанного шаблона могут быть выявлены следующие отношения:

hyponym(\temple", \civic building"),

hyponym(\treasury", \civic building").

Группа разработчиков во главе с Большаковой сформулировала язык для записи лексико-синтаксических шаблонов (LSPL) [2]. По её мнению, элементами шаблонов, для наиболее точного описания, могут быть:

- литералы, т.е. конкретные лексемы;
- определенные части речи;
- определенные грамматические конструкции;
- условия, уточняющие грамматические характеристики рассмотренных элементов.

Разработанные ее коллективом шаблоны применяются для анализа научно-технических документов. Для их обработки кроме традиционных словарей (терминологического и морфологического), используется словарь общенаучных слов и выражений, лексико-синтаксические шаблоны типичных фраз научной речи.

На пример, предложение:

«По результатам генерации форм, слова были разбиты на группы, названные профилями» с помощью разработанной методики формализации авторы записывают так:

Ng «,» Pa<«названный»> T<:case=ins>

Ng.gender=Pa.gender

Ng.number=Pa.number=T.number>

Метод, разработанный Кристофером Бревстером [20] основан на разработках Марти Хэрст, и в качестве элементов шаблонов предполагается использовать словарную форму, представляющую лексему в словаре (lexical item, lemma), part of speech, syntactic role.

Для построения более четкой онтологии, с помощью указанного метода, приходится накладывать ограничения на анализируемую область, что направлено на повышение эффективности процесса обработки текста.

Командой разработчиков из компании RCO был разработан модуль, позволяющий производить сравнение цепочек лексем, заданных своими описаниями.

Описание лексемы содержит набор предопределенных атрибутов (всего более 20-ти), например:

Token.Text - строка лексемы;

Token.Type - тип лексемы (известное/неизвестное слово русского языка, латинское слово, специальная конструкция);

множество грамматических характеристик - Morph.SpeechPart (часть речи), Morph.Case (падеж), Morph.Gender (род), Morph.Number (число) и т.п.

Ориентация, описанных шаблонов, направлена на выявление специфических объектов, таких как: даты, адреса, имена юридических организаций и т.п.

В целом, можно сказать, что лексико-синтаксические шаблоны как метод семантического анализа текста являются достаточно эффективным средством при условии большого объема шаблонов (разумеется, объем зависит от специфики задачи).

Также стоит заметить, что данная методика достаточно дорогостоящая в плане процессорного времени, потому как использует все уровни анализа естественного языка.

## 3 Новизна и уникальность работы

Рассмотрим данную работу в сравнении с приведенными выше исследованиями.

Выделим два критерия, по которым можно проводить анализ данной работы в отношении с другими работами. Это – цель семантического анализа и оценка качества полученных результатов.

Основным отличием является цель семантического анализа, который проводится в данной работе. В нашем случае – это построение онтологических конструкций, соответствующих тексту, на открытых языках представления знаний Semantic Web. Это предполагает использование на выходе системы форматов данных, схем и словарей, широко используемых Интернет сообществом, поддерживающим подходы Semantic Web и Linked Data. Также можно выделить отсутствие специализации методики семантического анализа на определенную задачу или предметную область, как например выделение фактографической информации в разработках RCO или анализ научно-технической прозы в работах Большаковой.

Соответственно отличается и использование результатов анализа, в нашем случае – это приложения Semantic Web, семантическая разметка ресурсов и др.

Если говорить об оценке результатов, то ввиду сложности самой задачи, в работах отечественных авторов этот вопрос не поднимается. В работах Шеффилдского университета оценка применения шаблонов и грамматик к тексту проводится. Однако они больше используются для автоматической аннотации текста, нежели построения онтологий. В данной работе оценка проводится опосредованно, что конечно вносит свое влияние, но при наличии фиксированной модели информационного поиска, на основе которого проводится оценка, можно однозначно судить о качестве построенных онтологий.

Автор считает, что использование лексико-синтаксических шаблонов как средства для автоматического построения онтологий является обоснованной методикой. Это обусловлено тем, что например, в работах Марти Хэрст получены качественные результаты выделения отношения гипономии из текста с помощью данной методики. Существуют и другие примеры, демонстрирующие состоятельность данной методики, реализованные также в работах Шеффилдского университета [11], однако в применении к автоматической аннотации.

## 4 Язык лексико-синтаксических шаблонов LSPL (ПГУ)

Автор занялся разработкой собственной системы формализации лексико-синтаксических шаблонов в 2006 г. [15], что по времени совпадает или на год позже разработок коллектива Большаковой. В тот момент также было выбрано название LSPL, однако о разработке Большаковой нам тогда известно не было.

Во избежание путаницы между данными системами формализации следует заметить, что далее речь пойдет только о данной разработке.

Формализация шаблонов производится на XML-подобном языке LSPL (Lexical-Syntactic Pattern Language). Тело шаблона состоит из входной и выходной схем. Входная схема – характерное описание части предложения, по которому в сочетании с входным текстом, можно однозначно построить выходную семантическую модель, соответствующую анализируемому тексту. Выходная семантическая модель представляется набором RDF триплетов, состоящих из субъекта, объекта и предиката.

Входная схема шаблона записывается в элементе <inputSchema>, который находится в теле основного тега <pattern>. Описание соответствующих элементов шаблона производится в теге <element>, дочернем относительно <inputSchema>. Атрибутами данного тега являются:

- type – тип элемента (literal, wordForm, partOfSpeech и syntacticGroup);
- id – идентификатор элемента в шаблоне по порядку;

В дочернем, относительно <element>, элементе <content> указывается содержание элемента шаблона. На данный момент поддерживаются следующие элементы:

1) Literal –слово, указанное внутри <content>: и, или, это

Пример:

```
<element type=”literal” id=”1”>
    <content>это</content>
</element>
```

2) wordForm – форма указанного слова

Пример:

```
<element type=” wordForm” id=”1”>
<content [грамматические значения] > место</content>
</element>
```

Рассматриваются любые или конкретно указанные формы слова внутри <content>

3) partOfSpeech – слово указанной части речи. На данный момент поддерживаются глаголы, существительные, прилагательные, числительные, предлоги, местоимения и наречия.

.Пример:

```
<element type=”partOfSpeech” id=”1”>
<content [грамматические значения] > noun</content>
</element>
```

4) syntacticGroup – синтаксическая группа, состоящая из нескольких слов, идущих подряд. В <content> содержится часть речи главного слова

Пример:

```
<element type=” syntacticGroup” id=”1”>
<content> noun</content>
</element>
```

5) punctualMark – под него подходит знак препинания. Если <content> пуст – то подойдёт любой знак препинания.

<element type=" punctualMark" id="1">
<content>,</content>
</element>

Вид выходной схемы описывается в теге <outputSchema>, дочернем для <pattern>. Элементы каждого триплета образца выходной модели записываются в теге <statement>, в котором указывается URI-ссылки субъекта, объекта и предиката.

Применение лексико-синтаксических шаблонов заключается в следующем:

1. на вход анализатора или интерпретатора языка LSPL подается входная схема шаблона и анализируемый текст;
2. анализатор, используя отмеченные в шаблоне свойства термов (часть речи, число, род, падеж, наклонение), последовательность слов, строит семантическую карту анализируемого текста в соответствии с выходной схемой шаблона

Пример:

| Анализируемое предложение | Студент - это человек, который учится в университете |
|---|---|
| Входная схема | <pattern> <inputSchema><br><element type="partOfSpeech" id="1"><br><content>noun</content><br></element><br><element type="punctualMark" id="2"><br><content>-</content><br></element><br><element type="literal" id="3"><br><content>это</content><br></element><br><element type="partOfSpeech" id="4"><br><content>noun</content><br></element><br><element type="punctualMark" id="5"><br><content>,</content><br></element><br><element type="wordForm" id="6"><br><content>который</content><br></element><br><element type="partOfSpeech" id="7"><br><content>verb</content><br></element><br></inputSchema> |
| Выходная схема | <outputSchema><br><statement><br><subject>http://result/subject/##1##</subject><br><object>http://result/object/##4##</object><br><property>http://result/property/#subClassOf</property><br></statement></outputSchema></pattern> |
| Полученные триплеты (за исключением вспомогательных триплетов языка RDFS) | http://result/subject/Студент<br>http://result/property/#subClassOf<br>http://result/bject/человек |

## 5 Интерпретатор языка лексико-синтаксических шаблонов LSPL

Для обеспечения семантики языка LSPL использовался синтаксический анализатор DictaScope [17]. Анализатор предоставлен компанией "Диктум" для некоммерческого использования и с согласия разработчика морфологического словаря А.Коваленко.

Для практического использования лексико-синтаксических шаблонов нами была разработана на языке Java библиотека PatternLib. Библиотека разделена на пакеты со следующей функциональностью

- взаимодействие с синтаксическим анализатором DictaScope
- обработка шаблонов и их применение к тексту
- визуализация полученных при анализе RDF графов с помощью библиотеки GraphViz [3]
- хранения шаблонов и анализируемых документов в базе данных.
- парсинг текстовых коллекций семинара РОМИП'2009 [16]

О последних двух пакетов подробнее будет сказано главе, посвященной участию в семинаре РОМИП.

Для использования возможностей библиотек DictaScope и GraphViz на платформе Java были разработаны две вспомогательные DLL-библиотеки, использующие для связи механизм JNI: LoaderLib.dll и GraphLib.dll соответственно.

Библиотека PatternLib сама не имеет пользовательского интерфейса и используется посредством следующих программам:

- программа редактирования и валидации шаблонов
- классы поисковой системы SEUS, отвечающие за индексацию документов, подробнее об этом

будет сказано в главе, посвященной поиску на основе семантики

- online версия анализатора, которая доступна на сайте проекта SEUS [1].

Алгоритм применения шаблонов к тексту выглядит следующим образом.

Берётся 1-ый элемент шаблона, после чего последовательно перебираются элементы предложения и уровни вложенности для синтаксических групп, до тех пор, пока не найдётся соответствующий элемент предложения.

После нахождения соответствия для 1-ого элемента аналогичная операция проводится для 2-ого элемента шаблона, но поиск ведётся с позиции, следующей за позицией первого элемента. Аналогично для следующих элементов шаблона.

Если на шаге N для N-ого элемента не нашлось лексикализации, то алгоритм поиска возвращается на уровень N-1.

Если найдено соответствие для последнего элемента шаблона (т.е. найдено соответствие для всего шаблона) – то формируются результирующий набор триплетов на основе выходной схемы шаблона..

После обработки выходной схемы алгоритм возвращается на уровень N-1 для поиска остальных возможных вариантов вхождения.

В дальнейшем планируется использовать поиск лексикализаций на основе би-деревьев.

## 6 Online демонстрация семантического анализатора на основе языка LSPL

Для демонстрации работы библиотеки PatternLib нами был разработан веб сервис анализатора. При разработке online сервиса использовались фреймворк Struts [12], и некоторые функции JavaScript из библиотеки компонентов YUI (Yahoo User Interface) [6].

Работа online версии анализатора проходит следующим образом:

1) На главной странице вводится текст или загружается файл с указанием кодировки.

2) Выбираются шаблоны, которые будут использоваться при анализе текста (по умолчанию используются все) – при выборе названия шаблона из списка, справа отображается его содержимое

3) При нажатии кнопки «Анализировать» производится анализ заданного текста и выводится набор полученных триплетов и уникальных RDF ресурсов, являющихся субъектами или объектами.

4) Результат анализа можно просмотреть в виде RDF-графа. Для этого нужно нажать кнопку «Граф».

Используемые анализатором шаблоны хранятся в базе данных. На сегодняшний день не предусмотрен Веб интерфейс для добавления, удаления или редактирования шаблонов в базе данных, эта операция выполняется администратором системы в ручную (через средства администрирования СУБД). В дальнейшем планируется реализовать веб интерфейс для управления шаблонами анализатора.

## 7 Разработка, хранение и валидация шаблонов на корпусе текстов

Для поиска новых шаблонов от разработчика требуется вручную анализировать тексты. Это занимает очень много времени, поэтому было разработано Web-приложение для редактирования и проверки шаблонов. Предполагается, что данный программный продукт (валидатор шаблонов – рабочее название Vallyweb) [7] сможет упростить создание, анализ и проверку шаблонов, описанных на языке LSPL.

Валидатор предназначен для работы с текстами, которые хранятся в виде HTML документов в файловой системе сервера. На сегодняшний день для работы используются коллекции документов, используемые в дорожках семинара РОМИП’2009.

Ниже представлен алгоритм работы валидатора.

Работа с валидатором строится следующим образом:

На вход программы подается шаблон (выбирается из списка). Текст шаблона записывается в соответствующее окно и становится доступным для редактирования.

Так же на вход подается анализируемый корпус текстов (указывается адрес до папки с файлами). Для получения содержимого указанных выше документов используется парсер PatternX3M, реализованный на компонентах библиотеки Lucene [5].

Для анализа выбирается выбранное случайным образом некоторое число документов;

Во время парсинга текст разделяется на отдельные предложения, каждое из которых анализируется с помощью библиотеки PatternLib.

Результат выводиться в виде таблицы из двух колонок. В строчках таблицы отображаются все лексикализации шаблона и соответствующие наборы триплетов. Например:

| Предложение | Соответствующая семантическая модель |
|---|---|
| Студент - это человек, который учится в университете | http://result/subject/Студент<br>http://result/property/#subClassOf<br>http://result/bject/человек |

В результате, пользователь может оценить полученные лексикализации шаблона и откорректировать шаблон соответствующим образом. Текст шаблона можно исправить прямо в программе.

При создании и валидации шаблона исследователь может ввести коэффициент доверия шаблона, который отражает адекватность работы шаблона. Данный коэффициент можно рассматривать, как вероятность успешной работы шаблона, то есть отношение количества выходных семантических моделей, полученных при применении шаблона, реально соответствующих

семантике предложения, к общему количеству лексикализаций шаблона.

На сегодняшний день интерфейс для вычисления коэффициента доверия шаблонов не реализован, его реализация планируется в дальнейшем.

Шаблоны, с которыми работает валидатор хранятся в виде XML файлов на сервере. В дальнейшем для хранения шаблонов планируется использовать ту же базу данных, в которой хранятся шаблоны, используемые online анализатором. Это позволит организовать централизованное хранилище шаблонов и избавиться от дублирования данных.

Так же в будущем планируется предоставить доступ к валидатору всем желающим, для того чтобы с помощью сообщества исследователей создавать и проводить валидацию новых шаблонов более быстро и эффективно. Для этого потребуется реализовать механизм разграничения прав доступа и контроля версий.

## 8 Информационный поиск на основе семантики

Наиболее распространенными моделями информационного поиска по текстовым коллекциям документов являются:

1. Статистические методы
2. Методы поиска по семантическим сетям
3. Комбинированные методы

Кратко рассмотрим метод поиска на основе метрик TF-IDF.

Для коллекции документов строится свой, особый «алфавит», в который входят все (за исключением стоп-слов и словоформ, отличающихся от нормальных) встречающиеся в данных документах слова (термы).

Затем для каждого терма определяется частота встречаемости его в каждом документе. Таким образом, для каждого документа можно построить вектор частот Dm(t1,t2,…,tn), где t1 – частота встречаемости терма 1 в документе m, t2 – частота встречаемости терма 2 в документе m, и т.д. m – уникальный номер документа в коллекции, n – количество известных термов.

В итоге, в индексе (матрице из векторов частот отдельных документов) поисковой машины хранятся частотные вектора всех документов. При обработке запроса, сначала выбираются все термы, которые присутствуют в тексте запроса и строится соответствующий вектор Q(tx,ty,…tz), где tx, ty, tz – частоты входящих в запрос термов.

После построения вектора частот запроса, вектора частот документов дополняются нулями для тех термов, которые входят в запрос, но не входят в алфавит, а вектор частот запроса дополняется нулями для всех термов из алфавита, которые не входят в текст запроса.

Таким образом, все вектора приводятся к одной размерности. В конечном итоге, вычисляется условный косинус угла между векторами запроса и документов. Чем меньше данная величина, тем более релевантным считается документ.

Статистические методы, в настоящий момент, являются наиболее распространенными методами информационного поиска. Основной их особенностью является качественная математическая модель, позволяющая получать хорошие оценки релевантности для документов коллекции. Поисковые машины, основанные на данных методах, отличаются простотой интерфейса. Основным минусом данного метода является то факт, что не учитывается смысловая нагрузка текста документов коллекции и текста запроса.

Отсутствие учета смысловой нагрузки текстов (документов и запросов), зачастую приводит к нерелевантным результатам. Примерами поисковых машин такого типа являются популярные поисковые машины Google, Yandex, Rambler, Yahoo и т.д.

Основная идея второй группы методов информационного поиска заключается в том, что все исходные данные представлены в виде объектов семантических моделей, а поиск представляет собой навигацию по графу онтологии. Данные методы, в отличие от статистических, учитывают смысловую нагрузку информации, поскольку информация изначально представлена в виде онтологии или ассоциирована с ней посредством семантической разметки документов. Однако данные методы имеют ряд недостатков:

- Сложность пользовательского интерфейса, требующая от пользователя дополнительных затрат на конкретизацию объектов и свойств.
- Большинство информации в Интернет представлено в виде HTML-страниц и не содержат семантического описания контента. А ручная семантическая разметка документов представляет собой огромный объем работы.

В качестве примера подобной системы можно рассматривать систему АСНИ (Автоматизированная система научных исследований) [13] или проект SHOE [18].

К третьей группе методов информационного поиска относятся методы, которые помимо статистических методов поиска используют методы семантического анализа текстов. Данная группа методов развивается в настоящее время наиболее интенсивно. Основным плюсом систем комбинированного типа является комбинация качественной статистической модели поиска и учета семантических конструкций.

Основные минусы подобных систем, существующих в настоящее время:

- Большое время отклика
- Мало где используются механизмы логического вывода
- Ограничения на структуру запроса (при использовании простого пользовательского интерфейса)

- Необходимость установки дополнительных параметров поиска (при использовании сложных пользовательских интерфейсов)

- Большинство систем подобного типа используют в качестве исходной информации стандартные тексты, проводя семантический анализ на конечном этапе задачи поиска, что приводит к медлительности данных систем.

В качестве примера такой системы можно рассматривать поисковую машину AskNet [19].

Несмотря на то, что третья группа методов наиболее полно отвечает требованиям. предъявляемым к системам информационного поиска на основе семантики, все системы данного типа имеют недостатки.

Была поставлена задача разработать метод поиска, который бы был основан на статистическом методе поиска, учитывал семантическую структуру текстов, а так же был лишен таких недостатков третьей группы методов, как большое время отклика, ограничения на структуру запроса, и сложный пользовательский интерфейс. Разработка получила название SEUS (search engine using semantics) [14]

Решение заключается в том, чтобы исходную информацию представить в виде семантической сети, и работать уже не с отдельными словами, а с элементами данной сети (RDF-триплетами). Для приведения исходных данных из текстов на естественном языке в семантическую сеть, предлагается использовать модель представления информации, основанную на автоматическом построении онтологии с использованием лексико-синтаксических шаблонов.

Суть данной модели заключается в следующем:

1. Из предложений исходного документа извлекаются триплеты, которые в совокупности составляют полную онтологию данного документа. Для этого используются 4 механизма получения триплетов:

- используя триплеты, заранее встроенные в HTML документы с помощью микроформатов, например RDF/A [10]

- используя лексико-синтаксические шаблоны

- используя логический вывод, реализованный на базе библиотеки для работы с онтологиями Jena [4]

- а также логический вывод, специально разработанный для информационного поиска

- предполагается, что также будут использоваться и RDF ресурсы, описанные в популярных онтологиях и словарях, таких VCard FOAF и т.д.

2. Полученные таким образом триплеты сохраняются в БД, в качестве Jena-моделей, имеющих уникальные идентификаторы и ссылки на документы, к которым они относятся.

3. После получения всевозможных триплетов, которые формируют онтологию документа, считаем, что содержимое данного документа – есть набор идентификаторов триплетов, каждый из которых будет считаться отдельным термом.

4. Таким образом, после проведенных преобразований, можно воспользоваться существующей моделью TF/IDF, для которой алфавит составят идентификаторы триплетов, входящих в документ.

Помимо преобразования документов к набору идентификаторов триплетов, такому же преобразованию должен быть подвергнут и запрос. То есть, используя механизм лексико-синтаксических шаблонов, можно получить набор триплетов запроса, для тех которые уже имеются в алфавите, указать соответствующие идентификаторы, а для тех, которых в алфавите нет – ввести отрицательную нумерацию идентификаторов, что позволит учитывать их при использовании методики TF/IDF.

Кроме того, поскольку результаты применения лексико-синтаксических шаблонов к тексту могут не всегда отражать реальную семантику текста, они обладают некоторыми экспериментально определенными коэффициентами доверия. Поэтому данный коэффициент доверия нужно учитывать при определении частот термов. Учтен он будет следующим образом: частота терма в документе и запросе будет умножена на соответствующий коэффициент доверия.

Таким образом, предложенный метод использует статистический метод поиска, учитывает смысловую нагрузку исходного текста, за счет того, что он представляется в виде набора RDF-триплетов.

Данный метод лишен недостатков систем, использующих комбинированные методы. Поскольку семантический анализ исходных текстов перенесен из последнего этапа задачи поиска в задачу представления исходной информации и проводится еще до этапа индексирования исходной коллекции документов, на конечном этапе задачи поиска существенно сокращается время отклика поисковой системы.

Если не удалось получить триплеты из текста запроса, система автоматически переключается на работу со стандартным статистическим методом.

Для того, чтобы использовался комбинированный метод поиска, необходимо из запроса получить хотя бы один триплет. Поскольку триплеты из запроса получаются с помощью механизма лексико-синтаксических шаблонов, то ограничения на запрос определяются лишь их распространенностью.

В качестве интерфейса поисковой машины, использующей данный метод поиска можно использовать обычную строку поиска.

## 9 SEUS на РОМИП 2009

На базе открытой библиотеки для полнотекстового поиска Lucene в рамках проекта SEUS нами был реализован программный комплекс,

который применяет к тексту лексико-синтаксические шаблоны, то есть получает соответствующие триплеты, а также строит семантический индекс, соответствующий модели, описанной в предыдущей главе. На данный момент реализована стандартная модель поиска Lucene и интерфейс в виде Веб приложения со строкой запроса [8].

На данный момент, описанный ранее механизм поиска не дает желаемых результатов. Это обусловлено тем, что набор лексико-синтаксических шаблонов сейчас достаточно мал (12 штук). Кроме того, в механизме предварительного анализа не реализовано использование триплетов, заложенных в исходные документы с помощью микроформатов. Механизм логического вывода, специфический для информационного поиска пока так же не реализован.

Участники проекта SEUS подали заявку на участие системы в семинаре РОМИП'2009. Однако приведенная модель поиска не использовалась при решении заданий семинара. Это связано с низким качеством полученных практических результатов. Вместо этого на семинар были представлены результаты, полученные на основе стандартной модели поиска библиотеки Lucene.

Имея таблицы релевантности для заданий семинара, полученные от экспертов, в дальнейшем планируется доработать все элементы системы SEUS согласно модели поиска с учетом семантики.

## Заключение

В рамках проекта SEUS были разработаны интерпретатор языка LSPL, а также online анализатор и валидатор шаблонов на его базе. Предполагается, что это станет значительным шагом к получению объема шаблонов, необходимого для качественного семантического анализа текстов на русском языке.

Предложенная модель поиска с учетом семантики требует более качественного семантического анализа, который ожидается, будет в связи с появлением анализатора и валидатора шаблонов будет получен в ближайшем будущем.

Также во время подготовки к участию в семинаре РОМИП авторы осознали, что поиск с учетом семантики сам по себе является обширной задачей и не может рассматриваться, лишь как метод оценки работы лексико-синтаксических шаблонов.

Дальнейшая работа над проектом SEUS будет двигаться в следующих направлениях:

- доработка и открытие в общий доступ валидатора шаблонов с целью привлечения заинтересованных специалистов
- получение количества шаблонов достаточного для качественного семантического анализа
- разработка собственного механизма логического вывода, специально предназначенного для информационного поиска
- получение результатов поиска с учетом семантики приближенных к стандартным моделям TF/IDF.

Автор благодарит компанию «Диктум», ее руководителя В.В. Окатьева и разработчика морфологического словаря А.Коваленко за предоставление синтаксического анализатора DictaScope. А также д. ф.-м. н., профессора заведующего кафедры компьютерных систем и телекоммуникаций Пермского Государственного Университета М.А. Марценюка за обсуждение работы и предоставление материально технической базы для проведения исследований. А также студентов кафедры за предоставление практических наработок, на базе которых написана статья.

## Литература

[1] Анализатор на базе лексико-синтаксических шаблонов http://seus.rabchevsky.name:8080/DemoServlet/

[2] Большакова Е.И., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006. Труды конференции в 3-х томах. М.: Физматлит, 2006. Т. 2. С.506-524

[3] Библиотека для визуализации графов GraphViz http://www.graphviz.org/

[4] Библиотека для работы с онтологиями Jena http://jena.sourceforge.net/

[5] Библиотека для полнотекстового поиска Lucene http://lucene.apache.org/

[6] Библиотека YUI (Yahoo User Interface) http://developer.yahoo.com/yui/

[7] Валидатор шаблонов Vallyweb http://seus.rabchevsky.name:8080/VallyWeb/

[8] Демонстрация поисковой системы SEUS http://seus.rabchevsky.name:8080/SEUS/

[9] Ермаков А.Е., Плешко В.В., Митюнин В.А. RCO Pattern Extractor: компонент выделения особых объектов в тексте. // Информатизация и информационная безопасность правоохранительных органов: XII Международная научная конференция. Сборник трудов - Москва, 2003. - С. 312-317. (http://www.rco.ru/article.asp?ob_no=237)

[10] Микроформат для внедрения RDF графов в HTML документы RDF/A http://www.w3.org/TR/xhtml-rdfa-primer/

[11] Обзор методов аннотирования в Semantic Web в работах Шеффилдского университета http://rabchevsky.name/sheffield

[12] Платформа Struts http://struts.apache.org/

[13] ПОДСИСТЕМА УТОЧНЯЕМОГО ПОИСКА СЕМАНТИЧЕСКОЙ ИНФОРМАЦИИ В

ФОРМЕ ГРАФОВЫХ МОДЕЛЕЙ АСНИ http://network-journal.mpei.ac.ru/cgi-bin/main.pl?l=ru&n=9&pa=12&ar=8
[14] Поисковая система с использованием семантики SEUS http://seus.rabchevsky.name/
[15] Рабчевский Е.А., Автоматическое построение онтологий // Научно-технические ведомости СПбГПУ № 4 2007 . – Санкт-Петербург: Издательство Политехнического Университета 2007.
[16] Семинар РОМИП http://romip.ru/
[17] Синтаксический анализатор DictaScope http://www.dictum.ru/?main=products&sub=dictascope
[18] Система поиска по семантически размеченным документам http://www.cs.umd.edu/projects/plus/SHOE/index.html
[19] Система полнотекстового поиска AskNet http://info.asknet.ru/technology.htm
[20] Christopher Brewster, Fabio Ciravegna и Yorick Wilks, User Centred Ontology Learning for Knowledge Management // Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems-Revised Papers, Pages: 203 - 207, Springer-Verlag London, UK, 2002.
[21] Marti A. Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora // Proceedings of the 14th conference on Computational linguistics - Volume 2, Pages: 539 - 545 , Nantes, France, Association for Computational Linguistics, Morristown, NJ, USA, 1992.

## Automatic ontology construction based on lexical-syntactic patterns for information retrieval

Evgeny Rabchevsky
Perm State University
seus@rabchevsky.name

The problem of automatic construction of ontology on a basis of semantic analysis of natural language is under discussion. The use of lexical-syntactic patterns is suggested. Syntax and semantics of language of lexical-syntactic patterns is considered. Developed software is able to

- store the patterns and text's corpora in Russian language database;
- edit and validate patterns on the Russian language on the corpora of Russian language texts;
- make semantic analysis of texts' corpora on a basis of patterns.

The results of the method is suggested to evaluate as a quality of information retrieval. The definition of term in the information retrieval model based on TF/IDF is been changed to RDF triple. The results of the given model application to ROMIP seminar tasks are discussed.

# Разработка метода семантической интеграции информации в сфере государственного и муниципального управления♣

©Ломов П. А. Шишаев М. Г.

Институт Учреждение Российской академии наук Институт информатики и математического моделирования технологических процессов Кольского научного центра РАН
lomov@iimm.kolasc.net.ru, shishaev@iimm.kolasc.net.ru

**Аннотация**

В данной работе предлагается подход к семантической интеграции данных в сфере государственного и муниципального управления с использованием разделяемого тезауруса, который позволяет устранить критичные для данной предметной области недостатки, присущие существующим подходам к интеграции. Представлена концептуальная модель тезауруса, механизм отображения в него онтологий, а также методика задания и сопоставления онтологических контекстов на основе набора общих атрибутов.

## 1 Введение

Проблема интеграции информации, заключающаяся в предоставлении единой точки доступа к распределенным и гетерогенным информационным ресурсам, характерна для многих предметных областей и сфер деятельности человека, в том числе и для сферы государственного или муниципального управления. Особенную актуальность данная проблема приобрела в настоящее время вследствие формирования так называемого электронного государства, которое предполагает создание разветвленной коммуникационной инфраструктуры, позволяющей государственным органам и гражданам взаимодействовать с использованием новых информационных технологий[1].

Целью работы является разработка метода интеграции данных, а также моделей информационных систем и программных средств, позволяющих производить интеграцию информации в сфере государственного и муниципального управления на основе ее семантики.

**Труды 11ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

## 2 Проблематика семантической интеграции информации

### 2.1 Применение онтологий для формального отражения семантики

Традиционно для представления знаний об определенной предметной области использовались такие формализмы, как семантические сети, фреймы. Однако, при всей своей наглядности такое представление не позволяло формально отразить значение того или иного термина или отношения, делая невозможным обработку таких знаний с помощью ЭВМ.

Оперирование семантикой стало возможным благодаря появлению и развитию моделей представления знаний, позволяющих в формальном виде отразить смысл некоторого информационного элемента. Это, в свою очередь, позволяло в определенной степени производить машинную обработку информации, подобно эксперту. Среди таких моделей можно выделить онтологии - формальные спецификации разделяемой концептуализации.

Способность определения формального смысла появилась, благодаря использованию дескриптивных логик, которые лежат в основе многих распространенных языков описания онтологий, таких как, OIL, DAML+OIL, OWL-DL, OWL-LITE.

Онтология – спецификация концептуализации [5], или явное, формальное описание предметной области. Онтологию можно представить в виде упорядоченной тройки конечных множеств:

$$O=\langle T, R, F\rangle, \quad (1)$$

где:

$T$ — термины предметной области, которую описывает онтология O;

$R$ — отношения между терминами заданной предметной области;

$F$ — функции интерпретации, определенные на терминах и отношениях онтологии O, имеющие следующий вид:

$$I(t) \subset dom, \quad (2)$$

$$I(R) \subset dom \times dom, \quad (3)$$

где $t$ – термин предметной области, $dom$ – множество объектов реального мира, $I \in F$.

Следует заметить, что функции интерпретации, явно в онтологии не присутствуют, эксперт, добавляя в онтологию аксиомы, определяет ограничения на интерпретации терминов и отношений, в соответствии со своим пониманием их смысла. Полученная в итоге совокупность аксиом и определяет формальную семантику элементов онтологии.

Для описания онтологий существует несколько языков, отличающихся выразительностью, наличием возможности полного логического вывода. Однако при разработке современных информационных систем является более предпочтительным использовать языки или технологии, прошедшие стандартизацию и рекомендованные к применению в промышленных проектах. Примерами таких технологий могут служить технологии Semantic Web, такие как RDF(Resource Definition Framework)[9], DAML+OIL, OWL(Ontology Web Language)[6], OWL2[7]. Основным назначением данных языков является формальное описание семантики данных в виде совокупности объектов и отношений между ними, что, в свою очередь, позволяет производить интеграцию информации, руководствуясь ее смыслом, а не форматом представления.

### 2.2 Рассмотрение существующих подходов к интеграции и их применения в области государсвенного и муниципального управления

При выборе того или иного интеграционного подхода необходимо учитывать специфические особенности предметной области, что может в определенной степени облегчить проведение интеграционных процессов.

В данном случае одной из таких особенностей является то, что большинство определений как субъектов и объектов, а также процедур, ситуаций, находит свое отражение в различных документах, которым должно следовать то или иное государственное образование. Данное обстоятельство облегчает задание формальных моделей как различных сущностей, так и взаимодействий между ними. Однако в тоже время существует тенденция внесения различных изменений в правовые документы, что должно находить свое отражение в изменениях моделях определенных сущностей.

Также в рассматриваемой предметной области данные об определенной сущности не хранятся централизованно, а разбросаны по различным источникам. Причем добавление новой информации или изменение существующей, должно происходить с учетом значений в других источниках, в противном случае может быть нарушена семантическая целостность информации об объекте.

Наряду с этим, немаловажной является обеспечение возможности гибкого регулирования доступа к атрибутам объекта в зависимости от задачи и решающего ее ведомства.

Исходя из данных особенностей, можно использовать централизованный подход к интеграции, который заключается в создании единой онтологии, постулирующей формальную семантику терминов предметной области. Однако разработка и поддержка достаточно объемной онтологии с большим количеством аксиом, является очень трудной задачей, и любая модификация будет требовать привлечения как экспертов предметной области, так и инженеров по знаниям. С ростом онтологии также появиться проблема, связанная с ее практическим использованием для определения семантики добавляемых информационных ресурсов, чей набор терминов не будет находить точного отображения в концептах онтологии. Данные недостатки отсутствуют у децентрализованного подхода, который не накладывает ограничений на создание онтологий отдельных информационных ресурсов, что позволяет точно отразить формальную семантику конкретных терминов. Его применение для рассматриваемой предметной области можно увидеть в работе[3]. Однако он требует определения способов отображения между концептами и отношениями различных онтологий, что является нетривиальной задачей. Также довольно сложным становиться централизованное установление прав доступа и получение совокупной информации об объекте из разных источников.

Неким компромиссом между рассмотренными подходами может являться гибридный подход. В сравнении с централизованным, он позволяет обеспечить гораздо большую выразительность при создании частных онтологий информационных ресурсов и, как следствие, более точное отражение семантики терминов. Наряду с этим, в отличие от децентрализованного, существенно облегчается задача установления различных отношений с терминами отдельных онтологий.

Перечисленные положительные свойства обеспечиваются благодаря использованию общего словаря, на основании которого строятся частные онтологические описания. Однако само построение словаря может производиться разными способами. Так, например, в работе[11] в разделяемом словаре содержатся термины-примитивы, которые, комбинируясь друг с другом, формируют лейблы описывающие отдельные концепты. Вследствие этого появляется возможность производить автоматизированное сравнение концептов, исходя из описывающих их лейблов. Недостаток этого подхода заключается в том, что выразительность описания того или иного информационного

элемента ограничивается выразительной мощностью общего словаря, и в ряде случаев приводит к усреднению семантических описаний. При расширении же словаря появляется проблема задания одного и того же лейбла, с помощью различных комбинаций терминов-примитивов, что приводит к проблеме неоднозначности формально заданной семантики термина.

В других подходах[2,10] общий словарь реализован в виде онтологии верхнего уровня, постулирующей общие концепты, которые уточняются частными онтологиями. Это позволяет устанавливать различные семантические отношения между концептами частных онтологий, относящихся, например, к одному классу верхнего уровня. Данный подход ориентирован на применение в рамках одной предметной области, где можно явно установить набор базовых классов. Однако представленные методы, оставляют без внимания проблему установления соответствия между экземплярами онтологий – моделями, отражающими основные свойства конкретных объектов реального мира. Данное обстоятельство является очень важным для рассматриваемой предметной области, и его учет требует применения особого подхода к семантической интеграции данных.

## 3 Подход с использованием общего разделяемого тезауруса

### 3.1 Определение тезауруса

Исходя из описанных характеристик различных подходов к семантической интеграции, было решено выбрать разновидность гибридного метода, предполагающую использование расширяемого тезауруса вместо общего словаря. Основными задачами тезауруса являются: централизованное хранение элементов отдельных онтологий с сохранением их семантики, установления различного рода связей между терминами. Сохранение семантики является одной из ключевых особенностей данного подхода, и невозможно при использовании общей онтологии, определяющей некоторый общий смысл для различных объектов и отношений в отдельных информационных ресурсах, к которому приводиться смысл каждого термина. При этом часть его семантики теряется или искажается.

Тезаурус можно определить как четверку множеств – объектов, связей, атрибутов и агентов:

$$TRS =< O_U, L_U, P_U, A_U >, \qquad (4)$$

Дадим формальные определения элементов тезауруса. Понятие некоторой предметной области представляется в тезаурусе соответствующим ему элементом тезауруса типа «Объект»:

$$O =< N_O, L_O, P_O, A_O >, \qquad (5)$$

где $N_O$ – символьное имя объекта $O$, соответствующее названию представляемого им понятия, $L_O$ – множество связей, в которых состоит объект $O$, $P_O$ - множество свойств, характеризующих данный объект, $A_O$ – множество агентов, использующих данное понятие в представляемых ими онтологиях.

Связь между объектами тезауруса представим в виде:

$$L =< TP_l, O_1, O_2, W_1 >, \qquad (6)$$

где $TP_l$ – тип связи $L$, $TP_l \in TP_U$, $O_1$- первый объект, входящий в связь, $O_1 \in O_U$, $O_2$.- второй объект, $O_2 \in O_U$, входящий в связь, $W_1$ – вес связи $(W \in \mathrm{N}) \& (0 \le W \le 100)$, $O_U$ - множество всех объектов тезауруса.

Множество типов связей между объектами, представляющими термины, в тезаурусе:

$TP_U = \{synonymOf, hyponymOf, associateWith\}$.

Атрибут объекта онтологии предметной области, будет представлен в тезаурусе соответствующим элементом типа «Свойство», которое представим в виде:

$$P =< N_P, O, A_P >, \qquad (7)$$

где $N_P$ – символьное имя свойства $P$, соответствующее наименованию атрибута объекта онтологии предметной области, $O$ – объект тезауруса, который характеризует данное свойство, $O \in O_U$, $A_P$ – множество агентов.

### 3.2 Отображение онтологий в тезаурус

Данный подход предполагает задание отдельных информационных моделей - онтологий для каждого информационного ресурса, это позволяет учесть и отразить различные особенности семантики элементов данных. Далее производится процесс отображения концептов и отношений отдельных онтологий в разделяемый тезаурус. В ходе этого процесса каждому концепту, свойству и отношению онтологии, ставиться в соответствие элемент тезауруса, которому также приписывается идентификатор агента – приложения выполняющего различные задачи по обработке информации в отдельном информационном ресурсе. При этом между терминами, уже находящимися в тезаурусе, которые являются семантически близкими добавляемым, формируются взвешенные связи.

В тезаурусе межклассовые отношения представляются связями гипонимии:

$$L_0 =< hyponymOf, O_i, O_k, 100 >, \qquad (8)$$

Связи между терминами различных онтологий, такие как синонимия и ассоциация формируются на основании трех оценок:

- сходства семантики символических имен терминов;
- структурного положение понятия и термина в онтологии и тезаурусе;
- степени сходства множеств необходимых и достаточных атрибутов.

Данные оценки определяются следующими функциями:

$Syneq(O,T) = x$, (9)

где $O \in O_U$, $T \in T_U$, $0 \le x \le 100$.

Функция принимает объект онтологии и элемент тезауруса в качестве аргументов и возвращать степень сходства семантики символических имен. Она включает такие методы, как сравнения токенов имен терминов, определения расстояния между ними, вычисление близости определений терминов, сравнения синонимов терминов. Введем также предельные значения функции (9). Если значение функции превышает предельное значение, то два ее аргумента считаются эквивалентными:

$1 \le UPSYN \le 100$, (10)

если $Syneq(O,T) \ge UPSYN$, то $N_O = N_T$, $N_O$ и $N_T$ – символические имена объекта тезауруса или понятия онтологии.

Оценка сходства положений в иерархии терминов будет осуществляться функцией:

$Poseq(O,T) = x$ (11)

где $O \in O_U$, $T \in T_U$, $0 \le x \le 100$.

Функция (11) содержит такие методы выявления подобия, основанного на таксономическом положении терминов, как сравнение связанных путей, правило над/под термина, определение числа схожих надтерминов.

Следует отметить, что в данном случае представленные оценки являются эвристическими и поэтому использование связей на их основе возможно только для задач не требовательных к точности результата. К таким задачам можно, к примеру, отнести семантический поиск, результаты которого будут так или иначе обрабатываться экспертом, способным выявить различные неточности.

Формальную оценку сходства понятий дает функция сравнения множеств необходимых и достаточных атрибутов терминов:

$Atreq(O,T) = x$ (12)

где $O \in O_U$, $T \in T_U$, $0 \le x \le 100$.

Предельное значение функции (12) будет иметь вид:

$1 \le UPATR \le 100$, (13)

если $Atreq(O,T) \ge UPATR$, то объект тезауруса $O$ и понятие онтологии $T$ имеют близкие интерпретации.

Определение и роль необходимых и достаточных атрибутов терминов описываются далее в данной работе.

Рассмотрим процедуру включения элементов онтологии в тезаурус по шагам.

Шаг 1. Зададим начальные значения переменных-счетчиков: $n = 1$, $k = 1$, $l = 1$, $u_l = 1$. Пусть начальными для рассмотрения объектом тезауруса - $TRT$ и понятием онтологии - $TRO$, будут соответственно объект и понятие «Сущность»:

$TRO =< "Сущность", \varnothing, \varnothing, D_U, L_U >$, (14)

$TRT =< "Сущность", L_U, A_U >$, (15)

Переходим к шагу 2.

Шаг 2. С помощью функции семантического сопоставления имен (9) и сравнения необходимых и достаточных атрибутов (12) производим сравнение каждого элемента множества гипонимов $HYPO$ (17) объекта $TRO$ с каждым элементом множества гипонимов $HYPT$ (16) объекта $TRT$:

$HYPT_{TRT} = \{O_i \mid i \in \mathrm{N}\}$, (16)

где для каждого $O_i$ существует $L_0 =< hyponymOf, O_i, O_{TRT}, W >$,

$HYPO_{TRO} = \{T_i \mid i \in \mathrm{N}\}$, (17)

где каждый $T_i$ является прямым подклассом $TRO$.

Понятия онтологии, для которых обе функции возвратили значения, превышающие пороговые (10) и (13), формируют множество схожих понятий онтологии - $EQ_l$, остальные понятия попадают во множество несхожих - $NEQ_k$:

$EQ_l = \{T_i \mid i \in N\}$, (18)

где для каждого $T_i : (T_i \in HYPO_{TRO}) \& (\exists TA_j : (TA_j \in HYPT_{TRT})$
$\&(Syneq(T_i, TA_j) \ge UPSYN))$
$\&(Poseq(T_i, TA) \ge UPPOS), i \in \mathrm{N}$

$NEQ_k = \{T_i \mid i \in N\}$, (19)

где для каждого $T_i : (T_i \in HYPO_{TRO}) \& (\exists TA_j : (TA_j \in HYPT_{TRT})$
$\&(Syneq(T_i, TA_j) < UPSYN))$
$\&(Poseq(T_i, TA) < UPPOS), i \in \mathrm{N}$

Переходим к шагу 3.

Шаг 3. Если $n > |NEQ_k|$, тогда переходим к шагу 3.3, иначе создаем в тезаурусе элемент типа «Объект» - $P_n$, соответствующий понятию $T_n$: $T_n \in NEQ_k$ и переходим к шагу 3.1.

Шаг 3.1. С помощью функций (9) и (12) производим сопоставление $T_n$ со всеми объектами тезауруса. Если одна из функций возвратила значение, превышающее пороговое, для каких-либо двух аргументов, то производиться оценка их близости, исходя из положения в иерархии с помощью функции(11). В итоге в тезаурусе между созданным элементом $P_n$ и элементом, отобранным с помощью функций, создается ассоциативная связь с весом - $W$, равным среднему арифметическому трех оценок:

$L =< associateWith, P_n, F, W >$,

где $P_n, F \in O_U$ и для

$P_n, F : (Syneq(P_n, F) \ge UPSYN)$
$\&(Poseq(P_n, F) \ge UPPOS)$ ,

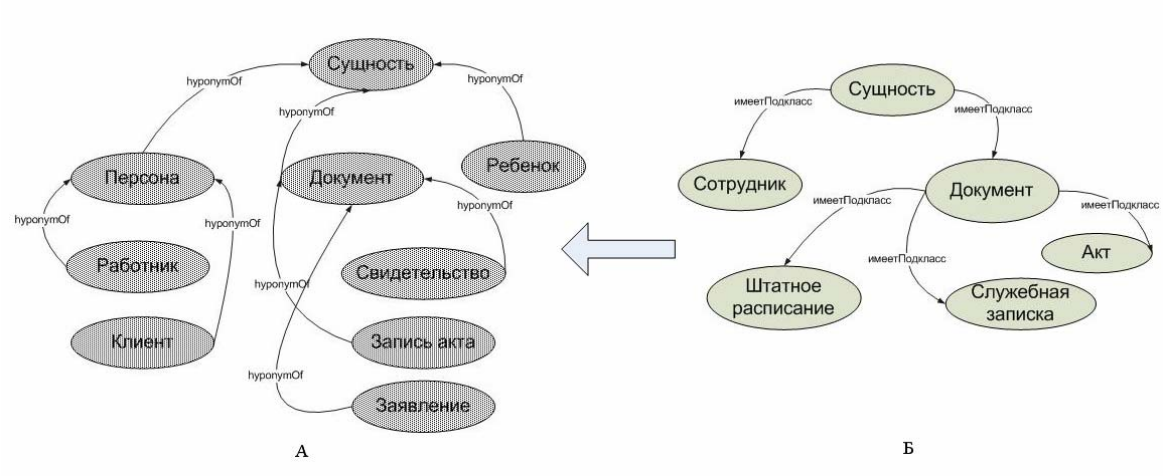



Рисунок 1. Расширение тезауруса терминами новой онтологии.

$$W = (s * Syneq(P_n, F) + p * Poseq(P_n, F) + a * Atreq(P_n, F)) / 3$$

где $s$, $p$, $a$ - коэффициенты от 0 до 1.

Переходим к шагу 3.2.

Шаг 3.2. Создаем связи гипонимии - $L$ объекта $P_n$ с объектами в тезаурусе, соответствующим его суперклассам в онтологии:

$$L =< hyponymOf, P_n, S, 100 >,$$

где $S$ объект тезауруса, представляющий понятие онтологии - надкласс для понятия $T_n$.

Далее инкрементируем счетчик $n$, переходим к шагу 3.

Шаг 3.3 Формируем новое множество $NEQ_{k+1}$, состоящее из понятий онтологии, являющихся непосредственными подклассами, понятий из множества $NEQ_k$:

$$NEQ_{k+1} = \{T_i \mid i \in N\},$$

где $T_i \in HYPO_H$, $H \in NEQ_k$

Далее инкрементируем счетчик $k$, счетчик $n$ сбрасываем в единицу. Переходим к шагу 3.4.

Шаг 3.4. Если $NEQ_k \neq \varnothing$, то переходим к шагу 3, иначе переходим к шагу 4.

Шаг 4. Если $l = 0$, то завершаем процедуру, иначе переходим к шагу 5.

Шаг 5. Если $u_l > |EQ_l|$, то декрементируем счетчик $l$, инкрементируем счетчик $u_l$, переходим к шагу 4, иначе добавляем агента, представляющего интегрируемую онтологию, к множеству агентов-представителей элемента тезауруса $H$, который признан синтаксически эквивалентным понятию онтологии $T_{ul}$, $T_{ul} \in EQ_l$:

$$H : Syneq(T_{ul}, H) \geq UPSYN$$

$$H : Atreq(T_{ul}, H) \geq UPATR$$

Переходим к шагу 6.

Шаг 6. Устанавливаем в качестве новых объектов для рассмотрения элемент тезауруса $H$ и соответствующие ему понятие онтологии $T_{ul}$:

$$TRO = T_{ul}, \text{ где } T_{ul} \in EQ_l$$

$$TRT = H, \text{ где}$$

$$H : (Syneq(T_u, H) \geq SYNEQ)$$

$$\&(Poseq(T_u, H) \geq UPPOS)$$

Переходим к шагу 7.

Шаг 7. Инкрементируем счетчик $l$, сбрасываем $u_l$ в единицу, переходим к шагу 2.

Основная идея алгоритма состоит в формировании новой ветви дерева терминов тезауруса, исходящей из вершины-корня, обозначающей предельную абстракцию «Сущность», если в тезаурусе отсутствует термин, вершина которого непосредственно связанна с корневой и который сопоставим с понятием онтологии, также непосредственно связанной с понятием «Сущность». В противном случае, то есть когда термин в тезаурусе признан эквивалентным понятию онтологии, их вершины сливаются. Далее по такому же принципу сравниваются их прямые потомки.

### 3.3 Пример работы алгоритма

Включение в тезаурус начальной онтологии является тривиальным, поэтому будем полагать, что в тезаурусе уже имеются термины какой-либо онтологии (рис. 1, А). Рассмотрим процедуру расширения тезауруса на упрощенном примере добавления в тезаурус новой онтологии (рис. 1, Б).

В начале работы алгоритма гипонимами термина «Сущность» в тезаурусе будут: «персона», «документ», «ребенок», а подклассами понятия «Сущность» в онтологии – «сотрудник» и «документ». В ходе сравнения семантики имен терминов и множеств необходимых и достаточных атрибутов с помощью функции (9) и (12), во множество схожих понятий онтологии (18) попадет – «документ», а несхожих (19) – «сотрудник».

Далее будут обработаны элементы множества несхожих понятий (19). В данном случае оно состоит из одного элемента – «сотрудник», который помещается в тезаурус, а далее с помощью функций (9) и (12) будет сравниваться с терминами

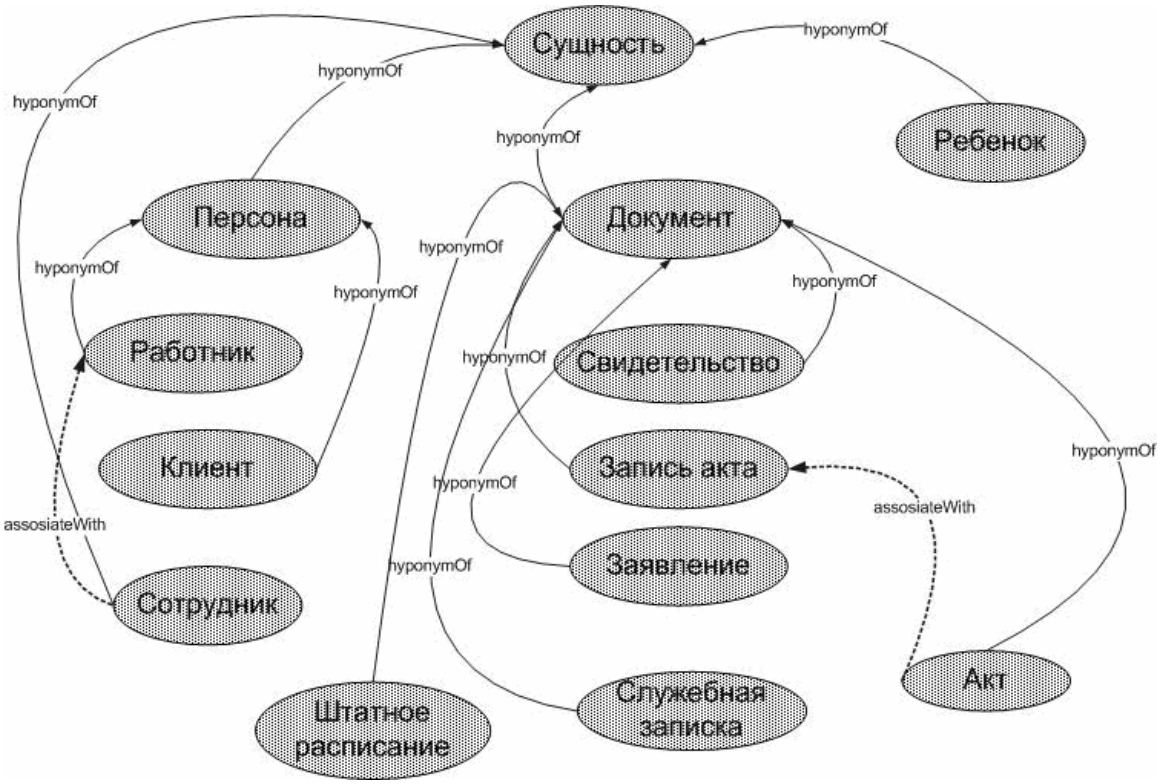



Рисунок 2. Тезаурус в результате работы алгоритма.

тезауруса. При достижении термина «работник», так как он является синонимом к «сотрудник», оба термина будут переданы в функцию (11). В зависимости от результирующей оценки между ними будет создана связь ассоциации с набранным весом или связь синонимии. Далее аналогичным образом обрабатываются гипонимы термина «сотрудник» и остальные элементы множества несхожих терминов, если таковые имеются. После этого будет обработан единственный элемент множества схожих терминов – «документ». Будут вновь определено множество гипонимов термина «документ» в тезаурусе, состоящее из элементов: «свидетельство», «запись акта», «заявление», и множество его гипонимов в онтологии: «штатное расписание», «служебная записка», «акт». В ходе сравнения с помощью функций (9) и (11) терминов данных множеств будут сформировано множество непохожих терминов(19), состоящее из элементов: «штатное расписание», «служебная записка», «акт», а множество похожих(18) в данном случае будет пустым. По рассмотренному ранее принципу будет обработан каждый элемент множества непохожих терминов, в результате чего все они будут включены в тезаурус, а между новым термином «акт» и «запись акта» будет создана связь ассоциации. Результирующий вид тезауруса представлен на рисунке 2.

### 3.4 Определение и использование набора идентификационных атрибутов в тезаурусе

Одной из ключевых проблем семантической интеграции является сопоставление моделей представление данных. В данном случае она частично решается на этапе интеграции онтологий в тезаурус. Однако это не позволяет избавиться от ошибок и неточностей, что довольно критично для таких задач, как проверка семантической целостности и получение совокупной информации об объекте из различных источников. Это связано с тем, что интерпретации концептов и отношений явно в онтологиях не заданы, поэтому получить и сравнить их не представляется возможным. Также часто складывается такая ситуация, что один концепт может иметь интерпретацию концепта другой онтологии, не противоречащую системе аксиом первой, но по-сути иметь иной подразумеваемый смысл. Например, рассмотрим следующие наборы OWL аксиом, двух различных онтологий (для записи используется манчестерский синтаксис[8]):

Class: Person

SubClassOf: owl:Thing that hasFirstName only string[minLength 1]

Class: Cat

SubClassOf: owl:Thing that hasName only string[minLength 1]

Формально эти два концепта эквиваленты, однако, как видно из названий классов, они подразумевают разные интерпретации, которые нельзя отличить, используя лишь данные наборы аксиом.

Разумеется, степень формальной выразительности зависит от сложности онтологии с точки зрения количества заданных аксиом, в большей степени типа Abox (assertional box), используя которые, можно с помощью машины

вывода получить формальную семантику. Но разработка сложных онтологий, а не просто таксономий терминов, требует от эксперта знания не только предметной области, но и принципов и языков онтологического моделирования, что, как правило, не выполняется.

Проблему установления соответствий можно отчасти решить применением онтологии верхнего уровня, определяющей абстрактные концепты и отношения, посредством привязки к которым можно установить соответствия между элементами разных онтологий. Однако здесь возникает вопрос об уровне абстракции концептов общей онтологии, так как их интерпретации могут быть довольно большими, что не позволить разнести по ним концепты отдельных онтологий. Если же добавлять в общую онтологию дополнительные аксиомы, сужающие множества интерпретации, то это будет накладывать дополнительные ограничения на определение концептов и отношений включаемых онтологий.

Альтернативным способом согласования может служить общий словарь терминов, постулирующий общую семантику. Однако то или иное определение, заданное в нем, может быть как неоднозначным, так и не полным, что вызовет проблемы в его формализации в виде модели объекта в онтологии. Проблемой также является будущее изменение словаря таким образом, чтобы новые определения не противоречили имеющимся.

Наряду с этим, как было отмечено ранее, остается нерешенной проблема конфликтов на уровне экземпляров – моделей наиболее конкретных сущностей. Определение какого-либо соответствия между концептами разных онтологий, например, эквивалентности, означает наличие этого отношения только между множествами их интерпретаций, а не их элементами.

Иными словами это означает, что отсутствует возможность установить соответствие между экземплярами разных онтологий, интерпретации которых представляют один и тот же объект реального мира. Для рассматриваемой предметной области это является довольно серьезной проблемой, так как в данном случае информационные источники часто содержат данные, описывающие одну и ту же сущность. Однако обработку этих данных необходимо вести совместно во избежание появления различного рода противоречий.

Для разрешения данной трудности предлагается задать множество общих атрибутов-идентификаторов у экземпляров в различных онтологиях. Это позволит устанавливать отношение эквивалентности между ними. Проблему определения атрибута и присвоения ему значения, уникального в рамках множества экземпляров всех онтологий, можно решить, если, проанализировав предметную область, выявить реальное идентификационное свойство, которое, так или иначе, уже заданно в контенте информационных ресурсов.

В данном случае предлагается использовать специфику области государственного и муниципального управления заключающуюся в том, что основные субъекты и объекты имеют заранее заданные в различных юридических документах наборы основных атрибутов, в том числе и идентификационных. При этом будут также выполнены основные требования к общезначимому атрибуту-идентификатору.

Использование общего идентификатора позволяет «склеить» различные кусочки информации для того, чтобы создать некое интегрированное представление определенного объекта реального мира.

Однако в рассматриваемой предметной области один и тот же идентификационный атрибут может использоваться для определения экземпляров, относящихся к разным классам. Например, индивидуальный номер налогоплательщика, может обозначать как гражданина, так и предприятие. В сущности, определяются два различных множества значений других атрибутов: в первом случае ФИО, адрес местожительства, год рождения, а во втором – название предприятия, юридический адрес. Для разрешения данной проблемы можно дополнительно обозначить общие описательные атрибуты, к которым предъявляется только требование общезначимости. Наличие их у экземпляра, можно считать достаточным условием для членства в определенном классе.

В результате, определив наборы общезначимых идентификационных и описательных атрибутов в тезаурусе, можно использовать их для задания концептов, а также для установления соответствия между ними с помощью функции (12). Также это позволяет устранить проблему семантических конфликтов и неопределенностей на уровне экземпляров и в то же время не накладывает ограничений на отдельные онтологические модели. Наряду с этим, задание общих атрибутов позволяет формально определить некие базовые классы в тезаурусе, которые можно конкретизировать в частных онтологиях, что позволяет сократить время разработки, задав общую модель определенных сущностей.

## 4 Текущие результаты и направления дальнейшей работы

В данной работе представлен подход к семантической интеграции данный в сфере государственного муниципального управления с использованием разделяемого тезауруса. На данный момент определена обобщенная структура системы интеграции и назначение ее функциональных модулей[4]. Задана концептуальная модель тезауруса, а также разработан алгоритм отображения онтологий в тезаурус, включающий оценки семантической близости концептов. Для

решения проблемы установления формального соответствия между экземплярами различных онтологий представлена методика определения и использования общих атрибутов.

Среди направлений дальнейшей работы можно выделить: имплементацию тезауруса в виде онтологии на языке OWL или в виде RDF респозитория, разработка прототипов онтологических моделей учреждений государственного и муниципального управления, выявление и определение в тезаурусе идентификационных атрибутов основных сущностей предметной области и задания прав доступа к ним, разработка языка запросов между агентами-интеграторами на основе языка запросов к RFD документам – SPARQL.

## Литература


[1] Богдановская И. Ю. Концепция «электронного государства», 2006. http://www.ifap.ru/pi/04/r02.doc

[2] Виттих В. А., Волхонцев Д. В., Горбенко А. В., Гриценко Е. А., Кистанов А. М., Светкина Г. Д., Скобелев П. О., Сурнин О. Л., Шамашов М. А., Мультиагентный Интернет-портал для интеграции ресурсов департаментов социального блока Самарской области, 2006. http://www.kg.ru/support/library/portal

[3] Виттих В.А., Волхонцев Д.В., Гинзбург А.Н., Караваев М.А., Скобелев П.О., Сурнин О.Л., Шамашов М.А., Распределенные онтологии и их применение в решении задач интеграции данных, http://www.kg.ru/support/library/dataintegration/

[4] Ломов П.А., Шишаев М.Г. Семантическая интеграция информационных источников для информационной поддержки управления микросистемой. – VII Всероссийская школа-семинар «Прикладные проблемы управления макросистемами». Апатиты, 31 марта - 4 апреля 2008г. / Материалы докладов. – Апатиты: изд-во КНЦ РАН , 2008. С.25-27.

[5] Gruber, T.R. (1993) A translation approach to portable ontology specifications. Knowledge Acquisition. Vol. 5

[6] OWL - Web Ontology Language. Overview, 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/

[7] OWL 2 - Web Ontology Language. Primer, 2009. http://www.w3.org/TR/owl2-primer/

[8] OWL 2 - Web Ontology Language Manchester Syntax, 2009. http://www.w3.org/TR/owl2-manchester-syntax

[9] Resource Description Framework, http://www.w3.org/RDF/

[10] Visser U., Stuckenschmidt H., Wache H., Vogele U., «Enabling Technologies for Interoperability» – Режим доступа: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=102D69688CD1F1200F311A2460DE6B5A?doi=10.1.1.21.5883&rep=rep1&type=pdf

[11] Wache H., Scholz T., Stieghahn H., Кцnig-Ries B., «An Integration Method for the Specification of Rule-Oriented Mediators.» In Proc. of 1999 International Symposium of Database Applications in Non-Traditional Environments (DANTE 99), Kyoto, Japan, November 1999


## Development of the method of semantic integration of the information in sphere of the state and municipal administration


Lomov P. A., Shishaev M. G.



This paper offers the approach of semantic integration of information in domain of the state and municipal administration with using shared thesaurus, which allows to eliminate some critical for the considered subject domain disadvantages of existing approaches. The conceptual model of the thesaurus was presented. The procedure of mapping concepts of ontologies in the thesaurus was described. Also, the technique of the definition and comparison of ontological contexts on the base of a set of the general attributes was introduced.



♣ Работа поддержана грантом РФФИ, проект № 08-07-00301-а

# Исследование и оптимизация параметров алгоритма Manifold Ranking на основе метрики автоматической оценки качества обзорного реферирования ROUGE-RUS

© С.Д. Тарасов

Балтийский Государственный Технический Университет им. Д.Ф.Устинова «ВОЕНМЕХ»

**Аннотация**

В статье рассматривается используемая в DUC метрика *ROUGE*, а также выполненная авторами ее модификация для русского языка *ROUGE-RUS.* Разработана система для автоматической оценки качества обзорного реферирования на основе *ROUGE-RUS* с Web-интерфейсом. Проведен эксперимент по составлению ручных аннотаций новостных кластеров и автоматической оценке качества обзорного реферирования по метрике *ROUGE-RUS*. Введено понятие базовой величины *ROUGE-RUS* на кластере. Рассмотрен алгоритм обзорного реферирования *Manifold Ranking*. Проведен анализ степени влияния выбора базовых параметров и начальной темы на результат работы алгоритма. На основе результата исследований влияния выбора темы на работу алгоритма разработан модифицированный алгоритм.

## Введение

Задача автоматического построения обзорных рефератов на сегодняшний день является очень актуальной. Это вызвано, в первую очередь, необходимостью в условиях постоянного роста информации знакомить специалистов и других заинтересованных людей с необходимыми им документами, представленными в сжатом виде, но с сохранением смысла. В обзорной статье [1] описывается современное состояние в области автоматического реферирования, а также основные направления и пути развития.

На текущий момент существует огромное количество различных методов получения обзорных рефератов. В традиционных методах реферирования чаще всего используются различные модификации подхода Г. Луна [3], известного с конца 50-х годов XX века, который заключается в отборе предложений с наибольшим весом для включения их в реферат. Вес предложения определяется как сумма частот, входящих в него значимых слов. В работе [7] описан метод, в котором в качестве значимых элементов выбираются не слова, а словосочетания. В работе [8] представлены методы обзорного реферирования с использованием концептов тезауруса. К наиболее перспективным можно отнести методы, описывающие связную модель текста документов с помощью формального математического аппарата. Данные методы, как правило, не привязаны к особенностям конкретного языка, не требуют большого количества лингвистических ресурсов. К таким методам относятся метод регрессии опорных векторов, метод ранжирования связных структур [5, 9], а также ряд других.

Не менее актуальной является и задача оценки полученных автоматических рефератов. Несомненно, наиболее правдоподобные оценки качества можно проводить в ручном режиме с привлечением большого числа экспертов. Однако такие ручные оценки являются чрезвычайно дорогими. Методики автоматической оценки качества реферирования не только делают этот процесс более доступным, но и позволяют в реальном времени производить настройку параметров работы определенного алгоритма, производить их оптимизацию.

## 1 Оценка качества обзорного реферирования

На сегодняшний день предложено огромное количество методов обзорного реферирования. Работа каждого метода определяется некоторым набором внешних условий; кроме того, каждый метод содержит набор параметров, подбор которых изменяет качество реферирования при заданных условиях в широком диапазоне. Эти параметры, как правило, определяются для заданных внешних условий эмпирическим путем. В связи с этим, одним из немаловажных вопросов является оценка качества обзорного реферирования. Это позволяет в реальном времени производить настройку параметров работы конкретного алгоритма, производить оптимизацию этих параметров, сравнивать эффективность разных алгоритмов, а также делать окончательный вывод о возможности

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

практического применения данного алгоритма автоматического реферирования.

Традиционные методы оценки качества обзорного реферирования включают в себя оценку обзорного реферата по ряду критериев специалистами-лингвистами. К основным критериям относятся связность, краткость (лаконичность), грамматическая правильность, сложность восприятия, содержание. Однако даже простая ручная оценка качества обзорного реферирования по нескольким критериям требует больших объемов человеческих ресурсов (согласно DUC, более 3000 часов работы лингвистов), что является очень дорогим. Кроме того, нет возможности проводить оценку качества в «реальном времени», например, при оптимизации работы некоторого метода реферирования. В связи с этим, вопрос о возможности автоматизации оценки качества является очень актуальным. За последние несколько лет было предложено несколько методик автоматической оценки качества обзорного реферирования. Все они основаны на автоматическом сравнении реферата, полученного с помощью метода автоматического реферирования, с одним или несколькими обзорными рефератами, составленными экспертами. В этом случае, так или иначе, критерием качества можно считать «схожесть» автоматического реферата с ручным. В качестве меры «сходства» в DUC были предложены «cosine similarity», «unit overlap(unigram or bigram)», «longest common subsequence».

## 2 ROUGE: метрика для автоматической оценки качества обзорного реферирования

Одной из наиболее удачных реализаций систем для автоматической оценки качества обзорного реферирования можно считать пакет *ROUGE* [6], используемый в DUC. Набор программ позволяет автоматически рассчитывать различные метрики *ROUGE* (*Recall-Oriented Understudy for Gisting Evaluation*): *ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, ROUGE-SU*. К наиболее часто используемым в DUC относятся *ROUGE-1* и *ROUGE-2*.

### 2.1 ROUGE-N

Метрика *ROUGE-N* представляет собой обобщенную статистическую меру, выражающую какой процент лексических единиц (*N-gram*,-последовательностей из *N* лексем), входящих в состав ручного, построенного независимым экспертом, реферата, повторяется в автоматическом реферате:

$$ROUGE-N=\frac{\sum_{S\in RefSum}\sum_{n-gram\in S}Count_{match}(n-gramm)}{\sum_{S\in RefSum}\sum_{n-gram\in S}Count(n-gramm)}.$$

В случае использования нескольких ручных рефератов для оценки автоматического, в [2] предлагается сравнивать автоматический реферат с каждым ручным по метрике *ROUGE-N*, а затем выбирать максимальное значение.

$$ROUGE-N_{multi}=\arg\max_i \mathrm{ROUGE}\text{-}\mathrm{N}(r_i, s),$$

где $r_i$ – *i*-й ручной реферат, *s* – оцениваемый автоматический реферат. Такая же процедура используется и для *ROUGE-L*.

Важно отметить, что *ROUGE* является метрикой полноты, и соответственно несимметричной относительно сочетаний ручной - автоматический реферат.

### 2.2 ROUGE-L

Для оценки степени совпадения автоматического реферата с ручным, также используется метод «наибольшей совпадающей подпоследовательности»[2]. Величина *LCS* (*Longest Common Subsequence*) представляет собой длину наибольшей подпоследовательности между двумя предложениями *X* и *Y*. В качестве элементов последовательностей выбираются лексемы. При вычислении величины *ROUGE-L* для автоматического реферата, содержащего *v* предложений (всего *n* слов) и ручного реферата, содержащего *u* предложений (всего *m* слов), производится вычисление объединенной *LCS* между каждым предложением ручного реферата $r_i$ и всеми предложениями автоматического $c_j$.

$$R_{lcs}=\frac{\sum_{i=1}^{u}LCS_{\cup}(r_i,C)}{m},\ P_{lcs}=\frac{\sum_{i=1}^{u}LCS_{\cup}(r_i,C)}{n}$$

$$ROUGE-L=F_{lcs}=\frac{(1+\beta^2)R_{lcs}\cdot P_{lcs}}{R_{lcs}+\beta^2\cdot P_{lcs}},$$

где $LCS_{\cup}(r_i,C)$ - длина наибольшей подпоследовательности между предложением ручного реферата $r_i$ и всеми предложениями автоматического реферата *C*[2]. В DUC полагается $\beta\to\infty$, таким образом, учитывается только $R_{lcs}$-составляющая.

В [2] на основе кластеров из DUC 2001, 2002, 2003, 2004 метрики *ROUGE* показали высокую степень корреляции (Коэффициент Пирсона до 0.99) с ручными оценками. Это позволяет использовать метрики *ROUGE* для автоматической оценки качества обзорного реферирования, для сравнительной оценки различных методов, а также для оптимизации работы отдельно взятого метода.

Пакет программ, реализующий автоматическую оценку *ROUGE*, является свободно распространяемым набором *Perl*-скриптов, однако ориентирован только на использование кластеров в формате DUC, а также на документы на английском языке. Это создает определенные трудности при попытке оценки по метрике *ROUGE* русскоязычных обзорных рефератов. Кроме того, «отличные» результаты, полученные в DUC для англоязычных

кластеров, вовсе не свидетельствуют о таких же результатах для кластеров на русском языке.

# 3 ROUGE-RUS

С учетом недостатков существующего пакета *ROUGE* была разработана модифицированная метрика *ROUGE-RUS*, обладающая следующими отличительными особенностями:

- Русская морфология, список стоп-слов для русского языка;
- Возможность учитывать синонимы (с использованием концептов тезауруса);
- Усреднение (а не максимум) значения при наличии нескольких ручных аннотаций. По полученным результатам это позволяет сделать метрику более устойчивой и использовать меньшее количество ручных рефератов.

Для расчета метрики *ROUGE-RUS* была разработана система с Web-интерфейсом, являющаяся частью системы "MDS Evaluation Framework"[9].

# 4 Эксперимент по оценке метрики ROUGE-RUS

Для изучения свойств метрики *ROUGE-RUS*, а также для исследования возможности использования ее для автоматической оценки качества обзорного реферирования был проведен следующий эксперимент.

## 4.1 Исходные данные

В качестве исходных данных были взяты новостные кластеры различной тематики («Россия», «Происшествия», «Спорт», «Культура» и др.) из системы «Google.News» за конец ноября – начало декабря 2008 года. Всего было обработано 67 кластеров из 613 документов, полученных из 21 источника («РБК», «РИА Новости», «Российская Газета» и т.д.).

## 4.2 Построение ручных рефератов

К построению ручных рефератов были привлечены студенты 5 курса БГТУ «ВОЕНМЕХ», обучающиеся по специальности «Автоматизированные и управляющие системы». Всего в исследовании приняло участие 67 человек. Каждому участнику было предложено составить одну аннотацию для каждого из случайно выбранных 50 кластеров. При составлении ручного реферата для каждого кластера участник должен был выбрать 4 различных предложения из всех документов кластера. В результате было получено 2385 ручных аннотаций.

## 4.3 Исследование метрики ROUGE-RUS на наборе ручных рефератов

Далее было отобрано $N=50$ кластеров, для которых имелось по $M>40$ ручных аннотаций, порожденных разными пользователями.

$$A_i \in A, 1 \le i \le N,$$

$$A_i^j \in A_i, 1 \le j \le M.$$

Для исследования метрики *ROUGE-RUS*, были произведены вычисления величин *ROUGE-1*, *ROUGE-2*, *ROUGE-3*, *ROUGE-4* и *ROUGE-L* для каждой пары аннотаций из множества $A_i$ по всем кластерам:

$$RR_i^{l,m} = ROUGE - RUS(A_i^l, A_i^m), 1 \le l, m \le M, 1 \le i \le N,$$

где $A_i^l$ выступает в роли ручной, а $A_i^m$ в роли автоматической аннотации, оцениваемой по метрике. Цель такого исследования – выявить особенности метрики, оценить распределение ее величины на множестве «заведомо хороших» ручных аннотаций, определить минимальное необходимое количество ручных аннотаций для стабильной оценки одной автоматической. Для произвольно взятой аннотации $A_i^j \in A_i$ распределение величины *ROUGE-RUS*( $A_i^l$ , $A_i^j$ ) имеет следующий вид (см. Рис. 1). Для всех кластеров распределение имеет примерно такой же вид. Из этого следует, что:

[1] Ручные рефераты, порожденные разными пользователями, слабо согласуются друг с другом.

[2] Использование одного ручного реферата для оценки недостаточно.

[3] В ручных рефератах, порожденных разными пользователями, практически отсутствует кластеризация. Если таковая и имеет место быть, то, как правило, не в области максимума, а где-то «посередине».

[4] Использование морфологии, списка стоп-слов и словаря синонимов положительно сказывается на пологости кривой, что обеспечивает меньший разброс величины для разных ручных рефератов в пределах одного кластера.

Отсутствие кластеризации, особенно в области максимального значения, говорит о том, что использование метода максимума, предложенного в [6], является неоправданным. В этом случае во внимание, фактически, принимается только одна ручная аннотация (наиболее близкая к автоматической), что при таком разбросе значения величины не является допустимым. Исходя из этих соображений, способ усреднения должен рассматриваться как основной для учета нескольких ручных рефератов при оценке одного автоматического.

Далее были сформированы выборки величины *ROUGE-RUS* для $K=1..10$, где $K$ – количество ручных аннотаций, принимаемых в расчет при оценке одной автоматической. В этом случае из

множества $A_i^j \in A_i$ выбиралась *1* аннотация, как автоматическая, подлежащая оценке, и *K* из оставшихся, как ручные. Были использованы следующие вычисления:

$$RR_i^{l,m} = ROUGE - *(A_i^l, A_i^m), 1 \le l, m \le 40, 1 \le i \le 50,$$

а затем вычислялся максимум и среднее для данного значения *K*. После этого были сформированы выборки из величин *ROUGE-RUS* для каждого *K*, и было произведено усреднение по всему множеству кластеров. Таким образом, были оценены такие параметры, как: зависимость среднего, минимального и максимального значения, дисперсии, ср. кв. откл. величины *ROUGE-RUS* от *K*. Результаты представлены в Таблице 1. Следует отметить, что дисперсия величины не рассматривалась как критерий качества самой метрики, однако ее минимализация необходима для того, чтобы метрика была более стабильной.

Таким образом, можно считать, что использование усреднения с учетом слабого согласия ручных аннотаций друг с другом, дает более стабильный результат. В этом случае для оценки одной автоматической аннотации достаточно 4-5 ручных.

## 5 Исследование и оптимизация параметров алгоритма обзорного реферирования на основе метрики ROUGE-RUS

Метрики автоматической оценки, к которым относится *ROUGE*, могут быть использованы, в первую очередь, для исследования влияния различных параметров на качество аннотаций, порождаемых алгоритмами обзорного реферирования, а также для оптимизации этих параметров. Авторами была произведена подробная оценка работы алгоритма *Manifold Ranking* для русского языка [9], а также сравнение результатов работы с *«Basic Lines»* по метрике *ROUGE-RUS* и исследование влияния различных параметров на работу алгоритма.

### 5.1 Алгоритм ранжирования связных структур для задачи обзорного реферирования

Алгоритм *Manifold Ranking*[5] позволяет описать связную структуру текста при помощи матриц. Изначально алгоритм предполагает выделение элементов (предложений) наиболее близких заданному (теме). Такая интерпретация характерна задаче информационного поиска. Для автоматического реферирования также выделяется набор предложений, наиболее близких заданной теме кластера, однако обязательным является применение алгоритма отсечения «похожих» предложений, что особенно актуально для многодокументного аннотирования.

Автоматическое реферирование набора документов с использованием алгоритма ранжирования связных структур состоит из двух этапов:

[1] Вычисление ранга каждого предложения. Этим решается задача ранжирования всех предложений в соответствии с их «близостью» заданной теме кластера.

[2] Применение алгоритма отсечения предложений, наиболее похожих на те, что уже попали в обзорный реферат. Этим решается задача исключения из обзорного реферата одинаковых или близких предложений.

Основной особенностью алгоритма ранжирования связных структур является учет внутренней связной структуры объектов, составляющих текст. Объекты должны быть представлены векторами в Евклидовом пространстве. В этом случае полагается, что «близость» двух объектов представленных векторами может быть вычислена, как Евклидова мера или скалярное произведение векторов. Целью алгоритма является упорядочить объекты, с учетом их внутренних связей между собой.

Формально, связная структура объектов представляется как некий взвешенный граф, вершинами которого являются сами объекты, а в качестве весов дуг задаются евклидовы расстояния между ними. Алгоритм ранжирования заключается в постепенном распространении объектами своего ранга на смежные объекты-вершины. Таким образом, ранг $f_i$ каждого предложения $x_i$ вычисляется не только с учетом «близости» его к эталонному объекту (теме кластера *T*), но и с учетом связной структуры текста, т.е. ранг «распространяется» по графу с учетом весов связей структур. В результате некоторое количество предложений с наибольшим рангом выбирается для результирующего реферата.

### 5.2 Сравнение результатов MR с результатами «ручная-ручная». Базовая величина метрики ROUGE-RUS на кластере

Для оценки результатов работы алгоритма по метрике *ROUGE-RUS* было введено понятие «Базовой величины метрики *ROUGE-RUS* на кластере». В качестве этой величины было использовано усредненное и максимальное значение *ROUGE-RU*S для данного кластера. Для оценки каждой автоматической аннотации, порождаемой алгоритмом было выбрано 10 ручных, построенных разными пользователями. Таким образом, результат работы алгоритма для заданного кластера оценивался относительно этих двух базовых величин (среднее и максимум).

При значении параметров из [5, 9] базовый алгоритм *Manifold Ranking* показал, в среднем, результаты, несколько худшие, чем базовые оценки ручных с ручными.

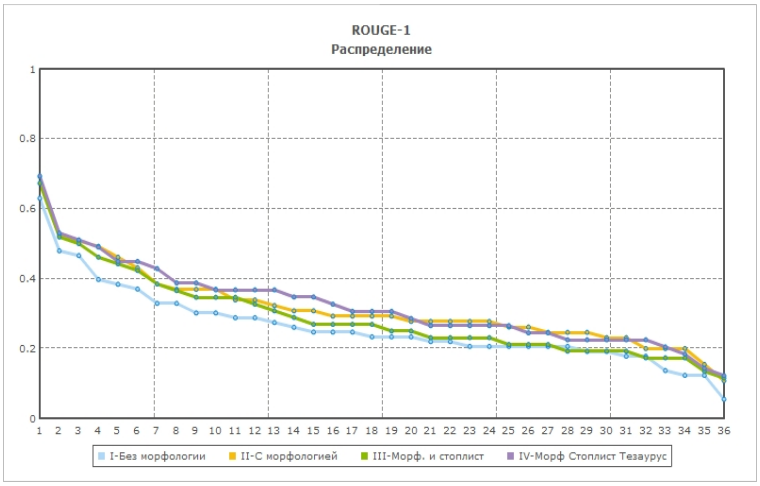



Рисунок 1 - Распределение величины *ROUGE-1* при сравнении одной ручной аннотации со всеми остальными для произвольного кластера. Значения отсортированы по убыванию.

Таблица 1 - Относительный разброс величин *ROUGE-RUS* при различных значениях *K*

| *K* | δ, Метод Максимума, % | | | | | δ Метод усреднения, % | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *ROUGE-1* | *ROUGE-2* | *ROUGE-3* | *ROUGE-4* | *ROUGE-L* | *ROUGE-1* | *ROUGE-2* | *ROUGE-3* | *ROUGE-4* | *ROUGE-L* |
| 1 | 42.18 | 79.31 | 108.69 | 127.03 | 45.90 | 42.18 | 79.31 | 108.69 | 127.03 | 45.90 |
| 5 | 28.52 | 41.47 | 49.00 | 56.47 | 30.25 | 25.10 | 42.57 | 57.64 | 67.85 | 26.59 |
| 10 | 24.58 | 33.24 | 37.91 | 43.38 | 26.19 | 22.05 | 35.39 | 47.55 | 56.23 | 23.07 |

### 5.3 Сравнение результатов MR с «Basic Lines»

Также были проведены оценки с псевдо-автоматическими аннотациями *«Basic Lines»*:

[1] **BL1** – 4 первых предложения 1-го документа,
[2] **BL2** – первые предложения 4-х первых документов,
[3] **BL3** – последние предложения 4-х первых документов,
[4] **BL4** – заголовки 4-х первых документов,
[5] **BL5** – 4 первых предложения последнего документата,
[6] **BL6** – последние предложения 4-х первых документов,
[7] **BL7** – последние предложения 4-х последних документов,
[8] **BL8** – заголовки 4-х последних документов.

### 5.4 Оптимизация общих параметров

К общим параметрам алгоритма можно отнести: $\alpha$ (определяет относительный вклад близких предложений в ранк текущего и начальный ранк каждого предложения), $\lambda_1$ (коэффициент учета веса связности предложений из одного документа), $\lambda_2$ (коэффициент учета веса связности предложений из разных документов), $\omega$ (коэффициент усечения сходных предложений). В среднем, по кластерам, были получены следующие значения оптимальных параметров: $\alpha$=0.9, $\lambda_1$=0.3, $\lambda_2$=0.8, $\omega$=10. Полученные значения несколько отличаются от использованных в DUC [5]. Это связано в первую очередь со спецификой взятых новостных кластеров, а также с особенностями русского языка. Кроме того, нет сведений о том, проводили ли авторы алгоритма подбор этих параметров или значения были взяты исходя из эмпирических соображений.

На имеющихся в наличии кластерах была выявлена сильная устойчивость алгоритма к значениям базовых параметров. С одной стороны, это хорошо, т.к. значительно упрощается задача подбора оптимальных значений этих параметров, с другой, нет возможности управлять работой алгоритма, изменяя в широком диапазоне значения параметров.

### 5.5 Ограничение длины документов

В отобранных новостных кластерах существует большое количество «очень длинных» документов, содержащих более 30 предложений. Учитывая новостной принцип «перевернутой пирамиды», предложения, настолько удаленные от начала документа, как правило, не несут в себе большой смысловой нагрузки и, как кандидаты для включения в обзорный реферат, не представляют большого интереса. Было проведено исследование влияния данного параметра на работу алгоритма. При этом документы усекались до 200, 50, 20, 15, 10, 7, 5, 4 предложений. В среднем, лучшие результаты были получены при укорачивании документов до 10 предложений (см. Таблицу 2).

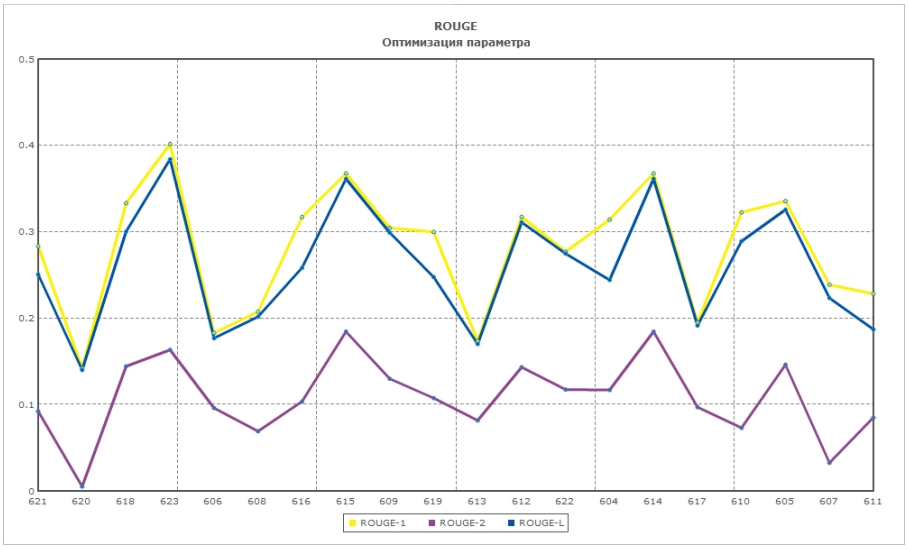



Рисунок 2 - Алгоритм *Manifold Ranking* является очень чувствительным к выбору темы. По горизонтальной оси отложены номера документов, заголовок которых был использован в качестве темы

Таблица 2 – Укорачивание длины документов

| **Кол-во предложений** | ***ROUGE-1*** | ***ROUGE-2*** | ***ROUGE-L*** |
|---|---|---|---|
| 200 | 0.32 | 0.11 | 0.27 |
| 50 | 0.32 | 0.11 | 0.27 |
| 20 | 0.36 | 0.12 | 0.30 |
| 15 | 0.36 | 0.12 | 0.30 |
| 10 | 0.42 | 0.18 | 0.39 |
| 7 | 0.40 | 0.17 | 0.37 |
| 5 | 0.38 | 0.16 | 0.36 |
| 4 | 0.38 | 0.16 | 0.36 |

Таблица 3 – Модификация алгоритма: выбор темы

| **Выбор темы** | ***ROUGE-1*** | ***ROUGE-2*** | ***ROUGE-L*** |
|---|---|---|---|
| Заголовок одного документа | 0.04-0.48 | 0.00-0.22 | 0.04-0.41 |
| Заголовки всех документов | 0.37 | 0.09 | 0.29 |
| Заголовки из первых двух документов | 0.31 | 0.07 | 0.23 |
| Заголовки из первых четырех документов | 0.18 | 0.06 | 0.17 |
| Заголовки из последних двух документов | 0.40 | 0.06 | 0.27 |
| Заголовки из последних четырех документов | 0.38 | 0.08 | 0.29 |

Таблица 4 - Результаты

| | ***ROUGE-1*** | ***ROUGE-2*** | ***ROUGE-L*** |
|---|---|---|---|
| Базовое значение метрики (среднее) | 0.28 | 0.11 | 0.25 |
| Базовое значение метрики (максимум) | 0.50 | 0.30 | 0.47 |
| Базовый алгоритм *MR* | 0.18 | 0.10 | 0.18 |
| *BL-1* | 0.33 | 0.16 | 0.32 |
| *BL-2* | 0.34 | 0.16 | 0.33 |
| *BL-3* | 0.28 | 0.08 | 0.24 |
| *BL-4* | 0.20 | 0.08 | 0.20 |
| *BL-5* | 0.01 | 0.00 | 0.01 |
| *BL-6* | 0.50 | 0.28 | 0.47 |
| *BL-7* | 0.30 | 0.15 | 0.25 |
| *BL-8* | 0.33 | 0.13 | 0.30 |
| **Модифицированный алгоритм** | **0.46** | **0.19** | **0.42** |

### 5.6 Исследование влияния выбора темы на работу алгоритма обзорного реферирования Manifold Ranking

Общеизвестным фактом является сильная чувствительность практически любого алгоритма обзорного реферирования к выбору начальной темы кластера. Так как алгоритм *Manifold Ranking* по определению является *«topic focused»*, то был получен действительно довольно большой разброс значений *ROUGE-RUS* для разных тем (Рис. 2).

Т.к. базовый алгоритм предполагает использование одной темы, то авторы попытались выявить зависимость величины *ROUGE-RUS* от выбора темы по следующим критериям:

[1] Дата публикации документа, откуда выбирается тема
[2] Кол-во слов в предложении темы
[3] Кол-во существительных в предложении темы

На имеющихся кластерах нам не удалось выявить различимой закономерности между выбором темы по вышеуказанным критериям и значениями величины *ROUGE-RUS*. Исходя из этого, проанализировав базовый алгоритм ранжирования абстрактных связных структур [6], авторы нашли возможность использовать несколько тем (предложений) как элементов, являющихся источником ранка.

$$y = [y_0, y_1, .. y_n]^T,$$
$$y_i = 1, i \in (0, n),$$

если $x_i$ – предложение, отмеченное как тема, и

$$y_i = 0, i \in (0, n),$$

для всех остальных предложений. Были рассмотрены несколько вариантов модифицированного алгоритма в отношении использования нескольких тем:

[1] Заголовки всех документов
[2] Заголовки из первых двух документов
[3] Заголовки из первых четырех документов
[4] Заголовки из последних двух документов
[5] Заголовки из последних четырех документов.

На имеющихся у авторов кластерах, в среднем, было получено, что наилучшую оценку дает использование всех тем и тем из последних документов; наихудшую – использование одной темы (нестабильно) и тем из первых документов (см. Таблицу 3). Кроме того, была предпринята попытка использовать в качестве тем не заголовки документа, а первое и второе предложения, т.к. в новостях заголовок зачастую призван привлечь внимание, и не всегда достоверно отражает суть вопроса. Однако никакие изменения в сторону повышения качества при использовании с учетом игнорирования тем при использовании заголовков получены не были.

### 5.7 Результаты подбора параметров

В результате подбора параметров и модификации алгоритма для использования нескольких тем авторам удалось получить результаты оценки мо метрике *ROUGE-RUS,* превосходящие базовую величину метрики на кластере. Результаты оптимизации параметров показаны в Таблице 4.

## Заключение

Задача автоматического построения обзорных рефератов на сегодняшний день является очень актуальной. Не менее актуальной является и задача оценки полученных автоматических рефератов. Несомненно, наиболее правдоподобные оценки качества можно проводить в ручном режиме с привлечением большого числа экспертов. Однако такие ручные оценки являются чрезвычайно дорогими.

Методики автоматической оценки качества реферирования не только делают этот процесс более доступным, но и позволяют в реальном времени производить настройку параметров работы определенного алгоритма, производить их оптимизацию.

Авторами была рассмотрена используемая в DUC автоматически вычисляемая метрика *ROUGE*. Адаптировав ее под русский язык и, немного видоизменив, авторы ввели метрику *ROUGE-RUS* и разработали систему для автоматической оценки качества обзорного реферирования на основе этой метрики с Web-интерфейсом.

Исследования метрики *ROUGE-RUS* показали возможность ее применения для оценки качества обзорного реферирования. Для получения стабильных результатов оценки необходимо использовать как минимум 4-5 ручных аннотаций, составленных различными пользователями.

Используя новую метрику *ROUGE-RUS*, была исследована степень влияния выбора базовых параметров и начальной темы на результат работы алгоритма. Алгоритм показал высокую степень устойчивости относительно выбора базовых параметров. На основе результатов исследования влияния выбора темы на работу алгоритма был разработан модифицированный алгоритм. Учет тем всех или нескольких документов позволил повысить качество работы алгоритма.

В результате подбора параметров и модификации алгоритма для использования нескольких тем были получены результаты оценки мо метрике *ROUGE-RUS* не хуже, чем результаты сравнения ручных аннотаций друг с другом. На ряде кластеров были получены результаты, превосходящие базовую величину *ROUGE-RUS* на кластере. Кроме того, удалось улучшить показания метода по сравнению с псевдо-автоматическими аннотациями *«Basic Lines»*.

Таким образом, проведенный эксперимент по составлению ручных аннотаций новостных кластеров различными людьми позволил на основе полученного материала, исследовать как саму метрику *ROUGE-RUS*, так и алгоритм обзорного реферирования *Manifold Ranking*, а также построить

модифицированный алгоритм, показавший значительно лучше результаты по сравнению с базовым.

## Литература


[1] H.T. Dang. Overview of DUC 2006. http://duc.nist.gov/pubs/2006papers/duc2006.pdf National Institute of Standards and Technology (NIST)

[2] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. Information Sciences Institute. University of Southern California 2004

[3] Luhn The Automatic Creation of Literature Abstracts (context) http://citeseer.ist.psu.edu/context/74679/0 1958

[4] MDS Evaluation Framework http://mdsevaluation.ru/

[5] Xiaojun Wan, Jianwu Yang and Jianguo Xiao. Manifold-Ranking Based Topic-Focused Multi-Document Summarization http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-467.pdf. DUC 2003. Institute of Computer Science and Technology Peking University, Beijing 100871, China

[6] Zhou et al., 2003b D. Zhou, J. Weston, A. Gretton, O. Bousquet and B. Schlkopf. Ranking on data manifolds. In Proceedings of NIPS'2003.

[7] Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии.

[8] Лукашевич Н.В., Добров Б.В., Автоматическое аннотирование новостных кластеров на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии - По материалам ежегодной международной конференции «Диалог». Периодическое издание. Выпуск 8 (15), 2009.

[9] С.Д. Тарасов. Автоматическое составление рефератов новостных сюжетов. Труды 10-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2008, Дубна, Россия, 2008.


## The research and parameter's optimization of Manifold Ranking Algorithm based on automatically summarization evaluation metric by ROUGE-RUS

S.D. Tarasov


In this article we review the ROUGE metric adopted by DUC for automatic summarization evaluation and also it modification for Russian language ROUGE-RUS implemented by author.

Developed prototype with WEB interface for automated evaluation of multi-document summarization performance based on ROUGE-RUS

Concluded the experiment with objective to compose the handmade annotations for news clusters and automated evaluation of multi-document summarization.

Defined the base variable ROUGE-RUS on cluster. Reviewed the algorithm of Multi-document summarization "Manifold Ranking".

Researched the impact of diversity penalty (base parameters and initial input) on algorithm final result. The algorithm was modified to compensate above diversity penalty.

# Using Fingerprints in n-Gram Indices

© Stefan Selbach

Lehrstuhl für Informatik II, Universität Würzburg
selbach@informatik.uni-wuerzburg.de

## Abstract

The major advantage of the n-gram inverted index is the possibility to locate any given substring in a document collection. Nevertheless, the n-gram inverted index also has drawbacks: If the collections are getting bigger, this index tends to be very large and the performance drops significantly. We propose a novel technique of enhancing the performance of an n-gram inverted index with the use of additional fingerprints for each n-gram. A fingerprint contains information about the positions of an n-gram. When combining two or more n-grams, these fingerprints also provide information about the positions of the combination. This can be used to reduce the complexity of merging the n-gram postings lists for a given search and improves the performance of the n-gram inverted index. Furthermore it is possible to freely scale the size of the fingerprints in order to adjust the performance of the index. The size of a fingerprint is neither dependent of the size of the document collection nor the number of n-grams.

## 1 Introduction

Text searching is regarded as one of the core subjects in Information Retrieval [1]. Many search engines (e.g. Lucene [2]) are building word based inverted indices for document retrieval. Thus the users of these search engines can only search for words. If only a substring of a term matches the given query, no results are being returned for this hit. Many word based search engines try to solve these problems by adding fuzziness to the index and to the query. While parsing documents every word is reduced to its radical. This is done by using stemmers [3,4] and by limiting the set of allowed characters. This procedure is also being applied to every query. This improves the recall, however it tends to reduce the precision of the search because words with different meaning can accidentally be mapped to the same radical. Even though this technique may perform well with documents containing conventional natural language it is not suited for terminologies. For example, if you think of the chemical substance "1,3-Cyclooctadiene" it should also be possible to find this occurrence using the query "Cyclooctadiene", because the usage of this specific isomer "1,3-Cyclooctadiene" is not common. So this word would be reduced to its radical "Cyclooctadiene". The prefix "1,3-" will be ignored. But since some users might however search for this specific isomer, this information has to be included in the index. That is where the inverted n-gram index [5,6,7] comes into play. It enables us to perform exact string matching, so that both queries mentioned above return the appropriate documents. An n-gram index divides the text of the document into overlapping substrings of the size 1 to n. For each n-gram the positions of all occurrences are stored in the index. The given query is divided into n-grams as well and the position lists of the affected n-grams are intersected to retrieve the positions of the query. Since the number of positions for an n-gram is always increasing while the document collection gets bigger, the performance of this index gets worse, because longer position lists have to be intersected (cp. [8]). First we tried to apply fingerprints to the n-grams, so that without storing any positions of the n-grams in the index decision could be made, whether the query does not hit a certain document, or might hit a document. In the second case verifications had to be made, if these documents match the query. Since this verification can worsen the performance, especially when many documents are being affected, we added additional information of the positions of the n-grams to the index, so that no verification was needed. With this technique we reduced the costs for intersecting the position lists of the n-grams. Another advantage is that we can adjust the performance of the index by changing the size of the fingerprints. This gives us the potential to scale the index according to the size of the document collection.

## 2 Related Work

A term based inverted index [9] consists of two major components: terms and posting lists (see Figure 1). Each term is linked to one posting list. A posting contains a document-identifier (*fileID*) and a position inside the document (*offset*). Besides, to get a fast access to the terms, an index such as a binary tree is

**Proceedings of the 11th All-Russian Research Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» - RCDL'2009, Petrozavodsk, Russia, 2009.**

created over the terms [10,11]. A query is processed by the binary tree and then the posting list is being returned. This setup can only be used, if a query contains terms. But if the query contains substrings of terms, an n-gram index should be used.

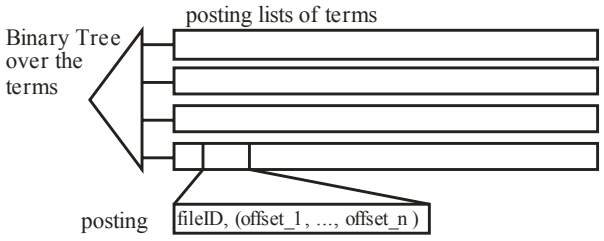



Figure 1: inverted index.

An n-gram index (see Figure 2) does not use terms for indexing. The documents are divided into overlapping substrings of the size 1 to n. For each of these n-grams a posting list containing the positions of all occurrences is stored in the index. In order to get a fast access to the n-grams, a trie or hashmap is used [12].

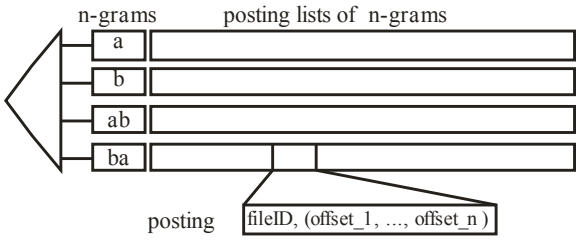



Figure 2: n-gram index.

Since all documents are covered by only including the uno-grams to the index, not all of the n-grams (n>=2) have to be included to the index. Adding more n-grams to the index increases the performance but also increases the index size.
The query is splitted in n-grams in a manner that the combination of these n-grams overlaps the query. After that the posting lists of these n-grams are being intersected. For example the positions for $pos(w \in \Sigma^n)$ the substring $w_1w_2$ can be calculated with the positions of the 1-grams $w_1$ and $w_2$:

$$pos(w_1w_2) = pos(w_1) \cap dec(pos(w_2))$$
$$dec(P) = \{x \mid x+1 \in P\}$$

If the document collection gets bigger, the posting lists get bigger as well. This significantly reduces the performance. An n-gram index grows multiple times faster compared to a word based inverted index. While in a word based index every additional term occurrence creates one new posting, in a n-gram index multiple postings have to be added because all posting lists of affected n-grams have to be updated. Kim et Al. [13] introduced a technique of adding a second m-gram layer to the index. The documents were at first splittet into m-grams and these m-grams were indexed using n-grams.

Choueka et Al. [14] presented a search index that uses bitmaps which act as an "occurrence" map. To each word a bitvector has been assigned. This vector has the same length as the number of documents and each bit identifies an occurrence of a term in a specific document. With this technique the offsets of the terms within the documents are discarded. Furthermore the size of this index increases rapidly for large document collections since the number of terms and the size of the bitvectors are getting bigger. Choueka et Al. [14] as well as Faloutsos [15] also introduced a hybrid index organization so that infrequent terms were treated differently to reduce the size of the index.

Signature files [1][9] are using a similar approach but the size of the bitvector is independent from the number of documents. Signature files are also word-oriented and are based on hashing. A hash function maps a word to a bit mask. The collection is split into blocks and to each block a bit mask is assigned by ORing the signatures of all the words in the text block. A search is carried out by comparing the signature of the query to the bitmasks of all blocks.

In this paper we want to propose a technique using fingerprints to increase the performance of an n-gram index. The size of these fingerprints can be freely selected like in signature files and each bit in the fingerprint gives information about the position of an n-gram like in the bitmap index. Furthermore these fingerprints can by combined with the approach that uses only posting lists in order to achieve a better performance.

## 3 n-Gram Fingerprints

Since the sizes of posting lists of n-grams are enormous when indexing large document collections we introduced a so called fingerprint for each n-gram that contains information about the positions of the n-gram and that can easily be compared to other fingerprints in order to give information about the positions of the combination of these n-grams. A fingerprint for an n-gram $w$ is a two-dimensional bit-matrix $B_w$ of the size $f$ Ч $o$. A bit $b_{i,j}$ is set to 1 if there is an occurrence $p \in pos(w)$ where $fileID(p) \bmod f = i$ and $offset(p) \bmod o = j$.

$$B_w = \begin{pmatrix} b_{0,0} & \cdots & b_{0,o-1} \\ \vdots & \ddots & \vdots \\ b_{f,0} & \cdots & b_{f-1,o-1} \end{pmatrix}$$

$$b_{i,j} = \begin{cases} 1 : \exists p \in pos(w) : fileid(p) \bmod f = i \wedge \\ \qquad\qquad offset(p) \bmod o = j \\ 0 : otherwise \end{cases}$$

Each bit-position is representing a collection of positions that match the above criteria (see Figure 3). Given two 1-grams $w_1$ and $w_2$ and their respective

fingerprints $B_{w1}$ and $B_{w2}$ we can approximate the fingerprint $B_{w1w2}$:

$$B_{w1w2} \approx B'_{w1w2} = B_{w1} \wedge B'_{w2}$$

$B'_{w2}$ is constructed by cyclic shifting each column of $B_{w2}$ by one position to the left. This is done in order to compensate the offset of the second 1-gram in relation to the first n-gram. The matrices $B_{w1}$ and $B'_{w2}$ are combined bitwise with AND ($\wedge$) to get $B'_{w1w2}$. We can guarantee that every bit that is set in $B_{w1w2}$ is also set in $B'_{w1w2}$. However $B'_{w1w2}$ may contain bits that are set in $B'_{w1w2}$ but not in $B_{w1w2}$.

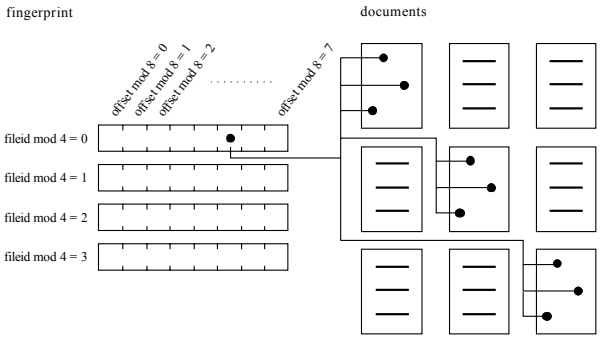



Figure 3: n-gram fingerprint.

To perform a search we have to split the query into n-grams. To get the best results we use as many n-grams as possible. But we can skip n-grams that are included by another n-gram because the bits set in their fingerprints are only supersets of the bits set in the superior n-gram. If we would index for example all 3-grams of a given document collection we would split the query in the same manner, so a query of length *n*, would be split in *n-2* 3-grams. This technique is different to the standard n-gram search where a query is splitted into as few n-grams as possible. After that, the fingerprints of these n-grams are being loaded from the index and the columns of these fingerprints are cyclic-shifted to the left by the relative position of the respective n-gram to the first n-gram of the query. These fingerprints are combined with AND. The resulting fingerprint represents a collection of positions where the query might occur. To verify the results, these positions have to be checked against the query. This is done by opening the involved documents and comparing the local context of the positions to the query. We tested this setup on the "Online Encyclopaedia of Dermatology from P. Altmeyer". This collection contains over 7500 documents. Furthermore we had the access to the querylog and could test the performance of our new setup with real user data. We created an index with $f$ = 1024 and $o$ = 128. Table 1 shows some of the resulting search speeds. The time for the combination of the bit-matrixes is only dependent of the length of the query. But the time for the verification of the positions depends on the number of bits set in the fingerprint of the search result. This can be problematic if a user is searching for a substring, which occurs multiple times in every document. The resulting bit-matrix of this query is very dense and thus many positions have to be verified, which leads to a poor performance.

Table 1: Search speed.

| Query | Bit-matrix | Time for verification | Hits |
|---|---|---|---|
| rhinolo | 219 ms | 94 ms | 18 |
| sanfilipo | 290 ms | 0 ms | 0 |
| itracon | 266 ms | 336 ms | 64 |
| oxyuria | 197 ms | 48 ms | 6 |

Since the size of the fingerprints can be adjusted by changing the parameters $o$ or $f$ and the number of n-grams can be limited (see section 2) the size of this index is independent from the number of documents in the collection. However it is recommend to increase the size of the fingerprints when indexing large collections. Large collections are leading to dense fingerprints which result in a low filtration ability.

## 4 n-Gram Fingerprints in Combination with Posting Lists

Analyses of user queries (see Figure 4) showed that most users do not search for very frequent terms. 10% of the queries hit 400 different terms with frequencies between 150 and 2000. The remaining 90% of the queries were almost equally distributed over the remaining 170.000 terms. However the most of the time of a search was needed for the verifications of hits.

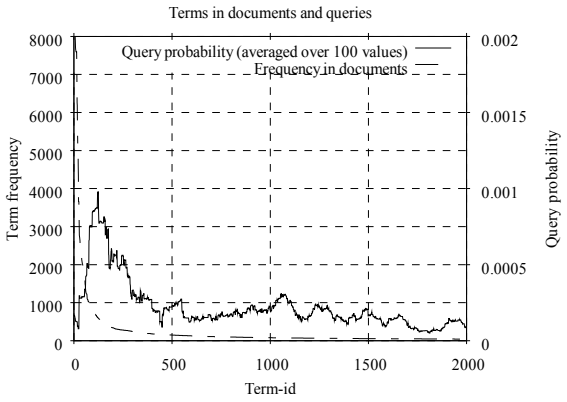



Figure 4: Term frequency in documents and queries.

In the next step we combined the standard n-gram search with our fingerprinting technique. When creating a fingerprint for an n-gram we assign a bit to every posting of an occurrence of the n-gram. Thus we also partitioned the postings into smaller subsets. If we do not discard these lists, we can use them for the verification process. For each n-gram the partitioned posting lists were added to the index (see Figure 5).

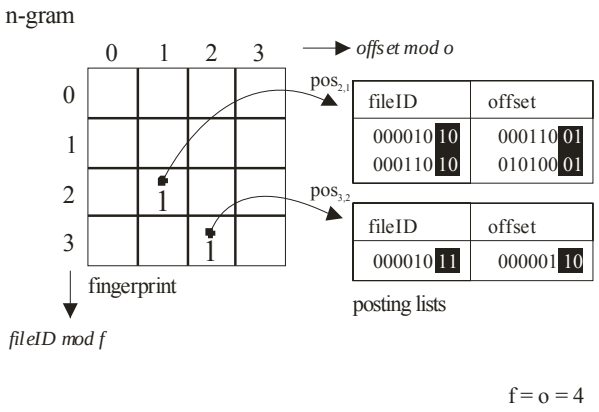



Figure 5: Combination of fingerprints and posting lists.

If $f$ and $o$ are multiples of 2 we do not have to include the last $log_2(f)$ bits of the *fileID* and the last $log_2(o)$ bits of the *offset* to the posting, because this information is already included by the residue class (see Figure 5).

The bitwise AND combination of the fingerprints of the affected n-grams for a query shows which subsets of posting lists have to be intersected. On the one hand we reduce the cost for the intersection of the postings since only a few of the smaller subsets have to be intersected instead of a complete posting list for an n-gram. On the other hand we have to deal with much more lists. Our test corpus has been indexed with 1024 residue classes for the *fileID* and 128 for the *offset*. As a result we get 131.072 subsets of posting lists for each n-gram. We included 14.000 different (1-3)-grams in our index, which made 18.350.080.000 different lists in total. Most of these lists are empty or have only a few entries. In order to reduce the overhead, which is necessary for handling this large number of posting lists, we used a file based hash table (see Figure 6). The hash value for a given list was computed as a function of the residue classes of *fileID* and *offset*:

$$h(pos_{i,j}) = j + i \cdot o$$

This way the posting lists of n-grams occuring nearby are also stored close to each other in the index. In this case $i$ (the residue class for the *fileID*) is constant and $j$ (the residue class for the *offset*) only differs by the distance (*mod o*) of the n-grams. Thus the hash value differs also only by the distance.

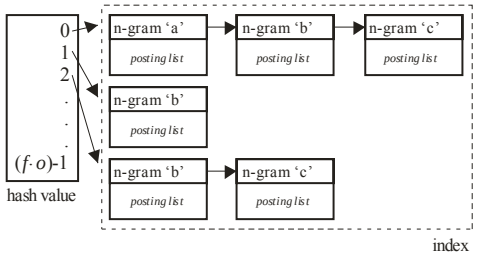



Figure 6: Managing the n-gram posting lists.

The main disadvantage of this technique is that the set of possible hash values in the hash table is rather small and the hashing function produces many collisions. In our index for the "Online Encyclopaedia of Dermatology from P. Altmeyer" we had 25 collisions in average. That means in average 12.5 lists had to be processed until the desired list was found.

If we take a look at the frequency of n-grams in the documents and in the queries of our collection we see, that many n-grams are rarely used in the queries (see Figure 7). We can use this information to define a ranking by which the posting lists sharing the same hash value can be sorted. This way we should get a faster access to the hash table for ordinary queries. In order to measure the performance of this methodology we used 5000 user queries to define the ranking of the n-grams. We then used 5000 different user queries to test the performance of the hash. We came to the result that in average only 5 lists had to be processed

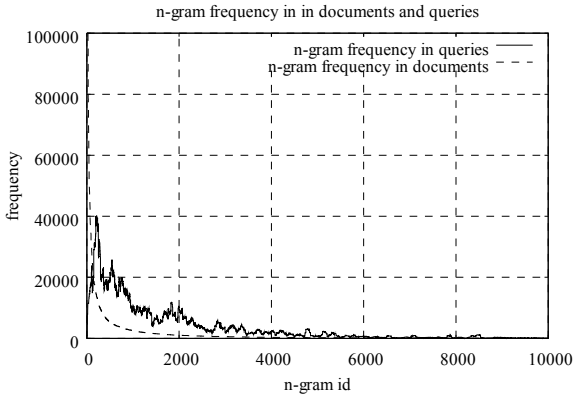



Figure 7: n-gram frequency in documents and queries.

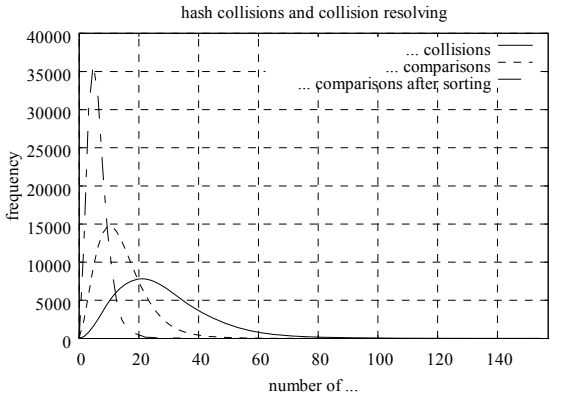



Figure 8: Hash collisions and collision resolution.

in order to retrieve the desired list (see Figure 8). This improved the performance of the hash by a factor of 2.5. This is a great result since we would need 4.6 comparisons in average to retrieve the requested list, if we would have random access to the different lists in

one hash entry (which requires more space for the index) and could perform a binary search. This is very close to our result, but we do not need the extra index space for the random access to the lists. We compared this setup with the search engine from section 3. In average the performance improved by 40%. Table 2 shows that the time needed for the verification reduced significantly. Besides it was more balanced.

Table 2: Search speed (fingerprints with position lists).

| Query | Bit-matrix | Time for verification | Hits |
|---|---|---|---|
| rhinolo | 230 ms | 10 ms | 18 |
| sanfilipo | 271 ms | 0 ms | 0 |
| itracon | 245 ms | 15 ms | 64 |
| oxyuria | 210 ms | 12 ms | 6 |

## 5 Fingerprint Compression

In order to compress the fingerprints we analysed their density. Over 70% of the fingerprints have a density less than 0.1, 20% are between 0.1 and 0.4 and 10% have a density greater than 0.4. Fingerprints, which have an extremely low or high density, do not contain much information. These fingerprints can be compressed by simply reducing their resolution. This equals to a convolution of the original fingerprint. Figure 9 shows an example for a 4 x 4 fingerprint. This fuzziness might in some cases lead to an incorrect fingerprint for a query. But since all bit-positions of the resulting fingerprint are being verified using the position lists of each n-gram no false positive hits are generated. It just slightly slows down the performance, because additional requests for position lists, that do not exist, have to be made.

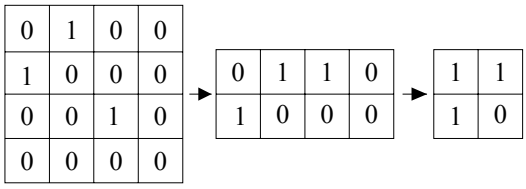



Figure 9: Convolution of fingerprints.

We tried different thresholds of densities for the convolution. If the density of a fingerprint was in the defined range, we reduced its size by factor 2. Table 3 shows the relative performance loss and the relative index reduction.

Furthermore we implemented a dictionary based compression. Each column of a fingerprint was added to a dictionary. The fingerprint now consists of multiple references to this dictionary. This methodology significantly reduced the size of the index. Especially dense and sparse bitvectors had been reused frequently.

Table 3: Fingerprint compression.

| Density threshold for convolution | Performance loss | Fingerprint index reduction |
|---|---|---|
| no convolution | 0 % | 0 % |
| 0-0,025 and 0.975-1 | 3.1 % | 23 % |
| 0-0.05 and 0.95-1 | 3.2 % | 27 % |
| 0-0.1 and 0.9-1 | 10 % | 29 % |
| 0-0.2 and 0.8-1 | 25 % | 31 % |

In combination with the convolution of fingerprints we were able to reduce the space needed to index the fingerprints by 50-60%. The dictionary based compression reduced the performance only by 1-3%.

## 6 Conclusion and Future Work

Our experiments have shown that n-gram fingerprints can be used to improve the performance and the scalability of n-gram inverted indices. We introduced techniques to optimize the index in a manner so that common user queries can be processed more efficiently. In standard n-gram indices a query is splitted into n-grams that overlap the query. The posting lists of all these n-grams have to be intersected to get the final search result. With our technique we split these posting lists into smaller lists. After combining the fingerprints of the involved n-grams we know which of these lists have to be intersected. Thereby we reduce the complexity of the intersection and gain more performance. We compressed the fingerprints that do not contain a certain level of information and we were able to reduce the fingerprint index size by 60% without major loss of performance.

In the future we would like to combine our indexing techniques with a word based inverted index in order to profit from the advantages of using both n-grams and words as indexing terms. We intend to split the text into terms and save the posting lists in an index like in the case of an inverted index. However we do not create a binary tree over the terms. The terms will be indexed in a second n-gram inverted index described in this work. This way we do not lose the ability of searching for any kind of substring in the document collection. Another advantage is that stop words, which tend to flood the posting lists of some n-grams, only occur once in the term list of the first inverted index and thus fewer postings for the affected n-grams have to be dealt with in the second index. Using a word based inverted index gives us further benefits. In n-gram inverted indices ranking the results of a search is a difficult task. Since n-grams generally do not hold semantic information, only TF/IDF or field based ranking methods can be used. Having words as index terms it is possible to define a precomputed rank for each tupel (*term*, *document*). Moreover index terms can be linked to each other in order to map thesaurus information to the index.

## References


[1] Baeza-Yates R. and Ribeiro-Neto B.: Modern Information Retrieval. ACM Press (1999)

[2] The Lucene search engine, http://jakarta.apache.org/lucene/ (2005)

[3] Porter M. F.: An algorithm for suffix stripping. Readings in Information Retrieval **14**(3) (1980) 130–137.

[4] Snowball stemmers, http://snowball.tartarus.org/ (2003)

[5] Yasushi O. and Toru M.: Optimizing query evaluation in n-gram indexing. In: Proceedings of International Conference on Information Retrieval, ACM SIGIR, Melbourne, Australia (1998) 367–368

[6] Brown M. K., Kellner A., and Ragget D.: Stochastic Language Models (N-Gram) Specification. W3C Working Draft (2001)

[7] Miller E., Shen D., Liu J., and Nicholas C.: Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System. In Journal of Digital Information **1**(5) (2000)

[8] Mayfield J. and McNamee P.: Single N-gram Stemming. In: Proceedings of International Conference on Information Retrieval, ACM SIGIR, Toronto, Canada (2003) 415–416

[9] Witten I. H., Moffat A., and Bell T. C.: Managing Gigabytes, second edition. Morgan Kaufmann. (1999)

[10] Zobel J., Moffat, A.: Inverted files for text search engines. ACM Computing Surveys **38**(2) (2006)

[11] Zobel J., Moffat A., and Ramamohanarao K.: Inverted Files versus Signature Files for Text Indexing. ACM Trans. on Database Systems **23**(4) (1998) 453–490

[12] Cohen J. D.: Recursive Hashing Functions for n-Grams. ACM Trans. on Information Systems. **15**(3) (1997) 291–320

[13] Kim M.-S., Whang K.-Y., Lee J.-G., and Lee M.-J.: n-Gram/2L: A space and time efficient two-level n-gram inverted index structure. In: Proceedings of VLDB. (2005) 325–336

[14] Choueka Y., Fraenkel A., Klein S., and Segal E.: Improved techniques for processing queries in full-text systems. In: Proceedings of the 10th ACM SIGIR Conference. ACM Press. (1987) 306–315

[15] Faloutsos C., and Jagadish H. V.: Hybrid index organizations for text databases. In: Proceedings of the International Conference on Extending Database Technology. LNCS, vol. 580, Springer. (1992) 310–327

# ПРИГЛАШЕННЫЙ ДОКЛАД

# INVITED PAPER

# IT Research Challenges in Digital Preservation

© Andreas Rauber

Vienna University of Technology, Vienna, Austria
rauber@ifs.tuwien.ac.at

## Abstract

Digital Preservation (DP) has evolved into a specialized, interdisciplinary research discipline of its own, seeing significant increases in terms of research capacity, results, but also challenges. With this specialization, however, core IT know-how that is needed to tackle the significant problems that we are facing in DP is not sufficiently present within the DP research community. This paper outlines some research challenges in DP, highlighting the need for the DP community to reach out to IT research in general to jointly develop solutions. It also shows some examples of integrating other computer science disciplines such as Information Retrieval / Machine Learning, or Software Engineering, to address DP challenges, concluding with a brief overview of activities at the Department of Software Technology and Interactive Systems at the Vienna University of Technology in this domain.

## 1 Introduction

Digital representation of information - once perceived as the solution to all problems of its analogue counter parts in terms of stability, replication, and thus long-term availability – has turned out to be more fragile and susceptible to total loss than expected. While digital objects can be replicated without any degradation in quality, it is the encoding and the dependency of the digital representation to be interpreted / rendered that is endangering accessibility. Any digital object, be it a document, a data file, or an application, requires some specific software application, such as an editor capable of interpreting ASCII/Unicode encodings, an office suite, or a database system etc., to be opened and rendered. These, in turn, rely on specific libraries, and a specific operating system, which, in turn, relies on a specific hardware environment to run. If any of these modules in the so-called view-path of a digital object is lost or defunct, the whole digital object is usually reduced to a meaningless bit-stream. Given the speed of evolution of file formats, versions of software applications, operating systems as well as hardware components, including the current drive for higher complexity at each of these levels via distributed objects, cloud computing and mesh-ups, digital objects are facing serious threats of becoming useless bit-streams within very short periods of time. At the same time an increasing amount of essential information is being produced and stored only in digital form, putting society at whole at risk.

To mitigate these risks, digital preservation has emerged as an active research discipline, combining expertise from a range of backgrounds including experts from cultural heritage and memory institutions, legal experts, scientists and engineers from a range of disciplines working intensively with scientific data, and – last, but not least – computer science and IT experts. In the last few years, numerous approaches have been analyzed, standards have been devised, and systems are being deployed to tackle the challenges. In this short time period, Digital Preservation (DP) has evolved into a research discipline in its own right, with experts collaborating intensively in an interdisciplinary manner, successfully driving both our understanding of the problem as well as the availability of solutions. However, due to the highly interdisciplinary nature of the challenges, as well as due to its evolution into an independent research field of its own, DP research runs the risk of excluding essential expertise and input from more traditional sub-disciplines in each of the various disciplines involved. This may be due both to the natural segregation happening with any sub-group forming, as well as due to the complexity inherent in interdisciplinary research, rendering both language as well as communication forms difficult to understand for the non-initiated.

This paper tries to shed some light on the complexities and challenges in DP research that require specific involvement from a range of core computer science disciplines. The list is by no means exhaustive, nor is its focus on computer science aspects meant to downplay the other disciplines involved. It shall merely act as a call to experts in the respective disciplines to consider devoting effort to these non-trivial challenges that are within their core expertise to help advancing the field and pushing it further in order to solve a problem that may well turn out to be the greatest disaster for an information society relying on digital information and processes if leaft unsolved.

**Proceedings of the 11th All-Russian Research Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» - RCDL'2009, Petrozavodsk, Russia, 2009.**

The remainder of this paper is organized as follows: The next section provides an overview of research agendas in the field of digital preservation. It also lists some of the EU-funded research projects launched recently in this domain. Section 3 collects a number of challenges in various sub-disciplines in IT, trying to motivate the need for research and potential directions. Section 4, finally, summarizes a number of efforts currently worked upon in our group. Section 5 summarizes the paper and points to some recent initiatives in terms of focussed DP education.

## 2 Research Agendas in Digital Preservation

Due to the pressing importance, a number of research agendas for Digital Preservation have been compiled in the last few years, and a series of research projects have been launched in Europe and all over the world to tackle these issues.

One of the most recent research agendas is the DPE Research Roadmap (DPE Project Consortium, 2007).

It is based on the analysis of a number of earlier research roadmaps in this domain, and – while acknowledging advances in certain domains such as specifically the creation of conceptual models and a common understanding of the problem domain – emphasizes the needs for interoperability and further standardization in numerous areas. Specifically, it recommends research in the domains of restoration and conservation; risk analysis and mitigation; understanding and handling of significant properties of digital objects and their context; interoperability amongst systems and automation of workflows, up to challenges in the general management of preservation activities, systems and organizations. Storage systems still pose numerous challenges. Last, but not least, a strong emphasis is placed on the need for controlled experimentation and evaluation to obtain a scientifically valid basis for decisions. It also provides an extensive review of earlier projects and national as well as international activities at that time.

Based on the need for further research and development, a range of activities was started, specifically within the 6th and 7th Framework Programmes of the European Commission. Two large initiatives started already within the 6th Framework Programme are the integrated projects Planets and Casper. Planets[1] (Preservation and Long-term Access through Networked Services) is a four-year project. It aims at building practical services and tools to help ensure long-term access to our digital cultural and scientific assets. It consists of a range of services embedded within an interoperability framework, comprising Preservation Actions, Preservation Characterization, a solid Preservation Planning workflow (H. Kulovits and A. Rauber, 2008), as well as a Testbed for service evaluation. CASPAR[2] (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval**)** is an Integrated Project aiming at implementing, extending, and validating the OAIS reference model (ISO14721:2003). It intends to design virtualisation services supporting long term digital resource preservation, despite changes in the underlying computing (hardware and software) and storage systems, and the designated communities, as well as to integrate digital rights management, authentication, and accreditation as standard features.

Under the 7th Framework Programme, a number of new initiatives have been launched[3], including specifically Keep[4], a project focussing on emulation as the core preservation strategy; PrestoPrime[5], focussing on the preservation of audio-visual media in the broadcasting domain; LiWA[6], tackling the challenges of long-term preservation of data in web archives, with a specific focus on the evolution of semantics across time; PROTAGE[7], which is relying on agent-based environments for the preservation of digital objects; as well as SHAMAN[8], which will implement large-scale European-wide collections with access services that support communities of practice in the creation, interpretation and use of cultural and scientific content, including multi-format and multi-source digital objects.

The progress that these projects make, as well as the increasingly challenging scenarios that they address, are good indicators of a maturing research discipline. While it is definitely a good direction for the community to establish itself, it also brings along some of the downsides of specialized sub-disciplines. Amongst them are increasingly closed circles of highly specialized experts working in the field, using specialized terminology, and building upon a large body of pre-existing knowledge that makes it hard for other disciplines to contribute in spite of the fact that specialized contributions from other sub-disciplines would be highly beneficial. One particular challenge, inherent in the interdisciplinarity of the field, stems from the fact that the most active communities for a long time came from the cultural heritage and archival sciences domains. As a consequence, we have found it harder to attract the interest of IT experts in different fields to address the challenges in digital preservation, and to understand the impact that their respective areas of expertise can have on this community.

One notable initiative to bridge this gap is the DPE Digital Preservation Challenge[9]. This competition series consists of a number of digital preservation tasks addressed specifically at computer and information science students. Examples include the identification and recovery of information from a binary data stream, such as e.g. recovering the sound played during the opening sequence of an old computer game; retrieving the address of a specific person from an obsolete database; or developing a solution for preserving early works of digital art. Faced with challenges like these, students immediately realized the difficulty and complexity of the problems – as opposed to the rather laissez-faire approach of suggesting that simple software updates and import/export functions could overcome all problems of IT obsolescence. It also made DP challenges more accessible and exciting to address, phrasing them as actual IT challenges, mitigating part of the community jargon problems.

The following section tries to highlight – without any claims of completeness – some of the challenges in DP and their relationship to core computer science disciplines, inviting further comments and contributions.

## 3 Digital Preservation and IT Research

Digital Preservation, by its very nature, is tightly bound to expertise in information technology and computer science. Nevertheless, formulating requirements and challenges using terminology such as a*uthenticity, archival workflows, ingest/deposit regulations, appraisal, significant properties,* and others does not make these obvious areas for computer science research. As a result, relatively few computer science experts are actively involved in research activities in this domain.

Amongst the more obvious IT topics within DP is storage. Preservation systems need to store massive amounts of data for long periods of time with high levels of data security. While these challenges are largely in-line with the requirements of traditional IT systems, DP adds some specific challenges. One of these is the need for a high stability of the storage medium in off-line conditions, i.e. without the need to be continuously operated or monitored. Also, self-checking and verification are essential. However, even more than the stability of the storage medium, the stability, simplicity and long-term availability of reading devices for the storage media is essential. With some tape storage devices, the durability of the media exceeds the availability of the technology necessary to read/write the tapes, rendering such media unreadable not due to media decay, but to HW obsolescence. While preservation actions such as data migration (conversion to new formats), which require periodic copying of data, mitigate this risk by allowing media migration at the same time, many institutions adopt a policy of keeping a copy of the original object ingested into a preservation system permanently and unchanged, calling for specific storage technology. Hybrid storage of both digital as well as analogue representations of objects, self-maintaining hierarchical storage systems, description of storage technology, large-volume transfer of several petabytes of information, error- and consistency checking across such large amounts of data, etc. all need to be addressed from the point of view of long-term preservation systems in times where replacement cycles of technology get shorter and shorter.

One of the core concepts of DP lies in the requirement of being able to provide authentic copies of digital objects in the future. Authenticity, in a nutshell, refers to the characteristic of an object to represent what it pertains to be, i.e. to show all characteristics, both technically as well as intellectually, that allow is usage. While ultimately this is all about trust and documentation of any changes that may have happened to an object, it also requires a consistent audit trail and software solutions to ensure that an object is not altered in an unauthorized and undocumented way, be it on purpose or accidentally. This poses high challenges in terms of IT security on preservation systems, requiring documentation and enforcement of access rules across long periods of time. Key management and the security of various generations of encryption technology need to be managed across several system replacement cycles. Revocation timestamps of keys need to be documented, and a complex management of users and roles needs to be in place and consistently monitored. Ensuring that an object remains encrypted for time periods of 70 years while allowing decryption – transformation – re-encryption to happen within a sealed-off black-box poses non-trivial challenges. Trustworthiness of such systems will depend on audit routines that require solid means for automatic verification of code, workflows, and designs.

This is further complicated by the fact that any preservation system built and audited against these requirements (which already constitutes a complex task) needs to undergo a series of revisions and re-implementations throughout its "life time" of 100+ years. The levels of documentation needed to understand and validate performance without compromising security are hardly analyzed in current DP research. In fact, the very need for specialized DP systems partially comes from the fact that operational IT systems lack the long-term perspective of other areas of systems engineering. Other disciplines, such as architecture, plan the whole life cycle of a system (bridge, industrial plant, etc.), including periodic maintenance routines up to its destruction and replacement already during the design phase. Similar maintenance and replacement cycles are planned as part of the design of cars, planes, infrastructure networks, and other complex systems. IT infrastructure and software systems, on the other hand, are often planned on a rather hic-et-nunc basis, with software maintainability often being reduced to a marginal conceptual requirement, and software system succession usually not considered at all. If these requirements were considered as an integral part of system modelling and software engineering processes, would then "OAIS compliance" be an integral characteristic of all IT systems? While obviously desirable, principles for such long-term levels of software life cycle management, costing, modelling and development principles still require intense research efforts.

The long-term aspect itself poses significant challenges, as the semantics of terminology used to describe system functions, actors, or requirements is likely to change across such large time horizons. Knowledge assumed to be generally available today may be obscure in the not-too-distant future, rendering descriptions of system architectures, HW components and algorithms incomprehensible. This is especially important considering the fast innovation cycles and the resulting fashion-like adoption of new, short-lived terminology.

While a lot of current efforts in DP focus on document type objects, some of the most valuable information is stored in databases of considerable complexity. While a clean separation of data and functionality has always been a core requirement in data base design, many real-life systems do not fully comply with this desired ideal. Databases not only host information, they act on it. Semantics is hidden in distributed tables; triggers and active code add further complexity. Design and documentation guidelines for databases that are preservation-aware would ease subsequent preservation of such systems and their valuable contents.

But even the rather simple concept of "digital objects" is likely to pose drastically new challenges in the near future. Most approaches for document-type objects such as images, office documents, spreadsheets, videos, etc. are based on the concept of (complex) files following specific file format definitions. However, recent development saw more and more file types evolve into generic containers that can include virtually any other file type container, hardly without limitations. This may ultimately lead to the death of the concept of a file format as a well-defined entity which restricts the technical characteristics of the bit-stream that is embedded. It poses drastic new challenges to the tools operating on such files, including a recursive decomposition of objects if they are supposed to deal with all objects of a certain type. Additionally, mash-up documents integrating distributed on-line sources, as well as increasingly active elements within files turn them into both programs as well as a network of dynamic content, rather than static objects that can be stored and handled. This will require drastically new approaches to both migration and emulation as the two dominant preservation strategies of current times.

Understanding what a digital object consists of, both technically as well as semantically, requires sophisticated concepts from data mining as well as information retrieval. While the challenges of analyzing and describing the structure of large amounts of files are only understood to a limited degree (C. Becker et al., 2008), those of semantically searching and mining the data are more obvious. As a rather obvious example we can see the challenges of providing full-text search within a Web Archive as the equivalent of operating a current state-of-the-art search engine on an index that does not only span the current Web, but also its entire history, requiring multiple server farms for storage and index handling (even though, very likely, for lower query traffic.) This will be complemented by new search technology, specifically in the multimedia domain, to enhance traditional text search. It will go beyond simple retrieval of objects, but addresses whole networks of objects and their semantics that may well be different from the semantics of the individual objects and at different periods of time.

This leads to another interesting aspect of digital preservation, where IT research closes the loop again to social sciences and ethics. While heritage institutions and research centres have always been collecting and maintaining data over long periods of time, these data usually were complicated to search through. Utilizing these assets was a complex research endeavour in its own right. With new search technology and assuming the availability of good long-term preservation solutions, masses of historic data can be analyzed efficiently, giving rise to concerns about the value of the concept of forgetting within a society. These issues are being addressed in the fields of data protection laws for operative databases. The impact of long-term availability and aggregation of data, for example within Web Archives, and its relation to advances in search technology is still far from being fully understood (A. Rauber, M. Kaiser, and B. Wachter, 2008).

The list of topics may be continued for quite some times. Further significant challenges that are essential for mastering the long-term preservation of digital objects include automatically identifying semantics from code, cross-compilation, abstraction from hardware layers, chip design and documentation, and others.
Many of these challenges are already subject of intensive research within the respective disciplines, independent of the DP research agenda. Connecting them and adopting them as standard best-practice principles may help moving computer science and information technology to the same level of maturity that other, older disciplines and specific sub-disciplines within IT may have reached already.

## 4 Selected DP Research Activities

This section briefly reviews some of the activities in digital preservation in our group. We will start by analyzing preservation planning as a specific type of Commercial-off-the-Shelf (COTS) component selection in Section 4.1. This is followed by a description of automating preservation services using rule-based decisions to provide preservation systems to small office/home office (SOHO) settings as part of the HOPPLA system in Section 4.2. While emulation has always been one of the dominant preservation actions, solidly evaluating and benchmarking emulators proves more challenging than expected. In Section 4.3 we will briefly review some of these challenges. In Section 4.4 we will switch to understanding the context of creation and usage of digital objects, relying on information retrieval and data warehousing principles to automatically establish context of usage. In Section 4.5 we will take a glimpse at using microfilm as a viable storage technique for binary data, before concluding in Section 4.6 by raising some ethical issues involved in archiving and analyzing data over long periods of time, specifically in the domain of Web Archiving.

### 4.1 Preservation Planning as COTS Selection

A range of different strategies, i.e. preservation actions, have been proposed to tackle the digital preservation

challenge. However, which strategy to choose, and subsequently which tools to select to implement it using which system configuration and which parameter settings, is a crucial decision. It must be based on a well-documented and profound analysis of the requirements and performance of the tools taken into consideration. The Planets Preservation Planning approach (Strodl et al., 2007) allows the assessment of all kinds of preservation actions against individual requirements and the selection of the most suitable solution. It enforces the explicit definition of preservation requirements and supports the appropriate documentation and evaluation by assisting in the process of running preservation experiments. It is based on work performed in the DELOS Digital Preservation cluster, first introduced in (C. Rauch, and A. Rauber, 2004). While the workflow developed for evaluating DP solutions may seem strongly DP-centric, it is, in fact, a highly repetitive scenario similar to commercial off-the-shelf component selection. Thanks to strictly formulated requirements and the highly standardized behavior of the components to be evaluated (basically, a pair of input and output objects, and system performance being evaluated against a large set of criteria, referred to as "objectives" in the preservation planning process) migration tools and emulators can be evaluated following procedures similar to those employed in COTS selection scenarios. Vice-versa, COTS selection may generalize the evaluation principles tested in DP settings to be utilized for generic COTS selection settings. (C. Becker and A. Rauber, 2009) Current work in this direction concentrates specifically on automating the evaluation of preservation actions by providing a sophisticated measurement framework to create quality-aware web services. (C. Becker et al., 2009)

### 4.2 HOPPLA: Automating Digital Preservation

Digital information is of crucial value to a range of institutions, from memory institutions of all sizes, via industry and SME down to private home computers containing office documents, valuable memories, and family photographs. While professional memory institutions have dedicated expertise and resources available to care for their digital assets, SMEs and private users lack both the expertise as well as the means to perform digital preservation activities to keep their assets available and usable for the future. The Hoppla (Home Office Painless Persistent Archiving) system provides digital preservation solutions specifically for small institutions and small home/office settings. (S. Strodl et al., 2008) It hides the technical complexity of digital preservation challenges by providing automated services based on established best practice examples. Appropriate preservation strategies and required tools for performing them are delivered via a web service, effectively outsourcing the required digital preservation expertise.

### 4.3 Evaluating Emulators

Within the preservation planning process, a range of different preservation actions, such as specific migration tools in different configurations and with specific parameter settings are evaluated, in how far they meet the requirements set out for the specific object collection at hand. While this is a difficult task for migration tools, it becomes even more complex in emulation settings. The reason is that, usually, for migration approaches "only" the final rendering of an object needs to be evaluated, with rendering including all characteristics of an object be it visible display, but also acoustic as well as structural aspects. In emulation settings, however, an important focus lies on the interactive aspects of a digital object. Behaviour of an object needs to be evaluated not only in a static sense, but it may also include a specific timeline. Furthermore, the results may depend on external factors such as random elements popular in computer games, or networked behaviour of objects. In order to understand, whether a specific behaviour of an object is correctly preserved or whether artefacts are introduced by the environment provided by the emulator is non-trivial. These aspects are further complicated by the fact that it is not obvious, at which level to compare the result of rendering a digital object in an emulation setting. Different options available include the representation on a set of output devices such as loudspeakers and the screen, the representation on internal memory such as the graphics card, or the internal process status in main memory, accounting for different artefacts introduced by components that may be beyond the control of an emulation system, such as the screen resolution and colour representation scheme available at a specific hardware platform. While ad-hoc evaluation may be sufficient in specific settings (M. Guttenbrunner et al, 2008), more systematic approaches are required, leading to specific design guidelines for emulators in digital preservation settings (M. Guttenbrunner, 2009)

### 4.4 Establishing Context of Digital Objects

The context of objects is essential for the interpretation of information entities, for establishing their authenticity as well as ensuring appropriate use. Thus, documenting the context of creation and use is an essential task in digital library and document management settings, for retrieval tasks as well as for digital preservation. Yet, context is notoriously difficult and labour-some to establish and document, and often missing or partially incomplete or incorrect when it has to be entered manually by the creator of the objects.

To address this challenge we are researching methods to (semi-)automatically determine the creation and usage context of digital objects (R. Mayer and A. Rauber, 2009). Various aspects of context in different dimensions are automatically detected, and different views at multiple levels of granularity allow the extraction of the most appropriate connections to other digital objects.

Context exists in several forms, ranging from a very low-level technical context in which the object was created, via its immediate context of use (people involved, the project or activity it is related to, etc.), to a wider sociological, legal or cultural context. All levels

of context are of importance for the authentic interpretation and usage of a digital object. However, we focus predominantly on the narrower focus of context that can be determined (semi-) automatically. We thus consider the detection and documentation of context of digital objects as a semi-automatic process along several partially orthogonal dimensions, each of which structures objects according to different aspects. We currently use the following dimensions in our first prototype:

- the time of object creation and modification
- the object type
- the people involved
- the content across different sub-categories, such as (a) topic (b) genre, and (c) acronyms, for example in project names.

The concept of using various dimensions as orthogonal views on the data is inspired by the concept of data warehouses and the data analysis method used therein, on-line analytical processing (OLAP) (R. Kimball and M. Ross, 2002). A central concept is the OLAP cube, which prepares the data for fast multi-dimensional queries and analysis. The analyst can pivot the data in various ways, e.g. see all the sales for a specific city for a certain product, and do this at various levels of aggregation, allowing easily to obtain a more detailed view on demand (`drill down') or a more abstract, summarised view (`roll up').
Establishing context along these and other dimensions in combination with appropriate tools for visualizing, grouping and exporting context information supports a range of different application scenarios, such as object ingest in digital repositories, disaster recovery (R. Mayer, R. Neumayer, and A. Rauber, 2009), or user support in information retrieval tasks within archival holdings.

### 4.5 Storage on Microfilm

Digital data is prone to decay on several levels, one being the storage of data (bit-level preservation), ensuring that the data is securely stored on data carriers that can be read with current technology. A second layer is the logical preservation: digital objects require specific software to be opened and read, which in turn require specific operating systems, device drivers, and, ultimately, hardware to run. Finally, semantic preservation is essential to facilitate correct interpretation of objects, similar to conventional analogue pieces of information.
Several solution are being implemented, usually relying on regularly migrating data both from old storage technology to current one, as well as format migration to current versions of file formats.
However, specifically the latter usually incurs changes to the objects, some of which may seem undesirable with respect to their future usage, especially since these changes accumulate over a series of migration steps.
While careful planning procedures try to limit this effect (Strodl et al., 2007), it cannot be avoided completely. Thus, most initiatives recommend to always maintain the original format version of an object to allow reconstruction and access e.g. via emulation if required. This, in turn, calls for a cost-effective strategy for bit-level preservation of digital data on a durable storage technology not requiring regular maintenance.
Additionally, most settings require a stable back-up copy to be maintained for all data (including migrated versions) in addition to on-line versions for continuous use.
Unfortunately, most digital storage techniques do not offer themselves for these purposes: hard disk RAID arrays require regular operation and need to be replaced every few years. Tape drives as long-term storage of massive amounts of data require regular re-winding of tapes to maintain them readable.
Further, with current development cycles the durability of the tapes has surpassed the support life-time for tape readers, rendering the respective tapes unreadable unless migrated to new types of tapes.

This has led to the revival of a rather unexpected storage technique for digital data, namely microfilming (C. Voges, V. Мдrgner, and T. Fingerscheidt, 2008). Microfilm, especially black/white film, has proven a very durable media, requiring no maintenance apart from appropriate storage conditions. Microfilm as storage media has a life span of more then 100+ years and has the advantage that a media migration has to be done less frequently. It is already used for long term storage of scanned images of paper documents.
Provided correct encoding schemas are used, microfilm can store both analogue representations (i.e. images) of objects as well as the digital data stream, offering the additional benefit of easy inspection and redundancy of representation forms, as it basically already includes a kind of "migrated" analogue representation in addition to the digital object. We are thus currently investigating different encoding schemas such as UUencode or XXE as well as 2-d barcodes as means to “print” digital data to microfilm, and to recover it by using scanning and OCR technology with subsequent decoding into binary data.

### 4.6 Ethical Issues in Web Archive Creation & Usage

A completely different type of challenge is posed by the accumulation of information across long periods of time in combination with advances in search technology, namely questions concerning ethically correct creation as well as usage and provision of archived data (A. Rauber, M. Kaiser, and B. Wachter, 2008). This is particularly prominent in the domain of Web Archiving. While Web Archiving initiatives rescue a massive amount of information on the Web from being permanently lost, the massive collection of Web data poses not only fascinating possibilities for accessing a wealth of information, as well as an invaluable resource for scientist wanting to understand the technological and sociological development of the Web and society at large. It also constitutes a new type of information on its own, posing numerous ethical challenges, specifically given the powerful techniques for analyzing and exploring the masses of accumulated information that

we will have available in the near future.
Being aware of this issue, most Web Archives currently strictly limit access to their holdings, or provide means to allow people having their content excluded from holdings to avoid the subsequent challenges, at the same time drastically limiting their value and usefulness. While the ultimate solution to the problem of what kind of access will be permissible will have to be a legal one, it is important to understand the detailed characteristics of the possibilities of "unethical" exploitation of a Web Archive's holdings, or simply types of usage that some people may feel uncomfortable with concerning the content that they have made available on the Web sometime in the past. It requires a profound understanding of the semantic and cognitive aspects and values of the information aggregated over time, as opposed to the individual pages it is based upon and that are currently searchable via conventional Web search engines. It will also require the development of technical means to counter these challenges. This requires researching the potential of new techniques of large-scale information retrieval including its semantic capabilities and, specifically, the drastically different level of semantic information that can be gleaned from the unprecedented collection of information that is available in Web Archive holdings.

### 4.7 Other activities

A number of other activities are currently being explored in our group, addressing issues such as archiving non-traditional environments on the Internet, with a specific focus on Virtual Worlds such as SecondLife. While approaches to completely preserve the data structures as well as rendering environment may provide the most comprehensive solution for preserving the world as such, the interaction happening in these worlds may be at least as important to capture. To this end we are investigating automated means of filming activities in certain areas of these virtual worlds while trying not to capture personal information about avatars and their users.
Another line of activity aims at recovering information from obsolete digital objects by analyzing the structure of an object's code in order to determine regular patterns that can help in its decoding.
We are also performing first experiments to model how "forgetting" may be implemented in a digital archive, specifically in settings where multiple versions of an object are archived incrementally, each of which may be available via a range of different migration paths, analyzing information growth between versions as well as the characteristics of complementary object representations in different formats.

## 5 Summary

Digital Preservation represents some tremendous challenges that need to be addressed within the near future if we want to ensure that the wealth of information that we are creating continuously is to remain accessible and usable in the near and far future. This affects all levels of society and all business domains, starting from cultural heritage institutions and science data centres, via industry and business, up to small and medium enterprises as well as home users.
A growing community of researchers has evolved in the last few years that are investigating the specific challenges in digital preservation from their respective backgrounds. However, in order to really solve the challenges, a wider range of experts particularly from core computer science disciplines needs to get involved and excited about the research challenges waiting in this field, if possible making DP an integral part of all IT systems, ultimately achieving sustainable computing at all levels.

The digital preservation community has become aware of this need. Specific initiatives such as the DPE Digital Preservation Challenge have been devised to attract the interest of Computer and Information Science students and demonstrate the hard tasks waiting to be solved.
Complementing this, specific curricula are being developed to train experts in the field of digital preservation. In this tradition we find the series of summer schools organized by DELOS[10] and nestor[11], as well as dedicated curricula initiatives such as DigCCurr[12], the German initiative for a training course for professionals in Digital Preservation, as well as an emerging initiative for a European Master in this domain aiming at providing a solid education for future experts to advance the field. Digital Preservation is also represented as a major field of specialization in the Digital Library curriculum developed under an NSF grant by Ed Fox and his team at Virginia Tech[13].
While the field is still young, it has matured to a level of considerable complexity and specialization. In order to solve the challenges ahead of us, however, the preservation community needs to ensure it remains open and manages to attract professionals from different backgrounds, including but definitely not limited to, computer science experts, to jointly address the challenges that our information society is facing.
On the other hand, Computer Science has to accept the need for achieving sustainable computing, considering system operation, maintenance and replacement as an integral part of the system design and development process. Once this level of IT system maturity is reached, DP will come "for free" as part of systems operation, rather than as a separate add-on. Until then, we will need to continue investing considerable efforts to mitigate the risks threatening the long-term availability of digital information.

## References

[1] C. Becker, H. Kulovits, M. Kraxner, R. Gottardi, and Andreas Rauber. (2009).
An Extensible Monitoring Framework for Measuring and Evaluating Tool Performance in a Service-oriented Architecture. In: Proceedings of the 9th International Conference on Web Engineering (ICWE 2009), LNCS 5648, Springer.

[2] Christoph Becker, Andreas Rauber. (2009). Requirements modelling and evaluation for digital preservation: A COTS selection method based on controlled experimentation. In: Proceedings of the ACM Symposium on Applied Computing (SAC'09), Track 'Requirements Engineering'. Honolulu, Hawaii, USA, March 9-12, 2009.

[3] Christoph Becker, Andreas Rauber, Volker Heydegger, Jan Schnasse, and Manfred Thaller. (2008). Systematic Characterisation of Objects in Digital Preservation: The eXtensible Characterisation Languages. Journal of Universal Computer Science, 14(18):2936-2952.

[4] DPE Project Consortium. (2007). Research Roadmap. Project Deliverable DPE-D7.2. http://www.digitalpreservationeurope.eu/publications/reports/dpe_research_roadmap_D72.pdf

[5] M. Guttenbrunner, C. Becker, A. Rauber, and C. Kehrberg. (2008). Evaluating strategies for the preservation of console video games. In Proceedings of the Fifth international Conference on Preservation of Digital Objects (iPRES 2008), 115-121.

[6] M. Guttenbrunner (2009). Evaluating the effects of emulation environments on rendering digital objects, Planets Deliverable PP/5-D2

[7] ISO14721. (2003). Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003).

[8] R. Kimball and M. Ross. (2002).The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition). Wiley.

[9] H. Kulovits and A. Rauber (2008). Preservation Planning with Plato. In: Proceedings of the All Russian Conference on Digital Libraries (RCDL 2008).

[10] R. Mayer and A. Rauber. (2009). Establishing Context of Digital Objects' Creation, Content and Usage. In Proceedings of the JCDL Workshop on Innovation in Digital Preservation (InDP 2009), Austin, Texas, USA. June 19 2009.

[11] R. Mayer, R. Neumayer, and A. Rauber. Data Recovery from Distributed Personal Repositories. In: Proceedings of the 13th European Conference on Digital Libraries, ECDL 2009, LNCS, Springer. Corfu, Crete, September 2009.

[12] A. Rauber, M. Kaiser, and B. Wachter. (2008). Ethical Issues in Web Archive Creation and Usage – Towards a Research Agenda. In: Proceedings of the 8th International Web Archiving Workshop, Aalborg, Дnemark.

[13] C. Rauch, and A. Rauber. (2004). Preserving digital media: Towards a preservation solution evaluation metric. In Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL 2004) (p. 203-212). Berlin, Heidelberg: Springer.

[14] S. Strodl, F. Motlik, K. Stadler, and A. Rauber. (2008). Personal & SOHO Archiving, In: Proceedings of the Joint Conference on Digital Libraries (JCDL 2008), June 16-20, 2008, Pittsburgh, Pennsylvania, USA.

[15] S. Strodl, C. Becker, R. Neumayer, and A. Rauber. (2007). How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure. In: Proceedings of the ACM IEEE Joint Conference on Digital Libraries (JCDL'07), Vancouver, British Columbia, Canada, June 18-23, 2007.

[16] C. Voges, V. Мдrgner, and T. Fingscheidt. (2008). Digital Data Storage on Microfilm - Error Correction and Storage Capacity Issues. In Proceedings of IS&T Archiving Conference, Bern, Switzerland, June 2008.

[1] Planets, http://www.planets-project.eu/
[2] CASPAR, http://www.casparpreserves.eu
[3] Digicult Projects FP7, http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-projects-fp7_en.html
[4] Keep, http://www.keep-project.eu/
[5] PrestoPRIME, http://www.prestoprime.eu/
[6] LiWA, http://www.liwa-project.eu/
[7] PROTAGE, http://www.protage.eu/
[8] SHAMAN, http://www.shaman-ip.eu/
[9] DPE Challenge http://www.digitalpreservationeurope.eu/challenge
[10] DELOS Summerschools on Digital Preservation, http://www.dpc.delos.info/ss08/
[11] nestor winter/spring/summerschools on digital preservation, http://nestor.sub.uni-goettingen.de/education/index.php?lang=en
[12] DigCCurr, http://ils.unc.edu/digccurr/index.html
[13] DL Curriculum, http://curric.dlib.vt.edu/

# БОРЬБА С ПЛАГИАТОМ, ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ В АТТЕСТАЦИИ НАУЧНЫХ КАДРОВ

# ANTIPLAGIARISM MEANS, DIGITAL LIBRARIES USE FOR SCIENTISTS ATTESTATION

# Внедрение системы Антиплагиат в Российской Государственной Библиотеке

© Романов Михаил Юрьевич Житлухин Дмитрий Анатольевич

ЗАО «Форексис»

mromanov@forecsys.ru dzitlukhin@forecsys.ru

**Аннотация**

В докладе приведено описание внедрения системы Антиплагиат в Российской Государственной библиотеке.

В рамках этого внедрения в РГБ поставлена система поиска заимствований по базе оцифрованных авторефератов и диссертаций, а также реализована поддержка проекта «научный поиск».

Приведено подробное изложение результатов внедрения, состав системы, вопросы безопасности и результирующие технические и скоростные характеристики.

## 1 Цели и задачи системы

### 1.1 О системе Антиплагиат

Интернет-сервис www.antiplagiat.ru (см. [1]) был создан в 2005 г. (см. [5]) для проверки текстовых документов на наличие заимствований из общедоступных сетевых источников. Изначально система разрабатывалась для внедрения в крупный коммерческий ВУЗ МИЭМП. Функциональное ядро системы Антиплагиат использует алгоритмы, разработанные сотрудниками ВЦ РАН и компании Форексис (см. [6]).

Стратегической задачей проекта Антиплагиат является повышение качества российского образования и научной деятельности преимущественно в тех случаях, когда требуется творческая работа по написанию рефератов, курсовых и дипломных работ, диссертаций и иных материалов. Эта задача решается путем побуждения обучающихся к самостоятельному написанию текстов, а не к созданию их, например, путем компиляции текстов, найденных в интернете и других источниках, касающихся заданной тематики.

В сентябре 2005 года была запущена в эксплуатацию тестовая версия интернет-сервиса www.antiplagiat.ru. Впервые использование системы Антиплагиат было включено в обязательный учебный процесс вуза в ноябре 2005 года. В 2006 году на пятом Конкурсе русских инноваций разработка получила приз Минсвязи РФ «За лучший проект в области телекоммуникаций». Летом 2006 года Советом Ассоциации негосударственных вузов РФ принято решение рекомендовать членам Ассоциации применение сервиса. В июне 2007 года использование в ВУЗах интернет-сервиса www.antiplagiat.ru было рекомендовано Советом по качеству образования при Рособрнадзоре РФ. К июлю 2007 года была разработана адаптированная под нужды Высшей аттестационной комиссии (ВАК) Минобрнауки РФ система для обязательной проверки на плагиат всех диссертаций и авторефератов и начата ее эксплуатация.

Система Антиплагиат постоянно развивается и расширяет функциональность. Постоянно происходит совершенствование алгоритмов поиска для выдачи более корректных результатов и для учёта ситуаций и особенностей, не обрабатывавшихся ранее. В 2006 г. введена возможность развёртывания системы на оборудовании держателей документов. В 2007 г. реализована поддержка распределённой проверки по совокупности коллекций документов. В 2008 г. введено использование для ВУЗов механизма «сигнальных статистик», подсказывающих преподавателю, что в проверяемом документе есть подозрение на попытку «обхода» системы. Постоянно развивается и дорабатывает пользовательский интерфейс.

В настоящий момент систему Антиплагиат используют такие вузы, как Высшая школа экономики, Московский институт экономики, менеджмента и права, Московский городской психолого-педагогический университет, Московский государственный педагогический университет, Современная гуманитарная академия и другие. Создан специальный сайт именно для работы с ВУЗами (см. [2]). А публичный интернет-сервис используют около 100 тысяч пользователей в России и за рубежом.

Внедрение системы в Российскую Государственную Библиотеку началось в первой половине 2008 г. Во второй половине 2008 г. открыт сайт для поиска по базе электронных документов РГБ (см. [3]). В настоящий момент продолжается активное расширение функциональности системы и построение новых сервисов на её основе.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

Система Антиплагиат разработана и поддерживается компанией Форексис.

### 1.2 Задачи системы в РГБ

Российская государственная библиотека (РГБ) является уникальным хранилищем диссертаций, защищенных в стране с 1944 года по всем специальностям, кроме медицины и фармации (см. [4]). Всероссийский фонд диссертационных работ был создан в 1944 году. Сейчас в фонде диссертаций хранятся свыше 900000 томов диссертаций, причём ежегодно поступает около 30000 диссертаций (20000 кандидатских и 10000 докторских). В 2003 году была начата работа по оцифровке этого хранилища документов.

Перед системой Антиплагиат стояла задача организации эффективного поиска заимствований по оцифрованной базе диссертаций, а также построения различных сервисов на основе поискового движка. К числу этих сервисов относится:

- предоставление ВУЗам и другим пользователям системы возможности проверки текстов по базе диссертаций Российской Государственной Библиотеки;
- реализация ядра системы «научный поиск»;
- предоставление сотрудникам РГБ аналитических средств, обеспечивающих работу с метаданными хранилища диссертаций.

## 2 Результаты внедрения

Первой и ключевой задачей системы было предоставление сервиса по проверке

### 2.1 Состав системы

Система Антиплагиат.РГБ состоит из следующих модулей.

- Хранилище документов (коллекция). Содержит набор документов, необходимые индексы и позволяет осуществлять поиск по ним.
- Модуль импорта из систем РГБ. Позволяет по расписанию либо вручную загружать документы из источников РГБ в коллекцию Антиплагиат. Поддерживает библиографический стандарт MARC. Обнаруживает и корректно обрабатывает изменения в уже загруженных документах (то есть позволяет работать в режиме обновления метаданных).
- Модуль доступа к внешним экземплярам системы Антиплагиат. Позволяет извне проверять документы по базе диссертаций, а также из системы Антиплагиат.РГБ проверять по внешним коллекциям.
- Пользовательский интерфейс в виде сайта antiplagiat.rsl.ru. Предоставляет функциональность, сходную с общедоступным сервисом www.antiplagiat.ru для пользователей читальных залов.
- Модуль предоставления функций системы Антиплагиат через WebService для внутреннего использования сотрудниками РГБ.
- Модуль для аналитической обработки метаданных документов. Предоставляет OLAP-куб с метаданными на Microsoft SQL Server 2005 для внутреннего использования сотрудниками РГБ. У них появляется возможность построения сводных отчетов по любым имеющимся в системе атрибутам с различными вариантами агрегации. Пример: вывести средний процент цитирования у всех загруженных в систему статей по теме «Психология» с недельной агрегацией по дате публикации.
- Модуль администрирования системы (веб-интерфейс). Позволяет управлять правами доступа и настройками системы (такими, как импорт).
- Модуль поддержки научного поиска. Предоставляет расширенную функциональность доступа к ядру системы для потребностей научного поиска.
- Модуль просмотра отчётов о проверке offline (Antiplagiat Report Viewer). Позволяет экспортировать результаты проверки в специальный файл (контейнер), включающий в себя исходный документ, отчёт о проверке, а также специальные ключи, обеспечивающие подлинность просматриваемого файла.

### 2.2 Качество и скорость проверки

Система Антиплагиат изначально разрабатывалась для проверки больших объёмов текста, загружаемых пользователями сети internet, в связи с чем она с самого начала оптимизировалась по времени работы - время проверки среднего документы ещё в 2005 году составляло несколько секунд. Со временем, благодаря открытой проверке на прочность системы в internet'e, открывались некоторые способы её "обхода" - способы автоматической или полуавтоматической переработки текста, после которых он не распознавался системой как списанный. Система шла в ногу со временем - при выявлении новых методов обхода они оперативно закрывались, что порой приводило к некоторому снижению производительности. В последних версиях был внедрён новый поисковый движок, уникальный индекс которого позволил решить проблемы производительности. Также теперь имеется возможность вручную управлять соотношением

«цена/качество» - можно сделать индекс маленьким, при этом проверка текстов будет выполняться быстро, но некоторые мелкие фрагменты могут быть не замечены, а можно большим, обеспечив высокое качество проверки, но потребуется больше дискового пространства и времени на проверку. Наконец, применение современных SSD-накопителей повысило скорость работы системы практически на порядок. Теперь время проверки составляет в среднем менее 1 секунды.

### 2.3 Интеграция различных коллекций документов

С развитием системы появилась возможность хранить отдельные независимые базы текстов - коллекции. Документ может проверяться как по одной, так и по нескольким коллекциям одновременно, что обеспечивает дополнительную гибкость при проверке. При подключении к системе различных организаций, для каждой из них заводится своя коллекция. После этого держатель коллекции может распоряжаться этой коллекцией по своему усмотрению - предоставлять доступ к ней другим пользователям, определять правила работы со своей коллекцией - можно ли будет им скачивать из неё документы или же использовать только для поиска цитирования без явной загрузки документов.

В Российской Государственной Библиотеке система работает на выделенном сервере, который находится в ведении ИТ-отдела РГБ. Таким образом, любой запрос на проверку по базе диссертаций автоматически направляется пользовательской системой на центральный шлюз Антиплагиат, который уже разбирает запрос и переадресует его на систему, расположенную в РГБ. Результаты проверки аналогичным образом через центральный шлюз доставляются до конечного пользователя.

В следующих трёх параграфах рассматриваются вопросы безопасности системы.

### 2.4 Сохранность текстовых баз

Система не реплицирует текстовые коллекции. При установке локальной версии у заказчика вся его текстовая база хранится на его сервере и никогда не реплицируется на другие машины в сети. Все проверки по данной коллекции обрабатывают сервера заказчика, что обеспечивает сохранность текстовой базы.

Систему можно настроить таким образом, чтобы она не предоставляла во вне целиком найденные источники заимствования. В этом режиме производится процедура «сокращения источника», то есть из него оставляются только те фрагменты, по которым найдено совпадение, либо можно запретить выдачу источников совсем. Это позволяет защитить коллекцию документов от распространения.

### 2.5 Безопасность протоколов

Взаимодействие систем www.antiplagiat.ru и Антиплагиат.РГБ осуществляется по шифрованному каналу с защитой по протоколу SSL. Авторизация системы проводится по специально подготовленным идентификационным ключам. Это гарантирует, что только один экземпляр системы может функционировать, как «библиотека документов РГБ» и никакая внешняя система не может получить несанкционированный доступ к пересылаемым данным.

Доступ к проверке по базе диссертаций имеют только те пользователи, которым администраторы системы явным образом открыли доступ к такой проверке. В настоящий момент поддерживается ограниченное предоставление доступа к проверке, причём ограничение может задаваться по числу проверяемых документов, по их суммарной длине и по времени использования доступа.

### 2.6 Подлинность отчётов

У каждого отчёта имеется несколько цифровых подписей. Необходимость нескольких подписей вызвана распределённостью системы и тем, что один отчёт строится сразу несколькими коллекциями, находящимися на разных серверах у разных подписчиков в сети.

Каждая коллекция имеет свои ключи и подписывает результаты своей работы - фрагмент отчёта, в дальнейшем называемый ревизией. Ревизия включает в себя контрольную сумму текста, вычисленную по алгоритму MD5 вместе со списком найденных источников и блоков цитирования. Опционно в ней могут содержаться сокращённые тексты самих источников (если включено сокращение источников, см. п. 2.4), а также атрибуты как отдельных источников, так и коллекции в целом. Наличие в ревизии контрольной суммы проверяемого документа гарантирует, что ревизия строилась именно для проверяемого документа, а не для какого-то другого текста (что исключает подмену текста проверяемого документа на другой, например пустой, с целью получения отчёта о полной оригинальности), а цифровая подпись самой ревизии гарантирует, что ревизия построена проверяющей коллекцией (исключается искажение самой ревизии).

После того, как все фрагменты отчёта построены, они через систему шлюзов отправляются к инициатору проверки. Например, если проверяемый документ находился в РГБ, то все фрагменты отправляются в РГБ, если с сайта www.antiplagiat.ru, то в хранилище данного сайта. Инициатор проверки собирает их в единый отчёт и подписывает его своим ключом, что гарантирует, что в отчёт включены все необходимые ревизии. Помимо этого, подпись отчёта позволяет однозначно определить создателя данного отчёта - при его просмотре посредством бесплатной

программы ReportViewer создатель отображается в нижней части экрана.

Таким образом, подделки отчётов полностью исключены.

### 2.7 Редактирование и просмотр отчётов

Несмотря на то, что система фильтрует незначащие заимствования, некоторые из них остаются в отчётах, т.к. значимость или незначимость некоторого фрагмента может оспариваться. Если невозможно точно определить, что фрагмент незначим, то он остаётся в отчёте и уже пользователь должен принять решение, значим он или нет. Для упрощения работы по анализу текста пользователю предоставлена возможность самостоятельно удалять некоторые найденные участки из отчёта, а также сохранять изменённые отчёты прямо в своей коллекции. Для защиты от злоупотребления удалением найденных блоков, при применении данной функции в отчёте появляется предупреждение, что некоторые блоки были удалены, а также возможность вернуть все блоки обратно. Также предусмотрена возможность экспортировать отчёты, в том числе и с удалёнными блоками, в виде отдельных файлов и просматривать их с помощью отдельного приложения, не требующего подключения к сети - AntiPlagiat ReportViewer.

### 2.8 Научный поиск

Сервис представляет собой средство сравнения более-менее обширного текста пользователя с каждым из текстов, имеющихся в хранилище. В отличии от системы "Антиплагиат" система "Научный поиск" в основном предназначена для обнаружения небольших совпадений, таких, как общие цитаты, имена лиц и учреждений, заглавия литературных произведений, устоявшиеся фразеологизмы и обороты речи. Предполагается, что получившаяся подборка документов позволит исследователю получить представление о направлении работ в смежных, параллельных и вообще как-то ассоциированных областях.

Сейчас область поиска системы «Научный поиск» совпадает с областью поиска системы Антиплагиат.РГБ, то есть поиск осуществляется по коллекции авторефератов и диссертаций по всем отраслям знаний с 1998 г. защиты.

Система возвращает список документов, которые похожи (содержат полные или частичные заимствования) на исследуемый.

Для каждого из них можно снова провести поиск - и получить список связанных уже с ним документов.

Список проверенных документов отражается в верхней части экрана. Всегда можно вернуться к любому из них. Если текущий документ имеется в электронном каталоге - можно перейти туда.

Отчёт содержит отрывки текущего документа, содержащие заимствования. Совпадающие фрагменты выделены жёлтым цветом. Стрелочки слева позволяют перемещаться от одного совпадающего фрагмента к другому. Номер в квадратных скобках - это порядковый номер документа в списке источников, из которого заимствован выделенный фрагмент.

### 2.9 Технические характеристики

Формирование хранилища

- Исходный объем PDF файлов – 2.5 TB, общий объем ANSI текстов - 57 GB
- Число документов – 260000, средний размер текста – 222 KB
- Общее время создания хранилища – 26 ч.
- Объем получившегося хранилища системы – 46 GB

Характеристики

- Оборудование: 2 двуядерных процессора Xeon 1.6 GHz, 4 GB Ram
- Реально используется только один жесткий диск емкостью 135 GB
- Время проверки документа по локальному хранилищу РГБ – не более 3 сек.
- Время проверки документа по 2-хранилищам РГБ и Антиплагиат одновременно – не более 5 секунд.

## 3 Особенности устройства системы

В данном разделе будет кратко описан основной модуль системы - коллекция, отвечающая за хранение документов, отчётов, а также поддержку индексов и выполнение поиска.

### 3.1 Хранение документов и атрибутов.

Коллекция хранит все данные в виде архивов - больших бинарных файлов, внутри которых подряд сохранены отдельные документы коллекции. Данных подход существенно экономит дисковое пространство за счёт отсутствия неиспользуемой ёмкости в конце кластеров файловой системы диска, а также существенно ускоряет поиск нужного текста по сравнению с хранением текстов в отдельных файлах. Для дополнительной экономии места поддерживается сжатие текстов различными архиваторами, на данный момент используются алгоритмы deflate (также применяется в архиваторе ZIP) и BWT (применяется в архиваторе BZ2). Для хранения данных определённого типа заводится свой архив, на данный момент их 6:

- Архив текстов документов. Обязателен, хранит тексты документов;
- Архив нормализованных текстов. Необязателен, в нём сохраняются видоизменённые тексты документов, подготовленные к поиску цитирования - все слова нормализованы (это и дало название данному архиву), т.е. приведены к начальной словоформе, буквы ё заменены на е и т.п.. Может быть перестроен по

архиву текстов, в случае его отключения нормализация текстов выполняется по мере необходимости, что замедляет процесс проверки документов;

- Архив документов. Необязателен, хранит исходные двоичные файлы документов (например *.doc; *.pdf). Нужен только в случае если пользователь захочет получить исходный двоичный файл документа, для поиска цитирования не используется;
- Архив ревизий. Обязателен, хранит построенные фрагменты отчётов;
- Архив атрибутов документов. Обязателен, хранит атрибуты каждого документа вместе с историей их изменения - имеется возможность проследить историю правок пользователями атрибутов указанного документа;
- Архив кэширования. Необязателен, хранит дополнительные двоичные данные для поискового ядра по часто используемым источникам. В случае его отключения ядро перестраивает эти данные при каждом обращении к источнику, что увеличивает время поиска.

Все архивы могут размещаться на отдельных физических дисках, что позволяет оптимизировать дисковый ввод/вывод и достигать оптимальной производительности дисковой подсистемы сервера.

### 3.2 Индексы.

Индексы - это структуры данных, позволяющие существенно ускорить поиск за счёт организации системы навигации по текстам и отсутствия необходимости перебора всей коллекции текстов для проверки факта наличия заданного фрагмента.

У коллекции есть два индекса, выполняющие одну и ту же задачу, но обладающие разными характеристиками:

- Постоянный индекс. Оптимизирован для хранения больших объёмов данных, обеспечивает быстрый поиск при любых объёмах коллекции, нечувствителен к её размеру. Относительно компактен, занимает мало места на диске. Добавление новых записей требует существенных временных затрат - приходится полностью пересматривать весь индекс.
- Временный индекс. Позволяет быстро добавлять новые документы, но с ростом объёма время поиска увеличивается. Предназначен для временной индексации документов, пока они не занесены в постоянный индекс.

Оба индекса поддерживают усечение, позволяющее обменивать качество проверки на снижение времени поиска и объёма индексов. Усечение индекса сделано таким образом, что существенное снижение его объёма приводит к несущественному снижению качества поиска. В частности, при принудительном усечении индекса в 4 раза средняя оценка оригинальности по тестовому корпусу, сформированному из загруженных пользователями документов выросла всего на 1% и составила 63%. При усечении индекса более, чем в 8 раз, качество начинает существенно снижаться. При усечении в 64 раза средняя оценка того же корпуса выросла на 14% и составила 76%. При применении неусечённого индекса оценка составила 62% оригинальности.

Потеря качества проявляется в основном на заимствованиях небольшой длины, поэтому если нужно отлавливать копирование только больших блоков (например, сразу по несколько страниц), можно использовать усечение вплоть до 64 раз, качество будет оставаться приемлемым.

При усечении индекса в N раз время поиска по данному индексу также снижается в N раз.

Все индексы позволяют дополнительно держать в оперативной памяти небольшой объём данных, ускоряющий обработку документов с большим процентом оригинальности. Данная надстройка работает по принципу hash-таблицы и осуществляет быстрый отсев фрагментов текста, заведомо отсутствующих в индексах на диске. Соответственно, отсеянные фрагменты искать на диске бесполезно, что позволяет экономить на обращениях к внешней памяти.

Существует возможность отключения индексов у коллекции. Отключение временного индекса позволяет очень быстро добавлять много документов в индекс, но поиск по ним начнётся только после перестройки постоянного индекса. Отключение обоих индексов делает невозможным поиск цитирования по данной коллекции. Целесообразно, если коллекция используется исключительно как хранилище документов, позволяет сэкономить немного памяти.

## 4 Перспективы системы

В ближайшее время планируется внедрение следующих компонентов:

- Приоритетов коллекций и отдельных документов с целью выявления первоисточников цитирования;
- Возможность задания параметров проверки для каждого документа индивидуально;
- Разработка дополнительных средств мониторинга использования текстовых баз и контроля отсутствия несанкционированного доступа;
- Добавление третьего промежуточного индекса в коллекцию;

### 4.1 Приоритеты коллекций

С ростом числа коллекций, а также их объёма, возникла проблема поиска первоисточников цитирования. Один и тот же фрагмент текста может содержаться в большом количестве источников, их число может достигать сотни и даже тысячи

документов. С целью экономии времени каждая коллекция (при конфигурации по умолчанию) ищет только один источник для каждого фрагмента текста. Если о документах ничего не известно, то выбрать документ-источник из множества документов, содержащих нужный фрагмент, можно только псевдослучайным образом - например, взять документ, попавший в коллекцию раньше других. К сожалению, данная стратегия иногда приводит к некорректным результатам - например, при цитировании текста закона или общеизвестного литературного произведения источником может быть объявлена другая диссертация, цитирующая тот же самый текст. Конечно, данное недоразумение будет разрешено при просмотре отчёта пользователем, но это заставляет пользователя затратить дополнительное время на редактирование блоков цитирования с целью переквалифицирования данного фрагмента.

При наличии у документов приоритетов, система сможет автоматически выбирать подходящий источник. Задание приоритетов потребует больше времени на грамотное составление хранилища, но существенно упростит работу с системой в дальнейшем. Вероятно, будет целесообразно назначить высокие приоритеты тем документам, которые заведомо являются первоисточниками текста, на их отбор и потребуется дополнительное время.

Помимо приоритетов документов вводятся также приоритеты целых коллекций, что позволяет регулировать значимость источников из различных хранилищ на этапе сборки отчёта из ревизий. При обнаружении фрагмента текста в источниках из разных коллекций, будет выбран источник из более приоритетной.

Например, для кого-то более важно заимствование из документов, находящихся в открытом доступе в сети internet, а для кого-то - из хранилища РГБ. Для достижения желаемых результатов достаточно будет поднять приоритет нужной коллекции.

Приоритеты коллекций могут также использоваться для исключения легитируемого цитирования из общей оценки документа - весь текст, найденный в высокоприоритетной коллекции не учитывается при вычислении итоговой оценки документа.

### 4.2 Индивидуальные параметры проверки

Каждая коллекция имеет множество настроек, определяющих, как она будет выполнять поиск. Сейчас все эти параметры задаются в конфигурационном файле коллекции и применяются при проверке любых документов.

Иногда возникает необходимость задания параметров индивидуально для каждого документа - например, для проверки одних документов достаточно грубой оценки оригинальности, но очень важна скорость проверки, для других, наоборот, в первую очередь важно качество. В частности, может потребоваться полный список источников для каждого фрагмента текста, построение которого может занять существенное время.

Предполагается сделать возможным включение в запрос на проверку дополнительных параметров, с их последующим сохранением в ревизии, с целью осуществления более гибкого применения системы и расширения сферы её применимости.

### 4.3 Контроль передаваемых данных

Несмотря на то, что система препятствует распространению текстовых баз, остаются некоторые вопросы по поводу невозможности извлечения текстов из коллекции сторонними людьми через её интерфейс, используемый для проверок документов. Действительно, невозможно дать гарантию того, что в кодировании системы безопасности не было допущено ошибок. С другой стороны, имеется гипотетическая возможность сознательного создания дырок для воровства чужих текстовых баз. Для контроля честности протокол системы сделан таким образом, что имеется возможность сохранять все передаваемые между модулями системы данные и обеспечить тем самым контроль за передачей текстов и отчётов.

Реализации контроля мешает сравнительно большое количество команд в протоколе системы, т.к. он рассчитан на широкую сферу применения. Даже в случае реализации модуля для анализа передаваемых данных с открытым кодом, в нём будет достаточно сложно разобраться.

Для решения данной проблемы предполагается выпустить модуль туннелирования, который отвечает за передачу данных между коллекцией подписчика и его шлюзом в упрощённой форме, с поддержкой необходимого минимума команд и без шифрования. В результате можно будет проанализировать каждый переданный системой байт данных и убедиться, что ничего лишнего передано не было.

### 4.4 Добавление промежуточного индекса в коллекцию

Как было описано ранее (см п.3.2) у коллекции имеется два индекса. К сожалению, при большой нагрузке по добавлению документов в индекс производительность коллекции падает - она либо будет постоянно занята перестройкой постоянного индекса, либо временный индекс станет большим и превратится в "узкое место" при поиске цитирования. Предполагается, что промежуточный индекс будет устроен аналогично постоянному, но по объёму будет существенно меньше его, что позволит выполнять его перестройку существенно быстрее.

## Литература


[1] Публичный сайт системы Антиплагиат. http://www.antiplagiat.ru.

[2] Сайт системы Антиплагиат для ВУЗов. http://corp.antiplagiat.ru.

[3] Сайт системы Антиплагиат в РГБ. http://antiplagiat.rsl.ru.

[4] Сайт Электронной Библиотеки Диссертаций РГБ. http://diss.rsl.ru/.

[5] Ю.И. Журавлёв и др. «Система распознавания интеллектуальных заимствований «Антиплагиат» // Доклады 12-й всероссийской конференции «Математические методы распознавания образов» (ММРО-12). Москва, 2005. С. 329-332.

[6] Ю.И. Журавлёв и др. «О проекте «Антиплагиат» // Доклады международной конференции «Интеллектуализация обработки информации» - 2006. Симферополь, 2006. С. 92-94.


## On integration Antiplagiat system in Russian State Library


Romanov Mikhail Yurievich
Zhitlukhin Dmitriy Anatolievich



In the article the description of the integration of the Antiplagiat system in the Russian State Library is given.

Within this integration the system of searching for adoptions over the dissertations and abstracts base is implemented in the RSL, and also the "scientific search" project support is developed.

The detailed exposition of the integration results, the system structure, the security matter and final technical and speed characteristics are given.

# Централизованная электронная библиотека результатов диссертационных исследований

©И. Д. Котляров

Северо-Западный институт печати
Санкт-Петербургского государственного университета технологии и дизайна
Санкт-Петербург
lrpg@mail.ru

**Аннотация**

Предлагается создать единую электронную библиотеку публикаций соискателей ученой степени для контроля качества диссертационных исследований. Сформулированы требования к данной библиотеки, описан алгоритм ее функционирования и дополнительные возможности, которые она должна предлагать научному сообществу.

## 1 Введение

Одной из важнейших задач при подготовке специалистов высшей квалификации (кандидатов и докторов наук) является обеспечение объективной и адекватной оценки достигнутых ими результатов научным сообществом. Для достижения этой цели Высшая аттестационная комиссия (ВАК РФ) регулярно формирует список ведущих российских рецензируемых журналов [2], в которых должны быть опубликованы результаты диссертационных исследований на соискание ученой степени кандидата и доктора наук (т. н. «список ВАК»). Соискатель степени кандидата наук (СК) обязан опубликовать в профильном (то есть рекомендованным экспертным советом ВАК по соответствующей специальности) журнале не менее одной статьи; от соискателя докторской степени (СД) требуется не менее семи статей. Внесенные недавно изменения в требования к формированию «списка ВАК» [1] позволяют, в случае их неукоснительного соблюдения, устранить ряд проблем, связанных с публикацией научных результатов диссертантов (новый «список ВАК», сформированный в соответствии с этими требованиями, должен вступить в силу предположительно с 01.01.2010):

1. Новые правила требуют обязательного рецензирования предоставляемых к публикации статей, высылки рецензий авторам, а также возможность запроса этих рецензий экспертными советами ВАК. В настоящее время весьма распространена ситуация, когда автор вместе со статьей должно также предоставить отзыв самостоятельно найденного им рецензента (сам журнал рецензированием не занимается). Изредка, но все еще встречается прием статей к публикации вообще без рецензирования. Наконец, «ваковские» журналы, издаваемые вузами, часто предлагают аспирантам, обучающимся в этих вузах, опубликовать в них статью без независимого рецензирования – только по отзыву своего научного руководителя. Для этих целей в данных журналах предусмотрены специальные разделы (например, «Страницы аспирантов и докторантов» в «Вестнике ИНЖЭКОНа»). Данное нововведение поставит барьер на пути этой практики;
2. Журналы из «списка ВАК» отныне обязаны иметь полнотекстовую сетевую версию в Интернете (при этом аннотации статей, сведения об авторах, ключевые слова и библиография должны быть в свободном доступе;
3. От журналов из «списка ВАК» теперь требуется предоставлять сведения об опубликованных в них статьях в систему Российского индекса научного цитирования, что позволяет оценивать востребованность результатов диссертационных исследований научным сообществом;
4. По косвенной информации можно сделать вывод о том, что в новом «списке ВАК» печатные и электронные издания будут уравнены в правах (до настоящего времени ВАК признавал статьи в зарегистрированных в Информрегистре электронных журналах в качестве обычных публикаций, а не публикаций в журнале из «списка ВАК», что в наше время представляется анахронизмом). Эта мера также будет способствовать упрощению доступа к результатам диссертационных исследований;
5. В соответствии с новыми требованиями, журналам из «списка ВАК» запрещается взимать плату с аспирантов за публикацию их статей (ранее ВАК не рекомендовал взимание платы). Это

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

требование очень важно: взимание платы за публикацию в ряде случаев ведет к профанации понятия «рецензируемый журнал», так как научное издание превращается из распространителя качественной и достоверной информации в продавца журнальных площадей и за плату публикует какие угодно материалы. Пример: громкий скандал с «Журналом научных публикаций аспирантов и докторантов» [5], который пропустил статью, представляющую собой случайно сгенерированный компьютерной программой текст на английском, переведенный на русский автоматическим переводчиком (рецензент при этом дал высокую оценку работе). Нужно отдать должное оперативности ВАК – в кратчайшие сроки было принято решение о исключении этого издания из списка рецензируемых журналов [4]. Другим примером могут быть «Известия РГПУ им. Герцена. Аспирантские тетради» (http://www.bookhouse.ru/?pagename=asptetr_price), которые предлагают услугу по срочной публикации статей (стоимость публикации одной авторской страницы составляет от 480 до 860 руб. в зависимости от того, насколько срочно автору нужно опубликовать свою работу). Разумеется, журналы, скорее всего, все равно продолжат взимать плату за публикацию (в частности, не запрещается ее взимание с соискателей и с докторантов; плату также можно брать за подписку, за авторский экземпляр и т.д.), поскольку даже уважаемым и престижным журналам эта плата нужна для окупаемости издания, однако новое требование ВАК, хочется надеяться, приведет к ликвидации откровенно коммерческих некачественных журналов.

На наш взгляд, необходимо разграничить требования, предъявляемые к публикации результатов собственно научных исследований и результатов, защищаемых в диссертациях (прежде всего – кандидатских). Диссертация, особенно кандидатская, является не только научной, но и квалификационной работой. Статьи многих аспирантов не дотягивают до уровня, требуемого этими журналами (известно высказывание руководителя одного из ведущих российских журналов по экономике, который сказал, что не хочет добиваться включения своего журнала в «список ВАК», чтобы не засорять его аспирантскими статьями). При этом для повышения уровня аспирантских публикаций есть объективные препятствия: аспиранты – это, как правило, молодые люди, только что закончившие вуз, и имеющие мало опыта как в проведении научных исследований, так и в публикации их результатов. Опыт этот, безусловно, нужно нарабатывать, но вряд ли тренировочной площадкой следует признать лучшие научные журналы страны. Разумеется, если статья аспиранта соответствует их уровню – ее, очевидно, нужно в них публиковать, но при этом существует необходимость в наличии альтернативной – помимо журналов из «списка ВАК» – площадки для публикаций работ СК и (в меньшей степени) СД.

Следует также указать на то, что обнародование и последующая оценка результатов диссертационных исследований заключается не только в публикации посвященных им статей в профильных рецензируемых журналах: СК (СД) должен выступить с докладом о своей работе на нескольких конференция, а непосредственно перед защитой автореферат его диссертации публикуют в Интернете (для СД – на сайте ВАК, для СК – на сайте учреждения, при котором функционирует диссертационный совет) и отправляют в другие вузы по списку рассылки, текст диссертации предоставляют официальным оппонентам, которые должны дать объективную и непредвзятую оценку работе диссертанта. Здесь также существует комплекс проблем:

1. Оппоненты, хотя и считаются назначенными советом, как правило, являются хорошими знакомыми научного руководителя (консультанта) соискателя, и получение отзывов на диссертации, по сути дела, происходит на основе неформальных договоренностей. В частности, хотя по положению ВАК защита диссертации разрешена даже в случае отрицательного отзыва одного из оппонентов и ведущей организации [3], обычно соискатели и их научные руководители стараются все организовать таким образом, чтобы все отзывы были положительными. Вряд ли такой подход можно считать объективной оценкой научных достижений соискателя;
2. На авторефераты, поступившие в другие вузы по списку рассылки, отзывы пишутся только в том случае, если адресатов об этом попросил научный руководитель соискателя. В противном случае конверты с авторефератами отправляются в мусорную корзину. Очевидно, что при таком подходе объективность полученных ДС отзывов представляется сомнительной;
3. Будучи формально обязательным, требование о публикации автореферата на сайте ДС (призванное резко расширить потенциальную аудиторию читателей автореферата) соблюдается не всеми вузами, кроме того, в отдельных случаях вуз ограничивается размещением объявления о предстоящей защите, не давая ссылки на автореферат, и даже не предоставляя информации о дате защиты (что, по сути, превращает защиту из публичного мероприятия в закрытое и является прямым нарушением требований ВАК). Например, для очень многих объявленных защит на сайте Санкт-Петербургского государственного инженерно-экономического университета даты защит найти невозможно. Наконец, если авторефераты представленных к защите диссертаций на соискание ученой степени доктора наук сравнительно легко (они выложены на сайте ВАК), то в случае, если у ученого возникает желание ознакомиться с авторефератами

кандидатских диссертаций, ему необходимо проверить сайты всех учреждений, при которых открыты диссоветы по соответствующей специальности, что зачастую затруднительно (кроме того, не все учреждения – в отличие от ВАК, на сайте которого сохраняются все авторефераты докторских диссертаций – поддерживают на своем сайте электронный архив авторефератов кандидатских диссертаций);

4. Конференции, призванные обеспечить апробацию полученных исследователями результатов, за редким исключением собирают малое число участников, а сборники материалов публикуются тиражом, равным числу участников этой конференции + список обязательной рассылки. Найти эти материалы в вузовских библиотеках невозможно (за исключением библиотек тех вузов, где происходили данные конференции). Это также затрудняет ознакомление других исследователей с результатами соискателя. В ряде случаев докладчики ограничиваются заочным участием и просто присылают тезисы или материалы для публикации в сборнике трудов конференции. Из-за этого никакой дискуссии по представленным в заочных докладах материалам не проводится, что превращает апробацию в фикцию.

В своем стремлении повысить качество диссертационных исследований ВАК активно обращается к потенциалу сетевого информационного пространства непоследовательны. Как уже упоминалось, есть вполне логичное в современных условиях требование о публикации авторефератов в Интернете, от рецензируемых журналов требуется наличие полнотекстовой версии в сети Интернет, и даже осуществляются Интернет-трансляции защит диссертаций. Отметим, правда, что последняя мера пока не представляется эффективной – пропускная возможность Интернета у большинства людей, потенциально заинтересованных в такой трансляции, недостаточна для ее просмотра.

Как нам представляется, наращивание использование возможностей сетевого информационного пространства позволило бы решить многие из отмеченных выше проблем. Ниже предлагается несколько вариантов такого использования.

## 2 Электронная библиотека научных публикаций

Как уже отмечалось, необходимо, наряду с журналами из «списка ВАК», предложить СК и СД альтернативный канал для публикаций результатов их работ. Этот канал должен удовлетворять следующим основным требованиям:

- Доступность публикаций для научной читательской аудитории – любой, желающий ознакомиться с публикациями того или иного соискателя, должен иметь возможность без проблем это сделать;
- Централизованность – публикации диссертантов должны быть сгруппированы в одном месте для упрощения ознакомления с ними;
- Независимый отбор – статьи, размещаемые в этом издании, должны проходить независимую экспертизу. В качестве критериев отбора должны выступать новизна полученных результатов и отсутствие плагиата, а также соответствие требованиям ВАК к оформлению публикаций в рецензируемых журналах (наличие ключевых слов, аннотации на русском и английском языках);
- Оперативность – срок между подачей статьи и ее публикацией (в случае принятия) должен быть минимальным: не секрет, что подлинно интересные результаты у диссертантов появляются к концу срока обучения в аспирантуре. В традиционных журналах с их длительными сроками рассмотрения и публикации всегда есть опасность не успеть опубликоваться до защиты. В предлагаемом издании эта опасность должна быть устранена;
- Бесплатность – статьи диссертантов должны публиковаться бесплатно (в том числе и без скрытых платежей наподобие платы за рецензирование, подписку или авторский экземпляр), причем речь идет о диссертантах всех категорий, а не только об аспирантах, как в нынешнем постановлении ВАК;
- Самостоятельность публикаций – принимаются только статьи, написанные единолично диссертантом. С одной стороны, это позволит лучше учитывать вклад соискателя в написание своей диссертации (что не всегда просто, если в списке работ по теме диссертации есть только статьи, написанные в соавторстве, что не редкость в наше время). С другой стороны, это хотя бы отчасти поможет поставить барьер на пути порочной практики, когда список трудов научного руководителя раздувается за счет бегло просмотренных им статей, написанных его аспирантами, вынужденных указывать его в качестве соавтора.

Можно возразить, что в отдельных направлениях современной науки (в частности, в экспериментальной физике высоких энергий) сложилась такая практика, что в качестве авторов статей, рассказывающих о результатах проведенных экспериментов, указываются все люди, причастные к их проведению. Соответственно, требование указывать только одного автора будет нарушением авторских прав других участников исследования. Однако это возражение легко снимается: как правило, в экспериментальные команды входит и некоторое число ученых мирового уровня, которым не составляет труда опубликовать статью (с коллективным авторством, и в числе авторов будет указан и диссертант) в ведущих научных журналах мира. Сам же диссертант должен издать статью, повествующую о его личном вкладе – и такая статья вполне может иметь одного автора;

- Рейтинг статей – необходимо, чтобы публикации в этом издании могли быть проранжированы по их популярности у читателей, так как это позволило бы дополнительно оценить важность излагаемых в них результатов. Разумеется, при этом нужно разработать строгий критерий ранжирования.

Легко убедиться, что этим требованиям удовлетворяет только публикация статей в специально предназначенном для этого сетевом издании. С учетом того, что такое издание должно удовлетворять потребности диссертантов всех специальностей, разумнее говорить не об электронном журнале, а об электронном архиве (или электронной библиотеке). Примерным аналогом такого архива можно было бы считать международный архив препринтов по физико-математическим специальностям arXiv (arxiv.org).

Создание предлагаемой библиотеки могло бы происходить по следующей схеме:

1. Федеральное агентство по образованию (возможно, в партнерстве с ведущими вузами страны, прежде всего – МГУ и СПбГУ, так как именно эти вузы могут в будущем получить право выдавать собственные дипломы о высшем образовании и самостоятельно присуждать ученые степени) – далее будем называть их учредителями – разрабатывают концепцию такого архива (включая механизмы отбора статей, рейтингования и финансирования). В качестве соучредителей проекта могли бы также выступить РАН и государственные отраслевые академии, прежде всего – Российская академия образования;
2. Учредители заключают договор с кем-либо из крупнейших Интернет-компаний России о предоставлении дискового пространства для этого проекта. Идеальным кандидатом на роль партнера является, на наш взгляд, компания Yandex, так как это позволило бы существенно оптимизировать поиск по базе публикаций (при этом, однако, Yandex придется усовершенствовать схему своего семантического поиска, что, в свою очередь, требует формирования у данной компании соответствующей заинтересованности в этой работе и в проекте в целом);
3. Запускается пробная версия архива. Вероятно, на этом этапе будут публиковаться только работы диссертантов из вузов-учредителей;
4. Если проект на практике подтверждает свою жизнеспособность, то запускается его рабочая версия, доступ к публикации в которой получают все соискатели, обучающиеся в российских вузах.

Процесс отбора и публикации статей мог бы происходить по следующему алгоритму:

1. На начальной стадии проекта в нем в качестве рецензентов регистрируются все ученые и преподаватели из вузов-участников, имеющие право руководить СК и выступать научными консультантами у СД. Каждый из этих специалистов в обязательном порядке указывает свой часто проверяемый электронный почтовый ящик и получает уникальные логин и пароль для входа в систему рецензирования архива. Каждый из этих специалистов, далее, должен указать коды специальностей (по номенклатуре ВАК), по которым он имеет право выступать руководителем и/или консультантом, а также код специальности, по которой он защищал свою докторскую и/или кандидатскую диссертацию. Возможно, на следующем этапе в качестве таких рецензентов могут быть зарегистрированы бывшие СК и СД, успешно прошедшие защиту своих диссертаций, оставшиеся работать в системе науки и высшего образования, и чьи публикации в архиве отличаются высоким рейтингом;
2. Аналогичным образом в системе регистрируются все СК и СД, проходящие в настоящее время обучение. Помимо своих специальностей, при регистрации они также указывают коды других специальностей, являющихся смежными или представляющими для них интерес (пример – основная специальность 08.00.13 «Математические и инструментальные методы экономики», дополнительные специальности – 05.13.18 «Математическое моделирование» и 01.01.09 «Дискретная математика и математическая кибернетика»);
3. Таким же образом в системе регистрируются руководители и ученые секретари всех диссертационных советов, а также заведующие отделами аспирантуры и докторантуры – для получения отчетов о публикациях прикрепленных к ним СК и СД;
4. На сайте библиотеки публикуются требования к оформлению работ, размещаемых в архиве;
5. Диссертант, желающий опубликовать свою работу, приводит ее в соответствие с требованиями к оформлению и отсылает ее в архив через специальную контактную форму, размещенную на его сайте. В этой форме обязательно указываются ключевые слова, название вуза, к которому прикреплен соискатель, ФИО его научного руководителя (консультанта) и код специальности, список статей, уже опубликованных в электронном архиве, на которые он ссылается в своей работе. Возможно, во избежание публикации статей третьими лицами (как это практикуется при написании диссертаций «на заказ»), отправка статей может быть осуществлена только с компьютера вуза, к которому прикреплен диссертант;
6. Система подтверждает диссертанту получение его статьи и производит проверку присланного текста на плагиат. Некоторые российские издания уже производят такую автоматизированную проверку при помощи системы «Антиплагиат», например, «Вестник Российской академии государственной службы» (http://oad.rags.ru/vestnikrags/index.htm). Этот пункт предполагает интеграцию системы проверки на плагиат с системой электронного архива;

7. Если в тексте был выявлен плагиат, диссертанту сразу же отправляется письмо с уведомлением об отказе в публикации и указанием причины. Аналогичное письмо отправляется его научному руководителю и в соответствующий диссертационный совет. Данное письмо должно быть включено в личное дело диссертанта. В случае повторной присылки статьи с элементами плагиата диссертант исключается из аспирантуры/докторантуры. При этом необходимо разработать также процедуру апелляции – нельзя исключать возможность машинных ошибок;

8. Если в тексте плагиата выявлено не было, статья отправляется по электронной почте трем случайным рецензентам из числа зарегистрированных научных руководителей/консультантов, чьи коды специальностей соответствуют коду специальности статьи. Число рецензентов должно быть разным для статей, написанных СК и СД. При этом рецензенту не сообщается, на получение какой именно степени (докторской или кандидатской) претендует автор статьи. Среди этих рецензентов, разумеется, не должно быть научного руководителя автора. Рецензирование анонимное – автору и рецензентам не сообщаются ФИО и названия вузов друг друга. Вероятно, имеет смысл предусмотреть возможность указания автором при отправке статьи ФИО рецензентов, к которым статью направлять не следует (необходимо для того, чтобы представители соперничающих научных школ не могли блокировать публикации противоположной стороны). Рецензентам могут быть присвоены разные веса в зависимости от их ученой степени и звания, членства в академиях наук, рейтинга вуза или НИИ, в которых они работают, и т. д.;

9. Эти рецензенты должны в течение двух недель оценить соответствие статьи определенным критериям (полнота изложения материала, стиль изложения, наличие научной новизны и т. д.) по десятибалльной шкале (оценивается не работа в целом, а каждый критерий в отдельности). Для выставления оценок на сайте электронной библиотеки предусматривается специальная форма. В этой форме также может быть специальное поле, в которой рецензент по желанию может указать наиболее существенные недочеты статьи;

10. После поступления оценок работы в систему электронных публикаций в ней автоматически выставляется итоговая оценка статьи по следующей формуле:

$$ИОС = \frac{\sum_{j=1}^{m} \frac{W_j \sum_{i=1}^{n} w_i Q_{ij}}{\sum_{i=1}^{n} w_i}}{\sum_{j=1}^{m} W_j}, \qquad (1)$$

*ИОС* – итоговая оценка статьи;

*n* – число критериев, по которым оценивается статья;

*m* – число рецензентов, которым статья была отправлена на рассмотрение;

$w_i$ – вес *i*-го критерия в совокупной оценке;

$W_j$ – вес *j*-го рецензента;

$Q_{ij}$ – оценка соответствия статья *i*-му критерию, выставленная *j*-ым рецензентом;

11. Тексты рецензий сохраняются в специальном закрытом отделе электронного архива и могут быть предоставлены по запросу ВАК;

12. Статья принимается к публикации в системе, если ее *ИОС* больше или равна некоторому минимальному значению. Это минимальное значение должно быть разным для «докторских» и «кандидатских» статей (для статей, написанных СД, $ИОС_{\min}$ должно быть выше);

13. К одному и тому же рецензенту нельзя обращаться более пяти раз в течение календарного года. При этом он обязан отрецензировать не менее трех присланных ему статей. Если он отрецензировал менее трех статей (из присланных пяти), статьи руководимых им СК и СД к публикации в архиве приниматься не будут;

14. В случае необходимости статья отправляется диссертанту на доработку. В течение недели он обязан либо прислать исправленную статью, либо отказаться от ее публикации. После этого она отправляется на повторное рассмотрение одному из рецензентов. В случае его одобрения (оценка по десятибалльной шкале от 5 и выше без повторной подробной рецензии) статья немедленно публикуется в архиве;

15. После публикации статьи информацию о ее размещении получают все соискатели, указавшие код соответствующей специальности в списке своих основной и дополнительной специальности. Такую же информацию получат все научные руководители по этой специальности;

16. На странице архива, на которой публикуется статья, размещается форма для отзывов. Эти отзывы (если они поступят) в последующем можно будет прочесть под текстом самой статьи. При этом будут существовать строгие правила составления текстов отзывов. За соблюдением правил будут следить модераторы (возможно, текст отзыва будет публиковаться только после его предварительного прочтения модератором);

17. Лица, желающие разместить свой отзыв о статье, должны будут указать в соответствующей форме код своей специальности, название вуза и ФИО. Это повысит ответственность автора отзыва и вынудит его с особым тщанием отнестись к подготовке текста;

18. Автор статьи получает на свой электронный почтовый ящик уведомление о поступившем отзыве и, при желании, может вступить с автором отзыва в дискуссию. При защите диссертации распечатки таких дискуссий

могут быть приложены к личному делу СК (СД) как подтверждение апробации его результатов. Как представляется, такая форма апробации является гораздо более эффективной, чем научные конференции;

19. Лица, допустившие неэтичные формулировки в текстах отзывов, на определенный срок лишаются права публиковать свои статьи в электронном архиве;

20. Система обеспечивает поиск публикуемых в ней работ по коду специальности, ФИО автора, ФИО научного руководителя, вузу и ключевым словам;

21. Перед защитой соискатель по специальному запросу получает от архива индекс цитирования своих статей (внутри системы) и информацию о количестве отзывов о них. Эти величины должны быть выше некоторого определенного порога (разного для разных специальностей – так как у разных научных направлений разная популярность, и статьи по ним будут интересны разному числу исследователей; соответственно число отзывов будет различаться). Введение представления о разных численных критериях для разных специальностей представляется исключительно важным, так как именно отсутствие таких различий является одним из основных поводов для критики наиболее известного в настоящее время индекса цитирования, рассчитываемого Институтом научной информации (г. Чикаго, США). Если число отзывов ниже этой пороговой величины, то соискатель не допускается до защиты, так как это означает, что результаты его исследований не вызвали интереса у коллег, а сама работа не соответствует критерию актуальности.

Этот этап имеет большое значение для оценки качества диссертационных исследований, так как в настоящее время формальные (имеющие количественное измерение) показатели актуальности отсутствуют. Предлагаемая модель позволяет дать количественную оценку актуальности работы.

Расчет этих пороговых значений может производиться по следующей формуле:

$$Q = a_0 N_{spec} + b_0 n_{spec} + \sum_{i=1}^{m} (a_i N_i + b_i n_i), \quad (2)$$

$Q$ – минимальное требуемое число отзывов на все статьи, опубликованные СК или СД в электронном архиве;

$N_{spec}$ – среднее число СК и СД по данной специальности по стране за нормативный срок обучения в аспирантуре (докторантуре);

$n_{spec}$ – среднее число научных руководителей по данной специальности по стране за нормативный срок обучения в аспирантуре (докторантуре);

$N_i$ – среднее число СК и СД по смежным специальностям по стране за нормативный срок обучения в аспирантуре (докторантуре);

$n_i$ – среднее число научных руководителей по смежным специальностям по стране за нормативный срок обучения в аспирантуре (докторантуре);

$m$ – число смежных специальностей;

$a_0$, $b_0$, $a_i$, $b_i$ – поправочные множители. Могут быть как одинаковыми для всех специальностей, так и различаться в зависимости от специальности, формы и срока обучения.

По аналогичной формуле будет рассчитываться минимально необходимое число ссылок на статьи диссертанта, с той разницей, что в этом случае из формулы будут исключены члены, учитывающие количество научных руководителей (так как в архиве публикуются только статьи диссертантов, и число цитирований учитывается только в рамках системы).

Для этих целей необходимо обеспечить справкам по индексу цитирования, предоставляемым архивом, необходимый правовой статус;

22. Во избежание чрезмерных затрат на поддержание архива, возможно, имело бы смысл ограничить объем одной публикации и число публикаций за время обучения в аспирантуре/докторантуре. На наш взгляд, для СК может быть достаточно пяти публикаций, для СД – пятнадцати. При этом СК обязан опубликовать не менее трех статей, а СД – не менее десяти. Публикация большего числа статей (сверх 5 и 15 соответственно) возможна при условии оплаты. Взимание этой платы нарушением требования бесплатности не будет, так как диссертанту предоставляется право бесплатной публикации определенного числа статей.

Предложенная схема способна обеспечить высокую прозрачность процесса подготовки специалистов высшей квалификации и снизить число т. н. «заказных» диссертаций.

Финансирование проекта может осуществляться за счет:

- бюджетных субсидий;
- отчислений от вузов, имеющих в своем составе аспирантуру и докторантуру. Размер отчислений зависит от количества фактически обучающихся СК и СД;
- добровольных пожертвований;
- грантов;
- платы за публикацию статей сверх предусмотренной квоты;
- размещения рекламы (при этом размер рекламных материалов не должен превышать определенной части от размера страницы, и рекламные материалы не должны мешать знакомиться с научными публикациями – в частности, не быть «всплывающими»). Партнерство с компанией Yandex в этом случае было бы идеальным, так как она могла бы предложить контекстную рекламу в рамках поиска по электронному архиву (ссылки на магазины, где

можно было бы купить соответствующую научную литературу, платные курсы и т. д.).

## 3 Публикация материалов конференций, авторефератов и монографий

Как уже отмечалось выше, материалы большей части конференций остаются недоступными широкой публике. В этой связи было бы логично принять требование о том, что конференция может претендовать на статус международной или всероссийской только в том случае, если у нее есть собственный сайт, на котором размещены тезисы и/или полные тексты всех прозвучавших на ней докладов. Было бы разумно, если бы одновременно эти сборники тезисов также размещались в специальном разделе предлагаемого электронного архива (а до проведения конференций на сайте публиковались бы их электронные письма). Это размещение должно происходить за определенную плату, включаемую во взнос, собираемый с участников происходящих в России международных и всероссийских конференций.

Помимо размещения авторефератов на сайте диссовета и (в случае СД) – на сайте ВАК, целесообразным представляется также их публикация в особом разделе предложенного выше электронного архива (вместе с извещением о защите). Таким образом, любой исследователь или преподаватель, желающий ознакомиться с авторефератом и прислать отзыв на него, сможет найти все авторефераты по интересующей его специальности в одном месте (а не сайтах разных вузов) и сразу же ознакомиться с основными публикациями соискателя по теме диссертации (которые размещены в том же электронном архиве). Это также позволило бы проверять тексты авторефератов на плагиат (поскольку, как уже предлагалось выше, модуль проверки на заимствования должен быть включен в систему электронного архива). В случае, если плагиат будет обнаружен, уведомления об этом поступают диссертанту, его научному руководителю и в диссовет, а диссертация снимается с защиты.

Допустим вариант взимания умеренной платы за публикацию автореферата на сайте архива (по аналогии с оплатой объявления о защите докторской диссертации).

После размещения автореферата на сайте электронного архива система автоматически рассылает уведомления о нем всем кандидатам и докторам наук по тем специальностям, которые диссертант при регистрации в электронном архиве указал в качестве основной и дополнительных. Ученый, пожелавший отправить отзыв, заполняет специальную форму на сайте архива, после чего текст отзыва поступает в диссертационный совет, в котором будет происходить защита.

В настоящий момент авторефераты и полные тексты диссертаций хранятся в электронном депозитарии РГБ, однако доступ к ним ограничен из-за того, что ознакомиться с ними можно либо непосредственно в самой РГБ, либо в ее виртуальных читальных залах, число которых в нашей стране крайне мало.

Можно было бы предложить модель подбора официальных оппонентов, совпадающую с моделью выбора рецензентов для статей в электронном архиве (то есть случайное определение из числа докторов и кандидатов наук по данной специальности из числа зарегистрированных в архиве). Такая модель, безусловно, позволила бы повысить качество оценки диссертационных работ (хотя бы по той причине, что в настоящее время предполагаемый оппонент до своего назначения имеет неофициальную возможность ознакомиться с работой диссертанта и отказаться от оппонирования до своего назначения, если качество диссертации его не устраивает; в предлагаемой модели отказ ученого от оппонирования будет зафиксирован в системе электронного архива и может послужить основанием для настороженности). Однако пока такой шаг представляется преждевременным: официальный оппонент должен ознакомиться не с авторефератом, а с полным текстом диссертации, и подавляющая часть исследователей не захочет высылать текст своей неопубликованной работы в электронном виде другим ученым.

Кроме того, электронный архив помог бы решить проблему с публикацией монографий, де-факто считающихся необходимым условием для защиты докторской диссертации. Не секрет, что монографии (за исключением написанных достаточно ограниченным кругом ведущих ученых) представляют собой коммерчески невыгодные проекты, и научные издательства соглашаются их публиковать только при условии полной или частичной компенсации затрат на издание. Авторы вынуждены либо самостоятельно изыскивать средства на финансирование публикации своей монографии, либо находить способы бесплатно их публиковать (например, пользуясь связями в РИО своих или дружественных вузов). Сами монографии издаются мизерными тиражами, не доходят в большинстве своем до потенциальной читательской аудитории и зачастую нужны только «для галочки» в личном деле СД.

Для экономии средств (как авторских, так и бюджетных) можно было бы применить депонирование монографий, но депонирование в научной среде имеет крайне низкий престиж и депонированные рукописи практически никогда не находят читателя.

Решением проблемы стала бы, на наш взгляд, замена депонирования и публикации печатной монографии размещением ее текста в специально для этого предназначенном разделе предлагаемого электронного архива. Это гарантировало бы донесение результатов этой монографии до заинтересованных читателей. Безусловно,

публикация монографий в архиве должна происходить на платной основе, но сама плата в любом случае будет ниже затрат на издание печатной версии работы. Разумеется, это решение требует уравнивания в правах электронных и печатных монографий.

Однако внесение платы за публикацию монографии не должно быть единственным условием для ее публикации в электронном архиве. Она должна пройти процедуру рецензирования, аналогичную предлагаемой для статей СД и СК, и может быть принята к публикации только в том случае, если ее итоговая оценка (рассчитываемая по формуле (1)) будет выше определенного порогового значения.

В соответствии с новыми решениями ВАК РФ, монография в отдельных случаях может быть засчитана в качестве «ваковской» публикации (т. е. вместо статьи в журнале из «списка ВАК»), однако критерии, по которым будет определяться, будет ли данная монография иметь «ваковский» статус, на сегодняшний день отсутствуют. На наш взгляд, таким критерием мог бы стать индекс цитирования данной монографии (определяемый по формуле (2)) – в том случае, если число ссылок на монографии меньше определенной величины (на момент представления диссертации к защите), то данная монография в качестве «ваковской» публикации засчитана быть не может.

## 5 Недостатки предлагаемой методики

К недостаткам предлагаемой модели следует отнести отсутствие решения следующих проблем:

1. Направленность на «формальных» соискателей, т. е. людей, проходящих курс обучения в аспирантуре (сразу после окончания вуза и при отсутствии значимых научных результатов) или докторантуре (по достижении определенной ступени в карьерном росте по научно-педагогической иерархической лестнице, но также при отсутствии существенных научных достижений). Очевидно, что именно этим категориям диссертантов особенно важно соответствие формальным критериям, в частности – требованию наличия обязательных публикаций. Те же ученые, которые уже добились важных результатов, но еще не получили степень, окажутся вынуждены в дополнение к своим публикациям в ведущих мировых журналах также размещать свои работы в предлагаемом электронном архиве (т. е., по сути, дополнительно доказывать уже очевидную для научного сообщества значимость своих результатов). Однако, на мой взгляд, этот недостаток не является существенным, так как пройти процедуру публикации в электронном архиве такому ученому будет легко, и не будет ничего плохого в том, что основные полученные им результаты окажутся опубликованы в доступном для российского научного сообщества электронном архиве, а не только в зарубежном журнале, подписка на который у большинства вузов и НИИ просто отсутствует;

2. Возможность обхода предлагаемой процедуры публикации на основе неформальных договоренностей (как это случилось, например, с результатами ЕГЭ). К сожалению, эта проблема неустранима, так как не существует схем контроля качества, которые нельзя было бы обойти, что, однако, не должно служить поводом для отказа от самой идеи использования схемы контроля;

3. Недопуск до защиты лиц с низким индексом цитирования в архиве – хорошо известно, что величина индекса цитирования зависит от отрасли науки, и от конкретной тематики, над которой работает исследователь. Если диссертант выбрал для своего исследования узкоспециализированную тему, над которой работает малое число ученых, то есть вероятность того, что индекс цитирования окажется мал по объективным причинам, и диссертант, даже получивший интересные и значимые результаты, до защиты допущен не будет. Однако эта опасность преувеличена: методика расчета индекса цитирования учитывает различия в количестве ученых по каждой специальности, а алгоритм работы электронного архива с обязательной рассылкой уведомлений о публикации статей ученым, работающих по той же или близкой специальности гарантирует, что все заинтересованные лица узнают о данной статье. В таких условиях отсутствие ссылок на статью будет означать недостаточную степень актуальности диссертационного исследование, т. е. его несоответствие одному из основных требований ВАК, в силу чего такую работу целесообразно не допускать до защиты.

## 6 Заключение

Более широкое использование возможностей Интернета с присущими ему сравнительно низкими издержками на публикацию научных работ, обеспечением свободного доступа к опубликованным материалом и возможностью открытого обсуждения достигнутых результатов позволило бы существенно повысить прозрачность процедуры подготовки диссертации и поставить барьер на пути написания диссертаций «на заказ».

Предлагаемый электронный архив статей СД и СК мог бы стать эффективным и общедоступным альтернативным каналом публикации результатов диссертационных исследований (в том числе и благодаря предусмотренной в нем системе оповещения о новых статьях и модуле проверки на плагиат). В случае же введения обязательной публикации на сайте архива авторефератов диссертаций (с предоставлением отзывов на них также через систему архива в электронном виде), размещения материалов международных и всероссийских конференций, а также информации о журналах из «списка ВАК» этот электронный архив может стать основой для формирования

информационного подпространства поддержки диссертационных исследований, выполняющего следующие функции:

- контрольная – обеспечение высокого качества диссертационных работ (путем проверки на плагиат, рецензирования связанных с ними статей и обсуждения этих статей научным сообществом);
- инновационная – канал для публикации статей, материалов конференций и авторефератов;
- информационная – публикация сведений о рецензируемых журналах, перечня диссертационных советов с их контактными координатами, пособия по подготовке диссертационных работ, перечни вопросов экзаменов кандидатского минимума и открытые пособия по подготовке к ним и т. д.;
- социальная – взаимодействие между членами научного сообщества (то есть создание на основе электронного архива научной социальной сети). Для начала каждый диссертант мог бы иметь возможность создать в системе свою собственную страничку с личной и рабочей информацией о себе.

В частности, такое пространство могло бы позволить решить следующую очень важную проблему. Должности профессорско-преподавательского состава замещаются на конкурсной основе, что теоретически должно способствовать отбору лучших кандидатов. Эти конкурсы формально являются открытыми, то есть объявления о них публикуются в средствах массовой информации, и заявление на участие в них может подавать любой научно-педагогический сотрудник соответствующей квалификации, ознакомившийся с этими объявлениями.

Однако очевидно, что в условиях низкой оплаты труда преподавателей вуз старается обеспечить им гарантию занятости и после истечения срока избрания по конкурсу. С другой стороны, вуз не всегда заинтересован в приходе преподавателей со стороны, так как потребуется некоторое время на их встраивание в корпоративную культуру соответствующего учебного заведения. Иными словами, вузы пытаются превратить конкурсный отбор в выборы единственного кандидата – как правило, человека, уже работающего на данной должности или же сотрудника, «под которого» эта должность была создана (обычно уже в момент конкурса хорошо известно, кто станет его победителем).

Эта схема порождает у преподавателей ощущение собственной несменяемости, препятствует академической мобильности, и – что еще более важно – превращает институт конкурса в профанацию, что неприемлемо.

Важным – и, возможно, ключевым, - элементом этой схемы является формальный подход к публикации объявлений о конкурсе. Как правило, их размещают в таких источниках, доступ к которым посторонних лиц невозможен или затруднен – например, в собственных вузовских изданиях (которые не распространяются за пределами соответствующего учебного заведения), или на сайте вуза (сторонний преподаватель, заинтересованный в работе в данном вузе, должен непрерывно отслеживать его сайт, что затруднительно), или даже просто вывешивается на доску объявлений внутри вуза. Это является надежной гарантией того, что информацию о конкурсе получат только свои.

В качестве средства противодействия такому формальному отношению предлагается создать федеральный веб-сайт, на котором в обязательном порядке должны публиковаться объявления о конкурсах на замещение профессорско-преподавательского состава (как минимум, начиная от должности доцента). Конкурс, который проводится без объявления на этом сайте, признается недействительным.

До даты проведения конкурса на сайте должны быть опубликованы досье всех соискателей, подавших заявления на участие в нем. Это позволит сторонним лицам ознакомиться с этими личными делами и прислать свои отзывы. В свою очередь, эти отзывы дадут возможность ученому совету вуза, проводящему конкурс, более объективно отнестись к кандидатам и не отдавать свое безоговорочное предпочтение собственным сотрудникам.

Превращение конкурса в открытую процедуру повысит мотивацию преподавателей к повышению собственной квалификации, разработке оригинальных курсов, активному участию в научной работе и т. д., что, в конечном счете, должно привести к общему повышению качества образования. Разумеется, эта мера должна сопровождаться соответствующим улучшением условий оплаты труда преподавателей.

Этот сайт было бы разумно интегрировать с предлагаемым электронным архивом публикаций аспирантов и докторантов.

До тех пор, пока такой сайт не будет запущен, было бы целесообразно в качестве временной меры размещать объявления о конкурсах на замещение должности профессора на сайте ВАК.

Потребность в таком информационном подпространстве настолько велика, что научное сообщество стихийно пытается формировать его (в качестве примера можно привести портал Phido.ru, для которого автор данной статьи подготовил перечень контактных координат журналов из «списка ВАК»). Тем не менее, без наличия государственной поддержки такие порталы будут всего лишь паллиативом и не смогут эффективно справляться с выполнением контрольной и информационной функций.

## Литература

[1] Информационное сообщение о порядке формирования Перечня ведущих рецензируемых научных журналов и

изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени доктора и кандидата наук (от 14.10.2008). http://vak.ed.gov.ru/ru/list/infletter-14-10-2008.

[2] Котляров И. Д. Сетевая публикация результатов диссертационных исследований./Информационные технологии, №7, 2009, стр. 69-77.

[3] Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные научные результаты диссертации на соискание ученой степени доктора и кандидата наук (редакция апрель 2008 года). http://vak.ed.gov.ru/common/img/uploaded/VAK/files_help_desk/per-04-2008.doc.

[4] Положение о порядке присуждения ученых степеней. Утверждено постановлением Правительства Российской Федерации №74 от 30.01.2002. http://vak.ed.gov.ru/ru/docs/?id4=212&i4=24.

[5] Решение президиума об исключении журнала из Перечня изданий (от 17.10.2008). http://vak.ed.gov.ru/ru/news/allnews/index.php?id4=1163.

[7] Российский научный журнал поймали на публикации заведомой ерунды//Грани.ру, 30.09.2008. http://grani.ru/Society/Science/m.142082.html.

## Electronic library for publication of results of Ph.D. and post-doctoral theses

Ivan Kotliarov

An electronic library of publications of results obtained by Ph.D. and post-doctoral students can be a good tool of control of scientific quality of doctoral and post-doctoral theses. A list of requirements this e-library should meet is proposed. An algorithm of this archive is described. Additional options this library should have are proposed.

# ОНТОЛОГИЧЕСКОЕ МОДЕЛИРОВАНИЕ–1

# ONTOLOGICAL MODELING–1

# Формальное представление метаинформации для некоторых подходов к согласованию онтологий♣

© Н. А. Скворцов

Институт проблем информатики РАН
nskv@ipi.ac.ru

**Аннотация**

Подходы, используемые средствами поддержки согласования онтологий, во многом основываются на методах, разработанных для интеграции схем баз данных. Однако этого недостаточно, так как онтологии определяют понятия, семантика которых может быть сходной при различных подходах к описанию их структуры. В [6] нами были описаны подходы для обнаружения сходств и различий семантики понятий при согласовании онтологий. В качестве продолжения этого исследования ставится задача разработать для этих подходов представление метаинформации в модели, определяемой языком Синтез, которая используется для концептуального моделирования предметных областей и в качестве канонической модели для интеграции неоднородных информационных ресурсов в спецификациях научных задач. Настоящая статья описывает, как могут использоваться метаонтологии, онтологии верхнего уровня, фундаментальные виды метасвойств понятий, экземпляры классов понятий для семантического согласования онтологий, и представляет спецификации этой метаинформации в модели, определяемой языком Синтез.

## 1 Введение

Онтология призвана задавать семантику понятий предметной области, она определяет онтологический контекст, в котором работает некоторое сообщество. Подходы к описанию структуры одних и тех же понятий в контекстах, описанных разными онтологиями, могут быть различными, что влияет на состав структурных спецификаций, их ограничения и степень детализации понятий.

**Труды 11ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

При взаимодействии сообществ, работающих в разных онтологических контекстах, возникают задачи согласования онтологий [3]. К таким задачам относятся отображение одной онтологии в другую, интеграция одной онтологии в другую как части, слияние онтологий для получения новой и другие. Основой решения этих задач является построение отображений понятий одной онтологии в понятия другой онтологии.

Информационные модели, используемые сегодня в качестве онтологических, либо неформальны, либо включают достаточно простые средства спецификации для возможности использования автоматического вывода, довольствуясь описаниями структурных спецификаций понятий и простых ограничений над ними. Поэтому большинство методов, используемых для согласования онтологий, предварительно связывают понятия по вербальной информации, и затем на основе полученных связей оперируют со структурными спецификациями, оценивая их близость и обнаруживая и устраняя разного рода конфликты. Результатом этого стало то, что для согласования онтологий применяются методы, уже наработанные в области интеграции концептуальных схем баз данных. Недавний обзор [5] методов отображения онтологий выявил именно такую тенденцию, что послужило поводом для настоящего исследования.

Очевидно, что семантика онтологической информации специфична и не ограничивается спецификацией схемы. Структурные спецификации онтологических понятий отражают не структуры для абстрактного представления информации, которую необходимо хранить и обрабатывать, а собственные свойства понятий и связи их между собой, образуя систему, вне которой каждое отдельно взятое понятие не может существовать.

Методы связывания по именам и вербальным спецификациям, методы структурной идентификации релевантных спецификаций и разрешения структурных конфликтов применимы как к схемам, так и к онтологиям. Однако важно, чтобы при таком согласовании гарантировалось сохранение семантики понятий. Кроме этого, необходимы механизмы, позволяющие находить сходства и различия в семантике понятий помимо той, которая определяется их структурными спецификациями.

В [6] нами был обоснован выбор для дальнейших исследований нескольких подходов к семантическому согласованию онтологий. Выбор исследуемых подходов производился не на основании качества применяемых методов анализа имён или структуры понятий, хотя такие подходы должны использоваться для предварительного связывания понятий. Основными критериями при выборе исследуемых подходов были формальность, то есть возможность формального вывода, и сохранение семантики понятий при отображении.

В настоящем исследовании ставится задача представления метаинформации, необходимой для онтологических спецификаций, используемых в выбранных подходах. В разделе 2 кратко объяснены исследуемые подходы к описанию семантики онтологических понятий и основанные на них подходы к согласованию онтологий. Раздел 3 посвящён описанию модели, определяемой языком Синтез [2], в которой описывается метаинформация для представления онтологических спецификаций. Последующие разделы описывают представление в данной модели метаинформации, используемой в конкретных подходах. Последний раздел 7 возвращается к схемам, говорит об особенностях концептуального моделировании предметных областей, требованиях к спецификациям концептуальных схем предметных областей и применимости подходов, описанных в статье, для их согласования.

## 2 Подходы к спецификации понятий и к согласованию онтологий

Если предположить, что онтологии достаточно точно и полно отражают семантику понятий предметной области, для проверки сохранения семантики описаний при отображении онтологических понятий может использоваться отношение уточнения спецификаций. Для рассуждения в терминах уточнения используются формальные методы, поэтому утверждение об уточнении можно доказывать. В разделе 3 описывается представление онтологических спецификаций в объектной модели языка Синтез, для которой разработано формальное отображение в средства доказательства уточнения [8].

Целью структурной спецификации понятий является определение их семантики через связи с другими понятиями. Однако подходы к определению семантики понятий не должны заканчиваться на этом. Ниже исследуются подходы к согласованию онтологий, независимые от сходств и различий в описаниях структуры онтологических понятий, но рассматривающие семантику понятий с других ракурсов. В качестве таких подходов предложено [6] использование общих метаонтологий и фундаментальных метасвойств (раздел 5), характеризующих понятия онтологий, а также анализ объектов, являющихся экземплярами понятий (раздел 6).

Семантика понятий, помимо структурных спецификаций, может определяться с помощью выражения их смысла в терминах другой более абстрактной модели. Одним из подходов описания теоретико-модельной семантики понятий является сопровождение спецификаций понятий онтологии описанием их в терминах метаонтологии. При согласовании онтологий противоречивость спецификаций согласовываемых понятий в терминах общей метаонтологии будет говорить об их разной семантике. И напротив, эквивалентность описаний понятий разных онтологий в терминах метаонтологии может означать близость этих понятий. Подробнее этот подход обсуждается в разделе 4.

Под фундаментальными метасвойствами понимаются такие свойства понятий по отношению к соответствующим им сущностям реального мира, как неотъемлемость, независимость, идентификация, различные обязательства сущностей, соответствующих понятиям, друг к другу. Подобными свойствами понятий, в свою очередь, определяются разновидности понятий – родовидовые, роль, категория и другие. В большинстве онтологических моделей средства для спецификации подобных особенностей понятий отсутствуют, хотя они немаловажны для определения семантики понятий. Их спецификация рассматривается в разделе 5. При согласовании онтологий фундаментальные свойства понятий могут помочь обнаружить разницу в их понимании, проверить корректность разного рода связей, установленных между онтологическими понятиями.

Анализ экземпляров при согласовании онтологий также важен. Ему посвящён раздел 6. Нахождение хотя бы одного экземпляра, который принадлежит экстенсионалу одного понятия, но не принадлежит эквивалентному понятию в другой онтологии, заставляет усомниться в корректности связи между понятиями. На анализе отношений экстенсионалов основан денотационный подход к определению семантики понятий, анализ соответствия экземпляров понятиям является его разновидностью. И хотя изначально онтологии не могут определяться спецификациями принадлежности понятиям каждого экземпляра в отдельности, для согласования онтологий применение конкретных экземпляров к согласовываемым понятиям часто выявляет скрытые различия в понятиях.

Все описанные подходы требуют дополнительных спецификаций к обычно имеющимся вербальным и структурным спецификациям онтологий, однако несут и дополнительные знания о согласовываемых онтологиях, помогающие разобраться в семантике понятий. Полезно сопровождать онтологии такими спецификациями уже при разработке онтологий.

## 3 Представление онтологий средствами абстрактных типов языка Синтез

Прежде чем можно будет начать действия по согласованию онтологий, их онтологические модели должны быть приведены к одной унифицированной модели. В качестве ядра для создания унифицированной модели мы используем расширяемый объектный язык спецификаций Синтез [2]. В основе языка используется язык фреймов, над которым построена объектная модель.

Любая спецификация в языке Синтез является фреймом с возможностью задавать его идентификатор, его слоты, значения слотов, а также метафреймы, метаслоты и метазначения. Семантика конкретных фреймов определяется принадлежностью фрейма метаклассам, перечисленным в специальном слоте `in`. В частности, метакласс `module` содержит фреймы, в которых определяются модули языка Синтез. Метакласс `type` содержит абстрактные типы данных, а метакласс `concept` – все онтологические понятия. В определениях понятий метакласс `type` можно опускать, так как указываемый метакласс `concept` является его подклассом.

Объектная модель языка позволяет выстраивать решётку типов, основанную на отношении тип/подтип. Кроме того, она определяет ещё одну иерархию, ортогональную первой. Это иерархия классификации, которая основана на отношении класс/экземпляр. На нулевом уровне классификации находятся экземпляры объектов типов, на первом – абстрактные типы данных, на втором – метатипы, то есть классы, экземплярами которых являются типы. И так далее. Эти две иерархии будут использованы в статье как средства выражения необходимой нам метаинформации.

В представленном ниже примере на языке Синтез определяются простые понятия звёздной астрономии:

```
{ StellarOntology;
  in: module;
  kind: ontology;

  type:

  { Star;
    in: type, concept;
    luminosityClass: LuminosityClass;
      metaslot
        inverse: LuminosityClass.ofStar
      end
  },
  { MainSequenceStar;
    in: type, concept;
    supertype: Star;
    luminosityInv:
    {in: predicate, invariant;
    { predicative; {
      all m/MainSequenceStar
      (m.luminosityClass = MainSequence)
    }}}
  },
  { LuminosityClass;
    in: type, concept;
    ofStar: Star;
      metaslot
        inverse: Star.luminosityClass;
      end
  };

  class_specification:

  { star;
    in: metatype;
    instance_type: Star
  },
  { mainSequenceStar;
    in: metatype;
    superclass: star;
    instance_type: Star;
  },
  { luminosityClass;
    in: metatype;
    instance_type: LuminosityClass;
  };
}
```

Онтология определяет понятия «Звезда», «Звезда главной последовательности» и «Класс светимости». Онтологические понятия определяются в онтологическом модуле (`StellarOntology`). Для спецификации интенсиональных свойств понятий используются абстрактные типы данных (`Star`, `MainSequenceStar`, `LuminosityClass`). Вербальные спецификации и подробные спецификации внутренней структуры типов понятий мы опускаем и не затрагиваем в статье. Экстенсиональные спецификации также являются частью спецификации онтологии, они описывают связанные с понятиями классы (`star`, `mainSequenceStar`, `luminosityClass`) как множества объектов-экземпляров понятий. Тип понятия является типом экземпляров данного класса. Любой фрейм, соответствующий данному типу, может стать экземпляром класса понятия. Например, «Главная последовательность» - конкретный класс светимости, и для неё определён фрейм (`MainSequence`), являющийся экземпляром понятия «Класс светимости».

```
{ MainSequence;
  in: frame, luminosityClass;
}
```

Формальный подход к отображению онтологических понятий разных онтологий в модели, определяемой языком Синтез, использует концепцию уточнения абстрактных типов данных. Установленное между абстрактными типами данных, отношение уточнения означает, что значение уточняющего типа гарантированно можно использовать вместо значения уточняемого типа, не замечая подмены. Данное отношение для абстрактных типов данных языка Синтез

определяется формально, поэтому утверждение об уточнении типов можно доказывать [9]. Спецификация отображения одного понятия в другое определяется в виде конкретизирующего типа, разрешающего конфликты типов, определяющих данные понятия. Корректность отображения обосновывается доказательством отношения уточнения между понятием и конкретизирующим типом. Разрешение конфликтов и доказательство уточнения понятий не входит в задачи статьи.

На практике для обоснования отношения уточнения абстрактных типов данных спецификации однозначно отображаются [8] в нотацию абстрактных машин и доказываются в специализированных средствах доказательства уточнения, использующих абстрактные машины в качестве входного языка. Важно отметить, что семантика объектной модели языка Синтез, заданная ее отображением в формальную модель, позволяет считать спецификации на языке Синтез формальными. Доказанное отношение уточнения понятий гарантирует корректность отображения онтологий. Однако при этом надо быть уверенным, что спецификации онтологических понятий изначально точно отражают их семантику.

## 4 Представление метаонтологий и онтологий верхнего уровня

Согласование онтологий производится, когда их предметные области, по меньшей мере, пересекаются. То есть, некоторые части согласовываемых онтологий имеют общую природу, общие понятия. Для согласования таких онтологий имеет смысл описывать их дополнительно в терминах некоторой общей онтологии более абстрактного уровня. Дополнительная спецификация в терминах более абстрактных понятий позволит обнаруживать родственные понятия в онтологиях и конфликты семантики близких понятий.

В модели, определяемой языком Синтез, для связывания согласовываемых онтологий с понятиями более абстрактной общей онтологии в зависимости от природы последней могут быть использованы два вида связи:

- отношение тип/подтип;
- отношение класс/экземпляр.

В случае использования отношения тип/подтип онтология верхнего уровня, содержащая понятия наиболее общего назначения, становится общим основанием согласовываемых онтологий. Понятия онтологий выстраиваются в продолжение иерархии понятий онтологии верхнего уровня. Данный подход требует серьёзных усилий по согласованию иерархий понятий рассматриваемых онтологий с онтологией верхнего уровня. Ниже показан пример спецификации понятия «Звезда» (`Star`), которое является подпонятием понятия онтологии верхнего уровня «Астрономический объект» (`AstronomicalObject`).

```
{ AstronomicalObject;
  in: concept;
}
…
{ astronomicalObject;
  in: metatype;
  instance_type: AstronomicalObject;
}


{ Star;
  in: concept;
  supertype: AstronomicalObject;
}
…
{ star;
  in: metatype;
  superclass: astronomicalObject;
  instance_type: Star;
}
```

Отношение класс/экземпляр используется для спецификации понятий онтологий дополнительно понятиями метаонтологии. Метаонтологией для данной онтологии является онтология, которая содержит обобщённое описание метамодели [7], на основе которой можно построить большинство онтологических моделей, либо онтология, описывающая наиболее общие категории понятий с общими свойствами, характерные для рассматриваемой предметной области [4].

Понятия онтологий, а также, при необходимости, отношения и атрибуты понятий становятся экземплярами понятий метаонтологии. Такой принцип построения спецификаций позволяет сделать независимыми друг от друга спецификации в терминах метаонтологии и собственно спецификации рассматриваемых онтологий, так как эти спецификации находятся на разных уровнях иерархии классификации, которая ортогональна иерархии абстрактных типов данных. Соответственно, благодаря независимости спецификаций, нет и ограничений на одновременное использование нескольких метаонтологий, рассматривающих онтологии с разных ракурсов предметной области. Ниже описано понятие «Класс светимости» (`LuminosityClass`), являющееся экземпляром понятия метаонтологии «Параметр измерения» (`MeasurementParameter`).

```
{ MeasurementParameter;
  in: concept;
}
…
{ measurementParameter;
  in: metatype;
  instance_type: MeasurementParameter;
}


{ LuminosityClass;
  in: concept, measurementParameter;
```

```
}
…
{ luminosityClass;
  in: metatype;
  instance_type: LuminosityClass;
}
```

Чтобы подчеркнуть независимость спецификации понятий в онтологии от спецификации их в терминах метаонтологии, последнюю можно убирать в метафрейм как дополнительную метаинформацию о понятиях.

```
{ LuminosityClass;
    metaframe
      in: measurementParameter;
    end
  in: concept;
}
```

Если подходящего понятия в метаонтологии нет, то элемент онтологии может становиться экземпляром специально созданного служебного понятия, являющегося выражением, описывающим в терминах понятий метаонтологии необходимую семантику. Служебные понятия являются подпонятиями понятий метаонтологии.

Описание понятий онтологии в терминах какой-либо метаонтологии является определением их теоретико-модельной семантики с точки зрения этой метаонтологии. Поэтому если понятия двух онтологий, оказываются в одном классе метаонтологии, то с точки зрения этой метаонтологии они имеют схожую семантику. Спецификации двух согласовываемых онтологий в терминах одной метаонтологии позволяют формализовать в рамках метаонтологии семантический поиск и проверку корректности отображения понятий между разными контекстами: отображаемое понятие должно находиться в одном классе понятия метаонтологии (или служебного понятия) или в его подклассах. В качестве формального вывода применимо доказательство уточнения типов понятий метаонтологии.

## 5 Представление метасвойств понятий

С онтологическим понятием может быть связан набор фундаментальных метасвойств, с точки зрения которых можно оценить любое понятие или отношение. Им уделяется особое внимание в подходе корректировки иерархий онтологий [1]. Это такие метасвойства, как:

- существенность – неотъемлемость свойства сущности;
- стойкость – существенное свойство в любом воображаемом контексте или мире;
- идентификация – является ли понятие идентифицирующим сущности;
- зависимость – может ли сущность существовать вне зависимости от других;
- единство – определяет экземпляры как целые сущности, объединяющими части по какому-либо признаку, и другие.

Подразумеваемая семантика конкретных понятий однозначно определяет значение этих метасвойств, и для данного понятия в данном контексте эти значения не могут быть другими. Значения метасвойств предполагают некоторые взаимные ограничения на метасвойства понятий. Такие ограничения могут использоваться и для проверки корректности отображения понятий друг в друга при согласовании онтологий.

Для представления набора фундаментальных метасвойств понятий используются метафреймы, содержащие информацию о метасвойствах конкретного понятия. Для определения точного формата такого метафрейма задаём специальный тип.

```
{ Metaproperties;
  in: type;
  essence: boolean;
  rigidity: {enum; enumlist: {‘rigid’,
‘non-rigid’, ‘anti-rigid’}};
  identity: {enum; enumlist: {‘own’,
‘identical’, ‘non-identical’}};
  dependency: boolean;
  unity: boolean;
  …
}
…
{ metaproperties;
  in: class;
  instance_type: Metaproperties;
}
```

С понятием онтологии связываются конкретные значения метасвойств. Например, понятие `Human` имеет свойства существенности, строгости, независимости, единства, имеет собственные критерии идентификации сущностей.

```
{ Human;
    metaframe
    in: metaproperties;
    essence: true;
    rigidity: ‘rigid’;
    identity: ‘own’;
    dependency: false;
    unity: true;
    end
  in: concept;
}
```

Формальное определение подобных метасвойств требует явного описания семантики метасвойств и ограничений, которые они накладывают на понятия онтологий. Однако определения некоторых метасвойств выразимы только в логике второго порядка. Формулы языка Синтез ограничены утверждениями первого порядка. Поэтому семантика метасвойств скрыта в самом типе `Metaproperties`, и приходится довольствоваться

только ограничениями на значения метасвойств в связанных понятиях.

Обнаружение конфликтов метасвойств будет означать некорректность построения отображения понятий. Вот примеры критериев для проверки корректности отображения. Для понятия q, являющегося подпонятием (или уточнением для случая отображения понятий) понятия p, верны следующие ограничения:

- если q стойкое для любых сущностей, то p также стойкое для любых сущностей;
- если q несёт свойство, идентифицирующее сущности, то и p тоже;
- если q не несёт единство, то и p тоже;
- всякая сущность должна быть экземпляром единственного наиболее общего понятия, несущего его идентификацию, и другие.

Следствием набора определённых значений метасвойств является выделение разновидностей понятий, таких как родовые понятия («Звезда»), категории («Звезда главной последовательности»), роль («Спутник»), разновидностей отношений часть/целое. Для подобных составных метасвойств также можно определять типы метаинформации и составлять связанные с ними ограничения, вытекающие из ограничений значений метасвойств. Например, сущность может быть экземпляром класса только одного родового понятия.

Фактически исследования фундаментальных метасвойств привели к созданию онтологии верхнего уровня [10], содержащей разновидности понятий, классифицированные по разным свойствам. Использование такой онтологии верхнего уровня может быть альтернативой представления с помощью специальных метафреймов. Оценка метасвойств понятий может помочь и в отнесении понятий к соответствующим понятиям онтологии верхнего уровня.

## 6 Представление экземпляров в классах понятий

Выше уже было показано, как фрейм, являющийся экземпляром класса понятия, представляется в языке Синтез. Например, «Сириус» является экземпляром понятия «Звезда»:

```
{ Sirius;
  in: frame, star;
}
```

С использованием экземпляров понятий, будь то сущности реального мира или информация, соотнесённая с понятиями онтологии, может быть реализован денотационный подход к определению семантики понятий для их согласования. Конфликты в отнесении конкретных сущностей и моделей реального мира экспертами из разных онтологических контекстов к понятиям своих онтологий будут служить сигналом к тому, что в отображении онтологий присутствуют ошибки.

Например, в одной онтологии понятие «Астрономический объект» включает все нерукотворные астрономические сущности, а в другой онтологии понятие «Источник» включает сущности, имеющие фиксируемое электромагнитное излучение. Понятия связаны, но не эквивалентны, и конфликт между ними можно обнаружить при проверке принадлежности обоим понятиям, в частности, какой-либо известной тёмной туманности, не излучающей электромагнитного излучения.

Методика согласования онтологий с помощью сущностей-экземпляров понятий предполагает работу не одного эксперта, а взаимодействие экспертов, представляющих каждую из согласовываемых онтологий. Данный экстенсиональный подход к отображению онтологических понятий «по образцу» может быть реализован формальным образом, так как существование хотя бы одного конфликтного экземпляра подвергает сомнению отношение уточнения понятий.

Другой подобный подход спецификации денотационной семантики понятий использует не конкретные сущности, но классы сущностей, которые могут применяться для проверки принадлежности понятиям обоих онтологий в качестве подпонятия.

## 7 Концептуальное моделирование предметных областей

До сих пор при упоминании схем речь шла о концептуальных схемах баз данных, специфицирующих структуры данных и ограничения целостности независимо от структур хранения и реализации. Структуры данных и ограничения схемы отражают соответствующие информационные аспекты сущностей предметной области.

Разница между онтологией предметной области и концептуальной схемой предметной области определяется назначением, направленностью на понятия или на информационные структуры. Способы спецификации могут быть разными в соответствии с разными назначениями. Однако для концептуальной схемы предметной области необходимо использовать выразительные с точки зрения семантики спецификации, чтобы точно определять каким понятиям предметной области соответствуют сущности, описанные в концептуальной схеме.

Что касается подходов к интеграции концептуальных схем предметных областей, то описанные в статье подходы к согласованию спецификаций равно применимы к ним, как и к онтологиям, и также полезны для семантически точного отображения схем.

В частности, одним из подходов к определению семантики спецификаций концептуальной схемы предметной области является семантическая

аннотация элементов схемы в терминах онтологии предметной области. Для этого используются средства, аналогичные тем, которые применялись в разделе 4 для аннотации спецификаций онтологии в терминах метаонтологии.

На языке Синтез аннотирование производится с помощью указания класса понятия онтологии в слоте `in` среди списка классов, в которые входит данный элемент спецификации. Другими словами, элементы спецификации схемы становятся экземплярами класса понятия. Выражения в терминах онтологии определяют новые служебные подпонятия, которые создаются в случае необходимости для описания точного смысла элемента спецификации схемы в терминах онтологии. Ниже представлен пример типа `Star` концептуальной схемы, описывающего звезду, с одним атрибутом `lc`, в котором задаётся значение класса светимости звезды от 1 до 5. При этом тип `Star` описывается как экземпляр класса `star` онтологии `StellarOntology`, а атрибут `lc` – как экземпляр класса `luminosityClass` этой же онтологии.

```
{ Star;
    metaframe
    in: StellarOntology.star;
    end
  in: type;

  lc: integer;
    metaslot
    in: StellarOntology.luminosityClass;
    end
  lcInv:
    {in: predicate, invariant;
      { predicative;
        { all s/Star
          ( s.ls >= 1 & s.ls <= 5)
    }}}
};
```

При согласовании схем предметной области релевантными будут считаться те отображаемые элементы схем, которые согласно аннотации соответствуют тому же понятию или одному из его подпонятий.

Также к концептуальным схемам предметной области применимы и подходы их согласования, основанные на метасвойствах типов и на экземплярах классов.

## Заключение

Приведённые в статье подходы к семантическому согласованию онтологий требуют дополнительного определения семантики онтологических понятий с разных точек зрения, но позволяют достигать формального и семантически обоснованного отображения понятий. Приведённые подходы к согласованию онтологий учитывают специфику онтологической информации, не ограничиваясь методами согласования структурных спецификаций онтологий.

Представленные методы применимы как для неавтоматизированного взаимодействия экспертов-представителей конкретных онтологий при ведении дискуссий в ходе согласования онтологий, так и для создания систем поддержки согласования онтологий, предоставляющей автоматизированные методы в помощь экспертам для построения формальных отображений онтологий.

## Литература


[1] Guarino, N. and Welty, C. An overview of OntoClean. In Staab, S. and Studer, R. (Eds), Handbook on Ontologies, Springer, Berlin, 2004, pp. 151-172

[2] Kalinichenko L.A., Stupnikov S.A., Martynov D.O. SYNTHESIS: a Language for Canonical Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments. Moscow: IPI RAN, 2007. - 171 p.

[3] J. Euzenat, P. Shvaiko. Ontology Matching. Springer-Verlag, New York, 2007

[4] И. Л. Артемьева. Сложно структурированные предметные области. Построение многоуровневых онтологий. "Информационные технологии " №1, 2009 – сс. 16-21

[5] Н. А. Скворцов. Вопросы согласования онтологических моделей и онтологических контекстов. Симпозиум «Онтологическое моделирование», М: ИПИ РАН, 2008

[6] Н. А. Скворцов. Специфика подходов согласования онтологических контекстов. «Знания и Онтологии *ELSEWHERE* – 2009», ICCS'09, Москва, 2009

[7] Н. А. Скворцов, С. А. Ступников. Использование онтологии верхнего уровня для отображения информационных моделей. RCDL'2008, Дубна: ОИЯИ, 2008 – сс. 122-127

[8] С. А. Ступников. Отображение спецификаций ядра канонической модели в нотацию абстрактных машин. Формальные методы и модели для композиционных инфраструктур распределённых информационных систем: Системы и средства информатики, специальный выпуск. – М: ИПИ РАН, 2005. -- сс. 69--95

[9] С. А. Ступников Моделирование композиционных уточняющих спецификаций. Диссертация на соискание степени кандидата технических наук. -- М: ИПИ РАН, 2006. – 195 с.

[10] DOLCE : a Descriptive Ontology for Linguistic and Cognitive Engineering. http://www.loa-cnr.it/DOLCE.html

# Formal representation of metainformation for some approaches to ontology reconciliation

N. A Skvortsov

Approaches used for support of ontology reconciliation are mostly based on the methods developed for schema integration. However it is not enough since ontologies define concepts with similar semantics using different approaches to structure description. At recent paper represented in the seminar “Knowledge and ontology *ELSEWHERE* - 2009”, approaches of discovering similarity and difference of concept semantics for ontology reconciliation. In continuation of that research we have developed representation of metainformation for these approaches in the model defined by the Synthesis language used in conceptual modeling of subject domains and as a canonical model for integration of heterogeneous information resources into scientific problem specifications. The paper describes how to use metaontologies, high level ontologies, fundamental kinds of metaproperties, and concept class instances for semantic reconciliation of ontological contexts, and represents specifications of this metainformation in the model defined by Synthesis language.

♣ Работа выполнена при поддержке РФФИ (грант 08-07-00157-а) и программы фундаментальных исследований Президиума РАН №3 «Фундаментальные проблемы системного программирования».

# Онтологическая семантика текста: форматирование лексики в семантическом словаре

© Г.В. Лезин

Санкт-Петербургский экономико-математический институт РАН
lezin@emi.nw.ru

**Аннотация**

В статье рассматривается задача выработки единого представления информации в семантической сети текста, лексиконе и в базовой онтологии. Предлагается дополнить стандартные описания классов в онтологии форматами многокомпонентных толкований, подготовленных для использования в лексиконе. Рассматривается семантика форматов и их связь с онтологией и лексиконом.

## 1 Введение

Интерпретация текста набором формальных утверждений, записываемых на предварительно строго определенном формальном метаязыке, является одной из центральных и не решенных в настоящее время задач автоматического анализа текста. Многообразие требований, предъявляемых к такому метаязыку со стороны различных приложений, делает практически невозможным его создание в законченном, раз и навсегда фиксированном виде. Но можно попытаться использовать в качестве метаязыка строго формализованную систему стандартных правил определения средств описания семантики текста для нужд конкретных приложений. Примером такого подхода, но для формирования описаний семантики данных во всемирной семантической паутине (Semantic Web) является предложенная W3C Консорциумом система формальных языков RDF/OWL [1, 2]. Опора на хорошо исследованную математическую модель (дескриптивные логики), наличие доступной и тщательно разработанной документации, имеющаяся уже в настоящее время программная поддержка (редакторы баз знаний, связи с языком программирования и современными базами данных) делают эту систему языков привлекательной и для экспериментов, нацеленных на получение формальной интерпретации текста.

Целесообразность интерпретации текста семантической сетью отношений между сущностями, упоминаемыми в тексте, стала общепризнанным фактом [3]. Сеть текста строится на основе описаний лексики языка, получаемых из семантического словаря (лексикона). Согласование описаний лексики в лексиконе с представлением семантических сетей текстов, формируемых на основе этих описаний, создание единой семантической основы, в терминах которой описываются как лексика языка, так и семантические сети текстов, написанных на этом языке – это задача, без решения которой говорить об автоматической интерпретации текстов просто не приходится.

Попытка определения такой общей семантической основы в виде сильно ограниченного универсального набора семантических примитивов так и не привела к созданию общепризнанного семантического языка [4]. Рассматривая известные проекты создания семантических словарей с формальными толкованиями лексики [5-8], можно отметить следующие характеристики семантической основы:

а) Определению формул толкования предшествует выделение подсловаря исходных (атомарных) понятий, используемых в толкованиях оставшейся части лексикона (например, в словаре В.А. Тузова [7] более 3000 атомарных понятий).

б) Важной компонентой толкования лексического значения слова является указание места лексемы в общей таксономической иерархии понятий лексикона.

в) Наличие развитого списка свойств (ролей, атрибутов), позволяющих связывать атомарные понятия в общую формулу толкования лексемы.

г) В общей таксономической иерархии лексикона выделяются классы лексем, формулы толкования которых имеют общие части. Исследованию связей между таксономической характеристикой слова и форматом его толкования посвящена монография [6].

Отмеченные характеристики лексикона, ориентированного на получение семантической интерпретации текста, позволяют надеяться на возможность оформления метаязыка системы лексикон/текст средствами RDF/OWL. Подчеркнем, что речь идет о создании лингвистического ресурса для решения прикладных задач, в частности задач извлечения информации из текста. Рассматриваемые в статье задачи и возможные

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

методы их решения относятся, главным образом, к инженерной технике совместного представления информации в лексиконе и лежащей в его основе онтологии.

Будем исходить из следующих предположений [9].

1) Семантическая основа лексикона оформляется в виде базовой OWL-онтологии, содержащей описания атомарных понятий.

2) Семантическая сеть анализируемого текста описывается конъюнкцией RDF-триплетов. Узлами семантической сети могут быть:

– имена классов базовой OWL-онтологии;

– имена классов – лексических значений слов текста, полученные из лексикона;

– имена-референты (индивиды и подклассы классов базовой OWL-онтологии, упомянутые в тексте) автоматически сформированные в процессе анализа текста;

– узлы-переменные (blank nodes), сформированные в соответствии с модельной семантикой RDF.

Ребрами семантической сети могут быть исключительно имена свойств, заданные базовой OWL-онтологией.

3) Каждое из лексических значений слова (лексема) в лексиконе толкуется как подкласс одного или нескольких классов из базовой онтологии. Возможно уточнение заданного общим классом толкования лексемы в виде высказывания на атомарных элементах OWL-онтологии. Толкования в лексиконе оформляются в виде RDF-сетей, подготовленных для включения в общую семантическую сеть текста. Явно выделены входной и выходные узлы сети толкования; узлы-переменные сети толкования согласованы с актантными переменными, указанными для синтаксических актантов лексемы.

Важно отметить, что базовая онтология в сочетании с лексиконом образует схему примитивных знаний о мире, нашедших отражение в лексике языка. Создана такая база может быть лишь в результате опыта, как результат исследований разных коллективов, но представляется принципиально важным, чтобы в ее основе лежали общепринятые средства описания знаний. Мы в нашей работе рассматриваем всего лишь возможный подход к решению этой задачи – общую технику описания лексики языка, позволяющую увязать в единый комплекс информацию в онтологии, лексиконе и анализируемом тексте.

Согласование информации в базовой OWL-онтологии, лексиконе и семантической сети текста связано с необходимостью поиска ответов на следующие вопросы:

а) Как определить в OWL-онтологии атомарные элементы, допустимые для использования в толкованиях лексикона (возможные ограничения, правила преобразования OWL-описания элемента в его же RDF-представление в составе толкования)?

б) Как увязать отобранные для толкования атомарные элементы в параметризованное высказывание (определение общей структуры толкования, выделение списка семантических актантов и их связь с синтактикой лексемы, возможность автоматизации процесса увязывания)?

в) Как организовать коллекцию общих форматов толкований (связь формата толкования с OWL-онтологией и его использование в конкретных толкованиях)?

В статье мы пытаемся найти возможный вариант ответа на поставленные вопросы.

## 2 Атомарные элементы онтологии.

Наша задача – оформление толкования лексемы в виде RDF-фрагмента семантической сети, точнее, в виде шаблона, переменные которого при подключении фрагмента к семантической сети текста, как правило, подлежат замене на константные значения. В RDF семантическая сеть интерпретируется как конъюнкция триплетов [13]. Логично потребовать, чтобы толкование лексемы представляло собой конъюнкцию атомарных элементов, получаемых из базовой OWL-онтологии и тоже оформленных в виде фрагментов семантической сети. (Мы ограничиваемся однозначными толкованиями значений слова. О возможности использования дизъюнктивно организованных толкований см. [14].)

Исходной информацией для построения атомарного элемента толкования служит описание класса в базовой OWL-онтологии. Согласно [15] описание класса содержит указание класса и список ограничений на свойства, сопоставляемые индивидам класса. Класс в описании может быть задан либо явно, либо как результат применения операций пересечения, объединения или дополнения классов. В любом случае описание класса позволяет выделить список свойств, явно сопоставленных классу указаниями ограничений на свойства. Полученный таким образом список свойств будем называть списком свойств класса.

Свойство *prop* может быть сопоставлено классу $C_I$ любым сочетанием из трех видов ограничений:

а) Ограничение "only V": любой индивид u из класса $C_I$ может иметь свойство *prop* с областью значений, ограниченной исключительно классом V, т.е.:

u *rdf:type* $C_I$; *prop* v влечет v *rdf:type* V.

б) Ограничение "some V": факт принадлежности индивида u классу $C_I$ свидетельствует о наличии у него некоторого значения v свойства *prop*, т.е. :

u *rdf:type* $C_I$ влечет $\exists$(u *prop* v. v *rdf:type* V)

в) Ограничения кардинальности: каждый индивид u из класса $C_I$ должен иметь в семантической сети значения свойства *prop* в количестве, удовлетворяющем ограничению (например, при *owl:minCardinality*=2 индивиду u в сети должно быть сопоставлено не менее двух значений свойства *prop,* при *owl:maxCardinality*=3 – не более трех значений этого свойства, а при *owl:cardinality*=2 – в точности два значения)

Предположим, что описание класса с именем ont:ClassId сопоставляет этому классу список из N

свойств: ont:prop1, …, ont:propN. Атомарным шаблоном класса ClassId будем называть RDF-граф вида

_:x rdf:type ont:ClassId; ont:prop1 _:x1; … ; ont:propN _:xN.

В определении использована нотация N3; _:x, _:x1, … , _:xN – обозначения переменных; ont: – префикс пространства имен OWL-онтологии. Шаблон получен в соответствии с правилами трансляции OWL-описания класса ont:ClassId в RDF-представление этого описания [15] с исключением из результатов трансляции ограничений на свойства, заданных в OWL-описании.

Пример 1.

В нашей базовой OWL-онтологии класс ont:Нахождение_Место описывает местонахождение объекта в заданном интервале времени.

```
ont:Нахождение_Место
    rdf:type owl:Class;
    rdfs:subClassOf ont:СостояниеОбъект;
    rdfs:subClassOf
        [rdf:type owl:Restriction ;
owl:onProperty ont:c_Уровнем_высоты;
owl:allValuesFrom ont:РазмерДистанции;
rdfs:subClassOf
        [rdf:type owl:Restriction ;
owl:onProperty ont:c_Уровнем_высоты;
owl:someValuesFrom ont: РазмерДистанции
        ] ;
    rdfs:subClassOf
        [rdf:type owl:Restriction ;
owl:onProperty ont:c_Местом ;
owl:someValuesFrom ont:Место_Объекты
        ] ;
    rdfs:subClassOf
        [rdf:type owl:Restriction ;
owl:onProperty ont:c_Местом;
owl:allValuesFrom ont:Место_Объекты
        ] ;
    rdfs:subClassOf
        [rdf:type owl:Restriction ;
owl:onProperty ont:c_Дистанцией;
owl:allValuesFrom ont:РазмерДистанции
        ] .
```

В приведенном описании класса ont:Нахождение_Место указаны лишь непосредственно заданные для этого класса ограничения на свойства. Ряд других свойств, также заданных ограничениями, наследуется от объемлющего класса ont:СостояниеОбъект (свойство ont:c_Объектом определяет субъекта, находящегося в определяемом состоянии; свойствами ont:c_Началом и ont:c_Концом – интервал времени существования состояния и свойством ont:c_Модальностью – фактивность состояния). В результате атомарным шаблоном класса ont:Нахождение_Место является граф:

```
_:x rdf:type ont:Нахождение_Место;
ont:c_Объектом _:x1;
ont:c_Местом _:x5; ont:c_Дистанцией _:x6;
ont:c_Уровнем высоты _:x7;
ont:c_Началом _:x2; ont:c_Концом _:x3;
ont:c_Модальностью _:x4.
```

Переменную, определенную на классе атомарного шаблона (в Примере 1 это _:x), мы будем называть переменной шаблона, прочие переменные – параметрами шаблона.

Отметим, что использование конкретных обозначений переменных в атомарном шаблоне, вообще говоря излишне. Можно было бы воспользоваться парой скобок '[ ]', считая каждую такую пару скобок указанием переменной с обозначением, отличным от других переменных. Мы будем использовать эту возможность в последующих примерах.

Отображение времени текста, регистрация временных отношений между ситуациями, описываемыми в тексте, является одной из основных компонент его семантической интерпретации. Учет времени текста осуществляется на основе информации, поставляемой толкованиями лексических значений слов [10]. Соответственно, базовая OWL-онтология должна обеспечивать необходимый набор атомарных понятий. Мы в нашем проекте следуем подходу, разрабатываемому в онтологии DOLCE [11].

Все классы OWL-онтологии делятся на две категории: времязависимые и постоянные. Любому из индивидов времязависимого класса сопоставляется интервал времени его существования. Отметим, что свойства любой онтологии, описанной на языках RDF или OWL по определению постоянны, не имеют индивидов и не могут, в свою очередь, иметь свойств. В то же время индивид, представляющий сущность из времязависимого класса, может в разное время иметь различные значения одного и того же свойства. (Например, какой-либо предмет может в разное время находиться в разных местах.) Отсюда возникает потребность сопоставлять время существования не только индивидам предметного плана, но и состояниям этих индивидов, процессам, участниками которых они являются. В общем классе времязависимых сущностей выделяется иерархия подклассов, позволяющих описать процессы и состояния с указанием времени их существования. Описание отношений между временными интервалами, их протяженности и положения на временной оси может быть выполнено, например, средствами временной онтологии [12].

## 3 Форматы (сценарии) толкований.

Достаточно широкому спектру лексических значений слов свойственна многокомпонентность толкований. Типична ситуация, когда некоторое множество лексем характеризуется наличием общей структуры толкования и именно эта общность служит основанием для объединения этих лексем в общий класс онтологии. Например, для глаголов, описывающих перемещение объекта из исходной

точки в конечную (*лететь, передвигать, идти* и др.), характерен сценарий:

– нахождение объекта в исходной точке;

– затем: движение объекта из исходной точки в конечную;

– затем: нахождение объекта в конечной точке.

Явно выделяются семантические участники этого сценария: Агенс, Объект, Исходная точка, Конечная точка, интервал времени, охватываемый сценарием.

Возникает задача дополнения базовой OWL-онтологии достаточно тесно увязанным с нею и тем не менее новым информационным ресурсом – описаниями форматов толкований. Одним из возможных способов подключения такого ресурса мог бы быть следующий:

а) Формат толкования сопоставляется классу OWL-онтологии. В качестве свойств этого класса, явно заданных в его описании или наследованных, указываются имена ролей, задающих параметры формата толкования.

б) Определяется набор шаблонов, образующих толкование. Шаблоны задаются указанием соответствующих классов базовой OWL-онтологии.

в) Нормализация (переобозначение) переменных в шаблонах. В результате этой операции переменные в шаблонах не имеют совпадающих обозначений.

г) Параметризация шаблонов. Переменным в шаблонах, обозначающим общие узлы графа толкования, присваиваются общие обозначения (слияние узлов графа, обозначенных этими переменными). Полученный в результате действий в) и г) словарь обозначений переменных представляет собой полный (в рамках заданной базовой OWL-онтологии) список семантических параметров формата – набор узлов графа толкования, допустимых для слияния с узлами общей семантической сети текста. Этот словарь аморфен в том смысле, что в нем отсутствуют деление на актанты и сирконстанты, а также указания коммуникативного ранга указанных нем параметров.

Пример 2.

Отмеченный выше сценарий перемещения может быть представлен следующим форматом:

```
<script>
 <header>
  _:x rdf:type ont:СценарийПеремещения;
    ont:c_Объектом _:Y1; ont:рольИточка _:Y2;
    ont:рольИдистанция _:Y3;
    ont:рольИвысота _:Y4; ont:рольКточка _:Y5;
    ont:рольКдистанция _:Y6;
    ont:рольКвысота _:Y7; ont:рольНачало _:Y8;
    ont:рольКонец; _:Y9".
</header>
<body>
  _:x1 rdf:type ont:Нахождение_Место;
    ont:c_Объектом _:Y1; ont:c_Местом _:Y2;
    ont:c_Дистанцией _:Y3;
    ont:c_Уровнем_высоты _:Y4;
    ont:c_Началом [ ]; ont:c_Концом _:Y8.

  _:x2 rdf:type ont:ДвижениеПроцесс;
    ont:c_Агенсом [ ]; ont:c_Объектом_:Y1;
    ont:co_Скоростью [ ]; ont:мимо [ ];
    ont:следом [ ]; ont:c_Началом _:Y8;
    ont:c_Концом _:Y9.

  _:x3 rdf:type ont:Нахождение_Место;
    ont:c_Объектом _:Y1; ont:c_Местом _:Y5;
    ont:c_Дистанцией _:Y6;
    ont:c_Уровнем_высоты _:Y7;
    ont:c_Началом _:Y9; ont:c_Концом [ ].
</body>
</script>
```

Формат имеет две явно выраженные части: заголовок и тело формата. Заголовком формата является шаблон класса общей OWL-онтологии. В Примере 2 таким классом является ont:СценарийПеремещения. Тело формата содержит RDF-граф, интерпретирующий класс, заданный заголовком. Формат оформлен в виде XML-контейнера.

RDF-граф тела формата представляет собой результат объединения шаблонов из заданного набора с предварительной нормализацией переменных шаблонов и последующей параметризацией полученного объединения. В Примере 2 заголовок интерпретируется набором из трех атомарных шаблонов: ont:Нахождение_Место, ont:ДвижениеПроцесс и ont:Нахождение_Место. Парами скобок '[ ]' обозначены места вхождения переменных с уникальными в пределах формата обозначениями, полученными в результате нормализации. Явно обозначенные Y-переменные отмечают результаты параметризации. В совокупности имена неявно обозначенных переменных и Y-переменных образуют словарь параметров формата. Имена _:x1, _:x2 и _:x3 использованы для обозначения переменных шаблонов.

Одно из основных назначений формата – многократное использование в толкованиях конкретных лексем. Принципиально важной является семантическая эквивалентность между заголовком формата и его телом. Содержательно эквивалентность заголовка и тела формата должна проявляться в следующем:

а) Указывая в толковании лексемы шаблон класса, которому сопоставлен формат, мы тем самым объявляем тело формата компонентой толкования лексемы.

б) Выявив в семантической сети текста фрагмент, из которого следует истинность высказывания, заданного телом формата, мы тем самым выявляем в тексте наличие примера сценария, которому сопоставлен этот формат.

Формальное определение эквивалентности основано на предположении, что полные списки параметров заголовка формата и его тела должны совпадать. Иначе говоря, мы должны иметь возможность сослаться на любой параметр тела формата в его заголовке. Отметим, что параметры в шаблонах идентифицируются свойствами,

значениями которых они являются. Обозначения переменных не могут служить идентификаторами параметров в силу локальности их использования и постоянной готовности к переименованию в процессе нормализации. При взаимно однозначном соответствии пары "свойство" – "параметр", выявленном в теле формата, соответствующее свойство удобно считать сопоставленным классу формата и явно не указывать в списке свойств его заголовка. В Примере 2 классу ont:СценарийПеремещения наряду со свойствами, перечисленными в заголовке, сопоставлены также свойства ont:с_Агенсом, ont:со_Скоростью, ont:мимо, ont:следом; при этом в заголовке формата обеспечено взаимно однозначное соответствие между параметрами формата и идентифицирующими их свойствами.

Пусть F – любой из классов общей OWL-онтологии, которому сопоставлен формат толкования; $X_F$ - общий список параметров формата F; $x_F$ - переменная, определенная на классе F; $V_F$ - набор переменных, определенных на классах шаблонов, образующих толкование (в нашем примере $V_F$={_:x1, _:x2, _:x3}); $\exists(V_F)$ - факт существования кортежа значений для переменных из набора $V_F$; $H_F(x_F, X_F)$ – RDF-высказывание, заданное заголовком формата F и $S_F(V_F, X_F)$ – RDF-высказывание тела F . Тогда заголовок формата F и его тело связаны отношением эквивалентности:

$$\forall(F)[\ \forall(x_F)[\ H_F(x_F, X_F) \rightarrow \exists(V_F)\ S_F(V_F, X_F)] \ \&$$
$$\forall(V_F)[\ S_F(V_F, X_F) \rightarrow \exists(x_F)\ H_F(x_F, X_F)]]$$

при взаимно однозначном соответствии между значениями $x_F$ и конкретными кортежами значений $V_F$.

Рассмотренное отношение имеет статус аксиомы, т.е. исходного требования, которому должен удовлетворять вновь определяемый формат. Следствием этой аксиомы является возможность сопоставления форматов в условиях, когда один формат отличается от другого лишь тем, что некоторые из шаблонов этого формата являются подклассами соответствующих шаблонов другого.

Пусть S1 и S2 – классы общей OWL-онтологии, причем каждому из них сопоставлен собственный формат толкования:

$S1 \leftrightarrow P_1, \ldots, P_i, \ldots, P_k$ и $S2 \leftrightarrow P_1, \ldots, C_i, \ldots, P_k$

где $C_i$, $P_1,\ldots, P_k$ – шаблоны классов OWL-онтологии, $C_i$ является подклассом $P_i$.

Тогда S2 – подкласс S1.

Рассмотрим произвольный кортеж $(p_1, \ldots, c_i, \ldots, p_k)$ значений переменных шаблонов формата S2: $(p_1, \ldots, c_i, \ldots, p_k) \in V_{S2}$. В соответствии с отношением эквивалентности ему сопоставлен единственный индивид s2 класса S2. Но $c_i$, будучи индивидом класса $C_i$, принадлежит и классу $P_i$. Следовательно, $(p_1, \ldots, c_i, \ldots, p_k) \in V_{S1}$, и кортежу сопоставлен также и индивид s1 из класса S1. Это невозможно ввиду требования взаимно однозначного соответствия между индивидами класса, заданного заголовком формата, и кортежами значений переменных шаблонов. Таким образом, $s1 \equiv s2$, т.е. любой индивид из класса S2 принадлежит также классу S1.

Определяя формат как средство описания семантики лексики, дополняющее общую OWL-онтологию, полезно рассмотреть вопросы, связанные с корректностью параметризации тела формата и попытаться выявить возможности наследования форматов для классов, связанных отношением "класс–подкласс".

В общем случае процедура параметризации тела формата сводится к попарному слиянию узлов графа, причем каждый из объединяемых узлов имеет в базовой онтологии собственное описание области возможных значений, заданное описанием шаблона. Объединяться могут как узлы, представленные переменными шаблонов, так и параметры шаблонов. В любом случае объединять можно лишь узлы с непротиворечивыми описаниями. Условия корректного слияния естественным образом следуют из семантики языков RDF/OWL:

а) Основным элементом описания узла является указание класса его возможных конкретных значений. Пусть узел $\_:x_i$ определен на непустом классе $C_i$, узел $\_:x_j$ – на непустом классе $C_j$, а $\_:y$ – обозначение результата слияния этих узлов. Тогда классом возможных значений узла $\_:y$ будет пересечение $C_i \cap C_j$. Слияние узлов $\_:x_i$ и $\_:x_j$ корректно, если это пересечение не пусто. В частности, при наличии в базовой онтологии указания на непересекаемость $C_i$ и $C_j$ слияние узлов $\_:x_i$ и $\_:x_j$ не корректно.

б) Параметры шаблона наряду с указанием класса возможных значений (only-ограничение) могут иметь также some-ограничение и явно заданную кардинальность. Some-ограничение для заданного в шаблоне свойства устанавливает факт существования значения свойства из заданного в ограничении класса (см. п.2). Если хотя бы один из объединяемых узлов имеет some-ограничение, результат слияния также имеет some-ограничение, но класс значений для этого ограничения – пересечение классов, заданных для значений объединяемых узлов. При объединении узлов, имеющих ограничения на кардинальность, для результата устанавливается сильнейшее из заданных ограничений.

Итак, слияние параметров шаблонов может привести к необходимости определения новых ограничений на значения свойств, зафиксированные в исходных шаблонах. В OWL-онтологии мы можем зафиксировать ограничение, которому должно удовлетворять конкретное значение v свойства p в условиях, когда это свойство сопоставлено конкретному классу $C_I$ (см. п.2). Таким классом в нашем случае является класс, интерпретируемый форматом и заданный его заголовком. Фиксация ограничения сводится к явному определению для этого класса свойства, значения которого обладают требуемым ограничением. В заголовке формата этому свойству сопоставляется соответствующий параметр.

Подводя итог, заметим следующее:

1) Полное множество классов OWL-онтологии может быть поделено на две категории: общие классы, полностью определенные средствами онтологии, и классы, которым наряду с их описанием в онтологии сопоставлен формат толкования – F-классы.
2) Шаблон F-класса является заголовком сопоставленного классу формата толкования и строится по общим правилам (см. п.2).
3) Список параметров шаблона F-класса совпадает со списком параметров, заданных телом его толкования. Ограничения на значения параметров в описании F-класса в онтологии не должны противоречить ограничениям на эти же параметры в теле толкования класса.
4) Список свойств F-класса должен обеспечивать в рамках описания класса взаимно однозначное соответствие между свойством и параметром, указанным в качестве значения свойства.
5) Вхождение шаблона F-класса в какое-либо RDF-высказывание (например, в состав толкования лексемы) равносильно ссылке на тело толкования в формате этого класса.

## 4 Структура толкований в лексиконе

При определении структуры описания лексического значения слова в лексиконе мы будем исходить из предположения, что в составе общей OWL-онтологии всегда имеется класс, к которому можно отнести определяемую лексему. Таким классом может быть либо класс, описанный стандартными OWL-средствами, либо F-класс, которому наряду со стандартными средствами описания сопоставлена также формула толкования. Выделив непосредственные описания смысла лексем в общую онтологию, мы в лексиконе можем сосредоточиться на определении условий выбора лексемы из набора возможных значений слова и на информации, поставляемой лексемой для решения задачи композиции семантической сети предложения и далее текста из фрагментов, поставляемых описаниями лексем.
Мы рассматриваем семантическую сеть предложения как результат трансформации синтаксической структуры этого предложения. Синтаксис предложения задается деревом сочинительно-подчинительных связей между его словами и словосочетаниями. Толкованиями лексических значений слов поставляются исходные заготовки, фрагменты семантической сети предложения. В процессе трансформации эти заготовки могут изменяться в соответствии с правилами объединения фрагментов в общую сеть [6]. В любом случае на фрагмент семантической сети толкования наложено следующее ограничение: фрагмент сети должен иметь единственный входной узел; входной и выходные узлы должны быть явно обозначены.
В Примере 3 показан фрагмент описания глагола ПЕРЕДВИНУТЬ. Фрагмент содержит две части: синтактику (описание валентности), в рамках которой выделены синтаксические актанты глагола при его использовании в предложении, и семантику глагола в значении "передвинуть физический объект с одного места на другое". Синтактика примера (в несколько упрощенном виде) взята из словаря В.А. Тузова [7].

Пример 3.
ПЕРЕДВИНУТЬ
Синтактика:
Z1: !Им,Z2: !Вин,Z3: НЕЧТО$1~!Откуда,Z4: НЕЧТО$1~!Куда)
Семантика лексемы:

```
<interpret  voc:input="_:x" voc:instance="_:x" >
  voc:ПЕРЕДВИНУТЬ rdf:subClassOf
                ont:СценарийПеремещения;
   _:x rdf:type voc:ПЕРЕДВИНУТЬ;
    ont:с_Агенсом :Z1; ont:с_Объектом_:Z2;
    ont:рольИточка _:Z3; ont:рольКточка _:Z4;
    ont:рольНачало _:Y8; ont:рольКонец; _:Y9".
   _:Y8 rdf:type ont:НачалоСитуации.
   _:Y9  rdf:type ont:КонецСитуации.
</ interpret >
```

Семантика лексемы в лексиконе оформляется в виде XML-контейнера, содержанием которого является текст RDF-графа, отражающего информацию, привносимую лексемой в семантическую сеть предложения. Атрибутами контейнера определяется входной узел графа. Набор его выходных узлов совпадает со списком параметров лексемы.
В толковании лексемы указываются:
– класс, подклассом которого является определяемая лексема (в Примере 3 – ont:СценарийПеремещения);
– параметры лексемы, увязываемые с ее синтактикой;
– возможные уточнения ограничений, накладываемых данной лексемой на параметры (в приведенном примере это указание на замкнутость интервала времени для ситуации описываемой лексемой).
В лексиконе каждому из лексических значений слова (его лексеме) сопоставляется собственная семантика. Синтактика может быть общей для нескольких лексем. Семантическими актантами лексемы принято называть список параметров ее толкования. В толковании набор семантических актантов лексемы совпадает со списком параметров ее класса, описанного в общей OWL-онтологии. В онтологии для каждого из параметров определен класс возможных значений и указано свойство (актантная роль) определяемое этим параметром. В составе описания лексемы упоминаются лишь те из ее актантных ролей, которые требуют согласования с ее синтактикой.
Таким образом, полная информация о семантике лексемы распределена по трем информационным ресурсам:
– общая OWL-онтология;
– форматы толкований;
– лексикон.
В задачу онтологии входит описание словаря основных понятий, в терминах которых

описываются смысловые значения слов естественного языка и обеспечение связи между описаниями лексики в лексиконе и форматами толкований.

Форматами толкований решается задача увязывания компонент толкования в единое целое. На этом уровне описаний лексики удается преодолеть статичность чисто онтологических описаний, отразить развитие ситуации во времени, определить пред- и постусловия реализации различных ситуаций. На этом же уровне достаточно обусловленным образом формируется список семантических актантов, свойственных той или иной ситуации. Формат дает схематическое описание смысла, образуя тем самым новое понятие, регистрируемое в онтологии как особый класс.

## Литература.


[1] OWL Web Ontology Language Guide. W3C Recommendation 10 February 2004 http://www.w3.org/TR/2004/REC-owl-guide-2004021.

[2] RDF Primer. W3C Recommendation 10 February 2004 http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

[3] Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. – М.: Изд. центр "Академия", 2006. 304 с.

[4] Перцов Н.В. К проблеме построения семантического метаязыка. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог 2006". – М.: Изд-во РГГУ, 2006. С. 419-425.

[5] Кустова Г.И., Падучева Е. В. Словарь как лексическая база данных // Вопросы языкознания, М., 1994, N 4. С. 96-106

[6] Падучева Е.В. Динамические модели в семантике лексики. – М. "Языки славянской культуры", 2004. 608 с.

[7] Тузов В. А. Компьютерная семантика русского языка. – СПб.: Издательство СПб ГУ, 2004. 400 с.

[8] Nirenburg S., Raskin V. Ontological Semantics. http://crl.nmsu.edu/Staff.pages/Technical/sergei/book/

[9] Лезин Г.В. Онтологическое представление семантики предложений в системе языков RDF/OWL. //Языковая инженененрия: в поисках смыслов. Доклады семинара "Лингвистические информационные технологии в Интернете". – СПб.: СПб ГУ, 2008. С. 5-21.

[10] Падучева Е.В. Семантические исследования: семантика времени и вида в русском языке. Семантика нарратива. – М.: "Языки русской культуры", 1996.

[11] Masolo C., Borgo S., Gangemi A, Guarino N., Oltramari A., Schneider L. DOLCE: a Descriptive Ontology for Linguistic and Cognitive Engineering // DOLCE documentation: http://www.loa-cnr.it/DOLCE.html.

[12] Time Ontology in OWL. W3C Working Draft. http://www.w3.org/TR/2006/WD-owl-time-20060927/

[13] RDF Semantics. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-mt-20040210/

[14] Урынсон Е.В. "Несостоявшаяся полисемия" и некоторые ее типы // Семиотика и информатика.– М.: "Яз. рус. культ., Рус. словари", Вып. 36,1998. С. 226-261.

[15] OWL Web Ontology Language. Semantics and Abstract Syntax. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-owl-semantics-20040210/


## Ontological semantics of the text: formatting of interpretation in the semantic dictionary


Lezin G.V.



In article the problem of development of uniform representation of the information in a semantic network of the text, a lexicon and in a base ontology is considered. It is offered to add standard descriptions of classes in an ontology with formats of the multicomponent interpretation prepared for use in a lexicon. Semantics of formats and their co-ordination with an ontology and a lexicon is discussed.

# ЧЕЛОВЕЧЕСКИЙ ФАКТОР

# HUMAN FACTOR

# Context-Based Retrieval in Digital Libraries: Approach and Technological Framework ♣

© Kurt Sandkuhl[1], Alexander Smirnov[2], Vladimir Mazalov[3], Vladimir Vdovitsyn[3],
Vladimir Tarasov[1], Andrew Krizhanovsky[2], Feiyu Lin[1], Evgeny Ivashko[3]

[1] School of Engineering at Jönköping University,
{kurt.sandkuhl, vladimir.tarasov, feiyu.lin}@ jth.hj.se
[2] St.Petersburg Institute for Informatics and Automation of the RAS (SPIIRAS)
{smir, aka}@iias.spb.su
[3] Institute of Applied Mathematical Research of the KarRC RAS (IAMR)
{vmazalov, vdov, ivashko}@krc.karelia.ru

## Abstract

Digital libraries face similar challenges as enterprise information sources and the Internet: a fast growing amount of digital content requires enhanced ways of supporting information seeking. This paper presents an approach to context-based retrieval in Digital Libraries (DLs). The proposed approach includes creation of a profile representing general information demand of a user (abstract context), and use of ontology matching to identify the documents relevant to the operational context representing the current information demand of the user. A profile represents the user's interests as a DL reader and after creation is dynamically updated based on the changes in the user's interests. The identification of documents relevance is carried out by matching the user profile ontology against the digital library ontology. Semantic distance calculation is based on the use of a thesaurus.

## 1 Introduction

This paper aims at contributing to an improved relevance of results retrieved from digital libraries by proposing a conceptual framework for context-based retrieval. Digital libraries face similar challenges as enterprise information sources and the Internet: a fast growing amount of digital content requires enhanced ways of supporting information seeking. Capturing and exploiting preferences and other information about a user's information demand have been proposed as one contribution addressing this challenge. The use of context information has been found promising for this purpose.

One of the goals of context-based retrieval in DLs is to assess the relevance of documents for user needs [1, 3]. Nowadays, user faces problems of management and sharing of huge amount of documents saved in the DLs. The work presented in this paper proposes methodology and technological framework allowing the user to be provided with a set of relevant documents based on context-based retrieval. The paper concentrates on formalizing information demand of the user by profiling and matching the profile against an ontology describing documents in the digital library. The purpose to be achieved in our approach is an access of the user to the documents that are considered to be relevant for him/her in a particular situation (context).

Prerequisites for the approach are an ontological model describing typical interest of a DL user and a set of available DLs as document sources. The approach proposes a methodology assuming three stages. The first stage aims at creation of a context representing the user's information demand. The context is dynamically updated during the second stage. The third stage focuses on identification of documents relevant to the context (user needs). At this stage, the user profile ontology is matched against the digital library ontology. During matching, semantic similarity between the context and the shared ontology fragments is determined. Metrics of semantic similarity, used for comparison of semantically related words, similar ontologies, etc., are addressed in a number of algorithms, like HITS algorithm [6] for searching Internet pages using a structure of hyperlinks, PageRank algorithm [2], etc.

The paper is structured as follows. The introduction is followed by earlier work on understanding and capturing information demand with context models. A brief description of the overall framework is given in chapter 3. The next section describes the procedure of identification of relevant documents based on formalized context. The last chapter presents concluding remarks.

**Proceedings of the 11th All-Russian Research Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» - RCDL'2009, Petrozavodsk, Russia, 2009.**

## 2 Information Demand

The background for the proposed framework for context-based retrieval is earlier work on understanding and capturing information demand with context models. The notion of information demand is closely related to work in two areas: information logistics and information retrieval. Information retrieval aims at retrieving relevant information meeting the needs of a user, which are expressed by a query. In this context the aspect of relevance of information is of high importance. Saracevic [11] considers several types of relevance, e.g. algorithmic, topical and cognitive relevance. The underlying concept for algorithmic relevance is the relation between the query features and the search result. Topical relevance is the relation between aboutness of content objects and query. These two relevance concepts are important for retrieving information meeting the demand of a user, but do not contribute to explaining information demand as a concept. Here, we consider cognitive relevance of higher importance, which is the association between perceived information need of the user and information presented to the user based on retrieval results.

The main objective of the research field information logistics is improved information provision and information flow [7]. This is based on information demands with respect to the content, the time of delivery, the location, the presentation and the quality of information. The research field information logistics explores, develops, and implements concepts, methods, technologies, and solutions for the above mentioned purpose. A core subject of information logistics is how to capture the needs and preferences of a user in order to get a fairly complete picture of the demand in question. Principal approaches for this purpose are user profiles, situation-based and context-based demand models.

User profiles have been subject to research in information systems and computer science since more than 25 years. User profiles are usually created for functionality provided by a specific application. They are based on a predefined structured set of personalization attributes and assigned default values at creation time. Adaptation of such profiles requires an explicit adjustment of the preference values by the user. A situation-based approach was proposed for implementing demand-oriented message supply. The basic idea is to divide the daily schedule of a person into situations and to determine the optimal situation for transferring a specific message based on the information value. This approach defines a situation as an activity in a specific time interval including topics and location relevant for the activity. Information value is a relation between a message and a situation, which is based on relevance of the topics of a message for the situation, utility of the message in specific situations and acceptance by the user. Details and examples from collaborative engineering are given in [9, 10].

A context-based approach was proposed for use in enterprises or networked organizations. The basic idea is that information demand of a person in an enterprise to a large extent depends on the work processes this person is involved in, on the co-workers or superiors of this person and on the products, services or machines the person is responsible for. This led to the proposal to capture the context of information demand [8], i.e. a formalized representation of the setting in which information demand exists, including the organizational role of the person under consideration, work activities, resources and informal information exchange channels available.

## 3 Context-based retrieval in DLs

The framework of the context-driven retrieval is based on the use of an ontology-based model of the digital library (DL) and user profile representing typical information demand of the user. The documents to be found in DL are described through a user request expressing the current information need [14]. The retrieval request is modeled by two types of contexts: abstract and operational. *Abstract context* is an ontology-based model integrating knowledge on a general DL user and information about typical preferences of a particular user. *Operational context* is concretization of the abstract context based on the current need provided by the user's information request. Operational context is used by an ontology matcher for matching it against the ontology representing DL resources to find relevant documents (Figure 1).

Knowledge to be integrated in the abstract context is stored in an ontology describing the model of a general DL user that is combined with preferences of a particular user. The preferences are based on the initial information provided by the user to create a profile and information obtained dynamically by tracing the user's activities. The latter is used to keep the profile up-to-date and to assign weights to user preferences. The request data are used to identify the user profile fragment that corresponds to the current information need. The resulting operational context is an ontology fragment (or slice). A DL collection is represented with three types of ontologies: a document ontology, DL ontology, and shared ontology. A document ontology represents content of an individual document; a DL ontology formalizes content of all the documents stored in the DL; the shared ontology integrates the ontologies for all the libraries in the DL collection.

The ontologies can be created either manually by experts or in a semi-automatic way (e.g., [5]). According to the selected formalism, ontology A is defined as:

A = <O, Q, D, C>,

where:

O – a set of object classes ("classes")

Q – a set of class attributes ("attributes");

D – a set of attribute domains ("domains");

C – a set of constraints used to model relationships occurring in ontology representation formats / languages.

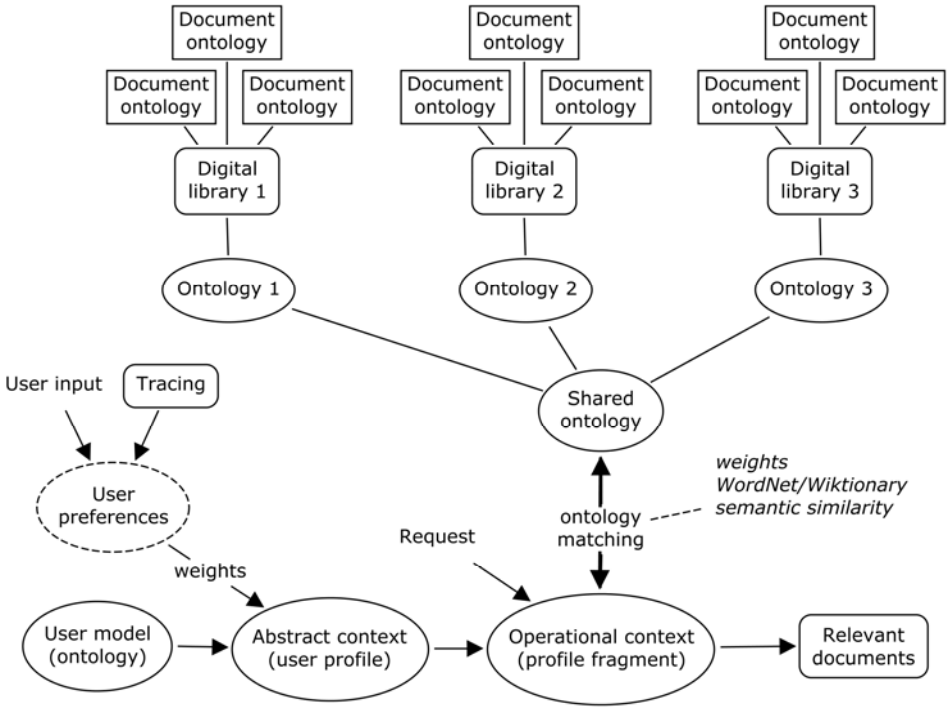



Figure 1. Conceptual framework of context-driven retrieval in DLs

For the DL user, the abstract context is a representational model of the user in a general view and the operational context is a current request description. For providing the DL users with documents relevant to the current request, a methodology and technological framework for context-driven retrieval in DLs have been developed. These methodology and technological framework have been developed within the conceptual framework of the context-driven information integration.

When abstract and operational contexts have been produced (i.e. the current situation has been described), the DL documents relevant to the current request are identified through ontology matching. WordNet and Wiktionary are used for improvement of semantic similarity algorithms during ontology matching. The resulting set of found documents with weights is presented to the user.

## 4 Context-based identification of relevant documents

In order to organize efficient identification of documents relevant to the current user request, the approach to context-sensitive access to information sources includes three stages (Figure 1): 1) creation of a profile representing general information demand of a user (abstract context); 2) updating the profile to adequately represent the current information demand of the user; and 3) use of ontology matching to identify the documents relevant to the operational context representing the current information demand of the user. A user profile is first created by ontological modeling of a DL user and then updated by behavioral modeling.

### 4.1 Constructing a profile of a DL user

Construction of a profile for a DL user draws upon our earlier work in competence modeling [15]. A competence model formalizes a person's skills and abilities, which are important for a certain task or situation. Competence models can be represented with ontologies. In a situation of document retrieval, the focus is on representation of the user's interests as a DL reader. These "reader's interests" can be described through professional interests and/or work role of the person in an organization.

As an example we can consider a user of a DL, which consists of scientific resources and aims at supporting workers in a research-oriented organization. A typical researcher would work in a research institute or university and would like to find documents relevant to their research interests. Hence, the main task is to represent research interests of the person. To do this, a user profile can be built based on papers published by the researcher and projects the researcher participated in. Each research paper/project can in turn be characterized by research fields relevant to the content of the paper/project. Additional research fields can be added by directly listing major research interest of the person and specifying fields connected to the scientific degrees or position.

To formalize research fields, different scientific taxonomies can be reused like the 1998 ACM Computing Classification System [16] or the Semantic Web Topic Hierarchy [12] in case of computer science.

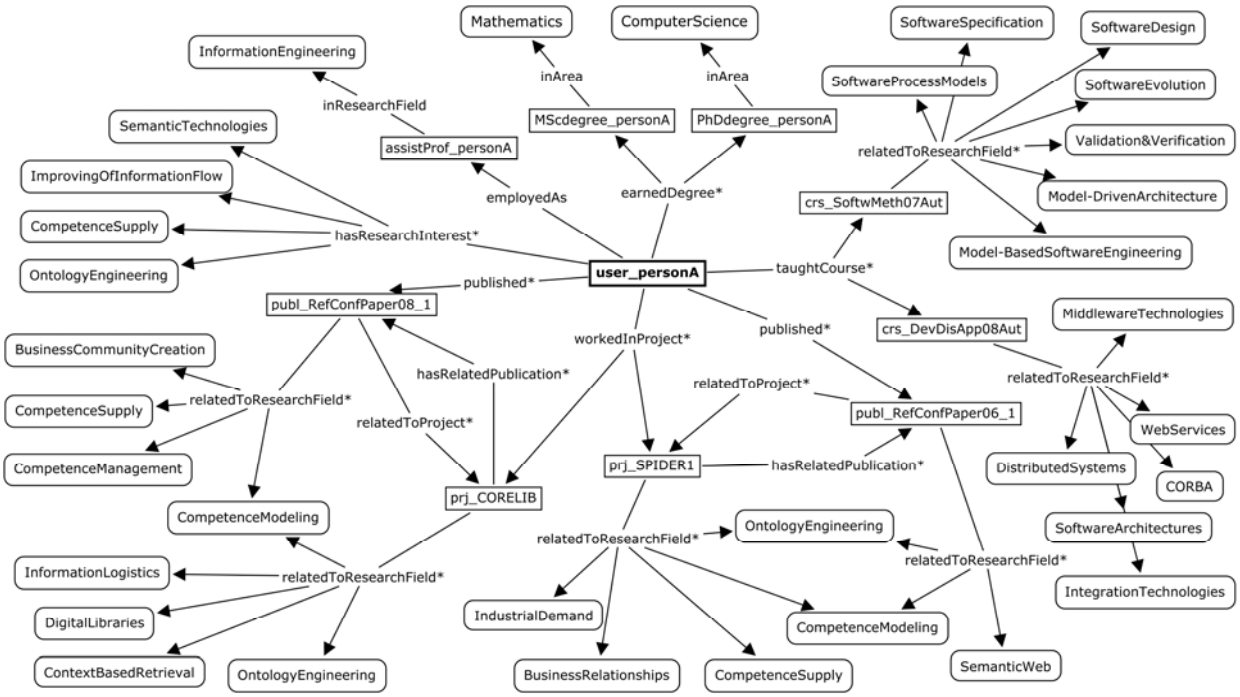



Figure 2. Example of a user profile for person A

If the person is engaged in teaching, then it is reasonable to include teaching topics in the model too. They can be included by connecting each taught course to the relevant research fields.

An example of a user profile is shown in figure 2. The profile includes research fields connected to two research papers, two projects, and two courses. Such a profile can be built manually or semi-automatically based on the general user information as well as the list of papers/projects for this person. Upon receiving a request from the user, the request can be mapped against the profile to identify the part (fragment) of the profile that corresponds to the user request.

### 4.2 Updating the user profile with behavioral modeling

After creation of a user profile describing topics of professional interest, the next task is to dynamically update the profile based on the changes in the user's interests. To trace these changes, a behavioral model is created representing current documents retrieved by the user. The behavioral model is constantly changed according to the user's activities. When there are many enough changes in the behavioral model, the user profile is updated to reflect the changes in the user's interests.

A behavioral model is built with the help of Markov chains [4]. The process includes tracing user searches to build a graph, constructing a classifier, and setting parameters. Using the profile shown in Figure 2, let us consider a possible behavioral model with respect to the project prj_CORELIB. Figure 3 depicts a simple example of a behavioral model, where each node represents transition from one topic of interest to another one. Every node is marked by the number of documents matching the appropriate attribute domain. Each topic can also be weighted based on the usage: access frequency, access date, etc. The transitions are extracted from a number of retrievals made by the user.

The example shows the difference between the original user's profile (see Figure 2, the part corresponding to prj_CORELIB) and the behavioral model (see Figure 3). The behavioral model related to prj_CORELIB contains two additional domains: CompetenceManagement and IndustrialDemand. But the domain IndustrialDemand appears only two times (this could show a weak interest that is it is not useful for the user in the context of prj_CORELIB) and CompetenceManagement – 6 times. This leads to the necessity to update the user's profile by adding the CompetenceManagement domain to the relation relatedToResearchField* (see Figure 4).

Thus, the creation of a behavioral model allows updating the user's profile according to the changes in the user's interests and therefore satisfying the user's current information demand.

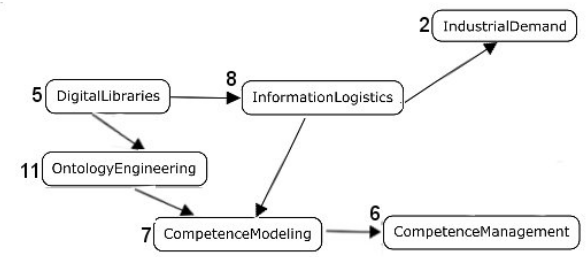



Figure 3. Example of a behavioral model of the user

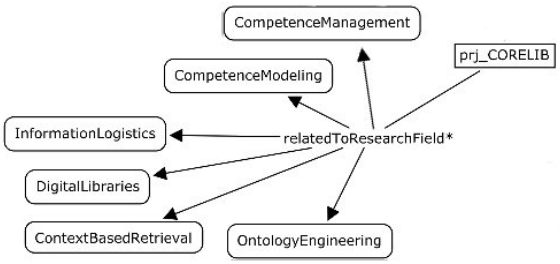



Figure 4. The updated part of person A's profile

### 4.3 Ontology matching based on WordNet and Wiktionary

Identification of DL documents relevant to the abstract and operational contexts is carried out at the stage of ontology matching. Since the current request is represented through the operational context, the aim is to sort documents, which are considered to be corresponding to the context, by relevance. Determining relevance is supposed to be based on measurement of similarity between the context and the shared ontology fragments, elements of which describe DL documents.

*Ontology matching*. A user profile is an ontology representing user preferences in terms of professional topics of interest and documents recently accessed. A digital library ontology describes the topics of documents stored in the library and relations between these topics. Hence, every document has one or several keywords or categories (like categories of wiki pages), which connect the document to the digital library ontology. After matching the user profile ontology against the digital library ontology, it is possible to predict potentially useful documents, which belong to the area of the user's interests.

This process consists of three steps:

1. Matching between the user profile ontology and digital library ontology. The result of this step is list *A* of entities of the digital library ontology corresponding to the user profile ontology.
2. "Closure" algorithm to find list *B*, which conforms to these three conditions:
    a. Entities of list *B* belong to the digital library ontology.
    b. Entities of list *B* do not belong to list *A*.
    c. Entities of list *B* are the closest elements of all the elements of the ontology to the entities of list *A*.
3. Enumerating a list of documents of interest for the user, which correspond to the entities of list B.

A set of ontology matching algorithms is based on thesauri (e.g., WordNet [17]). The comparison of different algorithms based on WordNet can be found in [18].

*Semantic similarity*. There are a lot of algorithms for semantic similarity, which are used for ontology matching. There is the following classification of ontology matching algorithms: internal and external [13]. An internal ontology matching algorithm exploits information that comes only with the source ontologies. An external ontology matching algorithm exploits external resources such as a domain ontology, corpus, thesaurus (e.g., WordNet, Wiktionary).

The Russian Wiktionary (the dump of the database as of January 2009) was parsed and the results were stored in a relational database (MySQL). Hence, the database of the parsed Wiktionary is the source data in the experiment [18].

The database of the parsed Russian Wiktionary has a better coverage than WordNet (247,580 words against 150,000). At the same time, WordNet consists of over 115,000 synsets while the total number of semantic relations in the database of the parsed Wiktionary is about 67,000 at this moment.

The experiment in [18] shows that the proposed method (Figure 5) is, in principle, capable of calculating a semantic distance between a pair of words in any language presented in Wiktionary (more than 200 in Russian Wiktionary). The comparison semantic distance between ontologies based on WordNet and Wiktionary raises an interesting question: whether the joint usage of Wiktionary and WordNet can improve calculation of the relatedness measure. This comparison is presented in [18].

## 5. Conclusions

The paper presents an approach to context-sensitive access to DL to help to identify documents relevant to a context (current situation). Capturing and exploiting preferences about a user's information demand have been proposed as one contribution. The approach includes three stages: (i) creation of a profile representing general information demand of a user (abstract context), (ii) dynamically updating the profile with behavioral modeling, and (iii) use of ontology matching to identify the documents relevant to the operational context representing the current information demand of the user.

The purpose of profiling (first stage) is to create a user profile by ontological modeling of a DL user. A profile represents the user's interests as a DL reader such as topics of professional interest and/or work role of the person in an organization. After creation the profile is dynamically updated based on the changes in the user's interests during the second stage. To trace these changes, a behavioral model is created representing current documents retrieved by the user.

The third stage focuses on identification of documents relevant to the current request (information demand of the user). The identification is carried out by matching the user profile ontology against the digital library ontology. The documents, which belong to the area of the user's interests, are sorted by relevance. Determining relevance is based on measurement of similarity between the context and the shared ontology fragments, elements of which describe DL documents.

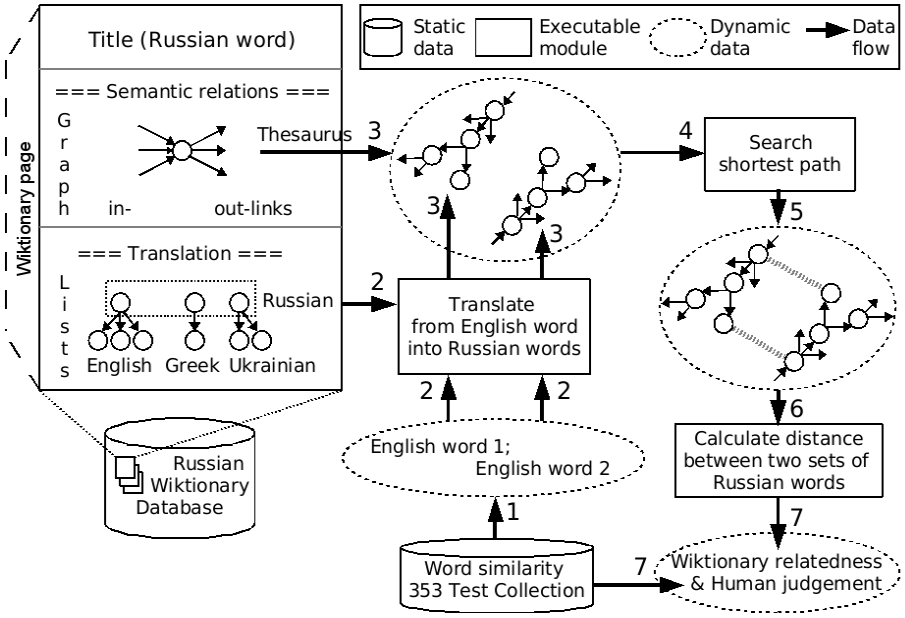



Fig. 5. Scheme of the experiment for calculating the semantic relatedness measure based on the Russian Wiktionary data

The shared ontology integrates the ontologies for all the libraries in the DL collection.

Our future work will focus on experiments with the proposed approach in a real setting. After initial modeling of several user profiles, experiments will be conducted in matching the digital documents against the user profiles by suing the semantic relatedness measure. The initial work in this direction has been started [18]. Further work with behavioural modelling to dynamically update user profiles is also an interesting direction that can enhance context-sensitive access to documents in DLs.

## References


[1] B. Aleman-Meza, P. Burns, M. Eavenson, D. Palaniswami, and A. P. Sheth, "An Ontological Approach to the Document Access Problem of Insider Threat", *IEEE International Conference on Intelligence and Security Informatics (ISI-2005)*, Atlanta, Georgia, USA, 2005.

[2] S. Brin, and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", 1998. URL: http://www-db.stanford.edu/~backrub/google.html

[3] M. Ehrig, and A. Maedche "Ontology-focused crawling of Web documents", *Proc. of the 2003 ACM symposium on Applied computing,* Melbourne, Florida, 2003

[4] S. Jha, K. Tan, R.A. Maxion. Markov Chains, Classifiers and Intrusion Detection. In the proceedings of Computer Security Foundations Workshop (CSFW), June 2001.

[5] V. Kashyap, C. Ramakrishnan, C. Thomas, and A. Sheth, "TaxaMiner: An Experimental Framework for Automated Taxonomy Bootstrapping", *International Journal of Web and Grid Services, Special Issue on Semantic Web and Mining Reasoning*, Inderscience, 1 (2), 2005, pp. 240-266.

[6] J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, Association for Computing Machinery (ACM), 46 (5), 1999, pp. 604-632. URL: http://www.cs.cornell.edu/home/kleinber

[7] Kurt Sandkuhl: Information Logistics in Networked Organizations: Selected Concepts and Applications. Enterprise Information Systems, 9th International Conference, ICEIS 2008. LNBIP, Springer.

[8] Lundqvist, M. (2005). Context as a Key Concept in Information Demand Analysis. In Proceedings of the Doctoral Consortium associated with the 5th Intl. and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-05), 63-73. Paris, France.

[9] Meissen U., Pfennigschmidt, S., Voisard, A., and Wahnfried, T. (2004). Context- and situation-awareness in information logistics. In Postproceedings of Workshops of the International Conference on Extending Database Technology (EDBT), LNCS, Berlin/Heidelberg, Springer.

[10] Meissen, U., Pfennigschmidt, S., Sandkuhl K., Wahnfried, T. (2004a) Situation-based Message Rating in Information Logistics and its Applicability in Collaboration Scenarios. In Euromicro 2004 Special Session on "Advances in Web Computing", August 31- September 3, IEEE Computer Society Press.

[11] Saracevic, T.: Relevance Reconsidered r96. In Ingwersen, P.; Pors, N. O. (eds.): Information Science: Integration in Perspective. Royal School

of Library and Information Science, Copenhagen, Denmark, pp. 201-218, 1996.
[12] Semantic Web Topic Hierarchy, 2008. http://semanticweb.org/wiki/Semantic_Web_Topic_Hierarchy
[13] P. Shvaiko and J. Euzenat, 'A survey of schema-based matching approaches', Journal on Data Semantics, (IV), 146–171, (2005).
[14] A. Singhal "Modern Information Retrieval: A Brief Overview", *In IEEE Data Engineering Bulletin*, 24 (4), 2001, pp. 35-43. URL: http://singhal.info/publications.html
[15] Tarasov, V., Lundqvist, M. (2007). Modelling Collaborative Design Competence with Ontologies. In International Journal of e-Collaboration (IJeC): Special Issue on the State of the Art and Future Challenges on Collaborative Design, ISSN 1548-3673. IGI Publishing: Hershey, USA. Pages 46-62.
[16] The 1998 ACM Computing Classification System, 2009. http://www.acm.org/about/class/1998
[17] WordNet, 2006) "WordNet", 2006. URL: http://wordnet.princeton.edu
[18] Krizhanovsky, A. and Lin, F., 'Exploiting WordNet / Wiktionary in Ontology Matching', submitted to RCDL2009, Petrozavodsk, Russia.

## Технология поиска в электронных библиотеках, основанная на контексте

К. Сенкюль, А.В. Смирнов, В.В. Мазалов,
В. Т. Вдовицын, В.В. Тарасов,
А.А. Крижановский, Ф.Лин, Е.Е. Ивашко

Электронные библиотеки в настоящее время сталкиваются с теми же проблемами, что и информационные системы предприятий, а также Интернет: быстро растущее количество электронных документов требует более совершенных методов поиска. В данной статье представлена технология поиска в электронных библиотеках (ЭБ), основанная на контексте. Предложенный подход предусматривает создание профиля, представляющего общие информационные потребности пользователя ЭБ (абстрактный контекст), и применения сопоставления на основе онтологии для распознавания документов, соответствующих операционному контексту, представляющему текущие информационные потребности пользователя ЭБ. Профиль представляет интересы пользователя как читателя ЭБ и после создания динамически обновляется на основе изменений интересов пользователя. Определение степени соответствия документов контексту выполняется через сопоставление онтологии профиля пользователя и онтологии ЭБ. Вычисление семантического расстояния основано на использовании тезауруса.

♣ The paper is based on research carried out as a part of several projects: project CoReLib supported by the Swedish Institute by grant # 01215-2007 and projects funded by grants of the Russian Foundation for Basic Research.

# Исследование пользовательских предпочтений для контроля и оптимизации Интернет-трафика в организации

© Леонова Ю.В., Федотов А.М.

Институт вычислительных технологий СО РАН
juli@ict.nsc.ru

**Аннотация**

В статье рассмотрены вопросы, связанные с оптимизацией потребления интернет-трафика и повышения эффективности работы интернет-канала, описаны проблемы управления информационными ресурсами и их защиты, основные принципы и понятия, методика поэтапного сокращения затрат на Интернет-трафик и потерь рабочего времени, связанных с нецелевым использованием сети Интернет, на примере научно-образовательной сети ННЦ СО РАН.

## Введение

В крупных городах многие провайдеры уже предоставляют возможность безлимитного использования интернет-канала за фиксированную сумму, однако стараются ограничивать либо пропускную способность предоставляемого канала, либо требуют соблюдения некоторых условий, направленных на то, чтобы удержать потребление клиентом интернет-трафика в определенных рамках. Кроме того, цены на безлимитный доступ в большинстве случаев все еще "кусаются".

Помимо стоимости услуг интернет-провайдеров, отдельно стоит вопрос состава потребляемого трафика. Редкий руководитель, подписывая очередной счет, не задавался вопросом: а за что, собственно, платятся деньги? Действительно ли оправданы затраты и есть ли способ их снижения?

Ниже мы рассмотрим вопросы оптимизации потребления интернет-трафика и повышения эффективности работы интернет-канала.

## 1 Проблемы управления информационными ресурсами и их защиты

Одной из особенностей Интернета является то, что на определенном этапе он развивался стихийно. Это, с одной стороны, обеспечило массовый характер его использования, а с другой — породило ряд проблем с серьезными последствиями.

1) Поскольку Интернет является каналом во внешний мир, он стал основным источником распространения вредоносного мобильного кода (вирусов, червей, троянских программ).

2) Глобальная сеть стала использоваться в качестве канала, через который осуществляются атаки на локальные вычислительные сети организаций, отдельные серверы и компьютеры. Многие Интернет-ресурсы включают в себя различный программный код — JavaScript, Flash, ActiveX и другие. Злоумышленники могут эксплуатировать этот код для организации атак на корпоративные сети и пользовательские рабочие места.

3) В настоящее время Интернет может рассматриваться как один из основных каналов утечки конфиденциальной информации. Например, информационные ресурсы компаний подвергаются серьезным угрозам из-за использования сотрудниками этих компаний бесплатных почтовых ящиков. Многие сотрудники различных компаний помимо внутренних корпоративных почтовых адресов активно используют бесплатные почтовые ящики, предоставляемые различными провайдерами. Имея доступ к Интернету со своего рабочего места и зная, что канал не контролируется, любой пользователь может беспрепятственно отправить за пределы организации любую конфиденциальную информацию. Но, даже понимая это, не каждая компания запрещает использовать бесплатные почтовые сервисы, тем самым, позволяя своим сотрудникам решать, как и какую информацию отправить за пределы компании.

4) Бесконтрольный доступ к Интернету значительно снижает производительность труда в коллективе. Простота освоения, легкость поиска необходимой информации и другие полезные качества Интернета — вот причины того, что данный сервис широко применяется, в том числе и для личных целей. Не секрет, что у многих уже давно появилась привычка начинать рабочий день с чтения новостей, просмотра сводок погоды и т.п.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

Сотрудники различных организаций и компаний используют Интернет в целях, не имеющих прямого отношения к их работе. Это и "походы" в Интернет-магазины, и сетевые игры, и просто поиск информации.

5) Наконец, еще одно следствие неконтролируемого использования Интернета — это снижение пропускной способности сети. Сотрудники организаций используют корпоративные ресурсы для просмотра видео, прослушивания аудиозаписей (через потоковые аудио- и видеоканалы), играют в сетевые игры, загружают файлы большого объема (например, файлы мультимедиа: графические, музыкальные файлы, фильмы и т.п.), что создает значительную нагрузку на локальные вычислительные сети.

Таким образом, проблемы управления информационными ресурсами вычислительных сетей и их защиты становятся все более актуальными для организаций.

# 2 Контроль и оптимизация Интернет-трафика

## 2.1 Аудит сети

Первым шагом для решения вышеперечисленных проблем является аудит сети организации, что позволит выявить "дыры" и узкие места в компьютерной системе организации, в том числе даст картину потребления интернет-трафика. В результате аудита можно получить не только данные о том, что происходит в сети организации, в каком состоянии находятся ее ресурсы, какова структура интернет-трафика, но и о том, чем конкретно занимаются сотрудники на рабочих местах.

## 2.2 Контроль доступа

Следующим шагом по сокращению затрат на интернет-трафик может стать контроль доступа к Интернет-ресурсам, который можно решить двумя способами. 1) Запрещение использования Интернета без необходимости, когда пользователям разрешается доступ только к строго определенным сайтам. 2) Контроль действий сотрудников, при этом сотрудник может свободно пользоваться ресурсами Интернета. Но если пользователь выполнит действия, противоречащие политике безопасности, это будет обнаружено и пресечено. Второй способ контроля является наиболее гибким и более распространенным, но именно при его применении возникают существенные проблемы, которые состоят в том, что практически невозможно однозначно определить, к какой информации следует запретить доступ. Необходимой составляющей решения этих проблем является разработка и внедрение политики безопасности сети и политики использования ресурсов.

### 2.2.1 Политики безопасности

Политика безопасности сети – это набор законов, правил и практических рекомендаций, на основе которых строится управление, защита и распределение защищаемых информационных ресурсов. Она должна охватывать все особенности процесса использования информационных ресурсов сети организации, определяя поведение системы в различных ситуациях. Ключевым шагом разработки политики безопасности является определение критичных для организации ресурсов и возможных угроз доступности, конфиденциальности и целостности этих ресурсов. При этом может применяться несколько подходов, в том числе ранжирование сетевых ресурсов по их стоимости, по вероятности реализации угроз и по серьезности их последствий для организации. Последняя не всегда связана с раскрытием конфиденциальной информации или выходом из строя дорогостоящих устройств, но также снижением производительности ресурсов, которые активно используются сотрудниками организации при исполнении служебных обязанностей. Поэтому выявление таких сетевых информационных ресурсов является важной задачей при разработке политики безопасности.

Существует несколько решений этой задачи, но наиболее эффективным представляется прослеживание истории сетевых взаимодействий путем накопления и анализа статистики обращений к сетевым серверам и предоставляемым ими сервисам. Сохраняя в базе данных структурированную информацию по сетевому трафику, извлеченную из заголовков передаваемых по сети пакетов, и обрабатывая накопленные данные с помощью автоматизированных алгоритмов анализа, можно составить четкую схему использования информационных ресурсов сети, необходимую для формирования сбалансированной политики безопасности сети, без выполнения трудоемкой и рутинной ручной работы. При этом рассмотренный механизм позволяет не только учесть сетевые взаимодействия внутри организации, но и обращения к информационным ресурсам из внешних по отношению к организации источников, в частности удаленный доступ к ресурсам сети, и таким образом выявить источники потенциальных угроз.

### 2.2.2 Политика использования Интернет-ресурсов

Для обеспечения гибкости контроля использования Интернет-ресурсов в организации вводится политика использования ресурсов. Эта политика может реализовываться на основе анализа и фильтрации веб-трафика. На сегодняшний день существует множество как коммерческих, так и некоммерческих решений. К наиболее распространённым коммерческим продуктам можно отнести: open-source систему Poesia [4],

коммерческие системы CyberPatrol [2], SurfControl [5], NetNanny [3] и др.

Можно выделить два основных признака систем фильтрации и анализа трафика – способ и время анализа трафика. По способу анализа все системы можно разбить на два больших класса: 1) анализирующие лишь общую (мета-) информацию о ресурсе; 2) анализирующие в том числе и содержимое (контент) ресурса.

По времени анализа все системы можно также разбить на два класса: 1) анализирующие информацию в реальном времени (онлайн), т.е. во время запроса пользователем Интернет-ресурса; 2) анализирующие информацию в отложенном режиме (оффлайн), т.е. после того, как пользователь получил доступ к ресурсу.

В данной работе рассматриваются системы масштаба локальных сетей, анализирующие метаинформацию в отложенном режиме.

Применение системы контроля использования Интернет-ресурсов нельзя представить без анализа событий, происходящих в системе. Администраторам необходимо оперативно получать информацию о текущем состоянии системы, а также сводные отчеты об использовании Интернет-ресурсов пользователями или группами пользователей. Такая информация позволяет не только контролировать использование Интернет-ресурсов, но и проверять эффективность политики безопасности и динамически адаптировать ее к изменяющимся условиям и задачам. Поэтому в большинстве существующих средств контроля использования Интернет-ресурсов есть возможность формирования статистических отчетов, а также интерактивного наблюдения за доступом к внешним ресурсам.

Существует несколько способов представления статистических данных о трафике. Первый способ предполагает использование внутренних возможностей продукта, то есть встроенной системы генерации отчетов. Как правило, в состав такой системы входят подсистема генерации отчетов и база данных, в которую в виде журналов записывается вся информация о событиях, а также некоторые запросы пользователей (например, команды POST). С помощью внутренних средств производятся SQL-запросы к базе данных, результаты которых дают наглядную картину трафика и действий пользователей. При этом могут создаваться типовые запросы (например, "100 часто загружаемых сайтов", "100 пользователей, переславших наибольшие объемы данных за указанный период", "100 самых активных пользователей" и т.п.) с изменяемыми параметрами (например, по дате и времени). Другой способ предполагает получение отчетов с помощью стандартных средств, таких как Crystal Reports, Oracle Reports и т.п. Эти средства интегрируются с системой контроля использования Интернет-ресурсов и тоже используют базу данных, которая создается в результате фильтрации трафика.

### 2.3 Установка кэширующих серверов и зеркал

Еще одним способом оптимизации интернет-трафика является использование кэширующих серверов и системы зеркал, на которые в «прозрачном» для конечного пользователя режиме перенаправляются HTTP-запросы пользователей. Использование кэширующих серверов и системы зеркал преследует две основные цели:

- Улучшение производительности: снижение нагрузки на каналы провайдера, используемые для выхода в интернет и уменьшение времени ожидания загрузки данных для пользователей.
- Сокращение затрат: размер трафика на канал в интернет после установки кэширующих серверов и зеркал уменьшится, что приведет к снижению платежей за передачу информации по этим каналам.

1) Создание системы кэширования интернет-трафика позволяет увеличить пропускную способность канала связи, одновременно снизив среднее время ожидания ответа на запрос пользователя. Кэширование минимизирует задержки при передаче файла, примерно в 5-200 раз. Суть кэширования www-трафика состоит в том, что запрос пользователя на получение документа перенаправляется на кэш-сервер, который сначала проверяет наличие документа в своем кэше, после чего продолжает обслуживание запроса. Если документ в кэше не найден, то кэш-сервер направляет запрос на сервер-источник документа или другому кэш-серверу.

Система кэширования с иерархической сетью создана, например, в Научном центре в Черноголовке, объединенной с кэш-серверами в Ярославле, Перми, Челябинске, МНФ (Москва), ИТФ им. Л.Д.Ландау (Москва), ИОХ им. Зелинского (Москва) [8].

2) Развертывание зеркал позволяет разместить наиболее востребованные данные «ближе» к пользователю. Зеркала обеспечивают максимальную скорость передачи данных от сервера к пользователю – при запросе файла с веб-сайта, его передает локальное зеркало. Под зеркалированием интернет ресурсов понимается создание полных или частичных копий (зеркал) этих ресурсов на географически удаленных серверах, обновление которых может производиться во время минимальной загрузки каналов, например, ночью. Прозрачное перенаправление на зеркало (незаметное для пользователя) реализуется посредством редиректора, который выполняет первоначальную обработку URL, и либо возвращает прежний URL для дальнейшей обработки прокси-серверу в случае, если все в порядке, либо возвращает тот, который, по его мнению, более правильный.

Естественно, что сам процесс зеркалирования создает определенную нагрузку на центральный

сервер и каналы связи (порой сравнимую, а иногда и превышающую выигрыш от зеркал). Зеркалирование увеличивает общую сложность системы (проблемы с администрированием, распределением прав, увеличением технического парка и т.д.). Можно утверждать, что необдуманное внедрение зеркал приведет к негативному результату. С другой стороны зеркала могут ощутимо повысить надежность и общую производительность системы.

Система зеркал была реализована при создании региональной научно-образовательной сети в Интернет центре Новгородского государственного университета [6], что позволило уменьшить внешний трафик организации.

В рамках данной работы было проведено исследование статистики обращений к веб-серверам, на основе которого рассмотрены вопросы, связанные с оптимизацией трафика и ускорением работы Интернет, в том числе задание и приложение правил обслуживания и учета трафика HTTP прокси-сервером, а также задание и реализация политики безопасности. Выработка и внедрение корпоративной сетевой политики, основной принцип которой сводится к тому, что пользователи научно-образовательной сети работают в первую очередь с научно-образовательной информацией.

## 3 Исследование Web-трафика

С момента своего появления технология веб стала предметом исследований [1]. Основная цель большинства исследований в вебе – это поиск таких свойств трафика, которые позволят совершенствовать саму технологию, увеличить скорость передачи информации к пользователю, уменьшить время загрузки нужного документа. Повышенный интерес к исследованиям веб-трафика вызван тем, что в настоящее время веб-трафик доминирует в общем трафике всех компьютерных сетей.

В одной из первых работ по исследованию веб-трафика [7] было замечено, что популярность документов в вебе распределена очень неоднородно. Большинство запросов приходят на очень небольшое количество документов, в то время как многие документы запрашиваются всего несколько раз. Для описания свойств популярности веб-документов очень удобно использовать технику ранговых распределений.

Рассмотрим информацию, которую можно получить на основе анализа логов прокси-сервера.

1. Информационный ресурс. Информационный ресурс представляет собой совокупность информационных объектов. Основные параметры информационного объекта: тип информации — текст, изображения, аудио-, видеоданные, потоковые данные, бинарные файлы, медиаданные; объем информации; в) приемлемая скорость доступа к объекту; полезность; частота модификации; потребность в объекте; права доступа на объект.
2. Потребитель ресурса — пользователи или
3. компьютеры. Основные параметры потребителя ресурса: текущее и потенциальное количество пользователей ресурса; интенсивность запросов к каждому информационному объекту, объем потребления, генерируемый трафик; удовлетворенность качеством доступа.
3. Канал передачи данных. Канал передачи данных между информационным ресурсом и потребителем. Основные параметры: полоса пропускания; загрузка канала (входящий/исходящий трафик); доля трафика ресурса в общей загрузке канала; стоимость работы по каналу.

Анализ полученной информации может быть использован для решения следующих задач.

1) **Оптимизация (уменьшение) трафика.** Как правило, наиболее «узким» местом является внешний канал научно-образовательной сети, когда большое количество пользователей одновременно работает с разнообразными Интернет-ресурсами и возникает перегрузка канала. Решение этой проблемы заключается в кэшировании и классификации наиболее важных и востребованных информационных ресурсов, например статей, с последующим размещением их для использования научным сообществом внутри локальной сети. В результате при уменьшении количества перекачек повышается надежность сети.

2) **Изучение информационных потребностей.** Данный анализ позволяет получить информацию о поведении пользователей локальной сети в Интернете, выявлять самых активных пользователей и смотреть, какие ресурсы они посещают, получать общее представление о распределении трафика по сайтам, дням недели и времени суток и многое другое. При обнаружении наиболее напряженных участков скачивания «важных» ресурсов может быть увеличена пропускная способность на данном направлении.

3) **Ограничение нецелевого использования.** Большой эффект по разгрузке канала дает ограничение трафика с нежелательным содержанием, например, порно-сайтов или развлекательных ресурсов типа «Одноклассники», различных «непрофильных» ресурсов аудио- и видео-серверов.

Установка или настройка существующих корпоративных прокси-серверов позволяет уменьшить внешний трафик организации и повышает качество работы с ресурсами. Для этого производится дополнительная настройка прокси-серверов: ограничивают доступ к непрофильным серверам; вводят ряд ограничений по пропуску типов файлов (avi, mp3 и т. д.); ограничивают пользователей по скорости доступа; при необходимости увеличивают размер кэша; при необходимости изменяют время хранения документов в кэше востребованных ресурсов.

| Обозначение | Кэш-сервер | Период | Число запросов | Число сайтов | Объем |
|---|---|---|---|---|---|
| ICT1 | proxy.ict.nsc.ru | 3 недели | 20,250,832 | 104,249 | 337,29 G |
| ICT2 | proxy.ict.nsc.ru | 2 недели | 11,402,797 | 63,190 | 240.09 G |
| NSU1 | proxy.nsu.ru | 2 недели | 34,040,909 | 113,324 | 276,52 G |
| NSU2 | proxy.nsu.ru | 2 недели | 32,908,553 | 121,999 | 253,42 G |

Таблица 1. Набор данных

### 3.1 Исследование наборов данных

Напомним основной принцип работы протокола HTTP. Для того, чтобы получить нужный документ, пользователь направляет запрос к веб-серверу, на котором находится этот документ. Веб-сервер в ответ возвращает пользователю требуемый документ. Кроме того, пользователь может посылать запрос не напрямую к веб-серверу, а на сервер-посредник, с которым у него имеется высокая скорость соединения (например, к прокси-серверу в его локальной сети). Веб прокси-сервер, как правило, имеет кэш и, если запрашиваемый документ находится в кэше прокси-сервера, то скорость получения этого документа значительно возрастает.

Таким образом, для исследователей имеется несколько способов получения информации о веб-трафике. Можно исследовать запросы, приходящие к отдельно взятым веб-серверам, можно собирать информацию о действиях отдельных пользователей или анализировать запросы пользователей к кэш-серверам. Основное отличие между этими способами состоит в том, что в первом случае мы получаем данные о трафике для очень небольшого подмножества веб, которым является множество документов на нескольких выбранных веб-серверах, а информация, полученная из машин пользователей или прокси-серверов, дает нам представление о трафике, создаваемым небольшой группой пользователей.

Поскольку нас интересовали исследование свойств для запросов второго вида, в качестве источника информации для исследования веб-трафика были взяты логи информация кеш-сервров. Для анализа использовались лог-файлы кэш-серверов сети ННЦ СО РАН: proxy.ict.nsc.ru (СО РАН) и proxy.nsu.ru (НГУ) – типичные прокси-серверы, обслуживающие запросы локальных пользователей организации. Мы проанализировали данные, собранные в течение одного месяца. Детальное описание наборов данных дано в таблице 1.

На рассматриваемых серверах установлено программное обеспечение Squid. Его лог представляет собой текстовый файл, в который записывается информация о всех запросах, поступивших на кэш-сервер. После получения очередного запроса кэш-сервер добавляет в лог-файл одну строку с информацией, характеризующей полученный запрос, например

1210274283.328 1010 194.226.177.55 TCP_HIT/200 67526 GET http://stats.iihf.com/Hydra/132/IHM132000_85K_6_0.pdf - DIRECT/80.231.19.71 application/pdf

Здесь 1210274283.328 обозначает время поступления запроса в формате UTC (Universal coordinated time), 1010 обозначает, сколько времени (в мс) заняла обработка запроса, 194.226.177.55 – IP адрес машины пользователя, пославшего запрос, TCP_HIT/200 – код результата выполнения запроса, 599 – размер запрашиваемого ресурса (в байтах), GET – метод протокола HTTP. Большинство запросов к кэш_серверу используют метод GET – метод получения нужного ресурса по протоколу HTTP.

http://stats.iihf.com/Hydra/132/IHM132000_85K_6_0.pdf – адрес запрашиваемого ресурса,

application/pdf – MIME-тип документа, в данном случае документ в формате PDF.

Далеко не все запросы пользователей к кэш-серверу благополучно им обрабатываются, например кэш-сервер может быть настроен таким образом, что он обрабатывает запросы только от определенной группы пользователей, а остальные запросы игнорирует. В другом случае, пользователь может сделать ошибку при вводе URL, запросить несуществующий документ или документ, для получения которого необходимо ввести пароль, который пользователь вводит неверно. Наконец, во время передачи данных может просто разорваться связь. Результат обработки запроса (код HTTP) кэш-сервер заносит в соответствующее поле лог-файла. Для того, чтобы достоверно судить о скачиваемости документов, мы будем анализировать только запросы, успешно обработанные кэш-сервером, имеющие код результата выполнения – 200. Таким образом, на втором этапе обработки данных мы оставляем только те записи в лог-файлах, у которых в поле результата выполнения запроса записано 200. На следующем этапе обработки данных мы выделяем из полей URL документа. Затем мы подсчитываем количество появлений каждого документа в лог-файле – $f_i$ и, сортируя документы по убывающим значениям $f_i$, мы получаем ранговое распределение популярности скачивания документов. Выделяя из поля URL название веб-сайта, аналогично можно определить популярность веб-сайтов. Аналогично определяется ранговое распределение объема скачивания документов. Далее для определения предпочтений пользователей выполняется категоризация полученных данных по областям деятельности в два этапа. На первом этапе категоризикация выполняется на основе классификатора каталога сайтов Яндекса.

| | | |
|---|---|---|
| ПК, Интернет, связь | Здоровье | Театры |
| Hardware | Мода и красота | Фотография |
| Интернет | Покупки | Музеи |
| Мобильная связь | Общество | Изобразительные искусства |
| Программы | Власть | Танец |
| Безопасность | Законы | Спорт |
| Сети и связь | НКО | СМИ |
| Интерфейс | Политика | Периодика |
| Работа | Религия | Информационные агентства |
| Учеба | Развлечения | Телевидение |
| Высшее образование | Игры | Радио |
| Курсы | Юмор | Бизнес |
| Среднее образование | Непознанное | Финансы |
| Школы | Личная жизнь | Недвижимость |
| Науки | Отдых | Строительство |
| Учебные материалы | Где развлечься | Производство и поставки |
| Дом | Туризм | Реклама |
| Квартира и дача | Хобби | Деловые услуги |
| Кулинария | Культура | Все для офиса |
| Все для праздника | Музыка | Справки |
| Семья | Литература | Транспорт |
| Домашние животные | Кино | Афиша |
| | | Авто |

Таблица 2. Классификатор Яндекса

Каталог Яндекса, содержащий описания сайтов русскоязычного интернета, систематизированных по тематическим категориям, построен на основе фасетной классификации. Такая классификация с одной стороны позволяет легко организовать поиск ресурсов не только по тематике, но и по типу информации, а с другой стороны предотвращает углубление рубрикатора и неоднозначность тематического отнесения ресурсов. На первом уровне дерева каталога имеется 13 тем, а число уровней в глубину не превышает четырех. Рубрики сгруппированы определенным образом. В первой группе темы «человек и его окружение»: дом, учеба, работа, общество, коммуникации. Вторая группа – «развлечения»: отдых, юмор, спорт, музыка и др.

Третья группа – «бизнес и экономика». Зато, помимо тем, в каталоге имеется ряд дополнительных признаков (фасет), позволяющих уточнить характер ресурсов, которые пользователь хочет увидеть в тематических категориях. Эти нетематические признаки характеризуют ресурсы по региону, сектору экономики, степени достоверности (источнику) информации, ее потенциальной аудитории (адресату информации), жанру (художественная литература, научно-техническая литература, и т. д.), цели (предложение товаров и услуг, интернет-представительство) и т. д.

Выше приведен классификатор Яндекса (табл.2).

Посредством программы-робота, запрограммированного на обход каталога сайтов Яндекса по различной глубине вложенности был сформирован классификатор, на основе которого выполнялась категоризация трафика по областям деятельности. На втором этапе выполняется категоризация оставшегося трафика на основе сигнатурного подхода, основанный на использовании экспертной базы знаний адресов Интернет-ресурсов. Такая база знаний содержит адреса ресурсов, с каждым из которых связан набор тем (категорий), к которым, по мнению экспертов, относится данный Интернет-ресурс. Для категоризации трафика был разработан классификатор доменных имен, с рубрикацией, аналогичной классификатору Яндекса. Полученные результаты исследования приведены на рис.1.

### 3.2 Обработка результатов

Можно заметить, что ранговое распределение предпочтений пользователей для различных кэш-серверов различаются, но близки друг к другу. Массовыми категориями, на долю которых приходится основной объем трафика являются для ИВТ СО РАН “ПК, Интернет, связь”, “Новости, СМИ“ – 58,61%, а для НГУ “ПК, Интернет, связь”, “Культура”, “Развлечения”, “Справки” – 74,15%.

Ясно, что для самых массовых категорий: “Культура”, “Развлечения” кэширование не является оправданным, поскольку наиболее популярными являются сервисы предоставляющие мультимедийные услуги, такие как просмотр флэш-, видеороликов, прослушивание радио и музыкальных файлов, использование других клиентских приложений, которые в своей работе используют передачу динамической информации. Категория “Новости, СМИ” также содержит в большей степени информацию динамического характера. В этом случае доля статической (кэшируемой) информации составляет малую часть общего веб-трафика. И эффективность использования кэширования – сводится к нескольким единицам процентов.

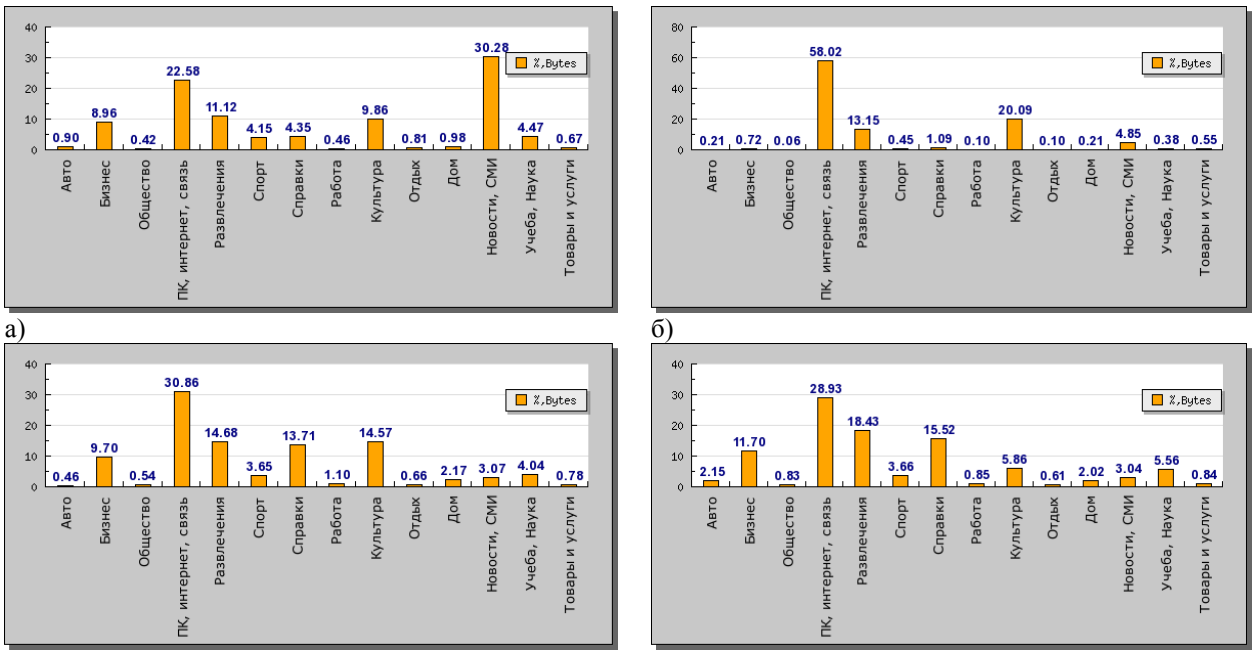



Рис.1 Ранговое распределение объема трафика по категориям для кэш-серверов
а) ICT1 б) ICT2 в) NSU1 г) NSU2

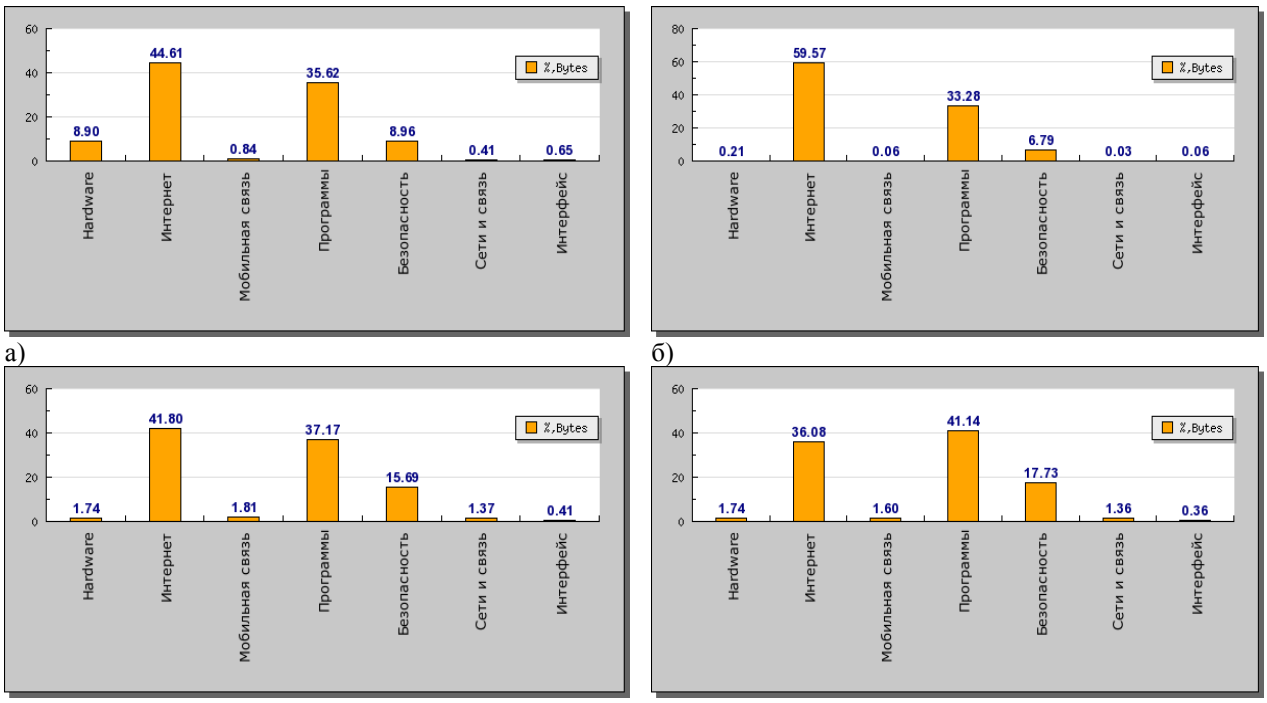
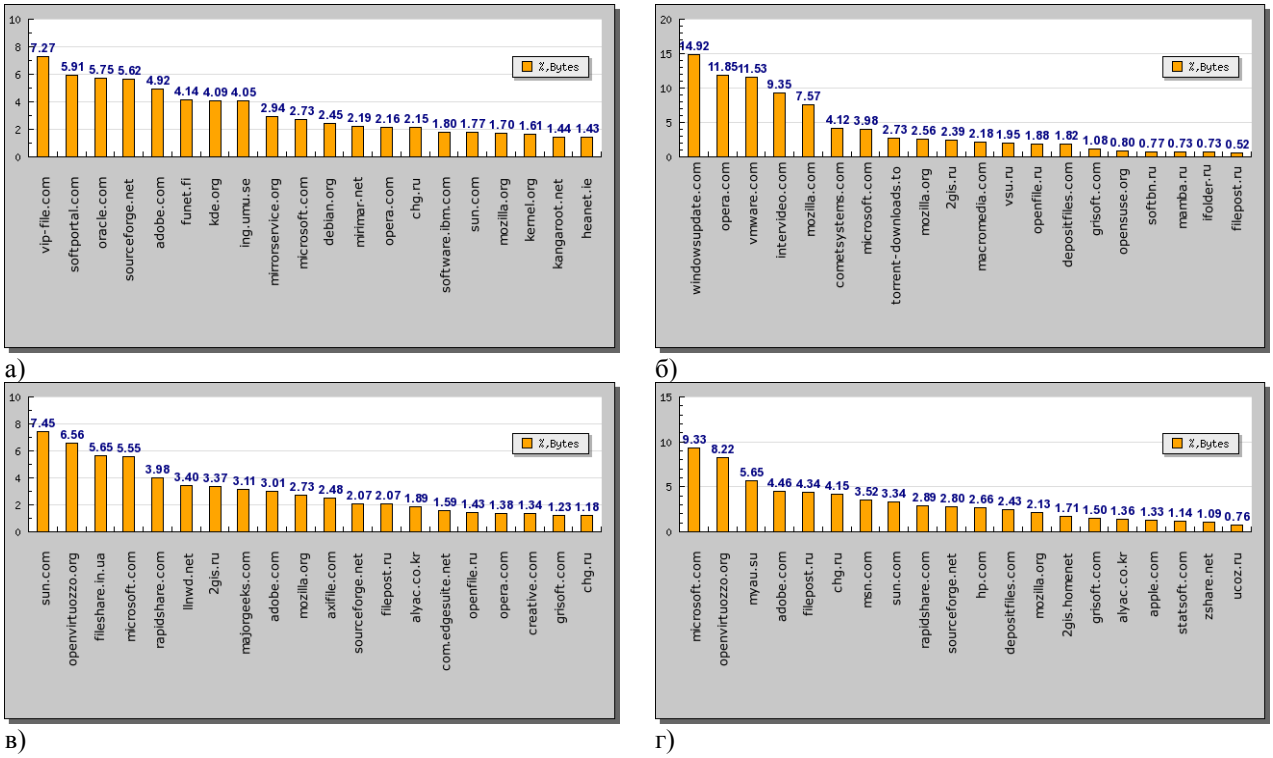



Рис.2 Ранговое распределение трафика в категории “ПК, Интернет, связь”
а) ICT1 б) ICT2 в) NSU1 г) NSU2

Очевидно, что существенной экономии трафика можно добиться кэшированием ресурсов статического характера, как архивы, полные тексты, программное обеспечение и т.п. Поэтому для дальнейшего анализа была выбрана наиболее массовая категория “ПК, Интернет, связь”, на долю которой приходится более 30% трафика. Проведем дальнейшую детализацию данной категории (рис.2).

Видно, что подкатегориями, на которые приходится основной трафик, являются “Интернет”, “Программы” и “Безопасность”. В подкатегории “Интернет” преобладают динамические ресурсы – электронная почта, запросы к поисковым системам, баннерные сети, счетчики и рейтинги. В подкатегории “Программы” преобладают статические ресурсы – программное обеспечение, а




а) б)

в) г)

Рис.3 Ранжирование сайтов по объему трафика (20 хитов)
а) ICT1 б) ICT2 в) NSU1 г) NSU2

подкатегория “Безопасность” также содержит статические ресурсы – антивирусное ПО, ПО для защиты от взлома и спама и т.п.

Дальнейший анализ выявил наиболее «объемное» скачиваемое ПО и его обновления:

- 2GIS
- Adobe Acrobat
- Adobe Macromedia flash
- Adobe photoshop
- CentOS
- Debian
- Eclipse
- FCKeditor
- JRE, JDK
- KDE
- Linux fedora
- Lotus
- Miktex
- Mozilla
- Nero
- Opera
- Oracle
- Pascal
- Safari
- Suse
- Tinymce
- VMWare
- Windows XP, Vista

На основе данного списка ресурсов будет приниматься решение о целесообразности создания зеркала для конкретного информационного ресурса, исходя из стоимости создания зеркала и поддержки его функционирования, возможностей зеркалирования ресурса, организационных, административных и юридических аспектов. В настоящее время в ИВТ СО РАН создан зеркальный сервер, содержащий обновления Windows XP, на который осуществляется «прозрачное» перенаправление пользователей.

Далее было произведено ранжирование сайтов подкатегории “Программы” по объему трафика (рис.3).

## 4 Заключение

Перечисленные методы сокращения затрат на интернет-трафик и потерь рабочего времени, связанных с нецелевым использованием сети интернет, могут быть использованы в любой компании. Эффект от их использования может быть различным. В одном случае затраты могут сократиться на 10%, в другом - в 2-3 раза.

Авторы благодарят рецензентов, высказанные замечания будут учтены в докладе на конференции.

## Литература

[1] A caching Relay for the World Wide Web. Proc. 1st International Conference on the World Wide Web, CERN, Geneva (Switzerland), May 1994. Elsevier Science, p. 69–76.
[2] CyberPatrol Internet Security Software. http://www.cyberpatrol.com/
[3] NetNanny Parental Control. http://www.netnanny.com/
[4] Open-Source Filtering Software. http://www.poesia-filter.org/
[5] SurfControl url and keyword-based Internet filtering and blocking software. http://www.surfcontrol.com/

[6] Герасимов В.В., Курмышев Н.В. Типовой проект создания регионального зеркала. // В сборнике научных статей "Интернет-порталы: содержание и технологии". Выпуск 3. / Редкол.: А.Н. Тихонов (пред.) и др.; ФГУ ГНИИ ИТТ "Информика". - М.: Просвещение, 2005. - С. 379-392.

[7] Крашаков С.А., Теслюк А.Б., Щур Л.Н. Об универсальности рангового распределения популярности веб-серверов // Вестник РФФИ – 2004 - № 1 - с. 46-66.

[8] Крашаков С.А., Щур Л.Н.. Кеширование информационных потоков и стратегия оптимизации маршрутов в распределенных системах. Тезисы докл. 2-ой Всерос. конф. “Научный сервис в сети Интернет”, Новороссийск, сент. 2000, с. 145-148.

## Research of the user preferences for the control and Internet traffic optimisation in the organisation

Leonova Yu., Fedotov A.

In article the questions connected with optimisation of consumption of the Internet traffic and increase of an overall performance of the Internet channel are considered, management problems by information resources and their protection, main principles and concepts, a technique of stage-by-stage reduction of expenses for the Internet traffic and the losses of working hours connected with no-purpose use of a network the Internet, on an example of a scientifically-educational network of NSC SB RAS are described.

# Do Digital Libraries satisfy Users' Information Demand? Findings from an Empirical Study♣

© Magnus Lundqvist[1], Vladimir Mazalov[2], Kurt Sandkuhl[1],

Vladimir Vdovitsyn[2], Evgeny Ivashko[2]

[1]School of Engineering at Jönköping University, Sweden
[kurt.sandkuhl, magnus.lundqvist]@jth.hj.se

[2]Karelian Research Center, Russia
[vmazalov, vdov, ivashko]@krc.karelia.ru

## Abstract

Digital libraries are an important information source of high quality information for various user groups in education, research and industry. With an exponential growing amount of digital content, digital libraries face the challenge of enhancing the support for information seeking. This paper takes the users' perspective and investigates whether the users of digital libraries perceive that their information demand is satisfied. The approach taken is an empirical qualitative study with various user groups in two different countries. From an information demand perspective, the main result is the support for the conjecture that there is a coupling between the user's context and the information demand. Furthermore, a usability questionnaire was used to identify shortcomings and propose improvements in the digital library systems applied at the two study locations.

## 1 Introduction

During the last decade, the amount of information available on the Internet, in digital libraries or in enterprise information systems has been growing exponentially. The main challenge of the information society is no longer that the needed information does not exist electronically [8], the challenge rather is to find and provide the right information. Among the research activities working on this challenge are approaches from information filtering and information retrieval [1, 4], context-based ubiquitous computing [2], context-based decision support and problem solving [9] and information logistics [6].

Digital libraries are an important information source of high quality information for various user groups in education, research and industry. Recent developments in this area aiming at meeting the challenge of the growing amount of digital content include the enhancement of meta-data, enrichment of content or meta-information systems. This paper takes the users' perspective and investigates selected aspects of the users' perception of digital libraries. The guiding question is: Do the users of digital libraries perceive that their information demand is satisfied? This subject can be divided into two aspects: (1) the users' awareness of the own information demand and (2) the usability of the retrieval tools.

The approach taken in this paper is an empirical qualitative study investigating the above questions in various user groups and two different application scenarios: the digital library at Jönköping University (Sweden) and the digital collections of the Karelian Research Center (Russia).

The remaining part of the paper is structured as follows: section 2 introduces the general design of the study. Section 3 presents results form an information demand perspective. Section 4 focuses on findings from a usability perspective. Summary of the work and conclusions are presented in section 5.

## 2 Study Design

The study consisted of two parts: the first part was performed in 2007 at Jönköping University in Sweden, included development of interview guidelines and a usability questionnaire as preparatory activities, and consisted of a pilot study and end user studies. The second part was carried out in 2009 at the Karelian Research Center in Petrozavodsk and used the guidelines and questionnaire from the first part.

### 2.1 Study at Jönköping University

The first part in Jönköping focused on a meta-information system used in the library of Jönköping University, called Samsök. This meta-information system offers a common interface for retrieving information in various "underlying systems", like library catalogues, online archives and full-text

**Proceedings of the 11th All-Russian Research Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» - RCDL'2009, Petrozavodsk, Russia, 2009.**

literature databases. Queries entered by the user in the Samsök user interface are transformed to the interfaces (i.e. query language/format and service interface) of the underlying systems, executed in these systems, and the results are presented in the common Samsök user interface with possibility to continue navigation into the underlying systems.

The purpose of the study was to investigate four main questions:

1. How does Samsök support the end-users, in particular in satisfying the end-user's information demand?
2. How does Samsök support the library's activities and services?
3. What are the results of evaluating Samsök from a usability perspective?
4. What improvement potential can be identified based on the results from the first 3 questions?

The scope of the paper is limited to the end-user perspective, i.e. question 2 will not be discussed and for question 3 and 4 only the end user related aspects are included. A complete account of the results is available in [5].

Based on the above questions, guidelines for data collection and a questionnaire were developed for use in end user studies consisting of sessions of the evaluator with one individual user (respondent) at the time. The guidelines had three purposes: to define the *tasks* to be performed by the respondents, to structure the *session* to be performed and to support the evaluator during the observation. The purpose of the questionnaire was to collect data about the usability and usefulness of Samsök from the respondents perspective. The questions used were a sub-set of the Questionnaire for User Interface Satisfaction (QUIS) [7]. The selection of questions was guided by two principles: 1) the question should be relevant in the context of Samsök and 2) it should be questions directed to end-users, i.e. the respondent should be able to answer them.

In the next step, guidelines and questionnaire were evaluated in a pilot study with two respondents. The results were documented by recording the screen events and recording of the users' oral comments while using the system according to the thinking-aloud approach [3]. Within this study, thinking-aloud means that the user was encouraged by the evaluator to say what he/she is thinking and doing when using the system. This leads to a richer set of data for the analysis work.
The results of the pilot study were used to improve both the interview guidelines and the questionnaire. The end user studies were performed with in total 12 users, 2 in the pilot study and 10 in the main study part. Among these 10 persons were 5 students, 1 researcher, 3 PhD candidates and 1 subject teacher. The objective was to observe a number of end-users with different roles and background in order to get a rich set of information regarding Samsök's application and potential improvements. It should be observed that the intention was to capture qualitative data, and not to collect data with statistical relevance.

### 2.2 Study at Karelian Research Center

The second part in Karelia used the second and improved version of guidelines and questionnaire from the Jönköping study (see 2.1). The study in Karelia was performed using two sets of digital collections located at the scientific digital library (http://dl.krc.karelia.ru) and at the section "Publications" of web-portal (www.krc.karelia.ru) KarRC RAS. These collections are the result of long-term researches developed at KarRC RAS.

The main questions of our study were aligned with questions 1, 3 and 4 of the Jönköping part:

1. What are the usability results of working with digital collections?
2. How does the digital collections' infrastructure meet the users' information demand?
3. How can we improve the digital collections' infrastructure?

The study was performed with 10 users. Among these users were 5 fourth-year students, 3 PhD students and 2 PhD researchers. Such scope of respondents should give us the different answers to our questions.

## 3 Information Demand Perspective

This section summarizes the results from the end user study from an information demand perspective.

For each respondent in the end user study, the session which was part of the end user study started with a pre-interview. In this pre-interview, the respondent had to briefly describe her/his role at the university, how familiar she/he was with using computers, and the information demand he/she has, which shall be the basis for the information searching. During the information searching, the screen events and the oral comments were recorded. After the use of the system, the respondent was asked to what extent her/his information demand was met by the results from the information searching. Due to the qualitative character of the study we chose to capture this "perceived" relevance rather than to evaluate the found information from a recall/precision perspective.

### 3.1 Observations from Jönköping

*Awareness of the own Information Demand*

In general, there is a clear tendency in the interviews that the research and teaching personnel has a more specific and better defined information demand as the students seem to have. Two statements taken from the interviews can serve as example. The first statement is from a student:

*I thought we should look into the area … We are currently developing a web-portal. I thought I could maybe search something about communities and usability. Some theories. Input to the theoretical frame, short theoretical presentations which I then might apply.* [Respondent 2]

The second statement is from a PhD candidate defining his information demand as follows:

*My research area is something called Discharge Care Planning, which is discharge from hospitals if someone has been patient at a hospital. In Sweden, the term coordinated care planning is also used. This is a quite specific track within nursing care. Currently, I am investigating how some sort of IT system or software – called Medics - is used for this purpose. Could be interesting to check whether there are some publications in the domain, but I have no idea which keywords to use as I didn't search for such material before.* [Respondent 6]

This obvious difference is not really surprising, since research and teaching personnel often has a quite well-defined, often narrow and specific work area, which makes it easier to define the information demand. Furthermore, the experience in using libraries is higher. But this difference illustrates the challenges to be met when improving usability of library systems. There seems to be a necessity to take the user's background into account in order to support information searching really well.

Another important aspect of the study was to investigate to what extent the information demand was met. Among the respondents, only a few perceived the support from the Samsök system as satisfactory for finding (enough) information meeting their information demand. This can be illustrated with some statements like:

*Ehhh.. No, you can't really say that. But I found a book. I wanted to have something that is connected to both, dialect and trust. And from Sweden. But, no. I am not really satisfied with what I found.* [Respondent 10]

*No, I didn't. I did not find any article, but I found a book.* [Respondent 4]

*Is work context important for information demand?*
From an information demand perspective, the main result from the study is the support for the conjecture that there is a tight coupling between the user's context and the information demand: The analysis of the data collected in the interviews and of the observations made during the system use shows a tight connection between the respondent's role (teacher, researcher, student, etc.) and the activities for which the searched information is needed (assignment, lectures, scientific work, etc.).

### 3.2 Observations from Karelia

From an information demand perspective the main results are the following.

First of all, we haven't a significant difference in the user groups' awareness of their specific information demand. Some of the students had a more specific and well-defined information demand because they tried to find the information helpful in their scientific work. Experienced scientists tried to find any information that they are interested in. It is also a consequence of specialization of our digital libraries because experienced scientists knew about specific areas of publications at both sites.

Another consequence of specialization of digital libraries is a quite big number of unsuccessful search queries. Some of the respondents tried to find an information in areas that haven't been studied by scientists of KarRC RAS.

## 4 Usability Perspective

From a usability perspective, the data collected during the end user study and the results of the usability questionnaire were evaluated. The next two sections will present the user study observations from Jönköping and Karelia, respectively. Section 4.3 will summarize the usability questionnaires.

### 4.1 Observations from Jönköping

The analysis of the data collected in Jönköping has been structured into different categories reflecting the activities to be supported by the Samsök system:

- Perform the selection (identify keywords for search, combine them, etc.)
- Interpret search results
- Get full-text version of publication

This section will summarize the above results.

**Perform Selection**

The study showed a number of problems when deciding about the keywords to use while searching, in what sequence to apply them during the search and how to express combinations (e.g. by using "and" or "or") of keywords.

Several informants point out the importance of knowing in advance how a search will be performed in order to achieve a good result. This, in combination with an understanding of the language (the syntax used for formulation the search condition), seems to be essential prerequisites for having real use of Samsök. Since the informants in this study received no training on the tool they lack such an understanding, resulting in an unmanageable amount of search hits (the bulk of them being irrelevant). As a consequence, the informants request functionality for filtering results based on language and date. Several informants express a need for higher competence and better support in the search process.

One problem that was observed with most of the informants was connected to preexisting knowledge about databases and different concepts used in the Samsök interface:

*"But if one searches such broad fields as this there is a risk that there will be to much, at least that is the feeling. But at the same time this perhaps is unavoidable and then one has to sort. And there is the possibility to narrow in, that I saw. If I get in to this different ones where I could choose databases one could exclude a lot. But that requires you to know what to exclude on beforehand"* [RE3]

The statement above exemplifies the library clients' need for pre-existing knowledge regarding the different databases used when searching. The observations also revealed some practical problems with the automatic

selection of databases i the quick search – some users did not realise that the search only was performed in a selection of databases.

The observations also revealed that the informants had problems with interpreting a number of terms in Samsök's interface. *Meta* search is not an intuitive word and thus means that it is not obvious to the user that this is where more advanced searches can be performed. The meta search is appreciated after some use but is perceived as unclear at a first glance. It is not clear how the left part, where the databases are chosen, is to be used. Most of the time the users click on the first list where "categories", "quick groups" and "combine" are instead of selecting a topic category in the lower list despite this being functionality they ask for. Functions for creating personal groupings of databases are requested.

The observations also revealed misinterpretations regarding the following terms; *Topic* terms was confused with *search* terms, the formulations *search database* and search *electronic magazine* were interpreted as searching content rather than on names of databases/publications.

To summarize the problems observed, we can group them according to the cause of the problem:

- *Database knowledge* – Knowledge of relevant databases and how to select the databases to be used during the search
- *Search competency* (general) - General knowledge about structured information search
- *Search competency* (Samsök) – Knowledge about Samsök and how to express queries
- *Functionality* – to be able to express the selection based on criteria important for the users (e.g. language or time interval)
- *Terminology* – regarding difficulties to understand the available choices for meta-search, including the term meta-search as such

**Interpret Search Results**

Several informants commented on the large amount of hits in the search results. This is directly connected to the difficulties with performing selection.

A lack of understanding of the link "view collected hits" in the quick search results in the abortion of the search. This is inconsistent with the interface in the meta search. When performing a meta search no indication of collected hits is showed in the result list until all results are collected.

Better support for the user when deciding on relevance is needed. This is connected to a lack of understanding of the underlying databases:

*"No, I found this a bit hard, that they are showed like this, OK, this magazine has so and so many hits and this has so many. It would have been much better to just get them listed in a row and not having to continue again by clicking in to an article or magazine because I do not know the magazines. If I had know that I might have been able to select in a different way but now it is just names to me. They could just as well been named 1234567 or blue, red, green because I have no idea what it is. It felt a bit, OK but which one should I choose? I take the one with most hits?"* [RE10]

In the part of the interface where the search results is listed a certain amount of problems regarding the navigation between different views were identified. When the view full post is shown use of the web browsers back functionality does not return the user to the previous page but rather to the previous post in the list, hence it is hard to return to the list view.

Conceptually it is also reasonable to question the use of the term weight with respect to search results as this is not a obvious term for describing relevance, something that contributes to confusion.

The problems observed in interpreting the search results can be categorized as follows:

- *Database knowledge* – lack of knowledge regarding the databases makes the interpretation of the search results difficult
- *Incomplete Searches* – users tend to misinterpret to what extent a search is "completed" when they start to look at the hits
- *Navigation between views* – some users had problems to navigate between the list of search results and the view showing details for one search result
- *Terminology* – respondents had difficulties to interpret certain system terms, like meaning of "weight" in search results, significance of different databases, meaning of "get more hits"

**Get Fulltext**

It is not obvious how one should go about to get a full text version of an article. Sometimes, this is done by following an ordinary hyperlink while in other cases it is done by means of the SFX screen. The symbol used for SFX is unintuitive and in some views its functionality is not explained and it is therefore consequently unused. Furthermore, the SFX screen gives no feedback on the existence of the article leading to the users using JULIA instead. Many of the problems seems to be related to the users' lack of understanding of the library domain, hence they do not understand the use of and need for "LIBRIS web search".

Sometimes when navigating to full text versions the user is transferred to external websites. This requires the user to interpret and understand additional environments to perform a successful search. As the appearance of these sites are not a part of Samsök this is hard to influence but there is nevertheless important to realise that these different systems are a part of the overall user experience.

To go from search result to full-text of a publication caused some problems for the respondents, which can be summarized in three categories:

*Unclear how to get full-text* – a quite general problem was to access the full-text versions of publications found during the information searching

*Terminology* – the users do not connect the SFX-symbol with the possibility to get the full-text. The users don't know the terms used in the SFX-window.

*Navigation between views* – Samsök offers and requires different ways to navigate to the full-text version. This confuses the respondents who would prefer one clearly defined way to go from search result to full-text

### 4.2 Observations from Karelia

There are three main categories of results from a usability perspective:

1. usability of the site;
2. performing the search;
3. interpretation of the search result.

Both studied sites estimated by respondents by a single mark if they haven't seen significant difference.

**Usability of the site**

Marks made by respondents show that the usability of both sites is good enough. The most part of respondents made good marks for convenience, usefulness, design and so on. These characteristics are important for stimulating users to further looking for needed information.

Terms used by the sites also didn't cause any doubts.

**Performing the search**

Characteristics related to performing the search have been estimated by respondents in different ways. Respondents hadn't a single opinion about complexity of the search system, functionality and flexibility. It is interesting that the more experienced users made the higher marks for these characteristics.

Opposite estimations made by respondents for time needed to learn about basic and additional search functions. There is also a similar difference between more experienced and less experienced users.

**Interpretation of the search result**

Respondents pointed that there is enough amount of visualized information of the search result, but the number of documents is very small. It is also a consequence of specialization of digital libraries. Any attempts to find out the areas that haven't been studied by researchers of KarRC RAS were not successful.

### 4.3 Results from the Usability Questionnaire

The results from the usability questionnaire are summarized in the following two tables. Table 1 reflects the answers regarding the general impression. Table 2 addresses the user interface impression.

The questionnaire results gave some indications regarding the users' impressions of Samsök, represented in table 1 below. It should be pointed out that the selection in the survey is too small to derive statistically valid conclusions about a larger population. Instead we view the results as an indication of how the users of Samsök perceive the application. The underlying reason for the not so positive remarks done by the informants (shadowed cells in the table) is according to our perception that most of them were unsuccessful in finding the type of material they were looking for. It should also be noted that there were users that valued the application as simple, powerful and rewarding despite the fact that they never before had used it.

| Perception | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1.1 Terrible - Wonderful | | 2 | 7 | 1 | |
| 1.2 Frustrat. - Rewarding | 3 | 3 | | 4 | |
| 1.3 Boring - Stimulating | 4 | | 4 | 2 | |
| 1.4 Difficult – Easy | 2 | 2 | 5 | | 1 |
| 1.5 Insufficient - Powerful | | 4 | 1 | 5 | |
| 1.6 Rigid – Flexible | | 7 | 2 | 1 | |

Table 1 – Respondents' general perception of Samsök

The following table describes results of the study at KarRC RAS. These results show that the search system meet the user's purposes.

| Perception | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1.1 Terrible - Wonderful | | | 2 | 8 | |
| 1.2 Frustrat. - Rewarding | | | 3 | 2 | 5 |
| 1.3 Boring - Stimulating | | 1 | 2 | 3 | 4 |
| 1.4 Difficult – Easy | 2 | 1 | 2 | 3 | 2 |
| 1.5 Insufficient - Powerful | | 1 | 1 | 5 | 3 |
| 1.6 Rigid – Flexible | | 2 | 7 | 1 | |

Table 2 – Respondents' general perception of KarRC RAS' digital libraries

The questionnaire also contained a number of questions regarding the design and learnability of Samsök, the results from which is listed in table 2 below. Parts of the table has been shadowed to point out the cases where opinions strongly various between different informants.

| Perception | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 2.1 Design | | 2 | 4 | 2 | 2 |
| 2.2 Terminology | | 3 | 1 | 3 | 3 |
| 2.3 Graphic symbols | 2 | 4 | 1 | 3 | |
| 2.4 System status | 1 | 3 | | 5 | 1 |
| 2.5 Feedback (content) | | 2 | 4 | 3 | 1 |
| 2.6 Feedback (visibility) | | 2 | 3 | 5 | |
| 2.7 Search results – amount of information | | 2 | 3 | 4 | 1 |
| 2.8 Learning - basic | | 1 | | 5 | 4 |
| 2.9 Learning - advanced | 1 | 1 | 3 | 3 | 2 |
| 2.10 Navigation | 1 | 1 | 2 | 5 | 1 |
| 2.11 Response time (search) | 2 | 3 | | 3 | 2 |
| 2.12 Response time (navigation) | 1 | 3 | 1 | 2 | 3 |

Table 3 – Respondents' impression of Samsök's user interface

The terminology in the interface was perceived as relatively clear while the graphical symbols was considered harder to interpret. Regarding the systems status, i.e. how easy it is to understand what the system is doing at the moment, the answers are polarised. This is most likely due to the users' different experience of using web applications. The feedback given by the system gets a vaguely positive judgment. The users generally think that it is relatively simple to learn the simpler parts of Samsök while the more advanced parts (meta search) is perceived as more difficult to understand. The navigation was by most perceived as relatively simple to handle while the response times when searching and navigating indicates certain problems. Especially the response time for searches hints that involved servers have different response times – some users have not experienced this as a problem while others have.

Themes that stands out in the survey is according to us that Samsök suffers from less than stable response times and that the users' impressions on a whole leans towards less positive judgements such as boring and frustrating.

The following table shows marks made by respondents of KarRC RAS.

| Perception | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 2.1 Design | | | 3 | 2 | 5 |
| 2.2 Terminology | | | 1 | 5 | 4 |
| 2.3 Graphic symbols | | | | 3 | 7 |
| 2.4 System status | | | 1 | 7 | 2 |
| 2.5 Feedback (content) | | | 1 | 5 | 4 |
| 2.6 Feedback (visibility) | | | 1 | | 9 |
| 2.7 Search results – amount of information | 2 | 1 | 2 | 3 | 2 |
| 2.8 Learning - basic | 2 | 1 | | 4 | 3 |
| 2.9 Learning - advanced | 2 | 1 | 3 | 3 | 1 |
| 2.10 Navigation | | | 3 | 5 | 2 |
| 2.11 Response time (search) | | | 2 | 3 | 5 |
| 2.12 Response time (navigation) | | 1 | 1 | 4 | 4 |

Table 4 – Respondents' impression of KarRC RAS' digital libraries user interface

The strict design with absence of superfluous elements of design and functions makes the interface convenient. But the lack of information makes users dissatisfied.

## 5 Conclusions

Six categories of experiences form the use of Samsök were discussed in section 3.3 Three of these are directly connected to different phases in the search process; *Selection*, *Interpretation*, and *Collecting full-text*. The remaining three categories are more connected to the overall use of Samsök; *General opinions, reasons for contacting the library*, and *proposals for further development*. Within each category a number of problematic themes have been generated from the interviews and observations. The following themes has been identified:

*Database knowledge* – a basic understanding of academic databases is required to utilise Samsök. This introduces problems to the activities, selection and interpretation.

*Search competence* (general) – information searching requires some general knowledge that many of the survey's informants do not have. An example of this is competence in evaluating the quality of different types of publications as well as the competence to, for a given problem, identify relevant topics and search terms.

*Search competence* (Samsök) – viewed as a tool Samsök requires its users to have some knowledge regarding how to formulate search terms and selecting suitable databases for the meta search. Several informants had problems with these parts.

*Unclear access to full-text* – there is at the moment several different ways to access full-text versions of articles, something that confuses users and in the worst case scenario means that they do not understand that a full-text version is available.

*Terms and symbols* – there is a number of terms and symbols used in Samsök that is hard to understand for the uninitiated users. One such term is meta-search, another is weight and the symbol used for SFX a third.

*Requested functionality* – some functionality, with respect to the users' information demand, is missing or hidden in Samsök. An example of this is the possibility to search based on language and dates. Other examples of the same problems identified by the informants are the possibility to search within search results as well as gaining simple access to search history.

*Unfinished searches* – it is possible to view collected hits despite the fact that the search still is ongoing. This is not obvious to the user in the current design. Furthermore, it is not obvious that more hits and then with higher relevance can be collected with the function "collect more hits".

*Navigation* between views – a number of problems in Samsök is related to the navigation. One such problem is that the use of the browsers back-button breaks the expected behaviour of taking the user back to the previous screen. Another is that the linking to external documents is inconsistent and unintuitive.

*Response times* – the response times of the system vary depending between the different observations, resulting in negative judgements from the informants.

*Overall perception* – The survey showed on a frustration amongst some of the informants regarding the use of Samsök. Some of them even perceived the system as a boring tool.

The study of the digital libraries of KarRC RAS highlighted a different main problem – the lack of content. On the one hand our digital libraries aims to make public the work of the researcher of KarRC RAS, but on the other hand the small amount of content makes users dissatisfied in looking for specific

information. The consequence of this is a small number of users who regularly use the digital libraries. With a growing amount of content, other usability issues might be raised, like for example navigation in large lists of hits for a query.

## 6 Summary

This paper investigates whether the users of digital libraries perceive that their information demand is satisfied. The approach taken is an empirical qualitative study with various user groups in Jönköping and Karelia. This study includes two aspects: the users' awareness of the own information demand and the usability of the retrieval tools.

From an information demand perspective, the main result from the study is the support for the conjecture that there is a coupling between the user's context and the information demand: The analysis of the data collected in Jönköping shows a tight connection between the respondent's role (teacher, researcher, student, etc.) and the activities for which the searched information is needed (assignment, lectures, scientific work, etc.). With respect to usability, there seems to be a necessity to take the user's background into account in order to support information searching really well. Furthermore, in the Jönköping study there is a clear tendency that the research and teaching personnel has a more specific and better defined information demand as the students seem to have. This observation from Jönköping that researchers seem to be more aware of their information demand was not confirmed in the Karelian part.

The usability questionnaire was helpful in identifying shortcomings and proposing improvements, both for the Samsök system in Jönköping and the digital collections in Karelia. However, the two systems are far too different regarding user interfaces, functionality and amount of content that a comparison of the findings should be considered. A commonality between both cases is that we observed that usability was graded worse by those users who were not successful in retrieving content meeting their information demand.

The main limit of the research presented here is the limitation to just two digital libraries/collections and to just groups of 10 end users in every part of the study. It would be worthwhile and interesting to include a larger number of both digital libraries and users.

## References


[1] Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM, 35 (12), 29–38

[2] Dey, A. K. (2000). Providing Architectural Support for Building Context-Aware Applications. PhD thesis, College of Computing, Georgia Institute of Technology.

[3] Lewis, C (1982) *Using the "thinking-aloud" method in cognitive interface design* (IBM Research Rep. No. RC 9265 [#40713]). Yorktown Heights, NY: IBM Thomas J. Watson Research Center

[4] Palme J. (1998). Information filtering. In Proceedings of the 12th Biennial ITS (International Telecommunications Society) Conference, Stockholm.

[5] Samsök: Handlingsbarhet, behovsinriktad användarstudie och tutorial. Rapport för projekt med utvecklingsbidrag från Kungl. biblioteket / BIBSAM; Högskolebiblioteket i Jönköping. Dnr. vid Kungl. biblioteket / BIBSAM: 63-612-2005.

[6] Kurt Sandkuhl: Information Logistics in Networked Organizations: Selected Concepts and Applications. Enterprise Information Systems, 9th International Conference, ICEIS 2008. LNBIP, Springer.

[7] Shneiderman, B, Plaisant C (2005) Designing the User Interface – Strategies for Effective Human-Computer Interaction. Addison-Wesley. ISBN 0-321-26978-0.

[8] Skjong, R., Johnsen, Ö., Weitzenböck, J., Brynestad, S., Mestl, T., 2004. Technology Outlook, p.109. Det Norske Veritas, Norway, ISBN 82-515-0300-0

[9] Smirnov, A., Pashkin, M., Chilov, N., and Levashova, T. (2005). Ontology–Based Knowledge Repository Support for Healthgrids. In Proceedings of Healthgrid 2005: From Grid to Healthgrid, (Solomonides T., McClatchey R., Breton V., Legré Y. and Nørager S. Eds.), 47—56. IOS Press.


## Удовлетворяют ли электронные библиотеки информационным запросам пользователей? Эмпирическое исследование


М. Ландквист, В. Мазалов, К. Сенкюль, В. Вдовицын, Е. Ивашко



Электронные библиотеки являются важным источником информации для различных групп пользователей в области промышленности, образования и науки. Взрывной рост объемов информации, представленной в цифровом виде, ведет к тому, что пользователи электронных библиотек все чаще сталкиваются с проблемами поиска информации. В статье представлено исследование, цель которого — оценить насколько хорошо пользователи могут удовлетворить свои потребности в информации с помощью электронных библиотек. Объектом исследования данной эмпирической работы являются различные группы пользователей двух стран. С точки зрения удовлетворения информационных запросов, основной полученный результат — это подтверждение наличия связи между контекстом поиска и информационными запросами. На основе специально разработанного опросника, среди

пользователей, участвующих в исследовании, было проведено анкетирование для определения недостатков и возможностей улучшения систем поддержки электронных библиотек.

♣ Part of the work was supported by RFFR (grant No 08-07-00085a). Another part was financed by The Swedish Royal Library in context of the Samsök project, The authors wish to thank their co-workers from Jönköping International Business School and Jönköping University Library, in particular Thomas Albertsen and Jonas Sjöström.

# ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА

# DIGITAL LIBRARIES TOOLS

# Архивная фактографическая система

© Марчук А.Г., Марчук П.А.

Институт систем информатики им. А.П. Ершова СО РАН, НГУ, Новосибирск
mag@iis.nsk.su, peter@iis.nsk.su

## Аннотация

В Институте систем информатики СО РАН создана и используется архивная система, построенная на фактографических принципах [1, 2]. Особенностями системы являются: распределенная база данных, реализованная в виде документов RDF, специальные средства поддержки распределенного редактирования и динамической синхронизации; использование формальных спецификаций схемы данных и ряда свойств данных через описания OWL, что обеспечивает сменность или расширяемость системы структуризации и наборов метаинформационных полей; предложена и реализована техника кассет, группирующих документные массивы и обеспечивающая экономный доступ к документам, перемещаемость опубликованных документов и общую схему публикации и архивации документов и базы данных; имеется совместимость с базами данных класса Semantic Web, реализованными средствами RDF и OWL, имеется возможность организации синхронизации с базами данных, реализованными средствами реляционных таблиц. Данный доклад описывает реализацию сформированного подхода и является фактически продолжением указанных работ. Работа выполнена при поддержки гранта РГНФ 08-03-12125в, программы РАН р 2/12.

## 1 Архитектура системы

Рассмотрим программные компоненты архивной системы и их взаимодействие Архивная система создана как модульная открытая система с возможностью гибкого конфигурирования и настроек на потребности групп пользователей. Объединяющие такой подход компоненты это ядро системы и базовая онтология. Формально говоря, базовая онтология не является «встроенной» в примененный подход, «встроенными» являются некоторые принципы формирования базовой онтологии. Это дает возможность подставить в архивную систему другую онтологию (построенную с учетом указанных принципов) и получить архивную систему с другой структуризацией данных. При этом, если подставленная онтология совсем другая, то мы утеряем возможность интеграции данных, основанных на этих разных онтологиях. Рекомендуемый способ изменения онтологии – расширение базовой онтологии до требуемой полноты описания конкретной предметной области. Тогда срез данных, соответствующих базовой системе понятий и отношений, будет основой для интеграции данных, построенных на разных онтологиях.

К другим компонентам текущей реализации архивной системы относятся:

- Универсальный редактор базы данных, выполненный в виде Web-приложения;
- Публичный интерфейс информационного массива «Фотоархив СО РАН»;
- Система обработки, архивации и публикации первичных документов;
- Универсальное фактографическое приложение «Фактограф», ориентированное на персональную и коллективную работу с фотодокументами и базой данных;
- Системный интегратор, позволяющий динамически синхронизовать изменения, порождаемые в разных активных моделях;
- Набор утилит, предназначенный для анализа и восстановления данных, импорта внешних массивов данных, резервного копирования и обеспечения целостности данных.

Общая архитектура архивной системы представлена на рисунке 1.

Архивная система реализуется как распределенная информационная система, в которой распределены данные (распределенная база данных), сервера обработки и доступа, а также абоненты системы. Распределенная база данных выполнена в виде опубликованных RDF/OWL файлов, клиентская часть может состоять из универсальных клиентов (браузеры) и специализированных клиентских приложений. Сервера обработки и доступа выполняют предписанные действия на собранных из элементов распределенной базы моделях. Модель представляет собой внутреннее представление

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

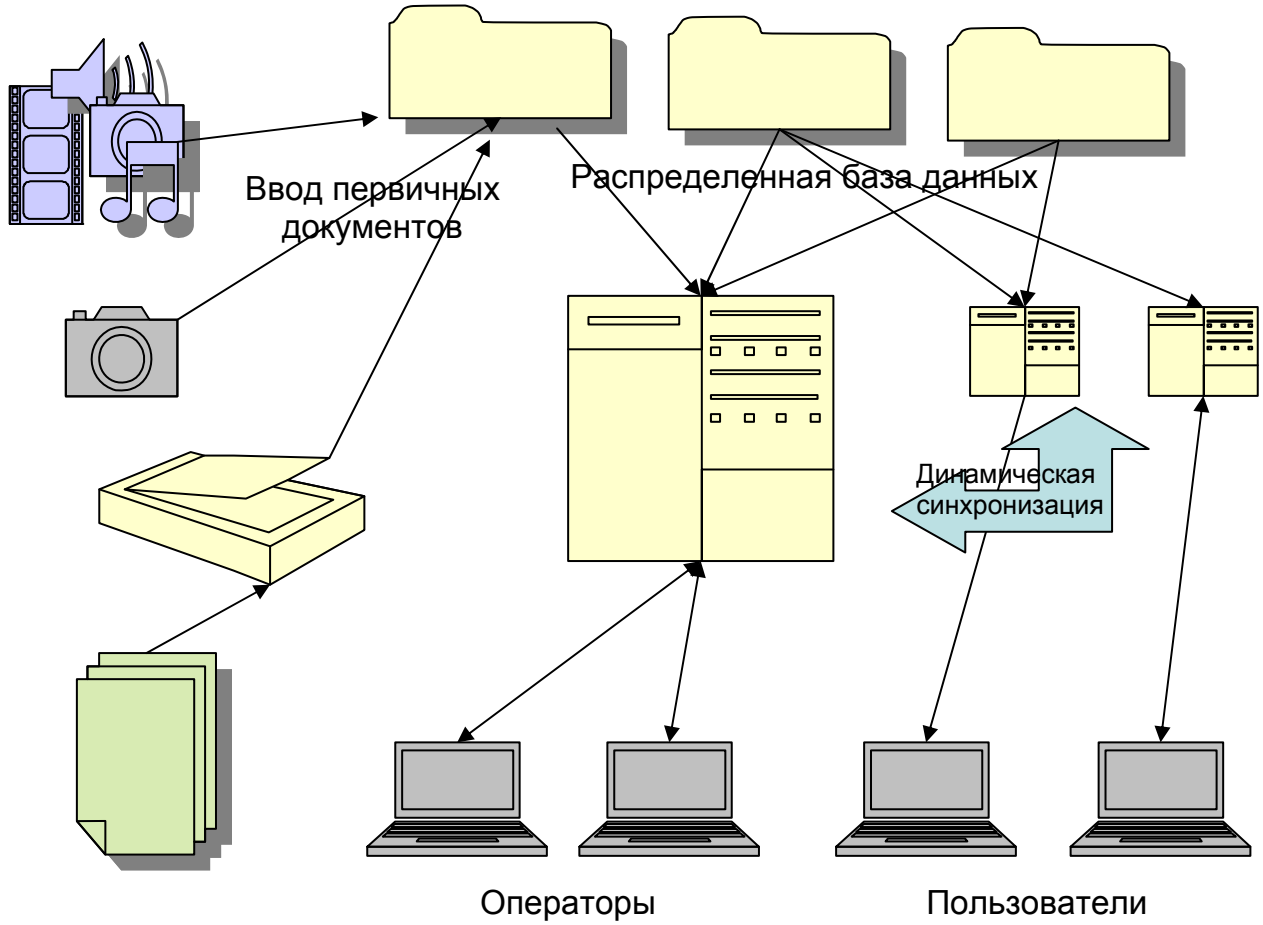



Рис. 1. Архитектура архивной системы

семантической сети и формируется из подмножества загруженных RDF-документов. Поскольку при функционировании системы отдельные модели могут изменяться, например средствами редактирования, необходимо производить динамическую синхронизацию, как с файлами документов, так и моделей между собой.

Отдельным слоем архивной фактографической системы является работа с документами, являющимися элементами базы данных. К документам относятся публикации, фотографии (фотодокументы), отсканированные образы страниц, аудио-видео файлы и т.д. Особенностью таких документов является то, что они фигурируют в базе данных в двух вариантах. С одной стороны – документы представлены в базе данных своими метаинформационными записями. С другой стороны, каждый документ это и конкретный информационный контент, например файл имиджа фотографии, который специальным образом структурируется, группируется и публикуется. Для этого была применена технология кассет данных.

Детали устройства и реализации отдельных слоев архивной фактографической системы будут изложены в следующих разделах.

## 2 Структура данных, онтология

Центральным местом структуризации данных в системе является онтология. Онтология задается в соответствии со спецификациями OWL Light [3] одним или несколькими OWL-документами. В настоящее время используется базовая онтология неспецифических сущностей, в некоторых случаях она несколько расширена для соответствия конкретной предметной области. Онтология определяет набор классов сущностей и отношений между ними, для реализации атрибутированных отношений используются классы псевдосущностей [2]. Например, отношение «работа» между персоной и организацией, нуждается в использовании атрибутов таких как: дата начала, дата завершения, должность и др. Соответственно, в онтологии введен класс participation. Этот класс не определяет объектов, а через объектные свойства participant и in-org порождает сложную связь между персонами и организациями. Такой подход делает нелогичным зафиксированный в стандарте OWL идентификатор thing в качестве корневого для всех классов, в нашей онтологии в качестве идентификатора корневого класса используется entity (сущность).

С некоторыми упрощениями, приведем основную часть используемой онтологии.

Спецификации в формализме OWL довольно громоздки и не достаточно наглядны, поэтому представлением в виде дерева, отражающего наследование классов. На диаграммах простые идентификаторы – идентификаторы классов, идентификаторы, следующие за прямоугольником – идентификаторы свойств. Причем если далее идут скобки, то это объектные свойства (ObjectProperty) и в скобках отмечен связываемый класс (range).

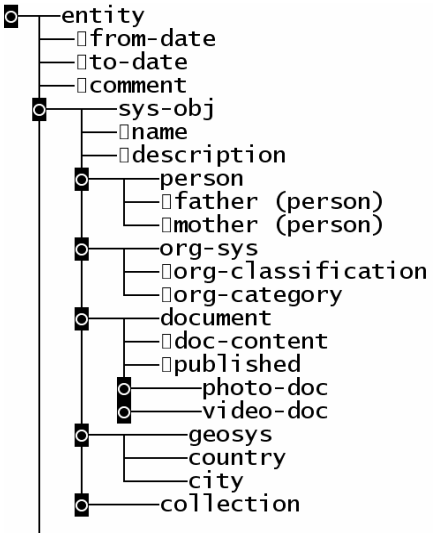



Рис. 2. Система объектов

На рисунке 2 приведены основные используемые в онтологии классы объектов. Для выделения свойства «объектности» класса, введен абстрактный класс системного объекта (sys-obj). Определено пять классов: персоны, организационные системы, документы, географические системы и коллекции. Для этих классов введен минимальный набор полей (DatatypeProperty) и объектных ссылок (ObjectProperty), используется наследование свойств.

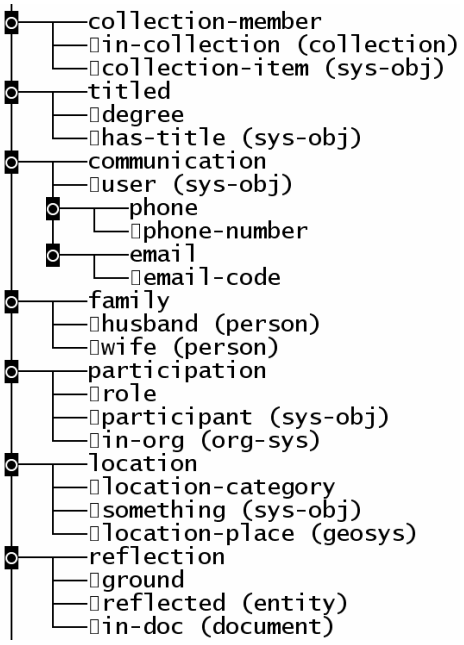



Рис. 3. Сложные (составные) отношения

Рисунок 3 представляет собой продолжение рисунка 2, на нем приведена система составных отношений между введенными объектами. Например, отношение collection-member устанавливает ассоциацию между объектом класса collection и системным объектом. Имеются не только бинарные отношения, но и унарные: titled, communication.

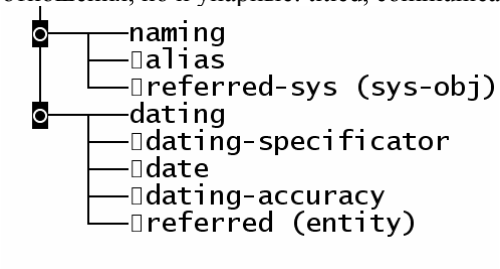



Рис. 4. Специальные отношения

Специальные унарные отношения представлены на рисунке 4, они реализуют более развернутым образом такие важные поля как name, from-date, to-date. Это дает возможность множественного указания имен и дат, а также записать уточняющую информацию. Например, девушка Иванова вышла замуж и сменила фамилию на Петрова. Соответственно, в базе данных была запись:

```
<person rdf:about="ivanova_4133">
   <name>Иванова</name>
</person>
```

Традиционным методом приведения базы данных в соответствие, является ее редактирование. Однако, если изменить значение name, то утеряется предыдущее значение, что является нарушением фактографических принципов. Применяя специальное отношение naming, мы можем отредактировать запись так, чтобы предыдущее значение не терялось, а было правильно оформлено. Для данного примера вместо одной записи появится две:

```
<person rdf:about="ivanova_4133">
   <name>Петрова</name>
</person>
<naming rdf:about="nmg28738">
   <alias> Иванова </alias>
   <referred-sys rdf:resource="ivanova_4133"/>
   <from-date>(дата рождения)</from-date>
   <to-date>(дата смены фамилии)</to-date>
</naming>
```

Аналогично, датирование помогает преодолевать проблемы связанные с неточностью знаний о дате и наличия нескольких версий дат.

## 3 Программное ядро системы

В архивной фактографической системе выделен общий модуль ядра системы, работающий с моделью данных. Этот модуль используется во всех программных решениях, использующих модель данных для своей работы. Такое инкапсулирование работы с данными позволяет оставлять возможность изменения отдельных технических решений,

связанных с форматами, пространствами имен, структурой и смыслом атрибутов документов и др. Кроме того, ядро системы сопряжено со слоем динамической синхронизации (см. далее).

В отличие от многих систем, работающих с RDF-данными, было принято решение не использовать реляционную СУБД в качестве движка, а создать прямую программную реализацию действий и доступов, соответствующую базовой модели RDF. Это значит, что объектами модели являются сущности и отношения, порождающие в совокупности семантическую сеть – ориентированный граф, состоящий из узлов и направленных дуг. Такое построение пока рассчитано только на полную загрузку модели в оперативную память исполняющего компьютера. Основанием к тому является быстрый рост объемов памяти не только на серверных, но и на пользовательских компьютерах. Оценка показывает, что подобное сетевое построение для 1 миллиарда фактов потребует порядка 100 Гб оперативной памяти, что не выглядит невыполнимым уже сейчас. При этом, баз данных общего назначения, содержащих миллиард фактов пока не создано, обычные объемы – сотни тысяч и единицы миллионов фактов.

Модуль ядра загружает RDF и OWL файлы, при этом используется технология кассет, см. ниже. Далее строится модель данных и модель онтологии. К модели данных и модели онтологии определен ряд методов доступа и манипуляций. Логически, модель данных представляется ориентированным графом, причем методы доступа позволяют «ходить» как по стрелкам (ссылкам), так и против. Еще одной особенностью представления модели является то, что сохраняется исходная структура документов. Это означает, что каждая информационная запись приписана к тому документу, в котором она находилась в файле RDF-документа. Имеется набор методов редактирования модели. При редактировании, сделанные изменения периодически записываются в RDF-документы, доступные компьютеру для записи. Если документ не доступен по записи, то посылается сообщение, в общем случае удаленному, сервису синхронизации детали работы которого будут изложены позже.

Еще один слой ядра поддерживает многоабонентскую работу. Это используется например в Web-приложениях, работающих в конфигурациях архивной фактографической системы. Также имеется менеджер документов и менеджер проектов.

Ядро поддерживает протоколирование (Log) выполненных действий, соответственно есть возможность реализации откатов и восстановления данных в случае порчи по непредвиденным обстоятельствам. Ядро снабжено Web-сервисом, работающем в логике синхронизации (см. далее) распределенной системы активных моделей.

Класс работы с онтологией расширяет класс работы с RDF-моделью специальными методами доступа к модели и ее различных представлений.

## 4 Публичный и операторский интерфейсы

В качестве основных средств работы, фактографическая система снабжена Web-приложениями, представляющими публичный пользовательский (front-end) и операторский (back-end) интерфейсы. Публичный интерфейс предназначен для доступа пользователей к информационным ресурсам системы. В общем случае, для достижения этой цели, для каждого большого проекта нужно разрабатывать свой вариант публичного интерфейса. Для проекта «Фотоархив СО РАН», такой интерфейс был создан и эксплуатируется с мая 2007 года. Скриншот типовой страницы сайта и некоторые детали его устройства имеются в [1].

Для наполнения фотоархива данными и их редактирования, был создан операторский интерфейс. Это Web-приложение, совместно с графическим обликом страниц, было спроектировано как универсальное решение, настраивающееся на предметную область через формальные спецификации – онтологию. Как универсальный компонент, операторский интерфейс был опробован и использован в целом ряде проектов, выполненных в ИСИ.

Операторский интерфейс построен на основе представления множеств записей в виде таблицы. Пусть нашей задачей является изображение нескольких однотипных записей, каждая из которых имеет вид (используется графовая модель RDF), изображенный на рисунке 5:

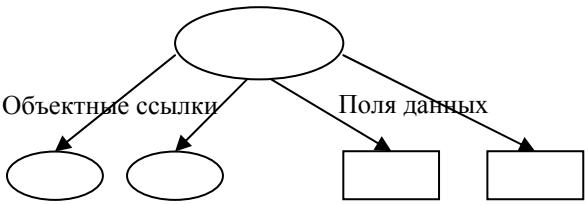



Рис. 5. Представление отдельной записи в виде фрагмента RDF-графа

При наличии заданной онтологии, однотипность означает единый набор возможных или имеющихся полей-данных и прямых ссылок. Сопоставим рассматриваемым записям таблицу, строками которой будут информационные поля записей, а колонками – поля и прямые ссылки. Поля-данные, помещаются в ячейки таблицы непосредственно, прямые ссылки логично поместить через их inline-

представление, например в виде гиперссылки.

| Метка типа | | | | |
|---|---|---|---|---|
| Метка поля 1 | Метка поля 2 | … | Метка прямой ссылки 1 | … |
| Значение поля 1 | Значение поля 2 | … | Inline-представление ссылки 1 | … |
| … | | | | |

Таблица 1

Обозначим (формулы приводятся в формализме XPath):

$type – тип элементов набора,

$type_id – идентификатор этого типа,

$items – набор записей этого типа,

$ontology – множество записей определения онтологии.

Дополнительно вычислим два промежуточных множества:

$dprops = $ontology/DatatypeProperty[domain/@rdf:resource=$type_id] – описание полей-данных,

$oprops = $ontology/ObjectProperty[domain/@rdf:resource=$type_id] – описание полей-ссылок.

Теперь рассмотрим каноническое представление информационной единицы (записи) в которой к записи добавляются через объектные ссылки еще другие элементы, оно изображено на рисунке 6.

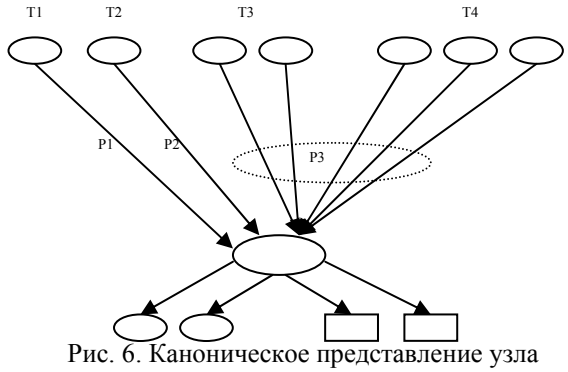



Рис. 6. Каноническое представление узла семантической сети RDF, соответствующего информационной единице (записи)

Легко видеть, что, если отсутствуют объектные ссылки узлов самих на себя, то любой узел семантической сети RDF представляется в представленном на рисунке 6 каноническом виде: в центре располагается опорная сущность, внизу имеются поля-данные (DatatypeProperty) и прямые объектные ссылки (ObjectProperty), сверху изображены объектные ссылки (отношения) других информационных единиц, ссылающиеся на данную. Обратные отношения для элемента сгруппированы по типу отношения и по типу класса элементов, ссылающихся на данный.

Получившиеся в результате разбиения подграфы укладываются в модифицированный вид рассмотренной таблицы (Таблица 2).

Где в первую строку перед меткой типа добавлено название общего для группы объектного свойства. Соответственно, переменная $oprop обозначает это свойство (ObjectProperty). Кроме того, для обратных отношений для колонок должно выполняться следующее условие:

$oprops[i] != $oprop.

Это означает, что не нужно выдавать информацию о записи «из которой» мы пришли к построению таблицы.

Имеющиеся в таблицы выражения теперь надо рассматривать в контексте групп записей, к которым таблица относится. Обозначим функцию, строящую таблицу по указанным формулам:

ViewTable(prop_id, type_id, items).

Если проанализировать каноническое представление информационной единицы, то в контексте вышесказанного можно увидеть, что целостное информационное представление информационной единицы состоит из N + 1 таблицы, где N – количество вариантов свойство-тип для обратных отношений. Первая таблица является вырожденной, имеющей одно и только одно описание информационной единицы – собственно базовое описание. И для этой таблицы не определено значение prop_id. Остальные таблицы могут изображать ноль или более элементов определенного типа, ссылающихся на базовую единицу через указанное отношение.

Это представление опишем формулами. Пусть имеется идентификатор item_id записи, которую целостно мы хотим изобразить. И пусть data_base есть множество записей нашей базы данных. Тогда первая таблица строится в виде:

```
$items = $data_base/*[@rdf:about=$item_id]
ViewTable(null, name($items), $items)
```

А остальные порождаются в цикле:

```
$range_set = $ontology/ObjectProperty[
    range/@rdf:resource=name($items)]
$inverse_links =
    $data_base/*/*[@rdf:resource=$item_id]
Foreach $inverse_prop in $range_set do
{
    $inverse_type_set =
    $inverse_prop/domain/@rdf:resource
    Foreach $inverse_type_id in
    $inverse_type_set do
{
       ViewTable( name($inverse_prop),
    $inverse_type_id,
    $inverse_links[name()=name($inverse_prop)
    and name(parent::*)=$inverse_type_id])
}
}
```

В этом псевдокоде не учтено наследование свойств для классов. Это сделано для того, чтобы не загромождать выражений и не меняет ситуации по существу.

По предложенной схеме был выполнен ряд разработок. В частности, рассматриваемый операторский интерфейс редактирования базы данных, используемый в проекте «Фотоархив СО РАН» имеет практически именно предложенный вид).

| $oprop/label : $type/label | | | | |
|---|---|---|---|---|
| $dprops[1]/label | $dprops[2]/label | … | $oprops[1]/label | … |
| $items[1]/*[name()=<br>$dprops[1]]/text() | $items[1]/*[name()=<br>$dprops[2]]/text() | … | ViewItemInline($items[1]/*[<br>name()=$oprops[1]]/@rdf:resource) | … |
| … | | | | |

Таблица 2.

## 5 Приложение Fact-o-graph для ведения малых (фото) документных архивов

Средства, представленные в предыдущем разделе, для пользователя достаточно громоздки для простых применений. Они требуют наличия Web-сервера, интерфейс универсален, а потому не достаточно лаконичен, средства работы с фотодокументами отделены от операторского интерфейса и составляют целую технологическую цепочку. В связи с этим, для простых проектов и каждодневного использования, было спроектировано приложение Fact-o-graph.

Фактограф работает на том же поле данных, что и другие компоненты архивной системы и можно его использовать для доступа и редактирования данных и в больших проектах и совместно с интерфейсами, описанными в предыдущем разделе. Собственно это определяется тем, что ядро (см. ранее) фактографа то же самое, что и в интерфейсах. Это несколько «утяжеляет» использование приложения на пользовательском компьютере, поскольку на нем формируется полная модель данных проекта, но в большинстве случаев, это не существенно.

Рассмотрим модель данных и их распределения, использованную при создании фактографа.

База данных состоит из отдельных файлов – документов RDF. Все документы RDF соответствуют единой спецификации – базовой онтологии. Некоторые подмножества документов могут иметь расширенную спецификацию. Онтология и ее расширения задаются средствами OWL-описаний. У каждого документа есть владелец и только владелец может вносить в документ изменения. Модель устройства и функционирования такой базы данных рассмотрена в работе [2]. Документы базы данных могут быть общими и приватными. Общие документы публикуются в открытом для внешнего интернетовского доступа месте, приватные – сохраняются на компьютере пользователя. В базу данных включены также произвольные файлы документов: фотодокументы, видео, аудио, doc, txt, pdf, rtf, html и прочие текстовые и гипертекстовые документы. Все документы, включая RDF-документы, группируются в кассеты (подробнее об этом – см. далее). Собственно признак общности или приватности относится к кассетам, а не к отдельным документам.

Интерфейс фактографа сформирован на основе достаточно традиционного трех-панельного интерфейса, применяемого например в Windows Explorer. Две панели свойств располагаются в левой части плоскости окна приложения, одна – в правой. Левая верхняя панель дает информацию о «рассматриваемой» записи, на правой панели располагается список простых и сложных отношений между записью и другими записями. Левая нижняя панель активируется когда выделяется конкретное отношение для выдачи свойств этого отношения. Список реализует традиционный формы представления списка и его элементов в виде таблицы, с большими или маленькими иконками, в виде детальной композиции.

Эта же форма интерфейса сохраняется для случая поисковых запросов. При этом, левая верхняя панель используется для формулирования запроса (задания имени объекта и некоторых его свойств), в правой панели мы получаем результаты поиска в виде списка найденных записей. При выделении элемента списка, расширенная информация о нем появляется в виде списка свойств в левой нижней панели.

Также как и для Windows Explorer, можно породить несколько экземпляров окна приложения, реализованы операции drag and drop. Интерфейс системы интуитивно понятен пользователю, легко осваивается и удобен в применении.

Существенным слоем фактографа является работа с документами. Любой файл из системной области Windows можно поместить в архив, заполнить его учетную карточку, связать его с элементами базы данных. При этом файл копируется в активную кассету, при необходимости производится его предобработка для обеспечения большей удобности работы с файлом в Интернете. Собственно архив представляет собой совокупность кассет, содержащих документы и часть базы данных. Обеспечение дальнейшего жизненного цикла кассет (сохранение, резервное копирование, общее администрирование) возлагается на произвольную систему сохранения данных (хранилище данных).

На рисунке 8 приведен скриншот окон фактографа с некоторыми элементами визуального интерфейса.

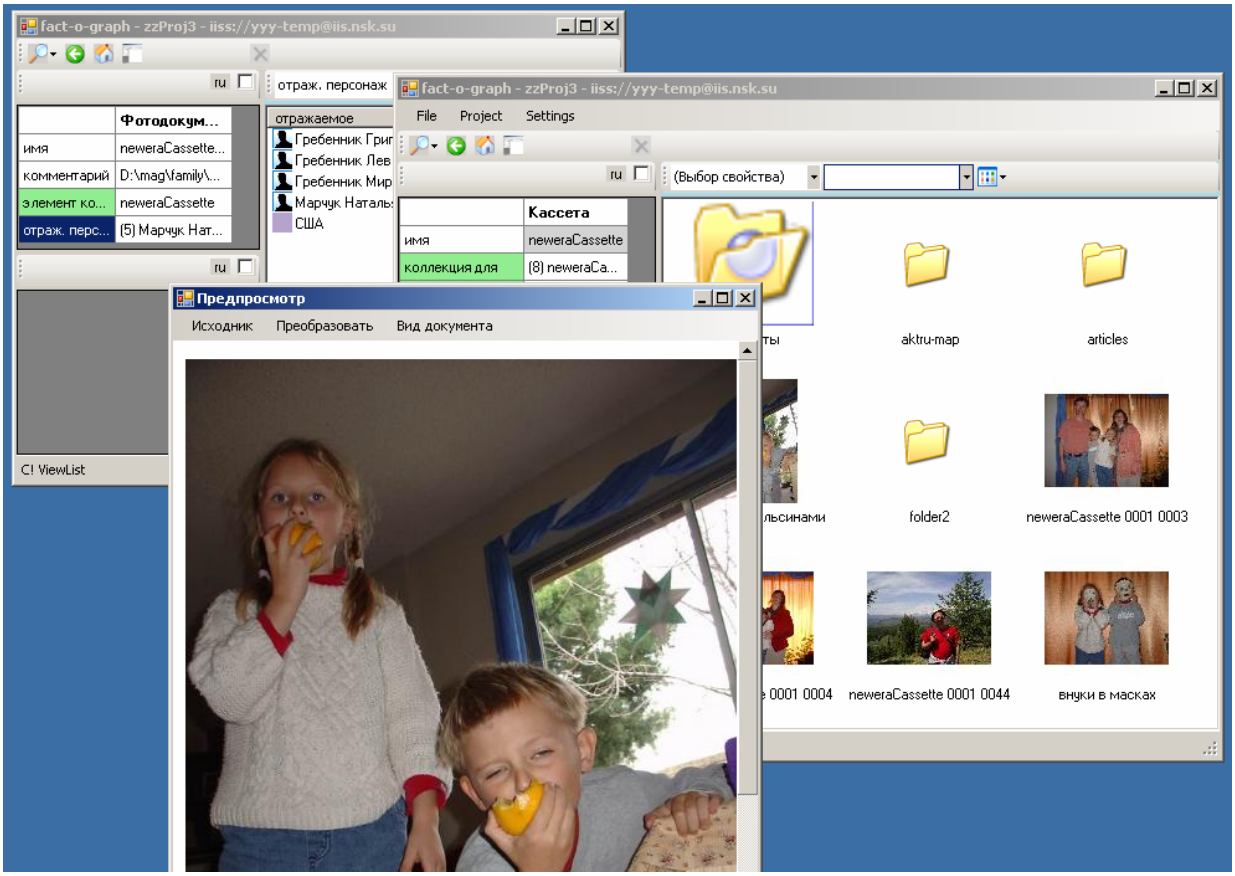



Рисунок 8

## 6 Обеспечение синхронизации редактирования в распределенной конфигурации

Как уже указывалось, в конкретной конфигурации одновременно могут быть запущены различные серверные и пользовательские приложения, разворачивающие и поддерживающие модель данных. Модель данных формируется из подмножества RDF-документов, соответственно одни и те же RDF-документы могут одновременно использоваться в разных местах. Это одновременное использование означает то, что при запуске системы документы используются приложениями для формирования фрагментов базы данных. Если документ не изменяется, то мы можем быть уверенными, что этот фрагмент идентичен для разных приложений. Теперь предположим, какой-то RDF-документ редактируется его владельцем. Ядро приложения будет обеспечивать синхронизацию фрагмента с документом путем записи новых состояний фрагмента в файл. Но теперь фрагмент перестанет быть идентичным для разных приложений. Производить слишком частую перезагрузку данных – решение неприемлемое, поэтому нужен другой механизм для обеспечения идентичности моделей.

Анализ показал, что мгновенная синхронизация, т.е. синхронизация с блокировкой процессов («остановка» времени), не требуется для рассматриваемых задач. Вполне достаточными и приемлемыми являются задержки в синхронизации на минуты и десятки минут для передачи изменений и одной модели в другие. В принципе, таймер для выполнения синхронизационных действий можно установить на малые интервалы времени, тогда можно реализовывать интерактивные сценарии типа обмена сообщениями и коллективных обсуждений.

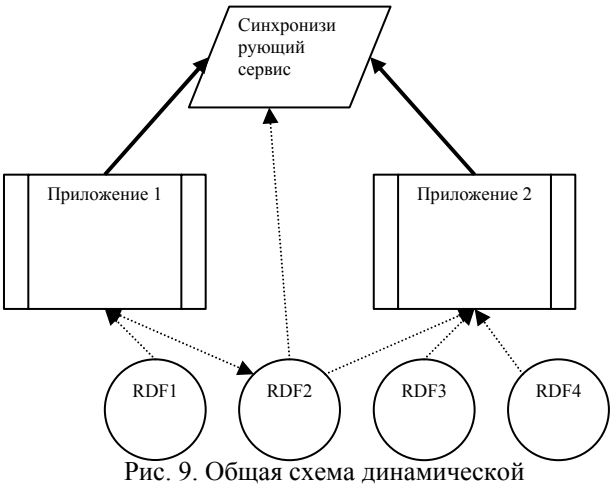



Рис. 9. Общая схема динамической синхронизации редактируемых моделей

Рассмотрим механизм динамической синхронизации моделей, содержащих одинаковые сегменты данных. На рисунке 9 изображены два

приложения, в которых имеется общий RDF-документ (RDF 2), причем приложение 1 этот документ может изменять. Пунктирными стрелками изображены связи приложений с файлами документов. Приложения читают файлы документов при инициации модели, владелец документа изменяет в приложении соответствующий образ документа и ядро приложения (через какое-то время) записывает образ в файл. Как уже указывалось, такая схема осуществляет слишком «медленную» синхронизацию – модели получают изменения только при загрузке. Поэтому вводится еще один активный участник – синхронизирующий сервис.

Сервис также, при инициации, строит у себя модель данных. Теперь пусть приложение 1 внесло изменения в образ документа 2. Это изменение сразу же пересылается «наверх» - синхронизирующему сервису. Посылается, естественно, не весь измененный документ, а только выполненное изменение. Сервис, в свою очередь, фиксирует произведенные изменения в своей модели, проставляя отметки времени на этих изменениях. Приложение 2, по таймеру или по другому «толчку», делает запрос сервису. Смысл запроса следующий: дай мне все изменения документа RDF 2, произошедшие с момента времени t. Где t, условно говоря (опустим детали) – время предыдущей синхронизации по данному документу. Синхронизирующий сервис посылает известные (пришедшие) ему изменения в данном документе, а приложение 2 фиксирует эти изменения в своей модели.

Реальный алгоритм синхронизации несколько сложней, но принципиально – он такой. Теперь достроим сервис возможностью посылать (транслировать) пришедшие от приложений изменения вышестоящему сервису и транслировать вышестоящему сервису запросы по тем документам, которые сервис «не обслуживает», соответственно возвращать ответ от вышестоящего сервиса своему клиенту. После такой достройки, можно создавать гибкие иерархические синхронизирующие конфигурации сервисов и приложений.

Данная схема синхронизации была реализована в архивной фактографической системе и используется в конфигурациях, где применяется приложение Fact-o-graph.

## 7 Технология кассет данных

Рассмотрим запись, соответствующую некоторому документу. Важнейшей составляющей записи является указание на собственно файл документа. Запись могла бы выглядеть следующим образом:

```
<document rdf:about="идентификатор
    документа">
    <coordinate>URL-документа</coordinate>
</document>
```

В этом варианте указания на файл документа, мы жестко фиксируем его координату и не позволяем альтернатив. Кроме того, все наши документы должны быть опубликованы и доступны только через серверные средства. Покажем как подход, изложенный в статье [2] может быть реализован для порождения более гибкого решения данной задачи.

Не умаляя общности, предположим, что все документы, зарегистрированные в базе данных, располагаются группами, которые мы назовем кассетами. Для каждой кассеты мы зададим ее уникальное логическое имя. А координата файла будет формироваться как конкатенация (в расширенном случае – как некоторая функция) расположения кассеты и относительного местоположения (адреса) файла в кассете.

```
file_coord = cassette_location +
    file_relative_position
```

Теперь, если мы заменим задание координаты в документе, на пару: имя кассеты, относительный адрес файла,

```
<document rdf:about="идентификатор
    документа">
    <cassette-name>имя кассеты</cassette-
    name>
    <file-relative-position>позиция
    файла</file-relative-position>
</document>
```

то мы легко вычислим текущую координату файла, имея информацию о текущей координате кассеты. Поэтому, в структуру настройки модели должны быть включены записи вида:

```
<cassette name="имя
    кассеты">cassette_location</cassette>
```

задающие таблицу текущего соответствия имен кассет их расположению.

Теперь любой, присутствующий в базе данных документ доступен из приложения, поддерживающего некоторую модель, если для этой модели имеется описание документа и если для кассеты этого документа зафиксирована координата.

Связь базы данных с пространством публикации документов будет не полной, если не дать механизма фиксации локализации кассет. В этом плане, проблему представляет множественность вариантов изображения координат файлов для разных случаев расположения файлов. Действительно, файлы могут располагаться в регулярном интернетовском пространстве, могут оказаться в локальном дисковом пространстве конкретного компьютера или на оптических дисках. Возможно даже инкапсулированное существование файлов с организацией доступа через серверный запрос. Кроме того, естественно, что к одному и тому же файлу могут вести различные адресные пути.

Рассмотрим задачу конфигурирования приложения. Точнее, нас будет интересовать только та часть, конфигурирования, которая относится к модели. Модели надо указать несколько моментов: какие RDF-документы должны быть загружены в базу данных модели, какие и как определены кассеты. Естественно, что максимально большую

часть конфигурационной информации надо погрузить в базу данных.

Во-первых, нам нужна декларация того, что существует некоторый RDF-документ, во-вторых, нам требуется путь доступа к нему, в-третьих, нужно указание необходимости или целесообразности загрузки данного документа в модель. Последнее вряд ли относится к содержимому базы данных, но остальные указания, можно превратить в запись в базе данных вида:

```
<RDF-doc rdf:about="идентификатор RDF-
    документа">
   Поля и ссылки элемента
</RDF-doc>
```

Если мы будем считать RDF-документы подклассом документов, то техника указания места публикации сохранится та же. Таким образом, если заложить такую особенность документов, как инициирование загрузки этого документа при его упоминании в базе данных, то можно считав один документ, всю остальную совокупность загружаемых документов определить по базе данных.

Тут есть одна проблема. При распределенном редактировании, мы позволяем отдельным записям «мигрировать» из одних документов в другие, это может породить нестабильный характер загрузки модели и ее неработоспособность. Другой проблемой является то, что не все RDF-документы могут оказаться совместимыми. Это, конечно, нарушает принципы единой распределенной базы данных, но практически может оказаться целесообразным. Соответственно, желательно прямо указывать конфигурационные наборы документов, требуемы для достижения тех или иных целей.

Кассеты и работа с ними, в достаточно целостном виде реализована и использована в приложении Fact-o-graph.

## 8 Заключение

Рассмотренная система компонентов и технологических решений обладает хорошей гибкостью и достаточно удобна для практического применения. Комбинируя компоненты в конфигурации, можно порождать проектные решения в спектре от индивидуальной малой архивной системы до большой системы, интегрирующей ряд проектов, имеющей протяженные пространственные и временные рамки. С помощью архивной фактографической системы или ее элементов были выполнены следующие проекты: Фотоархив СО РАН, Информационная система кафедры программирования ММФ НГУ, База данных Летних школ юных программистов, Хроника Сибирского отделения, Исторический портал ММФ, некоторые другие.

## Литература


[1] Марчук А.Г., Марчук П.А. Платформа интеграции электронных архивов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды девятой всероссийской конференции. Переславль, 2007, с. 89-94.

[2] Марчук А.Г. О распределенных фактографических системах // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды десятой Всероссийской конференции, Дубна, 2008, с. 93-102.

[3] Web Ontology Language (OWL). — http://www.w3.org/2004/OWL


## Archival Factographic System


A.G. Marchuk, P.A. Marchuk



New information system for archives was build in the Institute of Informatics Systems. System is based on factographic principles, its main features are: distributed database formed as a set of RDF-documents, special realization for distributed editing and dynamic synchronization, use of formal data specifications by OWL descriptions. System provides flexibility of data structuring, possibility of implementation of different ontologies, particularly, different metainformation set of fields. Several projects were fulfilled using proposed information system and proposed technilogies.

# Информационная система для разработки технологий организации сложной совместной деятельности♣

© Сергей М. Абрамов, Сергей В. Знаменский,
Институт программных систем имени А. К. Айламазяна РАН
abram@botik.ru svz@latex.pereslavl.ru

© Надежда С. Живчикова, Андрей В. Котомин, Елена В. Титова
УГП имени А. К. Айламазяна
ming@pereslavl.ru klein@pereslavl.ru andia@andia.pereslavl.ru

**Аннотация**

Такие сложные виды совместной творческой деятельности, как формирование программы научной конференции, экспертиза научных проектов и отчетов, разработка сложного программного обеспечения и управление образовательной активностью, могут качественно выиграть от добротной информационной поддержки.

Описан подход к созданию программной среды для поиска и разработки лучших технологий информационной поддержки организации сложной совместной творческой деятельности.

## 1 Идея и ее истоки

Сложной совместной деятельностью мы будем называть многопользовательскую активность по поддержанию и развитию информационного контента со следующими особенностями:

1. Структура контента и организация многопользовательского доступа могут потребовать непредсказуемых изменений.
2. Неожиданно может потребоваться доступ к прежним состояниям любой части контента.

Речь здесь, разумеется, идет не о немыслимом интеллекте информационной системы, сверхгибко подстраивающейся под потребности пользователей, а о включенности в эту активность группы разработчиков системы. Острая потребность в такой системе ощущается в некоторых сферах управления социальными процессами, например, при разработке инновационных технологий организации обучения. При этом полезны такие тесно взаимосвязанные инструменты совместной деятельности, как индикация изменившихся данных и важных дел, требующих участия пользователя, интерфейсы совместного доступа к данным и совместной выработки решений, алгоритмы персонализации информации и автоматического учета активности, продуктивности и дисциплинированности каждого участника. Эти инструменты очевидно не могут создаваться вне системы или независимо друг от друга и способны дать значимый эффект лишь в комплексе.

Задача создания такой системы, по-видимому, ранее не рассматривалась по причине её сложности. В случае успешного решения результат мог бы оказаться полезным и в других сферах, поскольку современные стандарты управления качеством ISO 9001:2008, ГОСТ Р ИСО 15489 ориентируют на:

- надежное сохранение истории и авторства информации,
- удобный эффективно авторизованный доступ к информации,
- возможность перестройки бизнес-процессов,

то есть, по сути, на сложную совместную деятельность в нашем понимании.

Важнейшая из сфер деятельности, где такая информационная система также обещает мощный эффект,— это разработка сложного программного обеспечения.

Информационная система, обещающая повышение эффективности творческого сотрудничества с участием разработчиков, обязана стать идеальной интегрированной средой для ее (системы) дальнейшей разработки и сопровождения.

.

### 1.1 Сохранение истории и авторства

Первые попытки организации полной истории разработки были связаны с задачами управления версиями программ и решались многочисленными системами управления ревизиями исходных кодов.

Эти попытки быстро распространились на сопроводительную документацию и на другие документы. Возможность анализа изменений и возврата к прежней версии стала основой функциональности многочисленных wiki-сайтов, включая Википедию.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

Системы отслеживания изменений стали закладываться и в основу корпоративных систем для обеспечения сохранности информации и безопасности. Например, первая в России биллинговая система с открытыми исходными кодами Nadmin [4] автоматически сохраняет в файлах под управлением RCS вместе с измененными данными дату и время изменений и аутентифицирующие пользователя данные.

### 1.2 Организация доступа к документам

В информационной системе с добротным контролем качества управление версиями должно интегрироваться с эффективными механизмами организации доступа. В wiki-системах проработаны механизмы совместного создания и редактирования документов, но мало поддержаны средства организации экспертизы и утверждения согласованной информации. Часто подобные механизмы надстраиваются над wiki-системами, оказываясь недостаточно удобными и малоэффективными.

Пятилетний опыт интенсивной эксплуатации системы Twiki (FOSwiki) [10,13], все эти годы остававшейся в списке лидеров wiki-систем, выявил полную непригодность заложенной модели структурирования информации и разграничения доступа к ней для построения на ее основе качественной информационной системы.

Главной проблемой стало оповещение эксперта, руководителя (и вообще пользователя) о появлении ожидающих срочного рассмотрения или утверждения документов или поправок.

Здесь возникли неожиданные сложности. Если право поставить визу имеет любой из группы пользователей, то все они должны видеть необходимость этого действия до тех пор, пока кто-то из них не сделает дело. Список текущих дел пользователя должен динамически отражать изменения его статуса и состояния дел в системе.

Оповещение о деле не должно отвлекать от работ по сути совершенно не связанных с выполняемой. Его видимость должна учитывать близость к текущей деятельности. В идеале у пользователя должны быть возможности контекстно влиять на видимость оповещений.

Задача организации динамического оповещения пользователя об изменениях в данных, учитывающего возможные изменения текущей организации доступа, настроек и текущей активности пользователя, оказалась трудной.

Проблема в том, что обычное сохранение данных одним пользователем может оказаться настолько важным для некоторых других пользователей, что их полезно известить об этом, не дожидаясь, пока они закончат чтение сложной страницы или заполнение большого текстового поля, родственного к этому измененному данному. Другая ситуация — практически одновременное изменение общего данного двумя пользователями в различное состояние. Если не дать знать о возникшем конфликте, то может потеряться ценная информация. Сообщения о системных неполадках и жалобы пользователей полезно оперативно и гарантированно доводить до соответствующих администраторов и разработчиков системы, причем любой получивший может выключить тревожный сигнал.

Стандартный подход передачи сообщений через рассылку неприемлем по двум причинам:

1. Если опасность уже миновала, то внимание к ней привлекать излишне.
2. Если у пользователя изменилась роль, то он должен сразу же увидеть ситуацию по-новому.

Качественный анализ этих и других возможных ситуаций при каждом обращении клиента за возможными уведомлениями требует или доступа к истории сохранения, или несоразмерных серверных ресурсов для своевременного анализа состояния данных.

Стало ясно, что какая-то часть работы должна происходить при каждом сохранении информации и учитывать текущую активность всех пользователей в системе.

Нам не удалось найти в литературе и в сети Internet ни разумных подходов к решению этой проблемы, ни ссылок на программные средства, пригодные для ее приемлемого решения.

### 1.3 Перестройки структур и процессов

Обостряющаяся проблема потерь доступа к старой информации при существенном обновлении информационных систем хорошо известна. Если при замене информационной системы заказчик поддается соблазну переноса накопленных данных в новую систему, то он почти наверняка теряет какую-то их часть и какие-то из старых интерфейсов к этим данным. Если не поддается, то он теряет больше — возможность совместной обработки старых и новых данных.

Описанные издержки становятся серьезной проблемой в ситуации, когда продолжительность актуальности данных многократно превышает среднюю продолжительность стабильности структуры организации доступа к хранимой информации. Такая ситуация особенно характерна для систем поддержки научно-образовательной деятельности.

Известные описания «эволюционирующих» или «устойчиво развивающихся» информационных систем, адресованные этой проблеме (см., например, [9–11]) указывают частные решения проблемы, поскольку в них будущее развитие проистекает в жестко фиксированных рамках.

## 2 Поиск архитектуры системы

Рассматриваемая модель построения информационных систем крайне существенно отличается от доминирующей MVC [12] и подавляющего большинства других «рациональных» моделей информационных систем: в ней организация хранения и обработки данных принципиально не может определяться фиксированной бизнес-логикой, извлечен-

ной из предпроектного обследовании. Архитектура системы должна соответствовать логике эффективного обеспечения взаимодействия сотрудников, информирования их о важных изменениях, а не тем, на какие изменения в данных направлена их совместная деятельность.

Остается один ориентир: возможность создания эффективного и дружественного пользовательского интерфейса. Рассмотрим те полезные качества, которые отсутствуют в большинстве существующих систем.

### 2.1 Контекстная автономность

Основным содержанием обсуждаемой системы является слабо структурированная информация (данные и исполняемый код). При разумной организации такой информации каждый элемент идентифицируется строкой, начало которой несет семантическую нагрузку. Мы будем называть контекстом совокупность данных с фиксированным началом строки идентификатора.

Примерами контекстов являются содержимое директории в файловой системе или на статическом сайте и множество слов, начинающиеся на 'множ' в словаре.

Назовем информационную структуру контекстно-автономной, если организация доступа к данным любого контекста (и, разумеется, сами данные) могут быть изменены независимо от остальной части системы без приостановки сервисов, в которых контекст не участвует.

Контекстная автономность характеризует гибкость структуры системы. По сути она обещает, что любую разумно выделяемую часть системы можно перестроить с ее бизнес-логики на любую другую без приостановки сервиса.

Отметим, что произвольно перестраиваемая иерархическая модель данных рассмотрена в [2].

### 2.2 Сохранение истории и доступ к ней

Поскольку перестройки структуры легко приводят к потере доступа к данным, контекстная автономность должна дополняться специфическими механизмами безопасности. В системе, отслеживающей изменения в информации, естественно организуется доступ на чтение ранее доступной информации.

В системе хранятся все версии данных и исполняемого кода (и рабочие ссылки на них) и при обработке запросов с указанием времени запроса вызывается действительная на момент запроса версия. Это гарантирует полный доступ на чтение к прежним состояниям системы. Разумеется, доступ должен быть авторизован в настоящем времени и не может фальсифицировать историю.

Вот как это может выглядеть со стороны пользователя:

Изображение странички с заданным url на экране, которое каждые пол-секунды уточняется (перерисовываются изменившиеся фрагменты).

Дополнительная верхняя строчка содержит панель управления «машиной времени»: табло с временем отображаемой странички, справа от табло кнопки ускорения и замедления прокрутки в широком диапазоне, а слева две кнопки возврата: на 5 секунд и на минуту назад. На кнопках возврата отображается время, которое появится на табло при нажатии на кнопку.

Такой просмотр (ускоренный, замедленный или в режиме реального времени) истории информационно наполненной страницы с цветными диаграммами и выделениями проблемных участков даст беспрецедентно наглядное и точное представление о динамике происходивших в данных изменений.

Добротная реализация отслеживания изменений в основе информационной системы позволит сделать возможными отмену действий пользователя (подобную undo в офисных приложениях) и быструю ликвидацию системным администратором ближайших последствий несанкционированного доступа или перехода на недоработанную версию исполняемого кода.

### 2.3 От блокировок к поточной обработке

Пользователю удобно иметь быстрый доступ к корректировке сопутствующей информации. Но чем больше доступных для изменения данных содержит страница web-интерфейса, тем больше вероятность потери никем не замеченных изменений, сохраненных одним пользователем во время работы другого над другими данными этой страницы.

Стандартный подход к исключению потерь информации это использование транзакций и блокировок. Если, например, два пользователя одновременно пытаются перевести по рублю один со счета А на счет Б, а другой со счета Б на счет А, то стандартный механизм подразумевает сохранение в логах запросов от каждого пользователя и блокировку обоих счетов при выполнении каждого из переводов. Если два разных процесса-обработчика одновременно сначала блокируют счет с которого, а потом счет, на который переводятся деньги, то получается вечная блокировка deadlock, наличие которой снижает производительность системы. Чем сложнее обработка, тем труднее своевременно выявлять и распутывать такие ситуации в автоматическом режиме. При сложных обработках для восстановления системы порой неизбежен перезапуск сервиса с потерей последних изменений.

Вместо тормозящих блокировок и слепого накладывания одних изменений на другие, используемых в целях разрешения конфликта, предлагается использовать немедленную передачу каждого законченного изменения маленького фрагмента информации на сервер с оперативной индикацией изменений (или, по меньшей мере, их наличия) у каждого из пользователей, видящих форму для изменения этих данных. Теперь мы видим, какая часть информации относительно стабильна, а какая только что изменилась другими, где другой пользователь

правит данные, а где его правки наложились на наши.

В случае, когда (как в примере с переводом денег) для конфликта нет основы, в совместных блокировках различных данных (а именно они порождают проблему) никакой нужды нет. Пользователи могут одновременно взять по рублю с одних счетов и положить на другие. При этом никто не пострадает, если гарантируется, что взятый рубль будет положен и что в момент, когда рубли взяты, но еще не положены, не произойдет вычисление суммы средств на счетах или другой операции, требующей согласованности этих данных.

Для неблокирующего разрешения обоснованных конфликтов полезны доступность история изменений каждого фрагмента информации, контекстный чат либо форум для обсуждения спорных ситуаций и разбиение большого документа на независимо изменяемые фрагменты.

Мелкая на первый взгляд особенность взаимодействия в корне меняет базовые требования к организации хранения, обработки и визуализации данных. Данные счетов в базе как и их отражение на экране браузера больше не должны отражать каждый момент консистентное состояние данных.

На самом деле стандартные строгие требования к безупречности логических взаимосвязей данных в любой момент времени ACID (см. [7]) не столь непреложны для динамически меняющейся информации. Пользователю вполне достаточно, чтобы в каждый момент времени каждая частица экрана отражала состояние, в котором соответствующие данные находились в очень близкий момент времени. Разница незаметна, если близость составляет десятую долю секунды и не раздражает даже составляя секунды, если она естественно отражает перегруженность сервера или сложность вычислений.

Аналогично и функция, обрабатывающая данные информационной системы (например, вычисляющая сумму средств на счету), просто не должна выполняться, пока все входные значения не окажутся в стабильном состоянии. Если система может гарантировать быструю актуализацию входных данных, то о последнем можно судить даже по давности их изменений. Но она в любом случае может фиксировать информацию о транзакциях (как в обычных логах) и вместо блокировки данных откладывать выполнение функции. При перегрузках сервиса это грозит тем, что сводные данные на экране пользователя будут обновляться реже, чем первичные, но такая потеря качества обслуживания намного лучше, чем перебои в сервисе.

Избежать хаотических дёрганий страницы web-интерфейса из-за незамедлительного показа изменений во время работы нетрудно. Для этого могут использоваться фрагменты фиксированного размера и иконки, сигнализирующие о наличии не просмотренных изменений.

### 2.4 Динамическая навигация

Будучи одновременно включен в сотни совместных дел, пользователь должен видеть, какие из них предполагают или ожидают его немедленного участия. Если, например, в одном из них изменено сохраненное им значение, то для него может быть важно не откладывая разобраться с конфликтной ситуацией.

Подобных ситуаций может быть много и они могут в разной степени волновать пользователя и настаивать на его участии.

В любом состоянии прокрутки на экране должна быть видна иконка, отражающая важность не просмотренных пользователем изменений в его делах с учетом его персональных настроек. Эта иконка должна открывать контекстное меню, оптимизированное под текущую ситуацию, в котором были бы отражены наиболее важные для пользователя дела, наиболее близкие к текущему контексту.

### 2.5 Выбор средств реализации

Попытки частичного сочетания описанной функциональности в системе поддержки проведения научных конференций отчетливо выявила проблему производительности: выявление обновлений системы и доведение их до пользователей вылилось в обработку больших объемов рассредоточенных данных. Обоснованные сомнения в возможности получения нужной производительности при использовании SQL-сервера из-за неизбежных блокировок привели к в выводу о целесообразности опоры на Berkeley DB, обещающей в десятки раз более высокую производительность [8]. Это однозначно определило и выбор языка реализации: Perl оказался по сути единственным скриптовым языком для высокопродуктивных (под Apache ModPerl2) веб-приложений с полноценной поддержкой Berkeley DB.

Однозначность выбора средств реализации привела к архитектуре, ориентированной на выбранные средства.

## 3 Архитектура и особенности реализации

Поскольку принципы и средства такой информационной поддержки должны уточняться и прорабатываться в ходе эксплуатации системы, то она базируется на идеях эволюционного прототипирования [9] и должна стать удобной интегрированной средой для саморазработки.

Клиентская часть системы представляет собой обычный современный браузер с включенной поддержкой javascript (при работе без javascript рекомендуются периодические перезагрузки страницы чтобы не упустить важную информацию).

Клиентские скрипты

- регулярно (несколько раз в секунду при измененных данных, раз в несколько секунд при длительном бездействии пользователя) пересылают на сервер асинхронный запрос,

содержащий идентификаторы фрагментов страницы и изменения в данных, не фиксированные в базе серверных данных;

- получив в ответе обновления, показывают их на странице без ее полной перезагрузки;
- получив хеш последних из зафиксированных на сервере в течение последней минуты значений записанных пользователем данных, сверяют их с посылаемыми серверу и удаляют принятые сервером из списка отправляемых.

### 3.1 Компоненты серверной части

***Загрузчик*** анализирует запрос и запускает код основной процедуры обработки конкретного запроса.

Задачей загрузчика является разбор параметров запроса к серверу и обращение к базе авторизации, которая должна вернуть имя той версии программного модуля из каталога, которую загрузчик должен вызвать.

Простота и неизменность кода загрузчика обеспечивают не только ретроспективную обработку с кодом и данными, актуальными на указанный момент времени, но и рабочее тестирование в той же системе следующей версии кода. Это облегчает и делает безопасными частые обновления системы и позволяет параллельно прорабатывать несколько веток кода.

***Репозиторий*** приложенных файлов хранит всю бинарную информацию. Это статическая директория, файлы которой доступны в сети Internet по именам, а листинг недоступен. Имена файлов содержат дайджест содержимого, что практически исключает возможность их угадывания неавторизованным пользователем.

***Хранилище*** текстовой и структурообразующей информации хранит текущее состояние и всю историю системы. Оно эффективно обрабатывает ретроспективные запросы с учетом пользовательских undo и учитывает записанные при тестировании данные только в тестовых запросах.

### 3.2 Базовая структура данных

Все данные, включая шаблоны и имена файлов исполняемого кода, логически хранится в базе типа BTree. Это означает, что записи в базе состоят из ключа и значения некоторого данного. Ключи записей логически образуют лексикографически упорядоченный список с двоичным деревом доступа к нему, позволяющим найти по строке первую запись, лексикографически не превосходящую строку запроса.

Ключ записи является конкатенацией строк, среди которых обычно есть

- префикс, идентифицирующий данное,
- строка обратного отсчета времени от конца XXI века.

Это позволяет очень быстро, за считанные (логарифмически растущие с числом записей в базе) микросекунды получить актуальное на любой заданный момент времени значение любого данного.

Значение данного является строкой, в которой может быть сериализована любая структура или код в Perl, ключ другой записи или имя файла с данными.

Значением может быть и директива пользовательского undo — отмены последних изменений. Директива делает неактуальными все сделанные пользователем изменения за период от указанного в ней момента времени до момента выдачи директивы и все их последствия. Среди таких изменений может быть директива undo, то есть неудачные изменения можно вернуть «по горячим следам».

Ключ записи может содержать тестовую пометку. При тестировании в системе учитываются все данные, и все изменения помечаются как тестовые и не отражаются на работе обычного пользователя, поскольку при обработке обычных запросов тестовые данные игнорируются. Это напоминает слои изображения в графическом редакторе тем, что слой тестовых данных как бы находится перед обычными слоями информации, но невидим для обычных пользователей.

Идентифицирующий данные префикс обычно является составным и формируется так, чтобы максимально использовать быстрый механизм двоичного дерева для проведения наследования данных при авторизации доступа к страничкам и выборе нужной версии шаблона.

### 3.3 Хранение и резервирование данных

Хранение данных осуществляется в базах типа BTree.

Архивная база A содержит все данные, сохраненные в системе до определенного момента. Большую часть времени она открыта только на чтение и представляет собой один большой файл, который после очередного пополнения данными архивируется стандартными утилитами работы с файлами.

Проверить сохранность старых данных может самостоятельный фоновый процесс, сверяющий сохраненные в последней и в старой резервной копии данные. Если не произошло порчи данных, то после создания очередной резервной копии можно спокойно удалить все предыдущие: все ее данные в неизменном виде содержатся в новой резервной копии. Таким образом, если учитывать место, требуемое для хранения резервный копий и логов в других информационных системах, сохранение полной истории изменений может оказаться даже более экономным в плане требований к размерам дискового пространства.

В любом случае хранение неактуальных версий файлов и другой информации безусловно связано с затратой ресурсов и естественно порождает проблему освобождения от лишней информации. Эта проблема автоматически решается выделением «белых пятен истории» (промежутков времени, история в которых недоступна) и удалением информации,

актуальной лишь в выделенных временных промежутках.

Для этого весь период времени с начала существования системы до суток назад равномерно по мере $dt/t^2$ разбивается например на 10000 промежутков и выделяется в качестве белых пятен десяток содержащих наибольшее количество записей. После этого уточняются границы белых пятен и фиксируются в особой записи базы. Все это может делать фоновый процесс, не взаимодействующий с сервером. Учитывая, что подавляющее большинство изменений вносится в периоды пиковой загрузки, пятна общей меры не превышающей один процент могут спрятать более 90% неактуальной информации.

Если в фоновом режиме составить список ключей базы, подлежащих удалению, то можно нацелить серверные процессы на чистку базы в периоды минимальной загруженности.

Если сверка в фоновом режиме содержимого актуальной и ранее сохраненной резервной копии показывает идентичность значений ключей, не покрытых белыми пятнами истории, то хранение старых резервных копий теряет смысл. Это облегчает задачу хранения резервных копий.

Буферная база B содержит данные, полученные в течение последней доли секунды или нескольких секунд и модифицируется всеми серверными процессами, обслуживающими запросы пользователей.

Перед чтением данных, переносимых в базу текущей информации C, буферная база блокируется и проверяется на управляемость. При сбое эта база не восстанавливается, а просто пересоздается, что не приводит к потерям ценной информации, поскольку скрипты браузера повторяют отправку на сервер измененных данных до получения подтверждения.

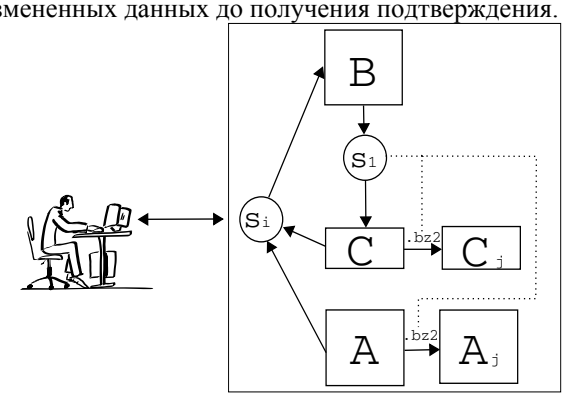



Рис. 1. Основные потоки информации

База текущей информации C содержит долю процента сохраненной в системе информации, что обеспечивает ее быстрое восстановление из резервной копии. Вероятность, что такое восстановление потребуется, сведена к минимуму тем, что модифицирует ее только один процесс и тот ничего не удаляет, а только чередует перенос подготовленных данных из буферной базы с резервным копированием базы. Резервные копии базы текущей информации реализуют редкое для баз данных инкрементное резервное копирование.

В случае гибели процесса в момент записи восстановление такой базы происходит в доли секунды и обычно даже не требует резервной копии. При необходимости просто организуется и репликация.

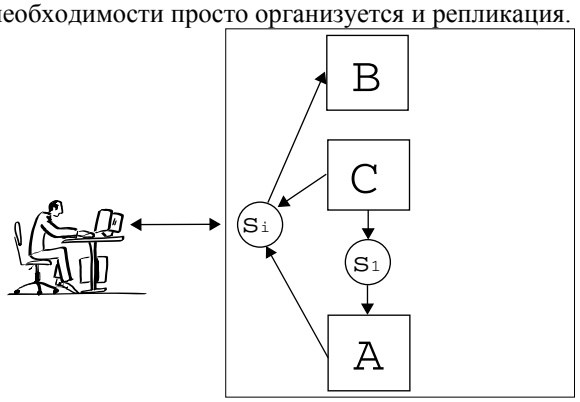



Рис. 2. Пополнение архивной базы.

### 3.4 Рабочий цикл серверного процесса

При получении запроса серверный процесс аутентифицирует пользователя и по командной строке запроса определяет, относится ли запрос к рабочей системе или к тестированию нового кода в ней. Если идет тестирование, то проверяется наличие в настройках пользователя указание на особую версию системы и подгружается либо активизируется соответствующая версия.

Далее проверяется вариант пользовательского доступа к запрошенной странице, считывается показатель загруженность сервера и показатель загруженности сервера запросами данного пользователя и выбирается вариант обслуживания. Из данных, сохраненных в базах C и A, формируется ответ (асинхронный запрос может остаться и без ответа).

Чтобы ответ на любой допустимый запрос отнял меньше ресурсов, при ответе никаких вычислений не происходит, только в сохраненные в базах шаблоны подставляются данные из этих же баз. Данные запроса сохраняются в буферной базе и ответ отправляется.

Если сервер достаточно свободен и пользователь не узурпировал слишком много серверных ресурсов, то кроме выдачи минимально возможного ответа процесс проводит очередную часть работы по уточнению пользовательского профайла и поиску интересной для пользователя информации и фиксирует в базе результаты поиска. Поскольку эта работа требует в десятки раз больше ресурсов, чем обычное обслуживание, то ее исключение амортизирует пики нагрузки, позволяя избежать привычно резкого падения производительности сервисов и катастрофического снижения надежности баз данных при пиковых перегрузках.

## 4 Заключение

Описанная система http://edu.botik.ru разрабатывается в научно-учебном комплексе ИПС имени А. К. Айламазяна РАН и УГП имени А. К. Айламазяна в г. Переславле-Залесском. С ее использованием с 2007 года систематически прово-

дятся конференции с перекрёстным рецензированием [1,5,6] и с 2008 года ведется трекинг студенческих проектов. Система ежегодно перерабатывается с сохранением функциональности. В 2009 году система перестраивается на описанных принципах с целью достижения нового качества организации учебного процесса.

## Литература


[1] Амелькин С. А., Знаменский С. В. Информационная поддержка организации сложной совместной деятельности. // Труды международной конференции “Программные системы: теория и приложения”, ИПС имени А. К. Айламазяна РАН, г. Переславль-Залесский, май 2009. Переславль-Залесский: Изд-во “Университет города Переславля”, Т.1, 2009. — с. 123–132

[2] Живчикова Н. С., Титова Е. В. Логическая модель изменчивых организационных структур. // Тезисы международной конференции "Системы проектирования, технологической подготовки производства и управления этапами жизненного цикла промышленного продукта (CAD/CAM/PDM - 2008)". М: Институт проблем управления РАН, 2008 - с.77-81.

[3] Знаменский С. В. Хорошо масштабируемое автономное администрирование доступа. // Международная конференция "Программные системы: теория и приложения", Переславль-Залесский, октябрь 2006, Наука,-Физматлит, М. Т. 1, 2006. - с. 155-169.

[4] Карлаш А. В., Абрамов С. М., Жбанов П. Г., Нестеров А. С., Ермилова Е. В., Шевчук Ю. В. Надмин - административно-расчетная система для региональных сетей. // Труды Всероссийской научной конференции " Научный сервис в сети Интернет ", 20-25 сентября 2004 г. Новороссийск, Изд-во МГУ, 2004. - М., с. 195 - 200. http://skif.pereslavl.ru/psi-info/rcmsregional.nets/regional.nets-publication/2004-rus/08-04.pdf.

[5] Коряка Ф. А.. Автоматизированная система управления вузом - UPIS. // XI научно-практическая конференция "Университета г. Переславля". Переславль-Залесский, апрель 2007, изд.-во "Университет города Переславля", Т.1, 2007. - с. 59-63. - эл. ресурс: http://wiki.botik.ru/up/pub/IS4UGP/StudConf/1-2/03-koryaka-p-59.pdf.

[6] Степанов Д. Н. Алгоритм назначения рецензентов как часть проведения научных конференций при поддержке информационной системы UPIS. // XII научно-практическая конференция Университета г. Переславля. Переславль-Залесский, апрель 2008, изд.-во "Университет города Переславля", Переславль-Залесский. Т. 1, 2008. - с. 155-169.

[7] ACID - // Материал из Википедии - свободной энциклопедии, электронный ресурс, http://ru.wikipedia.org/wiki/ACID.

[8] Berkeley DB // Материал из Википедии - свободной энциклопедии, эл. ресурс: http://ru.wikipedia.org/wiki/Berkeley_DB.

[9] Andersson Pontus; Birath David; .Serenity Patrik Willard: A Case Study in Developing Sustainable Information Systems. University essay from IT-universitetet I Gцteborg/Tilldmpad information-steknologi. Information Systems Development: Advances in Theory, Practice and Education. Edited by O. Vasilecas et al., Springer, 2005, эл. ресурс: http://www.essays.se/essay/17da5786ee/.

[10] Foswiki - The free and open source enterprise wiki. 2008-2009, эл. ресурс: http://www.foswiki.org/.

[11] Mart Roost, Karin Rava, Tarmo Veskioja, Supporting self-development in service oriented information systems, // Proceedings of the 7th Conference on 7th WSEAS International onference on Applied Informatics and Communications, p.52 - 57, August 24 - 26, 2007, Vouliagmeni, Athens, Greece.

[12] Model-View-Controller (MVC) // Материал из Википедии - свободной энциклопедии, эл. ресурс: http://ru.wikipedia.org/wiki/MVC.

[13] TWiki - the Open Source Enterprise Wiki and Web 2.0 Application Platform by Peter Toeny and contributing authors, 1998-2009, эл. ресурс: http://www.twiki.org.


## Information System for Complex Collaboration Technologies Development


S.M. Abramov, S.V. Znamenskij, N.S. Zhivchikova, A.N. Kotomin, E.V. Titova



Such creative collaboraton forms as scientific conference program development, project expertise, learning management and complex software development may benefit greatly from a good informational support.

The experimental approach to integrated development environment design for a best complex collaboration technologies research is briefly described.


---


♣ Работа поддержана грантом РФФИ № 09-07-00407.

# Управление контентом в крупных научно-технических Internet-библиотеках ♣

© Е.Е. Сальникова, ЦИТФорум elev@citforum.ru

С.А. Сальников, ЦИТФорум serg@citforum.ru

С.Д. Кузнецов ИСП РАН kuzloc@ispras.ru

**Аннотация**

Существующие системы управления Web-контентом не ориентированы на поддержку научно-технических интернет-библиотек, в которых текстовый контент может обладать произвольно большим размером. На основе многолетнего опыта поддержки и развития Internet-библиотеки CITForum формулируется набор функциональных требований, которым должна отвечать желательная система управления Web-контентом. Обсуждаются основные черты существующих систем, и отмечается их неудовлетворительность для использования в требуемых целях. Кратко рассматриваются готовые компоненты с открытым кодом, которые можно использовать при создании новой системы. Приводятся соображения относительно способов миграции контента существующих Internet-библиотек в новую среду. В заключение отмечаются другие области возможного применения разрабатываемой системы.

## 1 Введение

Полнотекстовые библиотеки научно-технической литературы являются популярными ресурсами российской части Internet. Во многих случаях эти библиотеки начали формироваться на заре Рунет, когда для публикации использовались подготовленные обычным редактором статические файлы HTML. По мере разрастания библиотек, повышения темпов их пополнения и увеличения уровня требований к качеству текстов эта унаследованная технология становится обременительной.

Желателен переход к современной технологии динамического формирования HTML-страниц на основе контента, сохраняемого в базе данных, обеспечения средств редактирования контента, выполнения требуемого администрирования и т.д. Однако, как будет показано в разд. 3, существующие системы управления Web-контентом (Web Content Management System, WCMS) не ориентированы на поддержку научно-технических интернет-библиотек, в которых текстовый контент может обладать произвольно большим размером. Требуется выполнение исследований, проектирования и разработки новой системы управления контентом.

Кроме того, сложной и трудоемкой задачей является перенос накопленного контента из унаследованной в новую WCMS. Обычно разработчики WCMS считают, что задача миграции контента должна решаться силами пользователей новой системы и не предлагают каких-либо инструментальных средств, хотя бы частично облегчающих ее решение. На Западе существуют отдельные компании, специализирующиеся на оказании подобных услуг (см., например, [15]). Естественно, в России такие услуги никем не предоставляются, и временами это приводит к результатам, печальным для пользователей соответствующих Internet-ресурсов.

В качестве наиболее близкого авторам примера можно привести Internet-библиотеку опубликованных материалов издательства «Открытые системы» [31]. С 1995 по 1998 гг. это издательство публиковало первый и до сих пор единственный в России журнал «СУБД», целиком посвященный тематике баз данных. Все выпуски журнала одновременно публиковались в бумажной форме и на сайте издательства. Опубликованные статьи были весьма востребованы читателями. Однако несколько лет назад издательство решило обновить свою систему управления контентом и, чтобы облегчить задачу миграции контента, оставило материалы журнала «СУБД» на старом сайте (несколько раз поменяв его имя). В результате интереснейшие статьи стали практически недоступны читателям, их невозможно найти через поисковые машины.

Чтобы спасти хотя бы наиболее ценные материалы (переводы ряда классических статей) в 2009 г. с согласия руководства «Открытых систем» мы заново отредактировали их и опубликовали на сайте CITForum [30], который сам остро нуждается в переходе на новую WCMS. Этот пример, во-

**Труды 11ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

первых, показывает важность надежно поддерживаемых Internet-изданий в современном мире, а во-вторых, демонстрирует нетривиальность процесса обновления их технологии.

Дальнейший материал доклада организован следующим образом. В разд. 2 формулируются функциональные требования к WCMS, пригодной для поддержки полнотекстовых HTML-ориентированных Internet-библиотек. В разд. 3. обсуждается состояние дел в области систем управления Web-контентом, и обосновывается их, как минимум, неполная пригодность для поддержки Internet-библиотек данного класса. В разд. 4 кратко описываются подходы, которые планируют использовать авторы для создания новой WCMS, удовлетворяющей поставленным требованиям. Разд. 5 посвящен обсуждению вопросов миграции контента на новую платформу. Наконец, в заключительном разд. 6 рассматриваются другие области потенциального применения новой системы.

## 2 Функциональные требования к системе управления контентом Internet-библиотеки

С 1996 г. авторы развивают и поддерживают Internet-библиотеку информационных технологий CITForum (www.citforum.ru). Исторически сложилось так, что хранимые в библиотеке документы (крупные и мелкие статьи, книги) сохраняются на сервере в файлах в формате HTML. Тщательно отслеживается, чтобы эти документы были доступны для просмотра в любом браузере.

Контент библиотеки покрывает большинство тем, относящихся к информационным технологиям: инженерия программного обеспечения, методы и средства программирования, операционные системы, базы данных, сетевые и Internet-технологии, информационная безопасность, аппаратные средства. Все разделы сайта постоянно пополняются новыми материалами (с обязательным сохранением ранее опубликованных материалов). В библиотеке публикуются материалы конференций, проводимых компанией ЦИТФорум, обзоры ведущих зарубежных журналов и т.д. На протяжении многих лет поддерживается ряд форумов: для системных администраторов, программистов и Web-мастеров и т.д.

Сайт пользуется устойчивой популярностью. В среднем каждый день его посещает 15-20 тыс. человек, в месяц – более 300 тысяч из России, Украины, Беларуссии, Казахстана, Литвы, Израиля, США, Канады и других стран ближнего и дальнего зарубежья (все материалы представлены на русском языке). На рассылку новостей библиотеки подписано 48 тыс. человек. Материалы библиотеки активно используются преподавателями и студентами университетов и ВУЗ'ов, о чем, в частности, свидетельствует значительный рост числа ее посетителей во время экзаменационных сессий.

Контент библиотеки в основном вторичен, т.е. включает документы, ранее опубликованные в печатных изданиях и/или в Internet и перепечатанные с разрешения владельцев авторских прав. Централизованное надежное хранение вторичного контента оказалось целесообразным и оправданным: к настоящему времени многие не утратившие своей ценности документы остаются доступными только в библиотеке CITForum (и на ее зеркалах). Однако нетривиальной задачей является импорт материалов, которые могут быть представлены в самых разнообразных форматах: HTML с разнообразными вариантами верстки, XML с таблицами стилей, TeX, Word различных версий, PDF и т.д. Другой очевидной (и решаемой, в лучшем случае, абсолютно вручную) проблемой в этом случае является обеспечение согласованности копии с потенциально изменяющимся оригиналом.

С другой стороны, в библиотеке CITForum размещается много оригинальных материалов, написанных или переведенных специально для публикации на этом сайте. Эти материалы обычно приходится тщательно редактировать (в ряде случаев в удаленном контакте с авторами). Без наличия средств коллективного редактирования делать это очень неудобно, громоздко и ненадежно. Кроме того, в соответствии с политикой редакции CITForum, любой материал, опубликованный в этой библиотеке, может быть перепечатан любым образом с разрешения редакции. Особенностью Internet-публикаций является потенциальная изменчивость текстов по инициативе авторов и/или редакторов. Очевидно, что в этом случае требуется поддержка версий опубликованных документов и истории их изменений.

В настоящее время программные средства, используемые для поддержки библиотеки CITForum (и большинства известных авторам других научно-технических Internet-библиотек, например, упомянутой ранее библиотеки издательства «Открытые системы» [31], насыщенной, но плохо организованной библиотеки компании «Интерфейс» [32]), не поддерживают эти и другие жизненно-важные функции. В результате, по мере расширения библиотеки,

- все труднее обеспечивать ее должное качество;
- велика трудоемкость подготовки первичных и вторичных материалов к публикации;
- трудно менять рубрикацию материалов, вводить новые разделы и т.д.;
- нелегко обеспечивать абсолютную гарантию сохранности и доступности ранее опубликованных материалов;
- все более сложной становится задача общего администрирования библиотеки.

Многолетний опыт авторов по развитию и поддержке библиотеки CITForum позволяет выдвинуть к желаемой системе управления контентом следующие функциональные требования.

### 2.1 Поддержка версий и истории изменений

Для любого хранимого в системе документа система должна:

- поддерживать требуемое число его версий,
- предоставлять данные о числе версий,
- обеспечивать доступ к каждой версии и
- выдавать подробный и понятный список изменений, обеспечивавших переход от предыдущей версии к последующей.

Другими словами, в WCMS должна иметься встроенная современная система управления версиями.

### 2.2 Удобные средства для работы с крупными документами

В полнотекстовых Internet-библиотеках предельная длина HTML-документа должна быть практически неограниченной (даже сейчас максимальная длина HTML-документа в библиотеке CitForum составляет 11 мегабайт). Для редактирования документа сейчас требуется его полная загрузка из серверного хранилища, а разбиение на страницы делается почти вручную.

В желаемой системе управления контентом:

- должна иметься возможность выполнять мелкие (локальные) изменения в хранимых документах без их полной загрузки;
- разбиение больших документов на страницы должно быть полностью автоматизировано;
- должна предоставляться возможность изменять разбиение документа на страницы, сохраняя корректность ссылок на внутренние точки документа.

### 2.3 Развитые средства форматирования

В целом требуемые средства форматирования должны отвечать общепринятым представлениям о системах управления контентом:

- в обязательном порядке должна поддерживаться вставка в отображаемый текст отдельно хранимых мультимедийных объектов (изображений, звука, видео);
- должны обеспечиваться удобные и мощные средства вставки математических формул (скорее всего, по образу основанного на TeX механизма Википедии [28]);
- менее важны, хотя и желательны, поддержка упрощенного синтаксиса разметки и наличие средства редактирования типа WYSIWYG (это менее важно, поскольку предполагается, что в данном случае с системой управления контентом будет работать ограниченное число редакторов и авторов с достаточной технической подготовкой).

### 2.4 Преобразование внешних источников к внутреннему представлению системы управления контентом

Система должна обеспечивать преобразование к своему внутреннему формату внешних документов, представленных в различных иных форматах (в частности, HTML, TeX, Word и т.д.). В системе должна обеспечиваться связь импортированного документа с первоисточником и автоматическая синхронизация при изменении первоисточника. Естественно, полностью автоматизировать преобразование из произвольного внешнего формата невозможно, поэтому следует распознавать случаи, когда изменения первоисточника требуют вмешательства редактора и изменения правил преобразования.

Подсистема импорта должна быть, прежде всего, опробована при переносе в новую систему текущего контента библиотеки CITForum, произведенного и поддерживаемого почти «вручную»..

### 2.5 Средства поиска

Поскольку речь идет о системах управления полнотекстовым контентом, в первую очередь, должны поддерживаться качественные средства полнотекстового поиска, обеспечивающие высокий уровень релевантности результатов. Кроме того, поскольку в системе управления контентом для каждого документа будут сохраняться различные метаданные, полнотекстовый поиск должен быть интегрирован с поиском документов по их метаданным.

### 2.6 Проверка распространенных ошибок

Система должна проверять разнообразные часто встречающиеся ошибки:

- синтаксические ошибки;
- опечатки; вероятно, в систему должен быть встроен спеллер;
- ошибки отображения,
  - в частности, вылезание части текста за пределы экрана на распространенных устройствах доступа;
  - соответствующее средство должно обновляться по мере распространения новых браузеров и операционных систем и регулярно запускаться по мере обновления как документов, так и самого средства;
- неработающие («битые») ссылки;
  - это средство также должно запускаться на регулярной основе.

### 2.7 Поддержка метаданных, полезных пользователям

Наряду со служебными метаданными, используемыми самой системой, должны поддерживаться метаданные с информацией, полезной пользователям:

- история документа (даты публикации, перевода, перепечатки и т.д.),
- авторы, их координаты и т.д.

По этим метаданным также должен поддерживаться поиск.

### 2.8 Поддержка тэгов как средства пользовательской навигации

В системе должны поддерживаться определяемые пользователями тэги («ярлыки») наподобие тех, которые используются на популярных сайтах типа livejournal [29], slashdot [22], gmail [5] и т.д. Наличие таких тэгов облегчает навигацию пользователя по материалам библиотеки.

### 2.9 Представления документов

Система должна обеспечивать различные внешние представления сохраняемых документов, например, в виде «одной страницы», в формате pdf, в форматах, предназначенных для разных размеров экрана и каналов в интернет (подстановка уменьшенных копий вместо картинок) и т.д.

При соблюдении перечисленных требований система управления контентом должна быть пригодна для эффективной поддержки Internet-библиотек любого размера – от библиотек масштаба CitForum до «сайтов одной книги».

## 3 Состояние дел в области систем оперативной подготовки и сопровождения Web-контента

Системы категории WCMS начали разрабатываться с середины 1990-х гг. Основное назначение этих систем состояло в том, чтобы обеспечить простые средства создания Web-контента (без потребностей в хорошем знании языков разметки и/или программирования), его редактирования, управления и сопровождения. К настоящему времени количество реально используемых WCMS составляет несколько десятков. В их число входят как коммерческие системы известных поставщиков (IBM Lotus Web Content Management [13], Microsoft Office SharePoint [20], EMC Documentum [10] и т.д.), так и многочисленные WCMS, произведенные в сообществе open source (к наиболее интересным системам можно отнести Drupal [8], Alfresco [1], Magnolia [18] и др.).

Следует отметить, что коммерческие WCMS в большой степени ориентированы на поддержку корпоративных Web-сайтов, Web-ориентированных систем е-бизнеса и т.д. Они интегрируются с системами управления документооборотом и/или бизнес-процессов, с системами управления связями с заказчиками и т.д. В коммерческих системах имеется много функциональных возможностей, совершенно не требуемых при управлении контентом Internet-библиотек, а с другой стороны, в них, как правило, не хватает ряда средств, упомянутых в разд. 2: возможности поддержки документов большого размера, коллективного редактирования и т.д.

Многие из WCMS категории open-source также в основном предназначены для поддержки корпоративных Web-ориентированных информационных систем и систем поддержки е-бизнеса. В частности, к таким системам относятся Alfresco и Magnolia. Ряд других систем этой категории, в частности, Drupal, направлен на содействие коллективному Web-творчеству. В них, безусловно, поддерживается коллективное редактирование, хотя бы примитивное управление версиями и т.д.

WCMS, поддерживающие онлайновую подготовку и коллективное редактирование контента, обычно называются «Wiki» [24]. Такие системы более близки к тому, что требуется для поддержки и развития Internet-библиотек. Возможно, при создании новой системы управления контентом стоит частично использовать код существующих Wiki – это еще не оценивалось, анализировались только возможности с точки зрения пользователя.

Тем не менее, использовать Wiki-подобную WCMS для целей поддержки полнотекстовой HTML-ориентированной Internet-библиотеки оказывается невозможно. В частности, нам неизвестны Wiki- WCMS, которые бы хорошо подходили для следующих сценариев.

### 3.1 Длинный документ (книга)

Длинным считается такой документ, который становится неудобно держать на одной странице. В единственной известной нам библиотеке Wikibooks [25], основанной на такой Wiki-подобной системе управления контентом, используется инструмент MediaWiki [12], предназначенный для поддержки энциклопедии – набора небольших статей, ссылающихся одна на другую. Насколько мы можем судить, редакторы собственно Википедии сталкиваются с заметными неудобствами, когда статья постепенно вырастает настолько, что ее части пора выделять в отдельные страницы.

В Drupal также имеется средство для подготовки, редактирования и публикации книг. Однако под книгой здесь понимается набор страниц, связанных в некоторую иерархию. Такая структура хорошо подходит для технических руководств, ответов на часто задаваемые вопросы и т.д., но совершенно неочевидно, что она будет пригодна для удобного размещения в Web произвольной книги, которая писалась в расчете на бумажное издание.

### 3.2 Документы, которые создаются в формате, не совпадающем с внутренним форматом

Практически во всех современных WCMS поддерживаются средства импорта/экспорта для

переноса контента из одной системы в другую. Однако нам не удалось обнаружить ни одной коммерческой или свободно распространяемой системы, в которую были бы интегрированы средства импорта документов, созданных вне этой системы в каком-либо внешнем формате. По всей видимости, такие средства просто не требуются в сценариях, для которых предназначены типичные WCMS. Как отмечалось выше, в WCMS, используемой для поддержки полнотекстовых HTML-ориентированных Internet-библиотек, возможность импорта внешних источников просто необходима.

В программном обеспечении Википедии имеется ряд конверторов документов, представленных во внешних форматах, в формат Wiki (в частности, из форматов HTML, Word, Open Office, LaTex) [26]. Безусловно, при разработке новой WCMS следует их оценить и, возможно, использовать. Однако нам не встречались упоминания о каких-либо инструментах, позволяющих отслеживать изменения в таких документах и помогающих соответствующим образом изменять документы, хранимые во внутреннем формате WCMS.

### 3.3 Проверка ошибок

В известных нам WCMS проверке ошибок в создаваемых или редактируемых документах уделяется незначительное внимание. И это понятно, поскольку инструменты проверки ошибок тем более актуальны, чем больше имеется возможностей поменять большое количество страниц, не посмотрев внимательно на результат. Поскольку Wiki-подобные WCMS предназначены для правки отдельных страниц, они таких инструментов и не содержат.

В онлайновых Internet-библиотеках авторами и редакторами могут править очень крупные документы, например, книги. Правки могут быть массовыми, затрагивающими несколько страниц. В этих условиях наличие развитых средств проверки различного рода ошибок становится обязательным.

## 4 Подход к созданию новой WCMS

Из всего сказанного выше легко заметить, что заметная часть нашей задачи ранее уже решалась. Перечислим крупные компоненты создаваемой системы, для которых кажется неестественным «изобретать велосипед», а можно воспользоваться готовыми решениями с открытыми кодами.

### 4.1 Системы управления версиями

«Индустриальным стандартом» современных систем управления версиями в настоящее время считается Subversion [34]. Эта система пришла на смену традиционной и широко распространенной системе CVS [4]. Наиболее привлекательные черты Subversion с точки зрения WCMS состоят в следующем:

- отслеживание изменений каталогов – под управление версиями попадают и файлы, и каталоги;
- атомарная фиксация изменений – каждый набор изменений либо попадает в хранилище целиком, либо не попадает туда вовсе;
- метаданные с версиями – у каждого файла и каталога имеется собственный набор метаданных, которые также находятся под управлением версиями;
- автоматическое (или полуавтоматическое) слияние параллельно выполняемых изменений;
- затраты ресурсов пропорциональны размеру изменений, а не размеру данных, затронутых изменениями;
- библиотеки для языков PHP, Python, Perl, Java позволяют встроить функциональность клиента Subversion в программы, написанные на этих языках.

Возможно, будут полезны эксперименты с более новыми распределенными системами Bazaar [2], Darcs [6], Git [11], Mercurial [19] и Monotone [21].

### 4.2 Библиотеки для работы со сложными форматами (языки разметки, форматы изображений и т.п.)

Очень большое количество модулей для работы со сложными данными содержится в библиотеке CPAN [3]. Эта библиотека предназначена для поддержки разработки приложений на языке Perl, хорошо поддерживается, легко доступна. Наличие в CPAN модулей, облегчающих разработку пользовательских интерфейсов, поддерживающих использование программ, написанных на других языках, обеспечивающих простой доступ из среды Perl к готовым коммерческим и свободно доступным системам стимулирует применение Perl как основного языка разработки новой WCMS.

### 4.3 Средства создания развитых Web-интерфейсов на языке JavaScript

Для разработки Web-интерфейсов в настоящее время принято использовать смесь технологий XML и JavaScript, получившую название AJAX. Имеется большой выбор реализаций AJAX с открытым кодом (см., например, [9]). Кроме того, доступен ряд редакторов rich text [7], которые позволяют в онлайновом режиме в стиле WYSIWYG редактировать контент Web-сайтов.

### 4.4 Верстка формул

Признанным средством верстки формул является TeX (см., например, [16]). TeX широко используется для написания математических и физических статей, и средства верстки формул годами отрабатывались с учетом требований ученых.

### 4.5 Wiki

Как отмечалось ранее, код существующих Wiki нами еще не изучался, но имеется надежда, что удастся использовать, по крайней мере, часть кода

- UseModWiki [23] – системы управления контентом, которая раньше использовалась в Википедии; в ее плагинах реализованы полезные функции, например, вставки формул и/или
- IkiWiki [14] – здесь привлекает представление разработчиков о хранении и публикации: документы хранятся в "настоящей" системе контроля версий – по умолчанию в Subversion, и компилируются в статический HTML.

## 5 Как будет производиться миграция на новую систему управления контентом и что это даст

Перенос отдельных документов стоит пытаться автоматизировать целиком или почти целиком – кроме отдельных документов, сверстанных совсем плохо.

Очевидно, нужно сделать и массовый инструмент, автоматизирующий обработку большинства документов с существующего сайта. Насколько это большинство будет подавляющим – зависит от того, насколько последовательно поддерживался исходный сайт. Для библиотеки CITForum, с учетом некоторого опыта автоматической обработки, можно предположить, что более 90% контента удастся перенести именно полностью автоматически. При этом нужно стремиться делать этот обработчик достаточно общим, чтобы можно было его применять и к другим сайтам.

Очевидно, что после завершения процесса миграции будет существовать заметный период доработки и тестирования (до нескольких месяцев), когда исходная библиотека будет жить в существующем формате и своей жизнью, а новая система управления контентом должна будет учитывать происходящие там изменения и без повторяющейся ручной работы соответствовать текущей версии.

Переход на новую систему (внутри и снаружи) позволит значительно сократить трудозатраты на публикацию документов. Возможно, станет целесообразно публиковать материалы, за которые раньше не брались из-за их чрезмерной сложности. (Полу)автоматическая синхронизация даст возможность во многих случаях избежать устаревания документов – в библиотеке CITForum в настоящее время это причиняет довольно серьезные неприятности. Упрощение и систематизация разметки, автоматическая проверка ошибок улучшит качество сайта. Поиск и тэги, альтернативные внешние представления – это просто новые возможности для читателей.

## 6 Заключение. Кому и как может быть полезна новая система

Прежде всего, как отмечалось выше, создаваемая система должна быть удобна не только для сайта с вторичным контентом, но и как один из вариантов системы, в которой контент создается. Здесь нам определенно пригодится опыт, полученный во время публикации книг [27] и [33].

Хорошее решение задачи будет полезно не только для сайтов, но и вообще для работы над документами, в том числе, в случае, когда публикация собственно в формате HTML не очень интересна.

Например, думается, что создаваемое решение пригодится для подготовки документов в тех областях, для которых сейчас принято использовать LaTeX с обычным текстовым редактором (научные работы по математике и т.п.). Существующему решению на основе TeX, очевидно, не хватает следующих возможностей, приблизительно в порядке важности:

- удобства совместной работы над документом;
- удобных средств централизованного хранения готовых документов и проектов в доступном для членов команды (или публично, в зависимости от конкретного применения) месте, включая поиск по этим документам;
- простоты установки и использования: при работе с TeX возникают разнообразные тонкие проблемы, особенно если переносить исходный текст на другую систему – в такой ситуации просто напрашивается серверное решение;
- функций типа WYSIWYG, особенно вставки изображений[1].

Полностью открытый характер нашего проекта и использование лицензии open source, безусловно, помогут найти и другие области полезного применения создаваемой системы управления контентом.

## Литература


[1] Alfresco. http://www.alfresco.com/

[2] Bazaar. http://bazaar-vcs.org/

[3] CPAN: Comprehensive Perl Archive Network. http://www.cpan.org/

[4] CVS – Concurrent Versions System. http://www.nongnu.org/cvs/

[5] Справка – Gmail – Ярлыки. http://mail.google.com/support/bin/topic.py?hl=ru&topic=12881

[6] Darcs. Distributed. Interactive. Smart. http://darcs.net/

[7] Don Albrecht. 5 Open Source Rich Text Editors, 2007. http://www.ajaxbestiary.com/2007/08/14/5-open-source-rich-text-editors/

[8] Drupal community. http://drupal.org

[9] Edmon Begoli. An Open Source AJAX Comparison Matrix, 2006.

http://www.devx.com/AJAXRoundup/Article/33209
[10] EMC Documentum. http://www.emc.com/products/family/documentum-family.htm
[11] Git – the fast version control system. http://git-scm.com/
[12] How does MediaWiki work? http://www.mediawiki.org/wiki/How_does_MediaWiki_work%3F
[13] IBM Lotus Web Content Management. http://www-01.ibm.com/software/lotus/products/webcontentmanagement/
[14] IkiWiki. http://ikiwiki.info/
[15] Kyle Short. No Small Task: Migrating Content to a New CMS, 2008. http://www.cmswire.com/cms/web-publishing/no-small-task-migrating-content-to-a-new-cms-002437.php
[16] LaTeX – A document preparation system. http://www.latex-project.org/
[17] LyX – The Document Processor. http://www.lyx.org/
[18] Magnolia. http://www.magnolia-cms.com/
[19] Mercurial. http://www.selenic.com/mercurial/wiki/
[20] Microsoft SharePoint. http://sharepoint.microsoft.com/Pages/Default.aspx
[21] Monotone. http://www.monotone.ca/
[22] Slashdot FAQ. http://slashdot.org/faq/
[23] UseModWiki. http://www.usemod.com/cgi-bin/wiki.pl
[24] What Is Wiki. http://wiki.org/wiki.cgi?WhatIsWiki
[25] Wikibooks. http://wikibooks.org/
[26] Wikipedia:Tools/Editing tools. http://en.wikipedia.org/wiki/Wikipedia:Tools/Editing_tools
[27] Алексей Федорчук. Zenwalk. Приобщение к Linux. 2008. http://citkit.ru/articles/892/
[28] Википедия:Формулы. http://ru.wikipedia.org/wiki/Википедия:Формулы
[29] ЖЖ: самое важное. http://www.livejournal.com/tour_rus/about_lj.bml
[30] Переводы классических статей по тематике баз данных, новая редакция, 2009. http://citforum.ru/database/classics/index.shtml
[31] Сайт издательства «Открытые системы». www.osp.ru
[32] Сайт компании «Интерфейс». http://www.interface.ru/home.asp
[33] Сергей Кузнецов. Базы данных. Вводный курс. 2008. http://www.citforum.ru/database/advanced_intro/
[34] Управление версиями в Subversion. http://svnbook.red-bean.com/

## Content Management in Large Technological Internet Libraries

E.E.Salnikova, S.A.Salnikov, S.D.Kuznetsov

Existing Web content management systems (WCMSs) are not designed for technological Internet libraries containing documents of unlimited size. Based on many years of experience in development and maintenance of the CifForum Internet library the authors define a set of functional requirements for a Web content management system appropriate for this task. Key features of existing systems are discussed, and their inadequacy for this kind of usage is justified. Then, open source components which can be used to build a new system are briefly considered. Finally, some considerations on possible ways to migrate content of old Internet libraries to the new environment and on other possible applications of the proposed system are presented.

♣ Работа поддерживается грантом РФФИ № 09-07-00282.

[1] Заметим, что здесь некоторое решение предлагает, например, LyX [17] – и, возможно, стоит учесть этот опыт. Впрочем, эта функция представляется ценной далеко не всегда, и, что более важно, предыдущие пункты все равно намекают на онлайновое решение, а таковые нам неизвестны.

# ДИССЕРТАЦИОННЫЙ СЕМИНАР–2

# PHD WORKSHOP–2

# Модели и принципы построения прототипа программной системы управления вузовской электронной библиотекой♣

© Зуев Д. С.

Казанский государственный университет
dzuev@ksu.ru

**Аннотация**

В статье рассматривается модель системы управления электронной библиотекой (ЭБ) вуза, описывается логическая структура и особенности реализации прототипа такой системы. Предлагается программное решение для объединения электронных коллекций разнородных информационных ресурсов.

## 1 Введение

Электронные библиотеки как новое направление научно-практической деятельности представляют собой область пересечения интересов целого ряда дисциплин, включая управление данными и документооборотом, информационный поиск, библиотечное дело, информационные системы, сети и телекоммуникации, обработка изображений, искусственный интеллект, человеко-компьютерное взаимодействие. Естественно, что в первые годы развития ЭБ, начиная с середины 1990-х годов, усилия исследователей главным образом были направлены на объединение на новой почве возможностей того инструментария, который был разработан в названных дисциплинах.

Известно, что основные задачи электронной библиотеки [4] – интеграция информационных ресурсов и эффективная навигация в них.

Под интеграцией информационных ресурсов будем понимать их объединение с целью использования (с помощью удобных и унифицированных пользовательских интерфейсов – желательно одного) различной информации с сохранением ее свойств, особенностей представления и пользовательских возможностей манипулирования с ней. При этом объединение ресурсов не обязательно должно осуществляться физически – оно может быть виртуальным. Главное – оно должно обеспечивать пользователю восприятие доступной информации как единого информационного пространства.

**Труды 11ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

Одну из важных ролей в производстве информационных ресурсов (ИР) играют вузы, которые всегда занимали передовые позиции в создании и распространении новых знаний и умений. На сегодняшний день основные вузовские информационные ресурсы, как правило, уже создаются в электронном виде, поэтому возникает проблема эффективного учета, хранения и использования таких ИР.

Сегодня большинство ЭБ – тематические и содержат в основном электронные аналоги печатных изданий, ЭБ же вуза содержит более широкий спектр информационных ресурсов. Это определяется хотя бы тем, что в вузе существует всегда не менее двух направлений деятельности – образовательная и научная, а в классических университетах, объединяющих множество научных направлений, научная ЭБ однозначно не может быть посвящена единственной тематике.

Основной особенностью информационных ресурсов ЭБ вуза является их неоднородность в различных аспектах – разнообразие сред представления (текст, числовые данные, статические изображения, видео, аудио, мультимедиа). При этом разнородные данные могут относиться к одному и тому же исследованию, однако фактически, учитывая их разнородность, не всегда могут храниться в одной и той же электронной коллекции. Таким образом, подобные информационные ресурсы должны иметь возможность объединения в коллекции по различным признакам, при этом один и тот же ресурс может состоять в нескольких коллекциях ЭБ.

Так, метаданные, описывающие электронные образовательные ресурсы (ЭОР), должны учитывать особенности предметной области ЭБ, т. е. включать элементы, специфичные для описания образовательных ресурсов. Таковой, например, является схема метаданных LOM, точнее, ее адаптация RUSLOM, учитывающая особенности российского образовательного процесса. ЭОР сами по себе также могут иметь достаточно специфичный вид и форму представления [3].

Если же рассматривать научную составляющую вузовской ЭБ, то здесь тоже существуют свои особенности. Специфика научных ЭБ подробно рассмотрена в [5,6], поэтому лишь заметим, что информационное наполнение научной составляющей

вузовской ЭБ достаточно разнородно. Практически во всех случаях основой информационных ресурсов ЭБ являются научные публикации в различных формах, однако помимо этого ЭБ так же может содержать библиографическую информацию, результаты различного рода экспериментов, наблюдений, измерений, моделирования исследуемой реальности, модели исследуемых процессов, явлений, феноменов, представленные в разнообразных формах, новостную, событийную информацию (календарь конференций и т.п.). Таким образом:

– научные ЭБ также обладают специфической предметной областью, и с этим связаны особые требования к ним;

– специфичны по содержанию и неоднородны их коллекции информационных ресурсов;

– научные ЭБ должны быть оснащены специфическими сервисами, благодаря которым они могут использоваться не только как источник информационных ресурсов, но и как полигон для непосредственных научных исследований [5].

При объединении всех электронных коллекций в одно целое (т. е. при создании новой информационной системы) нельзя не учитывать специфику вуза и изначально планировать интеграцию создаваемой электронной библиотеки в единую информационную среду образовательного учреждения.

Следовательно, возникают следующие требования, которым должна удовлетворять система вузовской ЭБ. Подробно требования к подобной системе и ее частям уже обсуждались в [2,3], поэтому лишь вкратце перечислим наиболее существенные. При разработке или внедрении готовой системы управления ЭБ нужно учитывать, что:

– система ЭБ может содержать разнородные электронные коллекции с различными профилями метаданных, каждая коллекция может иметь свой профиль метаданных, свое лингвистическое обеспечение, свое ПО.

– необходима возможность интеграции с внутренними информационными системами вуза, т. е. необходима разработка интерфейса, позволяющего из любого модуля информационной системы вуза формировать запрос к поисковой системе коллекции и передавать полученные результаты; для реализации подобного интерфейса дополнительно требуется разработка протокола передачи найденной информации между системами.

– необходима возможность интеграции с другими существующими электронными библиотеками и коллекциями, подобная задача, как правило, требует использования специальных протоколов для обмена информации и удаленного поиска; например, проект «Электронное полнотекстовое объединенное собрание» АРБИКОН (http://www.arbicon.ru/projects/epos/) предоставляет возможность подключения своего каталога для поиска по протоколу Z39.50.

– необходима возможность использования электронного каталога АБИС (в случае Казанского университета, достаточно предусмотреть возможность подключения к АБИС как к Z39.50 серверу).

## 2 Модель системы управления ЭБ

Целью работы является создание модели, алгоритмов и программного прототипа системы управления вузовской электронной библиотекой. Для достижения поставленной цели необходимо решить следующие задачи:

1. Разработать инфологическую модель системы управления ЭБ.
2. Проанализировать существующие информационные системы в этой предметной области.
3. Разработать логическую структуру программной системы ЭБ.
4. Создать прототип системы управления электронной библиотеки Казанского государственного университета (КГУ).

### 2.1 Инфологическая модель системы ЭБ

Как известно (см., например, [7]), инфологическая модель (или, иначе, ER-модель, ER-диаграмма) используется на ранних стадиях разработки проекта. Модель использует формализованный язык для описания и проектирования баз данных. Модель имеет однозначную интерпретацию, в отличие от некоторых предложений естественного языка, и поэтому здесь не может быть никакого недопонимания со стороны разработчиков.

Эта модель в наибольшей степени согласуется с концепцией объектно-ориентированного проектирования, которая в настоящий момент времени, несомненно, является базовой для разработки сложных программных систем.

В основе инфологической модели лежат следующие базовые понятия:

– сущность, с помощью которой моделируется класс однотипных объектов;

Сущность имеет имя, уникальное в пределах моделируемой системы. Так как сущность соответствует некоторому классу однотипных объектов, то предполагается, что в системе существует множество экземпляров данной сущности. Объект, которому соответствует понятие сущности, имеет свой набор атрибутов – характеристик, определяющих свойства данного представителя класса. При этом набор атрибутов должен быть таким, чтобы можно было различать конкретные экземпляры сущности.

– между сущностями могут быть установлены связи – бинарные ассоциации, показывающие, каким образом сущности соотносятся или взаимодействуют между собой.

Связь может существовать между двумя разными сущностями или между сущностью и ею же самой (рекурсивная связь). Она показывает, как связаны экземпляры сущностей между собой. Если связь устанавливается между двумя сущностями, то она

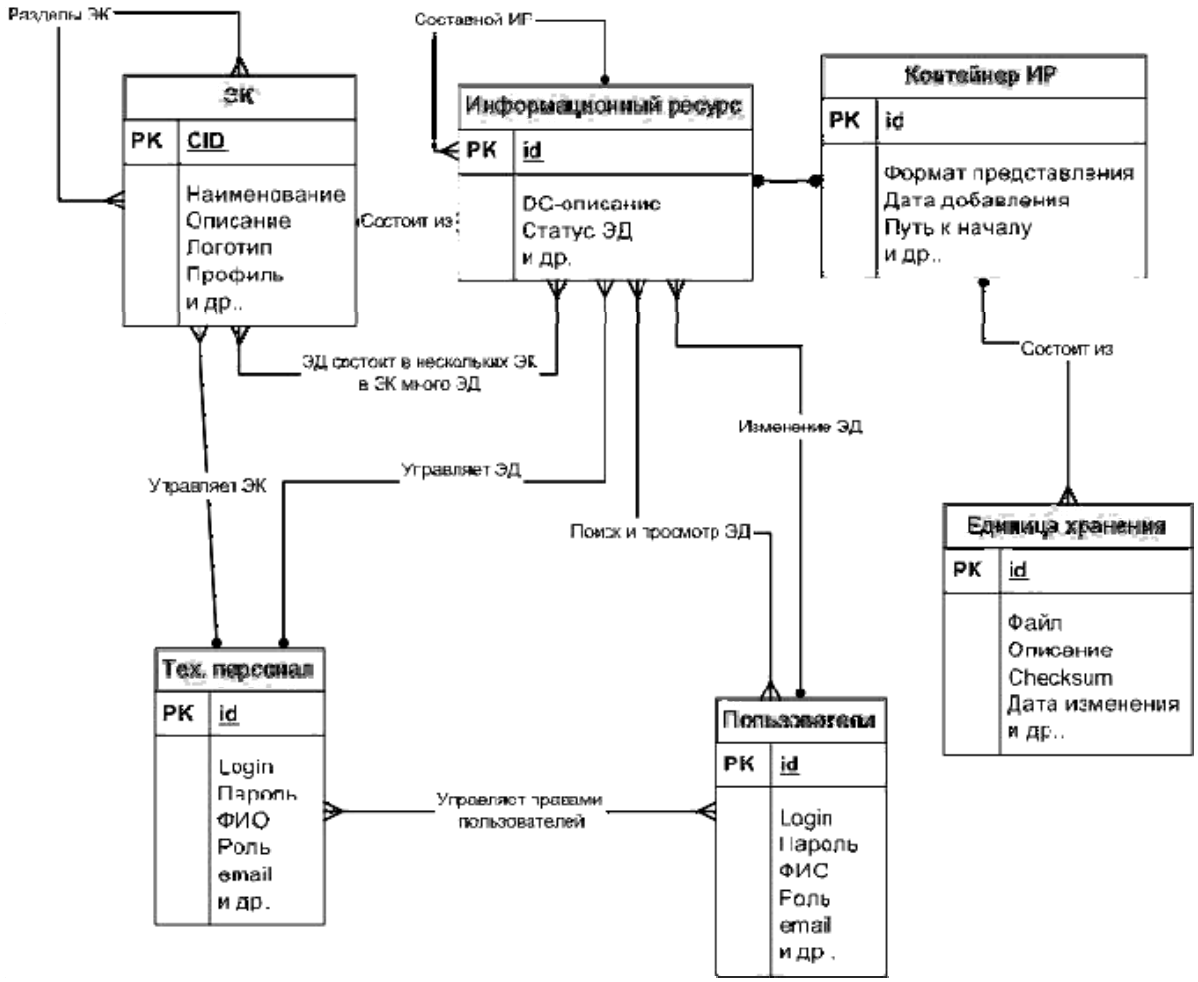



Рис. 1. Инфологическая модель

определяет взаимосвязь между экземплярами одной и другой сущности.

Рассмотрим инфологическую модель системы электронной библиотеки (рис. 1). Назовем основные сущности. Поскольку ЭБ состоит из коллекций, то разумно выделять сущность «ЭК». Она содержит уникальный идентификатор коллекции и ряд атрибутов, как-то: Наименование, Логотип, Создатель, Профиль метаданных коллекции и другие. Атрибуты коллекции должны отражать описания самой коллекции, общих свойств документов, содержащихся в ЭК, а также связи между документами и коллекцией.

Электронная коллекция состоит из разделов и электронных документов. Раздел коллекции должен содержит в точности такие же атрибуты, что и сама коллекция. Информационным ресурсом (ИР) будем называть основную единицу содержания ЭК, ИР состоит из электронного документа и его описания. Соответственно, необходимо рассматривать сущность «Информационный ресурс». Поскольку в каждой коллекции содержится множество ИР, то здесь присутствуют связи «один-ко-многим». С другой стороны, один ИР может содержаться в нескольких ЭК, т. е. ЭК и ИР связаны связью «многие-ко-многим». ИР можно представить как данные (собственно электронный документ) и метаданные, описывающие эти данные. С другой стороны, файлов с данными в электронном документе может быть несколько, однако нескольким файлам сопоставляется только одно описание ИР. Поскольку ЭБ – это еще и долговременное хранилище данных, то необходимо отслеживать все изменения не только описаний ИР, но и данных документов. Поэтому будем рассматривать сущность, назовем ее «Контейнер ИР». Каждому описанию ИР ставится в соответствие только один контейнер ресурса. Эта сущность помимо уникального идентификатора содержит ряд атрибутов, которые отвечают за целостность и изменение электронного документа (ЭД). В контейнере ЭД может содержаться несколько «Единиц хранения ИР». Это сущность, которая содержит информацию о конкретном файле или битовом потоке соответствующего электронного документа (ID, контрольную сумму, связи с другими частями ЭД, описание) и является неделимым информационным объектом. Информационный ресурс может в свою очередь может иметь также более сложную структуру, состоять из различных частей, например, журнал состоит из статей, книга состоит из отдельных глав и частей.

У электронной библиотеки существует свой круг пользователей. Поскольку у ЭБ помимо конечных пользователей практически всегда должен быть и

обслуживающий персонал, то будем выделять несколько сущностей, отображающих людей, взаимодействующих с ЭД. Первая сущность «Пользователь» – это все пользователи, использующие библиотеку. Она должна содержать информацию об уникальном идентификаторе пользователя в системе, о правах или ролях пользователя, а также некоторые другие атрибуты, связанные с пользователями системы. В зависимости от назначенной роли пользователь может управлять информационными ресурсами внутри коллекции (добавление/ изменение/удаление документов), а также выполнять поиск по коллекциям.

Помимо простых пользователей, у ЭБ, как и у любой информационной системы (ИС), должен быть технический персонал, который занимается поддержкой и развитием системы, – администраторы БД, проектировщики, системные администраторы. Совершенно ясно, что простому читателю и, например, администратору системы должен предоставляться совершенно разный функционал. Помимо этого существуют люди, которые не прикасаются к управлению функционированием системы, но должны серьезно влиять на качество предоставляемых услуг. Поэтому создадим еще одну сущность, «Технический персонал». Она содержит служебные данные об администраторах системы и другом обслуживающем персонале, о роли пользователя и правах доступа. Помимо управления коллекциями и документами администратор должен управлять всей ЭБ и ее пользователями.

Каждый человек, взаимодействующий с ЭБ, по сути, является пользователем ЭБ. Представим всех пользователей, взаимодействующих с системами электронных библиотек, следующими четырьмя различными категориями: конечные пользователи, редакторы, каталогизаторы и управляющий персонал ЭБ (администраторы, разработчики компоненты и т. п.). В зависимости от категории пользователю доступны различные функциональные возможности.

Для быстрой и корректной работы логическая модель БД, полученная из инфологической модели, должна быть приведена к нормальной форме. Технические моменты преобразования инфологической модели в нормальную форму и создания реляционной модели БД оставим вне поля зрения данной публикации, поскольку известны однозначные алгоритмы таких преобразований.

### 2.2 Жизненный цикл ИР в системе

Информационные объекты, прежде чем стать полноценной частью электронной библиотеки, должны пройти несколько стадий утверждения. Система должна отслеживать жизненный цикл информационных ресурсов, опишем его подробнее.

Автор создает новый документ – создается пустой документ, содержащий метаданные нового ИР, он получает статус «Предварительного описание». Далее производится загрузка электронного документа, так создается первичный ИР. Каталогизатор проверяет корректность описания ИР, а также соответствие самого документа и его описания требованиям библиотеки, предъявляемым к электронным ресурсам. Если качество документа не удовлетворительное, то каталогизатор может вернуть документ автору, если же качество ИР удовлетворительное, то ИР присваивается статус «Публичного Черновика», при этом может быть проведена необходимая доработка метаданных. После того, как документ получил статус «Публичного черновика», он проверяется редактором и при удовлетворительном результате получает статус «Опубликованный». С этого момента к документу предоставляется публичный доступ, и он становится полноценным элементом электронной коллекции. В противном случае документ возвращается на предыдущие шаги или вовсе удаляется из коллекции.

Администратор ЭБ создает архивные копии ИР, фиксирует его формат представления, контрольные суммы, месторасположение и т. п., также при необходимости управляет коллекциями и любыми документами в ЭБ.

### 2.3 Существующие разработки систем ЭБ

Для решения проблемы эффективности использования электронных ИР существует целый класс информационных систем – электронные хранилища, архивы, другие системы управления ЭБ. Примеры реализаций подобных электронных хранилищ существуют на практике, однако такие системы, как правило, являются строго вертикально выстроенными системами, предназначенными для хранения и управления сложными электронными объектами со строго определенным пользовательским интерфейсом (например, DSpace, ePrints, Greenstone и др.).

Для формирования более полного представления о предметной области и ее текущего состояния был проведен обзор аналогичных систем. Рассматривался ряд зарубежных аналогов, а также единственная, на данный момент, свободно распространяемая российская разработка ELSA (http://obs.ruslan.ru/?product:ELSA). Помимо собственного, для более полного анализа ряда систем использовались результаты работы групп исследователей, полученные в рамках европейского проекта DELOS [8 – 10]. собственного обзора аналогичных систем, в целом не противоречат результатам проекта DELOS. В итоге можно сделать следующие выводы о текущем состоянии дел в этой области.

У пользователей и организаций существуют специфические требования к различным системам ЭБ, особенно касающиеся функциональных возможностей отдельно взятой системы. Именно поэтому обычно отдельные электронные библиотеки создаются для специальных приложений и для конкретных целей и по этой причине не являются тиражируемым продуктом. Обычно они хорошо реализуют только тот функционал, который был необходим на момент создания ЭБ в организации, и совершенно

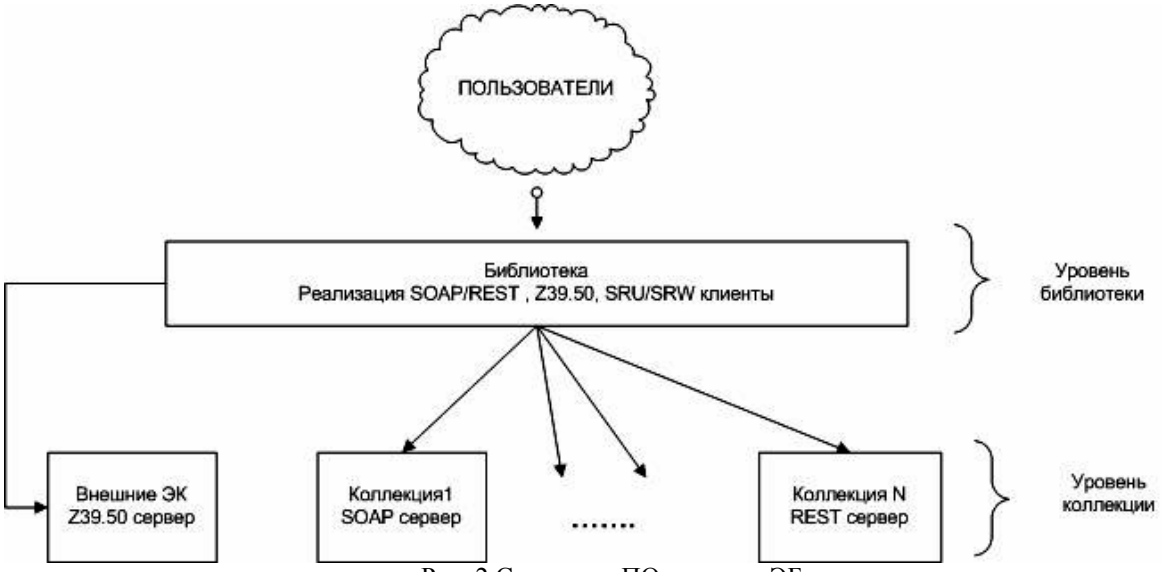


Рис. 2 Структура ПО системы ЭБ

не поддерживают весь спектр функциональных возможностей.

В настоящее время нет универсальной системы ЭБ, которая удовлетворяла бы все требования и ожидания конечных пользователей, хотя и существует много различных реализаций.

Анализ показывает разнородность систем на нескольких уровнях [9]. Из-за этой разнородности очень сложно сравнивать программное обеспечение (ПО) для ЭБ с точки зрения поддержи ими конкретных требований к ЭБ, не говоря уж о сравнительном анализе таких систем. Более того, при близком рассмотрении могут быть найдены существенные различия в системах одного и того же класса.

Системы главным образом реализуют поддержку специфических ИР и функциональных возможностей, добавление нового функционала становится очень сложной задачей.

При внедрении уже созданных систем электронных архивов или близких аналогов обычно требуется создание внутренней БД описаний электронных документов. Если организация планирует внедрение такой системы с «нуля», без какого-либо значимого ПО учета и поиска электронных документов, без имеющегося каталога электронных документов, то существующие разработки здесь подходят. Более того, они могут предоставить более полный функционал или более стабильную сборку ПО.

Однако при условии, что в организации уже существуют разнородные электронные коллекции (ЭК), то на этапе внедрения новой, пусть более мощной, но строго вертикально выстроенной системы для хранения и управления сложными электронными объектами в лучшем случае придется проводить работу по конвертации формата существующих БД ЭК в формат этой системы. В худшем же случае это повлечет за собой создание всех коллекций заново, что приведет фактически к двойной работе и лишним трудозатратам.

Таким образом, для отдельных электронных коллекций существующие готовые продукты могут подойти, однако, только представляя специфику конкретной коллекции, можно однозначно сделать вывод о приемлемости того или иного программного решения.

Например, для коллекции ЭОР или, сужая предметную область, для коллекции учебно-методических материалов большинство готовых систем электронных архивов или аналогов не подойдут, поскольку для данных коллекций необходимо использование профиля метаданных RUSLOM, а, например, пожалуй, наиболее распространенный в России электронный архив DSpace не позволяет использовать RUSLOM в качестве первичного профиля метаданных.

В Казанском государственном университете попытки создания ЭБ предпринимаются уже давно, существует достаточно много разнородных электронных коллекций, часть из них уже имеет свое собственное ПО. В КГУ созданы электронные коллекции, разнородные по своему (например, коллекция периодической печати 19-го века [1, 2], коллекция учебно-методических материалов, коллекция арабографических газет, коллекция трудов Казанских философов дореволюционного периода и т. п.), каждая из которых используется как для научных изысканий, так и для образовательной деятельности. Вдобавок в Научной библиотеке университета создан и активно пополняется электронный каталог автоматизированной библиотечно-информационной системы (АБИС). Поэтому была поставлена задача создать систему, которая позволила бы объединить существующие электронные коллекции, как собственные (со своим уже созданным ПО), так и приобретенные (внешние). К тому же было бы не разумно игнорировать каталог АБИС, который содержит описания книг, хранящихся в библиотеке вуза. Таким образом, система должна играть роль единой точки входа ко всем ИР, по крайней мере, в масштабе вуза.

## 3 Прототип системы управления ЭБ

Для реализации программных компонент системы предлагается сформировать следующую логическую структуру ПО ЭБ (рис. 2). Условно поделим все ПО для ЭБ на две части – уровень коллекции и уровень электронной библиотеки в целом. На уровне коллекции формируется ПО для отдельно взятой коллекции информационных ресурсов, на уровне библиотек производится объединение всех ЭК в одно целое.

Учитывая требования, которые предъявляются к системе, ИС электронной библиотеки должна иметь сервис-ориентированную архитектуру. Поскольку подобная система тесно связана с всемирной паутиной, получим, что разработка системы сводится к разработке набора различных веб-сервисов. Все интерфейсы, наборы точек доступа логично описывать с помощью языка WDSL, и в ИС должна быть реализована поддержка SOAP- или REST-архитектуры. При условии использования такого подхода ПО ЭК будет являться серверной частью, т. е. SOAP или REST-сервером, в то время как сервис ЭБ – SOAP или REST-клиентом.

Рассмотрим логическую структуру ПО на разных уровнях подробнее.

### 3.1 Уровень библиотеки

На уровне ЭБ должен быть реализован веб-сервис, главными целями которого являются одновременная трансляция одного и того же поискового запроса всем коллекциям и сбор результатов этого запроса. Помимо ЭК, ПО которых создано в рамках описываемой системы, в ЭБ могут входить и другие ЭК, ПО которых может быть построено по другим принципам. Для подключения коллекций, которые поддерживают идеологию веб-сервисов, достаточно иметь спецификацию точек входа.

Для более широкой интеграции с существующими электронными коллекциями Научной библиотеки КГУ, а также с учетом тенденций развития информационных технологий в российских библиотеках (они по-прежнему являются основными хранилищами электронных ресурсов) необходима поддержка использования специальных протоколов (Z39.50, SRU/SRW) для обмена информацией и удаленного поиска. В таком случае сервис представляет собой реализацию Z39.50-клиента и клиента SRU.

Разумеется, подобный сервис должен представлять инструментарий управления точками входа в ЭК, добавления и изменения ЭК в библиотеке, управления пользователями ЭБ.

Поиск по библиотеке проводится на основе единого коммуникативного формата метаданных (например, Dublin Core), формат описания ресурсов коллекций может отличаться, поэтому необходима система конвертации метаданных коллекций в коммуникативный формат и обратно там, где это необходимо.

Помимо трансляции запроса так же должна быть возможность перехода к поисковым формам отдельных коллекций. Такой веб-сервис является своего рода единой точкой входа в ЭБ, при необходимости предоставляющий переход к более узконаправленным частям системы, каковыми являются отдельные коллекции.

Необходимо заметить, что на уровне коллекции возможно использование уже существующих разработок электронных репозиториев, т. е., на наш взгляд, вполне жизнеспособна схема, когда ЭК реализуется с помощью уже готового стороннего ПО (например, Dspace).

У пользователя системы всегда есть выбор – либо произвести одновременный поиск по всем коллекциям, либо сразу перейти к определенной коллекции и использовать более широкие возможности по поиску и навигации ПО коллекции.

### 3.2 Уровень коллекции

В общем случае коллекции внутри различаются по типу хранимых электронных документов, по профилю метаданных и другим признакам.

В [1 – 3] были выработаны и обоснованы требования к ПО разнородных ЭК. Согласно этому:

- каждая коллекция имеет свой профиль метаданных;
- поиск внутри коллекций должен проводиться в соответствии с профилем метаданных этой коллекции, с использованием списков подстановок и авторитетных файлов, что влечет высокую релевантность поиска;
- создание программного обеспечения каждой электронной коллекции производится автоматически; для этого структура метаданных коллекций описывается на формальном машиночитаемом языке, причем первичный вариант структуры создается до формирования самой электронной коллекции; на основе созданного машинного описания структуры метаданных генерируется вся программная система: экранные формы, таблицы баз данных, поисковая система и т. д.

В качестве инструмента описания профиля метаданных коллекции используется XML Schema [2]. Таким образом, ПО ЭК представляет собой веб-сервис, который включает в себя реализацию REST-сервера, а также набор скриптов для анализа профилей метаданных на XML Schema и инструментарий для создания и редактирования ИР электронной коллекции.

Укажем примерный алгоритм создания ЭК. Администратор системы на вход подает файл с XML Schema профилем метаданных. Система анализирует схему, на ее основе создаются таблицы БД для хранения описаний ИР, а также поисковые формы. После этого система готова для импорта ИР или для создания описаний электронных документов. Авторизированным пользователям также предоставляется инструментарий создания новых описаний и загрузки электронных документов. Ресурсы публикуются в системе в соответствии с описанным жиз-

ненным циклом ИР. Как только ИР становится опубликованным в системе, он становится доступным для поиска и просмотра всем пользователям системы.

### 3.3 Основные результаты

На данный момент произведен анализ существующих систем электронных архивов и их аналогов, сделан вывод о неполном их соответствии поставленной задаче создания вузовской ЭБ на основе разнородных ЭК.

Разработаны инфологическая модель и логическая структура ПО системы управления ЭБ.

Разработаны принципы построения и алгоритмы работы прототипа такой системы, реализованы отдельные компоненты ПО ЭК. В качестве инструментов для программной разработки были выбраны связка СУБД MySQL и язык программирования PHP и платформа Zend Framework.

Названные результаты составляют основу кандидатской диссертации.

## 4. Заключение

Поскольку веб-сервисы – это реализация абсолютно четких интерфейсов обмена данными между различными приложениями, которые могут быть написаны не только на разных языках, но и распределены на разных узлах всемирной паутины, то при условии соответствия спецификации головной поисковый сервис ЭБ позволит подключить ЭК, реализованные с помощью совершенно другого ПО. Соответственно, ЭБ позволит объединять не только коллекции созданные на основе XML Schema, но и коллекции, реализованные с помощью ПО сторонних разработчиков.

Таким образом, прежде сильно интегрированные обособленные системы электронных архивов и обособленные разработки систем ЭБ трансформируются в набор отдельных модулей и сервисов, которые могут быть гибко объединены в различные многофункциональные системы.

## Литература


[1] Абросимов А. Г. Метаданные описания коллекции периодической печати [Электронный ресурс] // Электронные библиотеки: рос. науч. электронный журн. – 2005. – Т. 8, Вып. 2. – режим доступа: http://www.elbib.ru/ index.phtml?page=elbib/rus/journal2005/part2/Abrosimov, свободный.

[2] Абросимов А. Г., Зуев Д. С. Принцип построения программного обеспечения электронной коллекции периодической печати //Актуальные проблемы современной науки: Тр. 3-го Межд. форума (8-й межд. конф. молодых учёных и студентов). Естественные науки. Ч. 1, 2: Математика. Математическое моделирование. – Самара: Изд-во СамГТУ, 2007. – С. 78-83.

[3] Абросимов А. Г., Зуев Д. С. Научно-образовательная электронная библиотека вуза // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. 10-й Всерос. науч. конф. «RCDL'2008» (Дубна, Россия, 7 – 11 октября 2008 г.). – Дубна: ОИЯИ, 2008. – С. 374-379 – на рус. яз.

[4] Антопольский А. Б., Вигурский К. В. Концепция электронных библиотек. Электронные библиотеки: рос. науч. электронный журн. – 1999. – Т. 2, вып. 2. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/1999/part2/antopol

[5] Когаловский М. Р. Особенности научных электронных библиотек . Тезисы докладов научной конференции, посвященной 10-летию РФФИ "Электронные библиотеки и информационное обеспечение научной деятельности", Москва, 25-26 ноября 2002 г.

[6] Когаловский М.Р. Научные коллекции информационных ресурсов в электронных библиотеках. труды 1-й Всерос. науч. конф., С.-Петербург, 19–21 октября 1999. – СПб., 1999. – С : б.н. http://ict.edu.ru/ft/002340/scc_coll.pdf.

[7] Чен П. Модель «сущность-связь» – шаг к единому представлению о данных // СУБД. – 1995. – № 3. – С. 137 – 158.

[8] DELOS Workpackage 4 User Interfaces and Visualization. D4.1.1: Report on functional and non-functional digital library requirements, 2004. – http://delos.dis.uniroma1.it/docs/Delos_D4.1.1_v1.7.pdf.

[9] DELOS Workpackage 1. D1.4.1: Current Digital Library Systems: User Requirements vs Provided Functionality, 2005.

[10] DELOS Workpackage 1. D1.4.2 – Reference Model for Digital Library Management Systems, 2006.


## Models and features of prototype construction of digital library management system in institutes of higher education


Zuev D. S.



Model, logical structure and some features of creating digital library management system in institutes of higher education are discussed. Prototype of software solution to join together heterogeneous (based on different metadata profile) digital collections is suggested.



♣ Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (проект 07-01-12146)

# Информационные модели и технологии в организации работы научного сообщества по публикации и анализу коллекций исторических документов♣

© Кравцов И.В.

Петрозаводский государственный университет
ignat@drevlanka.ru

## Аннотация

Данная статья представляет собой краткое изложение основных идей и результатов диссертационного исследования, направленного на создание универсальной модели формализации информации, содержащейся в коллекциях текстов исторических документов, и построения информационной системы для упорядочивания и анализа накопленных знаний в рамках работы сетевого сообщества.

Также в работе делается попытка предложить методику построения целого класса информационных веб-систем, предназначенных для цифровой публикации документов культурного наследия нового типа – аналитических и динамических веб-публикаций.

С одной стороны – предметное поле работы достаточно широко и некоторые элементы информационных моделей и инструментов можно встретить во многих существующих Интернет-проектах. В то же время системообразующая модель организации данных и их взаимосвязей позволяет предлагать интерпретации многих существующих инструментов анализа данных, а также предлагать новые инструменты, которые сложно или невозможно построить на классических моделях организации данных в веб-системах.

## 1 Введение

### 1.1 Предметная область

Неуклонно растет объем цифровых веб-данных разных форматов, создаваемых и используемых научным сообществом. Прежде всего, это касается естественно-научных областей знаний. Необходимость соединять вычислительные мощности для обработки огромных объемов информации привела к появлению Grid-технологии и развитию e-Science – совокупности программных, технических и методологических средств для обеспечения территориально распределенных научных исследований.

В то же время гуманитарные науки также все чаще оказываются связанными с использованием больших объемов оцифрованных данных. В роли этих данных выступают коллекции текстов в корпусной лингвистике, изображения и тексты печатных источников или рукописей в истории и источниковедении, рисунки и фотографии предметов, привязанные к планам раскопок в археологии, аудиозаписи в устной истории и фольклористике – то, что уже давно получило название «массовые источники».

Достижения в области e-Science приводят к мысли об использовании тех же принципов организации распределенной работы в гуманитарных науках. Но главной задачей в этом случае выступает уже не совместное использование объединенных вычислительных мощностей, а территориальное распределение сбора и хранения самих данных, разработка стандартов для свободного обмена данными, а также сервисов, позволяющих с ними работать. В центре внимания оказывается социальная составляющая – организация работы сетевого научного сообщества.

До сих пор многие исследователи гуманитарии не хотят принимать новые информационные технологии, либо используют их в очень примитивном виде. На текущий момент одной из основных проблем использования информационных технологий является проблема выработки «единого» формата описания метаинформации и содержания текстового источника[11]. В то же время все чаще исследования по истории и лингвистике опираются на большие коллекции текстовых документов. Такие коллекции, представленные в Интернете, составляют основу для формирования сетевых сообществ исследователей, разделяющих между собой

**Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

форматы представления данных, а иногда и сами тексты для совместного изучения и редактирования.

### 1.2 Проекты сетевых публикаций

Всемирная паутина изначально создавалась как среда для научных публикаций, поэтому неудивительно, что веб-сайты, предоставляющие исходные данные для научных исследований, существуют в Сети. Самыми подходящими типами данных для веб-публикаций являются, безусловно, тексты и растровая графика, поэтому с середины 90-х годов широкое распространение получили веб-проекты, посвященные публикациям исторических документов [18-20], а также электронных копий печатных изданий [22]. Такие проекты были реализованы и в России [6,15].

Веб-сайты, посвященные научной публикации исторических документов, можно по способу представления исходной информации разделить на два типа. В одном случае публикация осуществляется в виде базы данных сканированных изображений с метаинформацией об источнике [18,19]. В другом случае сохраняются прежде всего электронные тексты источников в виде полнотекстовых реляционных баз данных или XML-документов [6,15,20]. Конечно, публикации второго типа имеют большие возможности для анализа информации, и, как правило, содержат некоторые инструменты для работы с текстами – хотя бы для отображения текстов на экране в виде, похожем на бумажную публикацию, или для организации полнотекстового поиска. Однако цели таких проектов, как правило, в обоих случаях археографические – сохранение культурного наследия, введение документов в научный оборот, апробация новых технологий публикации источников. Похожие по содержанию и целям веб-проекты объединяются в «консорциумы» [20,22], служащие основой для появления Интернет-сообществ исследователей. Но эти сообщества включают в себя не только тех, кто профессионально занимается публикациями текстовых источников, но и тех, для кого в центре внимания оказываются методы изучения текстов – исследования их структуры, выделения информации, формализации содержания, сравнения и классификации документов. Поэтому естественным направлением развития является переход от археографических к аналитическим веб-публикациям исторических документов.

В таком случае кроме самих коллекций необходимо предоставить сетевой инструментарий для работы с ними. Примером среды для совместной работы с текстами является разрабатываемая в Германии система TextGrid, ориентированная на историко-филологические исследования[23]. Для масштабного проекта «Монастериум»[21], начатого немецкими историками совместно с коллегами из Австрии, Венгрии и других стран, и посвященного созданию электронного архива документов из архивов монастырей Центральной Европы, в университете Кельна (Германия) разрабатывается специализированный редактор EditMom для совместной распределенной работы по оцифровке и ручному распознаванию текстов средневековых грамот.

Обобщение и развитие идей и успехов описанных выше проектов, а также участие в проекте создания системы «Источник» [7], предназначенной для организации работы сетевых сообществ исследователей текстовых исторических источников и реализуемой в Петрозаводском государственном университете, позволило сформировать автору работы концептуальный подход к решению информационных задач в данной предметной области.

Основная идея разработки системы «Источник» и других подобных систем - формировать открытые многофункциональные коллекции текстов, которые могут эволюционировать за счет деятельности организованного вокруг коллекций сообщества с одной стороны, и инструментария для поддержки разносторонней совместной деятельности этого сообщества.

Публикация коллекций документов вместе с методиками и результатами исследований, проведенных на основе этих документов[10], способна изменить традицию и приблизить методологию исторического и историко-филологического исследования к стандартам точных наук. Однако, несмотря на то, что исторические документы публикуются в Интернете уже давно, среди таких публикаций практически отсутствуют проекты, направленные на повышение объективности исследования путем представления его источниковой базы научному сообществу.

### 1.3 Технологии

Переход к эпохе «Web 2.0» с ее новыми формами взаимодействия пользователей и способами создания контента, а также параллельно развивающаяся XML-революция с повсеместным использованием самоописывающейся, семантической разметки текстов, не могли не затронуть научные веб-публикации текстовых документов. Во-первых, XML-технология оказалась очень удобной как основа для создания полнотекстовых баз данных для источников, не зависимая ни от аппаратного, ни от программного обеспечения пользователей. Создавать свои собственные коллекции текстовых источников, структурируя информацию с помощью XML-разметки, оказалось теперь по силам многочисленным специалистам, традиционно работающим с архивными документами. Если при этом используется какая-нибудь стандартная схема разметки, то пользователи получают возможность обмениваться созданными XML-документами и соединять их в большие коллекции. Подобная технология использовалась и раньше (SGML в проектах TEI и MEP)[20,22], просто сейчас она

стала общедоступной. Во-вторых, появившиеся в последние годы технологии организации работы сетевых сообществ позволяют предоставить возможность формирования веб-публикаций текстовых документов конечным пользователям – нетехническим специалистам в области истории, филологии, социологии и др.

### 1.4 Цели и задачи работы

Целью данной работы является попытка предложить модель организации многомерного пространства данных и знаний, необходимого для создания современной, аналитической и динамической сетевой публикации, а также архитектуру информационной системы (класса систем) с использованием этой модели. Эти абстракции охватывают собой комплексные методы и технологии автоматизации деятельности научных сообществ гуманитарных дисциплин (лингвистов, историков, источниковедов), способы сохранения исходных исследовательских источников (текстов) и результатов работы исследователей в онлайн пространстве, а также методы связывания исходной и извлеченной информации.

Решаемые задачи:

1) Разработка абстрактной модели описания структуры и семантики источников, а также окружающего их информационного поля;
2) Описание методов и технологий формализации и анализа текстов и коллекций исторических документов, отражение требований этих методов в модели системы;
3) Выработка концепции современной сетевой публикации коллекции исторических источников с учетом возможностей универсальной модели организации данных;
4) Разработка методологии и инструментария взаимодействия в сетевом сообществе;
5) Включение информационного поля сообщества в семантический веб, обеспечение связности с другими системами сети;
6) Проектирование открытой архитектуры информационной системы сообщества, состоящей из набора сервисов и информационных библиотек;
7) Проектирование хранилища данных для консолидации извлеченных из текстов знаний сообщества.

## 2 Тезисы, выносимые на защиту

### 2.1 Концепция аналитической публикации

С точки зрения предметной области традиционный подход оформления и издания научной публикации какой-либо коллекции исторических документов всегда связан с весьма продолжительным периодом времени, затраченным, как правило, одним исследователем на анализ большого количества текстовых источников. Большинство работы проделывается исследователем вручную, а в публикации фиксируется только окончательный вариант рассуждений без промежуточных выкладок. Доверие же к результатам исследований обычно базируется на авторитете автора. Кроме того, имеет место проблема практической невоспроизводимости и непроверяемости полученных результатов, так как для проверки необходимо затратить такое же или даже большее количество времени и обладать суммой знаний исследователя.

Для расширения доступности источников историки уже начали использовать пространство Интернет как площадку для размещения электронных копий документов. Публикация осуществляется в виде базы данных сканированных изображений с метаинформацией об источнике, либо сохраняются электронные тексты источников («транскрипции») в виде полнотекстовых реляционных баз данных или XML-документов. Цели таких проектов в обоих случаях археографические – сохранение культурного наследия, введение документов в научный оборот, апробация новых технологий публикации источников. По своему составу и функциональности электронные публикации на данный момент копируют бумажные аналоги или даже уступают им.

На наш взгляд, необходимо существенно шире использовать возможности информационных технологий в процессе подготовки публикаций, а также активнее использовать сетевой инструментарий для анализа оцифрованных текстов. Необходимо повсеместно использовать возможность распределенной удаленной, но в то же время, совместной работы в рамках сетевого сообщества. Сообщества, организованного вокруг специализированного инструментария, позволяющего не только проводить работу онлайн, но и дающего возможность проследить ход каждого исследования, вернуться на произвольный этап исследования, повторить цепочку анализа на примере чужого исследования. Тогда в центре внимания оказываются не сами источники, а методы изучения текстов – исследования их структуры, выделения информации, формализации содержания, сравнения и классификации документов. Меняется сама цель публикации – источники выкладываются в Интернет для обеспечения проведения исследований на их базе[3]. Тогда как оформленные результаты исследований могут стать отдельной цифровой и печатной публикацией.

### 2.2 Формализация текстов как основа сетевой публикации

Основой практически любого метода исследования текста является некоторая его формализация, то есть замена текста обобщенными

количественными показателями, качественными категориями, либо специальными моделями(графы, деревья)[13], отражающими структуру и тематику текста. Традиционным методом количественного анализа текста является контент-анализ. В этом случае текст формализуется с помощью вектора частот встречаемости составляющих его слов. Анализируя такие вектора, можно, например, решать задачи атрибуции текстов, то есть принадлежности их определенному автору, времени, литературному стилю и т. д. К распространенным методам качественного анализа относятся, например, задачи выделения в тексте ключевых слов, определения тематики текста, составления краткой аннотации документов.

На наш взгляд, использование глубокой разметки текстов с помощью технологий XML позволяет формировать произвольные графовые (сетевые) модели текстов, а также строить более емкие и комплексные инструменты анализа и проводить с помощью компьютера исследования над большими разнородными коллекциями. Кроме того, оформление в виде XML-документов различных интерпретаций или формализаций текстов позволяет считать тексты машиночитаемыми и строить на их основе инструменты автоматического семантического и структурного анализа. Это, например, построение онтологий предметных областей, семантический поиск, всевозможные классификаторы, интеллектуальные агенты, средства поддержки принятия решений на базе многомерных хранилищ данных.

Так как очевидно, что для разных методов анализа и областей применения формализованных текстов потребуются модели разного уровня сложности и детализации, необходимо предложить обобщенный способ описания необходимой единой модели формализации, используемой в системе.

### 2.3 Модель структурно-семантического пространства данных и знаний

Во многих гуманитарных дисциплинах довольно схожи методики анализа текстовых источников. Все они, так или иначе, пытаются формализовать текст, выделить необходимые категории, построить на уровне абстракции свои модели, провести с построенными моделями характерные исследования и попытаться интерпретировать результат. Для таких задач можно предложить универсальную модель описания формализованного текста и извлеченных из него знаний. Универсальная модель должна удовлетворять следующим требованиям:

- выделять произвольные единицы текста как обособленные объекты;
- формировать связь произвольного числа объектов;
- позволять строить произвольные иерархии объектов и связей;
- соотносить как объекты, так и связи с произвольными смысловыми категориями;
- привязывать к объектам и связям различные показатели (числовые, номинальные, вероятностные и пр.);
- позволять переходить от моделей текстов к моделям более высокого уровня (например, моделям коллекций текстов).

В работе предлагается в качестве такой универсальной модели концепция «структурно-семантического пространства»[9]. Данное пространство состоит из набора измерений, количество которых потенциально бесконечно, и точек в этом пространстве. Каждое измерение является фиксированным набором значений, определяет некоторую шкалу. Каждая точка отражает факт взаимосвязи значений на фиксированном наборе измерений.

При работе с текстом вводится понятие базового измерения – это отложенные на шкале слова, формирующие текст в порядке их появления. Практически любая точка в структурно-семантическом пространстве определяется хотя бы одним базовым измерением и некоторым набором других измерений. Например, для соотнесения элементов текста с определенными смысловыми категориями создается набор точек в двумерном подпространстве, где одно измерение базовое, а другое – шкала категорий объектов. Для создания связи между двумя объектами текста берется два базовых измерения и, если необходимо, измерение категорий связей, и в этом подпространстве ставится точка. Соответственно, для создания N-арной связи берется N базовых измерений.
Таким образом, задача преобразования информации о текстах из любой внутренней структуры хранения в обобщенную модель сводится к разложению информации на оси и точки в подобном пространстве.

### 2.4 Множественная разметка текстов

Система для научного сообщества исследователей будет обладать мощным научным потенциалом, если будет реализована возможность работы группы пользователей над одним историческим документом, в любых интересующих пользователей дисциплинах. Любое изменение и все исследования исходных исторических документов должны фиксироваться и сохраняться в системе. Таким требованиям вполне удовлетворяет глубокая и множественная XML разметка исходных документов[16, 17].

Все свои идеи и наблюдения исследователь выражает в виде разметки исходного документа, выбрав подходящую для конкретного изучения или же внедрив свою собственную разметку в систему. Размеченные документы сохраняются в базе данных системе в виде отдельных файлов, вариантов исходного текста.

Разметка считается множественной, так как наносится в несколько этапов. Такая разметка состоит из совокупности одноуровневых разметок, которые могут частично пересекаться между собой.

Простейшим этапом разметки является физическая разметка средневековой рукописи. Физическая разметка определяет границы и взаимное расположение словоформ относительно друг друга, а также специальные свойства источника: деление на строки, страницы, описание материала, места хранения, повреждений, встречающихся подписей и печатей и прочее.

На первичную физическую разметку накладывается вторичная и последующие, включающие в себя логические и семантические фрагменты текста. Например, одним из уровней, может быть разметка, выделяющая в тексте упоминания о персоналиях, о географических названиях (города, реки). Другим примером разметки могут выступать лексическая и синтаксическая разметки.

Фрагмент текста, разбитого на словоформы:

```
<doc id="pg002">
 <wf id="wf1">Доброродным</wf>
 <wf id="wf2">и</wf>
 <wf id="wf3">почестливым</wf>
 <wf id="wf4">паном</wf>
 <wf id="wf5">бурмистром</wf>
 ...
 <wf id="wf114">вашей</wf>
 <wf id="wf115">милости</wf>
 <wf id="wf116">листу</wf>
</doc>
```

Далее в этом же тексте выделены крупные блоки двух уровней. На первом уровне блоки «протокол» и «основной текст», на втором уровне сегменты различной семантической окраски(«тип 2», «тип 8»):

```
<doc id="pg002">
<blok type="protocol">
  <wf id="wf1">Доброродным</wf>
  ...
</blok>
<blok type="main_text">
 <seg type="2">
  <wf id="wf32">а</wf>
  ...
  <wf id="wf64">великого</wf>
  <wf id="wf65">короля</wf>
 </seg>
 ...
 <seg type="8">
  <wf id="wf103">и</wf>
  <wf id="wf104">мы</wf>
  ...
 </seg>
</blok>
</doc>
```

Далее пример выделения в тексте произвольной категории, например, персоналии:

```
<cat type="personality"> <wf id="wf102">
Ольбраха</wf></cat>
```

Еще вариант выделения персоналии:

```
<person><wf id="wf102">Ольбраха</wf>
</person>
```

А с добавлением идентификатора такое выделение становится индикатором упоминания в тексте определенного конкретного объекта, а не просто персоналии:

```
<person id="pers01"><wf id="wf102">Ольбраха
</wf></person>
```

Пример выделения индикатора события определенного типа, указание на то, что возможно в тексте речь идет о торговле:

```
<process type="trade"><wf id="wf46">товары
</wf></process>
```

Пример выделения в этом же тексте обращений к адресантам:

```
<doc id="pg002">
 <salutation>
  <wf id="wf1">Доброродным</wf>
  <wf id="wf2">и</wf>
  <wf id="wf3">почестливым</wf>
  <wf id="wf4">паном</wf>
  <wf id="wf5">бурмистром</wf>
 </salutation>
  <wf id="wf6">и</wf>
 ...
  <wf id="wf112">а</wf>
  <wf id="wf113">подлуг</wf>
 <salutation>
  <wf id="wf114">вашей</wf>
  <wf id="wf115">милости</wf>
 </salutation>
  <wf id="wf116">листу</wf>
</doc>
```

Объединении фрагментов, указывающих на один и тот же объект:

```
<doc id="pg002">
<text>
  ...
  <wf id="wf34">от</wf>
 <persona id="pers1">
  <wf id="wf35">господарь</wf>
  <wf id="wf36">наш</wf>
 </persona>
 <persona id="pers2">
  <wf id="wf37">освященный</wf>
  <wf id="wf38">великий</wf>
  <wf id="wf39">король</wf>
 </persona>
 <persona id="pers3">
  <wf id="wf40">его</wf>
  <wf id="wf41">милость</wf>
 </persona>
  <wf id="wf42">сказал</wf>
  ...
</text>
<links>
 <link id="l1" type="join" person_main="pers1"
person_relative="pers2">
 <link id="l2" type="join" person_main="pers1"
person_relative="pers3">
 ...
</links>
</doc>
```

### 2.5 Многомерное хранилище данных и многомерный анализ

При реализации системы возможно применение некоторых элементов технологии Хранилищ данных (Data warehouse), и тогда модель структурно-семантического пространства представляется в виде многомерной базы данных[9, 13].

В основе хранилищ данных лежит понятие гиперкуба, или многомерного куба данных, в ячейках которого хранятся анализируемые данные.

Факт в терминах хранилищ данных - это числовая величина, которая располагается в ячейках гиперкуба. Измерение - это множество объектов одного или нескольких типов, организованных в виде иерархической структуры и обеспечивающих информационный контекст числового показателя. Измерение принято визуализировать в виде ребра многомерного куба. Объекты, совокупность которых и образует измерение, называются членами измерений. Члены измерений визуализируют как точки или участки, откладываемые на осях гиперкуба.

В реляционном варианте реализации многомерной базы данные распределяются в таблицах двух видов.

Таблица фактов. Является основной таблицей хранилища данных. Как правило, она содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться.

Таблицы измерений. Содержат неизменяемые либо редко изменяемые данные. Каждая таблица измерений должна находиться в отношении «один ко многим» с таблицей фактов.

Схема таблиц, подходящая для разрабатываемой системы, называется «звездой» или «снежинкой». В этих схемах структура данных становится денормализованной, так как в нескольких таблицах дублируются идентификаторы таблиц измерений. Преимуществом же схем «снежинка» является сокращение время получения запросов к часто используемой информации, например, срез гиперкуба по конкретному измерению, по конкретной шкале категорий объектов. Для получения необходимой информации требуется анализ только таблицы фактов.

Измерения структурно-семантического пространства определяют размерность гиперкуба, а точки представляют ячейки гиперкуба. Основной объем данных хранится в таблицах фактов. Если рассматривать разметку текстов в терминах многомерной базы данных, в которой они хранятся, то схемы разметки будут представлены «таблицами измерений», а сама примененная разметка – «таблицами фактов». Ведущим измерением будет поле самого текста, разбитого на словоформы.

Удобство использования подобных структур в том, что при вводе новых измерений структурно-семантического пространства, необходимо лишь затронуть таблицу измерений, не изменяя всей остальной структуры данных. Также, если необходимо добавить какие-то оценочные, весовые и любые другие данные к разметкам текстов, необходимо добавить соответствующие поля в таблицы фактов (приписывание к конкретной реализации разметки текста дополнительных параметров).

В рамках системы многомерная база данных позволит консолидировать данные и знания, полученные через разметку и более удобно реализовывать сервисы анализа данных, сокращая время, требуемое на чтение и разбор XML-файлов.

Кроме того, построение хранилища данных позволит в перспективе применить к нему средства многомерного и интеллектуального анализа данных.

### 2.6 Информационная система – инструмент, образующий и поддерживающий сообщество

Очевидна необходимость при разработке информационных систем следующего шага – объединения простоты создания новой информации с помощью сетевых технологий и возможностью сделать информацию полезной и качественной с одной стороны, и с другой – способствовать формированию сообщества экспертов, модераторов накапливаемой информации на базе совместной работы в информационном пространстве. Можно сказать[4,12], развитие сообщества такой информационной системы – это постоянный процесс обучения, в котором его члены получают новые знания как в явном (explicit) формализованном виде представленной в системе информации и информационных сообщений друг другу, так и в неявном(tacit) виде - за счет совместного освоения инструментария и методик работы, передачи опыта между участниками сообщества. Работа в сообществе мыслится как работа в групповой сетевой операционной системе. Для каждого исследователя работает правило персонального управления знаниями: «то что я знаю, кого я знаю, и что знают те, кого я знаю». Сама же система является примером организации сообщества практикующих(community of practice).

### 2.7 Системы аналитической публикации в сети – это системы класса Metaweb

Среди современных тенденций развития Web как такового, то можно выделить две успешные ветви: развитие социальных сервисов, а также ветвь машиночитаемой информации Semantic Web. На наш взгляд, следующей ступенью развития будет являться соединение этих двух подходов в виде области так называемого Metaweb. Если изначальный Web соединял информацию, то Social software соединяет людей, а Semantic Web объединяет машинное знание. Metaweb будет представлять собой человеко-машинные решения для интеллектуальных задач (connects intelligence). Рассмотренный в работе подход как раз определяет описание информации и создание сервисов как

комбинацию человеко-машинного описания извлеченных из текстов знаний, а также использование ручных, автоматических и полуавтоматических инструментов работы.

### 2.8 Система современной сетевой публикации – это открытая и глобальная система

На текущий момент, даже если публикация готовится в специализированной информационной системе, такой как ИПС «Манускрипт» [15], то дополнительные возможности по работе со структурой и семантикой текстов возможны лишь в рамках этой системы, в большинстве случаев в локальном доступе. Извлечение текстов и информации о них из таких систем осложнено особенностями внутреннего формата хранения. Можно сказать, тексты являются полноправной частью системы, электронной библиотеки и не могут быть отчуждаемы. Поэтому система для сетевого сообщества обязательно должна быть открытой, а тексты и прочая информация - легко извлекаемы.

При соблюдении принципа открытости системы, она становится доступной для использования произвольными сторонними сетевыми сервисами. Особенно обширные возможности несут в себе стремительно развивающиеся инструменты Semantic Web, такие как семантический поиск.

Машинная обработка возможна в семантической паутине благодаря двум её важнейшим характеристикам:

- Повсеместном использовании универсальных идентификаторов ресурсов (URI). Традиционная схема использования таких идентификаторов в современном Интернете сводится к установке ссылок, ведущих на объект, им адресуемый (веб-страница, файл произвольного содержания). Концепция семантической паутины расширяет это понятие, включая в него ресурсы, недоступные для скачивания. Адресуемыми с помощью URI ресурсами могут быть, например, отдельные люди, города и другие географические сущности и т. д. К идентификатору предъявляются несколько простых требований: он должен быть строкой определённого формата, уникальной, а также адресующей реально существующий объект.

- Повсеместном использование онтологий и языков описания метаданных. Таких как семейство форматов, «Semantic Web family»: RDF, RDF Schema или RDF-S, и OWL.

Для создания в системе научного сообщества стандартной машиночитаемой разметки семантической публикации документа используется трансляция данных из XML-формата или хранилища данных в формат RDF.

Также в нашем случае, текстовый исторический объект или его фрагмент, имеющий уникальный идентификатор в рамках глобальной сети (URI, URN и пр.), через этот идентификатор обладает свойством распределенности. Один и тот же объект (источник) может входить в структуру нескольких сетевых информационных систем, изучаться и развиваться динамически усилиями разных людей в разных точках мира.

Понятно, что такое многообразие форм и свободное использование материалов требует регламентов работы и научного упорядочивания, но в то же время такой подход может стать инструментом источниковедения нового поколения.

### 2.9 Текст-ориентированная разработка

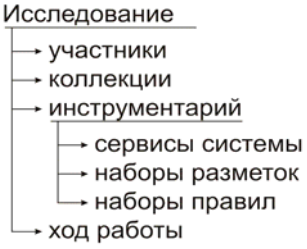
Если попытаться обосновать наш подход к разработке, то он находится в рамках парадигмы Model-driving engineering. Фактически текущим результатом работы является довольно детально описанная модель класса информационных систем. Сам процесс разработки опирается на моделирование структуры и семантики текстов и взаимосвязей текстов и прочих элементов или объектов системы. Моделью считается практически любая формализация текстов и информации сокрытой в текстах, которую мы называем знаниями. Основной технологией записи знаний и прочей информации является XML, потому моделирование определяется и частично ограничено возможностями XML. На данный момент существует большое количество различных проектов на базе XML моделей, но можно уверенно сказать, что в процессе их разработки не использовался подход, предлагаемый нами как основа разработки. Такой подход можно назвать концепцией «текст-ориентированной» разработки (text-driven), когда модули и сервисы системы проектируются так, чтобы передавать друг другу информацию в виде универсальных текстовых документов, файлов в XML-формате.

## 3. Примеры организации работы сетевых сообществ

### 3.1 Работа в системе «Источник»

Кроме описанного выше, в рамках системы «Источник» создается библиотека методик и результатов исследований коллекций текстов [8]. Цель такой библиотеки - организация хранилища аналитических публикаций, которые в своем составе содержат не только исходные источниковые материалы и описание полученного результата в виде научной статьи, но и сам инструментарий исследования с промежуточными выкладками.

Рассмотрим структуру информации в библиотеке об одном «типовом» исследовании.



Каждое исследование является фактом связывания набора составляющих его объектов и имеет собственный уникальный идентификатор и также является отдельным объектом системы.

В рамках одного исследования определяется состав участников, работающих над ним. Как правило, это руководитель исследования, который определяет состав остальных объектов исследования, и исполнители, которые получают возможность работать с определенными в исследовании объектами.

Исследование проводится на основе фиксированной коллекции документальных источников. Как правило, это полная обособленная коллекция документов.

Исследование заключается в применении к источниковому материалу специализированного инструментария: программных сервисов и алгоритмических модулей, фиксированных наборов структурных и семантических разметок, определенных на данных разметках наборов правил получения результатов.

Инструментарий применяется последовательно, согласно заранее описанному ходу работы. Например, сначала производится физическая разметка текстов (выделение словоформ), затем – структурно-семантическая разметка, после этого производится обработка полученной структуры с помощью заранее определенных правил (функций, преобразований) с целью получения новой разметки или новых правил, содержащих выявленные закономерности информации.

Результат всего исследования либо отдельной его стадии записывается в виде XML-документа, состоящего из трех частей: результата, правил, посылок. В качестве посылок (аргументов правил) выступают первично размеченные тексты, результатом является новая разметка или новые правила (закономерности).

В случае отсутствия заранее разработанного инструментария для анализа проводится «экспериментальное» исследование, результатом которого являются новые, выработанные исследователем, схемы разметки текстов или правила получения результатов.

Накопленные результаты могут быть рассмотрены как исходный материал для выдвижения и проверки новых гипотез и проведения новых исследований.

Кроме рассмотренного разреза «исследования», навигацию по библиотеке можно будет осуществлять на базе остальных составляющих ее объектов. Например, при просмотре с точки зрения участников можно будет видеть, в каких исследованиях они участвовали. Для коллекций текстов можно будет просмотреть примененные к ним разметки и правила, и, наоборот, для той или иной разметки получить примеры её использования.

### 3.2 Сообщество «Письменное наследие»

Кроме реализации в системе «Источник», планируется апробация полученных результатов при формировании сообщества исследователей древнерусских текстов. Проект предусматривает создание единого информационного портала "Письменное наследие"[1,11], в рамках которого должны быть подготовлены удобные инструменты для совместного решения учеными России и других стран актуальных сегодня технологических задач в области электронного хранения, представления в Интернете, исследования и популяризации древних и средневековых письменных памятников, для координации работ в области подготовки электронных полнотекстовых ресурсов, электронных описаний и каталогов и электронных словарей, в области выработки стандартов обмена данными. Кроме того, проект предусматривает создание организационной и технологической платформы для развития и поддержки единого исследовательского, учебного и информационного пространства, объединяющего текстовые ресурсы, справочные, аналитические и информационные материалы, рабочие группы, исследовательский инструментарий.

## 4. Заключение

Упомянутые выше, а также другие идеи и результаты, выносимые на защиту, докладывались на конференциях: RCDL (2003-2008), Современные информационные технологии и письменное наследие (2006, 2008), конференциях Ассоциации «История и Компьютер» (2006, 2008), Интернет и современное общество (2006, 2007), Научный сервис в сети Интернет (2007), Научных чтениях Даугавпилского университета (2008, 2009), и были опубликованы в работах [1-5,8-14,16,17].

## Литература


[1] Баранов В.А., Кравцов И.В. Интернет портал «Письменное наследие». Формирование сообщества исследователей древних текстов // Интернет и современное общество : Труды X Всероссийской объединенной конференции. – СПб. : Факультет филологии и искусств СПбГУ, 2007. – С. 57 – 60.

[2] Варфоломеев А.Г., Кравцов И.В, Москин Н.Д. Проект специализированного Интернет-ресурса для представления и анализа фольклорных песен // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Пятой Всероссийской научной конференции RCDL'2003. СПб, 2003. С.339-343.

[3] Варфоломеев А.Г., Кравцов И.В. Аналитические Web-публикации исторических документов // Научный сервис в сети Интернет: многоядерный компьютерный мир. 15 лет

РФФИ: Труды Всероссийской научной конференции. М.:Изд-во МГУ, 2007. С.389-390.
[4] Варфоломеев А.Г., Кравцов И.В. Приобретение и представление знаний в сетевом сообществе исследователей текстов // Вторая Международная конференция "Системный анализ и информационные технологии" САИТ-2007: Труды конференции. В 2 т. Т.1. М., 2007. С.104-106.
[5] Варфоломеев А.Г., Кравцов И.В., Филатов В.О. SVG-визуализация в цифровых библиотеках рукописных документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Девятой Всероссийской научной конференции RCDL'2007. Переславль-Залесский: Изд-во "Университет города Переславля", 2007. С.230-235.
[6] Древнерусские берестяные грамоты. Сайт проекта, 2009. http://gramoty.ru
[7] Источник. Сайт проекта, 2009. http://istochnik.karelia.ru
[8] Каргинова Н.В., Кравцов И.В., Москин Н.Д., Варфоломеев А.Г. Проект электронной библиотеки методик и результатов исследований текстовых коллекций для системы "Источник" // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Десятой Всероссийской научной конференции "RCDL'2008". Дубна: ОИЯИ, 2008. С.239-245.
[9] Кравцов И.В. Моделирование структуры и семантики текста в информационных системах для исследования исторических документов // Системы управления и информационные технологии. - № 1.1(31) – Москва-Воронеж : Научная книга, 2008. – С. 163 – 167.
[10] Кравцов И.В. О возможностях информационных технологий в подготовке публикаций и организации исследований комплексов исторических документов // Vēsture: avoti un cilvēki. XVIII Zinātniskie lasījumi. Vēsture XII. Daugavpils, 2009. P.110-114.
[11] Кравцов И.В., Багимова К.А. Модель обмена знаниями в системах гуманитарных исследований // Материалы международной научной конференции «Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам». Казань. 2008. С. 163-167.
[12] Кравцов И.В., Варфоломеев А.Г. Принципы организации информационного пространства сетевого сообщества исследователей рукописных текстов // Информационное общество. Интеллектуальная обработка информации. Информационные технологии. Материалы 7-ой международной конференции НТИ-2007. Москва : Изд-во ВИНИТИ, 2007. С.383-386.
[13] Кравцов И.В., Филатов В.О. Информационная система для работы с коллекциями рукописных исторических документов. // Информационные технологии моделирования и управления, 2007, №2(36). - С. 188-195.
[14] Кравцов И.В., Филатов В.О. Подходы к организации совместной работы научного сообщества в области публикации и исследования средневековых текстов // Интернет и современное общество. Труды IX Всероссийской объединенной конференции. – СПб: СПбГУ, 2006. – С.77-79.
[15] Манускрипт. Древние славянские памятники. Сайт проекта, 2009. http://manuscripts.ru
[16] Филатов В.О., Кравцов И.В. Технологии создания информационной системы для работы с полнотекстовыми базами данных исторических документов // Материалы международной научной конференции «Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам». Ижевск. 2006. С. 168-173.
[17] Филатов В.О., Кравцов И.В., Варфоломеев А.Г. Информационная система для работы с полнотекстовыми базами данных исторических документов на основе технологии XML // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Восьмой Всероссийской научной конференции (RCDL'2006). – Ярославль: ЯрГУ им. П.Г.Демидова, 2006. - С.337-344.
[18] Codices Electronici Ecclesiae Coloniensis (CEEC) Web site, 2009. http://www.ceec.uni-koeln.de
[19] Codices Electronici Sangallenses (CESG). Web site, 2009. http://www.cesg.unifr.ch
[20] Model Editions Partnership Web site, 2009. http://adh.sc.edu
[21] Monasterium Project Web site, 2009. http://monasterium.net
[22] Text Encoding Initiative Web site, 2009. http://www.tei-c.org
[23] TextGrid Project Web site, 2009. http://www.textgrid.de

## Information models and technologies for the web community of researchers in the field of historical documents publication and analysis

I. Kravtsov

This article represents a summary of the basic ideas and results of dissertational research. Researcher describe universal formal model of the information extracted from collections of texts of historical documents. Principles of construction of an intelligence system for marshaling and the analysis of the stored knowledge within the limits of operation of network community are besides described.

♣ Статья подготовлена в рамках проекта, поддержанного грантом РГНФ (проект № 08-01-12136в).

# Система генерации динамических web страниц

© Соломатов Владимир Юрьевич

Институт прикладных математических исследований КарНЦ РАН, Петрозаводск
solomatov@yahoo.com

**Аннотация**

Работа посвящена одной из актуальных проблем IT индустрии – автоматизации создания приложений. Автором разрабатывается система генерации динамических Web-страниц на основе реляционной структуры источника данных и шаблонов отображения данных. Система является расширяемой и не зависимой от источника данных.

## 1 Введение

В современном, быстро изменяющемся мире задача упрощения или автоматизации процесса создания информационных приложений является одной из наиболее востребованных. Тем более создания Web-систем доступа к базам данных. В российском и зарубежном сегментах глобальной сети представлено множество систем, упрощающих и ускоряющих разработку web-приложений. Большинство из них предлагают генерацию статических html страниц на основе документов известных форматов (например, MS Word или MS Excel). Данные решения больше подходят для создания домашних страниц или небольших web-отчетов. Так же существует ряд более продвинутых коммерческих предложений, например ASP.NET Maker (так же решения для PHP или JSP [1]) или ASP.NET Generator, для создания динамических web-приложений на основе сконфигурированных запросов к определенным источникам данных. Таким образом, пользователь может автоматически создать законченное web-приложение, определив лишь источник данных и сконфигурировав запросы к нему, что существенно убыстряет процесс создания приложения.

Целью работы является создание системы генерации динамических Web-приложений на основе шаблонов для реляционных источников данных с поддержкой доступа по средствам ADO.NET. В отличие от упомянутых систем автоматизации построения Web-приложений данная система является расширяемой, т. е. реализованы механизмы поддержки добавления элементов управлении и поддержка добавления новых элементов в контекст генерируемых страниц. Автоматически созданное приложение позволяет пользователю просматривать, изменять, производить поиск по записям и т. д.

## 2 Описание системы

Основной идеей является создание программной системы, способной автоматически создавать Web-приложения на основе шаблонов отображения и реляционных отношений между выделенными сущностями источника данных (Рис. 1 Схема работы системы генерации).

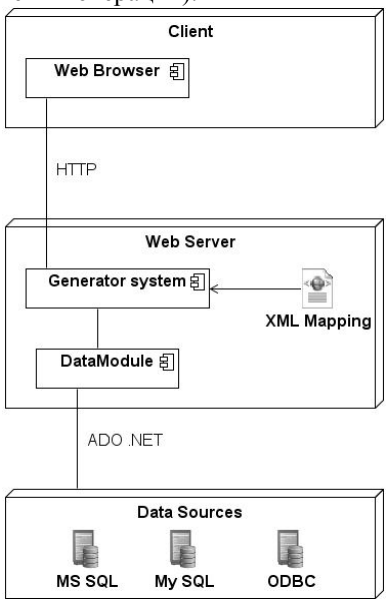


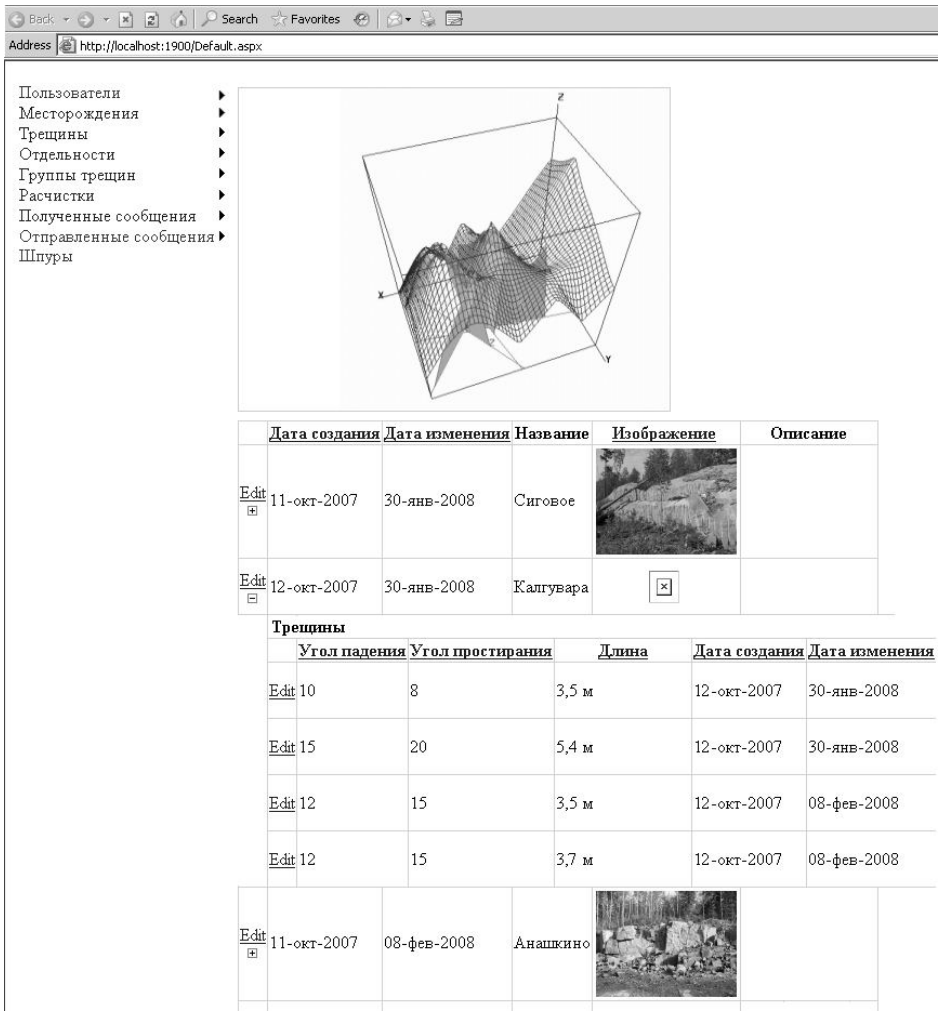

**Рис. 1 Схема работы системы генерации**

Приложение разрабатывается с использованием технологий ASP.NET и AJAX и является независимым от источника данных (используются источники данных, для которых реализована поддержка ADO.NET).

Автоматически полученная страница состоит из меню выбора нужной таблицы и самой таблицы Рис. 2. Реляционные отношения между таблицами сохранены в структуре меню и в отображаемой таблице. Для каждой выбранной таблицы, имеющей дочерние таблицы, можно выбрать непосредственно страницы по первичному ключу.

Названия пунктов меню может быть задано пользователем с помощью XML шаблона

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**




преобразования (файла конфигурации) с заданной структурой.

Для более удобного и более информативного отображения в таблице поддерживаются:

•автоматическая сортировка данных по столбцам;

•постраничное разбиение таблиц (с выбором метода разбиения);

•поиск по таблице в соответствии с выбранными колонками или по всем колонкам.

Подписи для колонок и скрытые колонки можно также определить в файле преобразований.

Возможность редактирования, удаления и добавления данных может быть определена так же в файле конфигурации. Редактирование осуществляется в соответствии с типом данных. Например, данные типа string редактируются с помощью стандартного элемента управления asp:TextBox, типа DataTime с помощью asp:Calendar и т.д. Элементы управления для каждого типа данных можно заменить на другие в соответствии с механизмами расширения. Редактирование данных из полей связанных с полями других таблиц, в целях ограничения целостности данных, осуществляется по средствам выпадающего списка. Объекты из родительских таблиц становятся элементами выпадающего списка для редактирования полей из дочерних таблиц. Данные могут быть представлены типом, не являющимся стандартным (например, Media – тип отображения видео). Для всех типов, используемых в текущем приложении, могут быть разработаны соответствующие пользовательские элементы управления отображения и редактирования данных. Добавление новых данных осуществляется аналогично редактированию.

## 3 Использование

На основе описанного генератора разрабатывается Web-система учета, моделирования и анализа систем трещин в горном массиве. Главной особенностью такой системы является то, что при моделировании трещин и их соотношений появляется возможность программными методами выделять участки, содержащие блоки правильной формы и оценивать их экономические характеристики. Основным критерием оценки в данной работе является – оценка выхода блоков заданных размером из рассматриваемого горного массива. В качестве инструмента для визуализации модели трещин в

горном массиве использовалась технология Java с поддержкой 3D [2].

## 4 Заключение

На данный момент разработан прототип данной системы, поддерживающий режимы отображения, удаления и редактирования данных с поддержкой использование расширенных элементов управления. Следующими основными шагами в создании системы будут: добавление поддержки Web-виджетов для создания расширяемого Web-портала [3] и доработка системы в соответствии с запланированными возможностями.

## Литература


[1] ASPMaker code generator site, http://www.hkvstore.com/

[2] Java 3D project site, https://java3d.dev.java.net/

[3] Zabir O. Building a Web 2.0 Portal with ASP.NET 3.5 // O'Reilly, 2007


## The system for generation of dynamic web pages


Solomatov Vladimir



The problem of simplifying software development is one of the most important in the IT industry. This paper describes a system for generating web-sites based on the relational structure of the data source and mapping schemas. This system features expandable architecture and enables users to include different custom controls for different data types into it. The Application is being developed using ASP.NET and AJAX and has no direct data source dependencies. The only requirement is compatibility of the data sources with ADO.NET.

# ПРИГЛАШЕННЫЙ ДОКЛАД

# INVITED PAPER

# Развитие электронных библиотек – путь к Открытой Науке ♣

© С.И. Паринов

Центральный экономико-математический институт РАН
sparinov@gmail.com

## Аннотация

Постоянное развитие концепции электронной библиотеки (ЭБ) формирует особую онлайновую среду для научных исследований. Успехи международной научной инициативы за открытый доступ к результатам исследований создают предпосылки для того, чтобы использование ЭБ и связанной с ними онлайновой научной среды стало повсеместным. Все вместе это порождает ряд новых обстоятельств для разработчиков ЭБ. Если в будущем ЭБ всех исследовательских организаций будут объединены в единое информационное пространство, то как научное сообщество может получить максимальные выгоды от использования возникающего системного эффекта? Как должна быть построена система фиксации и распространения результатов исследований, содержащихся в локальных ЭБ, чтобы обеспечивать их максимально возможное использование? Как мы должны организовать процесс использования результатов исследований и какие средства требуется создать, чтобы появились условия для сбора максимально полной и точной статистики о характере использования и научном влиянии результатов исследований? Как должны ЭБ и виртуальная научная среда собирать и обрабатывать статистику, чтобы показатели результативности исследований обладали существенно более высоким качеством по сравнению с текущими. Анализ возможных направлений развития ЭБ в данном контексте показал, что формируется комплекс условий для появления новой более эффективной организационной формы научно-исследовательской деятельности, которую мы назвали Открытой Наукой.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

## 1 Введение

В данной работе рассмотрение электронных библиотек (ЭБ) ограничено случаями их применения для обеспечения процесса **научных исследований**.

Современные процессы развития ЭБ создают особую виртуальную среду для научных исследований. В этих процессах можно выделить три логически связанные компоненты:

1) развитие программно-технических средств поддержки исследований (новый инструментарий);

2) использование новых инструментальных средств для совершенствования методов работы исследователей и их профессиональных взаимодействий (новые практики);

3) повышение эффективности организационных форм научно-исследовательской деятельности, использующее новый инструментарий и новые методы работы (новая форма организации сообщества исследователей).

В настоящее время формирование особой виртуальной среды для исследований наиболее заметно по многочисленным проектам интеграции содержаний ЭБ отдельных исследовательских организаций в профессиональные информационные пространства.

Наличие такого информационного пространства (как правило, на принципах федеративного объединения метаданных множества локальных ЭБ) и создаваемый им системный эффект являются отправными точками для нашего анализа.

Под профессиональным информационным пространством мы понимаем результат интеграции (на уровне метаданных) разнородных, организационно и географически распределенных источников научных данных. В минимальной конфигурации информационное пространство должно иметь единую систему навигации и поиск по накапливаемой структуре метаданных.

Кроме этого, содержание информационного пространства должно давать репрезентативное представление информационных ресурсов и деятельности соответствующего профессионального сообщества, а также иметь регулярную синхронизацию содержания

пространства с содержанием его локальных ЭБ как источников данных.

На уровне отдельных ЭБ подобные тенденции к интеграции проявляются в появлении в составе ЭБ так называемых "технических входов", предназначенных для доступа внешних программных роботов к переносу открытого контента ЭБ (как правило, только метаданные) в базы данных соответствующих информационных пространств.

В результате объединения в центральной базе метаданных достаточно большой доли производимых научным сообществом электронных материалов возникает своего рода "системный эффект", который должен быть использован в интересах данного сообщества.

Приложения, создаваемые для использования подобного системного эффекта, уже трудно классифицировать как ЭБ. Они создаются и развиваются как элементы онлайновой научной инфраструктуры во внешней по отношению к множеству ЭБ программно-технической среде.

Во втором разделе статьи рассматриваются 10 направлений наиболее важного, на наш взгляд, развития программно-технических средств поддержки исследований, которые создают и используют системный эффект профессиональных информационных пространств.

Существующее в международном научном сообществе мощное общественное движение за открытый доступ к результатам исследований популяризирует новые методы работы исследователей, прямо связанные с использованием современных ЭБ и профессиональных информационных пространств. Эти усилия постепенно формируют необходимые организационные предпосылки для того, чтобы использование ЭБ и связанной с ними онлайновой научной инфраструктуры стало повсеместным.

В третьем разделе статьи дается описание текущего состояния этих процессов.

В свою очередь, повсеместное использование учеными подобной виртуальной научной среды создает необходимые условия для реализации новой более эффективной организационной модели научной-исследовательской деятельности, которую мы назвали "Открытая Наука".

Открытая Наука возникает как результат реализации следующего комплекса программно-технических и организационных инноваций:

а) свободный доступ к результатам всех открытых научных исследований (кроме результатов, имеющих закрытый характер по коммерческим соображениям или связанных с безопасностью);

б) интеграция результатов исследований в онлайновую научную инфраструктуру, которая сконструирована для максимально широкого и полного использования этих результатов;

в) автоматический мониторинг онлайновой информационной активности ученых, формируемая на этой основе открытая онлайновая наукометрическая статистика и рассчитываемые на ее основе публичные показатели результативности ученых и исследовательских организаций;

г) использование онлайновых наукометрических показателей в процедурах принятия решений о финансировании научной деятельности, включая персональные надбавки ученых.

В четвертом разделе рассматриваются предпосылки формирования Открытой Науки, исходя из логики развития ЭБ и связанного с этим развития виртуальный среды для научной деятельности.

Обсуждаемые положения, где это возможно, иллюстрируются примерами из системы Соционет (http://socionet.ru/) и других российских и зарубежных проектов.

## 2 Современный контекст применения и развития ЭБ

Современные процессы развития ЭБ вносят существенный вклад (как прямой, так и косвенный) в формировании новой виртуальной среды для научных исследований. Общий контекст и основные направления развития этой среды можно, на наш взгляд, охарактеризовать следующими девятью факторами.

### 2.1 Трансформация ЭБ в CRIS

Наблюдается постепенная трансформация ЭБ в CRIS (CRIS – Current Research Information System, см. подробнее в [1-2]), которая часто проявляется в форме "глубокой" интеграции ЭБ с корпоративными научными информационными системами. Главной особенностью этого направления развития является лучшее приспособление ЭБ к комплексным потребностям организации, в которых они созданы. С одной стороны, это проявляется как рост ЭБ в "вширь" за счет обработки сервисами ЭБ новых типов информационных объектов, таких как информационные профили авторов, исследовательских организаций – мест работы авторов, описания научных проектов и другие типы, имеющие те или иные содержательные связи с традиционными для ЭБ научными статьями и материалами. С другой стороны, ЭБ обмениваются данными, а иногда и становятся органичной частью комплексных информационных систем управления научно-исследовательской деятельностью и автоматизации бизнес-процессов организации.

В контексте данной статьи главным положительным следствием этой тенденции является формирование в ЭБ более полного и точного информационного представления существующей в научном сообществе структуры деятельностей и действующих лиц. Например, отдельных исследователей – авторов статей, групп ученых – участников проектов, исследовательских организаций – направлений исследований и т.п.

Благодаря этому ЭБ становятся носителями информационных объектов, дающих комплексную и репрезентативную картину результатов деятельности и структуры соответствующих исследовательских организаций.

### 2.2 Интеграция локальных ЭБ

Наблюдается постоянный рост интереса к интеграции локальных ЭБ (т.е. принадлежащим отдельным организациям), друг с другом. Это относится, в первую очередь, к объединению контента ЭБ, а также к возможностям интеграции отдельных сервисов ЭБ. Интеграция, как правило, имеет форму конфедерации, т.е. предполагает сохранение полного контроля владельцев ЭБ над своими данными в объединенном информационном пространстве. Стимулом к объединению локальных научных информационных ресурсов является возникновение "системного эффекта", когда результирующее информационное пространство позволяет получить научному сообществу выгоды, недоступные от использования локальных ЭБ порознь. Данное направление развития проявляется в том, что в интерфейсах локальных ЭБ начинают создаваться "технические входы", предназначенные для программных роботов, которые в автоматическом режиме собирают через них метаданные об открытых для интеграции разделах ЭБ. Среди способов оформления "технических входов" в научной среде в настоящее время наиболее популярным является протокол OAI-PMH, разработанный международной инициативой Открытые Архивы [3]. Все большее количество ЭБ получают OAI-PMH интерфейс, что привело к появлению специальных названий "открытый архив" или "открытый репозиторий" (на Западе наиболее распространено название Institutional Repository или IR), обозначающих информационные ресурсы доступные в ЭБ через протокол OAI-PMH. В некоторых случаях OAI-PMH дополняется и другими подобными средствами для экспорта данных, например, RSS 2.0.

Примеры – Соционет (http://socionet.ru/), ЕНИП РАН (http://enip.ras.ru/), OAIster (http://oaister.org/), DRIVER (http://www.driver-repository.eu/), OpenDOAR (http://www.opendoar.org/) и др.

Как следствие данных процессов, растет общая открытость существующих научных материалов, и создаются благоприятные условия для их сбора и обработки в автоматическом режиме.

### 2.3 Появление информационных хабов

Среди ЭБ выделился особый вид, который получил название "информационный хаб" [4]. Данная разновидность ЭБ возникла в ответ на потребность обеспечить автоматический сбор и актуализацию разнородных метаданные из распределенных локальных ЭБ в единую стандартизованную базу данных, подготовив их для разнообразного использования всеми заинтересованными лицами. Для этого от информационного хаба требуется:

- развитая модель структуры данных, чтобы принять без потерь в центральную базу метаданных содержание большого количества разнородных ЭБ (в настоящее время, на наш взгляд, CERIF является наиболее подходящей моделью, см. http://www.eurocris.org/cerif/cerif-releases/cerif-2008/);
- возможность гибкой настройки на формат данных локальных ЭБ для переноса их открытого контента в центральную базу метаданных (решается созданием программных конвертеров под каждый конкретный случай);
- достаточно простой и гибкий способ экспорта (выгрузки) необходимых тематических выборок и коллекций из интегрированных стандартизованных метаданных для их последующей обработки (дать возможность выборок на основе популярных протоколов, например, RSS и OAI-PMH).

В последние два десятилетия происходила конкурентная борьба за модель информационного хаба между двумя основными подходами, один известен, как протокол Z39.50, второй – многим известен как подход RePEc (repec.org). К настоящему времени, на наш взгляд, подход физического копирования и синхронизации метаданных с сайтов-источников в общую централизованную базу данных, реализованный в проекте RePEc, выглядит предпочтительным и более востребованным.

Пример: система Соционет (http://socionet.ru/) является полноценным информационным хабом, поскольку предоставляет разработчикам средства не только для загрузки данных в базу (это делают все системы, перечисленные в примерах предыдущего подраздела), но и для выгрузки [4].

### 2.4 Создание информационных пространств

Профессиональные информационные пространства (см. определение в начале статьи) создаются на основе метаданных, собираемых информационным хабом через технические входы заданного множества ЭБ, и предназначены для визуализации текущей структуры и содержания объединенных информационных ресурсов. Они дают пользователям различные средства навигации и поиска по собранному контенту, а также предлагают средства для разнообразного использования собранных метаданных, включая создание новых информационных продуктов и услуг. Один из основных принципов подобных информационных пространств – создатель метаданных полностью контролирует их содержание. Посторонние лица не могут вносить изменения в структуру и содержание чужих метаданных, но могут создавать вторичные (обогащенные) метаданные на основе исходных и могут устанавливать связи от своих метаданных с любыми другими (см. ниже подраздел 2.6).

Примеры: все системы, перечисленные в примерах подраздела "Интеграция локальных ЭБ" создают профессиональные информационные пространства, но с разным набором функциональности для навигации и поиска.

### 2.5 Формирование онлайновой научной инфраструктуры

Наличие информационного хаба и визуализация его содержимого через информационное пространство создают для разработчиков ЭБ качественно новые условия. Все большая часть существующих разнородных и исходно распределенных метаданных, представляющих содержание научных информационных ресурсов, собирается в одном месте в стандартизованном виде. Собранные метаданные открыты для использования и обработки внешними программными средствами и сервисами. Результаты обработки исходных метаданных внешними программами и сервисами, если они являются научными информационными ресурсами, также могут передаваться (только метаданные) в информационный хаб и становятся доступны конечным пользователям в составе единого информационного пространства. Поскольку как первичные метаданные, так и вторичные, полученные в результате обработки первых, аккумулируются в одном месте, возникает эффект постепенного обогащения и развития научной информационной среды. Имеет место своего рода кругооборот создания "добавочной стоимости": разработчики взяли необходимые метаданные (М), обработали их сервисами (С), создали новые метаданные (М'), которые вернулись в информационный хаб, отобразились в информационном пространстве и стали "сырьем" для следующего цикла М-С-М'.

Здесь есть два важных новых обстоятельства:

а) Упомянутые в этом пункте "внешние программные средства и сервисы" уже не имеют явной привязки к определенным ЭБ, т.к. они работают с результатом виртуальной интеграции разнородных информационных ресурсов множества локальных ЭБ в массив стандартизованных метаданных. Поэтому, на наш взгляд, такие сервисы целесообразно классифицировать как часть онлайновой научной инфраструктуры. Информационные хабы и средства создания информационных пространств также скорее являются элементами онлайновой научной инфраструктуры, чем ЭБ.

б) Поскольку в схеме обработки М-С-М' циркулируют общедоступные метаданные, то при наличии необходимых программных средств их обогащение и развитие может быть выполнено уже не только создателями этих метаданных, но любым пользователем. Средства для развития/обогащения метаданных могут быть как частью некоторой ЭБ, так и входить в состав онлайновой инфраструктуры. Процесс развития метаданных, если он выполняется не их создателем, может быть организован как создание и включение в информационный хаб "улучшенных копий" (при этом оригинал улучшенных метаданных всегда сохраняется, т.к. он находится под полным управлением только своих создателей) или как создание связей между дополнительными информационными объектами и улучшаемыми метаданными.

Полноценный анализ направлений извлечения потенциального "системного эффекта" от интеграции научных метаданных еще впереди, но некоторые выгоды уже очевидны. Далее рассматриваются некоторые уже практически разрабатываемые примеры использования системного эффекта, создаваемого научным информационным пространством.

### 2.6 Формирование сетей связей между объектами информационного пространства

Одна из возможностей развития и "обогащения" метаданных, входящих в информационное пространство, сформированное на принципах федерации – создание инфраструктурных сервисов, позволяющих пользователям устанавливать связи между информационными объектами, собранными в центральную базу метаданных из множества различных ЭБ. Например, исследователь может создавать связи между своим профессиональным профилем, находящимся в ЭБ организации – места его работы, со своими статьями, находящимися в ЭБ других организаций (издательств журналов, конференций, ВУЗов и т.п.).

В самом общем виде это может означать создание сервисов для развития учеными своих профессиональных социальных сетей. Например, установление разнотипных сетей связей с метаданными родственных информационных объектов, развитие связей электронного цитирования, создание нового типа связей для визуализации профессионального влияния, и т.п.

Средства для поддержки целенаправленного формирования учеными профессиональных социальных сетей не являются принципиально новыми, т.к. в широком смысле для ученого это – постоянная деятельность по его профессиональной самоидентификации.

Центрами формирования такого рода связей являются следующие основные для научной деятельности виды информационных объектов: 1) карточка статьи/материала/проекта, состоящая из описательных данных (метаданные), 2) персональный профиль ученого, 3) профиль исследовательской организации или коллектива проекта.

В процедурах формирования профессиональных социальных сетей в системе Соционет между данными тремя видами объектов, а также и внутри отдельных видов создаются следующие наиболее важные конфигурации связей:

- между ученым (его персональным профилем) и организацией (профилем) – местом его

работы, это позволяет учитывать при расчете наукометрических показателей для организации соответствующие индикаторы ее сотрудников и наоборот;

- между ученым и его авторскими электронными материалами (в институтских ЭБ, научных журналах и т.п.), что позволяет при расчете наукометрических показателей ученого учитывать характеристики его материалов;
- между авторскими электронными материалами ученого и другими материалами, которые были им использованы (процитированы) при подготовке своего, что позволяет устанавливать содержательные связи между научными материалами и подсчитывать уровень их использования в научном сообществе;
- между персональными профилями ученых и статьями других ученых, как связи, отражающие индивидуальное профессиональное признание данными учеными выделенных ими статей своих коллег, что позволяет оценивать влияние (слабая форма использования) результатов исследований на научное сообщество.

Механизмы электронного депонирования материалов в системе Соционет позволяют автоматически включать в состав карточек статей/материалов (в метаданные) связи с профилем автора и его места работы. Таким образом, в процессе оформления и передачи материалов в институтскую ЭБ автоматически формируется определенная часть профессиональной социальной сети.

### 2.7 Развитие связей цитирования

Важной является задача развития и улучшения качественных характеристик связей между материалами, отражающими использование некоторым ученым результатов исследований, полученных другими учеными, в процессе создания им нового научного знания. В текущей научной практике эти связи устанавливаются с помощью общепринятых схем цитирования научных материалов или их фрагментов.

Одним из следствий развития в рамках ЭБ средств электронного депонирования (размещения статей и материалов в открытом доступе) является то, что традиционное научное цитирование постепенно обретает форму «электронного цитирования» с рядом новых возможностей. Схема электронного цитирования, если она реализована в ЭБ, может позволять автору электронной статьи уточнить в каком именно качестве он использовал чужие результаты исследований. Это может быть реализовано включением в связи цитирования определенных качественных атрибутов, которые в свою очередь должны быть сконструированы так, чтобы позволять их автоматическое распознавание и обработку для выявления различных характеристик использования научных знаний.

Если исходить из того, что такие характеристики должны давать данные для получения более точных индикаторов использования результатов исследований, то возможен следующий список характеристик электронного цитирования:

1. «основание для получения моих результатов», что означает прямое научное использование автором цитируемого результата;
2. «подтверждение цитируемого результата», т.е. результат автора подтверждает цитируемый;
3. «цитируемый результат подтверждает результат, ранее полученный автором», т.е. автор утверждает о своем приоритете по отношению к независимо полученному цитируемому результату;
4. «близкий или связанный результат», означающее, что автор в определенном смысле повторил цитируемый результат;
5. «иллюстрация моих выводов», означающее определенную логическую связь между результатом автора и цитируемым;
6. «объект для критики», означает, что автор подвергает сомнению цитируемый результат.

Необходимо уточнить, что возможны случаи, когда автор при оформлении связи цитирования может одновременно отметить несколько характеристик. Для разъяснения подобных случаев автор может использовать поле «комментарии» для каждой связи цитирования.

При наличии подобной статистики о характеристиках цитирования возможно создание (в рамках онлайновой научной инфраструктуры) процедур для автоматического построения индикаторов, которые позволят более точно, чем известные традиционные индексы цитирования, выявлять как для отдельной статьи, так и всего корпуса научных результатов:

1. какие результаты кем используются как основа научного вывода;
2. какие результаты подтверждают или подтверждаются другими результатами;
3. какие результаты повторяют уже известные;
4. упоминание результатов в качестве общих иллюстраций;
5. какие результаты кем критикуются и имела ли данная критика позитивные для науки последствия.

Подобная модель электронного цитирования превращается практику научного цитирования в механизм формирования между учеными профессиональной социальной сети. Одна из дополнительных неожиданных причин для этого - превращение электронных научных статей и материалов в «живые документы». См. ниже подраздел 2.9 о появлении у научных статей статуса «живых документов», что в сочетании с расширенными возможностями электронного цитирования имеет для научного сообщества ряд новых и важных последствий.

### 2.8 Необходимость нового типа связей "оценки профессионального влияния"

При размышлении, нет ли "подводных камней" в схеме цитирования, описанной в подразделе 2.7, возникает ощущение, что предложенная модель научного электронного цитирования может нарушить в "пользу" автора баланс сил в потенциальном конфликте интересов между учеными.

При наличии в научной среде потенциального конфликта интересов, возможности публичной качественной оценки чужих результатов, применяемые отдельным ученым по своему субъективному разумению, в обязательном порядке должны быть сбалансированы возможностями научного сообщества дать встречную публичную профессиональную оценку субъективных мнений отдельного ученого. Если научная истина и новое научное знание «рождаются в процессе борьбы мнений», то средства ЭБ по обеспечению научной деятельности должны давать равные возможности для всех участников этой «борьбы».

В статье [5] предлагается следующий набор качественных характеристик для субъективного оценивания учеными степени влияния отдельного научного результата/статьи на развитие соответствующих областей науки:

- очень интересный результат (particularly interesting);
- поворотный пункт для развития науки (landmarks);
- новаторский/революционный результат (groundbreaking).

На наш взгляд, этот список должен быть сбалансирован еще, как минимум, следующими характеристиками:

- результат, основанный на заблуждении;
- ненаучный подход к получению результата;
- результат с возможным опасным влиянием.

Процедуру формирования подобных оценок пользователями информационного пространства для выбранных ими статей и материалов предлагается организовать в виде создания нового типа связей. При создании таких связей ученый не только выбирает качественные характеристики (для одной статьи/результата можно одновременно указать несколько характеристик), отражающие его понимание степени влияния исходной статьи/результата на развитие науки, но и должен иметь возможность прокомментировать свое решение. Созданные таким образом профессиональные оценки, с одной стороны, связаны с профилем ученого, который их сделал (можно будет увидеть сводный список: какие статьи и как оценил соответствующий ученый), а с другой – с научной статьей/материалом, для которой сделана оценка (для каждой статьи можно будет увидеть какие профессиональные оценки она имеет и кому они принадлежат).

### 2.9 Мониторинг и поддержка сетей связей

Если ученый (или его представитель) депонирует в ЭБ свою статью, то у него остается возможность постоянно (т.е. на протяжении всей своей профессиональной жизни) редактировать и изменять текст этой статьи. Научные электронные статьи по определению превращаются в "живые" документы (в одном европейском проекте их назвали "текучими публикациями" - liquid publication [5]).

Выгоды и потенциальные проблемы от этого очевидны:

Выгода - научный результат, над которым долгое время работает ученый, локализован в одном и том же эволюционирующем информационном объекте. Его будет легче найти, а его цитирование гарантирует ссылку на текущую обновленную версию научного результата. Сохраняя старые версии статей, можно проследить историю развития научной идеи. И т.д. и т.п.

Потенциальные проблемы - очередное редактирование "живого" документа может нарушить цитаты из этой статьи, уже сделанные учеными. С учетом возможно развитой сети связей цитирования между "живыми" документами, нарушение связей цитирования с одним документом, может поставить под сомнение содержание и множества других документов в данной сети цитирования.

Решение - онлайновая научная инфраструктура должна иметь сервис мониторинга всех связей, существующих между объектами информационного пространства, включая связи цитирования. Подобный сервис, например, будет уведомлять:

- авторов исходной статьи - о том, что и кем процитировано из его статей, а также какие связи цитирования автор нарушает, когда вносит в свою статью изменения;
- авторов статей, цитирующих другие статьи - о том, что цитируемая статья была изменена и сделанные цитаты требуют проверки;
- читателей - о наличии или отсутствии обновления цитат в читаемой статье, если цитируемые статьи изменялись уже после "выхода" читаемой статьи.

### 2.10 Мониторинг развития информационного пространства и активности действующих лиц

Другое важное направление использования "системного эффекта" от интеграции метаданных локальных ЭБ в центральной базе связано с разработкой инфраструктурных сервисов для автоматического мониторинга изменения структуры и расширения информационного пространства, а также отслеживания параметров информационной активности ученых в рамках этого пространства. В первую очередь это – сбор качественной и количественной статистики о результативности работы ученых и исследовательских организаций (учет всех видов связей, отражающих

использование и профессиональное влияние их результатов исследований, подсчет количественных показателей и т.п.), а также сбор статистики о востребованности результатов исследований в открытом доступе.

Процедуры сбора статистики могут работать в непрерывном режиме, а формируемая ими статистическая база может наращиваться, например, ежедневными порция данных. Это означает формирование и обновление статистического портрета отслеживаемых процессов практически в реальном времени.

Подобный мониторинг открывает возможность автоматического формирования открытой онлайновой наукометрической статистики. На этой основе можно организовать автоматизированное построение публичных индикаторов результативности ученых и научных организаций. Сконструировать новую профессиональную сигнальную систему, улучшающую ученым ориентировку в текущих тенденциях и направлениях развития науки, а также упрощающую определения ими своего места в научном сообществе.

Если исследователи массовым образом используют описанные выше элементы виртуальной научной среды, то порождаемая в результате онлайновая статистика позволяет формировать статистический портрет ученого как набор, например, следующих данных:

- Персональные сведения об ученом, и история их изменений;
- Динамика роста количественных показателей активности ученого, в т.ч. числа статей, результатов исследований, материалов других типов и т.п.;
- Динамика количеств и структуры цитирования результатов данного ученого, а также цитирований чужих результатов, сделанные данным ученым;
- Параметры распределения качественных характеристик для цитирований, выполненных данным ученым, а также для результатов данного ученого, выполненных другими учеными;
- Динамика количеств, структуры и распределение качественных характеристик оценок профессионального влияния работ данного ученого, сделанных научным сообществом, а также чужих работ, сделанных данным ученым;
- Распределение цитирования и оценок влияния между различными результатами данного ученого.

Аналогичный статистический портрет может быть составлен и для исследовательских организаций на основе агрегирования статистических портретов ученых, которые работают в соответствующих организациях.

Онлайновая наукометрия безусловно является научным информационным ресурсом, и поэтому также должна быть в составе информационного пространства. Рассчитываемые на ее основе показатели использования, влияния и востребованности научных статей могут служить для обогащения/развития метаданных соответствующих информационных объектов. Технически это может решаться созданием дополнительного вида связей между метаданными определенного объекта и сформированным для него статистическим портретом. Примером подобного решения в Соционет являются ссылки "График суммарной статистики просмотров", которые автоматически включаются в веб-страницы, визуализирующие метаданные научных материалов. Подробнее эта тема обсуждается в публикациях [6, 7].

При определенных условиях собираемая таким образом онлайновая статистика будет содержать репрезентативные статистические портреты всех действующих лиц (ученых, проектов, организаций и т.п.) научного сообщества, а также фиксировать все значимые для научного сообщества процессы и их результаты. Фактически, это означает создание, и обновление практически в реальном времени своего рода информационной проекции или модели научного сообщества, объединенного соответствующим информационным пространством.

Вся собираемая онлайновая статистика является открытой и, следовательно, создаваемая на ее основе информационная модель сообщества также является открытой для всех членов научного сообщества.

С учетом того, что в реальной жизни научное сообщество представляет собой систему в существенной мере скрытую для прямого наблюдения и изучения, то появление информационной модели, которая при определенных условиях может быть достаточно точной, может оказаться ценными приобретением.

Представляется перспективным использование подобной информационной модели в сервисах, создаваемых разработчиками ЭБ для своих пользователей, т.к. это означало бы определенный учет "обратных связей". Такие адаптивные сервисы могли бы предоставлять пользователям услуги, которые учитывали бы текущее состояние научного сообщества, реакцию ученых на определенные события и другие параметры информационной модели сообщества.

Выводы к разделу: Современный контекст функционирования и развития ЭБ характеризуется стремлением к объединению содержания локальных ЭБ. Это проявляется в виде создания информационных хабов и профессиональных информационных пространств. Разработка "системного эффекта" от объединения метаданных в одной базе идет по двум направлениям: а) создание возможностей для улучшения содержательной связанности родственных информационных объектов (например, автора со всеми своими статьями, исходно размещенных в различных ЭБ, развитие связей цитирования, визуализация связей профессионального влияния, поддержка актуальности связей при превращении статей в "живые" документы и т.д.); б) комплексный

мониторинг информационной активности ученых, формирование на этой основе онлайновой наукометрии, создание индикаторов научной результативности ученых и исследовательских организаций, создание комплекса статистических индикаторов, выполняющих роль профессиональной сигнальной системы для научного сообщества.

## 3 Электронные библиотеки и открытый доступ к результатам исследований

Параллельно и отчасти независимо от потока программно-технических инноваций, являющихся развитием парадигмы ЭБ, в настоящее время в международном научном сообществе сформировалось мощное общественное движение за открытый доступ к результатам исследований [8].

Инициатива открытого доступа призывает все исследовательские организации создать открытые электронные репозитории, библиотеки или архивы для размещения в публичном доступе всех законченных результатов открытых исследований, проводимых в соответствующей организации. Именно откликом на эту инициативу объясняется рост количества ЭБ, имеющих "технический вход" для выгрузки метаданных внешними программными роботами (см. выше подраздел "Интеграция локальных ЭБ"). Масштабы и динамику этих процессов в реальном времени иллюстрирует "Реестр репозиториев открытого доступа" (ROAR, http://roar.eprints.org/index.php?action=browse).

Для создания мотивации у ученых к размещению своих результатов исследований в открытом доступе данная инициатива рекомендует организациям, финансирующим исследования, обязывать исследователей, которым они платят деньги, оперативно выкладывать в открытые институтские репозитории результаты соответствующих исследований. Отслеживание количества научных организаций, включая научные фонды, уже обязавших своих ученых депонировать все законченных результаты открытых исследований ведет еще один международный регистр Registry of Open Access Repository Material Archiving Policies (ROARMAP, http://www.eprints.org/openaccess/policysignup/).

Еще один важный аспект инициативы открытого доступа - увеличение количества научных коммерческих журналов, официально объявивших о согласии с размещением авторами в открытом доступе своих статей, которые были переданы на рассмотрение или уже опубликованы в таких журналах. Отслеживание ситуации в этой области ведется проектом ROMEO (http://romeo.eprints.org/publishers.html). Согласно этому источнику на апрель 2009 г. 97% зарегистрированных научных издательств объявили о согласии с этим положением (как правило, с определенными оговорками).

Идеи открытого доступа получили государственную поддержку в ряде стран в виде появления национальных программ по созданию электронных репозиториев открытого доступа, развитию открытых архивов, созданию на их основе научного информационного пространства, сбора онлайновой наукометрической статистики и ее использования при определении уровня финансирования науки.

Так, в Великобритании запущен специальный проект в поддержку репозиториев (Repositories Support Project, http://www.rsp.ac.uk/). Более важно, что в этой стране действует государственная программа Research Assessment Exercise (RAE, http://www.rae.ac.uk/), в которой в числе прочих показателей результативности ученых и исследовательских организаций используются элементы онлайновой наукометрии (индексы просмотров и скачиваний статей в Интернете и т.д.).

В Австралии действует государственная программа поддержки университетов в создании репозиториев открытого доступа Australian Scheme for Higher Education Repositories (ASHER), по которой выделяется 25.5 миллиона долларов на три года для поддержки создания и обновления цифровых репозиториев .

Выводы к разделу: Исходно идеи открытого доступа отталкивались от достижений в развитии ЭБ, поскольку ЭБ являются основным инструментом для обеспечения доступа к электронным публикациям. В настоящее время ситуация, на наш взгляд, несколько поменялась: инициативы открытого доступа к результатам исследований являются стимулом к дальнейшему развитию ЭБ и создают необходимые организационные предпосылки и мотивации, чтобы использование ЭБ и основанной на них новой виртуальной среды для исследований стало общедоступным и повсеместным.

## 4 Открытая Наука как следствие развития электронных библиотек

Повсеместное использование исследователями виртуальной научной среды, одним из основных элементов которой являются ЭБ, формирует необходимые условия для внедрения в практику научной деятельности новой модели ее организации, которую мы назвали "Открытая Наука". Для того чтобы оценить преимущества Открытой Науки перед существующим в настоящее время способом организации деятельности научного сообщества, рассмотрим ее основные положения.

Система организации научной деятельности может быть названа Открытой наукой, если она удовлетворяет следующим принципам:

1. Реализован повсеместный свободный доступ к результатам исследований через электронное депонирование учеными (или их представителями) всех результатов открытых исследований в ЭБ. Модель организации ЭБ

приближается к CRIS, что обеспечивает надлежащее информационное представление в ЭБ исследователей, организаций и других информационных объектов (см. подраздел 2.1).

2. Ученые используют электронную форму фиксации результатов своих исследований и научного приоритета, специально сконструированную в составе ЭБ для их максимально широкого и полного использования (см. подразделы 2.2 – 2.6).

3. Использование результатов исследований в виде их электронного цитирования или оценки их профессионального влияния выполняется по усовершенствованной схеме, включающей спецификацию качественных характеристик использования/влияния (см. подразделы 2.7 и 2.8).

4. Действует онлайновая научная инфраструктура, которая собирает содержание ЭБ, научных архивов и репозиториев отдельных организаций в единое научное информационное пространство (см. подразделы 2.2 – 2.5).

5. Действуют нормы по формированию в научном информационном пространстве репрезентативных информационных образов исследователей и научных организаций, а также правила по их поддержанию в актуальном виде и формированию на их основе профессиональных социальных сетей (подраздел 2.6). Действует система контроля правильности установленных связей с учетом превращения научных материалов в "живые" документы (подраздел 2.9).

6. В составе онлайновой научной инфраструктуры созданы средства электронного автоматизированного мониторинга за изменением качественных и количественных параметров научных архивов и информационных образов действующих лиц, обеспечено формирование на этой основе открытой наукометрической базы данных с ежедневным обновлением (подраздел 2.10).

7. Реализован автоматизированный расчет на основе наукометрической статистики индикаторов и показателей функционирования науки, включая комплекс показателей результативности работы ученых и научных организаций, размещение этих ежедневно обновляемых показателей в открытом доступе (подраздел 2.10).

8. Обеспечено использование этих показателей в принятии решений о финансировании научной деятельности, включая определение персональных надбавок ученых.

Из данного набора требований, если представить, что он полностью реализован, логически вытекают следующие преимущества Открытой Науки по сравнению с действующей системой научной деятельности:

- полный корпус современных результатов публичной науки в открытом онлайновом доступе, т.к. все результаты исследований, для которых нет ограничений доступа, своевременно помещаются в ЭБ, открытые архивы и репозитории исследовательских организаций;

- лучшие условия для доступа к научному знанию, и использования научных результатов, как следствие - ускорение научного кругооборота и повышение эффективности научной деятельности;

- лучшие условия для мониторинга результативности научной деятельности для отдельных ученых и исследовательских организаций;

- возможность создания прозрачной и проверяемой системы публичных индикаторов результативности ученых и исследовательских организаций, а также выявление и визуализация тенденций развития науки;

- работающие мотивации для ученых выкладывать все свои результаты в ЭБ, т.к. их персональные надбавки за научную деятельность зависят от соответствующих показателей.

Выводы к разделу: Различные направления развития, существующие в настоящее время в научном сообществе вокруг ЭБ, имеют заманчивую перспективу сформировать новую высокоэффективную модель научной деятельности под названием Открытая Наука.

## 5 Заключение

Описанные выше направления развития ЭБ способствуют, на наш взгляд, формированию некоторых важных характеристик научного сообщества:

а) открытость его информационного пространства для надлежащего представления как действующих лиц науки, так и результатов их исследований, или деятельности в общем случае;

б) открытость онлайновой научной инфраструктуры для создания разработчиками новых информационных продуктов и сервисов в интересах данного научного сообщества на основе содержания информационного пространства;

в) открытость информационной модели сообщества, которая позволяет всем желающим получить картину текущего состояния и происходящих в сообществе процессов, а также может быть использована для принятия более обоснованных решений всеми действующими лицами науки.

С точки зрения отдельного исследователя эти три перечисленные выше характеристики означают не только лучшие условия для его профессиональных взаимодействий с научным сообществом, но и новый уровень "прозрачности" или публичности результатов его профессиональной деятельности. С одной стороны, Открытая Наука дает ученому эффективную саморазвивающуюся среду для его профессиональной деятельности, ориентированную на максимальное использование полученных им результатов. С другой, деятельность ученого, полученные результаты, степень их использования

и влияния на науку, место ученого в науке и отношение к нему научного сообщества - все эти характеристики профессиональной деятельности ученого являются публичными в Открытой Науке.

Парадигма электронных научных библиотек переживает в настоящее время важное обновление. Развитие как самих ЭБ, так и связанных с ними программно-технических и организационных систем, составляющих комплексную среду для научной деятельности, проходит этап качественных изменений. На горизонте – Открытая Наука, возможность которой прямо вытекает из тенденций развития ЭБ и которая обещает ученым большое количество изменений в привычном научном мире. Надеемся, что данная статья будет способствовать формированию в научном сообществе убеждения в том, что отмеченные изменения олицетворяют собой прогресс.

## Литература


[1] Current Research Information System (CRIS) – http://www.eurocris.org/

[2] Kulagin M. V., Lopatenko A. S.. Current Research Information System and Digital Libraries. Needs for Integration. Сборник трудов конференции RCDL'2001, электронная версия сборника - http://rcdl2001.krc.karelia.ru:8002/papers/contents.ru.shtml, электронная версия статьи - http://rcdl2001.krc.karelia.ru/papers/papers/kulagin_lopatenko/kulagin_lopatenko_paper.rtf

[3] The Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

[4] Паринов С.И.. Информационные хабы. Соционет: электронный депонент, 2006, http://socionet.ru/publication.xml?h=repec:rus:mqijxk:9

[5] Fabio Casati, Fausto Giunchiglia, Maurizio Marchese. Publish and perish: why the current publication and review model is killing research and wasting your money, ACM Ubiquity 8 (3), Feb 2007. http://www.acm.org/ubiquity/views/v8i03_fabio.html (в тексте цитируется более новая версия этой статьи: Liquid Publications: Scientific Publications meet the Web, Version 2.3, October 1, 2007, http://liquidpub.org/attachment/wiki/WikiStart/LiquidPub%20paper-latest.pdf)

[6] Когаловский М.Р., Паринов С.И. Метрики онлайновых информационных пространств. // Экономика и математические методы, 2008, т. 44, №2, с. 108-120

[7] Паринов С.И. e-Science - онлайновое будущее науки. // Информационные технологии, №9, 2007, приложение.

[8] Stevan Harnad. The Implementation of the Berlin Declaration on Open Access. D-Lib Magazine, March 2005, V. 11 N. 3. Текущее состояние см. на http://www.eprints.org/openaccess/ и http://www.sherpa.ac.uk/


## Digital libraries development is a way to Open Science


Sergey Parinov



Step by step permanent development of digital libraries (DL) concept is forming a particular virtual research environment. Successfulness of international initiative for open access to research is creating conditions to use DL and related research environment as common research practice. It creates some important challenges for DL developers. If, in the future, DL of all research organizations are federated as universal research data and information space (DIS), how can research community exploit the emerged systemic effect to produce maximal benefits? How should we construct, in form and function, a system for shaping and sharing research results from local DL so as to provide maximal usage? How should we organize a process of research results usage, and design necessary tools to provide maximally comprehensive and accurate statistics on the uptake, usage and impact of research results? How should the DL and the research environment accumulate and process statistics to generate new online metrics sufficient for research assessment of higher quality, sensitivity, breadth, accuracy, reliability, and validity than current metrics? Analysis of possible in such context DL development trends shows an ability to appear a new efficient organization form for research activities. We called it as Open Science.



♣ Автор благодарен руководству ЦЭМИ РАН за поддержку данных исследований, а также М.Р. Когаловскому за многочисленные комментарии и пожелания, которые позволили улучшить текст этой статьи. Частично данные исследования поддержаны грантом РФФИ № 09-07-00378.

# ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ ТЕКСТОВ

# EXTRACTION OF INFORMATION FROM TEXTS

# Методы машинного обучения в задачах извлечения информации из текстов по эталону

© Алексеев С.С.
ООО «REHAU»
sergej.alexejew@rehau.com

© Морозов В.В.
ИАЦ Кортес
morozov@kortes.com

© Симаков К.В.
МГТУ им. Н.Э. Баумана
skv@ixlab.ru

**Аннотация**

Работа посвящена решению частного случая задачи извлечения информации из текстов – извлечению по эталону, при котором заранее известны эталонные (канонические) формы всех структур, подлежащих распознаванию в тексте. Основной акцент сделан на методах обучения, позволяющих снимать неоднозначности распознавания.

## 1 Введение

Задача извлечения информации из естественно-языковых текстов относится к классу задач распознавания. Извлечение заключается в выделении (распознавании) целевого текстового фрагмента, отвечающего определенным критериям, в сплошном тексте [7,24,27]. Задача извлечения информации возникает во многих областях, связанных с обработкой естественно-языковых текстов. Наиболее распространенным примером использования методов извлечения, является выделение в текстах участников событий заданного типа (например, событий, освещающих финансовые сделки между компаниями [10]).

Распознаваемые текстовые фрагменты являются результатом извлечения, а критерии, по которым они обнаруживаются – правилами извлечения. Составление полного перечня правил извлечения является весьма трудоемкой задачей независимо от формы их представления, поэтому параллельно с задачей извлечения решается задача автоматизированного составления правил распознавания [28]. Для этих целей обычно используют методы машинного обучения, позволяющие по набору обучающих позитивных и негативных примеров построить систему обобщенных правил.

## 2 Постановка задачи

### 2.1 Задача извлечения информации по эталону

Извлечение с использованием эталонной базы данных является частным случаем общей задачи извлечения информации из текстов. Несмотря на существование подходов к решению задачи в общем виде, извлечение информации по эталонной базе имеет свою специфику.

Положим, что в рамках некоторой предметной области существует эталонная база в виде реляционной таблицы $C$, содержащая N строк и M столбцов. В каждой строке $c_i$ хранится некоторая структура данных, полям которой соответствуют столбцы таблицы. Если положить, что каждая строка $c_i$ описывает некоторый объект, то таблица $C$ в этом случае описывает класс однотипных сущностей предметной области.

В общем случае все сущности предметной области могут быть описаны в форме онтологии, где кроме классов будут установлены и отношения между ними, однако в данной работе интерес представляет только объекты одного класса, заключенные в таблице $C$.

Таким образом, каждая ячейка $c_{ij}$ содержит значение j-ого свойства i-ого объекта. Положим, что каждое значение $c_{ij}$ может быть приведено к текстовому представлению, содержащему одно и более слов естественного языка. Это представление является эталонным, поскольку отражает каноническую форму записи значений, принятую в данной предметной области. Например, канонической формой словосочетаний, образующих именную группу из существительного и прилагательного, может являться форма именительного падежа обоих слов данной группы. В более сложных многословных значениях $c_{ij}$ каждое слово может употребляться в характерном для канонической формы падеже, роде, числе и т.д. Каноническая форма $c_{ij}$ также определяет порядок следования слов в каждой записи.

Положим, что в рамках предметной области существуют тексты $T = \{t_k\}$, в которых в тех или иных контекстах могут принимать участие объекты таблицы $C$ посредством употребления неканонических форм записи их свойств $c_{ij}$. Задача извлечения информации в данном случае заключается в установлении факта употребления конкретного объекта $c_i$ в заданном тексте, сопровождающегося выделением

**Труды 11ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

текстовых фрагментов, содержащих неканонические формы записи его свойств $c_{ij}$.

Тот факт, что в текстах употребляются неканонические формы записи целевой информации, приводит к необходимости решать проблему неоднозначности. Задача снятия неоднозначности заключается в следующем. Из множества $\{c_i\}$, являющегося результатом распознавания для заданного текста $t_k$, необходимо выбрать единственный верный элемент $c_r$.

### 2.2 Примеры задачи извлечения по эталону

Наиболее характерной задачей, решение которой возможно с использованием методов извлечения по эталону, является построение графа цитирования. Суть задачи заключается в следующем. Имеется коллекция из N статей, каждая из которых характеризуется следующими свойствами: название, Ф.И.О. авторов, год публикации и наименование сборника трудов, в рамках которого статья опубликована.

Требуется проанализировать текст каждой статьи и выявить в каждом из них ссылки на другие статьи данной коллекции. Несмотря на то, что обычно ссылки устанавливаются в конце каждой статьи в разделе «Литература», а также на то, что существуют соответствующие правила их оформление, распознавание ссылок на практике является нетривиальной задачей. Проблемы возникают из-за орфографических ошибок, например, в написании Ф.И.О. или в названии статьи, из-за неполноты приводимой информации, а также из-за нарушения порядка слов в некоторых полях библиографической ссылки.

Другим примером является задача распознавания в текстах почтовых адресов. Предположим, имеется N почтовых адресов, каждый из которых представлен набором полей (регион, район, город, улица и т.д.). Также предположим, что имеется поток документов, которые необходимо раскладывать по группам в зависимости от адреса, указанного в тексте документа. Форма употребления адресов в текстах может иметь произвольный характер, существенно отличающийся от канонической записи. Например, в адресе не всегда указывается регион и район, порядок слов в многословных наименованиях может отличаться от порядка их следования в канонической форме, некоторые слова в названиях могут записываться в тексте в сокращенной форме, либо с орфографическими ошибками (что вообще характерно для любых имен собственных, не подчиняющихся общей грамматике языка). В такой постановке, задача выявления адреса в тексте также становится нетривиальной, несмотря на то, что в наличии имеется эталонная база адресов.

## 3 Обзор методов извлечения

В настоящий момент в литературе не существует отдельных упоминаний о рассматриваемом подклассе задач, поскольку многие исследователи решают задачу извлечения информации из текстов в общем виде. В качестве основных направлений в данной области можно выделить символьный и численный подходы.

### 3.1 Символьный подход

В рамках данного подхода правила извлечения записываются на формальном языке, напоминающем язык регулярных выражений. Такие языки позволяют описывать свойства контекстов употребления целевой информации в виде морфологических признаков слов, окружающих извлекаемые текстовые фрагменты, синтаксические роли слов и групп слов, а также конкретные ключевые слова-сателлиты, регулярно встречающиеся в контексте целевой информации. Данные методы разделяются на два класса: пропозиционные [4,11,15,17] и реляционные [3,5,6,16] - в зависимости от выразительных возможностей языка и возможностей метода обучения.

В пропозиционных методах язык правил извлечения эквивалентен логике нулевого порядка. Полагается, что шаблоны правил, определяющие ограничения на связи между словами, задаются экспертом, а метод обучения автоматически подбирает ограничения, накладываемые на значения свойств слов-заполнителей этих шаблонов. В реляционных методах язык правил извлечения эквивалентен логике первого порядка [19], а методы обучения автоматически формируют как ограничения на связи между словами в тексте, так и ограничения на слова, принимающие участие в этих связях.

Обучение в символьных методах организуется по принципу дедуктивного вывода, либо по принципу индуктивного обобщения. В обоих случаях эксперт готовит обучающие примеры, представляющие собой тексты, в которых явным образом выделены целевые фрагменты, подлежащие извлечению. Далее к обучающей выборке применяется выбранный метод обучения, синтезирующий обобщенные правила распознавания целевой информации. Полученные таким образом правила, по сути, заключают в себе закономерности, характерные для употребления целевой информации в текстах обучающей выборки.

### 3.2 Численный подход

Методы данного класса полагают, что изначально имеется набор элементарных правил, определенный априори, а задача построения извлекателя сводится к подбору композиции этих правил, обеспечивающей заданную точность и полноту извлечения на исходной обучающей выборке. Композиция в разных методах может строиться по-разному, но общим у этих методов является то, что каждое элементарное правило включается в композицию с определенным численным весом.

Из наиболее распространенных представителей данного класса следует выделить Байесов классификатор [14], Скрытые Марковские Модели [2],

методы максимизирующие энтропию [1,9] и условные случайные поля [12]. Обучение во всех перечисленных методах сводится к подбору вероятностных коэффициентов, оценивающих вклад элементарного правила распознавания в рамках их полной композиции.

### 3.3 Применимость к поставленной задаче

Символьные методы естественным образом допускают использование ключевых слов в правилах извлечения, так что в общем случае правило извлечения может содержать прямое перечисление всех слов эталонной базы, однако при достаточно больших N (например, для базы почтовых адресов $N\sim10^6$) такое использование правил будет неэффективным. Вместе с тем, такое перечисление не дает ответ на вопрос о том, как распознавать в текстах неканонические формы перечисленных слов. Проблема неоднозначного распознавания в методах данного класса также не решается.

Численные методы позволяют решить задачу снятия неоднозначности распознавания, поскольку каждому из вариантов извлечения может быть поставлено в соответствие число в виде вероятности распознавания, так что выбор единственного варианта становится очевидным.

Общим свойством указанных подходов является то, что они запоминают структуру целевой информации и стремятся не запоминать конкретные ключевые слова контекста извлекаемых данных, чтобы добиться должного уровня общности синтезируемых правил. Однако в нашей задаче ситуация полностью противоположная – за счет того, что имеется полная эталонная база, появляется возможность не запоминать структуру извлекаемой информации. Поэтому порядок следования слов и распознаваемых полей в тексте становится неважным. Возвращаясь к описанным выше примерам, автор статьи может идти как перед ее названием, так и после нее, аналогично в записи почтового адреса все его поля могут быть записаны в любом порядке.

## 4 Метод извлечения

Наличие эталонной базы позволяет реализовать достаточно простой способ выделения целевой информации на основе технологии полнотекстового поиска.

Упрощенно данный метод может быть описан следующим образом. Для таблицы $C$ строится полнотекстовый индекс, представляющий собой словарь, в котором объединены все слова, задействованные в ячейках $c_{\text{ij}}$. Для каждого слова формируется инвертированный список, содержащий номера строк $c_i$, в полях которых встречается данное слово.

При анализе по обрабатываемому тексту скользит окно, размер которого соответствует длине максимальной строки $c_{\max}$ исходной таблицы. На основе каждого слова окна формируется поисковый запрос, в котором логическим ИЛИ объединены все его возможные словоформы, которые кроме морфологического словоизменения могут учитывать вероятные опечатки и варианты сокращения. Данный запрос выполняется на имеющемся словаре. Результатом такого поиска является список строк исходной таблицы, в каждой из которых встречается один или несколько вариантов словоизменения исходного слова. Дополнительному учету могут подлежать слова, по которым ничего не удалось найти. После превышения заданного порога по числу таких слов появляется возможность сделать вывод о том, что текущий текстовый сегмент не содержит целевой информации, после чего сдвинуть окно дальше, пропустив этот сегмент.

Найденные по всем словам окна списки пересекаются, в результате чего получается итоговый список $C_t = \{c_i\}$, содержащий номера записей исходной таблицы, в которых встречается большинство слов текущего окна. Таким образом, встает вторая задача – выбор одного единственно верного варианта распознавания $c_r$ из полученного множества $C_t$.

Для выполнения такого выбора необходима некоторая функция $\rho : T \times C \to [0..1]$, позволяющая для каждого варианта распознавания $(t, c_i)$ (где $t$ – распознанный фрагмент текста, $c_i$ - распознанный объект) поставить в соответствие число, оценивающее качество распознавания данного варианта. Имея такую функцию, выбор единственно наилучшего варианта распознавания выполняется согласно критерию максимизации значения $\rho$, т.е. $(t, c_r) = \arg\max_{c \in C_t} \rho(t, c)$.

## 5 Методы обучения извлекателя

Построение функции $\rho : T \times C \to [0..1]$ посвящена вторая часть работы. Чтобы получить функцию данного вида, необходимо располагать числовыми свойствами варианта распознавания. Пусть введено семейство признаков $\{f_j\}_{j=1}^{F}$, каждый признак представляет собой функцию вида $f_j : T \times C \to R$, т.е. позволяет дать количественную оценку варианту распознавания $(t, c) \in T \times C$. Тогда искомую функцию $\rho$ можно искать в виде композиции $\rho(t, c) = (f_1 \circ f_2 \circ \ldots \circ f_F)(t, c)$, где символ $\circ$ обозначает некоторую композицию функций.

Признаки $\{f_j\}_{j=1}^{F}$ могут описывать различные свойства варианта распознавания $c_{\text{i}}$ такие, как позиционную близость распознанных полей $c_{\text{ij}}$, количество распознанных полей $|c_{\text{ij}}|$, количество слов с опечатками, общее количество слов и др.

Композицию $(f_1 \circ f_2 \circ \ldots \circ f_F)(t, c)$ можно сконструировать, обладая набором обучающих примеров $T_{teach} = \{t_i\}$. Где каждый позитивный пример пред-

ставляет собой пару $(t_i, c_p^i)$ так, что для каждого текстового фрагмента $t_i \in T_{teach}$ явно выделен правильный вариант распознавания $c_p^i$. Негативными обучающими примерами для каждого $t_i$ объявляется множество $\{c_n^i\} = C_i \setminus c_p^i$, где $C_i$ - все варианты распознавания в рамках фрагмента $t_i$.

Положим, что $\rho(t_i, c_p^i) = 1$ и $\rho(t_i, c_n^j) = 0 : c_n^j \in C_i \setminus c_r^i$. Тогда обучающую выборку можно представить в виде таблицы следующего вида.

Табл. 1. Обучающая выборка для синтеза $\rho(t,c)$

| $\rho(t,c)$ | $f_1(t,c)$ | $f_2(t,c)$ | … | $f_F(t,c)$ |
|---|---|---|---|---|
| … | … | … | … | … |
| $\rho(t_i, c_p^i) = 1$ | $f_1(t_i, c_p^i)$ | $f_2(t_i, c_p^i)$ | … | $f_F(t_i, c_p^i)$ |
| $\rho(t_i, c_n^1) = 0$ | $f_1(t_i, c_n^1)$ | $f_2(t_i, c_n^1)$ | … | $f_F(t_i, c_n^1)$ |
| … | … | … | … | … |
| $\rho(t_i, c_n^j) = 0$ | $f_1(t_i, c_n^j)$ | $f_2(t_i, c_n^j)$ | … | $f_F(t_i, c_n^j)$ |
| … | … | … | … | … |

Эту таблицу далее будем назвать интерполяционной, поскольку она содержит значения целевой функции $\rho(t,c)$ в интерполяционных точках F-мерного пространства, где каждой j-ой координате соответствует признак $f_j$ вариантов распознавания. Далее обозначим число интерполяционных точек как $N_t$.

Каждый вариант распознавания характеризуется F-мерным вектором значений числовых признаков $\{f_j\}_{j=1}^F$, обозначив его как $x$, можем перейти от представления функции $\rho(t,c)$ к представлению $\rho(x)$. Через $x_j$ будем обозначать j-ую координату вектора, т.е. $x_j \equiv f_j(c,t)$.

Таким образом, задача обучения извлекателя сводится к задаче аппроксимации функции $\rho(x)$ по заданной интерполяционной таблице, где в качестве аргумента выступают F-мерные точки (векторы).

### 5.1 Наивный Байесов классификатор

Данный способ аппроксимации $\rho(x)$ был выбран в качестве отправной точки, относительно которой выполнялось сравнение остальных методов обучения в контексте поставленной задачи. Поскольку целевая функция в идеале может принимать только два значения 1 и 0, с точки зрения классификатора, они расцениваются как два класса.

Классификатор работает на основе формулы $p(\rho \mid x_1 \cdot \ldots \cdot x_F) = \dfrac{p(\rho) \cdot \prod_{j=1..F} p(x_j \mid \rho)}{\prod_{j=1..F} p(x_j)}$, где $p(\rho)$, $p(x_i \mid \rho)$ и $p(x_i)$ - распределения вероятностей, формируемые в результате обучения. Формула справедлива, при условии, что признаки $\{x_j\}_{j=1}^F$ независимы.

Основная проблема такого подхода заключается в том, что признаки могут принимать бесконечно большое число значений, тогда как на обучающей выборке может быть получено распределение только для тех значений, которые попали в выборку. Для решения этой задачи значения всех признаков были приведены к диапазону [0…1]. Этот диапазон был разделен на 10 частей так, что разные значения любого j-ого признака, попадающие в один и тот же диапазон, рассматривались в качестве одного и того же значения случайной величины $x_j$. Таким образом, распределение $p(x_i)$ аппроксимируется ступенчатой функцией вида

$$p(x_i) = \begin{bmatrix} p(0 \le x_i < 0.1) \\ \ldots \\ p(0.9 \le x_i \le 1) \end{bmatrix}.$$

Вероятности в правой части выражения определяются на обучающей выборке путем подсчета соответствующих относительных частот. Аналогичным образом формируются распределения $p(x_i \mid \rho)$.

Из достоинств данного подхода можно выделить простоту реализации. К недостаткам отнесем необходимость грубой аппроксимации $p(x_i)$ и $p(x_i \mid \rho)$ в виде ступенчатых функций, а также наличие условие независимости признаков $\{x_j\}_{j=1}^F$. В нашем случае за каждым признаком $x_j$, по сути, стоит функция $f_j(c,t)$, реализуемая алгоритмически, что фактически не дает информации о наличии/отсутствии зависимостей между ними.

### 5.2 SVM классификатор

Идея использования данного метода возникла из предположения о том, что два множества интерполяционных точек со значениями $\rho(x) = 1$ и $\rho(x) = 0$ в идеале являются выпуклыми непересекающимися множествами [23]. В этом случае, можно попытаться найти разделяющую их гиперплоскость $\langle w, x \rangle = w_0$, равноудаленную от границ этих множеств, а затем использовать ее в работе распознавателя (здесь $w$ - нормаль к гиперплоскости, $w_0$ - число, задающее ее сдвиг). Целевая функция $\rho(x)$ примет в этом случае вид:

$$\rho(x) = \begin{bmatrix} 1, & \textit{если} \quad \langle w, x \rangle \ge w_0 \\ 0, & \textit{если} \quad \langle w, x \rangle < w_0 \end{bmatrix}.$$

Метод опорных векторов (SVM) позволяет найти разделяющую гиперплоскость, даже если исходные множества линейно неразделимы, в этом случае итоговый классификатор будет работать с некоторой ошибкой. Для апробации этого метода интерполяционная таблица использовалась как есть без дополнительных преобразований. В качестве SVM реализации использовался проект SVM-Light [18].

### 5.3 МНК аппроксимация

Одним из способов синтеза $\rho(x)$ является ее представление в виде линейного разложения в функциональном базисе, т.е. $\rho(x)=\sum_{j=1}^{M_B} K_j \cdot \beta_j(x)$, где $\beta_j$ - j-ая базисная функция, $M_B$ - число базисных функций, $K_j \in R$ - коэффициенты разложения.

Для нахождения коэффициентов $K_j$ методом наименьших квадратов [22] отыскивается минимум суммы квадратов отклонений вида $E_1=\sum_{i=1}^{N_t}\left(\rho^i-\sum_{j=1}^{M_B} K_j \cdot \beta_j(x^i)\right)^2$, где $x^i$ - i-ая интерполяционная точка, а $\rho^i$ - значение целевой функции в этой точке. Для этого необходимо решить систему из $M_B$ линейных уравнений вида $\frac{\partial E_1}{\partial K_j}=0$. Итоговый вид j-ого уравнения следующий $\sum_{k=1}^{M_B} A_{kj} \cdot K_k = B_j$, где $A_{kj}=\sum_{i=1}^{N_t}\beta_k(x^i)\cdot\beta_j(x^i)$ и $B_j=\sum_{i=1}^{N_t}\rho^i\cdot\beta_j(x^i)\cdot$

В экспериментах в качестве базиса использовались полиномы следующего вида:

$$\begin{cases}\beta_k=1, & если \quad k=1\\ \beta_k=(x_j)^n, & если \quad k=F\cdot(j-1)+n+1,\ j=1..F, n=1..D_n\end{cases}$$

где $D_n$ варьировалась от 2 до 9. Также вместо обычных полиномов использовались полиномы Чебышева до 3 степени.

Основным недостатком этого подхода является отсутствие гарантий того, что после определения $K_j$ будет иметь место $\sum_{j=1}^{M_B} K_j \cdot \beta_j(x^i)=\rho^i$, а также отсутствие способов, позволяющих как-то влиять на итоговый результат, за исключением возможности выбрать сам функциональный базис. На практике это означает, что обученный извлекатель будет неправильно выбирать варианта распознавания $c_r$ из множества $C_t$ для обучающего примера $t_i \in T_{teach}$. Эта проблему далее будем назвать проблемой недостаточной обученности.

### 5.4 МГУА аппроксимация

Одним из подходов к решению проблемы недостаточной обученности является наращивание степени $D_n$ полиномов, в теории при $D_n=N_t$ МНК гарантирует совпадение синтезированной функции с ожидаемыми значениями в интерполяционных точках, однако при $N_t \in [10^2...10^3]$ реализация такого подхода становится непрактичной.

Отчасти данную проблему можно решить, используя метод группового учета аргументов (МГУА) [8]. В данной работе был реализован МГУА с линейными частными описаниями вида $\rho_{kj}(x)=K_0+K_k\cdot x_k+K_j\cdot x_j$. В качестве критерия регулярности в экспериментах использовалась выражение вида

$$\delta^2=\frac{1}{N_t}\left(\sum_{i=1}^{N_t}\left(\rho_{kj}(x^i)-\rho^i\right)^2+\sum_{i=1}^{N_t} err(x^i)\right),$$

где $err(x^i)=1$, если на этапе селекции на тестовом наборе частное описание $\rho_{kj}$ делает правильный выбор среди $C_t$ вариантов распознавания на тексте $t$, одному из которых соответствует точка $x^i$, в противном случае $err(x^i)=0$. Коэффициенты $K_k$ и $K_j$ на каждой итерации определяются методом МНК.

Достоинством данного подхода является возможность на каждой итерации учитывать семантику значений $\rho^i$. Так в нашем случае при выборе предпочтительных $\rho_{kj}$ из текущего ряда селекции, кроме среднеквадратичной ошибки учитывается привязка конкретных точек к вариантам распознавания одного и того же обучающего примера $t$. Это позволяет выбирать частные описания $\rho_{kj}$, которые не только в среднем мало ошибаются согласно $\delta^2$, но к тому же позволяют принимать корректные решения на максимальном числе обучающих примеров.

Недостатком данного подхода является отсутствие гарантий, что при наличии наложенных ограничений процесс обучения сойдется за конечное число итераций, а также то, что на некоторых итерациях могут быть ошибочно отброшены значимые переменные.

### 5.5 Деревья решений

Деревья решений, обычно, используются для классификации при помощи правил в иерархической, последовательной структуре, данное свойство позволяет использовать деревья решений для аппроксимации функции, имеющей конечное число дискретных значений, такой как $\rho(x)$ [20].

Известен ряд алгоритмов для построения дерева решений, таких как CART, ID3, C4.5 и некоторые другие [13,21], вне зависимости от деталей реализации, данные алгоритмы разбивают множество числовых признаков $X=\{x\}$ на подмножества, каждой из которых ассоциировано с одним из значений функции $\rho(x)$. Очевидно, что можно ожидать удовлетворительного результата только в случае, если количество таких подмножеств конечно и «хорошо» охватывается обучающей выборкой. Отдельно отметим, что из-за обрезки деревья решений не гаран-

тируют прохождение синтезированной функции через все интерполирующие точки $\{\rho^i\}$.

Вместе с тем, деревья решений могут обеспечить высокую точность распознания лишь на линейно разделимых множествах. Так как рассматриваемая в данной работе задаче не гарантирует линейной разделимости множеств $\{x:\rho(x)=1\}$ и $\{x:\rho(x)=0\}$, то представляет особый интерес экспериментальная апробация данного метода.

### 5.6 Нейронные сети

Для решения задач аппроксимации наиболее подходящими являются многослойные сети прямого распространения - многослойные персептроны и нейронные сети, использующие радиальные базисные функции [25]. Помимо входного и выходного слоя, нейронная сеть может содержать один или более скрытых слоев, количество которых выбирается на основе эмпирических критериев. Отметим, что использование персептрона без скрытых слоев возможно только в случае линейно разделимых множеств.

Для данной задачи была выбрана модель персептрона с одним скрытым слоем из 9 нейронов и 9 входных нейронов, что должно позволить данной модели обеспечивать точность на уровне модели МНК. Так как данные для обучения являются непрерывными, то в качестве функции активации использовалась сигмоидальная функция.

Для обучения многослойной сети обычно используется алгоритм обратного распространения ошибки. При этом в качестве критерия остановки обучения можно использовать критерий полного распознания всех примеров обучающего множества $\{\rho^i\}$, т.к. исходя из теоремы сходимости персептрона, можно обеспечить прохождение синтезируемой функции через все интерполирующие точки [26], что, однако, может потребовать неприемлемого большого времени обучения.

## 6 Экспериментальное сравнение методов обучения

### 6.1 Описание эксперимента

В рамках данной работы были проведены эксперименты, направленные на сравнение и анализ рассмотренных методов обучения на предмет их практической применимости в задаче извлечения почтовых адресов России в произвольных текстах в рамках on-line сервиса «Охотник за адресами» (http://www.ahunter.ru). За основу эталонной базы почтовых адресов был взят классификатор КЛАДР, для которого был построен полнотекстовый индекс в соответствии с положениями, изложенными выше.

В качестве признаков $\{f_j\}_{j=1}^{F}$ были выбраны:

- $f_1(t,c)$ - инвертированная сумма расстояний Левенштейна по всем распознанным адресным полям $c_{ij}$ и соответствующим им текстовым написаниям в $t$ (чем ближе каноническое написание $c_{ij}$ к его неканоническому представлению в $t$, тем $f_1(t,c)$ больше);
- $f_2(t,c)$ - количество слов в текстовом фрагменте $t$, не задействованных при распознавании $c$;
- $f_3(t,c)$ - количество распознанных числовых полей адреса (номер дома и пр.);
- $f_4(t,c)$ - количество верифицированных числовых полей адреса (номер дома и пр.);
- $f_5(t,c)$ - количество распознанных полей, содержащих тип адресного объекта (город, улица, бульвар, проспект и пр.);
- $f_6(t,c)$ - суммарное число слов во фрагменте $t$, задействованных при распознавании $c$;
- $f_7(t,c)$ - относительная позиция первого распознанного слова внутри фрагмента $t$;
- $f_8(t,c)$ - относительная позиция последнего распознанного слова текстового фрагмента $t$;
- $f_9(t,c)$ - количество полей в $c$ с устаревшими или синонимичными названиями.

Обучающая выборка была построена на основе анализа журналов работы сервиса по следующему принципу. Было отобрано ~320 текстовых фрагментов, на которых пользователи сервиса получили неоднозначные результаты. Эти тексты были вручную размечены, так, что в них посредством специальных тэгов явным образом были выделены поля извлекаемой адресной структуры. Пример такого текста приведен в таблице 2.

Табл. 2. Пример размеченного текста.

| Текст | Москва 2-ая Бауманская 5 |
|---|---|
| Разметка | <A:Region>Москва</A:Region><br><A:Street>2-ая Бауманская</A:Street><br><A:House>5</A:House> |

Каждый неразмеченный текст $t$ подавался на вход извлекателю в режиме обучения, при котором он выдавал все варианты извлечения $C_t$. Для каждого варианта $c_i \in C_t$ рассчитывался вектор значений признаков $\{f_j(c_t,t)\}_{j=1}^{9}$. Далее выполнялась покоординатная нормировка: значения каждого признака $f_j(c_i,t)$ у всех вариантов $c_i \in C_t$ делились на $\max_{c_i \in C_t} f_j(t,c_i)$. Вариант извлечения $c_p \in C_t$, совпадающий с показаниями разметки объявлялся позитивным и для него принималось значение $\rho(t,c_p)=1$, для остальных вариантов $c_n \in C_t \setminus c_p$ принималось $\rho(t,c_n)=0$.

Таким образом, была заполнена интерполяционная таблица (см. табл. 1). После устранения дублей, возникающих в результате покоординатной нормировки, в таблице было оставлено 520 интерполяци-

онных точек, на которых и проводились эксперименты.

В рамках эксперимента были подготовлены срезы полной обучающей выборки, содержащие 5, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90 и 100 процентов ее интерполяционных точек. Каждый метод обучения, таким образом, применялся к каждому из указанных срезов.

Тестирование каждой обученной модели проводилось на полной выборке в 520 интерполяционных точек. При этом подсчитывалось число ложных срабатываний, которые использовались для расчета точности обученной модели согласно выражению $P = \frac{N_s - E}{N_s}$, где $N_s$ - число исходных текстовых примеров (не интерполяционных точек) по отношению к которым применяется тестируемый метод снятия неоднозначности распознавания, $E$ - число текстовых примеров, на которых распознаватель допустил ошибку.

### 6.2 Результаты эксперимента

В таблице 3 приведены значения точности распознавания в зависимости от метода обучения и от размера обучающей выборки.

В первой колонке таблицы 3 приведены названия методов, принимавших участие в экспериментах для обучения распознавателя. Остальные колонки соответствуют размерам выборки, на которой проводилось обучение. Числа в прочих ячейках таблицы отражают значения точности распознавателя, обученного по методу, соответствующему строке ячейки, и на выборке, соответствующей колонке ячейки.

Как отмечалось, в обучении по МНК пробовались разные функциональные базисы (полиномы различных степеней и полиномы Чебышева), которые, как показали эксперименты, не влияют коренным образом на характер зависимости точности от размера выборки. Поэтому предпочтение было отдано базису $\beta_j(x) = x_j$, что фактически соответствует представлению $\rho(x) = \langle K \cdot x \rangle$, где $K$ - вектор коэффициентов разложения. В таблице 3 этому эксперименту соответствует строка с названием «МНК (линейный)». Также в таблице 3 приведены результаты экспериментов для МНК с разложением по базису полиномов Чебышева до 3-ей степени, этим результатам соответствует строка с названием «МНК (Чебышев)». На рис. 1 отражены диаграммы, демонстрирующие небольшую разницу между этими двумя видами МНК.

На рис. 2 приведено графическое представление полученных данных по всем методам из табл. 3, кроме МНК с разложением по базису полиномов Чебышева.

В случае с линейным МНК, использующим базис $\rho(x) = \langle K \cdot x \rangle$, представляет интерес его сравнение с SVM, поскольку они реализуют одну и ту же идею разделяющей гиперплоскости, но различными способами. Полученные показатели точности указывают на небольшое превосходство SVM, которое, однако, было достигнуто за счет опытного подбора управляющего параметра $C \geq 100$ (масштабный коэффициент, позволяющий задать компромисс между шириной зазора между разделяемыми множествами и суммарной ошибкой классификации по всем обучающим примерам [18]).

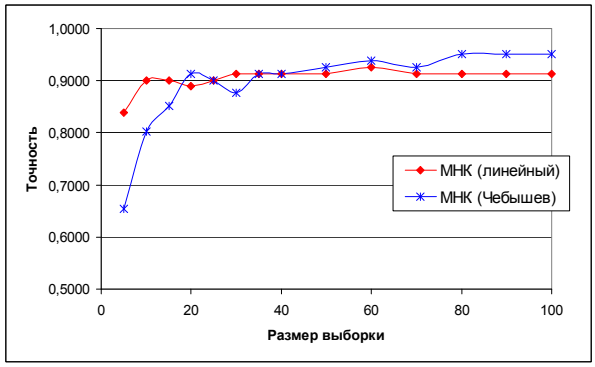



Рис. 1. Сравнение двух видов МНК.

При выполнении аппроксимации МГУА сходимость процесса обучения ограничивалась скоростью обучения так, что процедура завершалась, когда минимальное значение критерия регулярности на текущем слое селекции уменьшалось менее чем на 1% по отношению к этому же значению от предыдущего слоя. На каждом слое селекции оставлялось до 9 узлов с частными описаниями. Результирующие структуры, формируемые по данному методу, при обучении на разных выборках включали от 4 до 6 слоев.

Примечательным является тот факт, что для большинства методов (все кроме дерева решений и персептрона) рост размера выборки не гарантирует возрастания точности распознавателя. Например, для линейного МНК и SVM имеет место убывание точности между точками 15% и 20%.

У методов МНК, SVM, МГУА и классификатора Байеса общим свойством является недостаточная обученность, фактически не позволяющая гарантировать отсутствие ошибок распознавания на тех же данных, на которых обучался распознаватель.

Вместе с тем эти методы демонстрируют быструю сходимость процесса обучения. Так что, начиная с обучающей выборки, содержащей 35% примеров от тестового множества, возрастание точности распознавания практически прекращается. Данный факт говорит о хорошей способности к обобщению входных данных у этих методов обучения.

Перечисленные три свойства: быстрая сходимость, колебания точности и недостаточная обученность - можно объяснить ограниченными выразительными возможностями моделей, лежащих в основе этих методов. Фактически для этих методов характерно жесткое определение структуры обучаемой модели: для МНК и МГУА – линейная комбинация базисных функций, для SVM – уравнение гиперплоскости, для классификатора Байеса – ступенчатые функции распределения вероятностей.

Таблица 3. Точность обученных распознавателей на разных выборках.

| | 5% | 10% | 15% | 20% | 25% | 30% | 35% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **МНК (линейный)** | 0,8395 | 0,9012 | 0,9012 | 0,8889 | 0,9012 | 0,9136 | 0,9136 | 0,9136 | 0,9136 | 0,9259 | 0,9136 | 0,9136 | 0,9136 | 0,9136 |
| **МНК (Чебышев)** | 0,6543 | 0,8025 | 0,8519 | 0,9136 | 0,9012 | 0,8765 | 0,9136 | 0,9136 | 0,9259 | 0,9383 | 0,9259 | 0,9506 | 0,9506 | 0,9506 |
| **МГУА** | 0,8395 | 0,8889 | 0,9259 | 0,9259 | 0,9383 | 0,9383 | 0,9259 | 0,9383 | 0,9259 | 0,9259 | 0,9383 | 0,9630 | 0,9506 | 0,9753 |
| **Байес** | 0,3210 | 0,6914 | 0,7407 | 0,7531 | 0,7901 | 0,7778 | 0,7778 | 0,7654 | 0,7654 | 0,7778 | 0,7901 | 0,8025 | 0,8025 | 0,8025 |
| **SVM** | 0,9012 | 0,9506 | 0,9383 | 0,9136 | 0,9259 | 0,9383 | 0,9259 | 0,9259 | 0,9259 | 0,9259 | 0,9506 | 0,9506 | 0,9383 | 0,9383 |
| **Дерево решений** | 0,2963 | 0,6049 | 0,6296 | 0,6296 | 0,6420 | 0,6420 | 0,6420 | 0,6420 | 0,6543 | 0,6667 | 0,6667 | 0,6914 | 0,7531 | 0,7778 |
| **Персептрон** | 0,4321 | 0,4938 | 0,5432 | 0,6049 | 0,6914 | 0,7284 | 0,7407 | 0,7654 | 0,7778 | 0,8025 | 0,8148 | 0,8765 | 0,9259 | 0,9506 |

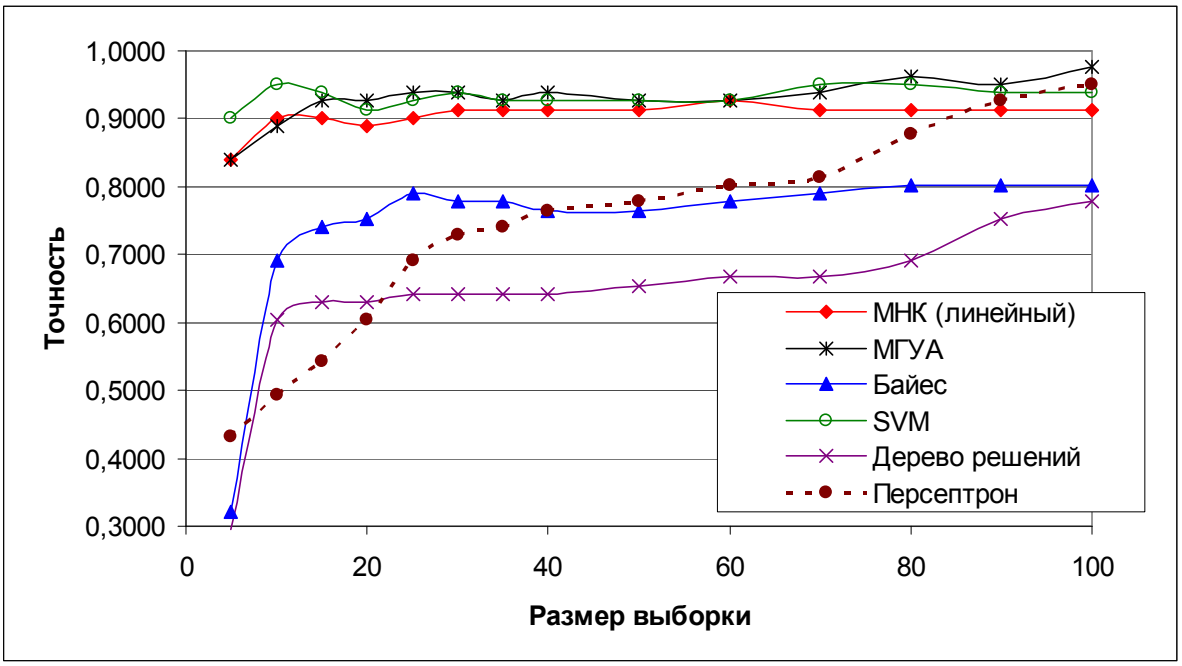



Рис. 2. Точность обученных распознавателей.

Информации в небольшой обучающей выборке оказывается достаточно, чтобы настроить все параметры данных моделей и достичь предела их выразительных возможностей. С дальнейшим ростом обучающей выборки возрастает количество информации, которое нужно учесть в параметрах модели, что фактически не приводит к качественному росту точности, а вызывает лишь небольшие колебания относительно достигнутого порога.

Так же следует учитывать, что с ростом объема обучающей выборки возрастает и ее зашумленность, обусловленная ошибками экспертов при подготовке выборки.

Деревья решений, как и ожидалось, показали наихудший результат, что связанно как с простотой данной модели, так и с тем, что к построенному дереву применялась процедура подрезки тех ветвей, уровень доверия к которым был менее 80%. Исключение процедуры подрезки ветвей могло бы несколько улучшить результат, но, вместе с тем, на больших массивах данных это могло бы привести к переобученности модели. Хорошо заметно, что точность распознания данного метода сравнима с классификатором Байеса, что связанно с тем, что в основе данных моделей лежит схожий вероятностный принцип выбора решения.

Нейронные сети, а именно персептрон с одним скрытым слоем, показал хорошую точность распознания в случаях, когда обучающее множество покрывало более 60% тестового множества. Плохие результаты при обучающем множестве менее 30% от тестового объяснимы тем, что для данной модели персептрона — 9 входных нейронов и один скрытый слой, содержащий 9 нейронов, эти выборки недостаточны для обучения и вызывают явление, известное как переобученность.

Проведенные эксперименты для модели персептрона без скрытых слоев показали, что данная модель даже на полной выборке обеспечивает невысокую точность, что связанно с линейной неразделимостью обучающего множества. Численные результаты по этим экспериментам в работе не приводятся.

### 6.3 Выводы по экспериментам

Для данной задачи была также выполнена ручная аппроксимация, которая заключалась в подборе экспертом эвристики, описывающей, с его точки зрения, функцию $\rho(x)$ наилучшим образом. Точность этой эвристики на тестовом наборе составила 0.93, поэтому целесообразность использования рассмотренных методов обучения оценивалась на основе сравнения этого значения с их показаниями точности.

Деревья решений, классификатор Байеса и однослойный персептрон показали свою неприменимость к решению исходной задачи. Несмотря на то, что данные модели обладают высокой скоростью работы, их точность оказалась ниже точности эвристической аппроксимации.

Методы МНК и SVM обеспечивают приемлемую точность распознавания, однако для них не существует способов управления процессом обучения, позволяющих влиять на итоговую точность. Метод МГУА показал лучший результат, причиной чего стала возможность управления процессом обучения, путем введения на этапе селекции релевантного для поставленной задачи критерия регулярности.

Многослойный персептрон продемонстрировал высокую точность распознания при представительной обучающей выборке (сравнимой с адаптированным к задаче методом МГУА) однако на данной выборке не удалось оценить обобщающие возможности этой модели.

## 7 Заключение

В данной работе предложен метод извлечения целевой информации по эталону. Полагается, что система извлечения обладает полной априорной базой так, что распознавание в текстах возможно только той информации, которая имеется в этой базе. Не смотря на кажущуюся ограниченность такого подхода, на практике он находит применение в различных областях таких, как распознавание Ф.И.О. и извлечение топонимов в произвольных текстах. Отсутствие сильной зависимости от формы текста является главным достоинством такого подхода, вытекающим из того, что не возникает необходимости запоминать структуру текста, в отличие от методов извлечения общего вида.

Авторами также разработан метод снятия неоднозначности распознавания, являющейся ключевой проблемой в задачах извлечения информации из текстов. Решение этой задачи рассмотрено с позиции машинного обучения на примерах. Проанализировано несколько вариантов такой реализации, каждая из которых проверена экспериментально на задаче выявления почтовых адресов России в произвольных текстах. Среди рассмотренных способов обучения наиболее предпочтительными являются те из них, которые оперируют многослойными структурами. В нашем случае к ним относится многослойный персептрон и МГУА-аппроксиматор.

Разработанные методы реализованы в виде прототипа системы извлечения по эталону, работоспособность которого можно проверить on-line по адресу: http://www.ahunter.ru.

## Литература


[1] Berger A.L., Della Pietra V.J., Della Pietra S.A. A maximum entropy approach to natural language processing // Computational Linguistics archive. – 1996. – Vol. 22, Issue 1, – P. 39–71.

[2] Borkar V., Deshmukh K., Sarawagi S. Automatic segmentation of text into structured records // Proceedings of the 2001 ACM SIGMOD international conference on Management of data. – 2001. – P. 175–186.

[3] Califf M.E., Mooney R.J. Bottom–up relational learning of pattern matching rules for information extraction // Journal of Machine Learning Research. – 2003. – Vol. 4. – P. 177–210.

[4] Chai J.Y., Biermann A.W., Guinn C.I. Two dimensional generalization in information extraction // In Proceedings of the Sixteenth National Conference on Articial Intelligence. – 1999. – July. – P. 431–438.

[5] Dejean H. Learning rules and their exceptions // The Journal of Machine Learning Research archive. – 2002. – Vol. 2 (March). – P. 669–693.

[6] Freitag D. Machine Learning for Information Extraction in Informal Domains // Machine Learning. – 2000. – Vol. 7. – P. 169–202.

[7] Grishman R., Sundheim B. Message Understanding Conference–6: a brief history // Proceedings of the 16th conference on Computational linguistics. – 1996. – Vol.1. – P. 466 – 471.

[8] Group Method of Data Handling [Электронный ресурс] – Режим доступа: http://www.gmdh.net/, свободный.

[9] Hai Leong Chieu, Hwee Tou Ng. A maximum entropy approach to information extraction from semi–structured and free text // Eighteenth national conference on Artificial intelligence. – 2002. – P. 786–791.

[10] Huffman S.B. Learning to extract information from text based on user–provided examples // Proceedings of the fifth international conference on Information and knowledge management. – Rockville, Maryland, (United States), 1996. – P. 154–163.

[11] Kim J., Moldovan D. Acquisition of Linguistic Patterns for Knowledge–Based Information Extraction // IEEE Transactions on Knowledge and Data Engineering archive. – 1995. – Vol. 7, Issue 5. – P. 713–724.

[12] McCallum A. An Introduction to Conditional Random Fields for Relational Learning / C. Sutton, A. McCallum // Introduction to Statistical Relational Learning / Edited by Lise Getoor and Ben Taskar. – MIT Press, 2007. – P. 95–130.

[13] Murthy S. Automatic construction of decision trees from data: A Multi-disciplinary survey 1997
[14] Pedersen T. A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation // Proceedings of the first conference on North American chapter of the Association for Computational Linguistics. – 2000. – P. 63 – 69.
[15] Riloff E. Automatically Constructing a Dictionary for Information Extraction Tasks // In Proceedings of the 11th National Conference on Artificial Intelligence (AAAI). – 1993. – P. 811–816.
[16] Soderland S. Learning information extraction rules for semi–structured and free text. Machine Learning. – 1999. – Vol. 34, Issue 1–3. P. 233–272.
[17] Soderland S. Crystal: Inducing a conceptual dictionary / S. Soderland, D. Fisher, J. Aseltine, We. Lehnert // In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. – 1995. – P. 1314–1319.
[18] SVM-Light Support Vector Machine [Электронный ресурс] – Режим доступа: http://svmlight.joachims.org/, свободный.
[19] Turmo J., Ageno A., Catala N. Adaptive information extraction // ACM Computing Surveys archive. – 2006. – Vol. 38, Issue 2. – Article No. 4.
[20] Venkatesan T. Chakaravarthy, Vinayaka Pandit, etc. Decision trees for entity identification: approximation algorithms and hardness results. //Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp 53 – 62, 2007.
[21] Yuan Y. and Shaw M. J. Induction of fuzzy decision trees - Fuzzy Sets Syst., vol. 69, pp. 125-139, 1995.
[22] Аппроксимация методом наименьших квадратов (МНК) [Электронный ресурс] – Режим доступа: http://alglib.sources.ru/interpolation/linearleastsquares.php, свободный.
[23] Карманов В.Г. Математическое программирование: учебное пособие. – 5-ое изд., стереотип. – М.: ФИЗМАТЛИТ, 2004. – 264 с.
[24] Кормалев Д.А. Система извлечения информации из текстов INEX / Д.А. Кормалев, Е.П. Куршев, Е.А. Сулейманова, И.В. Трофимов // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ–2004: Труды конференции. – М.: Физматлит, 2004. – Т.3. – С. 908–915.
[25] Осовский С. Нейронные сети для обработки информации. – М.: Финансы и статистика, 2002. – 344 с.: ил. С. 46-89.
[26] Розенблатт, Ф. Принципы нейродинамики: Перцептроны и теория механизмов. — М.: Мир, 1965. — 480 с.
[27] Симаков К.В. Модель извлечения знаний из естественно-языковых текстов / А.М. Андреев, Д.В. Березкин, К.В. Симаков // Информационные технологии. – 2007. – №12. – С. 57–63.
[28] Симаков К.В. Метод обучения модели извлечения знаний из естественно-языковых текстов / А.М. Андреев, Д.В. Березкин, К.В. Симаков // Вестник МГТУ. Приборостроение.–2007. – №3.– С. 75–94.

## Machine learning in information extraction having etalon database

Alexeev S, Morozov V, Simakov K

We describe a special case of task of information extraction from texts when a whole database of objects to extract is already exists. Such database includes only canonical representations of objects, so the task is to recognize them by their non-canonical descriptions in texts. To disambiguate the result of such recognition we research, test and compare a range of machine learning methods. The result of such comparison is also described.

# Извлечение информации из текста в системе ИСИДА-Т♣

© Д.А.Кормалев, Е.П.Куршев, Е.А.Сулейманова, И.В.Трофимов

Исследовательский центр искусственного интеллекта ИПС РАН
dkormalev@acm.org, epk@epk.botik.ru, yes@helen.botik.ru, itrofimov@km.ru

**Аннотация**

Статья посвящена методам и подходам к извлечению информации из текста на естественном языке, реализованным в системе ИСИДА-Т. Основной акцент сделан на представлении знаний и распознавании текстовых ситуаций.

## Введение

Значительная доля информации, доступной в электронном виде, представлена в виде текстов на естественном языке. Заключенная в них полезная информация не структурирована, а значит, ее невозможно обработать и проанализировать классическими вычислительными методами и средствами. Тексты могут быть прочитаны и поняты человеком, но для вычислительной машины они — всего лишь цепочки символов. Меж тем, машинная обработка информации существенно ускоряет любой рабочий процесс и обеспечивает качество результата. Объем накопленной текстовой информации заставляет задуматься о средствах автоматической обработки текстов.

Технология извлечения информации из текстов на естественном языке (ТИИ) [8] — это технология обработки текста, которая позволяет автоматически «просматривать» относительно большой объем текстов, содержащих относительно небольшое количество искомой информации. Обнаруженная в тексте информация преобразуется в структурированный формат: выявляются целевые факты, объекты, отношения в виде, пригодном для дальнейшей автоматической обработки (статистической обработки, визуализации, поиска закономерностей в данных и других).

Иногда ТИИ рассматривают как специфическую разновидность информационного поиска. Отличия ТИИ от информационного поиска заключаются в том, что «запросы» должны быть известны заранее, а результатом является не набор ссылок на документы, а построенные структуры данных, описывающие релевантные факты из набора документов.

Приведем несколько областей применения ТИИ:

- расширение возможностей информационного поиска (поиск не по ключевым словам, а по фактам, ситуациям, объектам, отношениям);
- построение досье на персон или организации из открытых текстовых источников;
- мониторинг сообщений СМИ (примеры событий, которые могут представлять интерес: слияния и поглощения компаний, появление новых игроков на рынке, выпуск новой продукции, теракты);
- извлечение специфической метаинформации из коллекций документов большого объема (например, построение по текстовой базе муниципальных нормативно-правовых актов, связанных с недвижимостью, реляционной базы данных с информацией о типах событий, объектах и субъектах).

Первоначально задача ТИИ формулировалась как выделение фрагментов текста, содержащих релевантную информацию, и, возможно, преобразование их в реляционную форму. Для решения задачи в такой постановке часто достаточно анализировать локальный контекст, используя ограниченный набор знаний предметной области. Назовем такую технологию *извлечением информации в «слабом» смысле*. Результаты извлечения информации в «слабом» смысле и характер их представления несколько ограничивают возможности дальнейшего использования добытых из текста данных. *Извлечением информации в «сильном» смысле* мы назвали бы переход от базы текстовых фактов к такому их представлению, которое можно было бы использовать как интеллектуальный информационный ресурс, своего рода базу текстовых знаний.

Наши исследования были направлены на усовершенствование методов и расширение возможностей ТИИ, что позволило бы подойти вплотную к решению задачи извлечения информации в «сильном» смысле. Полигоном для экспериментальной проверки идей и практического воплощения разработанных подходов стала система ИСИДА-Т[1], над которой мы работаем в течение нескольких лет.

Чтобы получить информацию из прочитанного фрагмента текста (понять текст), человек должен знать язык, на котором написан текст, и располагать некоторым объемом «фоновых» знаний. Аналогично, система извлечения информации из

**Труды 11ій Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

текста должна располагать двумя видами средств: средствами анализа естественного языка и некоторым объемом знаний предметной области. Однако прежде чем приступать к рассмотрению этих средств, остановимся на общей организации и инфраструктуре системы.

## 1 Общая организация системы

Краеугольным камнем системы ИСИДА-Т является точная настройка на предметную область и конкретную задачу извлечения. С одной стороны, это достигается за счет редактирования лингвистических ресурсов, ресурсов знаний, правил извлечения и правил трансформации. С другой стороны, настройка может потребовать включения в процесс обработки дополнительных специализированных методов обработки текста. Кроме того, для каждой задачи необходимо подобрать наиболее подходящие алгоритмические средства анализа из набора имеющихся. Эти аспекты требуют создания такой архитектуры, при которой легко могут добавляться и замещаться алгоритмические компоненты процесса извлечения.

Проблема конфигурирования на алгоритмическом уровне потребовала создания модульной архитектуры и декларативного подхода к определению процесса извлечения. Модули получили название обрабатывающих ресурсов в противовес лингвистическим ресурсам и ресурсам знаний. В конфигурации декларируется порядок обработки документа аналитическими модулями, потоки данных между ними, а также параметры их работы.

Обрабатывающие ресурсы можно разделить на следующие группы.

- *Ресурсы предобработки.* Сюда относятся средства определения кодировки документа, извлечения текста и стилевой разметки из документа, предварительной фильтрации.
- *Ресурсы лингвистического анализа.* Осуществляют разбор текста на отдельные слова, морфологический анализ (в том числе специализированные варианты для различных категорий имен собственных), поверхностный синтаксический анализ и определение границ предложений.
- *Ресурсы извлечения.* Осуществляют поиск в документе целевой лексики и синтаксических конструкций, а также первичное структурирование информации.
- *Ресурсы унификации знаний и вывода.* Осуществляют унификацию и отождествление элементов знаний, вывод производных знаний.
- *Ресурсы подготовки результата.* Осуществляют приведение извлеченной информации к определенному формату и передачу за пределы последовательности обработки (в БД, глобальный ресурс знаний, файл, приложение).

В целом средства конфигурирования выполняют те же функции, что и каркас (framework) в известных системах GATE [10] и UIMA [11]. Эти средства обеспечивают (1) расширяемость архитектуры, (2) управление потоками данных, (3) поддержку моделей разделяемой памяти, (4) настройку параметров обрабатывающих ресурсов и последовательности обработки, (5) упрощение и унификацию процесса разработки новых обрабатывающих ресурсов. Проблемой хранения результатов анализа в нашем подходе занимаются, преимущественно, сами обрабатывающие ресурсы. Обрабатывающие ресурсы мы реализуем в формате динамически загружаемых библиотек (или разделяемых объектов).

## 2 Средства анализа естественного языка

Средства анализа естественного языка, используемые в ТИИ, можно разделить на две большие категории: средства общего лингвистического анализа и предметно-ориентированные методы распознавания текстовых ситуаций.

Средства общего лингвистического анализа включают в себя графематический, морфологический и синтаксический анализ. Эти средства применимы практически во всех предметных областях, существует ряд реализаций с довольно высокими показателями качества, поэтому мы не будем останавливаться на них подробно.

Вторая категория — средства и методы распознавания текстовых ситуаций, характерных для решаемой задачи и предметной области. Распознавание текстовых ситуаций состоит в выделении фрагментов текста, описывающих объекты, и содержательных связей между этими фрагментами, основанных в той или иной мере на синтаксисе естественного языка. Можно рассматривать распознавание ситуаций как ориентированный на предметную область частичный, но точный синтактико-семантический анализ.

Распознавание опирается на сопоставление образцу, который задается при помощи правил на специализированном формальном языке. Правила определяют не только образец, но и действия, которые должны быть выполнены при успешном сопоставлении. Правила работают не с текстом как последовательностью символов, а со структурами, построенными «над» текстом и выражающими лингвистическую и предметную информацию о нем.

Для упрощения конфигурирования системы очень желательно, чтобы все модули использовали одинаковый способ представления информации о тексте (разметки текста). В системе ИСИДА-Т все модули, в том числе средства общего лингвистического анализа, используют структуры данных, описанные ниже.

### 2.1 Разметка текста и структуры данных

В различных системах обработки текста на естественном языке используется широкий спектр средств для представления лингвистической и предметно-ориентированной информации о тексте в целом или его фрагментах. Единого подхода к представлению разметки текста и информации о нем не существует.

В последнее десятилетие довольно широко используется способ представления информации о тексте, основанный на так называемых *аннотациях*, отличающийся простотой и высокой степенью универсальности [12]. На сегодняшний день многие системы обработки текста в той или иной степени используют идеи модели аннотаций. Эта модель используется и в нашем подходе.

*Аннотация* — объект, который приписывается фрагменту текста (например, слову, словосочетанию, предложению, ссылке на сущность предметной области и т.д.) и описывает свойства этого фрагмента. Аннотации разбиты на конечное множество классов. Каждый класс аннотаций описывает текст в определенном аспекте. Информация о фрагменте представлена значениями именованных атрибутов аннотации. Наборы классов и атрибутов аннотаций намеренно не специфицированы, чтобы можно было использовать произвольный набор обрабатывающих модулей и представлять необходимую лингвистическую и предметную информацию. Обмен данными между модулями тоже идет в терминах аннотаций: новые аннотации могут строиться на основании полученных на предыдущих этапах анализа.

Из способа представления информации с помощью аннотаций следует возможность разработки средств анализа текста, компоненты которых слабо связаны между собой. Не отражающееся на функциональных характеристиках сложной системы уменьшение числа зависимостей между ее составляющими облегчает ее понимание, разработку и поддержку. Слабая связность является существенным преимуществом, так как повышает возможность повторного использования компонентов и снижает риск критических сбоев, вызванных неправильным взаимодействием компонентов (например, из-за того, что в цепочке обработки какой-то компонент ошибочно не был зарегистрирован, или же частично нарушился порядок обработки).

Впрочем, базовая модель аннотаций не лишена недостатков. В частности, она не подразумевает средств проверки соответствия атрибутов и их значений. Атрибуты могут быть только атомарными. Отсутствуют возможности установления связей между отдельными аннотациями. Нет средств для контроля расположения границ аннотаций разных классов, в то время как для большинства классов аннотаций можно задать условия, описывающие их взаимное расположение. Например, аннотации, описывающие синтаксис предложения в терминах системы составляющих, не могут пересекаться – для них возможно только отношение строгого вхождения или совпадения.

В реализации системы ИСИДА-Т модель аннотаций была дополнена некоторыми полезными средствами. В частности, было снято ограничение на атомарность атрибутов и добавлена возможность устанавливать ссылки между аннотациями.

### 2.2 Язык правил распознавания текстовых ситуаций

Для распознавания текстовых ситуаций используется набор правил, описывающих характерные для конкретной задачи способы выражения ситуации в тексте. Эти правила задают образец для сопоставления и действия, которые должны быть произведены после успешного сопоставления. Качество работы (полнота и точность) ТИИ тесно связано с возможностями языка правил. Ряд современных систем извлечения информации (в том числе, система ИСИДА-Т) берут за основу различные диалекты языка CPSL [7]. Использование этого языка подразумевает разметку текста при помощи аннотаций.

Единицей трансляции языка правил является фаза. Правила, входящие в одну фазу, применяются в недетерминированном порядке. Результаты фазы — изменения, внесенные в набор аннотаций после работы правил — фиксируются после применения всех правил и становятся доступны в последующих фазах. Поэтому правило не может использовать результаты работы другого правила из этой же фазы. Можно рассматривать фазу как модуль для специфического анализа текста. Работа фаз может перемежаться применением произвольных обрабатывающих ресурсов.

Правило — основная единица языка. Правила представляются в виде «образец→ действие». Здесь «образец» — образец для поиска в терминах высказываний о взаимном расположении и значениях атрибутов аннотаций разных классов (левая часть правила); «действие» — набор действий, выполняемых при успешном сопоставлении (правая часть правила). По структуре левая часть правила во многом схожа с регулярным выражением, но существенное отличие состоит в том, что роль символов в правиле играют тесты. Тест представляет собой конъюнкцию высказываний (элементарных тестов) о значениях атрибутов аннотаций разных классов. Из тестов могут образовываться сложные конструкции с использованием следования, альтернативы, квантификаторов и скобок. Чтобы обозначить границы фрагментов текста, сопоставленных подвыражениям, используются метки. Метка — это идентификатор, которым помечается образец. В дальнейшем (при выполнении действий в правой части правила) можно использовать метку для ссылки на фрагмент текста, сопоставленный подвыражению.

Язык правил, используемый в системе ИСИДА-Т, является расширением CPSL. Предлагаемые нами расширения преследуют две цели: 1) обеспечить возможность описывать более сложные контексты, в которых встречается целевая информация, и 2) снизить объем рутинной работы при создании системы правил за счет более компактного описания контекста [5].

Отличия от других реализаций, например, JAPE [10] или диалекта CPSL, используемого в продуктах RCO [3] состоят в следующем.

- Для передачи информации между элементарными тестами, а также в правую часть правил могут использоваться именованные переменные, значения которых присваиваются явно в ходе сопоставления. Множество значений переменных входит в контекст сопоставления. Использование переменных позволяет компактно описывать отношения между атрибутами аннотаций, рассматриваемых в разных элементарных тестах. В частности, этот механизм обеспечивает компактное описание согласования языковых единиц, рассматриваемых в различных состояниях конечного автомата.
- Реализована встроенная поддержка расширенного спектра типов данных, в том числе, ссылок на аннотации и множественных значений. Данные этих типов могут использоваться в качестве значений переменных и значений атрибутов аннотаций.
- Логика работы интерпретатора правил приведена в максимальное соответствие поведению интерпретатора обычных регулярных выражений. Отличия от современной реализации JAPE и Montreal transducer [14] заключаются в поддержке «жадных» и «нежадных» квантификаторов и опережающей проверки.
- Поддерживаются кванторы существования (по умолчанию) и всеобщности, связывающие элементарные тесты. К кванторам может добавляться отрицание.
- Существуют языковые средства, позволяющие гибко проверять взаимное расположение аннотаций, рассматриваемых в контексте сопоставления, и прочих аннотаций во входной коллекции.
- В тестах могут использоваться функции для обращения к ресурсу знаний (раздел 3), например, проверки таксономической принадлежности элементов. Для более сложных запросов к ресурсу знаний используется предметно-ориентированный язык, совпадающий с языком описания левой части правил трансформации (подраздел 3.2).

Общая проблема средств распознавания текстовых ситуаций — при расширении функциональных возможностей этих средств резко падает производительность. Для решения этой проблемы мы использовали два основных способа оптимизации интерпретатора правил: предобрабатывать правила, анализируя потоки управления [9, 13], и сокращать перебор кандидатов при выполнении тестов [4]. Внедрение этих модификаций позволило ускорить интерпретацию правил в среднем в 6 раз в зависимости от конфигурации системы и качества входных данных (в отдельных случаях наблюдался прирост производительности до двух порядков). В большинстве случаев повышение производительности сопровождалось снижением расхода памяти на 20-40%.

## 3 Ресурс знаний

Практически в любой предметной области для точного извлечения требуются априорные знания о ней — знания о понятиях, объектах и отношениях, связанных с целями извлечения или являющихся целями. В свою очередь, извлеченная из текстов информация может нести в себе новые знания о предметной области и быть полезна для дальнейшей автоматической обработки текста. Тесная связь между априорной и извлеченной информацией, а также между предметными и лингвистическими знаниями сформировала потребность в унификации средств представления.

### 3.1 Представление знаний

Интегрированный ресурс знаний (РЗ) [1] системы ИСИДА-Т объединяет в себе базу априорных предметных знаний, хранилище фактографической информации и словарь. Предметные знания хранятся в РЗ в структурах, называемых *элементами знаний*. Элементы знаний делятся на 4 категории [6]: 1) концепты (CT), 2) экземпляры концептов (CI), 3) типы предметных отношений (RT), 4) экземпляры отношений (RI). Наш подход к представлению знаний использует элементы семантических сетей и систем фреймов.

Концепты и типы отношений служат для представления онтологической информации о предметной области и задаются априорно. Экземпляры концептов и отношений составляют базу фактов предметной области и могут быть как априорными, так и извлеченными из текстов.

Для каждого элемента знаний задается набор атрибутов. В списках атрибутов CT и RT хранятся пары «имя—ограничения на значение», в списках атрибутов CI и RI — пары «имя—значение». В терминах системы фреймов CT и RT выражались бы прототипами фреймов, а CI и RI — экзофреймами. Неявно определены два специальных (служебных) типа отношений: ISA и AKO. Их интерпретация такая же, как в системах фреймов.

Лингвистическая составляющая ресурса знаний — словарь. Словарь связан с базой предметных знаний посредством ссылок от дескрипторов к элементам знаний: дескрипторы словаря базовой лексики ссылаются на концепты, а дескрипторы словаря собственных имен — на априори известные экземпляры концептов из базы фактов. В отличие от тезауруса, дескрипторы в словаре базовой предметной лексики не связаны друг с другом никакими парадигматическими отношениями (последние выражаются с помощью отношений между соответствующими элементами базы предметных знаний).

Словарь предоставляет возможность указывать дополнительные ограничения на все словоформы, входящие в состав дескриптора и синонимов, чтобы увеличить точность распознавания словарных единиц в тексте.

Унификация априорных и извлеченных из текстов знаний удобна тем, что позволяет использовать одни и те же средства для работы с обоими типами знаний. Объединение лингвистических и предметных знаний в одном ресурсе, во-первых, облегчает первичное наполнение и последующую поддержку, а во-вторых, дает возможность использовать предметные знания уже на этапе первичной обработки текста правилами извлечения информации. Благодаря специально разработанному языку запросов к РЗ правила могут не ограничиваться словарной информацией, а обращаться в онтологию и базу фактов для проверки различных условий, требующих навигации по отношениям.

### 3.2 Трансформации

После извлечения информации из текста и помещения ее в хранилище фактографической информации часто требуется дополнительная обработка для ее унификации и уточнения. На основе такой обработки может решаться целый спектр задач:

- навигация по связанным объектам, фактам и ситуациям;
- определение и объединение тождественных элементов (некоторые случаи разрешения кореферентности);
- кластеризация сходных сюжетов;
- вывод имплицитной фактографической информации;
- генерация текстовых описаний фрагментов фактографической базы.

Для проведения экспериментов по преобразованию извлеченной фактографической информации был разработан язык трансформаций и выполнена экспериментальная программная реализация интерпретатора этого языка.

Трансформацию элементов ресурса знаний можно рассматривать как особый вид немонотонного вывода на знаниях. При трансформации происходит поиск образца ситуации в ресурсе знаний и выполнение указанных действий. Для описания ситуации можно задавать ограничения на типы элементов знаний, их атрибуты, наличие или отсутствие отношений того или иного типа между ними. Попытка выполнить действия производится для каждого набора элементов знаний, для которых выполняются условия, указанные в посылке правила трансформации. Набор действий включает в себя создание, удаление, модификацию элементов знаний, манипулирование их атрибутами.

Особенностью языка правил трансформации является сочетание декларативных и императивных элементов.

Язык трансформаций предназначен для описания правил, по которым выполняется преобразование элементов в хранилище фактографической информации. Правила языка схожи с продукционными правилами: каждое правило содержит образец для поиска (левая часть правила) и набор действий (правая часть), которые необходимо выполнить, когда образец был обнаружен.

При поиске образца используются следующие элементарные условия (перечислены только основные):

- проверка принадлежности элемента знаний к указанному классу;
- проверка наличия или отсутствия отношения указанного класса между элементами знаний;
- сравнение ссылок на элементы знаний;
- сравнение значений атрибутов элементов знаний.

В условиях левой части используются два вида переменных: переборные и присваиваемые. Для переборных переменных выполняется перебор возможных значений с означиванием переменных. Значения присваиваемых переменных устанавливаются явным образом, например, как результат функции или значение атрибута элемента знаний. Набор означенных переборных переменных и установленных присваиваемых переменных определяет контекст применения правила.

Все условия в левой части правила связаны конъюнкцией, соответственно, для выполнения условий правила должны выполниться все элементарные условия. После успешного означивания переборных переменных и выполнения всех элементарных условий происходит сохранение контекста применения правил для использования в правой части правила.

Правая часть правила представляет собой составной оператор простого императивного языка. В настоящее время реализованы следующие элементы:

- следование;
- составной оператор;
- условный оператор;
- присваивание значений переменным;

- вызов встроенных функций языка (создание, удаление, модификация, объединение элементов знаний).

Правила сгруппированы по фазам применения. Результаты применения правил и их побочные эффекты «незаметны» правилам, отнесенным к той же фазе — только правилам из последующих фаз. Если в результате ошибки записи правил или побочных эффектов других правил той же фазы выполнение всех действий правой части невозможно, происходит отмена эффекта частично выполненной правой части, после чего выполнение действий продолжается для других контекстов. Например, выполнение действий может быть невозможно, если элемент знаний, присутствующий в контексте одного правила, был удален в результате выполнения правой части другого.

Очевидно, что полный перебор всех возможных вариантов для означивания переменных в левой части правила неэффективен. Для повышения эффективности при поиске контекстов, в которых выполняется сопоставление, были разработаны алгоритмы предобработки правил трансформации и подготовки вспомогательных структур в ресурсе знаний, с которым будет идти работа. Цель предобработки — выделение минимально возможных множеств кандидатов для означивания переменных в правилах. Поскольку в пределах фазы (до начала выполнения действий) ресурс знаний не изменяется, можно однократно создать вспомогательные индексы и пользоваться ими при сопоставлении образцов всех правил, входящих в фазу. Индексы могут использоваться совместно всеми правилами фазы. С использованием индексов происходит исключение элементов знаний, которые заведомо не могут участвовать в означивании переменных.

После построения множеств кандидатов для означивания переменных происходит перебор кортежей декартова произведения множеств кандидатов для каждой переменной и окончательная проверка выполнения условий. Это необходимо, потому что не для всех условий возможна предобработка; кроме того, могут существовать зависимости между переборными переменными, которые можно определить только на этапе собственно сопоставления.

При успешном сопоставлении и означивании всех переборных переменных полный контекст сопоставления отправляется в хранилище результатов фазы. В дальнейшем это хранилище используется для выполнения действий каждого правила, входящего в фазу.

Правила трансформации позволяют унифицировать представление типовых ситуаций в хранилище фактографической информации и подготовить информацию для дальнейшей обработки, в том числе, для использования при анализе других текстов.

## 4 Результаты экспериментов

Чтобы читатель мог получить представление о качественных и технических характеристиках системы, рассмотрим задачу извлечения, которую мы решаем в настоящее время, и параметры системы, при которых эта задача решается.

Предметная область охватывает политику, межгосударственное взаимодействие и дипломатию; государственное и региональное управление; экономику, финансы, бизнес.

Целевые факты представляют собой события и состояния, участниками которых выступают целевые сущности.

К целевым сущностям относятся:

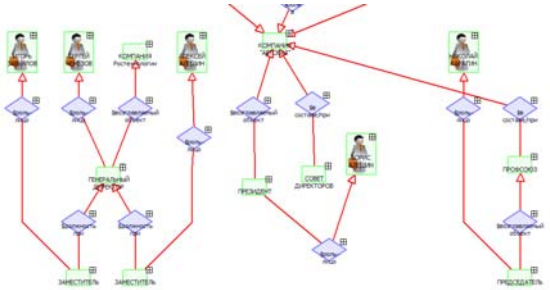
- лица;
- организации;
- роли лиц — должности, звания, род занятий, общие понятия принадлежности и иерархии (служащий, сотрудник, глава и т.п.), межличностные роли (родство, личное знакомство), членство и др.
- геополитические единицы.

Целевые факты описывают:

- отставки, назначения, пребывание в должности (роли);
- структурные отношения на множестве ролей лиц, организаций и геополитических единиц — должность при (лице, должности), должность в, должность во главе (организации или геополитической единицы), лицо во главе (организации или геополитической единицы), членство в организации, административно-территориальная принадлежность организации и др.;
- структурные отношения между организациями — в составе/при, часть-целое;
- отношения между лицами — родство (степень родства), знакомство и т.п.

Приведем характеристики системы на настоящий момент. В ресурсе знаний выделен 41 тип целевых и вспомогательных отношений (бинарных и тернарных) и 269 типов объектов предметной области (концептов), экземпляры которых могут стать участниками целевых ситуаций. Для решения задачи сейчас используется 156 контекстных правил извлечения информации, (42 фазы), а также 31 правило трансформации (6 фаз). Суммарная скорость обработки текста для такой конфигурации составляет порядка[2] 1 КБ/с на одном ядре процессора с тактовой частотой 2.4 ГГц.

Для примера на рисунке ниже приведен фрагмент результатов обработки новостной заметки [2].



## Заключение

Описанные в методы и подходы могут найти применение в технологических цепочках хранилищ знаний, для построения и наполнения ресурсов знаний разного рода, для повышения точности и обогащения результатов работы поисковых машин. Методы обработки текста и работы со знаниями, реализованные в системе ИСИДА-Т, создают основу для средств извлечения информации в «сильном» смысле. Такие средства не ограничиваются разметкой текста; они подразумевают переход от корпуса текстов к такому представлению фактографической информации, которое можно было бы использовать как интеллектуальный информационный ресурс, своего рода базу текстовых знаний.

## Литература


[1] Александровский Д.А., Кормалев Д.А., Куршев Е.П., Сулейманова Е.А., Трофимов И.В. Модель и реализация ресурса знаний в системе извлечения информации из текста // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2008, 28 сентября–3 октября 2008 г., г. Дубна, Россия): Труды конференции. Т. 2. — М.: ЛЕНАНД, 2008. — С. 201-209.

[2] Госкорпорация "Ростехнологии" и АВТОВАЗ создадут холдинг по производству автокомпонентов, 2008. http://quote.rbc.ru/stocks/news/2008/06/27/31997446.shtml

[3] Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в тексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. – Москва, Наука, 2004.

[4] Кормалев Д.А. Повышение производительности при распознавании текстовых ситуаций // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2008, 28 сентября–3 октября 2008 г., г. Дубна, Россия): Труды конференции. Т. 2. — М.: ЛЕНАНД, 2008. — С. 192-200.

[5] Кормалев Д. А., Куршев Е. П. Развитие языка правил извлечения информации в системе ИСИДА-Т // Труды международной конференции «Программные системы: теория и приложения». — Т. 2. — М.: Физматлит, 2006. — С. 365-377.

[6] Сулейманова Е.А. Классификация ресурсов знаний в системе извлечения информации из текста // Математические методы распознавания образов: 13-я Всероссиийская конференция. Ленинградская обл., г. Зеленогорск, 30 сентября - 6 октября 2007 г.: Сборник докладов. — М.: МАКС Пресс, 2007. — С. 625—628.

[7] Appelt D.E. The Common Pattern Specification Language: Technical report / SRI International, Artificial Intelligence Center. — 1996.

[8] Appelt D. E., Israel D. J. Introduction to Information Extraction. Tutorial // Sixteenth Int. Joint Conf. on Artificial Intelligence IJCAI'99, Stockholm, Sweden, 1999.

[9] Cooper K. D., Harvey T. J., Kennedy K. A Simple, Fast Dominance Algorithm. Software Practice and Experience, 2001.

[10] Cunningham H., Maynard D., Bontcheva K., Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[11] Ferrucci D., Lally A.UIMA by Example. IBM Systems Journal 43, No. 3,. 455-475 (2004).

[12] Grishman R. TIPSTER Text Architecture Design. Version 3.1. — New York: NYU, 1998.

[13] Lengauer T., Tarjan R.E. A fast algorithm for finding dominators in a flow graph. ACM Transactions on Programming Languages and Systems, 1(1):115120, July 1979.

[14] Plamondon L. The Montreal Transducer module for GATE. http://www.iro.umontreal.ca/~plamondl/mtltransducer/1_1/README.html


## Information extraction in ISIDA-T system


D.A. Kormalev, E.P. Kurshev, E.A. Suleimanova, I.V. Trofimov



The article discusses methods and techniques for information extraction from natural-language texts, as they are implemented in ISIDA-T system. Emphasis is made on knowledge representation and recognition of textual situations.



♣ Работа поддержана РФФИ, проект 09-07-00407, и программой фундаментальных исследований Президиума РАН №3, проект «Высокопроизводительные масштабируемые средства работы с фактографическими базами большого объема».


[1] Интеллектуальная система извлечения данных и их анализа (для обработки текстов).

[2] Скорость обработки текста и трансформаций сильно зависит от входного текста, а также объема накопленной фактографической информации.

# Использование методов извлечения информации при географической привязке текстов на русском языке

Прокофьев Петр Александрович

Компания «ЛАН-ПРОЕКТ»
p_prok@mail.ru

## Аннотация

В работе предлагается метод выделения фрагментов текста, позволяющих осуществлять географическую привязку текстов на русском языке. Настраиваются и оцениваются обучаемые методы MaxEnt, MEMM, CRF выделения объектов для разрешения неоднозначностей. Оценивается комбинированный метод выделения объектов. Методы сравниваются по показателям полноты и точности.

## Введение

Текст на естественном языке может содержать информацию о географических объектах, с которыми связан текст. Процедуру установления такой связи будем называть географической привязкой текста.

В более узком смысле под географической привязкой текста будем понимать установку связей между фрагментами текста и объектами географического справочника, содержащего информацию о названиях, типах и связях географических объектов. Таким справочником может быть электронная карта с именованными объектами, классификаторы по странам, регионам или адресам.

За рубежом задаче географической привязке текстов уделяется большое внимание. Периодически проводятся семинары: Geographic Information Retrieval (в рамках конференции Special Interest Group of Information Retrieval), GeoCLEF (в рамках конференции Cross-Language Evaluation Forum), Analysis of Geographic References (в рамках конференции North American Chapter of the Association for Computational Linguistics - Human Language Technologies). На этих семинарах представлен широкий спектр статей, описывающих методы и системы выделения географических объектов и разрешения неоднозначностей. Однако оценка этих методов и систем осуществлялась авторами преимущественно для англоязычных текстов.

В работе [10] описан обучаемый метод выделения именованных объектов, лежащий в основе метода разрешения неоднозначностей географической привязки, описанного в работе [4]. Разработчики системы MetaCarta в статье [8] описывают метод разрешения неоднозначностей на основе весов, назначаемых при анализе контекста.

Активно развиваются системы географического поиска и анализа географически привязанных текстов: MetaCarta [5], Google Maps [2]. Среди российских разработок можно выделить Yandex Карты [12]. Географическая привязка в этих системах используется на стадии индексации текстов для поиска по пространственным запросам.

Перед началом исследования на качественном уровне было оценено несколько систем выделения объектов в русскоязычных текстах [9], [11] и [14]. Оценка показала, что эти системы с географическими объектами работают на основе регулярных выражений и словарей названий. Эти методы часто не позволяют разрешать неоднозначности в текстах на естественном языке.

Ниже приведены неоднозначности, влияющие на географическую привязку текстов на русском языке, выявленные и классифицированные при анализе текстов новостных и энциклопедических статей.

1. Омонимия географических названий и имен нарицательных или их форм: «поселок Строитель», «город Чехов».
2. Географические названия включают имена или фамилии известных людей: город Энгельс Саратовской области.
3. Географические названия входят в состав названий организаций. Например: Администрация Воронежской области.
4. Синонимия различных названий:
«Россия» = «Российская федерация»,
«Париж» = «столица Франции»,
«Япония» = «родина японцев».
5. Исторические изменения географических названий.
6. Омонимия географических названий: Алтайский край и Алтайский район, город Железногорск Красноярского края и Железногорск Курской области.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

7. Контекстная зависимость относительной географической привязки: «50 километров от Москвы по Дмитровскому шоссе».

В работе описывается метод выделения и разрешения первых трех видов неоднозначностей. Для разрешения неоднозначностей используются обучаемые методы извлечения объектов и признаков в текстах.

Цель этого исследования – разработать метод выделения фрагментов русскоязычных текстов позволяющих осуществлять географическую привязки текстов.

Задачи исследования:

- разработать метод выделения фрагментов, не позволяющих однозначно осуществить географическую привязку;
- выбрать характеристики слов, используемые в обучаемых методах;
- сравнить качество разрешения неоднозначностей при использовании MaxEnt [10] [7], MEMM [6], CRF [3] и комбинированного методов с различными наборами характеристик;
- сделать вывод, какой метод лучше использовать для разрешения неоднозначностей при определении типа географического объекта, название которого употребляется в тексте.

## Метод выделения неоднозначностей

Географическая привязка текстов осуществляется путем поиска в тексте географических названий из справочника. Этапы выделения неоднозначностей привязки описаны ниже.

1. Перед поиском текст разбивается на лексемы, каждая из которых описывается морфологической и графематической информацией. Разбивка осуществляется с использованием морфологического модуля RML [13].

2. Каждой лексеме присваивается набор словарей, в которых найдена эта лексема. Используются словари имен, фамилий, отчеств, составленные по телефонному справочнику МГТС, родовые словари для географических объектов, «общий» словарь, составленный по набору художественных и научных произведений, а также словари географических названий, сгруппированные по типам объектов. Географические словари построены с использованием общероссийского классификатора стран мира (ОКСМ) и общероссийского классификатора объектов административно-территориального деления (ОКАТО). Используется 8 типов географических объектов:

- страны,
- федеральные округа,
- края,
- республики,
- автономные округа
- области,
- районы,
- города.

3. Для каждой лексемы проверяется выполнение предикатов, зависящих от лексемы и ее контекста. Далее такие предикаты будем называть запросами. Запросы формулируются на специальном языке, и их значения зависят от ранее вычисленных характеристик лексемы и лексем, входящих в ее контекст с учетом порядка. Запросы сгруппированы по типам географических объектов. Также выделены группы запросов для фамилий, имен и отчеств. Каждый запрос составлен для максимизации точности так, что если для лексемы выполнен только один запрос, то лексему можно однозначно классифицировать по типу географических объектов. В результате вычисления запросов, каждой лексеме присваивается набор групп, к которым относятся выполняющиеся для лексемы запросы. Примеры запросов приведены в таблице 1.

| | |
|---|---|
| Город | `город @%ГОРОД`<br>`г. @%ГОРОД`<br>`г @%ГОРОД`<br>`@%ГОРОД #2 столица %СТРАНА`<br>`столица %СТРАНА #2 @%ГОРОД`<br>`житель @%ГОРОД` |
| Страна | `житель @%СТРАНА`<br>`гражданин @%СТРАНА`<br>`гражданка @%СТРАНА`<br>`экономика @%СТРАНА`<br>`политика @%СТРАНА`<br>`бюджет @%СТРАНА`<br>`политика против @%СТРАНА`<br>`страна @%СТРАНА`<br>`граница @%СТРАНА` |
| Фамилия | `@%ФАМИЛИЯ %ИМЯ`<br>`%ИМЯ @%ФАМИЛИЯ`<br>`%ИМЯ %ОТЧЕСТВО @%ФАМИЛИЯ`<br>`*ИНИЦИАЛ . @%ФАМИЛИЯ`<br>`@%ФАМИЛИЯ *ИНИЦИАЛ .`<br>`господин @%ФАМИЛИЯ`<br>`госпожа @%ФАМИЛИЯ`<br>`%ЗВАНИЕ @%ФАМИЛИЯ`<br>`%ПРОФЕССИЯ @%ФАМИЛИЯ` |

**Таблица 1. Примеры запросов, взятые из разных групп. Синтаксис языка: @ – текущая лексема, % – словарь, * – регулярное выражение.**

4. Лексемы, для которых выполняются запросы из разных групп, или находящиеся в нескольких словарях, считаются неоднозначными и для их классификации используются обучаемые методы извлечения информации и метод максимального веса.

Таким образом, в тексте выделяются лексемы, которые согласно предварительной классификации по словарям и запросам относятся к нескольким классам. Эти лексемы будем называть «неоднозначностями». Их классификация осуществляется несколькими методами, описание и сравнение которых приведено ниже.

### Метод максимального веса

Запросам и словарям присваиваются веса. При классификации среди выполняемых запросов и содержащих слово словарей выбирается тот, у которого максимальный вес. Выбранный словарь или запрос определяют класс лексемы, исходя из группы, к которой принадлежит словарь или запрос.

## Обучаемые методы для разрешения неоднозначностей

Обучаемые методы, описанные ниже, решают одну и ту же задачу классификации лексем текста, поэтому можно ввести общую терминологию и обозначения.

Каждую лексему необходимо отнести к некоторому классу $c \in C, s = |C| < \infty$. Множество классов включает в себя набор типов объектов и служебные классы для обозначения «других» лексем (без типа), границ фрагментов и контекстных лексем (суффиксных и префиксных).

Перед классификацией для каждой лексемы $b$ вычисляется двоичный вектор длины $k$, координаты которого вычисляются, как значения характеристических функций $f_i(b), i \in \overline{1,k}$ и показывают, какие характеристики лексемы активны. Выбор характеристик влияет на результаты классификации, что будет показано ниже.

Обучение осуществляется по выборке $T = \left(\left(b^{(1)}, c^{(1)}\right), \left(b^{(2)}, c^{(2)}\right), \ldots, \left(b^{(h)}, c^{(h)}\right)\right) \in \left(B \times C\right)^h$, где $h$ - длина выборки, $B$ - множество лексем. Из обучающего набора вычисляется эмпирическое распределение $\tilde{p}(b,c) = \frac{n(b,c)}{h}$, где $n(b,c)$ – мощность $\left\{m \mid (b,c) = \left(b^{(m)}, c^{(m)}\right), m \in \overline{1,h}\right\}$.

При классификации выбирается класс, для которого условная вероятность $p(c \mid b)$ – наибольшая. Задача методов найти (оценить) распределение $p(c \mid b)$ по обучающей выборке.

### Maximum Entropy

Метод Maximum Entropy (MaxEnt) основан на принципе максимальной энтропии, и широко используется для извлечения информации и выделения именованных сущностей [10]. Метод строит вероятностную модель с явным заданием условной вероятности:

$$p(c \mid b) = \frac{1}{Z(b)} \prod_{i \in \overline{1,k}, j \in \overline{1,s}} u_{i,j}^{f_{i,j}(b,c)}, \qquad \textbf{(1)}$$

где $Z(b)$ – нормирующий множитель,

Множители $u_{i,j}$ вычисляются при решении оптимизационной задачи максимизации энтропии условного распределения модели:

$H(p) = -\sum_{c,b} \tilde{p}(b)\, p(c \mid b) \log\left(p(c \mid b)\right)$, то есть

$$p^* = \arg\max_{p \in P}\left(H(p)\right), \text{ где}$$

$P = \left\{p \mid M_p f_{i,j} = M_{\tilde{p}} f_{i,j}, i \in \overline{1,k}, j \in \overline{1,s}\right\}$ - ограничения на распределения, заданные через математические ожидания характеристик вычисленные по обучающей выборке.

Поиск решения можно осуществлять с помощью алгоритма Generalized Iterative Scaling (GIS-алгоритм) [1].

### Maximum Entropy Markov Model

Метод Maximum Entropy Markov Model (MEMM) [6] задает в явном виде распределение $p(c \mid b, c_{-1})$, где $c_{-1}$ – класс предыдущей лексемы. Распределение задается через совокупность распределений $\left\{p_{c_{-1}}(c \mid b), c_{-1} \in C\right\}$, каждое из которых описывается формулой (1) из метода MaxEnt. То есть модель можно также обучать с использованием GIS-алгоритма, но для каждого класса обучение проводится отдельно.

При классификации лексем последовательно определяют класс лексемы, в зависимости от результатов классификации предыдущей лексемы. В целом метод аналогичен MaxEnt, однако дает более точные результаты в задачах извлечения информации в текстах.

### Conditional Random Fields

Метод Conditional Random Fields (CRF) [3] аналогично MaxEnt определяет в явном виде $p(c \mid b)$, но кроме характеристик лексем используются также характеристики переходов классов.

$$p(c \mid b, c_{-1}) = \frac{1}{Z(b)} \prod_{i,j} u_{i,j}^{f_{i,j}(b,c)} \prod_{t=1}^{l} \eta_t^{g_t(b,c,c_{-1})}, \text{ где}$$

$g_t(b,c,c_{-1})$ – характеристика перехода, зависящая от текущей лексемы, а также классов предыдущей и текущей лексемы.

При тестировании метода использовались следующие характеристики переходом:

$$g_{i,j,r}(b,c,c') = \begin{cases} 1, \textit{если } f_i(b) = 1, \widetilde{n} = \widetilde{n}_j, c' = c_r \\ 0, \textit{иначе} \end{cases}$$

Аналогично MaxEnt для вычисления множителей использовался GIS-алгоритм.

### Комбинированный метод

В процессе обучения и классификации тестовой выборки осуществляется оценка результатов классификации. Оценка вычисляется с использованием показателей точности, полноты, F-меры, для каждого типа:

$$P_c = \frac{A}{A+B}; R_c = \frac{A}{A+C}; F_c = \frac{2P_cR_c}{P_c+R_c},$$

где $A$ – число лексем отнесенных к типу $c$, как в оцениваемом результате классификации, так и в тестовой выборке, $B$ – число лексем отнесенных к типу в оцениваемом результате классификации, но не отнесенных в тестовой выборке, $C$ – число

лексем отнесенных к типу в тестовой выборке, но не отнесенных в оцениваемом результате классификации. При подсчете показателей используются все лексемы (не только неоднозначные), отнесенные к какому-либо типу географических объектов либо в тестовой выборке, либо в оцениваемой выборке.

Комбинированный метод использует результаты классификации несколькими методами (MaxEnt, MEMM, CRF, максимального веса), а также оценки этих результатов. Если лексема отнесена к нескольким классам $\left\{c_{i_1}, c_{i_2}, ..., c_{i_d}\right\} =: C' \subset C$ по результатам разных классификаций, то ей присваивается класс $c_0 = \arg\max_{c \in C'} F_c$, выбранный методом, у которого показатель F-меры результатов классификации максимален для данного класса.

## Сравнение методов

Обучение и классификация осуществлялись на размеченном в полуавтоматическом режиме корпусе новостных сообщений по регионам России и странам мира. Каждый метод проверялся за 5 итераций, каждая из которых состояла из 2-х этапов: обучение на одной части корпуса и классификация оставшейся части корпуса. Для каждой итерации использовались различные деления корпуса на обучающий и тестовый наборы.

На каждой итерации производится вычисление интегральных показателей точности и полноты:

$$P^{(i)} = \frac{1}{|C|}\sum_{c \in C} P_c^{(i)}; R^{(i)} = \frac{1}{|C|}\sum_{c \in C} R_c^{(i)},$$

где $i$ – номер итерации. Для каждого метода выводились средние показатели, объединяющие результаты всех итерации тестирования метода.

### Обучающие выборки

Размеченный корпус состоит из 846 текстов, содержащих 372367 лексем (в то числе знаков пунктуации), из которых 5431 отмечены одним из 8 типов географических объектов.

При обучении на каждой итерации используется часть корпуса из 400 текстов, остальные 446 используются как тестовый набор.

### Характеристики слов и переходов

Тестирование осуществлялось с разными наборами характеристик. В качестве характеристик использовались функции, приведенные в таблице 2.

| | |
|---|---|
| 1) словарные | - принадлежность к словарю текущего, следующего или предыдущего слов;<br>- выполнение запросов в контекстах текущего, предыдущего или следующего слов;<br>- равенство предыдущего или следующего слова одному из часто употребляемых слов в контекстах помеченных в обучающей выборке слов. |
| 2) морфологические и графематические (вычисляются для текущего, предыдущего и следующего слова) | - графематические дескрипторы, например, «первая заглавная», «знак пунктуации»;<br>- морфологические дескрипторы, такие как падеж, число, род, часть речи. |

**Таблица 2. Характеристики, используемые при тестировании обучаемых методов.**

Первый набор характеристик содержит только словарные характеристики, второй – словарные, морфологические и графематические.

### Результаты сравнения

Отправной точкой для сравнения методов будем считать показатели качества метода максимального веса. Метод дал результаты: $P = 74,8\%; R = 83,4\%$. Показатели обучаемых методов приведены в таблице 3.

| | **Словарные (233)** | | **Сл. + морф. + граф. (285)** | |
|---|---|---|---|---|
| | ***P*** | ***R*** | ***P*** | ***R*** |
| **MaxEnt** | 87,9% | 87,7% | 86,6% | 84,5% |
| **MEMM** | 86,8% | 88,2% | 86,8% | 87,5% |
| **CRF** | 87% | 89,6% | 86,9% | 89,4% |
| **Комбин.** | *86,7%* | *88,2%* | *86,9%* | *87,5%* |

**Таблица 3. Показатели качества обучаемых методов при определении типов географических объектов. Рассмотрены 2 группы характеристик: только словарные и словарные с морфологическими и графематическими дескрипторами.**

Результаты сравнения показали, что комбинированный метод не оправдывает себя, поскольку дает показатели хуже, чем метод CRF.

Все методы дают показатели выше, чем метод максимального веса. Метод CRF показывает наилучшие результаты.

Кроме этого, использование морфологических и графематических дескрипторов ухудшает качество классификации. Возможно, это происходит по причине недостаточного объема размеченного корпуса.

Сравним методы, при использовании только словарных характеристик. При фильтрации характеристик по частоте встречаемости их в корпусе были получены наборы с различным числом характеристик. Результаты сравнения методов в зависимости от числа характеристик изображены на графике (см. рис. 1).

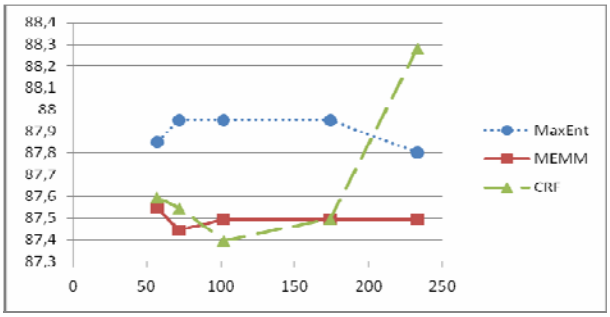



**Рисунок 1. Сравнение показателя F-меры качества классификации обучаемыми методами при использовании различного объема словарных характеристик.**

Стоит отметить, что метод MEMM не значительно зависит от числа характеристик в рассмотренном диапазоне. Уменьшение характеристик сильно сказывается на полноте в методе CRF. Метод MaxEnt имеет низкую полноту (ниже, чем у CRF и MEMM) и высокие показатели F-меры устанавливаются за счет показателя точности.

## Заключение

В работе представлен метод выявления неоднозначностей при поиске в текстах географических объектов и определения их типов. Для разрешения этих неоднозначностей использовались различные обучаемые методы, среди которых лучшим оказался метод CRF. Для метода CRF есть возможность подобрать характеристики переходов, что планируется сделать в будущих работах. Также в дальнейшем планируется провести анализ, какие характеристики лексем и переходов влияют на качество выполнения географической привязки.

При разрешении других указанных во вступлении неоднозначностей необходимы методы, учитывающие глобальный контекст и зависимости между фрагментами текста, такие методы предложены в работах [4] [8]. Описанные выше методы в том виде, в котором они используются в этой работе, не могут разрешить эти неоднозначности. В дальнейшем планируется использовать методы учета глобального контекста.

Работа показывает проблемы и возможные пути их разрешения при выполнении географической привязки текста. Использование географически привязанных текстов дает возможность совместно производить пространственный и тематический анализ и поиск данных, на что будут направлены дальнейшие исследования.

## Литература

[1] Darroch J. N. and Ratcliff D. Generalized Iterative Scaling for Log-Linear Models. The Annals of Mathematical Statistics, 43(5):1470-1480, 1972.

[2] Сервис «Google Maps». http://maps.google.ru.

[3] J. Lafferty, A. McCallum, F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML, 2001.

[4] Huifeng Li, Rohini K. Srihari, Cheng Niu, and Wei Li. InfoXtract location normalization: a hybrid approach to geographic references in information extraction, NAACL 2003 Workshop on the Analysis of Geographic References

[5] Solutions for government. MetaCarta, MC2006CB4-01, 2006. (www.metacarta.com).

[6] McCallum A., Freitag D., Pereira F. Maximum entropy Markov models for information extraction and segmentation. *Proc. ICML 2000* (pp. 591–598). Stanford, California.

[7] A. Ratnaparkhi. Maximum Entropy Models for Natural Language Ambiguity resolution. PHD thesis, Univ. of Pennsylvania, 1998.

[8] Erik Rauch, Michael Bukatin, Kenneth Baker. A confidence-based framework for disambiguating geographic terms. — 2003 — HLT-NAACL 2003 Workshop: Analysis of Geographic References.

[9] RCO Fact Extractor: персональная аналитическая система для поиска фактов в тексте. Компания RCO. www.rco.ru.

[10] Srihari, Rohini, Cheng Niu, and Wei Li. A Hybrid Approach for Named Entity and Sub-Type Tagging. In Proceedings of ANLP 2000, Seattle.

[11] Процессор SynSys Semantix. Компания «Синергетические системы». www.synsys.ru.

[12] Сервис «Яндекс – Карты». http://maps.yandex.ru.

[13] Автоматическая Обработка Текста. www.aot.ru.

[14] Описание технологии ИАС Арион. Компания «САЙТЕК». www.sytech.ru.

## Using the methods of information extraction in the geographic referencing of russian texts

Prokofjev Petr Alexandrovich

In this paper, a method of allocating text's fragments to geo-reference of russian texts. Set up and trained methods MaxEnt, MEMM, CRF and combined method for permits ambiguities. Methods compare of permission and recall

# Частичное обучение в логико-марковской сети в задаче извлечения временной информации из текста

© Фамхынг Д.К.

Волгоградский государственный технический университет
hungpdq@gmail.com

**Аннотация**

В работе описан метод использования логико-марковской сети для задачи извлечения временной информации в тексте на естественном языке. Предложенный метод является полным интегрированным решением, обеспечивающим строгие продукционные правила и правила с весом, указывающим уровень их доверия. Предложен новый алгоритм использования не аннотированных данных для повышения адекватности работы логико-марковской сети.

## 1 Введение

В связи с увеличением количества доступных электронных текстов требование к системам автоматической обработки и извлечения информации из документов на естественном языке для различных целей, таких, как добыча данных, значительно возросло. Например, нас, возможно, будет интересовать, когда и за сколько времени случилось некоторое событие. Эта задача связана с проблемой определения временных выражений в тексте, а также определения самого события. Например, дано следующее предложение:

*«С 1-ого января в Москве подорожает проезд в общественном транспорте».*

Исходя из этого предложения, мы можем определить, что «***подорожание»*** – это событие, и что оно случится с «***1-ого января»*.** После того как эта информация уже была извлечена, она может быть использована для создания более структурированной базы знаний, которую можно легко использовать в других системах обработки естественного языка (ОЕЯ), таких, как система поиска, система реферирования документов и вопросно-ответная система.

Задача извлечения временной информации из текста на естественном языке интенсивно исследуется в последние годы. Ее можно разделить на четыре подзадачи:

- Идентификация временных выражений;
- Идентификация событий в тексте;
- Связывание события с временным отчетом;
- Определение временных отношений между событиями в тексте.

Для решения этой задачи были проанализированы и используются различные подходы. Три основных подхода включают: традиционный подход, основанный на использовании правил, полученных от экспертов лингвистических знаний, подход, основанный на статистических методах, использующих какой-либо из алгоритмов машинного обучения, и гибридный подход. В работе [1] авторы поясняют преимущества и недостатки каждого из этих подходов. По сути, задача обработки естественного языка сложна из-за двусмысленности. Это вызывает трудности и ведет к неэффективности при применении статистических методов, так как большинство из них требует представить объект обучения в виде признаковых векторов. Аппарат логики, например, логика первого порядка или нечеткая логика, имеет широкие возможности для представления различных связей между явлениями в естественном языке, но он ограничен в способности обучения. Самые эффективные системы индуктивного обучения и вывода, такие как ALEPH [3], FOIL [4], Claudien, достигли не очень высокой адекватности.

В данной работе мы предлагаем новый метод для решения задачи извлечения временной информации с помощью аппарата логико-марковской сети (Markov logic networks), которая разработана в 2006 г. (ряд работ Домингоса и др. 2006-2008 гг.). Этот аппарат является вероятностным обобщением логики первого порядка, статистического обучения, позволяющего автоматически оценивать обоснованность выбранной модели явления и индуктивных правил, описывающих нестрогие зависимости между данными. Аппарат логико-марковской сети является самым удачным механизмом объединения мощности представления знания в традиционной двоичной логике и эффективности статистического обучения.

Статья состоит из четырех частей. Сначала мы описываем основные понятия марковской сети, логики первого порядка и логико-марковской сети. Далее будет подробно описано ее применение в

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

решении нашей задачи. Наконец, мы представляем меру близости для временных отношений и их интеграцию в Марковской сети.

## 2 Логико-марковская сеть (MLNs)

Логико-марковская сеть (Markov logic networks) комбинирует логику первого порядка и марковскую сеть. В обычной базе знания на основе правил продукции (логике первого порядка) типа «If X is E then Y is A», если хотя бы одно утверждение не выполняется, то полнота БЗ нарушается. MLNs можно рассматривать как обобщение логики первого порядка, при которой, когда одна формула не выполняется, то БЗ имеет меньшую вероятность существования.

### 2.1 Марковская сеть (Markov network)

Марковская сеть - неориентированная вероятностная графовая модель, использующаяся для представления совместных распределений набора нескольких случайных переменных *X*. Формально марковская сеть состоит из следующих компонентов:

- Неориентированный граф G = (V, E), где каждая вершина v □ V является случайной переменной в X и каждое ребро {u,v} □ E представляет собой зависимость между случайными величинами u и v.
- Набор потенциальных функций (potential function) $\varphi_k$ , одна для каждой клики в G. Функция $\varphi_k$ ставит каждому возможному состоянию элементов клики в соответствие некоторое неотрицательное действительное число.

Совместное распределение набора *X* в марковской сети вычисляется по формулам:

$$P(X = x) = \frac{1}{Z}\prod_k \phi_k(x_{\{k\}}),$$

где $x_{\{k\}}$ представляет собой состояние случайных переменных в k-ой клике и Z является коэффициентом нормализации

$$Z = \sum_{x \in X}\prod_k \phi_k(x_{\{k\}}).$$

### 2.2 Логико–марковская сеть (MLNs)

Логико-марковская сеть представляет собой множество пар $\{(F_i,w_i)\}$, где:

$F_i$ - формула логики первого порядка

$w_i$ - действительное число

$\{(F_i,w_i)\}$ вместе с набором констант C = $(c_1,c_2,\ldots,c_{|C|})$ используются как шаблон для создания марковской сети $M_{L,C}$ , содержащей:

- одну вершину для каждой возможной интерпретации (grounding) любого предиката в L. Данной вершине присваивается значение 1, если ее интерпретация верна, и 0 - в противном случае;
- один фактор для каждой интерпретации любой формулы $F_i$ в L с соответствующим весом $w_i$.

Для пояснения, как работает MLNs, мы построим примерную модель для задачи фильтрации спама. Здесь мы имеем три предиката, *H(l)* означает «заголовок письма *l* длиннее, чем его содержание», *S(l)* означает «*l* является спам-письмом» и *Ad(l,k)* означает «письма *l* и *k* приходят с одного адреса». Правила описаны в таблице 1.

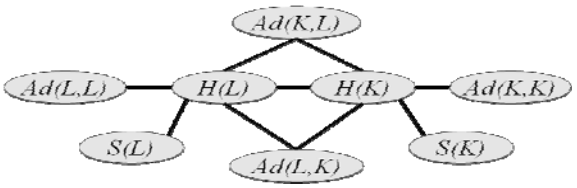



Рис 1. Марковская сеть, полученная из интерпретации формулы в таб. 1.

Распределение вероятностей возможного мира (БЗ) марковской сети $M_{L,C}$ определяется:

$$P(X = x) = \frac{1}{Z}\exp(\sum_i w_i n_i(x)) = \frac{1}{Z}\prod_i \phi_i(x_{\{i\}})^{n_i(x)}$$

где $n_i(x)$ является количеством верных интерпретаций (true grounding) формулы $F_i$ в $x$ , $x_{\{i\}}$ есть состояние (state) атомов, появляющихся в $F_i$ и $\phi_i(x_{\{i\}}) = e^{w_i}$ .

### 2.3 Вывод в MLNs

Допустим, что мы уже построили логико-марковскую сеть. Процесс вывода в ней часто требует и вероятностных, и детерминированных методов. Даны некоторые факты (evidences), требуется найти значения запрошенных предикатов для максимизации их частного распределения. Структура сети обычно бывает сложной, и в этом случае точный вывод является трудноразрешимой проблемой. Поэтому метод приблизительного рассуждения, такой как MCMC семплинг (Markov chain Monte Carlo sampling), станет хорошим выбором.

В связи с тем, что для решения нашей задачи мы используем и вероятностные правила, и строгие правила (deterministic dependencies), семплирование по Гиббсу ([2]), неэффективно обращается с детерминированными зависимостями, поэтому мы используем моделирование темперирования ([2]) для вывода.

Таблица 1- Спам-письмо

| Утверждение | Логика 1-го порядка | Форма высказываний | Вес |
|---|---|---|---|
| Если заголовок письма длиннее чем его содержание, то это спам | $\forall l \quad H(l) \Rightarrow S(l)$ | $\neg H(l) \vee S(l)$ | 1.1 |
| Письма, приходящие с одного адреса либо являются спамом, либо нет | $\forall l \forall k \ \mathrm{Ad}(l,k) \Rightarrow S(l) \Leftrightarrow S(k)$ | $\neg Ad(l,k) \vee S(h) \vee \neg S(l),$<br>$\neg Ad(l,k) \vee \neg S(h) \vee S(l)$ | 1.7<br>1.7 |

Таблица 2 – Предикаты для описания события

| Предикаты | Объяснение | Константы |
|---|---|---|
| Tense(e,!t) | Временная форма события e | {Прошедшее, Настоящее, Будущее} |
| Imperfect(e,!p) | Видовая форма | {Совершенный, Несовершенный} |
| Aspect (e,!a) | Аспектуальный класс | {Моментальное событие, Развивающееся событие, Процесс и др.} |
| Modal(e,!m) | Модальность | {Возможность, Вероятность, Обязанность} |
| Polarity(e,!p) | Полярность | {Положительное, Отрицательное} |
| VerbClass(e,!v) | Форма имени события | {Глагол (*в личной форме и инфинитив*), Краткое прилагательное, Отпредикатное имя, Причастие, Деепричастие} |

## 3 Модель извлечения временной информации в MLNs

В данной работе мы рассмотрим проблему определения временных отношений между событиями, описанными в источнике. Извлечение временных отношений заключается в определении взаимосвязей между событиями или между событиями и моментами времени [1]. В процессе вывода временных отношений возникает проблема, связанная с тем, что временная информация выражается в тексте на естественном языке явным или неявным образом. В любом случае остается неоспоримым тот факт, что события всегда связаны друг с другом или с моментом времени.

Как было отмечено в работе [1], на построение временного порядка влияет множество различных грамматических категорий, например, видо-временные формы глаголов, наречия времени, грамматические единицы (краткое прилагательное, отглагольное имя существительное). Кроме того, связь между событиями, отраженными в сложном предложении осуществляется с помощью союзов.

### 3.1 Предикаты

В нашей модели мы определяем различные предикаты для охвата этих явлений.

Для каждого события определены предикаты с разными грамматическими характеристиками: Tense(e,!t), Imperfect(e,!p), Aspect(e,!a), Class(e,!c), Modal(e,!m), Polarity(e,!p), VerbClass(e,!v),…

e – это событие,

знак «!» означает, что каждое событие e принадлежит только одному классу.

Мы также используем предикат HasVerb(e,+w) для указания, что событие e имеет слово w.

Для описания случая, когда специальные слова (союзы) присутствуют между событиями (мы считаем момент времени как специальное событие), мы приводим предикат HaveConnectedWord($e_1$,$e_2$,+c), где $e_1$, $e_2$ – события, с есть союзное слово, например, *перед тем как, после, во время, с тех пор, когда, пока, как, в то время как,…*

Два события, имеющие одинаковый объект, выражаются предикатом HaveSameObject($e_1$,$e_2$).

События, которые описываются в одном предложении, мы выражаем их предикатом InSameSentence($e_1$,$e_2$).

Главным предикатом запроса является InGroup($e_1$,$e_2$,g), означающий, что временное отношение между $e_1$,$e_2$ принадлежит группе *g*. Приняв классификацию временных отношений в работе [1], мы определяем девять групп временных отношений: {AFTER, BEFORE, DURING, INCLUDES, OVERLAPS, IS_OVERLAPPED, EQUALS, BEGIN, END}

В нашей работе мы рассмотрели два варианта описания предиката запроса: InGroup($e_1$,$e_2$,!g) и InGroup($e_1$,$e_2$,+g). Первый вариант означает, что временное отношение между $e_1$,$e_2$ принадлежит только одной из девяти возможных групп, в то время как второй разрешает мультигрупповой случай.

### 3.2 Формулы

Мы построили правила, чтобы описать нашу гипотезу о влиянии различных грамматических аспектов на временные отношения.

Для представления связи между видо-временной формой глаголов и временным отношением двух событий, мы предлагаем следующее правило:

Tense($e_1$,+$t_1$) ^ Tense($e_2$,+$t_2$) ^ Imperfect($e_1$,+$p_1$) ^ Imperfect($e_2$,+$p_2$) => InGroup($e_1$,$e_2$,g)

Для описания влияния аспектуального класса, модальности и полярности на временное отношение двух событий, мы определяем следующее правило:

Aspect ($e_1$,+$a_1$) ^ Aspect ($e_2$,+$a_2$) ^ Modal($e_1$,+$m_1$) ^ Modal($e_2$,+$m_2$) ^ Polarity($e_1$,+$p_1$) ^ Polarity ($e_2$,+$p_2$) => InGroup($e_1$,$e_2$,g)

Когда последовательность событий обозначается с помощью союзных слов, тогда с большой вероятностью можно сказать, что эти слова играют важную роль в построении временного порядка. Для этого случая мы используем простое правило:

HaveConnectedWord(e1,e2,+c) => InGroup($e_1$,$e_2$,g)

Причинно-следственное отношение часто приводит к тому, что одно событие является следствием другого события, и поэтому произошло после него. Например:

*«Свет выключился. Комната была темна».*

Для этого мы введем новый предикат Casual($e_1$,$e_2$), обозначающий, что событие $e_1$ является причиной $e_2$. Мы также добавим новое правило:

Casual($e_1$,$e_2$) => InGroup($e_1$,$e_2$, «BEFORE»)

Иногда сами имена событий, т.е. слова, обозначающие события, влияют на их временной порядок. Рассмотрим пример:

*«Я упал на пол. Кто-то меня толкнул»*

В этом примере между событиями нет никакого союзного слова. Но, исходя из знания о познаваемом мире, мы знаем, что падение всегда происходит после толчка, и сможем сделать правильное заключение об их временном отношении. Для описания этого появления мы определяем правило:

HasVerb($e_1$,+$w_1$) ^ HasVerb($e_2$,+$w_2$) => InGroup($e_1$,$e_2$,g)

**Детерминированные правила**

MLNs допускает строгие правила, и этим правилам будет присваиваться степень доверия (веса) с бесконечным значением. Это можно рассмотреть как продукционные правила, которые всегда выполняются. Мы сформулировали 5 групп экспертных правил для разных случаев проявления событий.

**Транзитивные правила**

В других работах, опубликованных ранее, транзитивные правила используются для повышения количества найденных временных отношений. Но во всех этих работах это происходит как пост-процесс только после процесса классификации [6]. Это приводит к несогласованности в базе знания. В работе [7] авторы предложили метод, гарантирующий глобальную согласованность в базе знания с помощью линейного целочисленного программирования, но ограничились только двумя возможными отношениями. С помощью MLNs мы интегрируем эти правила в процессе вывода, следовательно, гарантируются согласованность и эффективность при выводе.

Некоторые транзитные правила являются строгими, некоторые выполняются с большой вероятностью, поэтому имеет вес с большим значением, а другие являются «мягкими» правилами, например:

InGroup($e_1$,$e_2$, BEFORE) ^ InGroup($e_2$,$e_3$, BEFORE) => InGroup($e_1$,$e_3$, BEFORE)

InGroup($e_1$,$e_2$, BEFORE) ^ InGroup($e_2$,$e_3$, OVERLAPS) => InGroup($e_1$,$e_3$, BEFORE)

InGroup($e_1$,$e_2$, BEFORE) ^ InGroup($e_2$,$e_3$, DURING) => InGroup($e_1$,$e_3$, OVERLAPS)

## 4 Интеграция частичного обучения в MLNs

В задаче обработки естественного языка часто имеет место такая ситуация, что неаннотированных данных достаточно много и они легко собираются. Аннотирование этих данных очень дорого и требует колоссальных усилий. В данной работе мы предложили алгоритм, который использует неаннотированные данные для повышения эффективности работы MLNs.

Алгоритм основан на простой идее, что два близких объекта будут находиться в одной группе.

**Кластеризация**

На основе меры сходства мы используем метод кластеризации (k-NN и KNN-SVM [5]) для обучения корпуса, комбинирующего аннотированные и неаннотированные данные. За счет неаннотированного корпуса плотность данных повышается, и аккуратность кластеризации увеличивается. Обучаемые выборки, следовательно, разделятся на кластеры.

Применим этот алгоритм в решении нашей задачи. После кластеризации мы добавляем в базу знаний нашей MLNs новый предикат Label($e_1$,$e_2$,$e_3$,$e_4$), указывающий на то, что временное отношение между парами событий ($e_1$,$e_2$) и ($e_3$,$e_4$) принадлежит одному кластеру.

**Интеграция частичного обучения в MLNs**

Обучение и вывод в MLNs реализуются посредством максимизации псевдо-функции правдоподобия (pseudo-likelihood function), являющейся маргинальной вероятностью запрошенных предикатов при некоторых заданных свидетельствах. Интеграция частичного обучения в MLNs поэтому затруднена. Для этого мы добавляем правила:

Label($e_1$,$e_2$,$e_3$,$e_4$) ^ InGroup($e_1$,$e_2$,+g) => InGroup($e_3$,$e_4$,+g).

## 5 Заключение

В работе предложен новый подход к решению задачи извлечения временной информации с помощью аппарата логико-марковской сети. Интеграция априорных знаний всегда является большой проблемой в обработке естественного языка. Логико-марковская сеть дает естественный способ решения с поддержкой строгих правил и «мягких» правил, за счет которых смягчает условие согласованности базы знаний. Однако в логико-марковской сети никак нельзя влиять на процесс выбора обучаемой выборки, потому что вес для всех означенных формул (grounding formula) одной формулы зафиксирован. Предложенный в данной

работе метод использования меры близости дает возможность регулировать целевую функцию и поэтому влиять на определение значимости обучаемой выборки.

## Литература


[1] Заболеева-Зотова А.В., Фамхынг Д.К., Захаров С.С. Гибридный подход к обработке временной информации в тексте на русском языке // Труды одиннадцатой национальной конференции по искусственному интеллекту с международным участием - **КИИ** 2008.

[2] Richardson, M., Domingos, P. Markov Logic Networks. Machine Learning, volume 62 , pages 107–136, 2006

[3] Srinivasan, A. (2001). The Aleph manual. http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/.

[4] Landwehr, N., Kersting, K., & Raedt, L. D. Integrating Naive Bayes and FOIL. Journal of Machine Learning Research, volume 8, pages 481–507, 2007.

[5] Hao Zhang. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, Vol. 2, P. 2126-2136.

[6] Temporal Awareness and Reasoning Systems http://www.timeml.org/site/tarsqi/index.html

[7] Bramsen, P. and Deshpande, P. Inducing Temporal Graphs. In Proceedings of EMNLP-06, 2007.


## Semi-supervised learning with Markov Logic Networks and application to temporal information processing


Hung Pham D.Q.



This paper addressed the problem of temporal information extraction from text. We proposed a solution using Markov Logic Networks. The knowledge is presented in first-order logic production rules. We used both hard clauses and clauses that associated with weights to indicate their confidence. A new co-training algorithm was proposed that take benefit of unlabeled data to boost accuracy of Markov logic networks.

# ЛЕКСИКА, ПЕРЕВОД, СИНТАКСИЧЕСКАЯ РАЗМЕТКА ТЕКСТОВ

# LEXICOLOGY, TRANSLATION, SYNTACTIC MARKUP OF TEXTS

# Устранение лексической многозначности терминов Википедии на основе скрытой модели Маркова

© Турдаков Денис

ВМК МГУ, ИСП РАН
turdakov@gmail.com

**Аннотация**

В статье описывается способ автоматического устранения лексической многозначности терминов естественного языка, использующий открытую энциклопедию Википедию. Рассматриваются проблемы применения существующих алгоритмов, и предлагается собственный метод, основанный на скрытой модели Маркова, параметры которой вычисляются на основе словаря и ссылочной структуры Википедии. Также, предлагается эвристика для ускорения описанного алгоритма, и приводятся экспериментальные оценки точности на различных тестовых корпусах.

## 1 Введение

Задача устранения лексической многозначности (word sense disambiguation) возникла в 50-х годах прошлого века, в качестве подзадачи машинного перевода. С тех пор, исследователи предложили огромное количество методов ее решения, однако она остается более чем актуальной и по сей день. В общем случае, задача включает в себя два аспекта:

1) фиксирование всех различных значений для каждого слова, относящегося к тексту;
2) определение способа выбора подходящего значения для каждого экземпляра слова.

Рассмотрим каждый из них подробнее. Существует два основных подхода к определению списка значений слов. Большинство работ опираются на предопределенные значения: списки слов, найденные в словарях, переводы на иностранные языки, и т. п. Вторым подходом является анализ способов употребления слов в различных источниках и выделение значений на основе этого анализа. Однако до сих пор ведутся споры о том, что является значением слова. Кроме того, часто необходимо определить значение не отдельно стоящего слова, а группы слов, образующих термин. Тогда задача осложняется еще и поиском терминов и определением, какие значения могут соответствовать каждому термину.

Алгоритмы для выбора подходящего значения используют два источника информации: контекст слова - информацию, которая содержится в тексте, в котором слово встретилось; и внешние источники, такие как словари и базы знаний. Современные методы можно разделить на два класса: методы, основанные на обучении по размеченным корпусам, и методы, основанные на внешних источниках знаний (тезаурусы, машинно-ориентированные словари, лексиконы). Хороший обзор алгоритмов можно найти в [1, 7], также краткий обзор алгоритмов, использующих Википедию, будет дан во второй части статьи.

Еще одной важной проблемой является оценка методов и их сравнение. Так как снятие многозначности обычно используется для улучшения работы большей системы, существует два способа оценки: in vitro - на сколько хорошо работают методы сами по себе - и in vivo - как снятие многозначности улучшает работу системы в целом. Для оценивания самих методов обычно используют два коэффициента: точность и полноту. **Точность** - это число слов, размеченных правильно, по отношению к числу слов, обработанных системой. **Полнота** - число слов, размеченных правильно, по отношению к числу слов в тестовом множестве. Также часто вводят **F-меру**, значением которой является среднее гармоническое между полнотой и точностью. Для сравнения методов снятия многозначности английских слов были разработаны тестовые наборы и проводятся конференции Senseval-1,2,3 и Semeval[20]. Эти тестовые наборы используют заранее определенные значения многозначных слов, которые берутся из словаря WordNet [15], это накладывает ограничение на возможность их использования. Так, методы использующие словарь Википедии, нельзя напрямую сравнить с методами, использующими словарь WordNet, так как количество значений слов в Википедии намного превосходит аналогичное число в WordNet. В работе [13] авторы смогли отобразить используемые значения на словарь WordNet, однако в дальнейшем [14] отказались от такой процедуры. Это связано с тем, что Википедия растет и изменяется очень

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

быстро, а ее словарь на порядок превосходит словарь WordNet.

### 1.1 Скрытая модель Маркова и задача устранения лексической многозначности

Проблема снятия лексической многозначности может быть переформулирована как задача максимизации с использованием формализма скрытых Марковских моделей.

Пусть $\mathrm{T}$ - множество терминов, $\mathrm{M}$ - множество значений, соответствующих терминам. Для последовательности терминов $\tau = t_1,...,t_n$, где $t_i \in \mathrm{T}, \forall i$, задача максимизации состоит в нахождении наиболее вероятной последовательности значений $\mu = m_1,...,m_n$, где $m_i \in \mathrm{M}, \forall i$, соответствующей входным терминам и согласованной с ограничениями модели:

$$\hat{\mu} = \arg\max_{\mu} P(\mu \mid \tau) = \\ = \arg\max_{\mu}\left(\frac{P(\mu)P(\tau \mid \mu)}{P(\tau)}\right) \quad (1)$$

Так как вероятность $P(\tau)$ постоянна для фиксированной входной последовательности, задача редуцируется к максимизации числителя равенства 1. Для решения этого уравнения делается Марковское предположение, что значение *i*-го термина зависит только от конечного числа значений предыдущих терминов:

$$\hat{\mu} = \arg\max_{\mu}\left(\prod_{i=1}^{n} P(m_i \mid m_{i-1},...,m_{i-k}) \cdot P(t_i \mid m_i)\right) \quad (2)$$

где $k$ – порядок модели.

Множители уравнения 2 определяют скрытую Марковскую модель *k*-го порядка, где наблюдения соответствуют входным терминам, состояния соответствуют значениям терминов, $P(m_i \mid m_{i-1},...,m_{i-k})$ – модель перехода между состояниями и $P(t_i \mid m_i)$ – модель наблюдения, описывающая вероятность появления термина $t_i$ в каждом состоянии $m_i$.

Несмотря на то, что рассматриваемую задачу нетрудно формализовать с помощью скрытой Марковской модели, дальнейшее использование этого формализма связано с проблемой разреженности языка. Так, чтобы построить модель перехода для Марковской модели первого порядка, необходимо оценить вероятность каждой пары состояний, что для задачи устранения лексической многозначности является вероятностью того, что два термина в конкретных значениях встретились вместе. Если для задачи определения частей речи слов, параметры Марковской модели можно оценить на основе сравнительно небольшого размеченного корпуса, то для задачи устранения лексической многозначности проблема оценки параметров сильно усложняется. Это связано с количеством значений. Так, Википедия содержит более двух миллионов концепций, кроме того задача осложняется тем, что частота употребления терминов в тексте распределена не по равномерному закону, а по закону Зипфа (Zipf law). Учитывая эти факты, несложно заметить, что для обучения Марковской модели потребуется размеченный корпус огромного размера. Ниже мы предложим способ оценки модели перехода для поставленной задачи с помощью семантической близости концепций Википедии, вычисленной на основе графа ссылок.

### 1.2 Википедия

Википедия - это открытая энциклопедия, создающаяся пользователями Веба. Сейчас англоязычная Википедия содержит более 2.5 млн. статей, не считая специальных страниц. Граф Википедии (вершины графа - это страницы энциклопедии, а ребра - гипертекстовые ссылки между ними) обладает малым диаметром и высоким коэффициентом кластеризации[12], что структурно отличает ее от тезауруса WordNet. Кроме того, Википедия содержит огромное количество дополнительной информации, которая используется для различных исследований. Рассмотрим структуру открытой энциклопедии более подробно.

Большинство статей, которые видят пользователи, имеют заголовок, обозначающий **концепцию**, описанную в теле статьи. В теле статьи пользователи ставят ссылки на концепции связанные с данной. Структурно ссылки состоят из двух основных частей: концепции энциклопедии, на которую указывает ссылка, и **термина**, который видит пользователь. Кроме обычных ссылок, бывают специальные, например, "Смотри также", которые указывают на концепции, сильно связанные с данной.

Кроме обычных статей, существует еще несколько типов страниц. Страницы устранения многозначности содержат списки значений многозначных слов. Эти страницы создаются, как и вся Википедия, вручную пользователями и поэтому содержат не все возможные варианты. При этом, среднее количество вариантов намного превосходит аналогичное количество в WordNet (табл. 1).

| | **Википедия Октябрь'08** | **Википедия Март '09** | **WordNet** |
|---|---|---|---|
| Концепций | 2 500 000 | 2 800 000 | 150 000 |
| Многозначных терминов | 260 000 | 290 000 | 26 000 |
| Среднее кол-во значений | 5,4 | 5,47 | 2,95 |

**Таблица 1: Статистика Википедии и WordNet**

Каждая статья Википедии принадлежит одной или нескольким категориям. Сами категории также

могут принадлежать более общим категориям. При этом они не образует таксономию, так как граф категорий может содержать циклы.

Еще одним важным типом статей являются редиректы. С этих страниц происходит автоматическое перенаправление на обычные статьи. По сути, эти статьи содержат синонимы концепций. Кроме того, Википедия содержит еще много дополнительных структур (шаблоны, инфобоксы, списки, и т.д.).

В следующем разделе будет дан обзор наиболее близких методов устранения многозначности. Далее будет описан разработанный метод и приведены оценки его работы.

## 2 Обзор существующих работ

Условно можно выделить три этапа развития методов устранения лексической многозначности. С 50-х по 80-е года были разработаны основные подходы, однако из-за отсутствия хороших машинных словарей и баз знаний, в этот период созданы только "игрушечные" системы, покрывающие только некоторые слова языка. Следующий этап, пик которого пришелся на 90-е годы, был обусловлен вручную созданными крупномасштабными базами знаний, такими как WordNet и CyC [4] и сбалансированными корпусами документов(Brown [5], Penn TreeBank [21]). Алгоритмы, разработанные в этот период, использовали структуру баз знаний или обучались на общепризнанных корпусах. Исследователи получили хорошие результаты, однако сложность ручного создания и поддержки в актуальном состоянии больших структур ограничила область применения этих алгоритмов. В начале 21 века исследователей в области обработки естественного языка заинтересовала возможность использования сетей документов, таких как WWW и Wikipedia, связанных гиперссылками, и созданных огромным числом независимых пользователей. Ниже приводится краткий обзор работ, в которых для решения задачи снятия лексической многозначности использовалась модель Маркова или Википедия.

Для оценки нижней границы точности методов устранения лексической многозначности обычно используют наивный метод, который всем словам присваивает их наиболее частое значение (baseline algorithm). Заметим, что этот метод эквивалентен Марковской модели нулевого порядка, в которой все термины равновероятны для любой фиксированной концепции.

Для методов, основанных на внешних источниках, в качестве нижней границы часто используют результат работы алгоритма Леска. Алгоритм Леска - это классический алгоритм автоматического снятия многозначности, введенный М. Леском в 1986 году [8]. Алгоритм основан на предположении, что многозначное слово и его окружение относятся к одной теме. Простая реализация алгоритма Леска состоит из трех шагов:

- Выбрать многозначные слова и их контексты
- Взять их определение в некотором словаре
- В качестве значения многозначного термина выбрать то, которое максимизирует количество общих слов в словарном определении данного значения и определений терминов контекста.

В работе [10] для обучения модели Маркова использовался корпус SemCor, и было показано, что его недостаточно, чтобы обучить даже модель первого порядка. Чтобы устранить проблему разреженности языка, авторы предложили использовать категории WordNet для построения модели перехода в Марковской цепи. Это позволило немного повысить точность алгоритма, в сравнении с наивным методом.

Аналогичный результат был получен и в более поздних работах [18, 19], где использовалась специализированная модель Маркова, с состояниями, хранящими части речи слов. Авторы этой работы показали, что без дополнительной семантической информации из внешних источников, корпуса SemCor недостаточно для нахождения параметров марковской модели для задачи устранения лексической многозначности слов, и наилучшие результаты показывает наивный метод.

В работах, использующих Википедию, продолжают просматриваться парадигмы, заложенные в 90-х годах прошлого века. Условно их можно охарактеризовать как методы, основанные на статистическом подходе, и методы, основанные на внешних данных.

В работе [13] Википедия использовалась как аннотированный корпус для обучения Наивного Байесовского классификатора. Многозначные термины и их значения выделялись из ссылок ([[статья-значение|многозначный термин]]). Для каждого многозначного термина, представленного в Википедии, строился вектор признаков, составленный из

- части речи многозначного термина
- локального контекста из трех слов слева и справа многозначного термина с их частями речи
- глобального контекста из пяти самых частых специфичных для значения слов, встречающихся во всевозможных контекстах термина.

Авторы вручную отобразили часть терминов Википедии в термины WordNet, чтобы провести эксперименты на общепринятых эталонных тестах корпуса SENSEVAL. Эксперименты показали, что Наивный Байесовский классификатор, обученный на Википедии, показывает лучшие результаты, чем алгоритм Леска и baseline algorithm. Кроме того было показано, что распределение терминов в

статьях Википедии сильно отличается от тренировочного множества корпуса SENSEVAL - вручную сбалансированного британского национального корпуса. Также эксперименты подтвердили предположение, что с ростом Википедии улучшается точность снятия многозначности.

Исследование, описанное в [13], получило развитие в системе автоматического обогащения документов ссылками на статьи Википедии [14]. В статье, посвященной этой системе, описываются методы для автоматического извлечения ключевых слов и снятия лексической многозначности на основе Википедии. Для выделения ключевых фраз использовался словарь, состоящий из заголовков Википедии, расширенный морфологическими формами, которые встречаются во внутренних ссылках Википедии на соответствующие статьи не менее 5 раз. Выделение ключевых фраз происходит в два этапа:

- поиск кандидатов,
- ранжирование ключевых фраз.

На первом этапе выделяются всевозможные N-граммы, присутствующие в словаре, после этого на втором этапе выделенным терминам присваиваются веса на основе одного из трех методов: tf-idf, хи-квадрат, и информативность (вероятность использования термина в качестве текста ссылки).

На основании результатов проведенных экспериментов утверждается, что последний метод показывает наилучшие результаты. Во второй части работы приводится алгоритм снятия многозначности в выделенных ключевых фразах. Авторы использовали комбинацию алгоритма Леска [8] и статистического алгоритма, обученного на Википедии. В алгоритме Леска в качестве словарного определения термина бралась соответствующая статья Википедии, а в качестве контекста абзац, в котором встретился термин. В качестве статистического алгоритма использовался классификатор[13], описанный выше. Наконец, предполагая ортогональность этих методов, авторы использовали расхождения в результатах, как признак потенциальной ошибки и игнорировали такой результат. Оценка получившейся системы производилась на тестовом наборе из 85 случайно выбранных статей Википедии, специально размеченных вручную. Эксперименты показали, что статистический метод показывает большую точность (92.91%) и полноту (83.1%), чем алгоритм Леска (80.1% и 71.86% соответственно), а комбинирование алгоритмов увеличивает точность (94.33), но снижает полноту (70.51%).

Работа [3] является развитием алгоритма Леска, с учетом дополнительной информации, которую можно извлечь из Википедии. Для создания словаря и поиска возможных значений терминов использовались заголовки статей, редиректы, страницы значений многозначных слов и текст ссылок. Для каждого термина словаря собирался вектор признаков, состоящий из тэга категории, контекста (слова или термина встречающегося вместе с данным термином) и класса термина (Человек, Место, Организация, Остальное). В качестве значения многозначного термина выбирался кандидат, максимально похожий на контекст, где похожесть вычислялась как косинус между векторами признаков.

В работе [2] также рассматривается использование векторной модели для снятия многозначности имен собственных, однако, в отличии от [3], практически не уделяется внимания нахождению различных признаков, а, в дополнение к векторной модели, предлагается использовать иерархию категорий Википедии для обучения линейного классификатора, основанного на методе опорных векторов (Support Vector Machine, SVM).

В работе [11] для снятия лексической многозначности используется семантическая близость терминов [16]. Расстояние между терминами вычисляется на основе графа ссылок Википедии. Для каждого возможного значения термина вычисляется близость до терминов контекста и выбирается наиболее близкое. Авторы показали, что их алгоритм работает лучше, чем алгоритм Леска. Для оценки качества алгоритма использовалось тестовое множество, созданное на основе аннотаций ссылок из статей Википедии.

Для снятия многозначности авторы [17] использовали подход, основанный на машинном обучении в комбинации с семантической похожестью, описанной в [16]. В качестве тренировочного множества использовалось 500 случайных статей Википедии. Положительными примерами устранения многозначности служили термины, на которые указывали ссылки, а остальные возможные значения, как и в [11] полученные из ссылок Википедии, служили отрицательными примерами. В качестве признаков использовались вероятность значения многозначного термина, полученная просмотром всех ссылок Википедии, и расстояние до терминов контекста. Для вычисления расстояния до контекста использовалась та же мера, что и в [11]. Но, кроме того, терминам контекста придавался вес, посчитанный как среднее между информативностью данного термина и близостью термина к центральной нити документа, определенной как средняя близость между текущим термином и остальными терминами контекста. Еще одним признаком послужило качество контекста, определенное как сумма весов терминов, посчитанных на предыдущем шаге. Основываясь на данных признаках, авторы провели сравнительное тестирование нескольких алгоритмов машинного обучения (Naive Bayes, SVM, C4.5 с вариациями) на тестовом множестве из 100 случайных статей Википедии, и показали, что данный подход дает лучшие результаты (97.1%), чем описанный в [11].

В работе [22] так же, как и в [11] и [17] используется мера семантической близости терминов. Однако, для ее вычисления авторы

предложили использовать информацию о типах ссылок и давать им разный вес. Различные меры близости на взвешенном графе сравниваются между собой на примере задачи снятия лексической многозначности, и показывается, что коэффициенты Дайса и Жаккара дают наилучшие результаты. Мы используем аналогичный метод вычисления близости терминов для оценки параметров модели.

## 3 Снятие многозначности

Методы, предложенные в вышеперечисленных работах, основанные на внешних знаниях, имеют один общий недостаток. Они неявно используют предположение, что в тексте существуют однозначные термины, на основании значений которых впоследствии определяются значения многозначных терминов. Однако, было замечено, что с ростом Википедии словарь многозначных терминов увеличивается, причем дополнительные значения появляются у наиболее употребляемых терминов. Это приводит к тому, что в неспецифических сообщениях, таких как новостные статьи, все термины имеют более одного значения, либо встречающиеся однозначные термины мало связаны с основной темой документа. Это ухудшает точность алгоритмов, основанных на однозначном контексте, и ведет к необходимости разработки метода, лишенного такого недостатка.

Далее будет приведено описание такого метода, основанного на скрытой модели Маркова. Вероятности модели перехода мы оцениваем с помощью семантической близости концепций, способ подсчета которой описан в следующем подразделе. Вероятности модели наблюдений и априорная вероятность значений оценивается с помощью эмпирического распределения значений и терминов во внутренних ссылках Википедии. Кроме того ссылки, совместно со словарем Википедии и специальными статьями используются для создания словарей терминов и их значений (раздел 3.2). Далее приводится алгоритм и результаты экспериментов на различных корпусах.

### 3.1 Семантическая близость концепций

Ранее [22], мы разработали простую меру близости между статьями Википедии, которая может быть полезна для различных задач, в том числе для снятия многозначности. Меры близости между вершинами графа, описанные в литературе, можно разделить на два широких класса: меры, основанные на локальной информации, такие как косинус угла между векторами, коэффициенты Дайса и Жаккара, соцетирование и т. п., и меры, основанные на распространяющейся активации (spreading activation), например, SimRank [6]. В то время как меры, относящиеся ко второму классу, показывают более качественные результаты, их вычислительная сложность ($O(n^3)$ для SimRank [9]) не позволяет использовать их для работы с большими объемами данных. Кроме того, вычисление семантической близости между двумя вершинами требует построения полной матрицы близости для всех вершин графа. Поэтому в нашей работе мы используем меру, основанную на коэффициенте Дайса, часто используемую в системах информационного поиска.

Для двух статей Википедии мера Дайса определяется как удвоенное отношение числа их общих соседей к общему числу всех соседей обоих статей. Формально

$$sim(A,B) = Dice(A,B) = \frac{2\times|n(A)\cap n(B)|}{|n(A)|+|n(B)|}, \quad \textbf{(3)}$$

где $n(X)$ – множество статей, связанных ссылкой со статьей *X*.

Мы исследовали структуру Википедии и заметили, что некоторые типы ссылок чрезвычайно релевантны по отношению к семантической близости, в то время как другие могут привести к неверным результатам. Поэтому, в дополнение к основной мере, мы ввели схему весов, основанную на типах ссылок. Подробное описание способа вычисления семантической близости здесь не приводится, так как не относится к основной теме статьи, и его можно найти в работе [22].

### 3.2 Создание словарей

На данный момент Википедия содержит статьи, описывающие более 2.5 миллионов концепций. Мы используем названия статей для создания словарей, которые используется для поиска терминов в текстах. После того, как все термины в тексте найдены, они представляются как последовательность наблюдений, а их значения - как соответствующие состояния в скрытой модели Маркова.

Для формирования словаря терминов мы берем названия всех статей, описывающих соответствующие концепции и названия всех страниц переадресации на эти статьи.

Основным источником значений терминов является категория "Disambiguation pages". Статьи, входящие в эту категорию, содержат списки возможных значений и ссылки на страницы, описывающие эти значения. Однако в Википедии не существует четких правил для создания таких страниц, поэтому часто они содержат много лишних ссылок, напрямую не связанных с многозначной концепцией. Поэтому при обработке этих страниц мы выделяем только те значения, которые содержат в своем названии словоформы многозначной концепции, или для которых она является акронимом. Эта эвристика очень жесткая и отсеивает много хороших значений (мы добавляем их позднее при анализе ссылок). Например, для термина "*NATO*" будет найдено значение "***N**orth **A**tlantic **T**reaty **O**rganisation*" но пропущено значение "*Mora (plant)*" – растение, которое часто называют "нато".

Система заранее не имеет информации, какие тексты придется анализировать, следовательно, не должна быть чувствительна к регистру слов. Поэтому все слова в словаре мы приводим к верхнему регистру. Википедия, напротив, чувствительна к регистру, и одинаковые термины, написанные в разном регистре, могут указывать на различные концепции. Для решения этой проблемы, мы добавляем такие концепции в словарь значений терминов и выбираем нужной значение на этапе устранения многозначности терминов текста.

Кроме того, большое количество терминов содержит в названии уточняющие концепции, например *"Platform (computing)"*. Мы убираем текст в скобках и в случае коллизий применяем подход, аналогичный тому, который используется при приведении к одному регистру.

Во введении упоминалось, что источником значений многозначных терминов могут быть как словари, так и способы употребления терминов в текстах. В нашем случае дополнительным источником значений являются ссылки между страницами. Любая ссылка содержит две части: текст, который видит пользователь, и концепцию Википедии, на которую в действительности ведет ссылка. Мы анализируем употребление всех терминов из созданного словаря в качестве текста ссылок и добавляем в список значений этих терминов концепции, на которые указывали данные ссылки. На этом заканчивается формирование словаря значений терминов.

В итоге словарь терминов содержит более 5 500 000 терминов, соответствующих 2 500 000 концепций, из них многозначных терминов – 260 000 элементов, а среднее количество значений многозначных терминов равно 5,4.

### 3.3 Описание алгоритма

Во введении было показано, как использовать формализм скрытых моделей Маркова для решения задачи устранения лексической многозначности. Основная сложность использования этого формализма состоит в оценке параметров модели. Воспользуемся информацией Википедии, чтобы решить эту проблему.

Для оценки модели наблюдения воспользуемся ссылками Википедии. На основании способа построения словарей можно заметить, что термины, соответствующие синонимам концепции, могут появиться только из заголовка статьи, описывающей концепцию, названий редиректов на концепцию и терминов, совпадающих с текстом ссылок на концепцию. Исходя из этого, определим условную вероятность термина $t_i^j$, соответствующего значению $m_i$ через эмпирическую вероятность $\hat{P}(t_i^j \mid m_i)$:

$$P(t_i^j \mid m_i) = \hat{P}(t_i^j \mid m_i) = \frac{C(t_i^j, m_i)}{C(m_i)} \quad \textbf{(4)}$$

где $C(t_i^j, m_i)$ – количество ссылок на концепцию $m_i$, в которых термин которых совпадал с $t_i^j$, включая редиректы и название концепции, как специальный тип ссылок, а $C(m_i)$ – общее количество ссылок на концепцию.

Чтобы оценить модель перехода сделаем предположение, что вероятность значения $m_i$, при условии предыдущего контекста $m_{i-1}...m_{i-k}$ пропорциональна линейной комбинации близости значения к контексту и априорной вероятности этого значения.

$$P(m_i \mid m_{i-1}..m_{i-k}) = \hat{P}(m_i \mid m_{i-1}..m_{i-k}) = \alpha\left(sim(m_i; m_{i-1}..m_{i-k}) + \beta * P(m_i)\right) \quad \textbf{(5)}$$

Для модели первого порядка близость значения к контексту, соответствующему предыдущему значению, вычисляется через коэффициент Дайса, способом, описанным в разделе 3.1. Чтобы оценить близость значения к контексту из нескольких терминов, представим их в виде обобщенной концепции, объединяющей все входящие в нее значения, тогда

$$n(B_1 B_2 ... B_m) = \bigcup_{i=1}^{m} n(B_i), \quad \textbf{(6)}$$

и близость вычисляется так же, как и для двух обычных концепций.

Априорную вероятность значения будем оценивать на основе ссылок, способом аналогичным тому, который мы использовали при оценке модели наблюдения.

$$P(m_i) = \hat{P}(m_i) = \frac{C(m_i)}{\sum_i C(m_i)} \quad \textbf{(7)}$$

Коэффициент нормализации $\alpha$ в уравнении 5 не влияет на решение задачи максимизации, поэтому его можно не учитывать. Коэффициент $\beta$ на основании экспериментов мы определили равным 1.

После определения всех параметров модели задача максимизации решается с помощью алгоритма Витерби. Этот алгоритм использует замечание, что наиболее вероятный путь до каждого следующего состояния зависит только от наиболее вероятного пути через $k$ предыдущих состояний. Таким образом, количество сравнений на каждом шаге экспоненциально зависит от k и равно

$$\prod_{i=k-n}^{k-1} | m_i |$$

Чтобы сократить время работы алгоритма, мы выдвинули наивное предположение, что **наиболее вероятный путь до состояния $m_i$ зависит от**

$k$ **последних терминов наиболее вероятного пути до состояния** $\mathbf{m}_{i-1}$**.** В этом случае каждое состояние должно хранить дополнительную информацию не более чем о $k$ предыдущих терминах, и, таким образом, модель сведется к специализированной Марковской модели первого порядка. Наиболее вероятная последовательность состояний для такой модели находится так же, как и для обычной Марковской модели первого порядка, за исключением вычисления вероятности перехода между состояниями.

Конечно, это предположение в общем случае неверно, однако в рамках данной задачи, оно позволяет уменьшить порядок модели и, при этом, учесть контекст из нескольких терминов, тем самым не сильно ухудшить точность метода (табл. 3 и 4).

# 4 Эксперименты

## 4.1 Коллекция для тестирования

Как уже упоминалось, для оценки точности и полноты методов обычно используются тестовые коллекции Senseval-1,2,3 и SemEval, основанные на WordNet. Однако различия между Википедией и WordNet не позволяют использовать их напрямую. Более того, отображение концепций Википедии на словарь WordNet (что само по себе является трудоемкой задачей, также требующей оценки) не дает возможности корректно сравнить алгоритмы из-за взаимной неоднозначности такого отображения. Поэтому в методах, основанных на Википедии, в качестве тестового корпуса часто используются сами статьи Википедии, причем обрабатываются только термины, представленные в виде ссылок, а значениями этих терминов служат концепции Википедии, на которые указывают ссылки. Несложно заметить, что для таких тестовых корпусов методы, основанные на машинном обучении и обученные на Википедии, дают наилучшие результаты из-за схожести распределений обучающего и тестового множеств [17].

Для оценки нашего метода мы создали тестовое множество, выделив 500 случайных статей Википедии. Использовались только статьи, описывающие однозначные концепции, так как они наиболее близки к неструктурированным текстам. Кроме этого, для составления более полной картины, мы вручную разметили тестовую коллекцию из 131 документа, состоящую из новостных сообщений, взятых из различных источников, и нескольких научных статей. Характеристики коллекций представлены в таблице 2.

Среднее число значений многозначных терминов в обеих коллекциях сильно превышает аналогичное число во всем языке (табл. 1). Это происходит потому, что у часто употребляемых терминов значений больше. Кроме того, процент многозначных терминов в коллекции, размеченной вручную, намного превышает аналогичное число в коллекции, автоматически созданной из статей Википедии.

| | Новости и научные статьи | Статьи Википедии |
|---|---|---|
| Количество документов | 131 | 500 |
| Количество терминов | 8236 | 50974 |
| Многозначных терминов | 6952 | 39332 |
| Среднее количество значений | 22,34 | 35,34 |

**Таблица 2: Характеристики тестовых коллекций**

Также следует заметить, что с изменением Википедии приходится изменять и тесты, так как появляются новые термины, и увеличивается количество значений. И если создать тестовую коллекцию по Википедии можно автоматически, то тесты, созданные вручную, придется заново вручную переразметить.

## 4.2 Результаты

Результаты экспериментов представлены в таблицах 3 и 4. Все эксперименты проводились на снимке Википедии, полученном в октябре 2008г. Алгоритм применялся ко всем найденным терминам текста, поэтому точность и полнота совпадают.

| Порядок | Модель Маркова | ММ с эвристикой |
|---|---|---|
| 0 | 53.12 | 53.12 |
| 1 | 54.00 | 54.00 |
| 2 | 54.50 | 54.49 |
| 3 | 54,76 | 54.72 |

**Таблица 3: Результаты работы алгоритма на коллекции новостей и научных статей**

| Порядок | Модель Маркова | ММ с эвристикой |
|---|---|---|
| 0 | 91,34 | 91,34 |
| 1 | 91,64 | 91,64 |
| 2 | 92,40 | 92.37 |
| 3 | 92,51 | 92,41 |

**Таблица 4: Результаты работы алгоритма на коллекции статей Википедии**

Сравнительно низкие результаты, полученные на первом корпусе, связаны с тем, что мы считали заведомо неверным ответ алгоритма, данный для терминов, не имеющих правильного значения среди концепций Википедии. Такими терминами, в основном, являются имена людей и слова, входящие в устойчивые выражения, например, слово “lot” в выражении “a lot of time…”. Если не учитывать такие термины, точность алгоритма достигает 76,84%.

Для сравнения алгоритм, описанный в работе [22] показывает точность 43,41% и полноту 34,77% на первом тестовом наборе и, соответственно, 79,58% и 77,29% на наборе из статей Википедии. На снимке, сделанном в июле 2008г., этот алгоритм давал точность и полноту 59,19% и 49,60% на первой коллекции и 91,93% и 89,62% на второй. Эти результаты наглядно демонстрируют, как с ростом Википедии ухудшается точность алгоритмов, использующих однозначный контекст.

Наилучшие результаты были представлены в работах [14] и [17] (94,33/70,51 и 98.4/95,7) и получены на аналогичных коллекциях, основанных на статьях Википедии. Однако, эти алгоритмы так же используют однозначный контекст, что, несомненно, ухудшит их точность и полноту при использовании с новыми версиями Википедии.

## 5 Заключение

В работе предлагается метод устранения лексической многозначности терминов естественного языка, основанный на Марковской модели, параметры который вычислены с помощью данных Википедии. Проблема разреженности языка решается предположением, что апостериорная вероятность значения термина при условии предыдущего контекста пропорциональна линейной комбинации семантической близости соответствующих концепций Википедии и априорной вероятности значения. Для ускорения алгоритма предложена эвристика, которая, в рамках поставленной задачи, дает выигрыш по времени выполнения, при этом незначительно ухудшая точность результата.

Однако, анализ ошибок алгоритма позволил выявить существенный недостаток: метод неявно предполагает, что все термины в тексте имеют общий смысл. Однако, часто в тексте кроме основной семантической линии существует несколько параллельных, таких как место и время основных событий. Основываясь на этом замечании, мы пришли к выводу, что применение данного алгоритма необходимо комбинировать с методом, выделяющим семантически связанные цепочки терминов (lexical chains). В этом направлении мы планируем сделать следующий шаг работы.

## Литература


[1] Eneko Agirre, Philip Glenny Edmonds. Word Sense Disambiguation: Algorithms and Applications. Springer, 2006

[2] Razvan Bunescu, Marius Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento, Italy, April 2006

[3] Silviu Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In Proc. 2007 Joint Conference on EMNLP and CNLL, pages 708–716, Prague, The Czech Republic, 2007.

[4] Cycorp, Inc. Web site. www.cyc.com

[5] Francis, W. and Kucera, H. Brown Corpus Manual. http://icame.uib.no/brown/bcm.html

[6] Glen Jeh, Jennifer Widom, SimRank: a measure of structural-context similarity, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002

[7] Nancy Ide and Jean Vйronis. Word sense disambiguation: The state of the art. Computational Linguistics, 1998

[8] Michael Lesk, Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, ACM Special Interest Group for Design of Communication Proceedings of the 5th annual international conference on Systems documentation, p. 24 – 26, 1986.

[9] D. Lizorkin, P. Velikhov, M. Grinev and D. Turdakov. Accuracy Estimate and Optimization Techniques for SimRank Computation. In VLDB '08: Proceedings of the 34th International Conference on Very Large Data Bases, pages 422--433.

[10] C. Loupy, M. El-Beze, and P. F. Marteau. 1998.Word Sense Disambiguation using HMM Tagger. In *Proceedings of the 1st International Conference* on Language Resources and Evaluation, LREC, pages 1255–1258, Granada, Spain, May.

[11] Olena Medelyan, Ian. H. Witten and David Milne. Topic Indexing with Wikipedia. Proc. AAAI'08 Workshop on Wikipedia and Artificial Intelligence

[12] Menczer Filippo. Evolution of document networks. Proceedings of the National Academy of Sciences of the United States of America.

[13] Rada Mihalcea. Using Wikipedia for Automatic Word Sense Disambiguation. Proceedings of NAACL HLT 2007, pages 196–203, Rochester, NY, April 2007

[14] Mihalcea, R. and Csomai, A. (2007) Wikify!: linking documents to encyclopedic knowledge. In Proceedings of the 16th ACM Conference on Information and Knowledge management (CIKM'07)

[15] George A. Miller, WordNet: a lexical database for English, Communications of the ACM, v.38 n.11, p.39-41, Nov. 1995

[16] David Milne, Ian H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. Proc. AAAI'08 Workshop on Wikipedia and Artificial Intelligence

[17] David Milne, Ian H. Witten. Learning to Link with Wikipedia. Proceedings of the ACM Conference on Information and Knowledge Management, 2008
[18] Antonio Molina and Ferran Pla and Encarna Segarra and Lidia Moreno. Word Sense Disambiguation using Statistical Models and WordNet. Proceedings of 3rd International Conference on Language Resources and Evaluation, LREC2002, Las Palmas de Gran Canaria
[19] Molina, F. Pla, E. Segarra, WSD system based on specialized Hidden Markov Model (upv-shmm- aw), in: R. Mihalcea, P. Edmonds (Eds.), Senseval: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 171-174
[20] Senseval. Web site. www.senseval.org
[21] The Penn Treebank Project. http://www.cis.upenn.edu/~treebank/
[22] D. Turdakov, P. Velikhov. Semantic Relatedness Metric for Wikipedia Concepts Based on Link Analysis and its Application to Word Sense Disambiguation. In proceedings of SYRCoDIS, 2008.

## Sense disambiguation of Wikipedia terms based on Hidden Markov Model

Denis Turdakov

The paper presents a method for word sense disambiguation using external knowledge extracted from the open encyclopedia Wikipedia. We analyse the drawbacks of the existing word sense disambiguation algorithms and propose own algorithm, based on Hidden Markov Model, to overcome these drawbacks. HMM parameters are estimated by empirical probabilities derived from the Wikipedia dictionary and link structure. A heuristics for speeding up the computational aspects of the algorithm is proposed, and the evaluation of the algorithm for several test collections is provided.

# Некоторые особенности формирования электронного корпуса текстов с синтаксической разметкой

© А.А. Рогов, Ю.В. Сидоров, А. В. Седов, Г.Б. Гурин, А.А. Котов, М.Ю. Некрасов

Петрозаводский государственный университет
rogov@psu.karelia.ru

**Аннотация**

В данной статье описывается система, созданная авторами для проведения синтаксической разметки текстов. Система создается и совершенствуется в рамках гранта РГНФ №08-04-12105в (Рук. Рогов А.А.). Эта работа является продолжением разработок по созданию грамматически размеченного корпуса публицистических текстов XIX века [1-3].

## 1 Выбор синтаксического аннотирования

Существующие немногочисленные корпусы со встроенной синтаксической разметкой опираются либо на общепринятые классификации традиционной («школьной») грамматики (Хельсинкский аннотированный корпус русских текстов ХАНКО; http://www.slav.helsinki.fi/hanco/index.html), либо на доступные узкому кругу специалистов и требующие детального предварительного знакомства классификации, например разметка в терминах деревьев зависимостей и синтаксических отношений, принятых в теории «Смысл-Текст», как в Национальном корпусе русского языка (http://www.ruscorpora.ru). Эти пути аннотирования в целом решают разные задачи и имеют свои достоинства и недостатки. Опора на школьную классификацию существенно облегчает работу с корпусом и расширяет круг потенциальных пользователей до всех, кто получил школьное образование, однако пользователю придется смириться с недостатками традиционного подхода: нечеткостью понятий и, соответственно, разметки, множественностью и некоторой произвольностью синтаксического описания. Такой корпус скорее является удобным источником иллюстраций для преподавателей русского языка, переводчиков, он полезен для редакторов и самого широкого круга заинтересованных лиц. Классификации, принятые в рамках той или иной научной школы, заведомо осложняют процедуру овладения ресурсом, так как требуют тщательного знакомства с принципами разметки и единицами классификации, однако такое аннотирование в большей степени свободно от противоречий традиционного анализа. В создаваемом корпусе в основу синтаксической разметки положена идея структурной схемы в понимании Н. Ю. Шведовой и ее последователей, впервые отчетливо заявленная в «Грамматике современного русского литературного языка» [4], позднее наиболее полно отраженная и развитая в «Русской грамматике» [5]. С одной стороны, это обеспечивает достаточно широкий охват пользователей, так как знакомство с классификацией синтаксических образцов в терминах структурных схем предполагается стандартными вузовскими курсами синтаксиса на филологических факультетах, эти классификации описываются в целом ряде распространенных учебников, с другой стороны, анализ формы предложения позволяет объективировать и упорядочить, насколько это возможно, систему разметки. Создание полного списка структурных схем простого предложения (в корпусе размечаются предикативные клаузы) – отдельная научная проблема, не имеющая пока своего окончательного решения. На данный момент мы можем говорить о том, что в научном обороте существуют как минимум три списка структурных схем – различные как количественно, так и качественно: 1) список схем «Русской грамматики» (1980); 2) список «минимальных схем» В. А. Белошапковой; 3) список схем О. А. Крыловой и Е. Н. Ширяева [6]. Последние на основе достаточно убедительного теоретического обоснования значительно переработали и дополнили исходный список свободных структурных схем «Русской грамматики». Именно эта классификация является на сегодняшний день наиболее полной и точной, и с небольшими изменениями и дополнениями была взята за основу разметки настоящего корпуса. Этот выбор объясняется двумя причинами: во-первых, использование структурных схем для синтаксической разметки в корпусе имеет свою специфику, во-вторых, ситуация изучения вопроса

**Труды 11[ая] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

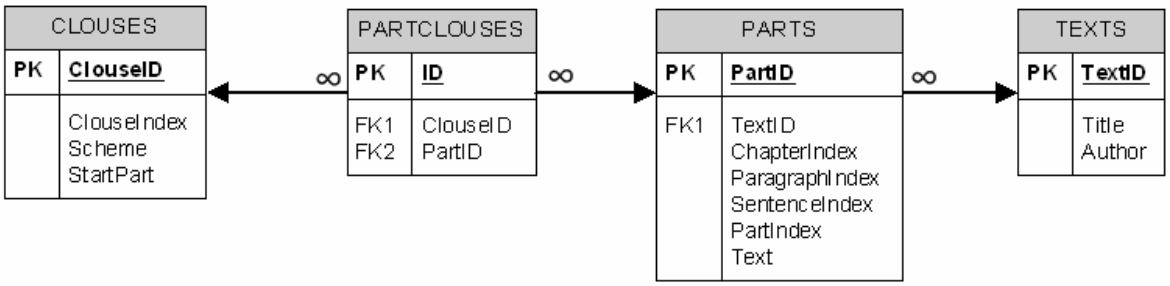



Рис. 1. Структура базы данных предназначенной для хранения синтаксической разметки

такова, что ни один из существующих списков структурных схем нельзя признать окончательно полным. Таким образом, на выходе мы получили наиболее полный и сбалансированный список структурных схем простого предложения, который будет использован для синтаксической разметки текстов.

## 2 Используемые структуры данных

Основой информационной системы синтаксической атрибуции является база данных, в которой хранится информация о синтаксических разборах текстов. Следовательно, первоначальной задачей была разработка структур данных для хранения информации о разборе текста. Основным структурным элементом синтаксической разметки, подвергаемым разбору, является клауза – минимальная предикативная единица, которая может выступать в качестве самостоятельного простого предложения или составной части сложного предложения. В составе сложной синтаксической структуры клауза может быть разбита одной или несколькими другими клаузами. Например, предложение «Вася пошел в бассейн, который открылся на днях, и плавал там до вечера», очевидно, делится на три части «Вася пошел в бассейн», «который открылся на днях», «плавал там до вечера», однако оно состоит из двух клауз: «Вася пошел в бассейн и плавал там до вечера», «который открылся на днях». То есть, как видно, 2 части предложения объединяются в одну клаузу. Для осуществления этого мы решили разделять понятия клаузы и части предложения. Приняв во внимание все написанное, структура текста у нас приобретает следующий вид: текст разбивается на главы, главы – на абзацы, абзацы – на предложения, предложения на клаузы, клаузы состоят из частей. Стоит отметить также, что одна часть предложения может принадлежать нескольким клаузам. Это возможно когда предложение содержит несколько однородных сказуемых, потому что сами схемы, прежде всего, различаются именно по структуре сказуемых, входящих в предложение. Например, предложение «Чудище обло, озорно, огромно, стозевно и лаяй» состоит из 5 клауз, причем слово «чудище» участвует во всех пяти. Предлагается структура данных, для которой это предложение разбивается на 6 частей:

P1=”Чудище”
P2=”обло”
P3=”озорно”
P4=”огромно”
P5=”стозевно”
P6=”лаяй”
Из этих частей составляются 5 клауз:
CL1={P1, P2};
CL2={P1, P3};
CL3={P1, P4};
CL4={P1, P5};
CL5={P1, P6}.

Каждая из 38 синтаксических схем кодируется числом от 1 до 38, поэтому при разборе каждой клаузе ставится в соответствие номер соответствующей схемы. Еще стоит отметить, что для клауз требуется хранение номера части, с которой эта клауза начинается.

На основе разработанной синтаксической разметки текстов были сформированы структуры таблиц и триггеров базы данных. Общая схема представлена на Рис.1. Для ее построения использовалась СУБД Interbase 6.0, и к настоящему времени она содержит 4 таблицы, и занимает объем 50 Мб.

Для каждой части текста определяется ее координаты: номер главы, абзаца, предложения, клаузы и ее содержание. Для каждой клаузы определяется ее номер, разбор этой клаузы и ее начальная часть. Части и клаузы соединяются связью «многие ко многим», так как одна клауза может состоять из нескольких частей, и, вместе с тем, одна часть может входить в состав нескольких клауз.

## 3 Программа синтаксического разбора

На вход описываемой программы подается текстовый файл формата .txt в кодировке unicode. На его основе создается рабочий проект, содержащий несколько файлов, с информацией о разбиении текста на структурные части, о синтаксическом разборе и прочие служебные файлы. Программа при открытии текста автоматически разбивает текст на структурные компоненты: главы, абзацы и предложения. Признаком новой главы является знак параграфа, расположенный первым на строке, признаком нового абзаца является табуляция, символами конца предложения являются точка, восклицательный и

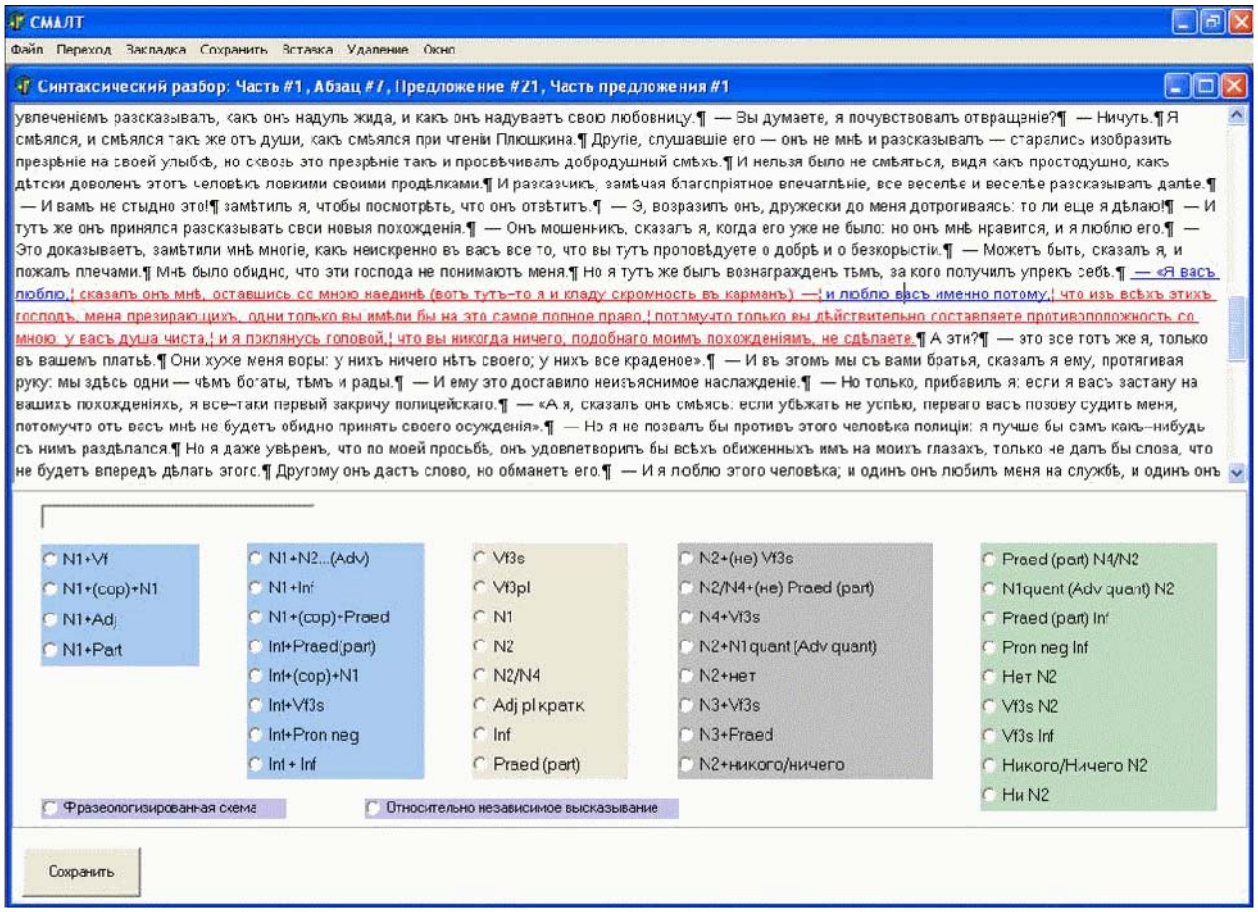



Рис. 2. Интерфейс программы синтаксического разбора

вопросительный знаки и т.д. На этом этапе могут возникать некоторые проблемы с автоматическим разбиением текста:

1. Существуют сложные знаки препинания, служащие концами предложения: «…», «?..», «!..» и пр. При автоматическом разборе каждый из этих знаков будет разделен на составные и получится, что в одном месте заканчиваются сразу 3 предложения. Эта проблема решена путем введения специальных соответствующих шаблонов.
2. Более сложная проблема связана с тем, что точка (как и остальные общепринятые признаки концов предложения) не всегда свидетельствуют о завершении предложения. Зачастую точка является признаком сокращения («…г. Волошин…»). Иногда вышеназванные знаки ставятся внутрь прямой речи, не являясь при этом признаками завершения предложения. Эту проблему можно решить путем анализа следующего слова: если оно начинается с прописной буквы, то можно сделать вывод о том, что рассматриваемый знак препинания не является концом предложения. Но если следующее слово начинается с заглавной буквы, это не гарантирует, что знак являлся концом предложения.
3. Стандартным признаком конца абзаца являются символы перевода строки. Однако перевод строки может использоваться и внутри текста, например в стихотворных отрывках. Выходом для большинства таких случаев будет проверка на то, заканчивалось ли предложение прямо перед переводом строки.

На этапе синтаксического разбора, который начинается сразу же за этапом разбиения текста, пользователь, если разбиение текста на предложения содержит ошибки и неправильно расставленные концы предложений/абзацев, может исправить вручную и имеет возможность редактировать разметку текста и разбивать предложения на части. Если клауза состоит из нескольких частей, пользователь может объединять их. В интерфейсе программы пользователю выводится пять или менее абзацев (текущий абзац, два предыдущих и два последующих), в которых возможна только покомпонентная навигация, то есть переходы только по клаузам. Всегда отдельно выделяется текущая клауза, которая разбирается в данный момент. Пользователь выбирает разбор (синтаксическую конструкцию) клаузы при помощи радиокнопок (т.н. «radio button»), путем выбора одной из схем. При сохранении разбора программа автоматически переходит на следующую клаузу. При помощи меню или горячих комбинаций клавиш пользователь может переходить на соседние клаузы/предложения/абзацы/главы. Кроме того, по двойному щелчку на участке текста, пользователь может переходить на выбранную клаузу. По

просьбам специалистов, работающих с программой, в ней реализованы различные варианты поиска: поиск следующей или предыдущей клаузы, кроме того, точный переход при выборе главы, абзаца, предложения и клаузы. Отметим, что в программе есть функция закладки: пользователь может поставить закладку на ту клаузу, которую разбирает в данный момент, выйти из программы, и при дальнейшей работе просто перейти по закладке на ту же часть текста. Фрагмент работы программы представлен на Рис.2.

Для работы программы синтаксической разметки требуется ОС Windows, объем оперативной памяти: больше 128 Мб, место на жестком диске: 10 Мб для программы, в среднем 5 Мб на хранение каждого рабочего проекта (в зависимости от объема текста). Было произведено наполнение базы данных текстами суммарным объемом более 36000 клауз из текстов, принадлежащие Ф.М. Достоевскому и его современникам, из журналов «Время» и «Эпоха», тексты В.И.Даля и ряд различных публицистических текстов XIX века. На данный момент размечено порядка 60 текстов.

## 4 Оптимизация работы программы

С целью оптимизации работы программы синтаксической атрибуции было решено провести ряд исследований, направленный на ускорение процедуры атрибуции текстов. Естественным вариантом ускорения работы видится автоматизация начального выбора синтаксической атрибуции клауз. Для ускорения процесса разметки проведен анализ различных вариантов его автоматизации на основании различных статистических методов. Критерием оптимальности является наименьшее время работы пользователя с программой при атрибуции текста. Для автоматической атрибуции было решено исследовать следующие эмпирические подходы:

1. Простейший статистический метод. Синтаксическая схема по умолчанию выбирается как наиболее часто используемая конструкция.
2. Анализ разбора клаузы, предшествующей текущей.
3. Анализ разбора двух предшествующих клауз.

Заметим, что все эти исследования можно провести, основываясь только на информации из базы данных уже имеющихся разборов. Действительно, взяв за основу клаузу с ее координатами в тексте, можно получить ее действительный разбор и разбор, прогнозируемый по каждому из вышеописанных подходов. Тут же надо оговориться, что внедрение слишком сложных интеллектуальных методов замедляет работу программы, и то преимущество во времени, которое может быть получено за счет автоматического выбора клауз, может быть нивелировано замедлением функционирования самой программы.

Существуют и другие подходы к оптимизации работы с программой. Можно, например, расставлять структурные схемы на панель выбора в соответствии с частотой их использования: самые часто используемые схемы поместить слева и, по мере уменьшения использования, располагать другие схемы все правее. Однако, как подтвердили специалисты работающие с программой, побочный эффект данного метода намного хуже, так как они привыкли к определенному расположению кнопок с выбором схем, как правило нажимают их уже не глядя, и постоянная перестановка этих кнопок сбивает специалистов и ощутимо замедляет время работы с программой. Это так называемый “эффект qwerty”.

**Статистический метод**

На основе разобранных текстов мы выбираем наиболее часто встречаемую схему, то есть схему, при помощи которой разбирается наибольшее количество клауз.

Результаты исследования:

Проведя анализ, мы выделили 7 наиболее употребляемых структурных схем. Они приведены в следующей таблице:

| Схема | Количество | Частота |
|---|---|---|
| $N_1 + V_f$ | 21195 | 58,5 % |
| $N_1$ + Adj | 3021 | 8,3 % |
| $N_1$ + (cop) + $N_1$ | 2344 | 6,5 % |
| $N_1$ + Part | 1474 | 4 % |
| $Praed_{(part)}$ Inf | 1160 | 3,2 % |
| $N_1$ | 1087 | 3 % |
| $N_1 + N_2 \ldots_{(Adv)}$ | 1038 | 2,9 % |

Табл. 1 Статистика частотности употребления структурных схем

Общее количество проанализированных клауз равно 36224. Как видно из результатов, при помощи схемы $N_1 + V_f$ разбирается больше половины клауз текста. При таком подавляющем преимуществе можно говорить о целесообразности выбора этой схемы как схемы по умолчанию.

**Анализ разбора предыдущей схемы**

Суть метода в следующем: мы пытаемся предположить, что существует связь между разборами двух рядом стоящих схем. Тогда мы предлагаем определять начальный разбор текущей клаузы на основе разбора предыдущей. Для этого мы анализируем для всех структурных схем разборы клауз, следующих за клаузами, разобранными этими схемами, и выбираем самый распространенный из них. В процессе атрибуции, при выборе структурной схемы для текущей клаузы, следующая клауза будет атрибутироваться автоматически.

Формально:

Есть структурная схема St;

Получаем набор клауз из текстов, атрибутированных ею $Cl = (cl_1, .. cl_n)$;

Получаем набор клауз, следующих за Cl : $Clp = (clp_1, .. clp_n)$;

Группируем эти клаузы в зависимости от их разбора: $Gr_{ij} = \{gr_{ij}\}$, i – номер группы, j – номер клаузы в группе;

Для каждой группы записываем соответствующую ей структурную схему: Stpi;

Считаем количество клауз в каждой группе: $N_i = \{n_1, n_2, .., n_k\}$;

Ищем номер наибольшей группы $i^* = \text{argmax}(n_1, n_2, .., n_k)$;

Тогда предполагаемый разбор клаузы, следующей за клаузой с разбором St, будет $St_i^*$.

Видно, что описанный метод является динамическим. То есть определять предполагаемый разбор клаузы необходимо непосредственно в процессе атрибуции, после того как разобрана предыдущая клауза, что приводит к замедлению разбора.

Можно рассматривать разборы не одной, а двух предыдущих клауз. Подробнее описывать этот метод не стоит по одной простой причине: как показало исследование этот и предыдущий методы абсолютно неэффективны. Учитывая, насколько одна структурная схема встречается чаще, чем все остальные, очевидно, что для любой схемы, следующей будет выбираться именно схема $N_1 + V_f$. То есть при применении методов анализа одной или двух предыдущих клауз, мы будем получать тот же результат, как и при применении статистического метода, только куда более алгоритмически сложно и ресурсоемко.

Из всего написанного можно сделать вывод, что наиболее эффективным будет применение статистического метода, выраженного в том, чтобы каждой схеме ставить начальным разбором структуру $N_1 + V_f$ . Этот метод будет давать верное предсказание примерно в 58,5 % случаев.

## 5 Представление результатов

Полученный в результате выполнения проекта размеченный корпус представляется конечным пользователям при помощи Web-ресурса, его адрес http://smalt.karelia.ru. На нём представлена возможность формирования собственного подкорпуса из предложенных проанализированных текстов и получения статистических данных, например, таких как частота встречаемости синтаксической конструкции в выбранном подкорпусе. На Интернет ресурсе предоставлена возможность нахождения требуемой синтаксической конструкции в выбранном подкорпусе текстов, для чего пользователь вводит при помощи дополнительного окна список искомых конструкций, выбирает тексты, по которым будет производиться поиск, и нажимает на кнопку «найти». В результате открывается html страница, содержащая набор предложений, разбитый на группы по 10 предложений. Переход между группами осуществляется при помощи кнопок навигации, а также при помощи ввода номера группы и непосредственный переход к ней. По желанию, пользователь может ввести слово, которое должно содержаться в данной синтаксической конструкции. Имеется возможность получения контекста данного предложения, а также переход к целиком произведению, в котором содержалось данное предложение.

Заметим, что работа по созданию системы синтаксической разметки является продолжением работы, в рамках которой ранее была создана система грамматической разметки. Эта разметка проводилась по более традиционной схеме: каждому слову ставился в соответствие ряд его грамматических параметров, как-то, часть речи, число, род и пр. Потребовалась синхронизация информации из обоих проектов в одну базу данных. Как следствие, на web-ресурсе стал возможен поиск не только отдельно по синтаксическим или по грамматическим параметрам, но также и смешанный поиск по обеим разметкам. В связи с этим встала проблема модернизации базы данных для оптимизации смешанного поиска. Отметим, что при синхронизации грамматической и синтаксической разметок возникал ряд сложностей. При грамматическом разборе каждому слову тоже ставились в соответствие координаты в тексте (глава, абзац, предложение, слово). Однако, так как разметки проводились разными людьми в разное время, структурное разбиение текста на части могло быть различным. Как следствие, приходилось проводить сканирование всех текстов для выравнивания разбиения текста на части. Для организации смешанного поиска было решено сделать отдельную таблицу, дублирующую ряд информации из других таблиц, но оптимизированную именно под поиск одновременно слов с выбранными грамматическими параметрами, находящихся в клаузах с выбранной синтаксической структурой. Каждая запись этой таблицы содержит слово из словаря, его грамматические параметры и ряд полей, содержащих информацию о том, встречается ли это слово в заданной синтаксической схеме.

## 6 Оптимизация базы данных

Возникла задача разработки такой структуры базы данных, чтобы минимизировать время выполнения запроса пользователя. Рассмотрим возможные критерии оптимальности предоставления и хранения данных. Среди них можно выделить среднюю скорость предоставления информации, объём хранимой информации, полноту информации, представительность.

**Оптимизация по времени поиска.** Рассмотрим вопрос разработки оптимальной структуры БД, предназначенной непосредственно для поиска. Для этого необходимо учесть структурные особенности поиска (поиск вхождений слова, грамматической или синтаксической конструкции и выдача результата согласной некой мере близости). Рассмотрим первый критерий оптимальности:

минимизация среднего времени поиска. На сервер периодически поступают запросы, необходимо максимально уменьшить среднее время поиска необходимой информации. Обозначим, через n число обращений, $t_i$ - время обработки запроса, тогда среднее время обработки всех запросов равно:

$$T_{ср} = \frac{\sum_{i=1}^{n} t_i}{n}$$

На нынешний момент в базе существует два вида разборов: грамматический и синтаксический. На их базе производятся различные виды поисков (по слову, по грамматическим признакам, по синтаксическим признакам, комбинированный поиск). Для каждого вида можно подобрать свою структуру. Пользователю в результате поиска предоставляется следующая информация:

- Получение одного слова, с грамматическими признаками, а также по возможности наборов синтаксических конструкций, в которые оно входит;
- Получение предложения целиком и характеристики каждого слова, в составе данного предложения.

Остановимся более подробно на определении среднего времени поиска. Поскольку слово может быть как в современной орфографии, так и в орфографии языка XIX века, поэтому приходится производить поиск в базе по двум полям (словоформа и современное написание). Это выполнимо за линейное время. Необходимо иметь доступ к предложению, содержащему данное слово, тогда таблица для поиска должна содержать ссылки каждого слова на предложения, где оно может быть найдено. Обозначим за $t_{11}$ - время поиска слова, а за $t_{12}$ - время поиска всех слов из того же предложения по таблице, состоящей из слов, их современных написании и «координат» слов в тексте. Тогда время обработки одного запроса $t_i$ будет равно: $t_i = t_{11} + t_{12}$.

В случае необходимости получения грамматических и синтаксических параметров для каждого слова можно хранить характеристики каждого слова в той же таблице, добавив соответствующие поля:

а) Добавив на каждый грамматический признак соответствующий столбец, и один столбец на синтаксический параметр. При этом время поиска данных в БД изменится. Тогда время обработки одного запроса $t_i$ будет равно времени поиска слова $t_{21}$ + времени поиска всех предложений содержащих слово $t_{22}$ + время, потраченное на поиск значения определённого параметра в соответствующей таблице $t_{23}$. Общее время поиска составит $t_i = t_{21} + t_{22} + t_{23}$;

б) Добавив 15 столбцов (по максимальному числу различных грамматических параметров + вид синтаксической конструкции). Тогда время обработки одного запроса $t_i$ будет равно времени поиска слова $t_{31}$ + время поиска всех предложений содержащих слово $t_{32}$ + время на поиск признака в таблице $t_{33}$ и расшифровку значения определённого параметра в соответствующей таблице $t_{рек.расш.}$. Общее время поиска: $t_i = t_{31} + t_{32} + t_{33} + t_{рек.расш}$.

Можно осуществлять хранение в одной таблице самого слова, начальной формы и грамматических параметров, а принадлежность к определённому предложению перенести в отдельную таблицу. Тогда время обработки одного запроса $t_i$ будет складываться из времени потраченного на поиск слова $t_{41}$ время на поиск предложений во второй таблице $t_{42}$ плюс время на поиск слов оставшихся слов в первой таблице $t_{43}$ плюс расшифровка параметров ($t_i = t_{41} + t_{42} + t_{43} + t_{рек.расш.}$). При данной структуре таблиц мы перестаём хранить избыточную информацию (если одно слово употребляется очень часто, то оно не копируется, а используется одна ссылка.) Преимущество - меньше записей в таблице - меньше время поиска.

В результате получаем несколько видов таблиц для различных видов поиска. Обозначим за $a_{ij}$ - среднее время j-го вида поиска при использовании i-го вида таблиц. Тогда введём $\Sigma_i = \beta_1\alpha_{i1} + \beta_2\alpha_{i2} + \ldots + \beta_{10}\alpha_{i10}$ - функция, характеризующая среднее время, которое тратится на поиск пользователем необходимой информации. Коэффициенты $\beta_j$ - определяют предпочтение определённого вида поиска. Изначально все коэффициенты $\beta_j$ приравнивались к $k = \frac{1}{число\ видов\ поиска}$ (использование различных типов запросов является равновероятным). В результате выбирается структура данных обеспечивающая наименьшее время.

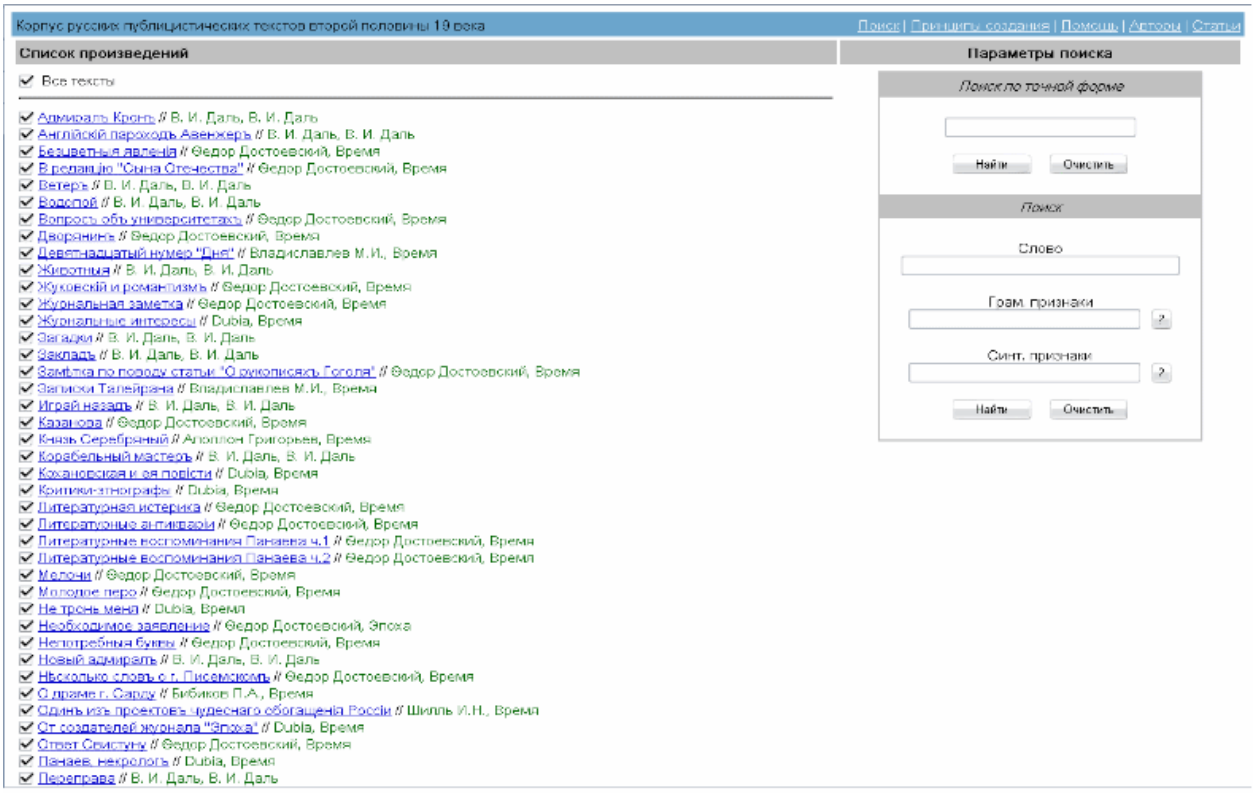



Рис. 3. Главная форма выбора параметров размеченного корпуса текстов.

Причина, по которой рассматривается среднее время поиска, состоит в следующем: предполагается, что изначально пользователя (специалиста) интересуют все слова одинаково. В дальнейшем, при сборе информации относительно "излюбленных" слов можно модернизировать БД с учетом статистики запросов по конкретным словам. То есть требуется построение и упорядочивание данных таким образом, чтобы наиболее часто искомые слова находились вначале.

**Оптимизация по объёму.** Время – не единственная характеристика для оптимизации БД. Важным параметром БД является её объём. Можно максимально сократить объём базы данных и при этом получить очень сложный поиск, и наоборот, создав очень быстрый поиск, мы сталкиваемся с проблемой не эффективного использования памяти. Поэтому для исследуемых баз данных рассматривали характеристику $Q = \Sigma_i * M$ - соотношение скорости получения информации к объёму БД. И на основании данной величины производился выбор подходящей структуры.

**Вывод информации на экран пользователя.** Обработанная информация будет предоставлена конечному пользователю при помощи веб-интерфейса. Необходимо ограничить число результатов предоставляемых пользователю одновременно. На наш взгляд, наиболее наглядным является предоставление информации на один экран, без скроллинга. Отсюда появляется задача определения среднего числа строк, занимаемых каждым из результатов поиска. На основании данного параметра можно будет определить необходимое число результатов, выдаваемое при поиске. Наиболее популярным разрешением экрана на данный момент является 1024x768 точек на дюйм.

Пусть $A$ - число строк, помещающихся на экране, $\overline{c}$ - среднее число строк после одного запроса, тогда величина $S = \frac{A}{\overline{c}}$ - число результатов, которые должны быть представлены пользователю. Величина $A$ -считается исходя из разрешения экрана, а $\overline{c}$ - число символов результатов запросов поделённое на число запросов и на количество символов, помещающихся в одной строке, округлённое вверх.

**Полученные результаты.** В результате проведённых исследований была выбрана база данных, содержащая слово, современное написание и зашифрованные значения параметров в одной таблице. Время поиска искомого слова по такой базе составил в среднем 0.602 секунды, поиск всех слов по заданным грамматическим и синтаксическим параметрам занял около 19 секунд.

Разработанный интерфейс предоставляет возможность начать поиск с главной страницы. При этом пользователю предоставляется возможность в отдельных окнах выбрать искомые параметры и ввести искомое слово. Вид экранной формы представлен на Рис. 3. При желании можно выбрать подкорпус текстов, в котором производить поиск. После чего предоставляется страница поиска, где результаты представляются блоками по 10 элементов в каждом.

## 7 Заключение

Полученный в результате выполнения проекта синтаксически размеченный корпус может быть использован при научных изысканиях в области истории, грамматики, лексикографии, а также при изучении соответствующих курсов студентами филологических специальностей. Кроме того, он может быть востребован специалистами по литературе XIX века.

Заметим, что создаваемая информационная система универсальна по отношению к языку текста и типу разметки. В дальнейшем пользователь сможет самостоятельно определять элементы текста и формировать список атрибутов для этих элементов. Для введенных атрибутов можно будет указать структурные связи. Предусмотрена возможность создания собственных правил для парсера текста при помощи определенного метаязыка (язык регулярных выражений, набор правил), поиск вхождений текстов, частей текстов. С использованием введенных атрибутов размеченный текст можно будет представить в виде графа.

## Литература


[1] Рогов А. А., Гурин Г.Б., Котов А.А., Сидоров Ю.В. Морфологически размеченный корпус по русской публицистике второй половины XIX века // Проблемы компьютерной лингвистики: сборник научных трудов. Вып. 3. Воронеж: Издательско-полиграфический центр Воронежского государственного университета, 2008. С. 209-219.

[2] Рогов А.А., Гурин Г.Б., Котов А.А. Некоторые особенности грамматически размеченного корпуса по русской публицистике второй половины XIX века. / Труды международной конференции «Корпусная лингвистика - 2008». – СПб.: С.–Петербургский гос. университет, факультет филологии и искусств, 2008. С. 326-333.

[3] Рогов А.А., Гурин Г.Б., Котов А.А., Сидоров Ю.В., Суровцева Т.Г. Программный комплекс «СМАЛТ». // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 10 Всероссийской научной конференции «RCDL-2008» (Дубна, Россия 7-11 октября 2008г.). – Дубна: ОИЯИ, 2008. С. 155-160.

[4] Грамматика современного русского литературного языка. М.: Наука. 1970.

[5] Русская грамматика. М.: Наука. 1980. Т. 2.

[6] Современный русский язык: Фонетика. Лексикология. Словообразование. Морфология. Синтаксис / Под общ. ред. Л. А. Новикова. Спб.: Лань. 2003. С. 631-644.


## Some features of formation of digital corpus of texts with syntactic markup


Rogov A.A., Sidorov Yu.V., Sedov A.V., Gurin G.B., Kotov A.A., Nekrasov M.Yu.



In paper we describe system of syntactical analysis. For marking we use 39 structural schemes. The main marking unit is the clause – simple sentence.

For marking was created application in Delphi. In it user can move from clause to clause and attribute them. We also analyze some paths to improve effectiveness of program.

The structure of dictionary database tables was developed. Query execution speed was constantly analyzed during all period of work. The results of the analysis were taken into account while improving the structure of tables.

For online access to the database the web-resource was created. Modules of the information system were realization on PHP 4. For providing the maintenance of the pre-revolutionary Russian alphabet symbols all texts and wordforms are stored in the coding Unicode. The type Palatino Linotype is used for representation.

At present, database is composed of works, which belong to F.M. Dostoevskij and his contemporaries from the magazines “Vremja”, “Epoha”, “Svetoch”, “Sovremennik”, “Molva”, “Biblioteka dlja chtenija”, “Zarja”, “Grazdanin”. It also contains texts by Dahl and some other publicistic texts.

# Корпоративная переводческая сеть с использованием специальных электронных библиотек

© В.Е. Абрамов
ЗАО СКБ «ТЭЛКА»
abramval@yandex.ru

Н.Н. Абрамова, А.А Карнацкая, В.М. Рожков
НИЦИ при МИД России
nabramova@mid.ru, akarnatskaya@mid.ru, vrozhkov@mid.ru

**Аннотация**

В статье описывается создание корпоративной переводческой сети, объединяющей несколько автоматизированных рабочих мест (АРМ) переводчика, на платформе IBM Lotus Domino/Notes. Разработка этой сети вызвана необходимостью автоматизировать труд переводчиков, что поможет соответствовать современным требованиям качества и скорости перевода. Система построена на принципах памяти переводчика, когда используется большой корпус текстов на четырех языках (русском, английском, французском и испанском), переведенных вручную. Описаны компоненты АРМ переводчика, приводятся примеры работы системы.

## 1 Введение

В настоящее время достигнуты определенные успехи в области машинного перевода. Однако профессиональные переводчики практически не используют системы машинного перевода, мотивируя это неудовлетворительным качеством перевода. На постредактирование переведенного материала иногда можно затратить больше времени, чем на перевод по старинке без помощи программ. Как известно, ни одна из ныне существующих в мире систем перевода не может обеспечить уровень перевода, сравнимый с уровнем человека-переводчика.

В то же время, современному переводчику необходимы средства автоматизации, облегчающие его труд. Такие средства, называемые «память переводчика», «накопители переводчика» или «накопители переводов» стали создаваться начиная с 80-х годов прошлого века. Этому способствовали большие объемы накопленных к тому времени параллельных текстов на разных иностранных языках, переведенных вручную, а также прогресс в области вычислительной техники, позволивший создать информационные системы для накопления, хранения и поиска информации.

Появилось направление машинного перевода, основанное на принципе памяти переводчика. За рубежом основоположником этого направления явился японский профессор М. Нагао [6], а в России идеологом стал профессор Белоногов Г.Г., под руководством которого была создана система фразеологического перевода Retrans [3].

На аналогичных принципах строятся системы автоматизированного перевода, выполняющие в отличие от систем машинного перевода не полный перевод текста, а его фрагменты без формирования связного текста, оставляя за человеком значительную часть по переводу, согласованию и редактированию текста. На сегодняшний день известно несколько часто используемых систем автоматизированного перевода, например, Trados [7], OmegaT [12], SDLX [9], Wordfisher [14], Metatexis [10], DejaVu [8], Transit [11], TermStar [13]. В обзорной статье [5] и докладе [4] можно познакомиться с общими характеристиками и возможностями, а также особенностями их архитектуры и принципами работы.

Помимо программных систем помощь переводчику оказывают автоматические словари, среди которых автор обзора [5] отмечает Translatelt, PROMT VER-Dict , ABBYY Lingvo, Мультитран, Контекст.

## 2 Необходимость создания корпоративной переводческой сети

Однако, несмотря на некоторые успехи в области автоматизированного перевода, в реальной жизни переводчики с большим трудом могут воспользоваться этими разработками. Многие переводчики действуют по принципу «Omnia mea mecum porto» (все свое ношу с собой), стараясь иметь на своем компьютере (насколько позволяют технические возможности) как можно больше различных словарей, глоссариев, параллельных

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

текстов, программных продуктов, облегчающих процесс перевода.

В условиях глобализации современного мира выдвигаются более высокие требования к переводу. Постоянно происходит рост требующих перевода на иностранные языки материалов, так как расширяются международные связи, и как следствие, растет количество договорных актов и соглашений, заявлений в СМИ и т.п. К качеству перевода международных документов всегда предъявляются повышенные требования. Кроме того, переводчики ограничены во времени, поскольку переводные материалы должны появляться достаточно быстро после печати оригинала и даже в одно время с ним. На перевод накладываются довольно жесткие требования к используемой терминологии: термины, впервые появившиеся в основополагающих международных документах, таких как резолюции ООН, международные конвенции и договоры и переведенные на иностранные языки, в последующих документах должны переводиться таким же образом. То есть речь идет не столько о предоставлении переводчику переводных эквивалентов для ускорения процесса перевода, сколько о стандартизации этого процесса. Естественно, в организациях, имеющих штат профессиональных переводчиков, необходимо иметь корпоративную сеть, которая давала бы возможность всем переводчикам обращаться в единую информационную базу, чтобы каждый отдельный переводчик не тратил силы на создание собственной базы, а отдал бы свои лексические богатства, накопленные за долгие годы переводческой деятельности, для создания общей корпоративной базы данных.

## 3 Основные требования

Что же нужно переводчику для работы на современном уровне? Прежде всего персональный компьютер с достаточным объемом оперативной и дисковой памяти и высокой скоростью обработки информации, оснащенный DVD и периферийными устройствами (сканер, принтер, web-камера) и имеющий доступ в корпоративную сеть и сеть Интернет.

На рабочей станции (персональном компьютере) переводчика должен быть установлен хотя бы минимальный набор программ, позволяющих проводить обработку документов на русском и иностранных языках, включающий текстовый редактор, систему оптического распознавания текстов, электронные переводчики, клиент-серверное программное обеспечение (ПО) для совместной работы. Дополнительно на рабочую станцию можно установить системы автоматического реферирования текстов [1] и системы распознавания голоса. Базы данных с электронными словарями и параллельными текстами на разных языках должны быть в общем пользовании и располагаться на сервере.

## 4 Состав и принципы работы корпоративной переводческой сети

### 4.1 Особенности платформы IBM Lotus Domino/Notes

В качестве среды для разработки баз данных используется платформа IBM Lotus Domino/Notes, так как уже существует корпоративная информационная система, разработанная на этой платформе, и накоплены значительные объемы информации для автоматизации переводов [2].

В соответствии с выдвинутыми выше требованиями создается корпоративная сеть, в которую объединены несколько АРМов переводчиков. Платформа Lotus Domino/Notes позволяет использовать не только сервер приложений для формирования и ведения баз данных, содержащих информацию для переводческой деятельности, но и почтовый сервер и Web–сервер для получения и обмена дополнительной информацией помимо имеющейся в корпоративной сети. Сервер Domino поддерживается самыми распространенными операционными системами (ОС), например, такими как Windows, Linux, Solaris, iSeries, AIX, z/OS, что позволяет легко переходить из одной ОС на другую или использовать несколько серверов под разными ОС.

Сервис репликаций позволяет синхронизировать состояние копий баз данных на разных серверах и клиентских машинах (в нашем случае это АРМ переводчика). Система репликаций дает возможность организовать коллективную работу над переводимым документом благодаря функции автоматического управления версиями документа, отслеживающей изменения оригинала документа на всех АРМах.

Платформа Lotus Domino/Notes поддерживает формат Unicode, что позволяет работать с многоязычными документами. Начиная с 7-ой версии, в Lotus Notes встроен текстовый редактор и сервис проверки орфографии. Кроме того, результирующий файл может быть представлен в формате ряда текстовых редакторов, например, Word.

На рис. 1. приведена схема корпоративной сети на базе Lotus Domino/Notes, объединяющая пять АРМов переводчика.

### 4.2 Комплекс баз данных «Перевод»

На сервере Lotus Domino находится комплекс баз данных «Перевод», в составе которого имеется три базы данных: «Тексты для перевода», «Результаты поиска» и «Память переводчика».

База данных «Память переводчика», содержит специальные электронные библиотеки - словари и параллельные тексты.

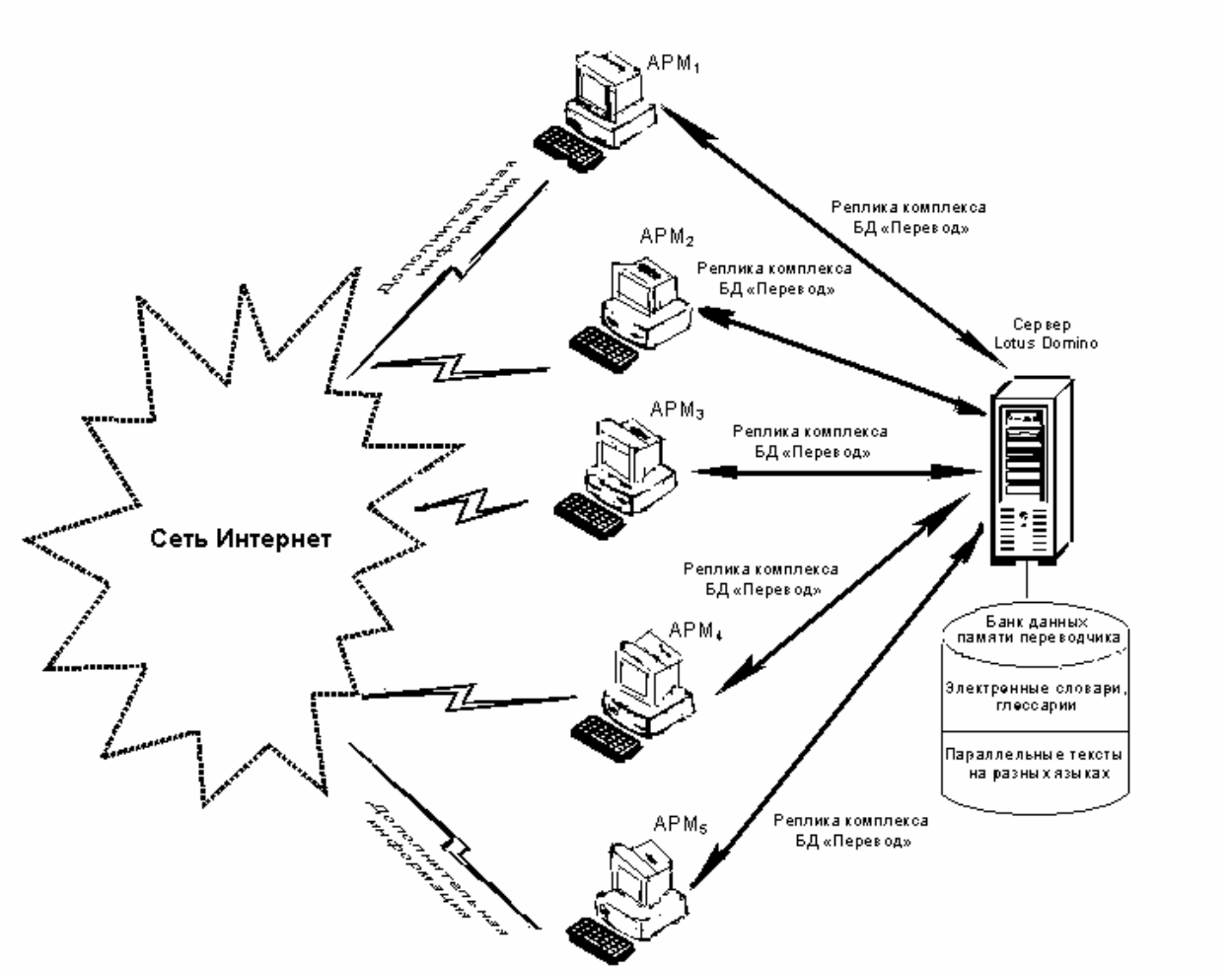



Рис. 1. Схема получения и обмена информацией в корпоративной сети, объединяющей АРМы переводчиков

Библиотека параллельных текстов формируется, в основном, за счет полных текстов документов международного характера (договоров, конвенций, меморандумов, резолюций ООН и т.д.).

Терминологические словари включают лексику из двуязычных глоссариев (русско-английских и русско-французских) по внешнеполитической деятельности, которые составлялись вручную, а также термины из многоязычного электронного словаря по внешней политике [2]. В настоящее время объем словаря составляет около 10 тыс. терминов, однако после перевода в электронную форму материалов на бумажном носителе объем словаря должен значительно вырасти ( примерно до 20-25 тыс. терминов).

В базе данных «Тексты для перевода» содержатся исходные документы, предназначенные для перевода, и результирующие (переведенные) документы.

База данных «Результаты поиска» содержит документы, найденные в «Памяти переводчика», в которые входят фрагменты, выделенные переводчиком в исходном тексте.

База данных «Память переводчика», доступна только на сервере, ее нет на АРМах, а реплики остальных баз имеются на каждом из АРМов.

### 4.3 Схема компонентов АРМ переводчика

На каждом АРМе должны быть установлены следующие средства:

- клиент-серверное программное обеспечение Lotus Notes;
- текстовый редактор Word Microsoft Office 2007;
- система оптического распознавания текстов Abby Fine Reader 8.0;
- электронный переводчик Promt ;
- электронный словарь Lingvo 9.0;
- реплика базы данных «Тексты для перевода»;
- реплика базы данных «Результаты поиска».

### 4.4 Подготовка текстов для ввода в базу данных «Память переводчика»

При вводе электронных терминологических словарей в базу данных памяти переводчика не возникает трудностей. Процесс осуществляется с помощью стандартных процедур импорта. Ввод параллельных текстов требует дополнительной обработки, которая заключается в выравнивании текстов на уровне абзацев. Такая опция присутствует во многих системах автоматизированного перевода, а также существуют специальные программы выравнивания текстов, доступные в бесплатном пользовании, например, bligner или bitext2tmx.

В нашей системе задача выравнивания значительно упрощается, так как тексты международных документов тщательно выверены и соблюдено полное соответствие их частей (статей, абзацев) на всех рабочих языках.

Как известно, в основе автоматизированной обработки информации на русском языке лежит морфологический анализ. Авторы использовали программу морфологического анализа, разработанную Абрамовым В.Е. на языке Borland C++ Builder 6.

Для подготовки импортируемых файлов разработана специальная программа, выполняющая разбивку параллельных текстов на абзацы и проведение морфологического анализа текста каждого абзаца на русском языке.

При вводе в базу данных для словарей и параллельных текстов предусмотрена одна общая форма, в которой имеются поля для записи названий документов и соответствующих абзацев текстов на двух языках, а также поля для результатов морфологического анализа текстов каждого абзаца.

На рис. 2 показано представление в базе данных «Память переводчика» русско-английского словаря, а на рис. 3 - русско-английских параллельных текстов. База данных содержит библиотеки параллельных текстов и терминологических словарей, сгруппированные для пар языков: русский - английский, русский - французский, русский – испанский.

### 4.5 Поиск информации в базе данных «Память переводчика»

Lotus Notes предлагает довольно полный объем средств поиска, которые позволяют найти указанные сведения в заголовках документов и в текстах документов по нескольким словам или фразам и с помощью поисковых запросов, составленных на основе булевской логики.

Например, можно составить запрос на поиск слов или словосочетаний, которые обязательно должны присутствовать в документе, кроме того, указать критерий их близости, регистры составляющих их букв, их веса.

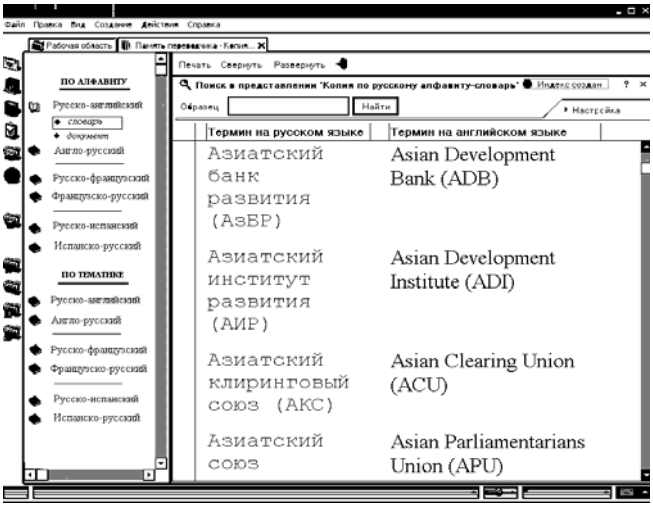



Рис. 2. Представление русско-английского словаря

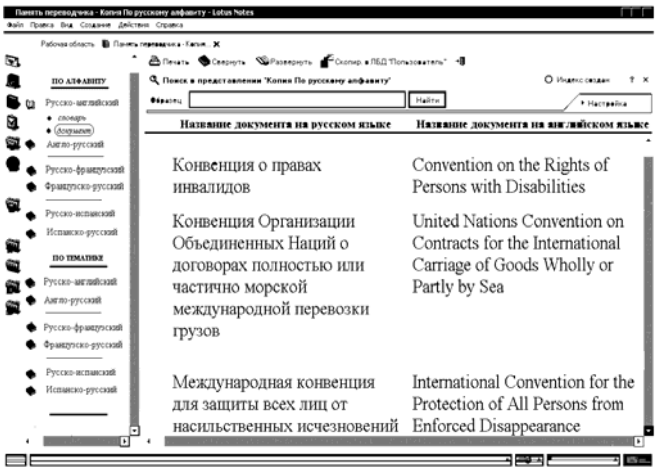



Рис. 3. Представление русско-английских параллельных текстов

Однако для решения поставленной задачи невозможно обойтись только стандартными средствами поиска. Это не слишком удобно для пользователя. Так, открыв найденный в результате поиска документ, нужно путем просмотра документа найти нужный абзац, опираясь на выделенные цветом ключевые слова из запроса. Кроме того, операторы поиска с учетом близости слов в предложении или абзаце работают не всегда корректно в силу того, что в алгоритмах поиска заложен не очень удачный критерий определения границ предложений и абзацев.

Не предусмотрена также операция выделения мышью какого-либо фрагмента текста и поиска его в базе данных.

Стандартные средства не позволяют найти близкие по смыслу к переводимому тексту или абзацу тексты в «памяти переводчика».

Для реализации указанных операций авторы системы разработали специальные программные средства на объектно-ориентированном языке программирования LotusScript .

Обращение из скрипта Lotus к программе морфологического анализа происходит при

выполнении морфологического разбора выделенного фрагмента переводимого текста. Затем в базе данных «Память переводчика» средствами Lotus находятся все документы, отвечающие запросу - текстовой строке, составленной из основ слов фрагмента. В каждом документе, найденном по запросу, ищутся абзацы, содержащие эту сформированную текстовую строку. Сравниваемая строка выбирается из результатов морфологического анализа текста абзаца, полученного на этапе подготовки текста для импорта в базу данных «Память переводчика». Все абзацы, соответствующие запросу, записываются в базу данных «Результаты поиска» вместе со ссылкой на полный текст документа.

### 4.6 Технология работы с базами данных

Исходные тексты для перевода можно разместить в базе данных различными способами:

- импортировать с помощью средств Lotus Notes;
- ввести с клавиатуры;
- скопировать через буфер обмена из Интернета или какого-либо текстового редактора.

Обязательными для заполнения полями являются поле «наименование», в которое вносится заголовок документа, и поле «исходный текст», куда непосредственно заносится текст документа. В поле «перевод текста» информация заносится в процессе работы. По желанию переводчика там может быть размещен текст, переведенный автоматически с помощью системы машинного перевода. Работая над переводом текста, переводчик может выделить мышью какой-либо фрагмент и выбрать электронную библиотеку (терминологический словарь или корпус параллельных текстов) на нужном языке, в которой будет проводиться поиск, щелкнув по соответствующей пиктограмме на панели действий (см. рис.4).

Результатом выполненного действия будет переход к представлению базы данных «Результаты поиска», в котором указывается текущая дата, фрагмент, по которому проводился поиск и название документа, содержащего этот фрагмент (см. рис. 5). Для удобства пользователей результаты содержат только те абзацы документов, которые включают найденные фрагменты. Документ открывается с помощью щелчка мышью по его названию. Текст располагается в двух колонках, причем каждому абзацу на русском языке соответствует его перевод на иностранном языке, найденный фрагмент выделяется цветом. Под названием документа находится значок гиперссылки, нажав на который переводчик может просмотреть полный текст документа (см. рис. 6).

Если фрагмент найден в терминологическом словаре, то искомый документ представляет собой словарную статью из соответствующего двуязычного словаря.

В том случае, когда фрагмент не найден в базе данных памяти переводчика, в специальном окне об этом выдается сообщение.

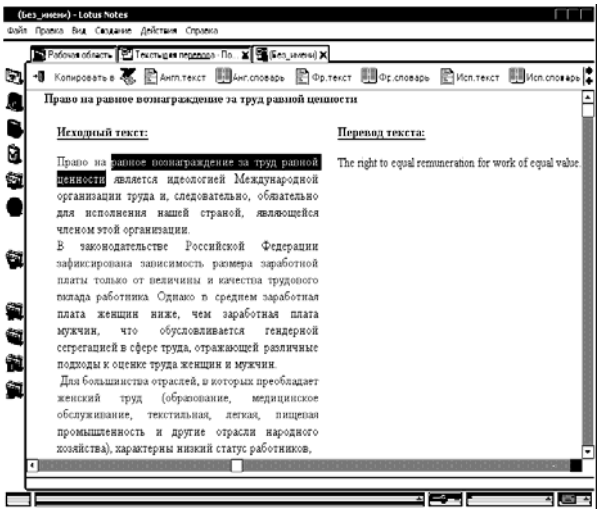



Рис. 4. Представление исходных текстов

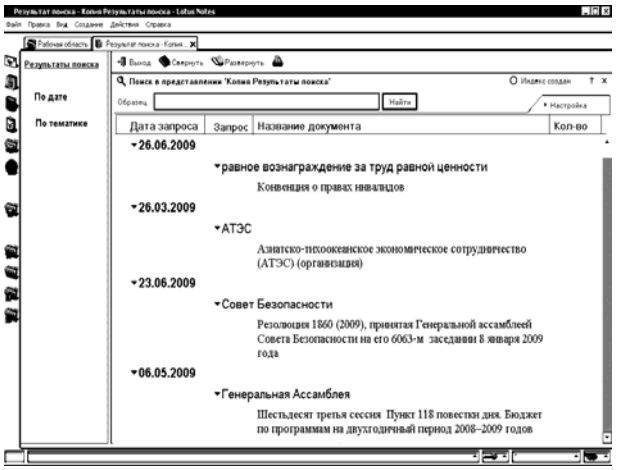



Рис. 5. Результаты поиска переводных эквивалентов

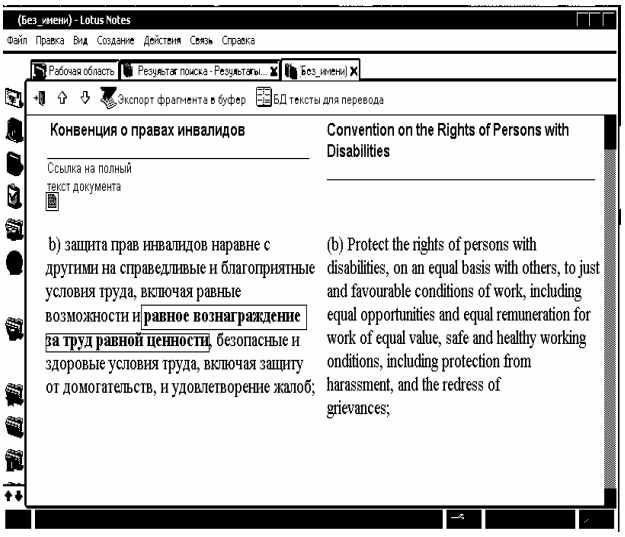



Рис. 6. Абзац из текста документа, содержащий переводные эквиваленты

Переводчик легко может сравнить переводные эквиваленты на двух языках и выделить фрагмент

на иностранном языке, соответствующий искомому фрагменту на русском языке. Нажав на кнопку «БД тексты для перевода», находящуюся на панели действий, он может вернуться в базу данных исходных текстов к документу, который им переводится. Затем в поле «перевод текста» можно вставить выделенный фрагмент на иностранном языке и продолжить процесс перевода.

Пользователь может воспользоваться также стандартными возможностями Lotus Notes, самостоятельно составляя поисковый запрос с учетом близости слов в предложении или абзаце.

## 5 Заключение

Описанная в данной работе экспериментальная система, разработанная на платформе IBM Lotus Domino/Notes, позволяет эффективно накапливать и осуществлять поиск информации в базе данных памяти переводчика. Пользователь может либо выделить любой фрагмент из переводимого текста, либо составить запрос с использованием языка Lotus Notes и провести поиск в терминологическом словаре или в библиотеке параллельных текстов на любом из рабочих языков.

По мере накопления терминологии и текстов в базе данных «Память переводчика» и следуя пожеланиям переводчиков, можно расширить возможности системы, включив функцию нечеткого поиска, которая имеется в некоторых системах, чтобы искать близкие по смыслу тексты, но допускающие некоторое перефразирование.

С этой целью можно использовать тезаурус. В настоящее время мы располагаем тезаурусом по общественно-политической тематике объемом 10 тысяч понятий. Если в исходном тексте, предназначенном для перевода, и текстах на русском языке из библиотеки параллельных текстов провести поиск слов и словосочетаний из тезауруса и найденные словарные единицы заменить на заглавные дескрипторы тезауруса, то можно (при достаточно хорошем покрытии текстов тезаурусом) решать проблему вариативности терминологии. Затем для каждого абзаца исходного текста можно находить в базе памяти переводчика абзацы, отвечающие некоторому критерию близости текстов, и выдавать пользователю соответствующие им абзацы параллельных текстов на нужном языке.

Однако все доработки системы можно будет проводить после некоторого периода эксплуатации системы.

## Литература


[1] Абрамов В.Е. Автоматическое рубрицирование и реферирование текстовой информации (в том числе на иностранных языках). Автореферат дис. канд. техн. наук.- М.: Стандартинформ, 2008, -27 с.

[2] Абрамова Н.Н., Косматова Л.В., Майорова Н.С., Матюшина Н.А., Шелимова И.Н. Многоязычный электронный словарь по внешней политике. – Тез. докл. 6-ой Международ. конф. «НТИ-2002» (Москва, 16-18 октября 2002 г.), 2002. – с.

[3] Белоногов Г. Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. – М.: Русский мир, 2004. – 248 с.

[4] Русина Л. Технические средства в работе переводчика: сравнительная характеристика. – Материалы регионал. конф. ProZ.com (Харьков, 18-19 октября 2008 г.). http://www.proz.com/conference/68?page=schedule&mode=details&session_id=2144

[5] Силонов А. Программы, помогающие переводчику.- PC Week/RF, N13, 2000. – с. 45. http://www.bntp.ru/home.asp?artId=23.

[6] Nagao M. A framework of a mechanical translation between Japanese and English by analogy principle // A. Elithorn, R. Banerji (eds.), Artificial and Human Intelligence, Edinburgh: North-Holland, 1984. Pp. 173-180.

[7] Интернет-сайт компании SDL Trados: http://www. trados.com

[8] Интернет-сайт компании DejaVu: http://www.atril.com

[9] Интернет-сайт компании SDL International: http://www.sdlintl.com

[10] Программа автоматизации перевода Metatexis: www.metatexis.com

[11] Система автоматизированного перевода Transit: http://www.star-portals.net/Transit/default.aspx

[12] Система автоматизированного перевода OmegaT: http://www.omegat.org/ru/reviews.html

[13] Система автоматизированного перевода Termstar: http://www.star-group.net/star-www/description/termstar/star-group/eng/star.html

[14] TM-программа (накопитель переводов) Wordfisher: www. wordfisher.com/wf4.htm


## Corporate translation Network using special digital Libraries


N.N.Abramova, V.E. Abramov, A.A. Karnatskaja, V.M. Rozhkov



The article describes an IBM Lotus Domino/Notes corporate translation network interconnecting several automated translator workstations (AWS). The development of the system was aimed at automating the translation process to meet the modern quality and speed standards of translation. The system is based on the translator memory principles when a great body of manually translated texts in four languages (Russian, English, French and Spanish) is used. Components of the AWS are described and examples of system's functioning are given.

# СЕМАНТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВЫХ КОЛЛЕКЦИЙ

# SEMANTIC ANALYSIS OF TEXT COLLECTIONS

# О возможности борьбы с дубликатами при запросах к разнородным библиографическим источникам♣

© Рубцов Д.Н., Барахнин В.Б.

Институт вычислительных технологий СО РАН,
Новосибирский государственный университет

roubtsov@academ.org, bar@ict.nsc.ru

## Аннотация

При запросах к нескольким разнородным библиографическим источникам возникает проблема выявления повторяющихся записей. В работе проанализированы проблемы, возникающие в процессе установления нечеткого соответствия между двумя записями. Рассмотрены существующие методы и алгоритмы решения задачи исключения дубликатов и, в частности, подходы к определению и вычислению функции похожести строк.

С учетом требований конкретной задачи – усовершенствования информационной системы "Научные сотрудники - математики СО РАН" - реализован метод решения, основанный на использовании в качестве функции похожести наибольшей общей подпоследовательности двух строк. Метод был протестирован на трёх базах данных публикаций СО РАН – Базе данных публикаций журнала "Вычислительные технологии", Базе данных публикаций сотрудников Института вычислительных технологий СО РАН и Базе данных публикаций системы "Web-ресурсы математического содержания". По итогам проведённого тестирования метод продемонстрировал высокую эффективность работы и был применён для системы "Научные сотрудники – математики СО РАН" и разрабатываемой в данный момент интегрированной системы удалённого доступа к разнородным ресурсам библиографической тематики.

## 1 Введение

При запросах к нескольким разнородным источникам зачастую возникает проблема повторяющихся записей, когда два различных источника содержат документы, описывающие один и тот же объект (сущность) реального мира. В информационных системах, работающих с библиографическими описаниями публикаций научной тематики, вероятность возникновения такой ситуации существенно повышается.

Так как библиографические информационные системы, как правило, разрабатываются и поддерживаются независимо, и в каждом конкретном случае разработчики руководствуются своими собственными подходами, то записи, относящиеся к одним и тем же документам, могут быть представлены по-разному. В частности, такие записи могут иметь различную степень полноты или не соответствовать друг другу по причине опечаток создателей записей. В результате этого может возникнуть неоднородность как на уровне модели и схемы данных, так и на уровне самих элементов данных. [10]

Для решения задачи интеграции разнородных источников возникает необходимость сопоставления, согласования и объединения различных представлений данных, а также исключения дублирующейся информации.

Процесс выявления и исключения дублирующейся информации может производиться как над двумя источниками одновременно, так и над уже интегрированным набором данных. Можно выделить следующие этапы:

- Приведение документов (записей), полученных из разнородных источников, к единой схеме данных;
- выявление (т.е. сопоставление) похожих записей, относящихся к одному и тому же объекту реального окружения;
- объединение похожих записей в одну, содержащую все соответствующие атрибуты без избыточности;
- удаление избыточных записей, содержащих менее полную информацию.

В данной работе мы рассматриваем алгоритм решения задачи исключения дублирующих записей, получаемых при запросах к разнородным библиографическим базам данных научной

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

тематики. При этом в понятие «библиографическая база данных» мы вкладываем более широкий смысл, чем «база данных, ведущаяся профессиональными библиографами», подразумевая, что речь может идти и о тех или иных библиографических списках, составленных без строгого соблюдения библиотечных стандартов.

В разрабатываемой системе «Научные сотрудники - математики СО РАН», приходится иметь дело с несколькими источниками, содержащими библиографическую информацию о публикациях сотрудников имеющими различную структуру схемы данных. Публикации интегрируются из трёх баз данных MySQL:

- База данных публикаций журнала «Вычислительные технологии»;
- База данных публикаций сотрудников Института вычислительных технологий СО РАН;
- База данных публикаций системы «Web-ресурсы математического содержания».

Главными атрибутами объекта «научная публикация» являются название, список авторов и другие выходные данные публикации, а также некоторая дополнительная информация (веб-ссылка на полный текст или, по крайней мере, аннотацию к статье и др.). Можно выделить основные проблемы, возникающие на уровне элемента данных:

- орфографические ошибки, транспозиции символов, измененный порядок слов и т.д.;
- несогласованность в написании фамилии автора;
- случай полного совпадения фамилии, имени и отчества двух авторов;
- другие проблемы, связанные с предметной областью (прим. - первая и вторая части одной и той же статьи, опубликованные раздельно).

Заметим, что дополнительная информация о публикациях для некоторых записей из вышеперечисленных источников может, вообще говоря, и отсутствовать. В частности, некоторые записи могут не содержать ссылок как на аннотацию, так и на полный текст статьи. Это сразу же ограничивает нас в применении алгоритмов сравнения документов, работающих с полными текстами. В нашем случае мы можем использовать для установления соответствия между записями только главные атрибуты, являющимися по своей сути строками, а также некоторые выходные данные публикации. Учитывая вышесказанное, первостепенной задачей требующей решения для обнаружения дубликатов публикаций становится выбор функции похожести или метрики для установления нечёткого соответствия двух строк.

## 2 Функции похожести строк и алгоритмы их вычисления

Расстояние между двумя объектами может быть вычислено с помощью различных мер близости, которые называют также метриками. Чем меньше это расстояние, тем более похожими считаются объекты сравнения.

Понятие метрики широко используется в различных областях, к примеру, в распознавании образов (букв, речи, изображений, лиц и т.д.).

Одними из наиболее часто встречающихся метрик для подсчёта расстояния в *n*-мерном пространстве являются меры Хемминга («манхэттенское расстояние»)

$$\sum_i |x_i - y_i|$$

и Евклида

$$\sqrt{\sum_i (x_i - y_i)^2} .$$

Для сравнения строк обычно используют метрики, оценивающие минимальное количество действий (операция редактирования), необходимых для преобразования одной строки в другую. К элементарным операциям редактирования относятся операции замены, вставки и удаления символа, последние две из которых иногда объединяют в одну.

Существует множество различных подходов к выбору функции похожести строк. Одной из классических мер является расстояние Левенштейна (также дистанция Левенштейна, функция Левенштейна, алгоритм Левенштейна). Согласно работам [2,6] функция Левенштейна – это мера разницы двух последовательностей символов (строк) относительно минимального числа элементарных операций редактирования, необходимых для перевода одной строки в другую в случае, когда операции имеют одинаковый вес. Существует также модификация расстояния Левенштейна – расстояние Левенштейна – Дамерау, где в множество элементарных операций включены транспозиции символов. При этом требуется, чтобы к транспонированным символам не применялись другие операции редактирования.

Если придать единичный вес удалению и вставке и удвоенный вес замене, мы получим «расстояние редактирования». Разрешив только операцию замены с единичным весом, мы приходим к расстоянию Хемминга, которое определяется, как количество позиций, в которых строки содержат различные символы. Оно пригодно для определения расстояния только в тех случаях, когда сравниваемые строки имеют одинаковую длину. В случае, когда разрешены только операции удаления и вставки с весом, равным единице, мы можем вычислить меру, которую называют наибольшей общей подпоследовательностью двух строк (LCS – Longest Common Subsequence).

Расстояние Джаро – Винклера [8] определяется по формуле:

$$d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right).$$

Здесь $m$ – число совпадающих символов, $s_1$ и $s_2$ – длины сравниваемых строк, а $t$ – число перестановок. Два символа считаются совпадающими, когда расстояние между ними не превышает

$$L = \left\lfloor \max\frac{(|s_1|,|s_2|)}{2} \right\rfloor - 1.$$

Каждый символ первой строки сравнивается со всеми совпадающими с ним символами второй строки. Число перестановок определяется, как число совпадающих, но идущих в неверном порядке символов, поделённое на два. Существует также модифицированный метод Джаро – Винклера, использующий веса отличные от 1/3.

Распознавание Рэтклиффа – Обершелпа [11] подсчитывает похожесть двух строк, как число совпадающих символов, поделённое на общее число символов в обеих строках. Совпадающие символы определяются в виде наибольшей общей последовательности, а также совпадающих символов в остальной части по каждую сторону от наибольшей общей подпоследовательности.

Среди других подходов можно выделить алгоритмы проверки похожести звучания слов с помощью фонетического кодирования (Soundex, Metaphone, NYSIIS и др.) [13]. Обычно, такие алгоритмы языкозависимы, и плохо работают в случае, когда строки различаются в первом символе или содержат пробелы.

Ряд подходов также основан на сопоставлении лексем (схожесть Джаккарда и др.). В них работа ведётся с векторной моделью документов, а текст представляется в виде набора слов. В некоторых случаях вместо слов в качестве лексем выступают $n$-граммы (общие подстроки фиксированной длины $n$). Основным недостатком этих методов, как правило, является не слишком высокая эффективность работы при сравнении коротких строк или при наличии орфографических ошибок в словах [5].

Расстояние Левенштейна может быть вычислено с помощью метода динамического программирования Вагнера – Фишера [14]. Идея метода состоит в том, чтобы последовательно оценивать расстояния между удлиняющимися на каждом шаге префиксами строк до получения окончательного результата. Промежуточные результаты вычисляются итеративно и хранятся в массиве размерности $(m+1)\times(n+1)$, что приводит к затратам времени и памяти $O(mn)$, где $m$ и $n$ – длины сравниваемых строк. Для нахождения значения расстояния требуется вычислить $mn$ элементов матрицы динамического программирования.

Согласно Смиту и Ватерману [12] можно дополнить понятие «расстояния редактирования», введя учёт пропуска последовательностей символов. Полученное расстояние обычно называется обобщённым расстоянием редактирования с аффинным штрафом за пропуски, и может быть вычислено с помощью метода динамического программирования использующего три матрицы. Однако сложность алгоритма в этом случае возрастает до $O(m^2 n^2)$. Такая же сложность возникает и при вычислении расстояния Левенштейна – Дамерау.

На основе метода динамического программирования было разработано множество алгоритмов, в частности алгоритмы Хиршберга, Ханта и Мак-Илроя, Ханта и Шиманского, Машека и Патерсона, Укконена и Майерса и других. Более подробное описание и исследование этих алгоритмов можно найти в работах [9,6].

Также для подсчёта длины наибольшей общей подпоследовательности существует ряд алгоритмов, основанных на использовании бит-векторов (AD, CIPR и другие) [7]. Они позволяют получить результат за время $O(mn/\omega)$, где $\omega$ – размер используемого алфавита.

## 3 Существующие решения

На данный момент существует достаточно большое число публикаций, посвящённых проблеме дублирующихся записей. Как правило, выбор решения в каждом конкретном случае сильно зависит от особенностей предметной области и постановки задачи.

В подавляющей части встретившихся нам работ для сравнения строк используется стандартная метрика Левенштейна, а также не производится никакой предварительной обработки данных.

В работе Цыганова Н. Л. [4] решается задача нечеткого сопоставления записей баз данных персонала. На первом этапе производится предварительная кластеризация данных, после чего применяются алгоритмы нечеткого сопоставления строк (значений) для всех полей записи, по которым осуществляется поиск. Для полей имен используется метод вычисления обобщенного расстояния редактирования с аффинным штрафом за пропуски, а для остальных полей - вычисление схожести, основанной на совпадении лексем. В заключении, строится вектор схожести отдельных полей для вычисления результирующей схожести записей.

В работе Гула А.Ю., Игнатенко А.П., Чадюка А.В. [3] рассматривается методика идентификации физических и юридических лиц в хранилищах данных. Плюсом работы является предложенный алгоритм нормализации данных, включающий в себя унификацию как структуры данных (загрузка данных из разных источников в таблицу единого формата), так и самих данных (перевод строк в

верхний регистр, удаление непечатных и повторяющихся символов, удаление пробелов в начале и конце строки). Для сравнения записей в случае юридических лиц предлагается использовать алгоритм сопоставления биграмм, а в случае физических лиц – модифицированное расстояние Левенштейна.

Помимо частных решений на данный момент существует ряд программных пакетов для очистки данных, в которых реализованы средства выявления и удаления дубликатов (DataCleanser DataBlade, ETI Data Cleanse, Integrity, Centrus Merge/Purge и другие). Ими поддерживаются различные подходы к согласованию атрибутов. Некоторые из них также позволяют интегрировать правила согласования, определённые пользователем.

## 4 Выбор алгоритма и его реализация

Рассмотрим более подробно процесс выявления и исключения дублирующихся публикаций для web-ресурса «Научные сотрудники - математики СО РАН». На первом этапе происходит интеграция данных из трёх источников, которые были перечислены выше. Информация извлекается с помощью метаданных. Ввиду того, что источники имеют различную структуру схемы данных, в местах, где это необходимо, производится слияние или расщепление соответствующих атрибутов. После этого приведённая к единообразному виду информация заносится во временную таблицу, над которой происходит весь дальнейший процесс оперирования данными. Временная таблица содержит следующие поля:

- Authors – Авторы
- Title – Название статьи (публикации)
- URL – Ссылка (на аннотацию к публикации)
- Description – Дополнительная информация о публикации (источник публикации)
- Year – Год публикации
- Priority – Приоритет

Название научной статьи по своей сути уникально, что позволяет добиться достаточно небольшого риска ошибочного определения дубликата. Однако использование только данного атрибута оказалось недостаточным для успешного выявления дубликатов. Как пример можно привести учебные пособия («Математический анализ» и другие), названия которых часто совпадают, при том, что в остальных атрибутах могут наблюдаться различия. Поэтому, для эффективного определения дубликатов мы использовали совокупное сравнение по нескольким атрибутам, основные из которых: список авторов и название статьи.

Исходя из априорной информации о большом количестве абсолютно идентичных записей, на предварительном этапе записи проверяются на полное совпадение по каждому из атрибутов и, при достижении положительного результата, автоматически классифицируются как дубликаты. В противном случае алгоритм переходит к проведению нечёткого сравнения.

Перед непосредственным сравнением двух строк на этапе предобработки данных происходит преобразование букв с акцентами, а также перевод обеих строк в нижний регистр. Это позволяет улучшить качество получаемого результата.

Пороговый показатель сходства для каждого из основных атрибутов подбирался путём тестирования как на реальных базах данных, так и на специально сгенерированной базе данных, содержащей всевозможные ошибки, и составляет 60% для атрибута 'Authors' и 80% для атрибута 'Title'. Было установлено, что для используемых баз данных эти показатели являются оптимальными для решения поставленной задачи – выявления всех дублирующихся записей. Во многом, этот результат достигается именно за счёт совокупного использования нескольких атрибутов [1]. При варьировании же установленных показателей, возможен пропуск некоторых дублирующихся записей, а также возникновение «лишних» пар дубликатов.

Для вычисления длины наибольшей общей подпоследовательность двух строк для сопоставления записей была выбрана одна из модификаций метода динамического программирования, предложенная Хиршбергом. Выбор данного метода был обусловлен достаточной эффективностью и относительной простотой реализации.

Затраты алгоритма относительно памяти и времени вычисления составляют соответственно $O(m+n)$ и $O(mn)$, где $m$ и $n$ - длины сравниваемых строк. Алгоритм реализован с помощью метода динамического программирования, основанного на рекурсии, на каждом шаге определяются длины наибольших общих подпоследовательностей у всё более и более длинных префиксов строк.

Обозначим их как $l(i,j)$, то есть:

$$l(i,j) = \left| lcs((x(1,i), y(1,j)) \right|.$$

Здесь функция $lcs(x,y)$ подсчитывает наибольшую общую подпоследовательность строк $x$ и $y$ соответственно. Так как длина наибольшей общей подпоследовательности любой строки и пустой равна нулю, значения границ массива задаются как $l(i,0) = l(0,j) = 0$. В позиции $(i,j)$, то есть когда рассматриваются префиксы $x(1,i)$ и $y(1,j)$, если $x_i = y_j$, мы получаем новое значение функции $lcs$, присоединяя этот символ к текущему значению $lcs$ префиксов $x(1,i-1)$ и $y(1,j-1)$, откуда $l(i,j) = l(i-1,j-1) + 1$. Иначе текущее значение $lcs$ берётся в виде максимума из предыдущих соседних значений: $l(i,j) = \max\{l(i-1, j), l(i, j-1)\}$.

Заметим, что для вычисления строки $i$ требуется только строка $i-1$. Для удобства введем вектор $ll(j) = l(m,j)$. Используется массив $h$ длины $2(n+1)$, в котором нулевая и первая строки

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| LCS | 99.10 | 98.19 | 98.73 | 80 | 100 | 83.3 | 74.19 | 80.52 | 96.55 | 98.67 |
| Levenshtein(PHP) | 98.20 | 96.36 | 97.47 | - | 100 | 70 | - | 76.62 | 93.10 | 97.33 |
| Similar_text(PHP) | 99.10 | 98.19 | 98.73 | 80 | 100 | 80 | 67.74 | 80.52 | 96.55 | 98.67 |

Таблица 1. Результаты сравнения записей по параметру ‘Authors’.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| LCS | 100 | 100 | 100 | 100 | 99.55 | 100 | 100 | 100 | 100 | 100 |
| Levenshtein(PHP) | 100 | 100 | 100 | 100 | 99.10 | 100 | 100 | 100 | 100 | 100 |
| Similar_text(PHP) | 100 | 100 | 100 | 100 | 99.55 | 100 | 100 | 100 | 100 | 100 |

Таблица 2. Результаты сравнения записей по параметру ‘Title’.

выступают в качестве строк *i*-1 и *i* массива *l*, соответственно.

Граничные условия по *j* от 0 до *n* задаются, как $h(1,j) = 0$.

Перед вычислением каждой новой “строки *i*” первая строка сдвигается вверх на место нулевой строки. Для этого используется цикл по *i* от 1 до *m* и по *j* от 0 до *n*, в котором $h(0,j)$ присваивается значение $h(1,j)$.

По *j* от 1 до *n* в позиции $(i,j)$ при $x_i = y_j$ полагаем $h(1,j) = h(0,1) + 1$. В противном случае полагаем $h(1,j) = \max\{h(1,j-1), h(0,j)\}$.

На последнем этапе по всем *j* от 0 до *n* происходит копирование результата $h(1,j)$ в выходной вектор $ll(j)$.

В таблицах 1 и 2 отражены результаты сравнения для десяти пар дубликатов по параметрам ‘Authors’ и ‘Title’ соответственно. Помимо разработанного алгоритма (LCS), в результаты теста также включены расчёты для двух стандартных функций PHP – Levenshtein и Similar_text.

Из полученных процентных значений видно, что большинство ошибок приходится на атрибут ‘Authors’, при практически стопроцентном соответствии заголовков статьи. Такая ситуация обусловлена различным представлением списка авторов в различных базах данных и, как следствие, возможными ошибками при интеграции. Кроме того, в некоторых случаях этот список может оказаться неполным.

Для разработанного алгоритма полученные результаты практически совпадают с результатами для функции Similar_text. Однако в столбцах 6 и 7 более высокий результат был получен за счёт лучшей обработки нашим алгоритмом ситуаций, когда расположение авторов в списке оказывается различным (случай перестановки слов). Таким образом, разработанный алгоритм продемонстрировал наилучшую эффективность работы.

Записи, для которых показатели сходства по каждому из основных атрибутов превышают пороговое значение, рассматриваются как потенциальные дубликаты, после чего происходит сопоставление по дополнительным атрибутам, таким как год публикации. При отсутствии информации о дополнительных атрибутах (пропущенные значения) записи трактуются как различные.

В случаях, когда приходиться иметь дело с разными частями одной и той же статьи или книги, вышеперечисленных методов может оказаться недостаточно для получения ответа на вопрос, являются ли две сравниваемые записи дубликатами или нет. В качестве примера приведём две следующие записи:

1) В. А. Ильин, В. А. Садовничий, Бл. Х. Сендов «Математический анализ. Часть 1», 2006
2) В. А. Ильин, В. А. Садовничий, Бл. Х. Сендов «Математический анализ. Часть 2», 2006

В данном примере, при полном совпадении параметров 'Authors' и 'Year', различие заключается только в параметре 'Title', при этом степень сходства очевидно превышает выбранный пороговый показатель, вследствие чего при отсутствии дополнительной проверки записи могут быть ошибочно определены как дубликаты. Для обработки таких уникальных случаев, используется алгоритм поиска всевозможных вхождений вида “(1)”, “[1]”, “Часть 1”, “Часть первая” и других, что позволяет избежать описанной выше ошибки.

Описанная стратегия обеспечивает выявление подавляющего числа дублирующихся записей в рамках решаемой задачи. По завершению процесса выявления дубликатов, из результата запроса исключаются дублирующиеся записи, содержащие менее полную информацию. Этот процесс происходит в соответствии с выставленными приоритетами (атрибут ‘Priority’). В нашем случае удалось единственным образом упорядочить источники по полноте, таким образом при выводе предпочтение отдается источникам, содержащим более полную информацию.

Предложенный алгоритм был применен при разработке web-ресурса «Научные сотрудники - математики СО РАН» [15], являющегося частью системы «Web-ресурсы математического содержания». Ресурс отражает информацию о научных сотрудниках – математиках СО РАН, а также ссылки на их научные труды и публикации.

## 5 Заключение

Был проведен анализ проблем и подходов к их решению в задаче исключения дублирующихся записей при одновременном запросе к нескольким библиографическим каталогам. На основе проведенного анализа реализован алгоритм исключения дублирующихся записей.

Также было проведено тестирование алгоритма на реальных базах данных публикаций СО РАН – Базе данных публикаций журнала «Вычислительные технологии», Базе данных публикаций сотрудников Института вычислительных технологий СО РАН и Базе данных публикаций системы «Web-ресурсы математического содержания».

В ходе тестирования были определены оптимальные параметры, необходимые для эффективной работы алгоритма. Алгоритм был применён для web-ресурса «Научные сотрудники - математики СО РАН».

## Литература


[1] Барахнин В.Б., Нехаева В.А., Федотов А.М. О задании меры сходства для кластеризации текстовых документов // Вестник НГУ. Сер. Информационные технологии. 2008. Т. 6. Вып. 1. С. 3-9.

[2] Бойцов Л.М. Классификация и экспериментальное исследование современных алгоритмов нечеткого словарного поиска // Труды Всероссийской конференции RCDL'2004. http://www.rcdl.ru//papers/2004/paper27.pdf

[3] Гула А.Ю., Игнатенко А.П., Чадюк А.В. "Задача идентификации физических и юридических лиц в хранилищах данных"// Шестая международная конференция по программированию УкрПРОГ'2008, 27-29 мая 2008 года, Киев, Украина. http://eprints.isofts.kiev.ua/416/

[4] Цыганов Н.Л. Методика поиска дублирующихся записей с помощью алгоритма нечеткого сопоставления строк // Научная сессия МИФИ - 2007. Сборник научных трудов. М.: МИФИ, 2007. Т.2: Технологии разработки программных систем. Информационные технологии. С. 159-160.

[5] Цыганов Н.Л., Циканин М.А. Исследование методов поиска дубликатов веб-документов с учетом запроса пользователя // Интернет-математика 2007: Сб. работ участников конкурса. Екатеринбург: Изд-во Урал. ун-та, 2007. С. 211-222.

[6] Graham A. "String Search" Technical Report TR-92-gas-01, School of Electronic Engineering Science, University College of North Wales (пер. Галкиной М.С., под ред. Дубнера П.Н.) http://infoscope.ws/string_search/Stephen-92/index.html

[7] Hyyro H. Bit-parallel LCS-length computation revisited // Proc. 15th Australasian Workshop on Combinatorial Algorithms (AWOCA 2004), 2004. http://www.cs.uta.fi/~helmu/pubs/pubs.html

[8] Jaro-Winkler distance http://en.wikipedia.org/wiki/Jaro-Winkler_distance

[9] Navarro G. A Guided Tour to Approximate String Matching. ACM Computing Surveys. 2001. V.33(1). P. 31-88.

[10] Rahm E., Hai Do H. Data Cleaning: Problems and Current Approaches // IEEE Data Engineering Bulletin. 2000. V. 23(4): P. 3-13.

[11] Ratcliff J., Metzener D. Pattern Matching: The Gestalt Approach // Dr. Dobb's Journal, page 46, July 1988.

[12] Smith T.F. Identification of Common Molecular Subsequences // Journal of Molecular Biology. 1981. V.147: P. 195-197.

[13] Soundex http://en.wikipedia.org/wiki/Soundex

[14] Wagner R.A., Fisher M.J. // The String to String Correction Problem. Journal of the ACM. 1974. V 21(1). P. 168-173.

[15] Web-ресурс "Научные сотрудники - математики СО РАН" http://pine.ict.nsc.ru/sbras/math_soran/


## On the possibility of duplicates struggle when performing queries to heterogeneous bibliographic sources


D.N. Roubtsov, V.B. Barakhnin



When performing queries to multiple heterogeneous bibliographic sources the problem of repetitive records arises. The problems appearing in the process of detection of fuzzy match between two records are analyzed in this paper. The existing methods and algorithms of duplicate elimination and in particular the approaches to determination and calculation of string similarity function are considered.

Taking into account the requirements of the concrete task of modernization of the information system "Mathematicians of SB RAS" the solution method was realized based on the use of longest common subsequence as string similarity function. The proposed method was tested on three SB RAS databases - Database of publications of Journal "Computational Technologies", Database of publications of employees of The Institute of Computational Technologies SB RAS and Database of publications of "Web-resources of the mathematical content". The method showed high efficiency on results of the testing and was applied for the information system "Mathematicians of SB RAS" and the integrated system of remote access to the heterogenous bibliographic resources which is being developed at the present moment.



♣ Работа выполнена при частичной поддержке РФФИ: проекты 07-07-00271, 08-07-00229, 09-07-00277, президентской программы «Ведущие научные школы РФ» (грант № НШ-931.2008.9) и интеграционных проектов СО РАН.

# Тематическое упорядочение текстов при формировании сводных документов

© Васильев Виталий Геннадьевич

ООО «ЛАН-ПРОЕКТ»
vvg_2000@mail.ru

**Аннотация**

В работе рассматривается новый подход к автоматизации процессов подготовки сводных документов, основанный на тематическом упорядочении текстов. Проводится описание и сравнительный анализ различных методов решения данной задачи на различных тестовых массивах.

## Введение

В настоящее время для автоматизации обработки потоков текстовых данных часто используются средства автоматической классификации текстов. Например, с их помощью осуществляется подготовка различных сводных и обобщенных справок путем отбора сообщений по интересующим тематикам. При формировании такого типа документов приходится сталкиваться с наличием повторяющейся информации и неупорядоченностью расположения тематически близких текстов в списках результатов. Возможная технология автоматизированного формирования сводных документов учетом решения указанных задач приведена на рис. 1.

Необходимо отметить, что так как в рамках приведенной технологии упорядочивание документов производится в рамках отдельной рубрики, то общее количество обрабатываемых документов оказывается относительно небольшим (порядка нескольких сотен документов). Данное свойство позволяет рассматривать более широкий спектр методов для решения задачи упорядочения документов

Реализация приведенной технологии также требует решения ряда дополнительных задач (сбор, классификация, выявление дубликатов документов), которые не рассматриваются в настоящей работе. С описанием применяемых подходов для решения ряда из них можно ознакомиться в дополнительной литературе. В частности, описание используемых методов классификации и выделения значимых фрагментов в текстах приводится в [17, 16].

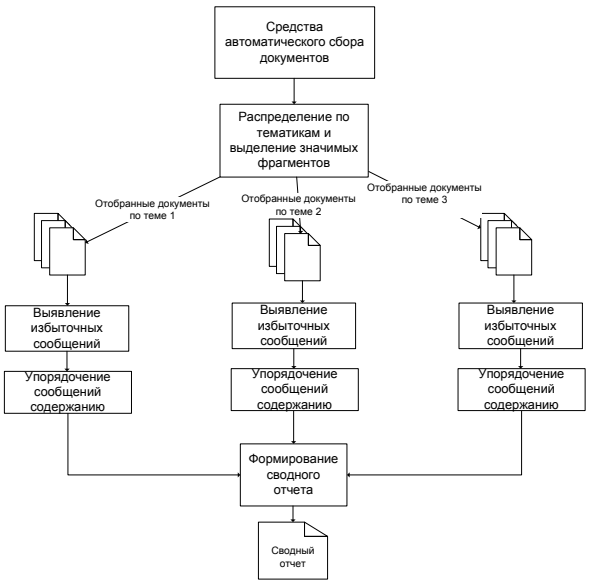



**Рис. 1.** Технология формирования сводных документов

Работа имеет следующую структуру. В первом разделе дается краткое описание моделей представления и мер близости текстов. Во втором разделе рассматриваются различные формальные постановки задачи тематического упорядочения документов и методы их решения. В третьем разделе приводятся результаты экспериментального исследования методов на различных тестовых массивах. В заключении приводятся общие выводы по результатам работы и описываются перспективные направления дальнейших исследований.

## 1 Модели представления и меры близости текстов

Существенным моментом при построении алгоритмов тематического упорядочения и выявления повторяющейся информации в текстах является выбор модели представления и меры близости текстов. При решении указанных задач различными авторами применяются как стандартные теоретико-множественные модели, используемые при решении задач поиска и классификации, так и специфические модели,

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

основанные на построении цифровых отпечатков, которые активно используются при выявлении дубликатов и копий документов.

В стандартных теоретико-множественных моделях в качестве информационных признаков обычно используются отдельные слова и словосочетания, а в качестве весов слов используются различные производные функции от следующих характеристик:

- частоты встречаемости слова в тексте,
- числа документов, содержащих слово,
- положения слова в тексте,
- присутствие слова в специальных словарях.

Методы выделения информационных признаков и вычисления их весов достаточно подробно описаны в литературе [17]. При этом наиболее распространенной схемой вычисления весов является TF-IDF, при использовании которой вес слова $t_j$, $j=1,...,m$, в тексте $d_i$, $i=1,...,n$ определяется как $w_{ij}=f_{ij}\log\left(\frac{n}{n_j}\right)$, где $n$ - общее число текстов в массиве, $f_{ij}$ - частота встречаемости слова $t_j$ в тексте $d_i$.

В моделях на основе цифровых отпечатков в качестве информационных признаков используются хэш-коды различных элементов текстов. В последнем случае удается получать более компактное представление текстов, а вычисление близости между текстами производить путем простого сравнения хэш-кодов. Рассмотрим более подробно методы, применяемые для построения цифровых отпечатков. Данные методы можно условно разбить на следующие виды:

- синтаксические - текст описывается с помощью набора хэш-кодов для цепочек последовательно идущих слов [1, 2, 6],
- лексические - текст описывается с помощью хэш-кодов для наборов слов, которые входят в некоторое множество [8, 10].

Одной из первых работ по синтаксическому подходу является [2]. В ней предлагается представлять текст в виде множества хэш-кодов всех последовательностей соседних слов длины $k$, которые названы шинглами. Поскольку число шинглов обычно достаточно большое (примерно равно числу слов в документе), то для сокращения их числа авторами работы предлагается несколько эвристических подходов:

- отбираются шинглы равные 0 по некоторому модулю;
- отбирается шинглов с минимальными значениями хэш-кодов;
- отобранные шинглы объединяются в группы - мегашинглы.

В ряде работ предлагается использовать также и другие подходы к выбору последовательностей в текстах. В работах [1, 18, 12] в качестве элементов для вычисления хэш-кодов рассматриваются последовательности слов входящие в отдельные предложения, в $k$ соседних предложений с перекрытием и без, в блоки предложений заканчивающиеся предложениями с хэш-кодом равным нулю по некоторому модулю, а также в текст целиком. В частности, в работе [13] текст представляется в виде последовательности хэш-кодов $k$-грамм символов. Для ее построения предлагается лексический алгоритм Winnowing, в котором по последовательности хэш-кодов $h_1,...,h_n$ всех $k$-грамм перемещается окно размера $w$, где $n$ - число $k$-грамм выделенных в тексте. В каждом окне $h_i,...,h_{i+w-1}$, $i=1,...,n-w+1$, отбирается один элемент с минимальным значением.

В работах [5, 8, 10] используется лексический подход к построению цифровых отпечатков. Он основан на построении функции, которая осуществляет отображение вектора весов признаков $x$ размерности $m$ в битовый вектор $z=(z_1,...,z_f)$ ($f$ обычно 32 или 64) следующим образом:

$$z_i=\begin{cases}0, & y_i<0,\\ 1, & y_i\geq 0,\end{cases} \qquad (1)$$

где $y_i=\sum_{j=1}^{m}x_j(-1)^{s_{ji}}$, $s_j=(s_{j1},...,s_{jf})$ - битовый вектор длины $f$ для информационного признака с номером $j=1,...,m$, полученный с помощью стандартной хэш-функции (MD5 или SHA-1). Особенностью получаемого вектора $z$ является то, что близкие вектора признаков получают близкие значения данного вектора.

В работах [4, 18] рассматривается алгоритм Imatch, который также основан на лексическом подходе. В нем из текста документа отбираются заданный процент таких терминов, которые не встречаются в слишком большом или в слишком маленьком числе текстов обрабатываемого массива. Цифрой отпечаток получается путем вычисления хэш-функции SHA-1 от строки, составленной из отсортированного списка отобранных терминов.

Для вычисления степени близости текстов $x$ и $y$ наибольшее распространение получили следующие меры:

$sim_{jaccard}(x,y)=\frac{|S(x)\cap S(y)|}{|S(x)\cup S(y)|}$ - мера Жаккарда, используется при описании текста в виде множества хэш-кодов [2, 1, 6, 18], где $S(x)$ и $S(y)$ - множества информационных признаков в текстах $x$ и $y$;

$sim_{contain}(x,y)=\frac{|S(x)\cap S(y)|}{|S(x)|}$ - мера включения $x$ в $y$, также обычно используется при описании текста в виде множества хэш-кодов [2], где $S(x)$ и $S(y)$ - множества информационных признаков в текстах $x$ и $y$;

$sim_{\cos ine}(x, y) = \frac{x^T y}{\|x\|_2 \|y\|_2}$ - косинусная мера близости, используется при описании текста в виде вектора весов информационных признаков [9], где $x = (x_1, ..., x_m)$ - вектор весов признаков в тексте $x$ и $y = (y_1, ..., y_m)$ - вектор весов признаков в тексте $y$, $m$ - общее число различных признаков во всех документах;

$sim_{ham\min g}(x, y) = \sum_{i=1}^{m} |x_i - y_i|$ - мера Хэмминга, используется при представлении текстов в виде битовых векторов в лексических моделях цифровых отпечатков [8, 10, 11], где $x_i \in \{0,1\}$ - битовый признак с номером $i = 1,...,m$ у текста $x$, $y_i$ - битовый признак с номером $i = 1,...,m$ у текста $y$.

## 2. Методы тематического упорядочения текстов

Задачу тематического упорядочения текстов заключается в нахождении такой перестановки заданного множества текстов, что тематически близкие тексты будут находиться рядом друг с другом, а тематически отличные - далеко. Исходя из приведенного определения, она является близкой по своему содержанию к задаче коммивояжера, задача одномерного размещения элементов, задаче иерархического кластерного анализа, а также задачам ранжирования документов при поиске текстов.

Рассмотрим сначала использования методов решения задач коммивояжера и одномерного размещения элементов [7, 15] для тематического упорядочения текстов.

В рамках задачи коммивояжера задача тематического упорядочения текстов формально определяется следующим образом. Требуется найти такую перестановку $(l_1, ..., l_n)$ номеров текстов $\{1, ..., n\}$, на которой достигается максимум функции

$$C_{TSP}(l_1, ..., l_n) = \sum_{i=2}^{n} sim(x_{l_{i-1}}, x_{l_i}),$$

где $sim(x, y)$ - мера близости между текстами $x$ и $y$.

В рамках задачи одномерного размещения элементов формальная постановка задачи тематического упорядочения будет следующая. Требуется найти такую перестановку $(l_1, ..., l_n)$ номеров текстов $\{1, ..., n\}$, на которой достигается минимум функции

$$C_{PPP}(l_1, ..., l_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} |j - i| sim(x_{l_i}, x_{l_j})$$

Обе приведенные задачи являются NP-полными, что приводит к необходимости использования приближенных алгоритмов для возможности обработке наборов из нескольких сотен текстов. Приближенные алгоритмы можно разбить на следующие типы (в скобках указаны названия наиболее распространенных алгоритмов):

"жадные" алгоритмы – основаны на последовательном построении решения путем выбора локально-оптимального решения на каждом шаге (алгоритм ближайшего соседа);

методы локальной оптимизации – основаны на итерационном изменении решения с помощью преобразований из заданного множества до тех пока нельзя будет улучшить имеющееся решение (метод двойного выбора, метод тройного выбора, алгоритм Лина-Кернигана и д.р.);

методы случайного поиска – основаны на использовании методов статистического моделирования (генетические алгоритмы, прямое моделирование "Монте-Карло", эволюционные алгоритмы, моделирование отжига).

В настоящей работе остановимся на использовании наиболее быстрых приближенных методов, вычислительная сложность которых не более $O(n^3)$:

- метод ближайшего соседа;
- метод двойного выбора;
- метод перестановки смежных элементов;
- генетический алгоритм.

В методе ближайшего соседа получение решения $(l_1, ..., l_n)$ осуществляется следующим образом. Номер первого текста $l_1$ произвольным образом (обычно $l_1 = 1$). Номер $l_{i+1}$ текста определяется путем нахождения ближайшего текста к $l_i$ среди тех текстов, номера которых не присутствуют во множестве $(l_1, ..., l_i)$.

В методе двойного выбора для нахождения решения задачи коммивояжера сначала строят случайную перестановку текстов $(l_1^0, ..., l_n^0)$, которая рассматривается как замкнутый путь (гамильтонов цикл) в графе, у которого вершинами являются тексты. Далее рассматривают все возможные пары не смежных ребер $(l_a, l_{a+1})$, $(l_b, l_{b+1})$ и производится их замена на ребра $(l_a, l_b)$ и $(l_{a+1}, l_{b+1})$, если это приводит к улучшению решения. Данная процедура повторяется до тех пор, пока нельзя будет улучшить имеющее решение.

В методе перестановке смежных элементов начальное решение строится случайным образом. Текущее решение $(l_1, ..., l_n)$ модифицируется путем перестановки соседних элементов $l_i$ и $l_{i+1}$, если новый порядок имеет меньшую стоимость. Пусть $L_j = \sum_{s=1}^{j-1} sim(x_{l_s}, x_{l_j})$ - сумма весов элементов расположенных слева от элемента $l_j$,

$R_j = \sum_{s=j+1}^{n} sim\left(x_{l_s}, x_{l_j}\right)$ - сумма весов элементов расположенных справа от $l_j$. Перестановку $l_i$ и $l_{i+1}$ можно выполнить, если $L_i - R_i + R_{i+1} - L_{i+1} + 2sim\left(x_{l_i}, x_{l_{i+1}}\right) < 0$.

В методе на основе генетического алгоритма [19] на первом шаге формируется набор (популяция) $T_1$ из некоторого числа случайных перестановок текстов. На последующих шагах из имеющегося набора $T_k$ производится формирование нового набора перестановок $T_{k+1}$ путем объединения и преобразования имеющихся перестановок с помощью операторов скрещивания и мутации, и отбора перестановок с максимальным значением целевой функции.

Рассмотрим теперь использование агломеративных методов иерархического кластерного анализа [17] для решения задачи тематического упорядочения текстов. В данных методах производится последовательное объединение имеющего набора текстов $x_1,...,x_n$ во все более крупные классы. Схема типичного агломеративного алгоритма имеет следующий вид.

**Схема агломеративного иерархического алгоритма**

1. Положить $t = 0$, $\Omega^{(0)} = \left\{\omega_1^{(0)}, ..., \omega_n^{(0)}\right\}$, где $\omega_i^{(0)} = \{x_i\}$, $\delta_{ij}^{(0)} = \rho\left(x_i, x_j\right)$, $i, j = 1, ..., n$.
2. Построить разбиение $\Omega^{(t+1)}$ путем объединения классов $\omega_l^{(t)}$ и $\omega_r^{(t)}$ в разбиении $\Omega^{(t)}$, где $(l,r) = \arg \min_{i,j=1,...,n-t,\, i \neq j} \delta\left(\omega_i^{(t)}, \omega_j^{(t)}\right)$.
3. Если $t = n-1$, то завершить работу алгоритма, в противном случае положить $t = t+1$ и перейти к шагу 2. ■

Для нахождения приближенного решения воспользуемся тем фактом, что на каждом шаге $t = 1, ..., n-1$ работы агломеративного алгоритма разбиение $\Omega^{(t+1)}$ получается путем объединения двух ближайших классов $\omega_s^{(t)}$ и $\omega_r^{(t)}$ в разбиении $\Omega^{(t)}$, т. е. $(s,r) = \arg \min_{i,j=1,...,n-t,\, i \neq j} \delta\left(\omega_i^{(t)}, \omega_j^{(t)}\right)$, где $\delta$ – функция расстояния между классами.

В качестве приближенного решения будем использовать такую перестановку $\left(l_1^*, ..., l_n^*\right)$, в которой для любого $t = 1, ..., n-1$ элементы классов $\omega_s^{(t)}$ и $\omega_r^{(t)}$, которые входят в один больший класс, располагаются непосредственно друг за другом. Причем порядок следования элементов классов $\omega_s^{(t)}$ и $\omega_r^{(t)}$ должен быть таким, чтобы обеспечить максимум функции $C\left(l_1, ..., l_n\right)$.

Рассмотрим теперь рекурсивную процедуру нахождения перестановки $L^* = \left(l_1^*, ..., l_n^*\right)$. Пусть для элементов классов $\omega_s^{(t)}$ и $\omega_r^{(t)}$, объединяемых на шаге $t$, уже построены перестановки $L_s^* = \left(l_{s1}^*, ..., l_{sn_s}^*\right)$ и $L_r^* = \left(l_{r1}^*, ..., l_{rn_r}^*\right)$ соответственно, где $n_s$ – число элементов в классе $\omega_s^{(t)}$, $n_r$ – число элементов в классе $\omega_r^{(t)}$. Для произвольной перестановки $\left(u_1, ..., u_n\right)$ оператор $R(u,p)$, $p \in \{0,1\}$, определяется следующим образом:

$R\left(\left(u_1, ..., u_n\right), 0\right) \equiv \left(u_1, ..., u_n\right)$,

$R\left(\left(u_1, ..., u_n\right), 1\right) \equiv \left(u_n, u_{n-1}, ..., u_1\right)$.

Перестановка $L_{s\cup r}^*$ для объединенного класса $\omega_{l\cup r} = \omega_l \cup \omega_r$ получается следующим образом:

$\left(p_s^*, p_r^*\right) = \arg \max_{p_s, p_r \in \{0,1\}} sim\left(R_{n_s}\left(L_s^*, p_s\right), R_1\left(L_r^*, p_r\right)\right)$,

$L_{s\cup r}^* = \left(R\left(L_s^*, p_s^*\right), R\left(L_r^*, p_r^*\right)\right)$.

Несложно заметить, что вычислительная сложность данной процедуры с учетом проведения иерархического кластерного анализа составляет $O\left(n^2\right)$, где $n$ – число текстов.

Рассмотрим теперь использование методов ранжирования текстов, применяемых в поисковых системах Интернет. Одним из наиболее распространенных методов является PageRank [3]. На его основе рядом авторов предложены алгоритмы для упорядочения текстов из произвольного множества по степени близости к некоторому эталонному тексту или набору текстов [14, 20]. В частности, в работе [14] предлагается следующий алгоритм.

Пусть $W$ - матрица размера $n \times n$, где $w_{ij}$ - мера близости текстов с номерами $i$ и $j$, $D = diag\left(d_1, ..., d_n\right)$ - диагональная матрица, $d_i = \sum_{j=1}^{n} w_{ij}$, $S = D^{-1/2} W D^{-1/2}$, $f^{(0)} = \left(f_1^{(0)}, ..., f_n^{(0)}\right)$ - вектор весов текстов, задающий начальное ранжирование текстов, $\alpha \in (0,1)$ - параметр, $y = \left(y_1, ..., y_n\right)$ - вектор эталонных текстов (в настоящей работе полагаем, что $y = (1,0,...,0)$, т.е. упорядочение производится относительно первого текста), $t_{\max}$ - число итераций.

Вектор итогового ранжирования текстов $f^* = \left(f_1, ..., f_n\right)$ находится с помощью следующей итерационной процедуры:

$f(t+1) = \alpha S f(t) + (1-\alpha) y$, $t = 1, ..., t_{\max}$.

Несложно показать, что последовательность весов текстов $f(t)$ сходится к $f^* = (1-\alpha)(I - \alpha S)^{-1} y$. Для получения итоговой перестановки текстов производится упорядочение элементов вектора $f^*$ по убыванию.

## 3. Эксперименты

Для экспериментального исследования эффективности различных подходов к решению задачи тематического упорядочения текстов были использованы массивы, приведенные в следующей таблице.

**Таблица** 1. Тестовые массивы для оценки тематического упорядочения

| Название массива | Число текстов | Число рубрик | Комментарий |
|---|---|---|---|
| Yandex News | 256 | 21 | Массив построен путем ручного отбора и коррекции документов из тематически различных сюжетов в системе Яндекс Новости. Результаты автоматической группировки новостей в данном случае не учитывались. |
| Google News | 511 | 24 | Массив построен путем ручного отбора групп новостей по определенным тематикам с помощью поисковой системы Google. Группы организованы в виде двухуровневого дерева. Например, в группу сюжетов "НАТО Афганистан" попали сюжеты про события в населенных пунктах Гельменд, Лагман, Саидхель и др. Результаты автоматической группировки новостей в данном случае не учитывались. |
| Reuters 21578-6 | 935 | 6 | Подмножество из 6 рубрик ("gold", "gnp", "gas", "nat-gas", "ship", "sugar") массива Reuters-21578. |
| ROMIP 2004 | 2000 | 173 | Обучающая выборка для дорожки классификации нормативно-правовых документов РОМИП 2004 (первые 2000 текстов). |
| Reuters 21578 | 5000 | 142 | Массив Reuters-21578, http://www.daviddlewis.com (первые 5000 текстов). |

Для оценки качества работы алгоритмов с точки зрения конечного пользователя произведем оценку расположение эталонных номеров классов текстов в итоговой перестановке с помощью следующего показателя

$$C_{\text{var}}(l_1,...,l_n)=\sum_{j=1}^{k}\left(\frac{1}{n_j}\sum_{i\in\omega_j}\left(l_i-\bar{l}_j\right)^2\right)$$ - суммарная дисперсия, которая представляет собой сумму разбросов элементов эталонных классов $\omega_1,...,\omega_k$ в полученной перестановке $(l_1,...,l_n)$, где $n_j$ - число элементов в классе $\omega_j$, $k$ - число эталонных классов. Заметим, что минимальное значение $C_{\text{var}}(l_1,...,l_n)$ достигается в том случае, когда тексты из одного класса расположены рядом. Так как $\sum_{i=1}^{n_j} i^2=\frac{1}{6}n_j(n_j+1)(2n_j+1)$, получаем, что для произвольной перестановки $(l_1,...,l_n)$ справедливо следующее неравенство

$$C_{\text{var}}(l_1,...,l_n)\geq\frac{1}{12}\sum_{j=1}^{k}\left(n_j^2-1\right).$$

В результате можно определить следующий относительный показатель «нормированная дисперсия»

$$C(l_1,...,l_n)=\frac{\sum_{j=1}^{r}\left(n_j^2-1\right)}{12\sum_{j=1}^{r}\left(\frac{1}{n_j}\sum_{i\in\omega_j}\left(l_i-\bar{l}_j\right)^2\right)},$$

который принимает значения из промежутка от 0 до 1.

Рассмотрим теперь результаты экспериментов. В табл. 2 приводятся оценки времени работы алгоритмов на тестовых массивах. При этом используются следующие обозначения: Hier – алгоритм на основе иерархического кластерного анализа, 2-opt – алгоритм на основе двойного выбора, NN – алгоритм ближайшего соседа, GA – генетический алгоритм, PPP – метод перестановки смежных элементов, Rank – алгоритм упорядочения текстов PageRank. Для представления текстов использовалась теоретико-множественная модель, а для вычисления расстояний косинусная мера близости. Необходимо отметить, что время вычисления расстояний между всеми парами текстов в данном случае не учитывалось.

**Таблица** 2. Время работы (сек)

| Время (сек) | HIER | 2OPT | Rank | NN | GA | PP |
|---|---|---|---|---|---|---|
| Yandex News | 0.06 | 0.08 | 0.02 | 0.03 | 19 | 2.7 |
| Google News | 0.1 | 0.65 | 0.1 | 0.09 | 34 | 3.5 |
| Reuters 21578_6 | 0.22 | 4.72 | 0.36 | 0.24 | 64 | 8.67 |
| Romip 2004 | 1.7 | 86 | 2.5 | 1.8 | 133 | 90 |
| Reuters 21578 | 13 | >1500 | 36 | 6 | 398 | >1500 |

В табл. 3 приводятся оценки показателя "нормированная дисперсия" для различных алгоритмов, моделей представления текстов и методов вычисления расстояний между текстами. В таблице используются следующие обозначения:

2OPT_TAIL – модификация алгоритма 2OPT, в которой производится циклический сдвиг полученной перестановки элементов таким образом, чтобы концевые элементы $l_1$ и $l_n$ были максимально не похожи друг на друга;

TFIDF – для представления текстов используется стандартная теоретико-множественная модель, а для вычисления близости используется косинусная мера;

KGRAMM – текст представляется хэш-кодами всех последовательностей слов длины $k$, в качестве меры близости используется мера включения;

KGRAMM TFIDF – используется комбинированная мера близости, получаемая в результате объединения матриц с косинусной мерой, вычисленной на основе стандартных теоретико-множественных признаков, и мерой включения, вычисленной на основе к-грамм слов.

Для удобства анализа результатов в таблице в каждом столбце жирным шрифтом выделено наилучшее значение показателя, а курсивом следующие два наилучших показателя.

**Таблица** 3. Нормированная дисперсия

| **Алгоритм** | **Модель** | **Yandex News** | **Google News** | **Reuters 21578_6** | **Romip** |
|---|---|---|---|---|---|
| PPP | TFIDF | 0.0176 | 0.0148 | 0.0375 | 0.0014 |
| | KGRAMM TFIDF | *0.0246* | 0.0186 | 0.0375 | 0.0014 |
| 2OPT | TFIDF | 0.0184 | 0.0163 | 0.0377 | 0.0015 |
| | KGRAMM | 0.0034 | 0.0104 | 0.0359 | 0.0014 |
| | KGRAMM TFIDF | *0.0254* | 0.0189 | 0.0377 | 0.0015 |
| 2OPT TAIL | TFIDF | **0.0350** | **0.0979** | 0.0387 | 0.0015 |
| | KGRAMM | 0.0036 | 0.0088 | 0.0372 | 0.0014 |
| HIER | KGRAMM | 0.0048 | 0.0148 | 0.0363 | 0.0015 |
| | TFIDF | 0.0160 | *0.0519* | **0.0427** | 0.0015 |
| | KGRAMM TFIDF | 0.0158 | 0.0492 | 0.0427 | 0.0015 |
| NN | KGRAMM | 0.0037 | 0.0104 | 0.0348 | **0.0016** |
| | TFIDF | 0.0072 | 0.0506 | 0.0392 | 0.0013 |
| | KGRAMM TFIDF | 0.0081 | *0.0515* | *0.0406* | 0.0015 |
| RANK | TFIDF | 0.0049 | 0.0372 | *0.0417* | 0.0016 |
| GA | TFIDF | 0.0087 | 0.0089 | 0.0347 | 0.0014 |

Таким образом, проведенные эксперименты показывают, что наиболее эффективным с точки зрения качества работы являются методы двойного выбора (2OPT-TAIL) и на основе иерархического кластерного анализа (HIER). Наилучшие результаты, как правило, достигаются при использовании косинусной меры близости, либо комбинированной косинусной и к-граммной меры близости (меры включения). Методы, применяемые для ранжирования результатов поиска в поисковых системах Интернет в данном случае оказываются малопригодными.

## Выводы и направления дальнейших исследований

Таким образом, в настоящей работе проведен сравнительный анализ эффективности использования различных методов для решения задач тематического упорядочения и устранения избыточной информации в текстах. Разработаны два новых алгоритма тематического упорядочения текстов, один основанный на использовании иерархических методов кластерного анализа, а другой основанные на модификации алгоритма приближенного решения задачи коммивояжера.

В качестве перспективных направлении дальнейших исследований можно выделить следующие:

- анализ эффективности процедур тематического упорядочения для обработки других типов документов: электронной почты, законодательных актов, служебных документов, научных работ и т.п.
- анализ эффективности использования других методов для тематического упорядочения текстов (например, методов на основе нейронных сетей Кохонена, методов проецирования данных на плоскость, методов ранжирования страниц в поисковых системах Интернет);
- анализ эффективности использования в алгоритмах устранения избыточной информации результатов выделения значимых фрагментов при классификации текстов;
- разработка эффективных алгоритмов устранения повторяющихся фрагментов в различных текстах из заданного массива.

Работа выполнена при поддержке гранта Президента Российской Федерации для государственной поддержки молодых российских ученых МК-12.2008.10.

## Литература

[1] Brin S., Davis J., Garcна-Molina H. Copy detection mechanisms for digital documents // SIGMOD Rec. 24, 2, 1995. - 398-409. DOI= http://doi.acm.org/10.1145/568271.223855

[2] Broder A., Glassman S., Manasse M., Zweig G. Syntactic Clustering of the Web // DEC SRC Technical Note 1997-015, 1997. – 13 p.

[3] Brin S., Page L. The anatomy of a large scale hypertextual web search engine. In *Proc. 7th International World Wide Web Conf.*, 1998.

[4] Chowdhury A., Frieder O., Grossman D.A., McCabe M.C., Collection statistics for fast duplicate document detection // ACM Transactions on Information Systems, 20(2), 2002. – pp. 171-191

[5] Charikar M. Similarity estimation techniques from rounding algorithms // Proc. 34th Annual Symposium on Theory of Computing (STOC 2002). pp. 380-388.

[6] Forman G., Eshghi K., Chiocchetti S. Finding Similar Files in Large Document Repositories // KDD'05, 2005. – 8 p.

[7] Gutin G., Punnen A.P. The Traveling Salesman Problem and Its Variations (Combinatorial Optimization). Kluwer Academic Publishers, 2004. – 850 p.

[8] Henzinger M. Finding near-duplicate web pages: a large-scale evaluation of algorithms // Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, 2006. – pp. 284-291. DOI= http://doi.acm.org/10.1145/1148170.1148222

[9] Hoad T. C., Zobel J. Methods for identifying versioned and plagiarized documents // J. Am. Soc. Inf. Sci. Technol. 54, 3, 2003. - 203-215. DOI= http://dx.doi.org/10.1002/asi.10170
[10] Manku G. S., Jain A., Das Sarma A. Detecting near-duplicates for web crawling // Proceedings of the 16th international Conference on World Wide Web, 2007. – pp. 141-150. DOI= http://doi.acm.org/10.1145/1242572.1242592
[11] Seo J., Croft W. Local Text Reuse Detection // SIGIR'08, July 20–24, 2008, Singapore. – 8 p.
[12] Shivakumar N., Garcia-Molina H. Finding near-replicas of documents on the web // Proceedings of Workshop on Web Databases (WebDB'98), 1998. pp. 204-212.
[13] Schleimer, S., D.S. Wilkerson and A. Aiken "Winnowing: Local Algorithms for Document Fingerprinting", SIGMOD 2003, Jun. 9-12, 2003.
[14] D. Zhou, J. Weston, A. Gretton, O. Bousquet and B. Schlkopf. Ranking on data manifolds. In Proceedings of NIPS'2003.
[15] Ахо А.В., Хопкрофт Д.Э., Ульман Д.Д. Структуры данных и алгоритмы. – М. Вильямс, 2007. – 384 с.
[16] Васильев В.Г. Комплексная технология автоматической класси-фикации текстов. Компьютерная лингвистика и интеллек-туальные технологии: по материалам ежегодной международ-ной конференции "Диалог" (Бекасово, 4-8 июня, 2008). Вып. 7(14). – М. РГГУ. ISBN 978-5-7281-1022-4. с. 83-90.
[17] Васильев В.Г., Кривенко М.П. Методы автоматизированной обработки текстов. – М. ИПИ РАН, 2008. – 304 с.
[18] Зеленков Ю.Г, Сегалович И.В Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2007, Переславль-Залесский, Россия, 2007. – 9р.
[19] Корнеенко В.П. Методы оптимизации. Учебник. – М.: Высшая школа, 2007. – 664 с.
[20] Тарасов С.Д. Автоматическое составление обзорных рефератов новостных сюжетов // Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

## Thematical arrangement of texts for creating digests

© Vitaly Vasilyev
vvg_2000@mail.ru

In this work new approach for digest creating is suggested. It is based on thematical arranging texts collection in such way that thematically similar texts are placed close to each other in the resulting list of documents. Different methods for solving this task are proposed and experimentally evaluated on wide range of collections.

# Поиск неестественных текстов

© Е.А. Гречников, Г.Г. Гусев, А.А. Кустарев, А.М. Райгородский

Яндекс, Лаборатория комбинаторных и вероятностных методов,
{grechnik, gleb57, kustarev, raigorodsky}@yandex-team.ru

## Аннотация

В работе описывается метод определения неестественного происхождения документа, основанный на изучении статистики встречаемости пар соседних слов в тексте. Тестирование показывает, что метод может быть использован как отдельно, так и для существенного улучшения результатов уже известных методов определения спама по контенту.

## 1 Введение

### 1.1 Постановка задачи

Требуется построить алгоритм, определяющий, написан ли данный документ человеком или же является автоматически сгенерированным либо модифицированным. Под модификацией документа понимается следующее:

- текст является результатом работы синонимайзера – программы, заменяющей отдельные слова на синонимы, или иной системы уникализации контента;
- текст является результатом работы автоматического переводчика с иностранного языка на русский.

Эта задача актуальна, в первую очередь, для проблемы нахождения поискового спама. Многие сайты используют системы уникализации контента и бессмысленную накачку документов ключевыми словами для повышения собственных позиций в поисковой выдаче.

Отличить автоматически сгенерированные или модифицированные тексты от написанных человеком на глаз обычно несложно. Трудность заключается в поиске автоматического алгоритма решения задачи.

Говоря менее формально, нужно на машинном уровне научиться определять степень «бредовости» текста. Идея решения состоит в исследовании корреляций соседних слов в исходном документе.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

### 1.2 Работы по схожей тематике

Методы определения поискового спама можно довольно грубо разделить на две части: анализ контента самой страницы и анализ входящих и исходящих ссылок, топологии сети в окрестности документа. Предлагаемый нами метод, очевидно, относится к первой группе.

В ряде статей неоднократно рассматривались методы обнаружения спама, основанные на анализе содержания страницы. В работе [5] был построен классификатор спама, использовавший несколько таких признаков, как сжимаемость документа, средняя вероятность триграмм, доля частых слов в документе и другие – которые затем были собраны в дерево решений при помощи алгоритма машинного обучения C4.5[6].

Довольно близко к нашей работе по теме и методу находится работа [1], в которой анализируются пары слов, находящихся на данном расстоянии друг от друга, после чего по всем таким парам строится общая гистограмма частоты.

В работе [2] анализировались частоты встречаемости в тексте слов, имевших большую коммерческую и рекламную привлекательность.

Наконец, в работе [4] метод определения спама основан на сравнении моделей языка в исходном документе и в документе, его цитирующем – то есть находится на пересечении двух рассматриваемых нами областей.

## 2 Описание метода

Мы будем работать с множеством 2000 наиболее распространенных слов русского языка. Рассмотрим матрицу $(A_{ij})$ размера 2000 на 2000, в которой на пересечении $i$-й строки и $j$-го столбца стоит частота встречаемости в языке пары слов с номерами $i$ и $j$. Частота встречаемости пары вычисляется по фиксированной базе текстов, в данном случае использовалась база *ruscorpora,* объем которой – 41298 документов [8].

Пусть $A_i$ и $A_j$ – суммы по строкам и столбцам матрицы $(A_{ij})$ соответственно. Определим функцию $Cor(i, j)$ по формуле $Cor(i, j) = (A_{ij}/A_i) + (A_{ij}/A_j)$. Функция Cor(i, j) измеряет степень «сочетаемости» слов с порядковыми номерами $i$ и $j$.

Была выдвинута следующая гипотеза: в неестественном тексте должно быть нарушено распределение пар в тексте по функции Cor. Более

точно, количество редких, нехарактерных для языка пар должно быть завышено по сравнению со стандартом, а количество частых пар – занижено.

Эта гипотеза подтверждается следующей таблицей. В ней четыре столбца с числами соответствуют четырем текстам – один является оригинальной газетной статьей, а три других – его синонимизированными модификациями. Были использованы три различных программы-синонимайзера, найденных в Интернете.

| | | | | |
|---|---|---|---|---|
| Cor≥0.1 | 115 | 92 | 87 | 76 |
| 0.1>Cor≥0.01 | 502 | 350 | 317 | 309 |
| 0.01>Cor≥0.001 | 341 | 291 | 219 | 290 |
| 0.001>Cor≥0.0001 | 98 | 148 | 73 | 159 |
| 0.0001>Cor≥0.00001 | 12 | 18 | 19 | 39 |
| 0.00001>Cor | 2 | 3 | 2 | 6 |

Основываясь на данных таблицы, объявим теперь пары с $Cor<10^{-4}$ редкими, а пары с $Cor>10^{-2}$ – частыми и заметим, что в данном примере искомая гипотеза тогда подтверждается. Стоит также отметить, что во всех других рассмотренных нами случаях (всего было аналогичным образом проанализировано 45 текстов) число редких пар в синонимизированном дубликате оказалось завышенным по сравнению с исходным текстом.

## 3 Применение и результаты

Мы приводим два метода использования полученных данных. В первом число редких пар в тестируемом тексте сравнивается с данными, полученными из заведомо хороших текстов, после чего делается вывод о качестве тестируемого документа. Второй метод использует машинное обучение при помощи алгоритма TreeNet [3,7], в котором в качестве факторов используется число пар в тексте, Cor которых лежит в том или ином диапазоне. Мы также сравниваем эффективность наших факторов с классическими факторами, использованными в работе [5] для определения спама по контенту страницы.

### 3.1 Сравнение с нормальными текстами

Рассмотрим базу заведомо качественных и возможно более разнообразных текстов русского языка (в данном случае использовалась база *ruscorpora* [7]). Если на вход дан тестируемый текст T, то найдем 10 ближайших к нему по длине текстов из базы качественных документов. Затем определим три числа: N(T) – число редких пар в тексте T, M(T) – среднее арифметическое числа редких пар в 10 ближайших по длине к T текстах, D(T) – квадратный корень дисперсии набора чисел редких пар из 10 ближайших к T текстов. В этом случае число (N(T) – M(T))/D(T) измеряет «степень неестественности» текста T.

Тестирование проводилось на выборке из 165 текстов, полученных вручную с помощью трех синонимайзеров из Интернета, а также 41298 оригинальных текстов из базы *ruscorpora (*для контроля ошибки метода). При тестировании текста из базы *ruscorpora* обязательно выбирались 10 текстов, ближайших к нему по длине, но не совпадающих с ним.

Результаты применения этого метода оказались следующие: условие (N(T) – M(T))/D(T)>3 позволяло корректно идентифицировать 41.5% некачественных текстов. При этом ошибка на базе текстов *ruscorpora* составила немногим более 2.3% (978 текстов из 41298 были признаны неестественными).

### 3.2 Машинное обучение

Гораздо лучшие результаты были получены при применении машинного обучения, позволившего подобрать нужную формулу для проверки качества текста автоматически.

Тестирование проводилось на обучающей и тестовой выборках, не содержащих совпадающих текстов. Обучающая выборка состояла из 2000 заведомо оригинальных документов, а также из 250 некачественных документов, из которых 25 автоматически сгенерированных и 25 синонимизированных были найдены в Интернете, а 200 были изготовлены вручную при помощи трех различных программ-синонимайзеров. Тестовая выборка состояла из 500 заведомо оригинальных и 245 некачественных текстов, из которых 25 автоматически сгенерированных и 25 синонимизированных были найдены в Интернете, а 145 были изготовлены вручную при помощи трех программ-синонимайзеров.

В качестве алгоритма машинного обучения был выбран TreeNet ([3],[7]) с потенциалом «сумма логарифмов вероятностей ошибки классификации». Число итераций алгоритма выбиралось с тем, чтобы минимизировать ошибку на тестовом множестве. В качестве других параметров были выбраны: шаг регуляризации – 0.01, доля обучающей выборки, по которой шло обучение на каждом шаге – 0.5. Результатом работы TreeNet являлась не бинарная классификация, а число, определяющее вероятность нахождения элемента в данном из двух классов.

В качестве факторов использовалось число пар в тексте, Cor которых лежал в данном диапазоне: от 0 до $10^{-7}$, от $10^{-7}$ до $10^{-6}$ и так далее до диапазона Cor>1. По окончанию работы TreeNet фиксировалось два порога, соответствующих ошибке в 1% и 5% на тестовом подмножестве из 500 оригинальных документов. Затем для этих порогов рассматривалась полнота на соответствующем подмножестве неестественных документов.

Результаты оказались следующие:

- при ошибке в 1% было правильно опознано 77.95% неестественных текстов,

- при ошибке в 5% было правильно опознано 90.61% неестественных текстов.

Для сравнения эффективности метода с методами, уже описанными в литературе, были реализованы 10 контентных факторов, описанных в статье [5]. Для машинного обучения снова использовался алгоритм TreeNet. Результаты классического метода таковы:

- при ошибке в 1% было правильно опознано 90.61% неестественных текстов,
- при ошибке в 5% было правильно опознано 96.73% неестественных текстов.

После этого был проведен тест на «совместную эффективность»: к классическим 10 факторам, показавшим хороший результат, были добавлены еще два новых фактора – число «редких» и число «частых» пар в определениях п.2. Результаты оказались следующие:

- при ошибке в 1% было правильно опознано 93.06% неестественных текстов,
- при ошибке в 5% было правильно опознано 97.95% неестественных текстов.

## 4 Заключение

Подводя итог проведенным тестам, можно сделать вывод, что описанный нами метод может успешно использоваться для поиска неестественных текстов. Хотя он существенно проигрывает классическому методу в эффективности, добавление новых факторов позволяет улучшить результаты классического метода на несколько процентов (в нашем случае – на 2.45% и 1.22% соответственно). Иными словами, в проведенном нами тестировании около четверти еще не пойманного классическими методами спама подпадает под сферу действия новых факторов.

## Литература


[1] J. Attenberg, T. Suel. Cleaning search results using term distance features. In *Proceedings of AIRWeb 2008*, pages 21-24, ACM.

[2] A. Benczur, I. Biro, K. Csalogany, and T. Sarlos. Web spam detection via commercial intent analysis. In *Proceedings of AIRWeb 2007*, pages 89-92, New York, NY, USA, 2007, ACM.

[3] J.H.Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29(5):1189-1232, 2000.

[4] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proc. of the 1st Int. Workshop on Adversarial Information Retrieval on the Web*, pages 1-6, 2005.

[5] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. 15th WWW*, pages 83-92, 2006.

[6] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan-Kauffman, 1993.

*[7]* G. Ridgeway. Generalized Boosted Models: A guide to the gbm package. *http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf*

[8] Национальный корпус русского языка – *www.ruscorpora.ru*.


## Detection of Artificial Texts


Evgeny A. Grechnikov, Gleb G. Gusev,
Andrei A. Kustarev, Andrei M. Raigorodsky.



We present a method of artificial text search based on analysis of frequency of word pairs. This method can be used either for improving results of well-known content spam classifiers or independently.

# ВЕБ-ТЕХНОЛОГИИ

# WEB TECHNOLOGIES

# Метод обнаружения поискового спама, порожденного с помощью цепей Маркова

© Павлов А.С.

Факультет Вычислительной математики и кибернетики
МГУ имени М.В.Ломоносова
pavvloff@gmail.com

© Добров Б.В.

Научно-исследовательский вычислительный центр
МГУ имени М.В.Ломоносова;
АНО Центр информационных исследований
dobroff@mail.cir.ru

**Аннотация**

В работе предложен метод обнаружения поискового спама, порожденного генераторами текста. Данный метод основывается на анализе частотных распределений стилистических и жанровых особенностей текста. Для автоматического обнаружения спама на основе выделенных характеристик используются методы машинного обучения. Проведены эксперименты, в которых показана возможность обнаружения текстов, порожденных генераторами на основе цепей Маркова, с помощью предложенного подхода.

## 1 Введение

В связи с большим количеством информации в сети Интернет пользователи чаще всего используют веб поиск для нахождения интересующих их данных. В настоящее время одной из основных проблем информационного поиска является распространение поискового спама.

Поисковый спам создается в результате намеренных действий, направленных на завышение оценки страницы в поисковой системе, по сравнению с ее истинной ценностью. В соответствии с современными оценками поисковый спам составляет около 22% всего содержимого сети Интернет [11]. На настоящий момент поисковый спам остается эффективным методом выведения сайта на верхние позиции в выдаче поисковых систем. Поисковый спам ухудшает качество поиска и мешает нормальной работе поисковых систем.

Поисковый спам наиболее эффективен при массовом автоматическом создании спам-страниц. Одним из распространенных способов автоматического создания большого количества текстов является генерация текстов на основе цепей Маркова. При использовании генерации текстов на основе цепей Маркова сначала на отобранных текстах производится обучение, затем можно породить большое количество в целом бессмысленных, но локально связных текстов. Учитывая то, что в качестве исходных текстов часто берутся релевантные определенной тематике документы, то и результаты генерации текстов также отражают статистические тематические характеристики.

Отметим, что в настоящее время не существует полной теории, описывающей законы порождения связных осмысленных текстов. Как следствие отсутствуют в общем случае методы порождения текстов, не отличимых от созданных человеком.

Тем не менее, известны многие закономерности, характерные естественным текстам - единство стиля, следование законом жанра, локальная связность, глобальная тематическая связность и т.п.

Цепи Маркова позволяют моделировать лишь локальную связность текста и общие тематические характеристики.

Основная идея настоящей работы состоит в том, чтобы с помощью учета статистических характеристик стилистических и жанровых особенностей естественных текстов обнаруживать тексты, обладающие локальной связностью, но нарушающие другие свойства естественных текстов.

## 2 Обзор существующих методов

### 2.1 Методы детектирования поискового спама

Применимость простых статистических характеристик для определения поискового спама изучалась в работе [14]. При этом наибольшее внимание уделялось ссылочным характеристикам, в то время как статистические характеристики текста практически не рассматривались.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

Исследование лингвистических характеристик для обнаружения поискового спама исследовалось в работе [19]. Данный подход основывается на применении специального словаря, на основе которого вычисляется более 200 статистических признаков. Применение словаря означает, что генераторы текстов могут обойти предлагаемые методы.

Еще один интересный подход к анализу содержимого документов для обнаружения поискового спама предлагается в [9]. Эта работа основывается на определении коммерческой направленности текстов по нескольким статистическим атрибутам. Данные атрибуты в основном основаны на анализе логов поисковых запросов или логов систем контекстной рекламы. Использование характеристик, ограниченных некоторым словарем, позволяет генераторам на основе цепей Маркова обходить предложенные алгоритмы.

Подходы, не зависящие от конкретной лексики и тематики документов, предлагаются в работах [18, 21]. Первая работа посвящена обнаружению спама в блогах, и использует особенности формата блогов, например, наличие комментариев, что ограничивает применимость данного метода. Вторая работа основана на анализе стилистических особенностей HTML-кода страниц, в то время как текстовое содержимое не учитывается в принципе.

### 2.2 Использование несловарных характеристик при анализе текстов

Существует постоянный интерес исследователей к анализу статистических, трудно контролируемым автором, характеристик естественного текста.

Метрики читаемости (readability, также употребляется термин «читабельность») текста подробно описаны в [12, 13]. Применение простых статистических характеристик, таких как длина предложений и длина слов, широко используется в США для оценки простоты восприятия текста.

Задача определения жанра текста по простым статистическим характеристикам решается в [10]. В этой работе показано как небольшое число характеристик текста позволяют с неплохой точностью определять наиболее распространенные стили и жанры.

Определение авторства текста, точнее специфического авторского стиля, также может основываться на глобальных статистических закономерностях текстов. Большое количество работ в данном направлении основывается на статье [6]. Анализ статистики употребления частиц, предлогов, а также длин предложений и слов позволяет формулировать критерии принадлежности текста конкретному автору.

### 2.3 Определение дубликатов

Задача обнаружения дубликатов текста является смежной с задачей обнаружения текстов, созданных цепями Маркова, так как в зависимости от длины цепи порожденный текст может копировать большие куски документов-образцов.

Обзор методов обнаружения дубликатов приведен в [3]. В статье [15] описано применение шинглирования для обнаружения спам-текстов, порожденных из отрывков естественных текстов.

## 3 Генераторы текстов на основе цепей Маркова

### 3.1 Методы порождения текстов

Создание поискового спама сопряжено с созданием большого количества текстов для автоматического наполнения сайтов. В настоящий момент существует несколько подходов к созданию текстов для спам-сайтов [16]:

- Создание текстов вручную;
- Копирование текстов из других источников;
- Автоматическая генерация текстов;
- Автоматическая модификация существующих текстов.

Создание текстов вручную является трудоемким и дорогостоящим процессом, поэтому редко применяется для массового поискового спама. Копирование содержимого других сайтов является довольно распространенным явлением, но в настоящее время существуют достаточно эффективные способы определения скопированного текста, например, на основе шинглирования [3].

В итоге на данный момент наиболее эффективными являются методы, которые позволяют автоматически получать уникальные тексты.

Генератор текста — компьютерная программа, способная генерировать последовательности символов, внешне похожие на текст, но при этом, как правило, лишённые смысла. Такие тексты не представляют никакой ценности для пользователей поиска. При генерации текста спамеры также стараются оптимизировать его под некоторый набор запросов, чтобы повысить вероятность попадания сайта с этим содержимым в выдачу поисковой системы.

### 3.2 Цепи Маркова

Распространенным видом генераторов текста являются генераторы текста на основе цепей Маркова. Цепью Маркова с дискретным временем называется последовательность случайных величин, для которой условное распределение каждой величины зависит только от значения предыдущих величин.

Цепь Маркова описывается множеством значений случайных величин, которое называется пространством состояний; а также матрицей переходных вероятностей между состояниями. Матрица переходных состояний определяет вероятность перехода в следующее состояние, с учетом текущего. В случае если матрица

переходных вероятностей не зависит от шага, она называется однородной, именно однородные матрицы чаще всего применяются для порождения текстов.

### 3.2 Порождение поискового спама с помощью цепей Маркова

Когда цепи Маркова применяются для порождения искусственных текстов, пространством состояний становится множество всех слов и знаков препинания. Переходная матрица обычно формируется по некоторому множеству текстов-образцов. По образцу оценивается вероятность порождения нового слова после последовательности уже порожденных слов. Последовательность событий, произведенная такой цепью Маркова, представляет собой набор слов и знаков препинания, внешне напоминающий связный текст.

Важной характеристикой таких генераторов является порядок цепи Маркова – то есть количество слов, учитывающихся при порождении следующего слова. С ростом порядка цепи растет длина локально связных фрагментов текста, в то же время с ростом длины цепи генератор начинает повторять все большие куски исходного текста.

Тексты, созданные с помощью цепей Маркова, обладают рядом свойств, благодаря которым этот метод порождения текстов стал популярен при создании поискового спама. Во-первых, порожденный текст содержит ту же лексику, что и исходный образец. Это позволяет использовать в качестве образца существующие тексты, которые высоко ранжируются поисковыми системами, например, брать образцы текстов из сниппетов поисковых систем, и получать на выходе тексты, оптимизированные под конкретные запросы. Во-вторых, порожденный текст является с высокой вероятностью уникальным. Это затрудняет обнаружение таких текстов методами обнаружения дубликатов.

В качестве примера приведем фрагмент текста, порожденного по данной статье:

«*Генераторы текстов из отдельных документов. Для проверки возможности обнаружения дубликатов Задача определения текстов, порожденных генераторами текстов из рассматриваемых генераторов текстов. Большое количество слов, начинающихся не являющаяся листом, помечена номером признака и авторства. Проверялась эффективность данного метода опорных векторов позволяет формулировать критерии принадлежности спаму или неспаму.*»

Применение автоматических генераторов текстов на основе цепей Маркова часто используется в таком виде спама как дорвеи. Функция дорвея перенаправить пользователя на некоторый целевой сайт, при этом само содержимое такого сайта никакой ценности для пользователя не несет. Дорвеи должны попадать в выдачу по популярным запросам, поэтому эффективное обнаружение такого вида спама может сократить количество спама в выдаче поисковых систем.

## 4 Предлагаемый метод

В основе предлагаемого подхода лежат методы, ранее использовавшиеся для определения жанра текста и авторства.

В рамках данной работы выделяется набор трудно контролируемых автором статистических признаков текстового документа.

Затем, на основе полученных характеристик и машинного обучения строится автоматический классификатор, который позволяет обнаруживать неестественные тексты.

Данный метод основывается на предположении, что по данным характеристикам тексты, полученные с помощью генератора на основе цепей Маркова, будут отличаться от естественных текстов.

### 4.1 Рассматриваемые характеристики

Рассматриваемые характеристики можно условно разделить на следующие пересекающиеся группы:

- Признаки, связанные с читабельностью текста;
- Стилистические особенности текста;
- Жанровые особенности текста [10];
- Глобальные статистические характеристики;
- Морфологические особенности слов текста (для анализа морфологической информации применялся парсер mystem [4]);
- Статистика употребления знаков препинания.

Признаки выбирались из условия, что автору-человеку сложно проконтролировать их значение. Каждому автору и каждому жанру свойственен особенный стиль, который отражается на значении некоторых характеристик.

Предположительно, генераторы текстов на основе цепей Маркова не способны эмулировать некоторые из этих характеристик. Таким образом, поисковый спам можно выделить в качестве отдельного жанра документов, и рассматривать задачу идентификации этого жанра.

Важным свойством выбранных характеристик является то, что они не основываются на тематике или лексике текстов, таким образом, они позволят определять искусственно порожденные тексты вне зависимости от их тематики.

Ниже приведен полный список выделявшихся характеристик, некоторым пунктам соответствует несколько характеристик:

1. Среднее количество слов в предложениях;
2. Среднее количество символов в словах;
3. Среднее количество слогов в слове;
4. Доля слов длиннее 7 символов;

5. Доля слов более чем из 7 слогов;
6. Доля слов из слога;
7. Доля слов из двух слогов;
8. Минимальное количество слогов в одном предложении;
9. Максимальное количество слогов в одном предложении;
10. Количество частиц «бы»;
11. Количество частиц «ну», «вот», «ведь»;
12. Среднее количество знаков пунктуации на предложение;
13. Среднее количество знаков экспрессивной пунктуации («!», «?», «…»);
14. Среднее количество слов, начинающихся с заглавной буквы;
15. Доля различных частей речи:
    a. Доля глаголов среди слов;
    b. Доля прилагательных среди слов;
    c. Доля существительных среди слов;
    d. Доля числительных среди слов;
    e. Доля порядковых числительных среди слов;
    f. Доля наречий среди слов;
    g. Доля частиц среди слов;
    h. Доля предлогов среди слов;
    i. Доля частиц среди слов;
    j. Доля междометий среди слов;
16. Дисперсии количества различных частей (из п.15) речи по предложениям;
17. Доля местоимений первого лица;
18. Доля местоимений второго лица;
19. Доля глаголов по временам:
    a. Доля глаголов настоящего времени;
    b. Доля глаголов прошедшего времени;
    c. Доля глаголов не прошедшего времени;
20. Доля существительных по родам:
    a. Доля существительных мужского рода среди слов и среди существительных;
    b. Доля существительных женского рода среди слов и среди существительных;
    c. Доля существительных среднего рода среди слов и среди существительных;
21. Сжатие текста различными алгоритмами:
    a. bz2;
    b. zlib2;
22. Частотность употребления слов (оценка распределения по гистограмме частотностей).

Всего исследуется 61 характеристика текста. Каждая характеристика представляет собой положительное вещественное число. Таким образом, каждому документу ставится в соответствие вектор признаков из 61 элемента.

### 4.2 Методы машинного обучения

Для определения текстов, созданных с помощью генераторов, предлагается объединить выделенные характеристики в классификатор с помощью алгоритмов машинного обучения.

Предлагается обучать классификатор с использованием тренировочного набора, состоящего из естественных документов и документов, которые предположительно созданы с помощью генераторов текстов. В случае если классификатору удастся построить зависимость между выделяемыми характеристиками и методом порождения документа, данный классификатор можно будет использовать для обнаружения поискового спама.

В данной работе рассматривались два распространенных алгоритма машинного обучения:

- Классификатор на основе машины опорных векторов с использованием линейного ядра SVMLight [17];
- Классификатор на основе деревьев решений Dtree.

#### 4.2.1 Алгоритм машинного обучения SVMLight

Алгоритмы классификации на основе метода опорных векторов строят гиперплоскость, разделяющую разные классы объектов в пространстве признаков. При этом метод опорных векторов позволяет максимизировать зазор между классами, что способствует более качественной классификации.

В данной работе использовалась одна из распространенных реализаций метода опорных векторов SVMLight [17].

#### 4.2.2 Алгоритм машинного обучения DTree

В основе используемого метода лежит алгоритм построения деревьев решений C4.5 [20]. Каждое дерево решений представляет собой двоичное дерево. Каждая вершина, не являющаяся листом, помечена номером признака и значением, по которому происходит разбиение набора документов на две части. Листы дерева помечены вероятностями принадлежности документа спаму или неспаму.

Дерево строится с корня. Вначале, в корень дерева помещается часть тренировочного набора. Затем, в каждом листе выбирается такой признак и такое значение разбиения, которые минимизируют информационную энтропию в наборах, полученных после разбиения. В случае если энтропия в наборах, полученных после разбиения, меньше, чем в исходном наборе, для данного листа строится левые и правые поддеревья, и лист помечается номером соответствующего признака и порогом разбиения. Затем набор распределяется по левому и правому поддереву в соответствии с выбранным разбиением.

После построения дерева для каждого листа вычисляется вероятность того, что документы, попавшие в этот лист, являются спамом или

| | DTree | | | SVMLight | | |
|---|---|---|---|---|---|---|
| | Точность | Полнота | F-мера | Точность | Полнота | F-мера |
| Markov2 | 91,30% | 93,27% | 92,27% | 72,9% | 74,83% | 73,85% |
| Doorway_Su | 92,11% | 89,98% | 91,03% | 69,41% | 68,24% | 68,82% |
| Rusadult | 99,19% | 99,26% | 99,23% | 89,65% | 89,83% | 89,74% |

Таблица 1. Точность и полнота обнаружения текстов, порожденных различными генераторами текстов

неспамом. Для этого документы распределяются по листам построенного дерева, затем для каждого листа вычисляется доли спам и неспам-документов, попавших в данный лист, которые и записывается в лист дерева. Чтобы минимизировать эффект переобучения на тренировочном наборе дерево строится на одной части тренировочного набора, а вероятности вычисляются по другой.

При обучении по одному и тому же набору строится несколько деревьев решений. При построении каждого дерева тренировочный набор произвольным образом делится пополам. Первая половина используется для построения дерева, вторая используется для вычисления вероятностей спама и неспама в каждом листе дерева.

Деревья объединяются в один классификатор с помощью простой процедуры голосования. При классификации документа вычисляется, в какой лист он попадает в каждом дереве. После этого вычисляется сумма вероятностей принадлежности спаму и неспаму по всем деревьям. Документу присваивается та метка, сумма вероятностей которой наибольшая.

# 5 Эксперименты

Эксперимент проводился на коллекции веб страниц ROMIP By.Web [1]. Из документов коллекции была удалена вся HTML разметка, включая содержимое тегов <script>.

В первом эксперименте оценивалась возможность обнаружения текстов, порожденных цепями Маркова, а также популярными генераторами поискового спама.

Во втором эксперименте исследовалась возможность обнаружения реального поискового спама.

## 5.1 Рассматриваемые генераторы текстов

В первом эксперименте было взято три генератора текстов:

- Собственная реализация генератора текста на основе цепей Маркова с длиной цепи 2 (далее - markov2);
- Генератор дорвеев Doorway.su [2] (далее - doorway_su);
- Генератор дорвеев Rusadult [5] (далее - rusadult).

Каждый документ, представленный в виде вектора признаков, может принадлежать одному из классов: {spam, good}. К классу spam относятся документы, порожденные генератором текста, к классу good относятся документы, из набора ROMIP By.Web. Изучалась возможность обнаружения текстов, созданных каждым из рассматриваемых генераторов.

## 5.2 Построение тренировочных наборов

Тренировочные наборы для классификаторов строились следующим образом:

1. Из набора By.Web были взяты 20000 произвольных документов, помеченных как good (GoodSet);
2. Из набора By.Web были взяты другие 20000 произвольных документов (ExampleSet).
3. Используя набор документов ExampleSet в качестве образца для алгоритма генерации текстов, были порождены следующие наборы:
   a. 20000 текстов были порождены алгоритмом markov2 и помечены как спам (SpamMarkov2);
   b. 20000 текстов были порождены генератором дорвеев Doorway.Su и помечены как спам (SpamDoorwaySu);
   c. 20000 текстов были порождены генератором дорвеев Rusadult и помечены как спам (SpamRusadult).

Затем каждый из наборов GoodSet, SpamMarkov2, SpamDoorwaySu и SpamRusadult был разбит пополам – первые 10000 документов из каждого набора использовались в качестве тренировочных наборов, вторые 10000 документов использовались при тестировании построенных классификаторов. Для проверки возможности обнаружения каждого из рассматриваемых генераторов по отдельности было составлено три обучающих множества, каждый состоит из 10000 документов с пометкой good и 10000 документов, порожденных одним из генераторов.

Каждый классификатор тестировался на 10000 документах, помеченных good, и 10000 документах, порожденных соответствующим генератором.

## 5.3 Применение методов машинного обучения

При обучении классификатора SVMLight все параметры брались по умолчанию. После процедуры обучения производился подбор порога классификации таким образом, чтобы сбалансировать точность и полноту классификатора на тренировочном наборе.

| | DTree | | SVMLight | |
|---|---|---|---|---|
| | Полнота по реальному спаму | F-мера | Полнота по реальному спаму | F-мера |
| Markov2 | 65,66% | 91,57% | 44,33% | 71,69% |
| Markov2+RealSpam | 87% | 92,61% | 49,33% | 72,46% |

Таблица 2. Полнота обнаружения реального спама

При классификации с использование деревьев решений для каждого дерева строилось не более 50 вершин, 100 деревьев объединялись в один классификатор голосованием.

В ходе эксперимента измерялась точность, полнота и F1-мера при обнаружении спама. Результаты эксперимента для трех генераторов текстов и двух алгоритмов классификации приведены в таблице 1.

### 5.4 Эксперимент по обнаружению реального спама

Целью второго эксперимента было проверить возможность обнаружения реального спама, при условии отсутствия образцов таких документов. Для этого вручную было отобрано 600 документов, являющихся поисковым спамом, предположительно порожденным с помощью цепей Маркова (RealSpam). Данные документы были обнаружены в Поиске по блогам компании Яндекс [8], по коммерческим запросам.

Обнаруженный вид спама ориентирован на попадание в выдачу по низкочастотным запросам, например:

«*Deo автомобиль Deo*
*Продажа автомашин. Каталог цен! автомобильная акустика Большой выбор, хорошие цены Модель Daewoo Nexia представляет собой последнюю модификацию модели Opel Kadett E. В 1986 году лицензированное производство этого автомобиля началось в автомобильный видеорегистратор Купля-продажа авто в Тюмени Много частных объявлений о продаже автомобилей в Тюмени на Новые и. автомобили.*»[1]

По всей видимости, для порождения данного текста спамеры использовали несколько текстов на автомобильную тематику. Целью спамеров было привлечь посетителей на данную страницу, с расчетом, что те перейдут по ссылкам, размещенным на ней.

В ходе эксперимента было рассмотрено два тренировочных набора:

- Набор, состоящий из 10000 документов из коллекции ROMIP By.Web, и 10000 документов, порожденных алгоритмом markov2;
- Набор, идентичный предыдущему, к которому было добавлено 300 образцов реального спама.

Тестовый набор состоял из 10000 документов из коллекции ROMIP By.Web, 10000 документов, порожденных алгоритмом markov2 и 300 других документов из набора RealSpam.

Также как и в первом эксперименте использовалось два алгоритма машинного обучения: DTree и SVMLight. В ходе эксперимента измерялась полнота обнаружения реального спама в тестовом наборе, а также общая F1-мера классификации. Результаты эксперимента приведены в таблице 2.

## 6 Выводы

На основании проведенных экспериментов можно сделать вывод, что анализ трудно контролируемых автором признаков может быть использован для обнаружения поискового спама, порожденного с помощью различных генераторов текста. Классификатор на основе деревьев решений показывает лучшие результаты при использовании большого количества разнородных характеристик.

На основании полученных результатов можно сделать некоторые выводы относительно алгоритмов порождения текстов, применяемых в рассмотренных генераторах поискового спама.

Скорее всего, генератор Doorway.Su использует цепи Маркова для порождения текстов, так как точность обнаружения такого рода спама близка к точности обнаружения текстов, созданных модельным алгоритмом цепей Маркова.

Результаты второго эксперимента показывают, что, обучаясь на искусственно созданных образцах поискового спама, можно обнаруживать реальный спам. Как следствие, применение данного подхода для обнаружения спама потребует меньше трудозатрат на составление адекватного тренировочного набора.

При этом добавление даже небольшого количества образцов реального спама (в данном случае 300 документов) ведет к значительному улучшению полноты обнаружения данного типа спама, без значительных изменений в общем качестве обнаружения спама.

Подчеркнем, что обнаружение текстов, порожденных цепями Маркова, возможно даже в том случае, когда образцы, по которым они созданы, неизвестны.

[1] Данный пример можно найти по адресу: *http://eltec15.livejournal.com/6698.html*

## 7 Заключение

В данной работе представлен подход к обнаружению поискового спама, созданного с помощью генераторов текста. Подход основан на выделении характеристик текста, применявшихся ранее для задач жанровой классификации и определения авторства.

Проверялась эффективность данного метода при обнаружении текстов, созданных с помощью генератора текстов на основе цепей Маркова, и двух распространенных в российском сегменте Интернета генераторов поискового спама. Эксперименты показали применимость предлагаемого подхода для задачи обнаружения поискового спама.

В дальнейшем планируется исследовать другие подходы к анализу текстов, а также исследовать возможность обнаружения других распространенных видов поискового спама.

## Литература


[1] Веб коллекция BY.Web, http://romip.ru/ru/collections/by.web-2007.html.

[2] Генератор дорвеев Doorway.Su, http://doorway.su/.

[3] Зеленков Ю.Г., Сегалович И.В., Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль, Россия, 2007. – Том 1, С. 166-174.

[4] Парсер mystem http://company.yandex.ru/technology/mystem/.

[5] Серверный генератор дорвеев от RUSADULT.com, http://doorways.rusadult.com/ru/.

[6] Фоменко В.П., Фоменко Т.Г., Авторский инвариант русских литературных текстов, 1981.

[7] Чжун Кай-лай, Однородные цепи Маркова. Перев. с англ. — М.: Мир, 1964. — 425 с.

[8] Яндекс.Поиск по блогам, http://blogs.yandex.ru/.

[9] Benczúr, A. A., Bíró, I., Csalogány, K. and Sárlós, T. Web Spam Detection via Commercial Intent Analysis. In Proceedings of the 3rd international workshop on Adversarial Information Retrieval on the Web, Banff, Alberta, Canada, May 8th, 2007. Pages: 89–92.

[10] Braslavski P. Document Style Recognition Using Shallow Statistical Analysis. In Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP, Nancy, France, 2004, p. 1–9.

[11] Castillo, C., Donato, D., Becchetti, L., Boldi, P., Leonardi, S., Santini, M., Vigna, S. A Reference Collection for Web Spam. ACM SIGIR Forum Volume 40, Issue 2 (December 2006) Pages: 11–24.

[12] Dale, E. and J. S. Chall. 1949. "The concept of readability." Elementary English 26: 23.

[13] Dubay, W.H.. 2004. The Principles of Readability. Costa Mesa, CA: Impact Information

[14] Fetterly, D., Manasse, M., Najork, M. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In Proceedings of WebDB'04, New York, USA, 2004.

[15] Fetterly, D., Manasse, M., Najork, M. Detecting phrase-level duplication on the World Wide Web. In Proceedings of SIGIR'05, pages 170–177, New York, NY, USA, 2005. ACM.

[16] Gyöngyi, Z. and Garcia-Molina H., Web Spam Taxonomy. In Proceedings of AIRWeb 2005, May 2005.

[17] Joachims, T. Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schцlkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[18] Mishne, G., Carmel, D., and Lempel, R. Blocking blog spam with language model disagreement. In Proceedings of AIRWeb 2005, May 2005.

[19] Piskorski, J., Sydow, M., Weiss, D., Exploring Linguistic Features for Web Spam Detection: A Preliminary Study. In Proceedings of the 4th international workshop on Adversarial Information Retrieval on the Web, Beijing, China, Pages 25-28.

[20] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[21] Urvoy T., Chauveau E., Filoche, P. Tracking Web Spam with HTML Style Similarities. ACM Transactions on the Web, Vol. 2, No. 1, Article 3.


## Detecting Web Spam Created With Markov Chains Text Generators


Anton S. Pavlov, Boris V. Dobrov



In this paper we introduce an approach to detection of web spam generated by text generators. This method is based on text style and genre analysis. Machine learning algorithms are applied to automate spam detection, using the extracted features. Experiments prove possibility of proposed method to be applied to detect texts, created by Markov chains text generators.

# Применение метода опорных векторов для обнаружения ссылочного спама

© Шарапов Руслан Владимирович © Шарапова Екатерина Викторовна

Муромский институт (филиал) Владимирского государственного университета
info@vanta.ru

**Аннотация**

В статье рассматриваются подходы к выявлению ссылочного спама методами машинного обучения. Приводится обзор существующих методов борьбы с поисковым спамом. Анализируются значимые признаки, способствующие выявлению ссылочного спама. Дается алгоритм выявления спама на основе метода опорных векторов и приводятся результаты его работы.

## 1 Введение

С ростом популярности сети интернет повышается интерес к поисковым системам, как средствам быстрого поиска нужной информации. Вместе с тем, увеличивается число попыток манипулирования поисковыми системами посредством поискового спама.

Поисковый спам можно разделить на две большие группы: спам содержания (контента) и ссылочный спам [13]. К спаму содержания относятся методы искусственного добавления ключевых слов на страницу (в заголовки, метатеги, тексты ссылок, названия URL и текст страниц). Ссылочный спам заключается в формировании ссылочных структур, способных повлиять на алгоритмы работы поисковых систем с целью достижения более высоких позиций в результатах поиска по пользовательским запросам.

Поисковые системы активно используют ссылки. Большинство систем, так или иначе, учитывают ссылки на страницы для более эффективного ранжирования результатов поиска. В основе этого лежит постулат о том, что ссылка является воплощением желания поделиться полезной информацией с другими людьми, своего рода голосом за ресурс, на который ведет ссылка. Поэтому сайт, на который ведет много ссылок, вероятно, будет более полезен и интересен пользователям, чем сайт, на который никто не ссылается. Кроме того, ссылки с известных и популярных ресурсов считаются более весомыми, чем с никому не известных сайтов. Все это используется современными алгоритмами поисковых систем (PageRank, HITS, индекс цитирования), чтобы предоставлять пользователям более нужную и полезную информацию по поисковым запросам.

Этими же принципами пользуются при размещении ссылочного спама – намеренно размещая большое число ссылок на сайтах с возможностью простого добавления информации (форумах, гостевых книгах, комментариях в блогах и т.д.). Такие ссылки предназначены в первую очередь для поисковых систем, а не для человека. В результате набираются искусственные “голоса” в пользу сайтов, на которые ведут эти спам-ссылки и сайты начинают лучше “искаться” поисковыми системами, оттесняя качественным и интересные ресурсы на второй план.

Существует несколько способов размещения большого количества ссылочного спама. Несколько лет назад основными способами являлись обмен ссылками и создание ферм ссылок. Методам борьбы с ними посвящено множество алгоритмов, которые успешно используются поисковыми системами. В настоящее время на смену им пришли автоматизированные средства массового размещения ссылок. К таким средствам относятся специализированные программные продукты, позволяющие автоматически добавлять ссылки в каталоги, гостевые книги, форумы, блоги и т.д. Например, с помощью программы Allsubmitter можно за несколько часов поместить ссылки на десятках (а то и сотнях) тысяч сайтов. С такими ссылками можно бороться путем выявления сайтов с возможностью свободного, немодерируемого добавления информации (ссылок). Еще большей проблемой являются системы пакетной покупки ссылок через рекламных брокеров. Такие системы могут размещать ссылки на миллионах страниц. Например, самая популярная система купли-продажи ссылок – Sape.ru имеет возможность размещать ссылки на более чем 73 миллионах страниц. В прошлом году это число составляло 35 миллионов страниц. Рост аудитории более чем в 2 раза свидетельствует о повышающейся популярности этой системы. Система MainLink.ru также размещает ссылки на 40 миллионах страниц, LinkFeed.ru – на 14 миллионах.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2009, Петрозаводск, Россия, 2009.**

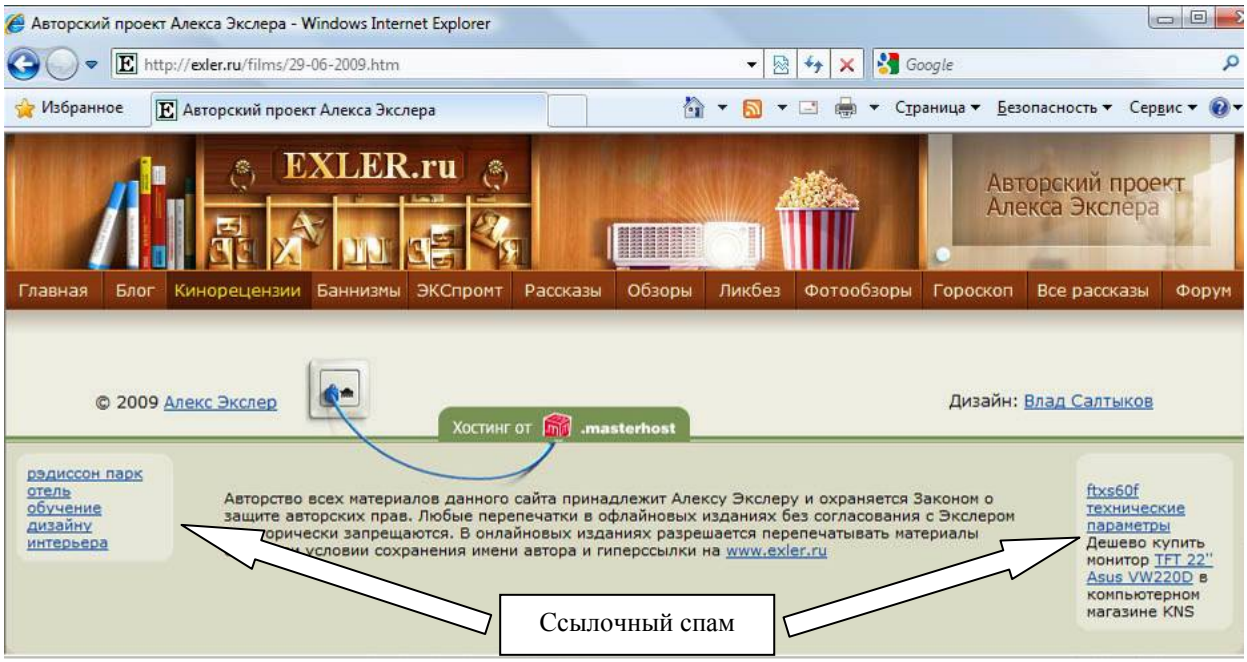



Рис. 1. Пример ссылочного спама на сайте Алекса Экслера

Таблица 1. Характеристики рекламных брокеров

| Система | Страниц | Сайтов |
|---|---|---|
| Sape.ru | 75 702 512 | 206 325 |
| MainLink.ru | 39 253 465 | 92 929 |
| Xap.ru | 32 608 976 | 80 000 (?) |
| LinkFeed.ru | 14 068 832 | 35 302 |
| SetLinks.ru | 7 600 136 | 28 991 |

Среди сайтов, размещающих у себя ссылки через рекламных брокеров имеется множество популярных и авторитетных сайтов. Например, один из рекламных брокеров (Prospero.ru) размещает ссылки на сайтах с PR=7 и индексом цитируемости до 27000. В списке доступных для размещения площадок у него мы нашли такие сайты, как http://www.auto.ru, http://www.foto.ru, http://www.vkontakte.ru, http://www.interfax.ru/, http://news.rin.ru и т.д.

Влияние спам-ссылок на алгоритмы поисковых систем может быть существенным, учитывая их массовый характер. Особую опасность представляют ссылки с авторитетных ресурсов. Поэтому выявление ссылочного спама является актуальной задачей.

Ранее [27] мы предлагали алгоритм, основанный на эмпирических методах определения ссылочного спама. Сущность метода заключалась в обнаружении ряда признаков ссылочного спама, за каждый из которых ссылкам назначались штрафы. При превышении суммы штрафов некоторого порога, ссылки признавались спамом. Недостатком алгоритма является ручной подбор значений штрафов и некоторые неточности в работе.

В текущей работе рассматривается возможность применения методов машинного обучения для задач обнаружения ссылочного спама.

## 2 Текущее состояние проблемы

Вопросу выявления ссылочного спама посвящено несколько направлений исследований.

В работе [10] предлагается статистический анализ для выявления автоматически сгенерированных страниц со спамом. О спаме может свидетельствовать: отклонение от нормального распределения различных свойств страниц, включая имена и IP-адреса, входящие и исходящие ссылки, содержание страницы и норму изменения.

Множество работ посвящено анализу ссылочной информации – в первую очередь взаимосвязях страниц, объединяемых ссылками и текстам самих ссылок. Ряд разработчиков предлагают алгоритмы, построенных на основе PageRank.

В [16] рассматривается алгоритм Anti-Trust Rank. Алгоритм основан на ручном отборе страниц с и без спама. Дальнейший анализ структуры вэб-графа, построенного на основе ссылочных структур, позволяет выявить страницы, использующие спам. Алгоритм показывает высокую точность обнаружения спама, в том числе для страниц с высоким PageRank.

В [2] предлагается алгоритм SpamRank. Алгоритм основан на понятии персонализированного PageRank и обнаруживает страницы с незаслуженным высоким значением PageRank без использования любого вида белых или черных списков или других средств вмешательства человека.

В работе [11] описывается алгоритм TrustRank для борьбы со спамом. Принцип TrustRank строится на том, что “хорошие” страницы обычно ссылаются на “хорошие” страницы и редко используют ссылки

для спама. Сначала выбирается набор “хороших” страниц и им назначается высокий вес. Далее используется подход, аналогичный PageRank: вес разделяется на исходящие ссылки к другим страницам. Наконец, после конвергенции, страницы с высоким весом принимаются за хорошие страницы. Авторы считают, что использование алгоритма TrustRank дает более качественные результаты, чем PageRank.

В работе [23] предлагается анализировать вэб-граф для определения ссылочного спама (в частности, ферм ссылок). Алгоритм основан на анализе входящих и исходящих ссылок сайтов. В случая обнаружения пересечения входящих и исходящих ссылок больше определенного порога, страницам назначается штраф. Эта операция выполняется для всех страниц.

В [12] рассматривается алгоритм определения страниц, повышающих свой PageRank с помощью ссылочного спама. Используется понятие массы спама, меры воздействия спам-ссылок на ранг страницы. Рассматриваются вопросы оценки массы спама. Для определения спама активно используется ссылочная структура вэб-графа.

В работе [8] предлагается алгоритм HostRank (PageRank, вычисленный по графу хостов), который более гибок по отношению к ссылочному спаму. Алгоритм позволяет сократить число сомнительных сайтов в результатах поиска, что достигается уменьшением веса, получаемого сайтами от ссылочного спама.

Еще одно направление работ – применение методов машинного обучения.

В работе [5] делается попытка определять ссылочный спам (“непотистский” спам). Для решения задачи используется дерево решений C4.5. Всего авторы выделяют 75 свойств, используемых для классификации. Эти свойства позволяют определять: совпадение заголовка и описания страницы, описание пересекается с текстом страницы, совпадение имен хостов, совпадение доменов, совпадение адресов страниц без доменов, совпадение некоторых частей IP адресов, одинаковые контактные E-mail домены и т.д.

В работе [18] для определения страниц со спамом так же применяется дерево решений C4.5. Авторы считают, что деревья решений имеют преимущество перед нейронными сетями, системами, основанными на правилах и методам опорных векторов. В качестве свойств для классификации используются: число слов на странице, число слов в заголовке, средняя длина слов, количество текста в ссылках, процент видимого содержания, величина сжатия страницы, процент страницы, описанный в списке популярных слов, независимая вероятность n-грамм и т.д.

В работе [19] для задач классификации используются дерево решений C4.5, входящее в пакет Weka 3.4.4. В качестве основы для классификации используются две группы свойств – связанные с содержанием и со ссылочной структурой. К первой группе относятся: число слов на странице, средняя длина слов на странице, процент слов из списка популярных слов, процент видимого содержания страницы, число слов в заголовке страницы и т.д. Во второй группе относятся: процент страниц на наиболее популярном уровне, число входящих ссылок на страницу, число исходящих ссылок на страницу, отношение числа входящих и исходящих ссылок, число ссылок с главных страниц, процент входящих ссылок на наиболее популярные страницы, процент исходящих ссылок на наиболее популярные страницы, перекрестные ссылки на страницу, средний уровень страниц на сайте и т.д.

В [3] также применяется дерево решений C4.5. Как и в [19] выделяются две группы свойств. Ссылочные свойства включают: 16 свойств степени близости (входящие и исходящие ссылки, число взаимных ссылок и т.д.), 11 свойств, основанных на PageRank (различные меры, связанные с PageRank страницы и PageRank со ссылающихся на нее страниц), Truncated PageRank и т.д. 24 свойства, зависящие от содержания, аналогичны [18]. Предложенный авторами алгоритм позволяет определять 88,4% спама. Еще одно применение дерева решений C4.5 описано в [1]. В качестве основы для классификации используются ссылочные структуры вэб-графа. Работа является продолжением исследований [3].

Работа [6] посвящена обнаружению ссылочного спама. Задача сводится к разбиению страниц на два класса – “spam” (спам) и “ham” (не спам). Для этого используется метод опорных векторов (пакет SVM-light со стандартными параметрами). Для каждой страницы выделяются 89 свойств и TF-IDF вектор. Авторы упоминают о следующих свойствах, используемых для классификации: число слов в метатегах keyword, description и заголовке, редирект на страницу, число входящих и исходящих ссылок, число символов в URL и домене, число поддоменов в URL, длина страницы, домены в зонах .edu, .org, .biz, .com, одинаковые IP-адреса, одинаковый размер страниц и т.д. Для обнаружения одинаковых страниц авторы предлагают использовать MD5 хэш-коды. Авторы анализируют различные варианты ядер SVM и выявляют наиболее важные для классификации свойства.

В работе [21] рассматривается один из подходов борьбы с поисковым спамом. Задача сводится к классификации страниц с использованием метода опорных векторов (SVM). За основу берется функция линейного ядра. В качестве базиса для классификации используется 360 свойств (зависящих и независящих от запросов). Среди независящих от запроса свойств выделяются: свойства страницы (статический ранг, самый частый терм, число уникальных термов, общее количество термов, число слов в адресе страницы, число слов в заголовке), свойств домена (ранг домена, среднее число слов, уровень домена), популярность (число

посещений домена, пользователи домена, число посещений страницы, пользователи страницы), время (дата посещения пауком, время затраченное на получение страницы, последняя дата изменения) и т.д. К зависящим от запроса свойствам относятся: количество слов запроса в заголовке, частота слов запроса в документе, частота слов запроса во всех документах, количество документов, содержащих слова запроса, n-граммы по словам запроса / документу и т.д. Эти свойства вычисляются отдельно для каждой пары (запрос, страница).

В [20] рассматривается подход к обнаружению E-mail спама и спама блогов (как ссылочного, так и спама контента) с помощью ослабленного онлайн SVM (Relaxed Online SVM). Демонстрируются возможности работы с большими наборами данных и обсуждаются возможности существенного снижения вычислительной нагрузки.

В [15] также описывается применение SVM для определения ссылочного спама в блогах. Большое внимание уделяется локальным свойствам, зависящим только от содержания каждой конкретной страницы. Проводится сравнение классификации на основе локальных свойств для линейного ядра, ядра радиальной базисной функции и сигмоида. Далее рассматриваются глобальные свойства и проводится сравнение классификации для линейного, полиномиального ядра, ядра радиальной базисной функции.

В [24] кроме содержания страницы, учитываются временные характеристики и ссылочная структура вэб-графа (на основе HITS-алгоритма) для определения спама в блогах. Для классификации используется SVM с полиномиальным ядром.

В [7] проводится сравнение метода опорных векторов, деревьев решений C4.5 и алгоритма Rocchio. Авторы указывают, что SVM работает быстрее, чем C4.5, как при обучении, так и классификации.

Анализ показывает, что существующие алгоритмы базируются на анализе структуры сети ссылок, выявлении спам-страниц и сайтов и т.д. Но они практически не предназначены для обнаружения “хороших” и “спам” ссылок на каждой отдельной странице.

Цель нашего исследования – определение спам-ссылок на любых веб-сайтах, в том числе авторитетных. На каждой отдельной странице могут присутствовать и обычные и спам-ссылки.

## 3 Метод исследования

Анализ показал, что среди методов машинного обучения высокие перспективы имеем метод опорных векторов [4, 22]. Этот метод позволяет добиться высокого качества в области классификации.

Работа метода опорных векторов включает в себя два основных этапа – обучение на тренировочных данных и непосредственно классификация.

Для работы метода необходимо определение пространства признаков, по которым будет проходить выявление ссылочного спама.

## 4 Признаки ссылочного спама

Рассмотрим признаки ссылочного спама, которые можно выделить на основе анализа содержания страницы или всего сайта. Признаки, основанные на свойствах вэб-графа (ссылочной структуре, образуемой сайтами интернет), в рамках данной работы рассматриваться не будут.

Из применяемых в настоящее время признаков (см. пункт 2), львиная доля предназначены для выявления страниц и сайтов со спамом. Для идентификации на странице конкретных спам-ссылок такие признаки использоваться практически не могут. По этой причине нами были выделены новые признаки, способные справиться с поставленной задачей. Надо заметить, что признаки не претендуют на полноту и их перечень будет расширяться по мере развития и совершенствования работы.

Все признаки ссылочного спама мы разбили на две группы [9, 17, 26, 27].

**Группа 1. Свойства ссылки.**

*1.1. Тематическая близость ссылки и страницы.*

Указывает, насколько отличается текст ссылки от тематики страницы, на которой ссылка расположена.

*1.2. Тематическая близость сайта, на который ведет ссылка ссылки и страницы, на которой ссылка расположена.*

*1.3. Тематическая близость соседних ссылок.*

Сигнализирует, совпадает ли тематика ссылки с тематикой соседних ссылок (в случае наличия блока ссылок).

*1.4. Расположение ссылки в блоке ссылок.*

Указывает, является ли ссылка одиночной, либо расположенной в области с повышенной плотностью ссылок на небольшом участке страницы (блоке ссылок).

*1.5. Место расположения ссылок.*

Указывает положение ссылки на странице – в начале или конце страницы, по центру, в левом или правом столбце и т.д. Также указывает на расстояние ссылки от основного содержания страницы.

*1.6. Пометка ссылки как рекламного объявления.*

Сигнализирует, есть ли в окрестностях ссылки пометки "Реклама", "Спонсоры", "Наши Партнеры", и т.д.

*1.7. Наличие похожих ссылок на сайте.*

Указывает, встречаются ли еще на сайте ссылки, похожие на анализируемую ссылку.

*1.8. Наличие ссылки в спам-списке.*

Спам-список [27] содержит ссылки, отобранные вручную и определенные раннее как спам.

*1.9. Признак размещения ссылки рекламным брокером*

Используется способ, описанный в [25]. Суть его заключается в том, что ссылки от рекламных брокеров устанавливаются для определенных страниц, и чаще всего с помощью одного и того же кода. Соответственно, рекламный брокер узнает о том, какой код разместить на странице, анализируя строку адреса страницы, например, http://www.site.ru/index.php?cat=1&page=11. Тогда, передав дополнительный параметр (например, http://www.site.ru/index.php?cat=1&page=11&**aa=bb**) можно ввести в заблуждение рекламного брокера, и он не установит рекламные ссылки на страницу. Сравнив содержание страницы в первом и втором случае, появляется возможность выявить такие ссылки.

**Группа 2. Свойства страницы/сайта.**

*2.1. Наличие спам-ссылок на сайте.*

*2.2. Наличие спам-ссылок на странице.*

*2.3. Сайт продает ссылки.*

Указывает, если ли на сайте информация о том, как можно купить ссылки.

*2.4. Наличие на сайте признаков кода рекламных брокеров.*

Многие автоматизированные системы установки ссылок (биржи, обменники, брокеры) устанавливают код автоматически по шаблону. Наличие блока идентичных по коду ссылок может указывать на их спамерское происхождение.

*2.5. Наличие на странице признаков кода рекламных брокеров.*

*2.6. Наличие на сайте ссылки на рекламного брокера.*

*2.7. Наличие на странице ссылки на рекламного брокера.*

*2.8. Отношение числа внешних ссылок на странице к среднему числу внешних ссылок на сайте.*

*2.9. Процент контента страницы, занятого внешними ссылками.*

*2.10. Совпадение IP-адресов сайтов.*

Указывает, совпадают ли IP-адреса сайта, на котором размещена ссылка, с сайтом, на который она указывает.

*2.11. Совпадение контактных E-mail сайтов.*

Указывает, совпадает ли контактный E-mail (указанный при регистрации домена) сайта, на котором размещена ссылка, с контактным E-mail сайта, на который ссылка указывает.

## 5 Набор данных

В качестве тестовых наборов мы использовали собственную коллекцию RV, коллекции narod.ru и By.Web семинара РОМИП. В каждой коллекции были выделены ссылки, для которых установлены метки “спам” и не “спам”.

В коллекцию RV вошли ссылки с 20 сайтов, размещающих спам-ссылки (информация о местах размещения платных ссылок были предоставлены нам владельцами сайтов). Число страниц на каждом сайте – от 100 до 5000 [27]. Всего было размечено (в автоматическом режиме) 23000 спам-ссылок и 8000 обычных ссылок.

В связи с тем, что коллекция narod.ru содержит сайты 2003 года, когда ссылочный спам только начинал свое массовое распространение (первая биржа ссылок clx.ru появилась в середине 2002 года), в ней отсутствуют некоторые признаки, присущие современному ссылочному спаму. Мы произвольно выбрали из коллекции набор страниц, на которых вручную провели разметку ссылок. Всего было размечено 2000 ссылок, из которых спам-ссылок 500, обычных ссылок 1500.

Коллекция By.Web оказалась более современной и интересной. В ней ссылочный спам представлен достаточно ярко и разносторонне. Из-за ограниченности в ресурсах, мы выбрали по 3500 спам и обычных ссылок.

## 6 Результаты исследований

В качестве основы для исследования был взят пакет SVM-Light [14]. Мы использовали линейное ядро с параметрами по умолчанию.

Для коллекции RV мы выбрали 4000 ссылок для обучения (по 2000 спам и не спам). Для классификации было использовано 21000 спам и 6000 не спам ссылок.

Для коллекции Narod.ru мы выбрали 200 ссылок для обучения (по 100 спам и не спам). Для классификации было использовано 400 спам-ссылок и 1400 не спам.

Для коллекции By.Web мы выбрали по 1750 спам и не спам ссылок для обучения. Для классификации было использовано также по 1750 ссылок (всего 3500).

Таким образом, для обучения во всех трех коллекциях использовалось одинаковое число положительных и отрицательных примеров (в нашем случае, спам и не спам ссылки).

Для оценки качества работы алгоритма использовалась следующие метрики [27]:

$$\text{Precision} = \frac{\text{Число спам - ссылок, отмеченных как спам}}{\text{Число ссылок,отмененных как спам}}$$

$$\text{Recall} = \frac{\text{Число спам - ссылок, отмеченных как спам}}{\text{Общее число спам - ссылок}}$$

$$\text{FalseSpam} = \frac{\text{Число обычных ссылок, отмеченных как спам}}{\text{Общее число обычных ссылок}}$$

$$\text{FalseNotSpam} = \frac{\text{Число спам - ссылок, отмеченных как не спам}}{\text{Общее число спам - ссылок}}$$

Значения метрик для обеих коллекций приведены в таблице 2.

Качество определения спам-ссылок для коллекции RV значительно лучше, чем для коллекций Narod.ru и By.Web. Разницу можно объяснить существенным различием в виде

тестовых данных. В коллекции RV преобладают ссылки с явно выраженными признаками спама (сайты массово размещают ссылки через рекламных брокеров). В коллекции Narod.ru преобладают обычные ссылки. Невысокое значение Precision (0.53) объясняется ошибочным отнесением хороших ссылок в разряд спама. В коллекции By.Web значение Precision несколько выше (0.72). В тоже время количество ссылок, ошибочно отнесенных к разряду спама в этой коллекции несколько выше (FalseSpam = 0.3).

Таблица 2 – Результаты работы

| | RV | Narod.ru | By.Web |
|---|---|---|---|
| Precision | 0.95 | 0.53 | 0.72 |
| Recall | 0.87 | 0.77 | 0.8 |
| FalseSpam | 0.13 | 0.20 | 0.3 |
| FalseNotSmap | 0.13 | 0.23 | 0.2 |

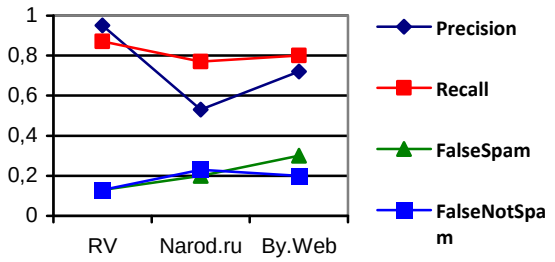



Рис. 2. Сравнение результатов для разных коллекций.

Таким образом, предложенный метод обнаружения ссылочного спама демонстрирует приемлемые результаты. Продолжение работы мы видим в расширении пространства признаков, исследовании их влияния на качество классификации, оптимизации параметров SVM-Light.

## Литература


[1] Becchetti L., Castillo C., Donato D., Leonardi S., Baeza-Yates R. Link Analysis for Web Spam Detection. ACM Trans. Web 2, 1, 1-42, 2008

[2] Benczur A. A., Csalogany K., Sarlos T., Uher, M. Spamrank - fully automatic link spam detection. In First International Workshop on Adversarial Information Retrieval on the Web, 2005.

[3] Castillo C., Donato D., Gionis A., Murdock V., Silvestri F. Know Your Neighbors: Web Spam Detection Using the Web Topology. SIGIR'07, May, 2007.

[4] Cristianini N., Shawe-Taylor J. "An introduction to Support Vector Machines", Cambridge, 2000.

[5] Davison B. D. Recognizing nepotistic links on the web. In AAAI-2000 Workshop on Artificial Intelligence for Web Search, Austin, TX, pages 23–28, July 30 2000.

[6] Drost, I., Scheffer, T. Thwarting the Nigritude Ultramarine: Learning to Identify Link Spam. in 16th European Conference on Machine Learning, (Porto, 2005).

[7] Drucker H., Wu D., Vapnik, V. Support vector machines for Spam categorization. IEEE-NN 10(5):1048–1054, 1999.

[8] Eiron N., McCurley K. S., Tomlin J. A. Ranking the web frontier. In Proceedings of the 13th International World Wide Web Conference (WWW), pages 309–318, New York, NY, USA, 2004. ACM Press.

[9] Enge E. 15 Methods for Paid Link Detection http://www.stonetemple.com/blog/?p=167

[10] Fetterly D., Manasse M., Najork M. Spam, damn spam, and statistics – Using statistical analysis to locate spam web pages. In Proceedings of the 7th International Workshop on the Web and Databases (WebDB), Paris, France, 2004.

[11] Gyongyi Z., Garcia-Molina H., Pedersen J. Combating web spam with TrustRank. In Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), Toronto, Canada, 2004.

[12] Gyongyi Z., Berkhin P., Garcia-Molina H., Pedersen J. Link Spam Detection Based on Mass Estimation. In: 32nd International Conference on Very Large Data Bases (VLDB 2006), September 12-15, 2006, Seoul, Korea

[13] Gyongyi Z., Garcia-Molina H. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005), May 10-14, 2005, Chiba, Japan.

[14] Joachims T. Making large-scale support vector machine learning practical // Advances in Kernel Methods: Support Vector Machines / B.Scholkopf, C. Burges, A. Smola (eds.) - MIT Press, 1998.

[15] Kolari P., Java A., Finin T., Oates T., Joshi A. Detecting Spam Blogs: A Machine Learning Approach. AAAI '06, 2006.

[16] Krishnan V., Raj R. Web Spam Detection with Anti-Trust-Rank. In the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '06), August 2006.

[17] Nash T. How to find a paid link? http://paymentblogger.com/2007/10/07/how-to-find-a-paid-link/

[18] Ntoulas A., Najork M., Manasse M., Fetterly D.. Detecting spam web pages through content analysis. In WWW, pages 83–92, Edinburgh, Scotland, May 2006.

[19] Qingqing Gan, Torsten Suel. Improving web spam classifiers using link structure. Proceedings in Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '07), May 2007, Banff, Alberta, Canada.

[20] Sculley D., Gabriel M. Wachman. Relaxed Online Support Vector Machines for Spam Filtering, SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference, 2007

[21] Svore, K., Wu, Q., Burges, C. and Raman, A. Improving Web Spam Classification using Rank-time Features. Proceedings in Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '07), May 2007.
[22] Vapnik V. "An Overview of Statistical Learning Theory", IEEE Transactions on Neural Networks, 1999.
[23] Wu B., Davison B. D. Identifying link farm pages. In Proceedings of the 14th International World Wide Web Conference (WWW), 2005.
[24] Yu-Ru Lin, Hari Sundaram, Yun Chi, Jun Tatemura, Belle Tseng. Splog Detection Using Content, Time and Link Structures,Proc. International Conference and Multimedia Expo (ICME) 2007, July 2007, Beijing, China.
[25] Детектор продажных ссылок, 2008. http://venality.name/
[26] Кравцов Алексей. Ссылочный спам: найти и обезвредить http://www.kravcov.ru/2007/03/11/nnueiiue-niai-e-eae-n-iei-aidhiouny/
[27] Шарапов Р.В., Шарапова Е.В. Обнаружение ссылочного спама // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Десятой Всероссийской научной конференции «RCDL'2008» (Дубна, Россия, 7-11 октября 2008 г.). - Дубна: ОИЯИ, 2008. С. 191-196.

## The Using of Support Vector Machines for Link Spam Detection

Sharapov R.V., Sharapova E.V.

In article approaches to detecting of Link Spam by methods of machine learning are considered. The significant signs helping Link Spam detection are analyzed. The algorithm of detecting a spam-links on basis of Support Vector Machine is given and results of its work are considered.

# Технологии управления разнородным естественнонаучным контентом на основе семантического веба♣

© А. М. Елизаров, Е. К. Липачев, М. А. Малахальцев
НИИ математики и механики им. Н. Г. Чеботарева
Казанского государственного университета
amelizarov@gmail.com

**Аннотация**

В работе рассматриваются подходы к формированию электронных коллекций, организации хранения и поиска данных на основе XML и других технологий семантического веба, реализованные в проекте «Научная электронная библиотека eLibrary.ru» и электронном журнале «Lobachevskii Journal of Mathematics» (http://ljm.ksu.ru). Обсуждается возможность создания электронных хранилищ нового типа, характеризующихся наличием динамических интеллектуальных связей между документами разных типов, на основе спектра языков разметки, широко применяемых в математических, химических, биологических и других предметных областях.

## 1 Введение

В настоящее время во всем мире, в том числе в России, идет активная работа по созданию электронных хранилищ научных документов, в частности, создаются и развиваются разнообразные электронные научные коллекции. Разработаны основные принципы организации таких коллекций и соответствующее программное обеспечение (см. обзор [1]). На этих принципах организовано большинство электронных хранилищ (например, [2 – 5]). Как правило, научные электронные коллекции представляют собой набор документов (в основном текстов статей и книг) и их «библиографических» описаний, построенных на основе языка XML. Поэтому научные электронные коллекции уже сейчас позволяют организовать поиск не только по текстовой информации, но и по XML-описанию.

Одним из первых в России в области управления электронной научной информацией был проект «Научная электронная библиотека eLibrary.ru» (см., например, [5]). В рамках этого проекта в части разработки методов управления электронным научным контентом нами были предложены технологические решения, базирующиеся на идеях семантического веба [2], [6]. Эти подходы были реализованы в системе управления математическим контентом электронного журнала «Lobachevskii Journal of Mathematics» (http://ljm.ksu.ru) [2].

## 2 Управление научным контентом в НЭБ

Для структурирования макетов печатных изданий в электронной коллекции, созданной в рамках проекта «Научная электронная библиотека elibrary.ru» (НЭБ), была разработана программная среда, в основу которой положен алгоритм выделения элементов текста и присвоения им меток полей собственного XML-формата, названного Sarcticle (подробности см. в [5]). Отличительными особенностями этого формата являются: вложенность полей, возможности описания любого количества информации одним файлом, проверки правильности составления файлов описаний на стороне издательств, использования файлов описаний для наполнения собственных сайтов издательств и совместимости с другими форматами обмена метаданными, основанными на XML. Основные блоки формата – информация о журнале, выпуске и статье (основная информация файла). Большинство полей может дублироваться на нескольких языках с целью более удобного представления для разных пользователей конечной информации в электронной библиотеке.

Задача подготовки библиографических материалов, включаемых в различные индексы научного цитирования, решается с помощью программного модуля, производящего автоматическое структурирование (разбор по полям формата) списков литературы и сносок. В этом модуле учтены требования ГОСТ 7.1-84 «Библиографическое описание документа».

Названное программное обеспечение позволяет работать с большинством форматов (в том числе, html, PageMaker, .pdf, .doc) и максимально автоматизирует процесс структуризации текста макетов. Одной из основных задач, которая была решена при разработке программы разметки текстов электронных журналов, состояла в организации обработки и отображения математических и химических формул, диакритических, математических и других специа-

**Труды 11ⁱᵉ Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

лизированных знаков и символов.

Использование формата Sarcticle позволило создать программный комплекс, поддерживающий расширенный поиск статей по журнальной базе, в частности, возможен поиск по следующим параметрам: авторы, названия статей, аннотации (рефераты), ключевые слова, слова из полных текстов статей, библиографические описания источников.

Программное обеспечение разметки электронных журналов прошло успешное тестирование в ряде редакций научных журналов. По разработанной технологии в Научной электронной библиотеке eLibrary.ru было размещено более 200 научных журналов по разным специальностям, в том числе «Успехи физических наук», «Успехи химии», «Биология моря», «Ученые записки Казанского университета», «Известия вузов. Авиационная техника», «Известия вузов. Математика», «Известия вузов. Радиофизика», «Казанский медицинский журнал», «Тихоокеанская геология» и ряд других. Большинство из названных журналов до этого момента времени не было представлено в интернете.

Указанные технологии и возможности НЭБ характерны для современных электронных библиотек, однако развитие информационных технологий делает возможным переход в организации электронных коллекций на качественно новый уровень на основе технологий семантического веба.

## 3 Технологии семантического веба и электронные хранилища нового типа

Согласно программным документам консорциума W3C [7], семантический веб – это «... расширение традиционного веба в направлении существенно лучшего определения смысла информации, позволяющего компьютерам и людям эффективнее выполнять совместную работу. Мы хотим, чтобы данные в вебе были определены и связаны ссылками так, чтобы их можно было легче находить, интегрировать, автоматизировать и повторно использовать в различных приложениях, ... чтобы данные были разделяемыми и могли обрабатываться как автоматизированными средствами, так и людьми». Конечная цель этого проекта состоит в создании такой среды, в которой программные агенты смогут динамически обнаруживать и опрашивать ресурсы, а затем взаимодействовать с ними. Такие агенты должны уметь справляться с возникающими виртуальными проблемами интеллектуализированной среды, обнаруживать новые факты и выполнять изощренные задания, получаемые от людей (см., например, [8]).

Основные цели семантического веба, приведенные выше, естественным образом переносятся на электронные коллекции и электронные библиотеки, позволяя говорить об *электронных хранилищах нового типа.* С нашей точки зрения, можно выделить следующие характеристики, присущие таким электронным хранилищам.

1) *Разнородность контента* – понимается как разнообразие предметных областей (математика, физика, химия, биология, геология и т. д.), разнообразие как типов документов (научные статьи, результаты наблюдений и экспериментов, программные продукты), так и форматов (текстовый, графический, звуковой, видео). Кроме того, структура документов хранилища может быть отличной от традиционных научных статей. Например, документ по астрономии может включать одновременно текст, математические расчеты, данные наблюдений и программы обработки этих данных.

2) *Возможности интеллектуального поиска по специализированным документам*. Сегодня для подавляющего большинства коллекций электронных документов основная задача состоит в предоставлении пользователю информации по запросу, который формирует сам пользователь, например, указав предметную область и ключевые слова. В электронном хранилище нового типа должны быть установлены *динамические интеллектуальные связи* между документами различных типов, позволяющие предоставлять пользователю информацию из разных областей и разных типов. Простой пример: при запросе о дифференциальном уравнении определенного типа должна выдаваться информация не только о математических результатах, связанных с этим уравнением, но и сведения о его применениях, вычислительных экспериментах, связанных с ним, и т. д. Подчеркнем, что в документах, где данное уравнение используется, оно может быть не упомянуто как ключевой термин и даже не названо, однако система электронного хранилища сама должна установить необходимые связи. Поэтому в целом система управления электронным хранилищем должна предоставлять пользователю *структурированный комплекс документов*.

3) *Новый тип интерфейса*. Интерфейсы имеющихся хранилищ ориентированы на взаимодействие с человеком. Хранилище нового типа, согласно принципам семантического веба, должно иметь интерфейс, приспособленный для взаимодействия на программном уровне («машинно-ориентированный подход» [9]).

Электронные хранилища нового типа соответствуют задачам современной науки, где на первый план вышли междисциплинарные исследования. Создание таких хранилищ возможно благодаря тому, что к настоящему времени для большинства естественнонаучных предметных областей уже созданы специализированные языки разметки.

Язык разметки химических формул CML (Chemical Markup Language) разработан как часть проекта Open Molecule Foundation. С помощью CML записывается информация о молекулярных структурах, химических реакциях, спектрах, неорганических кристаллах, объектах квантовой химии. Для создания и обработки CML-файлов можно использовать уже созданные программные средства, предназначенные для работы с информацией в формате XML [10 – 12].

Для представления математических формул в

рамках семантического веба используется технология MathML [13]. Разработка этого языка ведется консорциумом W3C с 1998 года. Фактически он уже стал стандартом представления математической информации в электронной форме [14, 15].

Для хранения и электронного обмена математическими моделями применяется язык CellML (Cell Markup Language). В частности, он широко используется в биологическом моделировании. Отметим, что CellML поддерживает спецификацию MathML.

Для описания свойств материалов можно использовать язык Materials Markup Language (MatML). Подробное его описание можно найти на сайте www.matml.org.

Географическим сообществом используется язык Geography Markup Language (GML) (информацию о нем можно найти на сайтах http://www.opengis.net/gml/ и http://www. opengeospatial.org/standards/gml, http://schemas.opengis.net/gml/).

Отметим, что созданы языки разметки и для других предметных областей, причем названия языков, как правило, отражают их назначение: Business Rules Markup Language (BRML), Geography Markup Language (GML), Finite Element Modeling Markup Language (femML), Ink Markup Language (InkML), Mathematics Education Markup Language (MeML), Materials Markup Language (MatML), Numerical Data Markup Language (NDML), Relational-Functional Markup Language (RFML), Robotic Markup Language (RoboML), Voice Extensible Markup Language (VoiceXML). В работе [16] предложена классификация уже созданных языков разметки в виде карты языков XML.

Важно, что технологии семантического веба обеспечивают стандартную процедуру создания языка разметки, адаптированного к определенной предметной области [12], что позволяет гибко настраивать структуру хранилища для включения информации из новых предметных областей.

Структура естественнонаучного электронного хранилища нового типа основана на имеющихся в настоящий момент времени языках разметки, широко применяемых в математических, химических, биологических и других предметных областях. Общая природа этих языков позволяет использовать уже отлаженные технологии семантического веба для управления разнородным контентом естественнонаучного электронного хранилища.

**Заключение**

Таким образом, имеющиеся на сегодняшний день технологии семантического веба, в частности, разработанные языки разметки естественнонаучной информации позволяют решить задачу управления разнородным контентом в электронных хранилищах.

**Литература**


[1] Когаловский М. Р. Тенденции развития технологий управления информационными ресурсами в электронных библиотеках// Тр. 8-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, Россия, 2006.

[2] Веселаго В. Г., Елизаров А. М., Липачёв Е. К., Малахальцев М. А. Формирование и поддержка физико-математических электронных научных изданий: переход на технологии семантического веба// В кн. «Научно-исследовательский институт математики и механики им. Н. Г. Чеботарева Казанского государственного университета. 2003 – 2007 гг.». Кол. монография под ред. А. М. Елизарова. – Казань: Изд-во Казан. ун-та, 2008. – С. 456-476.

[3] Бархатов А. В., Вдовицын В. Т., Луговая Н. Б., Сорокин А. Д. Электронные научные информационные ресурсы для поддержки инвестиционной деятельности в регионе. – www.kareliainvest.ru/file.php/id/f3770/name/STATIA_Inf_res.doc.

[4] Голосов Ю. И., Брагина Г. А., Пржиялковская М. Н. Электронные документы научно-технической информации в системе ВНТИЦ// Тр. 10-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

[5] Глухов В. А., Елизаров А. М. Проект «Научная электронная библиотека eLibrary.ru» и российские электронные журналы: новый этап развития //Тр. 8-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, Россия, 2006. – С. 203-207.

[6] Елизаров А. М., Липачев Е. К., Малахальцев М. А. Технологии Semantic Web в практике работы электронного журнала по математике //Тр. 8-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, Россия, 2006. – С. 215-218.

[7] W3C Semantic Web Activity Statement. – http://www.w3.org/2001/sw/Activity.

[8] Hendler J. Agents and the semantic web// IEEE Intelligent Systems J. – 2001. – V. 16, No 2. – P. 30-37.

[9] Berners-Lee T. Semantic web road map. – http://www.w3.org/DesignIssues/Semantic.html; Рус. перевод: http://gridclub.ru/library/publication. 2007-04-23.2195467714/view.

[10] Jirat J. Chemical Markup Language 1.0 reference with examples. – http://www.zvon.org/xxl/CML1.0

[11] Amies A. Tools for working with Chemical Markup Language. – http://www.medicalcomputing. net/cmltools.html.

[12] Елизаров А. М., Липачёв Е. К., Малахаль-

цев М. А. Языки разметки семантического веба. Практические аспекты. – http://www.ksu.ru/fpk/docs/lip_mal.pdf.
[13] Mathematical Markup Language (MathML) Version 2.0 (Second Edition). – http://www.w3.org/TR/2003/REC-MathML2-20031021/.
[14] Елизаров А. М., Липачёв Е. К., Малахальцев М. А. Основы MathML Представление математических текстов в Internet. Практическое руководство. – Казань: Изд-во Казан. матем. общества, 2004. – 60 с. – www.ksu.ru/.
[15] Елизаров А. М., Липачёв Е. К., Малахальцев М. А. Основы MathML Представление математических текстов в Internet. – Казань, 2008. – 101 с. – http://www.niimm.ksu.ru/data/preprints/.
[16] Лозовюк А. Комета по имени XML. – http://www.marketer.ru/.

## Management technology for multi-discipline scientific content based on Semantic Web

A.M. Elizarov, E.K. Lipachev, M.A. Malakhaltsev

In the present paper we consider approaches to formation of electronic collections, organization of data storage and search on the base of XML and other Semantic Web technologies, which were used in the project «Scientific electronic library eLibrary.ru» and in the electronic journal «Lobachevskii Journal of Mathematics» (http://ljm.ksu.ru). We propose to design new type electronic storages which are characterized by dynamical knowledge-based relations between documents of different types with the use of the variety of markup languages widely applied in mathematics, chemistry, biology and other areas of science.

♣ Работа поддержана РГНФ (проект № 07–01–12146) и РФФИ (проект № 09–07– 12059-офи_м)

# Являются ли сайты конференций RCDL научными веб-коммуникаторами?♣

© А.А. Печников, Н.Б. Луговая

Институт прикладных математических исследований КарНЦ РАН
{pechnikov, nataly}@krc.karelia.ru

**Аннотация**

В авторской терминологии модель, называемая схемой научного Веба, конструируется из четырех основных компонент, являющихся непересекающимися подмножествами научных сайтов. Эти компоненты называются административным каркасом, научным подмножеством, ближайшей окрестностью и веб-коммуникатором. При всей кажущейся очевидности того факта, что сайты научных конференций являются средством коммуникации ученых, проведенные исследования показывают, что сайты конференций RCDL веб-коммуникаторами не являются.

## 1 Схема научного Веба

К актуальным направлениям вебометрики [1], - одного из развивающихся направлений информатики, - относятся исследования гиперссылок (аналогичные термины – «ссылка», «веб-ссылка»), являющиеся единственным способом взаимодействия между сайтами. Практическая применимость этих исследований успешно демонстрируется реализацией алгоритмов информационного поиска таких популярных систем, как Google и Яндекс [2,4]. Теоретические исследования показывают, что изучение гиперссылок имеет достаточный потенциал как в смысле новых источников информации и коммуникации, так и ценности самих веб-страниц [3-6].

В 2008 году в Институте прикладных математических исследований КарНЦ РАН началась работа по проекту «Вебометрические исследования научных интернет-ресурсов российского Интернета». В рамках проекта разработаны поисковый робот для сбора исходящих с сайтов гиперссылок (LPR – аббревиатура от Link, Page и Robot) и база данных, предназначенная для их хранения и обработки (БД ВГ – База Данных Внешних Гиперссылок). Программный комплекс, состоящий из LPR и БД ВГ, создан на языке PHP, работает под управлением веб-сервера Apache с интегрированным модулем PHP и СУБД MySQL и находится в стадии постоянного совершенствования.

Информацию о ходе проекта можно найти на сайте «Вебометрика. ИПМИ КарНЦ РАН» [7]. По данным на март 2009 года (именно на этих данных основан дальнейший материал), проведено сканирование 280 официальных сайтов организаций и учреждений. Российской академии наук (РАН). Обработано более миллиона html-страниц, найдено и сохранено 660000 различных внешних ссылок, из которых 81000 уникальных.

В результате проведенных исследований построена теоретико-графовая модель множества научных сайтов Рунета, получившая название схемы научного Веба. Схема научного Веба представляет собой ориентированный граф, множество вершин которого соответствует исследуемым сайтам, а дуги отражают гиперссылки, существующие между сайтами (считается, что дуга существует тогда и только когда, существует хотя бы одна гиперссылка с одного сайта на другой).

Показано, что в схеме научного Веба можно выделить четыре компоненты.

Первая из них, – административный каркас, – отражает ссылки между сайтами, соответствующие иерархической подчиненности организаций.

Вторая, – множество научных подмножеств, где научное подмножество представляет связи между сайтами родственных организаций.

Третья компонента - это множество ближайших окрестностей официальных сайтов. Ближайшие окрестности содержат вершины, сайты которых имеют имена ddd.nnn.ss∈ nnn.ss, где nnn.ss – доменное имя официального сайта.

И, наконец, четвертая компонента называется множеством научных веб-коммуникаторов и соответствует множеству сайтов, выполняющих коммуникационные функции между официальными научными сайтами, то есть научные сайты имеют много входящих ссылок с сайтов этого множества и много исходящих ссылок на них.

Веб-коммуникаторы, в свою очередь, можно классифицировать по трем типам как «посредник», «индуктор» и «коммутатор». Краткое описание

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

посредника - «много входящих ссылок, много исходящих ссылок», коммутатора – «мало входящих, много исходящих», а индуктора – «много входящих, мало исходящих».

Упрощенный вариант схемы научного Веба изображен на рис. 1.

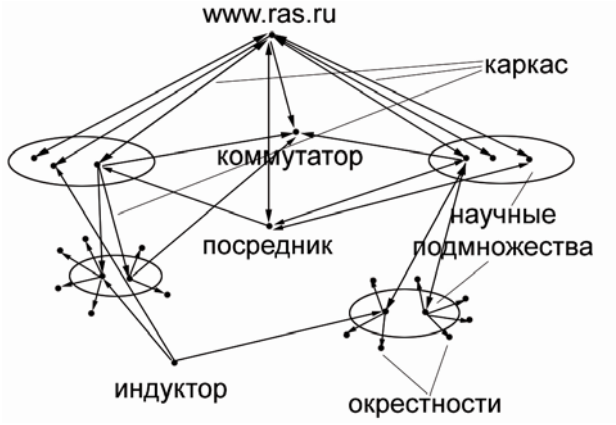



**Рис.1 Схема научного Веба.**

## 2 Цель и целевое множество исследования

Основной вопрос данной публикации заключается в следующем: являются ли сайты научных конференций научными веб-коммуникаторами?

В качестве объекта исследования были взяты хорошо знакомые авторам конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (английская аббревиатура – RCDL).

Исследования проводились на множестве сайтов конференций RCDL и научных организаций РАН, являющихся организаторами этих конференций, а также сайтов поддерживающих организаций. Сведения о веб-ресурсах всех 11 конференций приведены на сайте [10], являющемся общим сайтом RCDL и также включенным в число исследуемых сайтов. С помощью LPR было проведено сканирование 10 сайтов конференций из 12. Исключениями являются ресурсы «RCDL 2000: Протвино» и «RCDL 2003: Санкт-Петербург», поскольку они представлены не отдельными сайтами, а директориями на сайтах других организаций (LPR «не умеет» сканировать с произвольной страницы сайта).

Список всех учреждений и организаций РАН, представители которых хотя бы один раз входили в состав организационного или программного комитета, содержит 22 наименования учреждений РАН. По ряду причин (отсутствие сайта, отсутствие на сегодняшний день самой организации и др.) нам пришлось ограничиться 16 сайтами. Список составлен на основе информации, представленной на сайтах конференций и приведен ниже (рис. 2).

Единственной организацией, не входящей в состав РАН, регулярно поддерживающей конференции, является Российский фонд фундаментальных исследований, сайт которого также включен в исследуемое множество.

## 3 Результаты исследования

Операция БД ВГ, которая называется ПОСТРОЕНИЕ МАТРИЦЫ, позволяет построить матрицу смежности для 16 сайтов учреждений и организаций РАН; матрицы приведена на рис. 2. Элемент матрицы (i,j)=1, если существует хотя бы одна гиперссылка с сайта i на сайт j , и (i,j)=0 в ином случае.

Соответствие номеров вершин матрицы и организаций:

0 Библиотека по естественным наукам РАН
1 Дальневосточное отделение РАН
2 Карельский научный центр
3 Вычислительный центр имени А.А. Дородницына РАН
4 Институт астрономии РАН
5 Институт вычислительных технологий СО РАН
6 Институт космических исследований РАН
7 Институт математических проблем биологии РАН
8 Институт прикладной математики им. М. В. Келдыша РАН
9 Институт прикладных математических исследований КарНЦ РАН
10 Институт проблем информатики РАН
11 Институт систем информатики им. А.П. Ершова СО РАН
12 Институт химической физики им. Н.Н. Семенова РАН
13 Институт цитологии и генетики СО РАН
14 Объединенный институт геологии, геофизики и минералогии
15 Специальная астрофизическая обсерватория РАН

Определяя силу связности научного подмножества сайтов CFC (Community Force of Connectivity) как отношение реального количества дуг к потенциально возможному, для построенной матрицы имеем CFC=0,133. Если же удалить строки и столбцы с номерами 1 и 15 (нулевые), то получаем CFC=0,205. Следует сказать, что для академических сайтов это очень высокий показатель (эксперименты с замерами на научных подмножествах, сформированных по различным признакам принадлежности, дают значения в CFC интервале от 0 до 0,35).

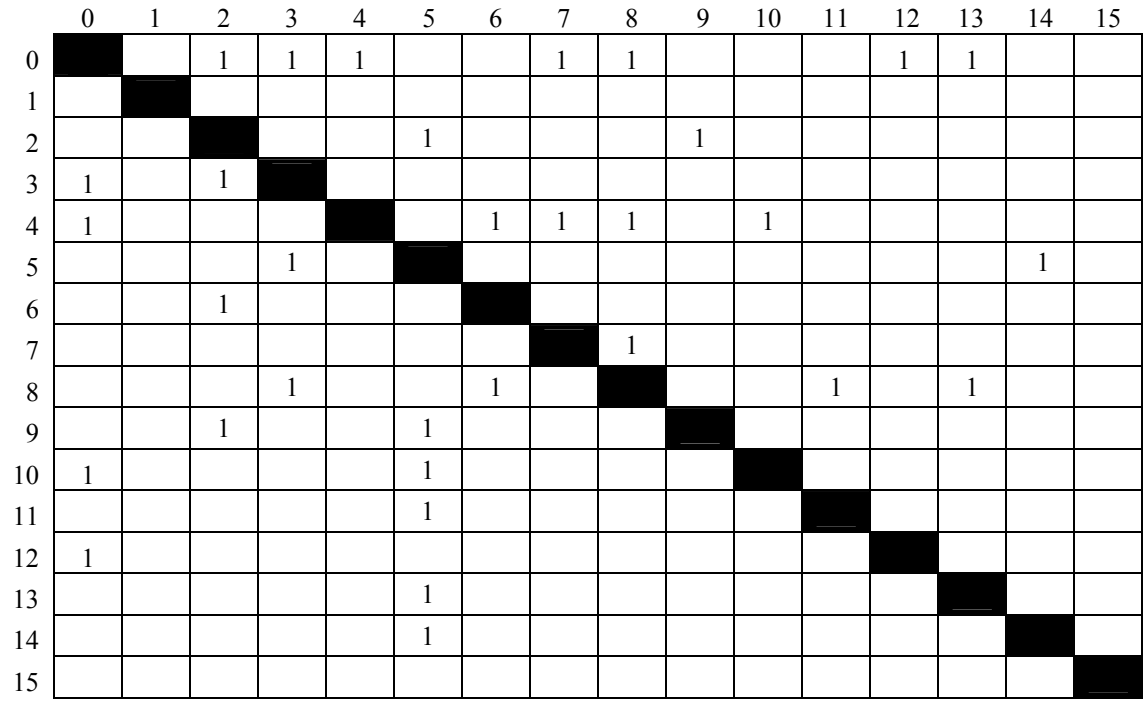


Рис. 2. Матрица смежности веб-графа сайтов организаторов конференций RCDL.

На фоне такого хорошего результата тем более удивительными выглядит картина гиперссылок с сайтов конференций на сайты организаторов выбранного нами множества (и наоборот), приведенная в табл. 1.

| Конференция | Ссылки на организаторов | Ссылки с организаторов |
|---|---|---|
| RCDL: | 1 | 0 |
| RCDL 1999: | 0 | 0 |
| RCDL 2001: | 1 | 2 |
| RCDL 2002: | 1 | 0 |
| RCDL 2003: | 1 | 1 |
| RCDL 2005: | 0 | 2 |
| RCDL 2006: | 2 | 2 |
| RCDL 2007: | 2 | 3 |
| RCDL 2008: | 2 | 3 |
| RCDL 2009: | 3 | 3 |
| ВСЕГО | 13 | 16 |

Табл. 1. Количество связей между сайтами конференций и организаторов.

В табл.1 значение, равное трем, на пересечении строки с названием «RCDL 2005: Ярославль» и столбцом «Ссылки на организаторов» означает, что с сайта конференции RCDL 2005 существуют три ссылки на различные сайты организаторов конференций (не обязательно данной конференции). Соответственно значение, равное двум, на пересечении строки с названием «RCDL 2005: Ярославль» и столбцом «Ссылки с организаторов» означает, что существуют два сайта организаторов, имеющие ссылки на сайт этой конференции. То есть за 11 лет проведения конференций с сайтов институтов, сотрудники которых (или хотя бы руководство) не могут не знать об этих конференциях, поставлено всего 16 ссылок. И если можно еще понять отсутствие ссылок на конференцию RCDL 1999: Санкт-Петербург (давно это было!), то отсутствие ссылок на общий сайт не поддается объяснению. Правда, и с основного сайта конференций сделана лишь одна ссылка на организаторов.

После этого вряд ли стоит удивляться, что из остальных академических сайтов (а их 265) ссылки на сайт какой-либо конференции RCDL сделаны лишь с шести. Соответственно, и с сайтов конференций ссылки отсылают лишь к 2 академическим сайтам (из тех же 265).

Интересно заметить, что на сайт РФФИ 13 из 15 сайтов организаторов конференций RCDL содержат гиперссылки (и не по одной, а в среднем по 8), не говоря уже о сайтах самих конференций RCDL.

Не хуже, чем сайты конференций, в качестве веб-коммуникатора для сайтов-организаторов конференции выглядит сайт РОМИП (Российский семинар по Оценке Методов Информационного Поиска) [9], на который сделаны ссылки с 2 сайтов, а он ссылается на 3.

Частичный анализ гиперссылок, сделанных с некоторых сайтов конференций RCDL, позволяет классифицировать исходящие ссылки следующим образом: четверть всех ссылок сделаны на другие сайты конференций RCDL, 17% - на сайт РОМИП, 19% - на сайты, рассказывающие о городах, в которых проводятся конференции, 8% - на Яндекс, 7% - на основного организатора текущей конференции и 5% - на РФФИ.

С помощью средств Google мы проверили, какие же сайты все-таки ссылаются на сайты конференций RCDL, если на них не ссылаются сайты организаций-учредителей из числа институтов РАН. Оказалось, что Google обнаруживает 144 ссылки, из которых 30% приходятся на сайт DELOS an Association for Digital Libraries, 30% - на все сайты RCDL вместе взятые, 10% - сайт Института информационных технологий НАН Азербайджана, по 5% - сайт журнала "Электронные ресурсы в библиотеках" и сайт РОМИП; остальные 15 сайтов с долями меньше 5%.

## Заключение

Научные конференции являются коммуникационными площадками для ученых. Проведенные исследования показывают, что сайты конференций RCDL на сегодня с такой ролью справляются не в полной мере. Наверное, на это стоит обратить внимание организаторов конференций и разработчиков соответствующих веб-ресурсов.

## Литература


[1] Almind T., Ingwersen P. Informetric analyses on the World Wide Web: Methodological approaches to "webometrics"" // Journal of Documentation. 1997. №53 (4). P. 404-426.

[2] Brin S., Page L. The Anatomy of a large scale hypertextual web search engine // Computer Networks and ISDN Systems. 1998. №30 (1-7). P. 107-117.

[3] Cronin B., Snyder H.W., Rosenbaum H., Martinson A., Callahan E. Invoked on the web // Journal of the American Society for Information Science. 1998. №49 (14). P. 1319-1328.

[4] Flake G. W., Lawrence S., Giles C. L., Coetzee, F. M. Self-organization and identification of web communities // IEEE Computer. 2002. №35. P. 66-71.

[5] Thelwall M. Extracting macroscopic information from web links // Journal of the American Society for Information Science and Technology. 2001. №52 (13). P. 1157-1168.

[6] Thelwall M. What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation // Information Research. Vol. 8. №3, April 2003. [Электронный ресурс] - 2003. - Режим доступа: http://informationr.net/ir/8-3/paper151.html.

[7] Вебометрика. Институт прикладных математических исследований КарНЦ РАН. [Электронный ресурс] - 2009. - Режим доступа: http://webometrics.krc.karelia.ru.

[8] Индекс цитирования. [Электронный ресурс] - 2008. - Режим доступа: http://help.yandex.ru/catalogue/?id=873431.

[9] Российский семинар по Оценке Методов Информационного Поиска (РОМИП). [Электронный ресурс] - 2009. - Режим доступа: http://romip.ru.

[10] Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции. [Электронный ресурс] - 2009. - Режим доступа: http://www.rcdl.ru.


## Are the RCDL conferenses sites scientific web-communicators?


A.A. Pechnikov, N.B. Lugovaya



In author's terminology the model named the scheme of a scientific Web, is designed from four cores a component which are not crossed subsets of scientific sites. These components are called as an administrative skeleton, the scientific subset, the nearest environs and a web-communicator. At all seeming evidence of that fact that sites of scientific conferences are a communication medium of the scientists, the conducted researches show that sites of conferences RCDL a web-communicators are not.



♣ Работа выполнена при финансовой поддержке РФФИ (проект № 08-07-00023а)

# ИНТЕГРАЦИЯ ИНФОРМАЦИОННЫХ РЕСУРСОВ

# INTEGRATION OF INFORMATION RESOURCES

# Планирование запросов над множеством неоднородных распределенных информационных ресурсов в архитектуре средств поддержки предметных посредников ♣

© Вовченко А.Е., Крупа А.В.
ИПИ РАН
itsnein@gmail.com, akrupa@gmail.com

**Аннотация**

Планировщик запросов играет одну из важнейших ролей в архитектуре средств поддержки предметных посредников. Планировщик отвечает за декомпозицию и выполнение пользовательского запроса на множестве распределенных неоднородных информационных ресурсов. Задача декомпозиции в архитектуре посредников всегда имеет множество решений, поэтому одной из важнейших задач планировщика является выбор наиболее эффективного решения.

В статье приводится описание планировщика, который способен эффективно выполнять запросы, сформулированные над информационными ресурсами, содержащими значительные (по сравнению с пропускной способностью каналов связи) объёмы данных и требующие продолжительного выполнения. Также в статье приводится описание требований к адаптерам ресурсов, для возможности построения эффективного плана выполнения запроса.

## 1 Введение

В различных областях науки наблюдается экспоненциальный рост объема получаемых экспериментальных (наблюдательных) данных. Например, в астрономии текущий и ожидаемый темп роста данных от наземных и космических инструментов удваивается в течение периода от шести месяцев до одного года. Сложность использования таких данных увеличивается еще и вследствие их естественной разнородности. Число организаций, получающих данные наблюдений в отдельных областях науки в мире, велико. Разнообразие (информационная несогласованность) получаемой информации вызывается, в частности, не только большим числом организаций, производящих наблюдения, и их независимостью, но и разнообразием объектов наблюдения, непрерывным и быстрым совершенствованием техники наблюдений, вызывающим адекватные изменения структуры и содержания накапливаемой информации. Это приводит к необходимости использования неоднородной, распределенной информации, накопленной в течение значительного периода наблюдений технологически различными инструментами.

Основной идеей в инфраструктуре доступа к множественным неоднородным информационным ресурсам является введение промежуточного слоя между ресурсами, и потребителями информации. Основными компонентами промежуточного слоя являются предметные посредники [1,5], существующие независимо от информационных ресурсов. Использование предметных посредников в среде неоднородных информационных ресурсов представляет интеллектуальный подход к интеграции информации, который может быть эффективно использован при создании новых средств поддержки электронных библиотек.

Архитектура средств поддержки посредников разрабатывалась для обеспечения интеграции широкого класса информационных ресурсов, различающихся возможностями выполнения запросов. Она призвана обеспечить интеграцию разнообразных видов информационных ресурсов, включая:

- базы данных (в том числе объектные) в СУБД, обладающих возможностью создания временных коллекций;
- базы данных (в том числе и объектные) в СУБД с доступом только для чтения;
- веб-сервисы;
- простые таблицы (HTML, excel), текстовые коллекции, коллекции объектов, без встроенных средств выполнения запросов.

Система исполнения запросов в такой среде должна корректно функционировать вне зависимости от объемов содержащейся в коллекциях информации. Предполагается, что каждый ресурс может взаимодействовать с

**Труды 11[ой] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2008.**

каждым. В случае если не требуется передачи данных от ресурсов к ресурсам, взаимодействие происходит только между посредником и ресурсами. Здесь рассматриваются вопросы адаптивного планирования запросов[11] в среде посредников для обеспечения ее приемлемой эффективности.

Одним из основных требований к планировщику запросов является предсказуемость времени вычисления запросов. Планировщик должен построить эффективный план и выдать для него оценку времени исполнения. Поскольку окружение, в котором исполняется запрос, непостоянно, планировщик должен следить за выполнением запроса, изменяя прогноз в зависимости от изменений в окружении. В случае существенных изменений в окружении планировщик должен уметь изменить план, не прерывая выполнения запроса, если это возможно.

В посреднике различаются два уровня представления информации: локальный уровень, представляющий метаинформацию о разнородных коллекциях, и федеративный уровень, поддерживающий глобальную метаинформацию на языке канонической модели, формируемую на основе подхода GLAV [3]. Канонической моделью данных [2] в посреднике является информационная модель языка СИНТЕЗ [6]. В качестве языка запросов в этой модели в настоящее время используется подмножество языка формул СИНТЕЗа - язык Syfs (Synthesis Formula Subset), а точнее его алгебраическая форма – Asyfs[6].

После получения запроса, переписанного в терминах локальных схем, выраженных однородно в канонической модели, планировщик осуществляет его декомпозицию во множество локальных запросов и составляет план их выполнения. Глобальный запрос формулируется в терминах федеративной схемы посредника. Локальные запросы формулируются в терминах локальной схемы для конкретной коллекции. Посредник инициирует, а планировщик должен контролировать процесс выполнения запроса согласно составленному плану.

Целью настоящей работы является обсуждение подхода к реализации эффективного алгоритма планирования в среде предметных посредников. В следующем разделе рассматривается реализованный алгоритм планирования. Затем в третьем и в четвертом разделе обсуждаются ключевые моменты алгоритма планирования. Затем в пятом разделе представлен ряд требований к ресурсам, а точнее к адаптерам ресурсов, которые должны выполняться для эффективной работы алгоритма, после чего описана архитектура разработанного адаптера ресурсов а также способности адаптеров (capabilities) в соответствии с выработанными требованиями.

## 2 Алгоритм планирования запросов

Описываемый планировщик является частью архитектуры средств поддержки посредников и отвечает в ней за эффективное выполнение запросов. Следующие свойства характеризуют класс рассматриваемой архитектуры:

- выполнением запроса управляет клиентское приложение, представляющее собой интерфейс взаимодействия с посредником;
- средства координации выполнения запросов между клиентами отсутствуют;
- координация выполнения запросов между адаптерами ресурсов реализуется исключительно клиентским приложением;
- планировщик взаимодействует непосредственно с ресурсами;
- ресурсы не предоставляют информации о других клиентах их использующих;
- информация о внутреннем устройстве ресурса, способе и скорости выполнения в нем операций, отсутствует.

В такой системе невозможно построить план выполнения запроса [8-10] со строгим указанием выделяемого времени того или иного адаптера как вычислительного ресурса каждому из запросов для выполнения определенной операции. Планировщик в этой системе формулирует для каждого информационного ресурса набор независимых заданий [7], которые могут выполняться одновременно, а могут быть поставлены в очередь, и передает их адаптеру. Если выполнение следующего задания требует завершения предыдущего, то планировщик должен ждать завершения первого, и только потом делать запрос на выполнение второго. Задания от разных клиентов также независимы. Адаптерам ресурсов дана почти полная свобода в выборе порядка исполнения всех заданий. В общем случае, они могут выполняться параллельно, если это позволяет ресурс. Задания могут быть выполнены не полностью, а по частям. Это позволяет адаптерам достаточно гибко планировать и распределять нагрузку и другие ограниченные вычислительные ресурсы между всеми заданиями.

Задача планировщика в системе - формулировать и раздавать задания адаптерам и координировать их выполнение в пределах одного запроса. Любой запрос, обращающийся более чем к одному информационному ресурсу, может быть разбит на задания/подзапросы более чем одним способом. Выбор того или иного варианта разбиения может существенно повлиять на время выполнения запроса. Поиск эффективного варианта разбиения – важная задача планировщика в архитектуре посредников. Рассматриваемая реализация планировщика (в системе допускается одновременная работа любого количества различных реализаций планировщиков) использует для поиска эффективного разбиения эвристические предположения и имитацию выполнения запроса [4]

на небольших подмножествах данных там, где предположения не работают. В процессе имитации измеряются такие параметры, как:

- Скорость передачи данных;
- Скорость выполнения определенных заданий;
- Время ожидания в очереди.

Для оценки таких характеристик, как, например, селективность операций выборки по условию, используются данные, полученные в результате имитации. Таким же образом оцениваются результаты выполнения операций соединения. В результате, выбирается определенное разбиение запроса на задания. В процессе выполнения запроса планировщик контролирует, соответствуют ли измерения на реальных данных оценке, полученной на этапе имитации. Результаты измерений во время выполнения могут существенно отличаться по многим причинам, например, из-за существенного изменения нагрузки на тот или иной ресурс. Если планировщик обнаружит, что зафиксированные изменения существенно влияют на ожидаемое время выполнения запроса, то он, не прерывая исполнения запроса, попытается найти более эффективный способ выполнения оставшейся его части. Таким образом, алгоритм обработки запроса выглядит следующим образом:

1. Обработка запроса:
   - Разбиение на элементарные операции;
   - Оптимизирующие преобразования на основе правил;
2. Разбиение множества невыполненных операций на задания:
   - Оценка различных вариантов разбиения на основе эвристических предположений;
   - Имитация в случае, когда предположений недостаточно;
3. Выполнение плана:
   - Выполнение операций согласно плану;
   - Контроль соответствия результатов имитации/предположений реальной ситуации;
   - Возврат к шагу 2, если различия оказывают существенное влияние на прогноз времени выполнения запроса.

Эффективность такого способа планирования запросов, в сочетании с наличием прогноза времени выполнения запроса до начала выполнения достаточна для практического применения на запросах, требующих продолжительного исполнения и использующих большие объемы данных.

При реализации планировщика, функционирующего данным способом, очень важно правильно выбрать пространство планов, в котором будет осуществляться поиск, а также четко сформулировать правила адаптации плана и условия, при которых изменение плана считается необходимым.

## 3 Пространство возможных планов

Языки запросов Syfs и Asys поддерживают операции Join и Union. По грубой оценке, одно это обеспечивает рост количества возможных планов для данного запроса, как O(N!), где N, это количество задействованных в запросе коллекций. На практике, оно может быть больше и зависит также от количества обращениям к функциям в запросе. В планировщике применяются два распространенных способа сужения пространства поиска.

Во-первых, все операции Union переставляются в корень запроса, т.е. запрос представляется в виде объединения нескольких подзапросов, каждый из которых не содержит операций Union. Каждый подобный подзапрос планируется отдельно.

Во-вторых, при планировании подзапросов, не содержащих операций Union, рассматриваются только те планы, которые удовлетворяют следующим условиям:

- план не должен выполнять операцию перекрестного соединения, если этого не требуется пользователю;
- правый подзапрос-операнд каждой из операций Join не должен содержать операций Join.

Перечисленным ограничениям удовлетворяют только те планы, которые допускают вычисление посредством конвейерной обработки. Рост количества планов, удовлетворяющих заданным ограничениям можно оценить как O(NI), что приемлемо при использовании на практике.

## 4 Итерационное выполнение

В процессе вычисления запросов, для выполнения очередной операции Join часто возникает необходимость передать данные из одного ресурса в другой. В случае наличия альтернативных ресурсов, на которых может быть выполнена операция, выбор наилучшего варианта осуществляется посредством имитации выполнения подпланов на произвольных частях реальных данных. Адаптеры ресурсов накладывают ограничения на объем данных, который они могут единовременно загружать из других адаптеров или посредника. Объем данных, которые могут быть задействованы в запросе, теоретически не ограничен (в реализации есть ограничение в 2 в 63-ей степени байт). Для обработки больших массивов данных на запросе открывается ридер, из которого данные можно брать блоками любой заданной длины. Каждый из блоков обрабатывается отдельно и независимо и проходит все последующие этапы выполнения.

Планировщик контролирует время выполнения оставшейся части плана (и каждого этапа) для каждого из блоков данных. Если в определенный момент время выполнения становится существенно больше полученного во время имитации, то для всех

последующих этапов, начиная с данного, заново запускается поиск плана, и производится переоценка. Если более эффективный план найден, то оставшиеся блоки данных будут обрабатываться в соответствие с новым планом.

Таким образом, при выполнении запросов, использующих большие объемы данных, планировщик адаптирует план под изменяющиеся условия. При выполнении запросов, для выполнения которых нет необходимости в разбиении данных на блоки, изменений плана во время выполнения не происходит. Тем не менее, на таких запросах применимы и работают оценки на основе имитации.

# 5 Требования к адаптерам ресурсов

## 5.1 Традиционная архтектура адаптера

Спецификация интерфейса адаптеров неоднородных информационных ресурсов, применяемых в настоящее время в архитектуре средств поддержки предметных посредников, включает два метода:

```
public interface Adapter
{
        long executeAsyfsQuery(Plan plan) throws
AdapterException;
        void getAsyfsResult(long id, VOClassWriter
writer) throws AdapterException;
}
```

Таким образом, осуществляется возможность выполнения запроса на информационном ресурсе, и получение результата этого запроса. Отметим основные функциональные недостатки применяемых адаптеров:

1) Не поддерживаются сессии. Поскольку адаптеры функционируют в многопользовательских интеграционных средах, наличие сессий, изолированных друг от друга, с изолированными ресурсами является необходимым.

2) Не поддерживается возможность повторного использования результата запроса. В случае, когда одни и те же данные требуются в нескольких частях плана, запрос, получающий эти данные, придется выполнять несколько раз. Подобный подход неприемлем для эффективного планирования.

3) Не поддерживается возможность управления загрузкой данных в адаптер супервизором, а также возможность материализации запросов в адаптерах. В текущей реализации адаптерам передается план, в котором содержится URI и ID результата, по которому можно забрать результат. Для эффективного планирования требуется, чтобы момент загрузки данных и их объем определялись планировщиком динамически, а не заранее построенным планом.

4) Не поддерживаются учет способностей (Capabilities) адаптеров. Ресурсы обладают различными возможностями, например способностью хранить временные данные, выполнять соединения, и.пр. Для эффективного планирования необходимо учитывать все такие особенности.

5) Не поддерживаются оценочные запросы, необходимые для имитации. Для эффективного планирования необходимы оценки объема результата.

Все описанные недостатки были устранены в новой версии адаптера.

## 5.2 Перспективная архитектура адаптера

Спецификация интерфейса новых адаптеров является более сложной:

```
public interface RemoteAdapter
{
   boolean isSameResource(RemoteAdapter ra);
   void echoRequest() throws ConnectionException;
   void remoteEchoRequest(RemoteAdapter adapter);
   long getAdapterCapability(AdapterCapability
capability);
   boolean adapterSelectExtendedCapability(boolean
opOnSchemaData,String op, boolean
left_op_const,String left_op_type,boolean
right_op_const,String right_op_type);
   Module getAdapterScheme();

   int openSession();/*returns SessionID*/
   void closeSession(int sessionID);
   void keepAlive(int sessionID);
   long getSessionCapability(int sessionID,
SessionCapability capability);

   int registerQuery(int sessionID,Plan plan);  /*returns
QueryID*/
   void closeQuery(int sessionID,int queryID);

   int openReaderOnQuery(int sessionID, int
queryID,String rowIdAttributeName);
   /*returns ReaderID*/

   int materializeReader(int sessionID,int readerID);
   int materializeReader(int sessionID,int readerID,long
maxBytes,long maxRows);
   /*returns DataSetID*/

   int loadRemoteData(int sessionID,RemoteAdapter
adapter,int remoteSessionID,int remoteReaderID);
   /*returns DataSetID*/
   int loadRemoteData(int sessionID,RemoteAdapter
adapter,int remoteSessionID,int remoteReaderID,long
maxBytes,long maxRows);
   /*returns DataSetID*/

   void removeData(int sessionID,int dataSetID);


   long freeSpaceAvailable(int sessionID);   /*bytes*/
```

```
   String getData(int sessionID, int readerID, long
maxBytes, long maxRows);
   /*    *returns VOCLASS    */

   long dataAvailable(int sessionID, int readerID);
   long rowsSent(int sessionID,int readerID);
   long bytesSent(int sessionID,int readerID);
   void closeReader(int sessionID, int readerID);

   /*Query Evaluating Functions*/
   long estimateRows(int sessionID,int queryID);
   long estimateDataSize(int sessionID,int queryID);
   long averageRowSize(int sessionID,int queryID);
   int getSampleDataReader(int sessionID, int
queryID,long maxBytes, long maxRows, String
rowIdAttributeName);
   /*returns ReaderID*/

}
```

Функции *isSameResource, echoRequest, remoteEchoRequest, getAdapterScheme* необходимы для корректной работы с удаленными ресурсами.

Функции *getAdapterCapability, adapterSelectExtendedCapability, getSessionCapability, freeSpaceAvailable* осуществляют поддержку возможностей ресурсов, обладающих различными способностями.

Функции *openSession, closeSession, keepAlive* отвечают за поддержку сессий.

Функции *registerQuery, closeQuery* необходимы для передачи запросов ресурсам.

Функции *openReaderOnQuery, getData, dataAvailable, rowsSent, bytesSent, closeReader* осуществляют поддержку многократного использования результата запроса. На запрос открывается *reader*, посредством которого данные могут быть считаны как целиком, так и по частям. Кроме того предоставляются возможности получения статистики (сколько данных считано, остались ли данные).

Функции materializeReader, loadRemoteData, removeData отвечают за работу с временными коллекциями. В качестве источника для временных данных могут выступать как данные, полученные от других удаленных адаптеров, так и собственные данные, являющиеся результатом некоторого запроса, над которым открыт *reader*. Временные коллекции могут участвовать в запросах, как если бы это были внутренние данные ресурса.

Функции *estimateRows, estimateDataSize, averageRowSize, getSampleDataReader* предоставляют некоторые оценочные данные, на основе которых строится эффективный план.

### 5.3 Способности адаптеров (capabilities)

Различают два вида способностей (capabilities): способности сессии, и способности адаптера. Способности сессии описывают ресурсы выделенные под конкретную сессию. Способности адаптера описывают принципиальные возможности адаптера, как например возможность выполнения некоторых операций. Планы, которые не удовлетворяют ограничениям по ресурсам сессии и ограничениям способностей адаптера отбрасываются планировщиком во время поиска последовательным перебором возможных планов в заданном пространстве.

Рассмотрим подробнее все возможности.

**Описание способностей сессии.**

```
public class SessionCapability {
   public final int value;
      /*IntValues*/
   public static final SessionCapability
iMaxRowsInDataSet = new SessionCapability(0);
   public static final SessionCapability
iMaxDataSetSize = new SessionCapability(1);
   public static final SessionCapability
iMaxDataSpacePerSession = new SessionCapability(2);
   public static final SessionCapability
iMaxDataSetsPerSession = new SessionCapability(3);
   public static final SessionCapability
iMaxRegistredQuerysPerSession = new
SessionCapability(4);
   public static final SessionCapability
iMaxOpenReadersPerSession = new
SessionCapability(5);
   public SessionCapability(int v)
   {      value=v;   }
}
```

iMaxRowsInDataSet – результатом вызова getSessionCapability должно быть максимальное количество кортежей в одной временной коллекции или 0, если временные коллекции не поддерживаются данным адаптером.

iMaxDataSetSize – результатом вызова getSessionCapability должен быть максимальный размер одной временной коллекции в байтах или 0, если временные коллекции не поддерживаются данным адаптером.

iMaxDataSpacePerSession - результатом вызова getSessionCapability должен быть максимальный суммарный размер временных коллекций в пределах данной сессии, или 0, если временные коллекции не поддерживаются данным адаптером.

iMaxDataSetsPerSession - результатом вызова getSessionCapability должно быть максимальное количество одновременно существующих временных коллекций в пределах данной сессии или 0, если временные коллекции не поддерживаются данным адаптером.

iMaxRegistredQuerysPerSession - результатом вызова getSessionCapability должно быть максимальное количество одновременно существующих зарегистрированных запросов в пределах сессии. Минимальное значение 1.

iMaxOpenReadersPerSession - результатом вызова getSessionCapability должно быть максимальное количество одновременно открытых ридеров в пределах сессии. Минимальное значение – 1.

На всех остальных значениях getSessionCapability должна возвращать 0.

**Описание способностей адаптера.**

```
public class AdapterCapability {
   public final int value;
   public static final AdapterCapability
UserDataSetsSupport = new AdapterCapability(0);
   public static final AdapterCapability
UserDataSetsComplexQuery = new
AdapterCapability(1);
   public static final AdapterCapability
RemoteUserDataLoad = new AdapterCapability(2);
   public static final AdapterCapability
CanExecuteSchemaJoin = new AdapterCapability(4);
   public static final AdapterCapability
CanExecuteTempJoin = new AdapterCapability(5);
   public static final AdapterCapability
CanExecuteMixedJoin = new AdapterCapability(6);
   public static final AdapterCapability
CanExecuteSchemaUnion = new AdapterCapability(7);
   public static final AdapterCapability
CanExecuteTempUnion = new AdapterCapability(8);
   public static final AdapterCapability
CanExecuteMixedUnion = new AdapterCapability(9);
   public static final AdapterCapability
CanExecuteSchemaSelect = new
AdapterCapability(10);
   public static final AdapterCapability
CanExecuteTempSelect = new AdapterCapability(11);
   public static final AdapterCapability
SupportsExtendedSelectCapability = new
AdapterCapability(18);
   public static final AdapterCapability
CanExecuteSchemaSetTypeOperations = new
AdapterCapability(12);
   public static final AdapterCapability
CanExecuteSetTypeOperationsOnTempData = new
AdapterCapability(13);
   public static final AdapterCapability
CanAcceptTempSetTypeData = new
AdapterCapability(14);
   public static final AdapterCapability
AllowSchemaMethodsOnTempData = new
AdapterCapability(15);
   public static final AdapterCapability
CanExecuteTempAppendID = new
AdapterCapability(16);
   public static final AdapterCapability
CanExecuteSchemaAppendID = new
AdapterCapability(17);
   public AdapterCapability(int v)
   {      value=v;      }
}
```

UserDataSetsSupport - Если возвращается значение отличное от нуля, это значит, что адаптер поддерживает временные коллекции. Это в свою очередь означает, что в этом адаптере обязательно должны быть реализованы функции materializeReader, removeData, freeSpaceAvailable.

UserDataSetsComplexQuery - Если возвращается значение равное нулю, это означает, что использование временных коллекций в запросах сильно ограничено. Единственным допустимым запросом с использованием временных коллекций при этом будет запрос, который запрашивает без каких либо изменений все данные, содержащиеся в одной конкретной коллекции.

RemoteUserDataLoad - Если возвращается отличное от нуля значение, это означает, что адаптер реализует и поддерживает методы loadRomoteData и remoteEchoRequest. Фактически, это означает, что адаптер поддерживает загрузку данных, поступающих от других адаптеров.

CanExecuteSchemaJoin - Если возвращается отличное от нуля значение, это означает что поддерживается операция Join над двумя и более коллекциями, объявленными в схеме источника (не временными).

CanExecuteTempJoin - Если возвращается отличное от нуля значение, это означает, что на адаптере допустимо выполнение операции Join над двумя или более временными коллекциями.

CanExecuteMixedJoin - Если возвращается отличное от нуля значение, это означает, что адаптер способен выполнить операцию Join над парой операндов, один из которых является результатом выполнения каких-либо допустимых на этом адаптере операций над коллекциями, объявленными в схеме, а второй над временными коллекциями. Над результатом выполнения подобной операции должны быть выполнимы любые операции, которые разрешено выполнять над временной коллекцией.

CanExecuteSchemaUnion - Если возвращается отличное от нуля значение, это означает что поддерживается операция Union над двумя и более коллекциями, объявленными в схеме источника (не временными).

CanExecuteTempUnion - Если возвращается отличное от нуля значение, это означает, что на адаптере допустимо выполнение операции Union над двумя или более временными коллекциями.

CanExecuteMixedUnion - Если возвращается отличное от нуля значение, это означает, что адаптер способен выполнить операцию Union над парой операндов, один из которых является результатом выполнения каких-либо допустимых на этом адаптере операций над коллекциями, объявленными в схеме, а второй над временными коллекциями. Над результатом выполнения подобной операции должны быть выполнимы любые операции, которые разрешено выполнять над временной коллекцией.

CanExecuteSchemaSelect - Если возвращается отличное от нуля значение при вызове с этим параметром, это означает что поддерживается операция Select над данными в составе схемы (и любыми данными, которые можно получить в результате допустимых операций над данными в схеме) в полном объеме.

CanExecuteTempSelect - Если возвращается отличное от нуля значение при вызове с этим параметром, это означает что поддерживается операция Select над временными данными в составе

схемы (и любыми данными, которые можно получить в результате допустимых операций над временными данными) в полном объеме.

SupportsExtendedSelectCapability - Если функция с этим параметром возвращает ненулевое значение, то в адаптере должна быть корректно реализована функция adapterSelectExtendedCapability.

CanExecuteSchemaSetTypeOperations - Если возвращается отличое от нуля значение при вызове с этим параметром, это означает что поддерживаются операции над множествами, над данными объявленными в схеме источника. Если в схеме источника содержатся множества, то операции над множествами обязаны поддерживаться.

CanAcceptTempSetTypeData - Если возвращается отличное от нуля значение при вызове с этим параметром, это означает, что во временных коллекциях могут содержаться атрибуты - множества.

CanExecuteSetTypeOperationsOnTempData - Если возвращается отличное от нуля значение, допустимо использование операций над множествами над временными данными

AllowSchemaMethodsOnTempData - Если возвращается отличное от нуля значение при вызове с этим параметром, это означает что все методы, объявленные в схеме источника, могут быть использованы, в том числе и над временными данными.

CanExecuteSchemaAppendID - Если возвращается ненулевое значение при вызове с этим параметром, это означает, что адаптер может выполнять операции Append с Id предикатом над данными, объявленными в схеме источника. В виде Id предикатов будут представлены все операции преобразования типов атрибутов, а также арифметические операции (могут быть логические, над булевыми аргументами).

CanExecuteTempAppendID - Если возвращается ненулевое значение при вызове с этим параметром, это означает, что адаптер может выполнять операции Append с Id предикатом над временными данными. В виде Id предикатов будут представлены все операции преобразования типов атрибутов, а также арифметические операции (могут быть логические операции над булевыми аргументами).

## 6 Заключение

Предложенное применение оценок на основе имитации выполнения планов, а также итерационное выполнение запросов, позволяет уже в достаточно простой задаче интеграции информационных ресурсов, такой, как задача поиска далеких галактик, описанная в публикации [1], для решения которой необходимо интегрировать несколько объемных каталогов астрономических объектов, снять ограничения на объем задействованных в выполнении запроса данных и, в зависимости от загрузки задействованных в её решении ресурсов, повысить эффективность выполнения запросов.

## Литература


[1] Брюхов Д.О., Вовченко А.Е., Захаров В.Н., Желенкова О.П., Мартынов Д.О., Скворцов Н.А., Ступников С.А. Архитектура промежуточного слоя предметных посредников для решения задач над множеством интегрируемых неоднородных распределенных информационных ресурсов в гибридной грид-инфраструктуре виртуальных обсерваторий. Информатика и ее применения, Т. 2, вып. 1, 2008, стр. 2--34.

[2] Захаров В.Н., Калиниченко Л.А., Соколов И.А., Ступников С.А. Конструирование канонических информационных моделей для интегрированных информационных систем. Информатика и ее применения, Т. 1, вып. 2, 2007, стр. 15--38.

[3] Briukhov D.O., Kalinichenko L.A., Martynov D.O. Source Registration and Query Rewriting Applying LAV/GLAV Techniques in a Typed Subject Mediator. Proc. of the Ninth Russian Conference on Digital Libraries RCDL'2007.

[4] Christian Wiesner. Query Evaluation Techniques for Data Integration Systems. 2004. http://www.opus-bayern.de/uni-passau/volltexte/2004/40/pdf/QETechniquesForDISystems.pdf

[5] Kalinichenko L.A., Briukhov D.O., Martynov D.O., Skvortsov N.A., Stupnikov S.A. Mediation Framework for Enterprise Information System Infrastructures. Proc. of the 9th International Conference on Enterprise Information Systems ICEIS 2007. -- Funchal, 2007. -- Volume Databases and Information Systems Integration. -- P. 246--251.

[6] Kalinichenko L.A., Stupnikov S.A., Martynov D.O. SYNTHESIS: a Language for Canonical Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments. Moscow: IPI RAN, 2007.

[7] Mostafa Elhemali, Cйsar A. Galindo-Legaria, Torsten Grabs, Milind M. Josh. Execution Strategies for SQL Subqueries. Proceedings of the 2007 ACM SIGMOD international conference on Management of data. P. 993 – 1004.

[8] Surajit Chaudhuri. An Overview of Query Optimization in Relational Systems. Symposium on Principles of Database Systems. 1998. P.. 34 - 43.

[9] Volker Markl, Vijayshankar Raman, David Simmen, Guy Lohman, Hamid Pirahesh, Miso Cilimdzic. Robust Query Processing through Progressive Optimization. Proceedings of the ACM SIGMOD international conference on Management of data. P. 659 – 670. 2004.

[10] Yannis E. Ioannidis. Query Optimization. ACM Computing Surveys, Volume 28, Issue 1. March 1996. P. 121 – 123.

[11] Zachary G. Ives, Daniela Florescu, Marc Friedman, Alon Levy, Daniel S. Weld An Adaptive Query Execution System for Data Integration. Proceedings of the 1999 ACM SIGMOD international conference on Management of data. P. 299 – 310.

## Query planning over heterogeneous distributed information resources in the architecture of the subject mediators

© Vovchenko A.E., Krupa A.V.

Query planner plays one of the major roles in the architecture of subject mediators. Planner is responsible for decomposition and execution of the user query over a set of heterogeneous information resources. The decomposition problem in the mediator architecture implies a set of acceptable decisions, therefore one of the major tasks of the planner is the choice of the most effective one.

In the article the description of the planner which is able to effectively execute the queries formulated over information resources, containing large-scale sets of data (in comparison with the bandwidth of communication channels). Such queries usually require long time for execution. Also in the article the description of requirements to resource wrappers is presented. A wrapper architecture was developed to provide the planner with a possibility of construction of effective execution plans.

♣ Работа выполнена при частичной финансовой поддержке РФФИ (грант 08-07-00157-а) и Программы фундаментальных исследований Президиума РАН № 3 «Фундаментальные проблемы системного программирования» (проект «Исследование методов и средств промежуточного слоя предметных посредников, обеспечивающего решение задач над множеством неоднородных распределенных информационных ресурсов»).

# Формирование выражений взглядов в задаче регистрации ресурсов в предметных посредниках♣

© Рябухин О.В. Брюхов Д.О. Калиниченко Л.А.

Институт проблем информатики РАН
{ovr, brd, leonidk}@ipi.ac.ru

**Аннотация**

Целью данной работы является развитие технологии регистрации информационных ресурсов в предметных посредниках. Основное внимание уделено заключительному этапу данного процесса – формированию выражений взглядов. Этот этап рассматривается как полуавтоматический процесс, подразумевающий участие эксперта в ходе принятия решений. В работе рассмотрена возможная его организация, позволяющая системе регистрации предложить эксперту возможные решения.

## 1 Введение

### 1.1 Концепция предметных посредников

В настоящее время в мире наблюдается экспоненциальный рост количества информации, накапливаемой в различных предметных областях в неоднородных распределенных информационных ресурсах (базах данных, сервисах, процессах). При решении задач в этих областях появляется необходимость совместного использования таких ресурсов, их интеграции. Концепция предметных посредников ориентирована на поддержку решения задач над множеством неоднородных информационных ресурсов. В качестве канонической информационной модели в рассматриваемой архитектуре предметных посредников используется язык СИНТЕЗ [6]. Занимая место между пользователем и ресурсами, предметный посредник предоставляет общий интерфейс доступа к ним; использование канонической информационной модели позволяет преодолеть модельную неоднородность ресурсов, а исчисление спецификаций дает возможность проведения интеграции спецификаций ресурсов в спецификациях посредника. Концепция интеграции ресурсов рассматривает два подхода, позволяющих решать задачи над множеством неоднородных ресурсов: подход, движимый задачей (приложением), и подход, движимый ресурсами. В подходе, движимом ресурсами, глобальная схема строится как композиция множества схем имеющихся информационных ресурсов. Недостатком такого решения является плохая масштабируемость, т.к. при появлении очередного ресурса глобальная схема требует изменений. В основе концепции предметных посредников лежит движение от задачи к ресурсам. При подобном подходе, глобальная схема (концептуальная схема предметной области) строится независимо от конкретных информационных ресурсов и утверждается экспертным сообществом в данной предметной области. Также в архитектуре предметных посредников используются техники виртуальной интеграции, при которых виртуальные классы глобальной схемы представляют собой взгляды [7] над классами ресурсов (техника Global-as-View (GAV)), либо классы ресурсов являются взглядами над виртуальными классами глобальной схемы (техника Local-as-View (LAV)). Подходу предметных посредников, в котором глобальная схема не зависит от схем ресурсов, органично соответствует техника LAV, при которой классы ресурса выражаются как композиция классов посредника. Следует, однако, особо отметить, что при интеграции спецификаций неизбежно возникают различного рода конфликты. Конфликты могут возникать как из-за разных прикладных областей, так и из-за различного представления спецификаций семантически близких сущностей. Для устранения рассогласований используются методы структурных преобразований, а также метод использования функций разрешения конфликтов, описанных на языке высокого порядка. Очевидно, что такие функции должны соответствовать движению результатов запросов от ресурса к посреднику, и задавать соответствующие преобразования. Следовательно, функции разрешения конфликтов удобно разместить в теле GAV – правила. С учетом этого, использование техники GLAV (Global-Local-as-View) предпочтительно. GLAV соединяет в себе преимущества техник GAV и LAV, предоставляя возможность использования конструкций,

**Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

соответствующих классам ресурса, в голове LAV – правила, позволяя, таким образом, формировать выражения GAV.

### 1.2 Процесс регистрации

Выражения взглядов формируются в заключительной стадии так называемого процесса регистрации информационных ресурсов в предметном посреднике. Процесс регистрации [4] представляет собой последовательность согласованных действий эксперта и системы регистрации ресурсов, в ходе которой происходит поиск ресурсов-кандидатов для регистрации по нефункциональным требованиям (поиск по метаданным), выбор среди них ресурсов, онтологически релевантных посреднику, сопоставление спецификаций ресурса и посредника, и, наконец, формирование выражений GLAV взглядов. Процесс рассматривается как полуавтоматический: на каждом его шаге система регистрации ресурсов предлагает эксперту возможное решение (например, список пар онтологически релевантных типов), после чего эксперт подтверждает (либо корректирует) предложенное решение.

### 1.3 Формирование выражений взглядов

К моменту написания данной работы, заключительный этап процесса регистрации – шаг формирования выражений взглядов – подразумевал полную его реализацию экспертом (в частности, в работе [3] приведены примеры взглядов, определенных экспертом вручную, по соотношениям между классами ресурса и посредника). В данной работе процесс формирования выражений взглядов (аналогично другим этапам процесса регистрации) также рассматривается как полуавтоматический, описывается возможная его организация, позволяющая системе регистрации предложить эксперту возможные решения.

## 2 Общие понятия

Конечной целью процесса регистрации, помимо формирования выражений взглядов, является достижение отношения уточнения между спецификациями ресурса и посредника. Говоря неформально, можно утверждать, что спецификация A уточняет спецификацию B, если пользователь не замечает подмены при использовании A вместо B. Очевидно, что не каждая спецификация типа ресурса уточняет спецификацию соответствующего типа посредника, поэтому в процессе регистрации необходимо произвести поиск необходимых фрагментов. Специальное композиционное исчисление спецификаций [5], которое формализует процесс композиции и декомпозиции их фрагментов, позволяет решать задачи в таких условиях. Важнейшими здесь являются понятие отношения уточнения, понятия редукта типа и наиболее общего редукта двух типов, а также операции композиции типов.

**Определение.** Редукт $R_T$ типа T определяется как подсигнатура сигнатуры типа, при этом множество атрибутов редукта является подмножеством множества атрибутов типа, множество функций редукта является подмножеством множества функций типа, множество предикатов редукта является подмножеством множества предикатов типа.

**Определение.** Наиболее общий редукт $R_{MC}(T_1,T_2)$ для типов $T_1$, $T_2$ есть редукт $R_{T1}$ типа $T_1$ такой, что существует редукт $R_{T2}$ типа $T_2$ который является уточнением редукта $R_{T1}$ и не существует другого редукта $R'_{T1}$ такого, что $R_{MC}(T_1,T_2)$ является редуктом $R'_{T1}$, $R'_{T1}$ не равен $R_{MC}(T_1,T_2)$ и существует $R'_{T2}$ являющийся уточнением редукта $R'_{T1}$. Редукт $R_{T2}$ называется сопряженным по отношению к $R_{T1}$.

**Определение.** Операция meet. Операция образует тип T как пересечение спецификаций типов-операндов. Тип T образуется слиянием двух наиболее общих редуктов типов $T_1$ и $T_2$ : $R_{MC}(T_1,T_2)$ и $R_{MC}(T_2,T_1)$. Слияние двух редуктов включает в себя объединение множеств их операций. Если в объединении имеется 2 операции находящиеся в отношении уточнения, то только одна из них – более абстрактная – включается в результирующий тип T. Инвариант результирующего типа формируется как дизъюнкция инвариантов, взятых из спецификаций наиболее общих редуктов.

**Определение.** Операция join. Операция образует тип T как объединение спецификаций типов-операндов. Тип T образуется слиянием спецификаций типов $T_1$ и $T_2$. Общие элементы спецификаций типов $T_1$ и $T_2$ включаются в результирующий тип только один раз. Общие элементы определяются посредством слияния сопряженных редуктов двух наиболее общих редуктов типов $T_1$ и $T_2$: $R_{MC}(T_1,T_2)$ и $R_{MC}(T_2,T_1)$. Слияние сопряженных редуктов включает в себя объединение множеств их операций. Если в объединении имеется 2 операции находящиеся в отношении уточнения, то только одна из них – более точная – включается в результирующий тип T. Инвариант результирующего типа формируется как конъюнкция инвариантов, взятых из спецификаций типов операндов $T_1$ и $T_2$.

## 3 Процесс регистрации

В общих чертах, процесс регистрации сводится к следующим основным шагам:

1. Поиск информационных ресурсов по нефункциональным требованиям к ним (поиск по метаданным)
2. Поиск среди найденных кандидатов для регистрации ресурсов, онтологически релевантных посреднику
3. Идентификация типов посредника, онтологически релевантных типам ресурса.

4. Построение наиболее общих редуктов для каждой пары релевантных типов. Если при этом возникают конфликты между спецификациями, на этом шаге формируются функции разрешения конфликтов

5. Формирование взглядов

На этапах 1-3 осуществляется отбор ресурсов-кандидатов для регистрации в посреднике, при этом используются методы онтологического поиска, а также поиск по метаданным. Далее в работе подразумевается, что для регистрации выбран конкретный информационный ресурс, и, с помощью онтологических методов, получен набор попарно релевантных типов ресурса и посредника.

В работе [1] показаны возможные пути автоматизации шага 4 процесса регистрации. Приведен алгоритм построения наиболее общего редукта, основанный на проверке формального определения уточнения для спецификаций типов. Рассмотрены методы автоматизации разрешения конфликтов, основанные на применении правил структурных преобразований и использовании функций разрешения конфликтов, описываемых на языке высокого порядка. Техника применения правил структурных преобразований основывается на методе релевантных путей. Функции разрешения конфликтов описываются на языке Syfs (Synthesis Formula Subset), который является вариантом типизированного языка логики первого порядка.

## 4 Формирование выражений взглядов

Процесс формирования выражений взглядов представляет собой последовательность согласованных действий эксперта и системы регистрации ресурсов. Используемое далее в описании структуры процесса понятие дерева типа заимствовано из работы [1] и вводится следующим образом:

**Определение**. Дерево типа *T* строится по следующим правилам:

1. корнем дерева является тип T
2. вершинами дерева уровня *i+1* являются типы $\{T_{i+1}\}$ непосредственно связанные с произвольным типом $T_i$ уровня i посредством атрибута-ссылки ($T_i$–*a*→$T_{i+1}$) и не содержащиеся в пути между типами *T* и $T_i$
3. ребрами дерева являются атрибуты-ссылки
4. листьями дерева являются атрибуты-значения типов

При построении дерева для устранения возможных циклов, если очередной тип, возникающий при построении дерева, уже встречался в пути к нему от вершины дерева, то новой вершины не вводится, и соответствующий атрибут-ссылка не рассматривается. Если же тип еще не встречался в пути, независимо от того, встречается он в дереве или нет, вводится новая вершина (являющаяся, таким образом, дублирующей, если такой тип уже встречался в дереве). Т.е. каждый раз при появлении очередной вершины вводится новый набор ее вершин-потомков, что исключает образование циклов.

Дополнительных пояснений требует формирование наиболее общих редуктов, при котором некоторым элементам спецификации типа $T_1$ требуется ставить в соответствие специальные функции разрешения конфликтов, включаемые в спецификацию сопряженного редукта. Таким образом происходит преобразование сопряженных редуктов в конкретизирующие.

Далее структура процесса формирования взглядов рассматривается с точки зрения системы регистрации, действия эксперта и комментарии выделены курсивом (во всех случаях вмешательства эксперта без особых указаний предполагается, что его действия включают подтверждение или корректировку предложенного системой решения).

Для каждого класса ресурса:

1. Взять тип экземпляров данного класса (тип A)
2. Найти релевантные ему типы посредника, которые являются типами экземпляров каких-либо классов посредника
3. Выдать результат эксперту

   *Полученный экспертом набор релевантных типов используется для идентификации релевантных классов так, что если тип ресурса релевантен типу посредника, и данные типы являются типами экземпляров каких-либо классов, то эти классы возможно релевантны*
4. Для каждой пары таких типов (тип A – релевантный ему тип посредника), формировать выражения взглядов, определяющие соответствующий типу ресурса класс ресурса через соответствующий типу посредника класс посредника:
5. Построить деревья типа ресурса (типа A) и релевантного ему типа посредника (типа B)
6. В построенных деревьях типов для каждой пары релевантных типов построить возможный вариант их наиболее общего редукта, сформировать возможные функции разрешения конфликтов, выдать результат эксперту

   *Правильность автоматического построения наиболее общего редукта, а также функций разрешения конфликтов в случае структурных рассогласований целиком зависит от правильности установления отношения релевантности между типами ресурса и посредника, что требует тщательной проверки экспертом. Автоматическое построение функций разрешения конфликтов в случае сложных зависимостей не представляется возможным.*
7. Для последующей типизации голов GAV и LAV правил [2], сформировать композиционный тип соединением построенных ранее редуктов (композиционный тип C) и соответствующих им конкретизирующих редуктов (композиционный тип D). При построении композиционного типа использовать переименование атрибутов во

избежание конфликтов имен. Выдать результат эксперту.

*Полученный экспертом композиционный тип должен определять совокупность общих элементов спецификаций в деревьях типов ресурса и посредника, находящихся в отношении уточнения.*

8. Сформировать голову правил GAV и LAV, типизированную полученным композиционным типом.

9. Сформировать тело GAV – правила:

10. Указать в нем соответствующий класс ресурса, типизированный композицией конкретизирующих редуктов (тип D), с добавлением соответствующих функций разрешения конфликтов

11. Сформировать тело LAV – правила:

12. Для каждого атрибута композиционного типа, не вошедшего в редукт типа посредника (типа B), произвести поиск пути к соответствующему атрибуту в его дереве. Выдать результат эксперту.

13. Если для класса ресурса получилось несколько GLAV – правил, произвести их композицию, выразив, в конечном итоге, класс ресурса через композицию классов посредника. Выдать результат эксперту.

Описанный процесс иллюстрируется на сокращенной адаптации примера, рассматриваемого в [2]. Предметной областью являются задачи поиска далеких галактик [2], которые были сформулированы в контексте подхода, развиваемого на основе виртуальных обсерваторий (Virtual Observatory, VO). Далее в примере рассматривается единственный информационный ресурс SDSS (Sloan Digital Sky Survey), содержащий обзор части видимого звездного неба.

Схема посредника включает в себя следующие типы (здесь и далее используется нотация языка СИНТЕЗ [6]):

```
{CoordEQJ; in: type;
   ra: real;
   de: real;
}
{CatalogData; in: type;
   name: string;
   coord: CoordEQJ;
}
{Magnitude; in: type;
   magValue: real;
   magError: real;
   filter: string;
}
{OpticalCatalogData; in: type; supertype: CatalogData;
   colorIndexURG: real;
   deltaColorIndexURG: real;
   magnitude: {set; type_of_element: Magnitude;};
};
```

Тип CatalogData описывает данные, накапливаемые в различных каталогах астрономических объектов, атрибут name определяет имя объекта, атрибут coord – координаты: атрибут ra типа CoordEQJ соответствует прямому восхождению, de - склонению. Подтип OpticalCatalogData типа CatalogData описывает данные, накапливаемые в оптических каталогах. Атрибуты colorIndexURG и deltaColorIndexURG имеют семантику показателей цвета. Тип Magnitude соответствует понятию звездной величины.

Посредник содержит единственный класс (instance_section определяет тип экземпляров данного класса):

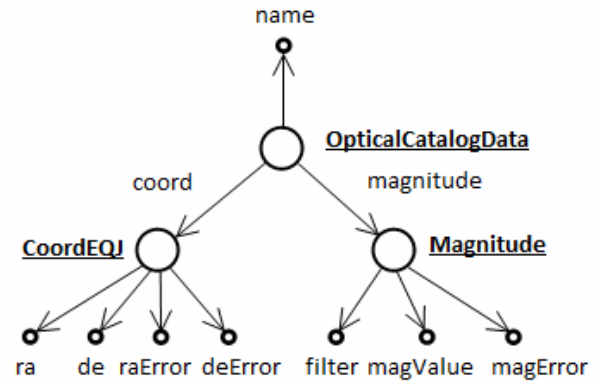
```
class_specification:
   {opticalCatalogData; in: class; instance_section: OpticalCatalogData;
   };
```

SDSS содержит фотометрический обзор неба в фильтрах u, g, r, i, z, его схема описана следующим образом:

```
{PhotoPrimary; in: type;
   objID: integer;
   ra: real;
   dec: real;
   u: real;
   err_u: real;
   g: real;
   err_g: real;
   i: real;
   err_i: real;
   r: real;
   err_r: real;
   z: real;
   err_z: real;
};
class_specification:
{photoPrimary; in: class;
   instance_section: PhotoPrimary;
};
```

Атрибут objID представляет собой идентификатор астрономического объекта, ra и dec определяют его координаты, остальные атрибуты соответствуют фотометрическим параметрам.

Опуская описание предыдущих шагов процесса регистрации, положим, что тип экземпляров класса photoPrimary ресурса PhotoPrimary, релевантен трем типам посредника: OpticalCatalogData, CoordEQJ, Magnitude. Из этих трех типов, тип OpticalCatalogData является типом экземпляров класса посредника opticalCatalogData. Следовательно, единственный GLAV – взгляд будет выражать класс посредника photoPrimary через класс ресурса opticalCatalogData. Дерево типа PhotoPrimary состоит из вершины, соответствующей данному типу, и вершин, соответствующих атрибутам. Дерево типа OpticalCatalogData выглядит следующим образом:




В данных деревьях типов имеем три пары релевантных типов: PhotoPrimary ~ OpticalCatalogData, PhotoPrimary ~ CoordEQJ, PhotoPrimary ~ Magnitude. Заметим, что построение деревьев типов и поиск в них пар релевантных типов позволяет локализовать поиск общей информации между классами ресурса и посредника.

Далее для каждой пары релевантных типов строятся наиболее общий и соответствующий конкретизирующий редукты. Заметим, что для этого применяются уже упоминавшиеся выше алгоритм построения наиболее общих редуктов, методы структурных преобразований, задание функций разрешения конфликтов.

Наиболее общий редукт типов OpticalCatalogData и PhotoPrimary выглядит следующим образом:

```
{R_OData_SDSS; in: reduct;
 metaslot of: OpticalCatalogData;
 taking: {name, coord, magnitude};
 c_reduct: CR_OData_SDSS;
 end;
};
```

Ключевые слова of, taking, c_reduct указывают, что R_OData_SDSS является редуктом типа OpticalCatalogData, в нем остаются атрибуты name, coord и magnitude, и CR_OData_SDSS является соответствующим конкретизирующий редуктом.

```
{CR_OData_SDSS; in: c_reduct;
 metaslot of: PhotoPrimary;
 taking: {objID};
 reduct: R_OData_SDSS;
end;
simulating: {
 R_OData_SDSS.name ~ CR_OData_SDSS.objID,
 R_OData_SDSS.coord ~
CR_OData_SDSS.get_coord,
 R_OData_SDSS.magnitude ~
CR_OData_SDSS.get_magnitude;
 get_coord: {in: function; params:{ -returns/
 {set; type_of_elements: CoordEQJ;}};
 {{ returns' = this/CR_Coord_SDSS }}
};
 get_magnitude: {in: function; params:{ -
 returns/ {set; type_of_elements:
 Magnitude;}};
 {{ returns' = this/CR_Magnitude_SDSS }}
};
};
```

Здесь секция simulating указывает на отношение уточнения между соответствующими элементами (например, атрибут objID ресурса соответствует атрибуту name посредника). Аналогично строятся остальные редукты: R_Magnitude_SDSS, R_Coord_SDSS, и соответствующие им конкретизирующие редукты CR_Magnitude_SDSS, CR_Coord_SDSS. Далее необходимо выделить наибольшую совокупность элементов спецификаций, находящихся в отношении уточнения в построенных ранее деревьях типов. Такой наибольший фрагмент может быть описан с помощью специального композиционного типа, полученного путем применения операции *join* к полученным ранее редуктам и соответствующим им конкретизирующим редуктам с соответствующими переименованиями.

```
CT_OData[name, coord, coord_ra, coord_de,
magnitude, magnitude_magValue,
magnitude_magError, magnitude_filter] =
R_OData_SDSS[name, coord, magnitude] |
R_Magnitude_SDSS[magnitude_magValue:
magValue, magnitude_magError:magError,
magnitude_filter:filter] |
R_Coord_SDSS[coord_ra:ra, coord_de:de]

CT_SDSS[objID, get_coord, get_magnitude, u,
err_u, g, err_g, i, err_i, r, err_r, z,
err_z, get_magValue, get_magError,
get_filter, ra, dec] = CR_OData_SDSS[objID,
get_coord, get_magnitude] |
CR_Magnitude_SDSS[u, err_u, g, err_g, i,
err_i, r, err_r, z, err_z, get_magValue,
get_magError, get_filter] |
CR_Coord_SDSS[ra, dec]
```

Идентификатор типа с добавленными квадратными скобками специфицирует редукт данного типа, включающий атрибуты, перечисленные в скобках. Вертикальная черта обозначает операцию *join* композиции типов.

Выражения взглядов формулируются с помощью формул на языке Syfs, который является вариантом типизированного языка логики первого порядка. Предикаты в формулах соответствуют классам и функциям. Правила имеют вид:

$$Q(v/T_v) :- C_1(v_1/T_{v1}),..,C_n(v_n/T_{vn}),F_1(t_1,y_1),..,F_m(t_m,y_m),B$$

где $C_i$ – предикат класса, $F_i$ – функциональный предикат, $B$ – ограничение. Правила представляют собой конъюнкцию предикатов и функций. Переменные $v, v_1, \ldots, v_n$, типизированы $T_{v1}, \ldots, T_{vn}$ соответственно.

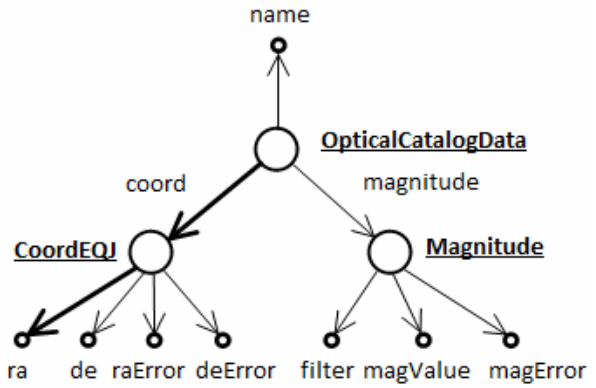
Определим голову взглядов GAV и LAV: она типизируется композиционным типом CT_OData:

```
v_SDSS_OData(x/CT_OData[name, coord,
coord_ra, coord_raError, coord_de,
coord_deError, magnitude,
magnitude_magValue,
magnitude_magError, magnitude_filter])
```

Для формирования тела GAV – взгляда допишем в него класс ресурса, типизированный редуктом композиционного типа CT_SDSS, включающим в себя все необходимые атрибуты, а также перечислим функции разрешения конфликтов, полученные при построении конкретизирующих редуктов.

```
v_SDSS_OData(x/CT_OData[name, coord,
coord_ra, coord_raError, coord_de,
coord_deError, magnitude,
magnitude_magValue,
magnitude_magError, magnitude_filter])
:- photoPrimary(x/CT_SDSS[name:objID,
coord_ra: ra, coord_de: dec])
& get_coord(x, coord)
& get_magnitude(x, magnitude)
& get_magValue(x, magnitude_magValue)
& get_magError(x, magnitude_magError)
& get_filter (x, magnitude_filter)
```

Для построения тела LAV – правила, допишем в него класс посредника, типизированный редуктом типа его экземпляров (в данном случае класс opticalCatalogData, тип его экземпляров OpticalCatalogData). Для каждого атрибута композиционного типа, не вошедшего в тип OpticalCatalogData, ищем путь к соответствующему атрибуту в дереве типа. Например, для атрибута coord_ra типа CT_OData:




Следовательно, в редукте типа OpticalCatalogData атрибут, соответствующий coord_ra, может быть найден по пути coord.ra. В итоге:

```
v_SDSS_OData(x/CT_OData[name, coord,
coord_ra, coord_de, magnitude,
magnitude_magValue, magnitude_magError,
magnitude_filter]):-
opticalCatalogData(x/OpticalCatalogData
[name, coord, magnitude,
coord_ra:coord.ra, coord_de:coord.de,
magnitude_magValue: magnitude.magValue,
magnitude_magError: magnitude.magError,
magnitude_filter: magnitude.filter])
```

## 5 Заключение

Концепция предметных посредников позволяет создать инструментарий для решения задач над множеством неоднородных информационных ресурсов. В данной работе были затронуты вопросы, связанные с технологией регистрации ресурсов в посредниках. Заключительному шагу данного процесса уделено особое внимание. Формирование выражений взглядов определено как полуавтоматический процесс, представляющий собой последовательность согласованных действий эксперта и системы регистрации ресурсов. В результате выполнения определенных шагов система предоставляет эксперту возможное решение, которое может быть подтверждено, или скорректировано. Для построения возможных решений системой рассмотрен подход, в основе которого лежит использование соответствий между спецификациями ресурсов и посредника, полученных в результате построения конкретизирующих редуктов. Для локализации поиска релевантных типов в рамках релевантных классов и поиска путей при формировании LAV – правил используются деревья типов. Предлагаемый подход позволяет включить процесс формирования выражений взглядов в общий процесс регистрации для согласования действий эксперта и системы регистрации ресурсов.

## Литература


[1] Брюхов Д.О. Конструирование информационных систем на основе интероперабельных сред информационных ресурсов. -- Москва: ИПИ РАН, 2003. -- 158 с.

[2] Брюхов Д.О., Вовченко А.Е., Захаров В.Н., Желенкова О.П., Калиниченко Л.А., Мартынов Д.О., Скворцов Н.А., Ступников С.А. Архитектура промежуточного слоя предметных посредников для решения задач над множеством интегрируемых неоднородных распределенных информационных ресурсов в гибридной грид-инфраструктуре виртуальных обсерваторий. Информатика и ее применения -- 2008. -- Т. 2, Вып. 1. -- С. 2--34.

[3] Briukhov D.O., Kalinichenko L.A., Martynov D.O. Source Registration and Query Rewriting Applying LAV/GLAV Techniques in a Typed Subject Mediator. Proc. of the Ninth Russian Conference on Digital Libraries RCDL'2007. -- Pereslavl-Zalesskij: Pereslavl University, 2007. -- P. 253--262.

[4] Briukhov D.O., Kalinichenko L.A., Stupnikov S.A. Compositional approach for heterogeneous sources registration at a subject mediator. Emerging Database Research in Eastern Europe: Proc. of the Pre-Conference Workshop of VLDB 2003. -- Cottbus: Brandenburg University of Technology, 2003. -- P. 5-11.

[5] Kalinichenko L.A. Compositional Specification Calculus for Information Systems Development Advances in Databases and Information Systems: Proc. of the Third East European Conference. LNCS 1691. -- Berlin-Heidelberg: Springer-Verlag, 1999. -- P. 317--331.

[6] Kalinichenko L.A., Stupnikov S.A., Martynov D.O. SYNTHESIS: a Language for Canonical Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments. Moscow: IPI RAN, 2007. - 171 p.

[7] Jeffrey D. Ullman, Information Integration Using Logical Views, Proceedings of the 6th International Conference on Database Theory, p.19-40, January 08-10, 1997

## Views expressions construction at information resource registration in typed subject mediator

Ryabukhin O. V., Briukhov D. O., Kalinichenko L.A.

This paper is focused on evolution of information resources registration process in a typed subject mediator. Final phase of this process - the process of views expressions construction is specially emphasized. This phase is considered to be a semi-automatic process with expert participation required. Possible organization of this process, which allows the resource registration system to suggest possible solutions, is described.

---

♣ Работа выполнена при финансовой поддержке РФФИ (проект 08-07-00157-а) и Программы фундаментальных исследований Президиума РАН № 3 "Фундаментальные проблемы системного программирования" (проект "Исследование методов и средств промежуточного слоя предметных посредников, обеспечивающего решение задач над множеством неоднородных распределенных информационных ресурсов").

# Обзор некоторых направлений интеграции гетерогенных ресурсов в электронных библиотеках

© Новицкий А.В.

Институт программных систем НАН Украины
alex@zu.edu.ua

**Аннотация**

Работа посвящена интеграции информации. Сделан обзор текущих проектов по интеграции информации в электронных библиотеках. Проблемы, которые возникают при интеграции данных, связны с тем, что ресурсы описаны метаданными различным способом с различной семантикой. Семантическое аннотирование ресурсов это будущее Интернета и электронных библиотек. С недавнего времени существуют ряд технологий Semantic Web, которые способны автоматически решать проблему интеграции аннотированных веб-документов. В роботе сделано обзор текущих проектов по семантической интеграции в электронных библиотеках, а также показано пример преобразования семантической аннотации RDFa веб-документа в RDF для ЭБ.

## 1 Введение

В информационной интеграции можно выделить следующие проблемы: интеграция схемы [19], хранилищ данных, интеграция данных (также известна как интеграция информации предприятия, EII enterprise information integration) и интеграция каталога.

Подход к интеграции данных с использованием онтологий называется интеграция данных на основе онтологий.

В общем, есть ряд шагов, которые необходимо выполнять при интеграции информационных систем с использованием онтологий. К ним относятся:

- интерпретация запроса в терминологию общей онтологии;
- выявление соответствия между семантически связанными сущностями в локальной и общей онтологии;
- перевод соответствующих данных из локальных информационных источников (участвующих в обработке запроса пользователя) в формализм представления знаний системы интеграции информационных систем;
- согласования результатов, полученных из различных локальных информационных источников, а именно выявление и устранение, например избытка, дублирование и др.

Семантическая гетерогенность [13], как правило, отличается от синтаксической гетерогенности и структурной гетерогенности в семействе баз данных [3, 11, 10, 12].

Синтаксическая неоднородность связана с неоднородностью форматов данных. Стандартизация форматов данных принимается в качестве подхода к решению проблем синтаксической неоднородности. Например, XML используется в качестве стандартного формата для всех видов доступных Web данных.

Структурные неоднородности связаны с различными моделями данных, структур данных или схем, например, реляционных и объектно-ориентированные модели базы данных. Примером решения проблемы структурной неоднородности является использование RDF, который основан на синтаксисе XML и обеспечивает унифицированный способ структуры источников информации.

Следует обратить внимание, несмотря на то, что в электронной библиотеке информация может быть представлена в различных видах, семантика этой информации представляется с помощью текстовых метаданных, соответственно мы будем сосредотачиваться на интеграции семантических метаданных.

Когда два информационных источника смоделированы в одном и том же формате данных с применением одной и той же модели данных, как и раньше, могут возникать проблемы семантической неоднородности [20]:

- семантические конфликты. Различные разработчики моделей не испытывают точно такой же набор объектов реального мира, но вместо этого они представляют наборы которые пересекаются (включение или перекрытие элементов набора). Например, "Студент" объект класса может возникнуть в

**Труды 11[ii] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

одной схеме, в то время как более ограниченный объект класса "Студенты специальности информатика" находится в другой схеме. При интеграции двух схем класс "Студент специальности информатика" будет интегрирован как подкласс класса "Студент".

- описательные конфликты. Описательные конфликты относятся к конфликтам именования вследствие омонимов и синонимов.
- структурные конфликты. Структурные конфликты отличаются от структурной неоднородности. Даже если два разработчика моделей, используют одну и ту же модель данных, они могут выбирать различные конструкторы для представления объектов реального мира. Например, в объектно-ориентированной модели, когда разработчик описывает компонент объекта типа О, он оказывается между выбором, создания нового типа объекта или добавить атрибут к О.

Каждый домен использует локальные онтологии, которые являются результатом концептуализации домена. Поскольку процесс концептуализации не является однозначным, то это порождает гетерогенность источников. Для того чтобы их объединить, необходимо сделать больше, чем простой механизм маркировки соответствия объектов, классов или содержания. На самом деле, часто возникает ситуация, когда понятия не совсем совпадают, поскольку они могут иметь различия по свойствам в видовой или родовой классификации [18]. В целом можно выделить несколько видов сопоставления онтологий:

- расширение: предусматривает определение онтологии домена, связывая некоторые понятия между двумя выходными онтологиями. Две концептуальные модели дополняют друг друга, например концепты первой онтологии уточняются во второй через дополнительные атрибуты, которые не указаны в первой.
- гармонизация: предполагает семантическую эквивалентность между доменом и прикладными онтологиями, касается одного и того же онтологического обязательства. В этом случае текущий домен можно рассматривать как специализация в другом домене, который является более общим или расположен на абстрактно-формальном уровне.
- выравнивание: предполагает обобщение онтологии домена через общие понятия и аксиомы. Обе модели имеют (много/несколько) общих совместных концептов.

Для электронных библиотек решения проблемы семантической гетерогенности можно решать на двух уровнях: на уровне метаданных и на уровне контента. Для того чтобы показать текущее состояние дел, сделаем краткий обзор проектов по интеграции электронных библиотек. Следует обратить внимание, что рассмотрены различные аспекты проблематики интеграции данных. Такими аспектами могут быть особенности поискового механизма или архитектурные особенности. Тем не менее, данный обзор предоставляет, в некотором смысле, общий взгляд на текущее состояние проблемы.

Цель обзора - прояснить особенности интеграции информации в семантической среде. Следует обратить внимание, что если получиться представление контента (метаданных) в модели RDF, то проблему интеграции для контента (метаданных) можно считать решенной, или решение проблемы будет преведено к задаче мапинга онтологий.

В разделе 2 будет попытка сделать обзор проектов семантической интерпарабельности с позиций принципов интеграции и поисковых механизмов, затронуты и сопутствующие вопросы.

В разделе 3 приводится пример, который показывает, насколько простым может быть решение интеграции контента (метаданных) если изначально наш контент имеет семантическую аннотацию.

## 2 Краткий обзор семантической интерпарабельности в Европейских проектах.

### 2.1 Проект SWHi

Рассмотрим европейские проекты по внедрению Semantic Web в электронные библиотеки. Онтология SWHi [6] разработана для электронной библиотеки с точки зрения, когда наши основные источники данных в репозитории описаны метаданными. Эти метаданные отображаются и хранятся в онтологии, которая базируется на онтологии схемы. Для обогащения онтологии, также добавляют новую, связанную информацию из выбранных веб-документов.

Поиск в этой системе реализован в двух формах, простой и сложной. Система использует SeRQL как язык запросов к RDF [2]. Процессор генерирования запросов SeRQL сталкивается по крайней мере с двумя проблемами. Во-первых, он не знает, в каком классе или свойстве могут быть найдены слова. Чтобы избежать этой проблемы, прикладное программное обеспечение Semantic Web, такое как OpenAcademia [16], требует от пользователей вводить ключевые слова в соответствующее поле (автор, название или год) в ее расширенный поисковый интерфейс. Во-вторых, существуют некоторые ограничения в подстроках соответствия SeRQL при использовании символа общности '*'. Эту проблему можно решить с помощью информационно-поисковых программ, таких как Lucene. Другие семантические системы, такие как KIM, используют машину поиска Lucene для индексации и поиска семантически аннотированных документов. Одновременно OWLIR и Swoogle используют индексно-поисковую Haircut для

индексирования и поиска RDF документов, используя механизм N-грамм в качестве терминов индексации.

Помимо самого поиска, важным также является вопрос представления результатов поиска. Одним из направлений является визуализация поиска. В Semantic Web визуализация становится все более важной. Существуют случаи сложных взаимоотношений между ресурсами, которые не могут быть представлены с помощью простого списка. Кроме того, как правило, отражается только небольшое количество результатов поиска. К документам находящимся в хвосте результата поиска, скорее всего, никогда не будут обращаться.

Общая архитектура системы SWHi состоит из трех уровней: система управления знаниями (Kms), семантических веб-приложений (SWA), и уровень пользовательских интерфейсов (Web UI).

Очевидно, онтологии играют центральную роль в SWHi. Для развития SWHi онтологии, повторно используют имеющиеся онтологические ресурсы для структурирования и сохранение исторической информации, а именно: PROTON базовая онтология, таксономия предметной классификации NewsBank/Readex, Дублинское Ядро и словарь FOAF Vocabulary. Эти онтологии сохраняется с использованием Sesame2.

Экземпляры для этой онтологии брались с данных ранней Американской истории 1639-1800. Метаданные состояли из 36305 записей, которые подробно описаны (название, автор, дата публикации и т.д.) в MARC21. В будущем планируется расширить источники данных из других исторических электронных журналов, включив их полные тексты.

### 2.2 Проект eCulture

eCulture это семантическая поисковая система, которая позволяет одновременно искать в нескольких коллекциях учреждений культурного наследия [15]. Работает путем переноса коллекций в RDF, связывая объекты коллекций как экземпляры классов через общедоступные словари, тем самым создавая большой RDF граф.

Основным механизмом поиска является использование Prolog [21]. Запросы прикладной логики выражаются как Prolog цели на необработанных данных RDF и/или модулях суждения RDFS/OWL.

В eCulture разработана методология портирования культурных хранилищ для Semantic Web и RDF. Эта методология основана на том , что мы можем рассчитывать, как правило, на два типа данных:

-метаданные, которые описывают культурные объекты и фотографии;

-локальные словари, которые используются в некоторых метаданных.

Основное внимание в проекте уделено конвертации XML в RDF. Обращая внимание на то, что в отличие от ранее предложенных подходов, где целевая модель RDF следует из источников данных XML, в eCulture такой подход неприемлем. Трансформация XML в RDF основана на правилах свойств, правилах очередности, замены значений и других.

### 2.3 Проект IPISAR (Image Preservation, Information Systems, Access and Research)

Проект IPISAR исследует распространение, изучение и рациональное использование культурного наследия, а также попытки представить решения общих проблем в этих областях в рамках Semantic Web (SW) [14]. В рамках проекта предложено ряд идей, которые возможно, упростят интеграцию информации.

В проекте разработано приложение «Pescador», которое будет хранить каталогизированные данные в устойчивых тройных хранилищах (чьи функции будут такими же, что и реляционные базы данных в традиционных системах).

Pescador использует модель SW для каталогизации, где каждому формату записи будет соответствовать отдельный вид структурированного графа, в который включена специальная лексика и правила, с ссылкой на специализированную прикладную логику.

Одним из направлений Pescador является обеспечение интерфейсов программирования для пользователей, которые выполняют дизайн, моделирование и программирование каталогов. Для достижения этой цели была предложена семантическая компонентная архитектура (SCA). То есть, адаптация компонентов архитектуры в соответствии с принципами SW, в которых данные, структура и правила прикладной логики тесно связаны между собой. SCA должен координировать подключаемые "компоненты", которые "обернуты" оболочкой и могли бы взаимодействовать со следующими типами: схемами; ограничениями; правилами вывода; онтологиями; определениями путей; программным кодом; спецификациями вывода; информацией о конфигурациях Abox; ссылками к внешним источникам данных.

SCA должна включать средства определения модели пути. Предварительный обзор существующих механизмов определения путей показывает, что SPARQLeR [9] расширение SPARQL [4] может быть лучшим кандидатом для адаптации SCA. SPARQLeR предназначена для запросов по семантический ассоциации. Запрос в SPARQLeR сосредоточен на построении шаблонных путей, включая в себя неориентированные и направленные, пути направления которых, задаются определенными свойствами.

### 2.4 Проект EPOCH и AMA для библиотеки культурного наследия

EPOCH представляет собой сеть из более ста европейских культурных институтов, объединение

их привело к повышению качества и эффективности использования информационно-коммуникационных технологий для культурного наследия.

При интеграции ЕБ возникает ряд проблем.

Первая проблема заключается в том, что каждый справочник имеет свою поисковую систему и использует свою грамматику метаданных для описания и индексации данных, в частности, она никогда не будет работать на других системах. Ни одна из этих систем метаданных не может проанализировать всю информацию на веб-сайте, если мы не будем делать их доступными через машинно-читаемую форму с использованием RDF [17].

Вторая проблема касается непосредственно информации: огромное количество различных форматов, используемых для индексирования данных, является большим препятствием на пути к интеграции, и должны быть серьезно проанализированы. Даже если мы ограничиваем наши усилия исключительно для архивов культурного наследия, (например, базы данных музеев и коллекций, археологические раскопки, отчеты, доклады и другие неструктурированные данные), мы вынуждены признать, что информация, также является гетерогенной.

Чтобы создать единый концептуальный слой, семантическая информация должна быть взята из базы данных, HTML-страниц, описательных текстов, метаданных и представлена в стандартном формате, с целью получения концептуального содержания информации, создав концептуальный мапинг[1]. Как только концептуальный слой для данных и метаданных готов семантическая информация будет храниться в контейнере, основанном на RDF и онтологии.

Для упрощения и доступности процесса отображения в проекте AMA было разработано программное обеспечение AMA Mapping Tool. Этот инструмент способствует сопоставлению различных моделей данных с разной структурой, в том числе и робота с неструктурированными данными (текстовое описание). Этот мапинг основывается на известной онтологии CIDOC CRM.

Для неструктурированных документов используется ПО AMA TextTool.

Большая часть информации для наполнения CIDOC CRM-онтологии получается из текстов вручную. Для этого в рамках проекта разработано ПО AMA TextTool, которое предназначено для полуавтоматического кодирования археологических текстов в CIDOC-CRM. Даное ПО работает на понятиях и методиках компьютерной лингвистики. AMA TextTool реализует KWIC (ключевое слово в контексте). ПО используется для поиска слова, фразы или шаблонов слов и, возможно, XML-разметки в тексте. Пользователи могут затем проанализировать текст и знаки, которые представлены в следствие KWIC и отметить согласованные элементы. Система затем вставляет соответствующие отметки в тексты файла (ов).

### 2.5 Мапинг данных культурного наследия в CIDOC-CRM.

На данный момент несколько описательных стандартов уже отражено в CIDOC-CRM. Например, MIDAS стандарт данных Великобритании для информации об исторической среде, разработанной в интересах Форума информационных стандартов в области охраны наследия (FISH).

В введении к справочнику CIDOC CRM его авторы отмечают, что "поскольку прогнозируемая сфера применения CRM является подмножеством реального мира, и поэтому потенциально бесконечна", то модель была разработана для расширения посредством связей с совместимыми внешними типами иерархий.

В этом смысле "совместимость расширения с CRM означает, что данные, структурированные в соответствии с расширением, должны также оставаться правильными, как экземпляры класса CRM ".

В документации к CRM-CIDOC описано целый ряд процедур, которые можно использовать для расширения, придерживаясь выше поставленных требований:

- существующие классы высокого уровня могут быть расширены через подкласс или динамически с использованием типа иерархии.
- существующие свойства высокого уровня могут быть расширены с помощью структурированных подсвойств, а в некоторых случаях, динамично, с использованием атрибутов свойств, позволяющих подтипы.
- дополнительная информация, которая выходит за рамки семантики формально определенной в CRM, и может быть записана как неструктурированные данные, используя E1.CRM_Entity. P3.has_note: E62.String.

Начальные и целевые структуры, возможно, не всегда совпадают: в этом случае новая модель (источник), которая согласовывается с CIDOC CRM-иерархией будет создана путем явной декларации некоторых понятий, которые не явные в источнике, в соответствии с аксиомами/путями, описывающих структуру и иерархию CIDOC-CRM. Это своего рода процесс анисоморфизм[2] (anisomorphic), который изменяет очевидную структуру первоначального источника.

[1] метод для представления знаний в виде графов

[2] различия в семантический сфере применения терминов, относящихся к реальной жизни: например, английский и русский языки являются анисоморфизм в том, что касается терминологии цвета. Английский определяет светло-голубой (light blue) и темно-синего цвета (navy blue), как оттенки одного цвета, но русский трактует как не связные оттенки различных цветов

Это показывает, что отображение не является простым вопросом, или линейным процессом и требует дисциплинарной компетенции, а также глубокое понимание неявных предположений о модели исходного источника.

В рамках выполнения проекта было осуществлено гармонизацию с CIDOC-CRM доменных онтологий AMICO, DC, EAD, FRBR, TEI. Также было осуществлено расширение и специализация CIDOC-CRM с дополнительными прикладными онтологиями X3D и MPEG7, а также осуществлено отображение к CIDOC-CRM онтологии задач MIDAS, English Heritage, ICCD, PERSEUS. Для осуществления отображения использовалось ПО AMA (Archaeological Mapping Tool) [7].

## 3. Семантическая аннотация

Для научных исследований предложенные решения для интеграции библиотек, несомненно, важны и интеграция результатов научных экспериментов со знаниями, которые представлены в электронных библиотеках есть перспективным направлением. Опорой в этом направлении мы считаем технологию Semantic Web, а именно аннотация контента.

Как следует с обзора, обязательным этапом интеграции информации в электронной библиотеке есть перенос коллекции в RDF. В случаи структурированных кратких описательных метаданных (например, в рамках ДЯ) этот процесс возможно автоматизировать. Но для возможности автоматически анализировать содержания документа, таких аннотаций явно недостаточно. Поэтому в последнее время большое внимание уделяется более подробному раскрытию смысла контента через аннотации. Другими словами для анализа научных данных интеграция схем метаданных является не достаточной.

При анализе научных данных необходимо подавлять разного типа гетерогенность. Интеграция позволит объединить основные сведения из различных электронных архивов и других научных источников, которые могут быть пересмотрены и в которых возможно осуществить поиск, как в единой целой электронной библиотеке.

Хотя с момента появления Semantic Web прошло уже около 10 лет, тем не менее, должной популярности в широких массах эта технология не набрала. Виной тому множество рекомендаций и стандартов, которые существуют в данном направлении. С одной стороны нет удобных приложений, которые работали бы с RDF, с другой стороны отсутствуют RDF данные и онтологии.

Следующим этапом мы считаем более глубокое проникновение семантических технологий в электронные библиотеки, тому есть несколько причин. Во-первых, сейчас уже создано достаточное количество онтологий в различных предметных областях, например, Basic Formal Ontology [http://www.ifomis.org/bfo/], CIDOC Conceptual Reference Model [http://cidoc.ics.forth.gr/], Open Biomedical Ontologies [http://www.obofoundry.org/] и т.д. Во-вторых, разработано ряд приложений, которые способствуют внедрению Semantic Web на практике. Важным этапом на пути интеграции информации в Semantic Web есть принятия в качестве рекомендации языка запросов SPARQL (W3C Recommendation, January 15, 2008) и рекомендации по повторному использованию RDF-данных в XHTML - RDFa (W3C Recommendation, October 18, 2008).

В электронных библиотеках очень хорошая среда для наполнения Semantic Web посредством семантического аннотирования. Электронная библиотека хороша тем, что данные в ней структурированы с помощью метаданных. Однако метаданные хоть и представляются в машиночитаемом формате, но не дают полного представления об контенте информационного ресурса которые они описывают.

Семантическая разметка или аннотирование представляет собой явное описание семантики контента ресурса при помощи понятий семантической модели (онтологии или словаря). Такое явное описание семантики выполняется указанием четкого соответствия между определенной частью контента ресурса и его семантикой, описанной в семантической модели. Аннотирование при этом базируется на RDF.

Сегодняшние Web-ресурсы разрабатываются по большей части для использования людьми. Несмотря на постепенное появление в сети данных, предназначенных для машинного восприятия, эти данные в основном распространяются отдельным файлом в определенном формате. Притом соответствие машинной версии человеческому представлению весьма ограничено. Как следствие, Web-браузеры могут обеспечить пользователей лишь минимальной поддержкой в анализе и обработке сетевых данных, ведь браузеры только представляют информацию. Технология RDFa [18] позволяет сопроводить графические данные машиночитаемыми подсказками с помощью набора XHTML-атрибутов. RDFa — это способ выражения RDF-данных в XHTML, в рамках которого данные, предназначенные для человека, используются повторно.

Примером использования RDFa может служить закладывания фрагмента кода, который описывает название и автора статьи, которая расположена в электронной библиотеке. При описании используется схема метаданных Дублинского Ядра (xmlns:dc=http://purl.org/dc/elements/1.1/).

```
<?xml version="1.0" encoding="UTF-8"?>
<html xmlns="http://www.w3.org/1999/xhtml"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<head profile="http://www.w3.org/2003/g/data-view">
<title> Доповідь про http://oai.org.ua</title>
</head>
<body>
  <h1>Ресурс http://oai.org.ua</h1>
```

```
        <dl about="http://eprints.zu.edu.ua/2648/">
        <dt>Назва доповіді</dt>
        <dd property="dc:title">Інтеграція наукових
електронних бібліотек України: всеукраїнський портал
збору та пошуку метаданих http://oai.org.ua</dd>
        <dt>Автор</dt>
<dd property="dc:creator">Новицький, О.В.</dd>
        </dl>
 </body>
</html>
```

**Рис. 1 Фрагмент XHTML кода ЭБ с разметкой RDFa.**

Сам по себе механизм RDFa был бы малоинтересен, хоть определяет семантику контента. Необходимым условием является возможность извлечение семантической аннотации со страниц. Такой механизм к счастью разработан и носит названия GRDDL - **G**leaning **R**esource **D**escriptions from **D**ialects of **L**anguages (http://www.w3.org/TR/grddl/).

При помощи GRDDL возможно однообразно извлекать микроформатированный контент. Спецификация GRDDL определяет разметку на основе существующих стандартов для объявления о том, что XML документ включает в себя данные совместимые с RDF, а также ссылку на алгоритм (как правило, представленный в XSLT), для извлечения данных из документа.

Разметки содержат определения пространства имен общего назначения для XML-документов, а также ссылку на профиль отношений для использования в валидных XHTML документов.

Ниже представлен фрагмент XHTML кода в соответствии с GRDDL.

```
<?xml version="1.0" encoding="UTF-8"?>
<html xmlns="http://www.w3.org/1999/xhtml"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<head profile="http://www.w3.org/2003/g/data-view">
<link rel="transformation" href="RDFaRDF.xsl"/>
<title> Доповідь про http://oai.org.ua</title>
</head>
<body>
  <h1>Ресурс http://oai.org.ua</h1>
        <dl about="http://eprints.zu.edu.ua/2648/">
        <dt>Назва доповіді</dt>
        <dd property="dc:title">Інтеграція наукових
електронних бібліотек України: всеукраїнський портал
збору та пошуку метаданих http://oai.org.ua</dd>
        <dt>Автор</dt>
        <dd property="dc:creator">Новицький, О.В.</dd>
        </dl>
 </body>
</html>
```

**Рис. 2 Фрагмент XHTML кода с разметкой RDFa и GRDDL**

При обработке преобразования данного фрагмента средствами XSLT будет получена модель данных и представлена в RDF с помощью XML Рис. 3.

```
<rdf:RDF xmlns:h="http://www.w3.org/1999/xhtml"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
        <rdf:Description rdf:about="">
<transformation xmlns="http://www.w3.org/1999/xhtml"
rdf:resource="RDFa2RDFXML.xsl"/>
        </rdf:Description>
        <rdf:Description
rdf:about="http://eprints.zu.edu.ua/2648/">
        <dc:title
xmlns:dc="http://purl.org/dc/elements/1.1/">Інтеграція
наукових електронних бібліотек України: всеукраїнський
портал збору та пошуку метаданих http://oai.org.ua</dc:title>
        </rdf:Description>
        <rdf:Description
rdf:about="http://eprints.zu.edu.ua/2648/">
        <dc:creator
xmlns:dc="http://purl.org/dc/elements/1.1/">Новицький,
О.В.</dc:creator>
        </rdf:Description>
</rdf:RDF>
```

**Рис. 3 RDF представлен с помощью XML**

Стоит обратит внимание, что GRDDL имеет возможность преобразования разметки RDFa для которой, например, используется схема данных Дублинского Ядра, непосредственно в другие схемы метаданных, таких как CIDOC-CRM.

Такой подход применим только к веб-документам, но возможно получится применение данной технологии к мультимедиа форматам.

Еще одним применением данного подхода может быть процесс который предложен в [8].

Пример автоматического внесения документов (с возможностью распределенности) и построения индексов. Идея заключается в GRDDL обработке источников документов и извлечения встроенного RDFa для подключения в хранилища RDF. Далее SPARQL запросы выбирали бы с этого хранилища соотвествующие результаты, которые были бы представлены в виде автоматически генерируемой веб-страницы Рис. 4.

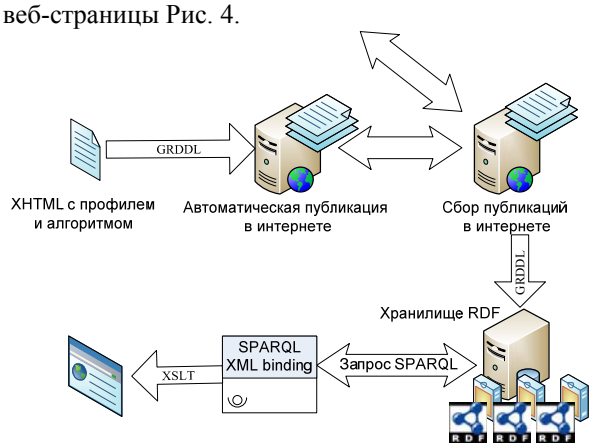



**Рис. 4 Пример RDFa и GRDDL**

## Литература


[1] *Mapping, Embedding and Extending: Pathways to Semantic Interoperability The Case of Numismatic Collections.* . **Andrea D' Andrea, Franco Niccolucci.** Tenerife : б.н., 2008. Semantic Interoperability in the European Digital Library, Proceedings of the First International Workshop. стр. 63-75.

[2] *SeRQL: A Second Generation RDF Query Language.* **Broekstra J, Kampman A.** Proc SWAD-Europe Workshop on Semantic Web Storage and Retrieval 2003.

[3] *Answering XML queries over heterogeneous data sources.* **Daniela Florescu, Ioana Manolescu, and Donald Kossmann.** б.м. : Morgan Kaufmann. 27th International Conference on Very Large Data Bases (VLDB 2001). стр. 241-250.

[4] **Eric Prud'hommeaux Andy Seaborne.** SPARQL Query Language for RDF. *W3C.* [В Интернете] 2008 г. http://www.w3.org/TR/rdf-sparql-query.

[5] **Ismail Fahmi, Junte Zhang, Henk Ellermann, Gosse Bouma.** SWHi System Description: A Case Study in Information Retrieval, Inference, and Visualization in the Semantic Web. *The Semantic Web: Research and Applications, 4th European Semantic Web Conference.* Innsbruck, Austria : Springer, 2007, стр. 769-778.

[6] **Felicetti A. M. Ioannides, D.Arnold, F. Niccolucci, K. Mania (eds.).** *MAD – Management of Archaeological Data.* Budapest : б.н., 2006,. стр. 124 – 131, The e-volution of Information Communication Technology in Cultural Heritage – Project papers.

[7] *Semantic Maps and Digital Islands: Semantic Web technologies for the future of Cultural Heritage Digital Libraries.* **A. Felicetti, H. Mara.** Tenerife, Spain : б.н., 2008. SIEDL 2008: Semantic Interoperability in the European Digital Library. стр. 51-62.

[8] **Gandon, Fabien.** Digital library example. *Institut National de Recherche en Informatique et en Automatique / Centre de recherche Sophia Antipolis - Mŭditerranŭe.* [В Интернете] 2009 г. http://www-sop.inria.fr/acacia/personnel/Fabien.Gandon/tmp/grddl/rdfaprimer/PrimerRDFaSection.html.

[9] *SPARQLeR: Extended Sparql for Semantic Association Discovery.* **Krys Kochut, Maciej Janik.** 2007. 4th European Semantic Web Conference (ESWC2007). http://www.eswc2007.org/pdf/eswc07-kochut.pdf.

[10] *Data integration: a theoretical perspective.* **Lenzerini, Maurizio.** New York : ACM Press, 2002. 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2002).

[11] *Locating data sources in large distributed systems.* **Leonidas Galanis, Yuan Wang, Shawn R. Jeffery, and David J. DeWitt.** б.м. : Morgan Kaufmann. 29th International Conference on Very Large Data Bases (VLDB 2003). стр. 874–885.

[12] *Combining artificial intelligence and database for data integration.* **Levy., Alon Y.** б.м. : Berlin/Heidelberg, LNCS 1600, Springer. In Artificial Intelligence Today: Recent Trends and Developments. стр. 249–268.

[13] **Lin, Yun.** *Semantic Annotation for Process Models: Facilitating Process Knowledge Management via Semantic Interoperability.* б.м. : Department of Computer and Information Science Norwegian University of Science and Technology. 7491.

[14] *Modular, Best-Practice Solutions for a Semantic Web-Based Digital Library Application* . **Martinez, Andrew Russell Green and Jose Antonio Villarreal.** Tenerife, Spain : б.н., 2008. Proceedings of the Workshop on Ontologies: Reasoning and Modularity (WORM-08).

[15] *Porting Cultural Repositories to the Semantic Web.* **Omelayenko, B.** Tenerife, Spain : б.н., 2008. Proceedings of the First Workshop on Semantic Interoperability in the European Digital Library (SIEDL-2008). стр. 14-25.

[16] **OpenAcademia.** [В Интернете] www.openacademia.org.

[17] **RDF Core Working Group.** Resource Description Framework (RDF). *Resource Description Framework* . [В Интернете] W3C. http://www.w3.org/RDF/.

[18] RDFa Primer. Bridging the Human and Data Webs. **W3C Working Group.** [В Интернете] W3C. http://www.w3.org/TR/xhtml-rdfa-primer/

[19] **Shvaiko, Pavel.** *Iterative schema-based semantic matching.* Informatica e Telecomunicazioni. Trento : University of Trento, 2006. Technical Report DIT-06-102.

[20] *Model independent assertions for integration of heterogeneous schemas.* **Stefano Spaccapietra, Christine Parent, and Yann Dupont.** 1, 1992 г., VLDB Journal, Т. 1, стр. 81–126.

[21] **Wielemaker, J., Hildebrand, M., Ossenbruggen, J.R. Van.** Using Prolog as the fundament for applications on the semantic web (2008). *Proceedings of the 2nd Workshop on Applicatiions of Logic Programming and to the web, Semantic Web and Semantic Web Services.* Porto, Portugal : б.н., 2007.

## A review of some of the integration of heterogeneous resources in digital libraries

O. Novytskyi

This paper is devoted to the integration of information. An overview of current projects for the integration of information in digital libraries. Problems that arise when integrating data, coupled with the fact that the resources described in the metadata different way with different semantics. Semantic annotation of resources is the future of the Internet and digital libraries. Recently, there are a number of technologies Semantic Web, which can automatically solve the problem of integration of annotated web documents. In the robot made a review of current projects on semantic integration in digital libraries, and shows an example of the transformation of semantic web annotations RDFa document to the RDF for the DL.

# ОНТОЛОГИЧЕСКОЕ МОДЕЛИРОВАНИЕ–2

# ONTOLOGICAL MODELING–2

# Реляционная база данных как структурированное хранилище многоязычного глоссария терминов по аналитической химии. Разработка лингвистической онтологии ♣

© Колотов В.П., Широкова В.И., Аленина М.В.

Институт геохимии и аналитической химии им. В.И.Вернадского РАН,
119991, ГСП-1, Москва, ул. Косыгина, 19
shirokova@geokhi.ru

**Аннотация***

В качестве первого шага к разработке онтологии по аналитической химии создана база данных ключевых понятий в виде двуязычного русско-английского глоссария (ее расширение для других языков подразумевается). Разработана структура базы данных, проведено ее наполнение данными из различных официальных документов и выполнено ранжирование данных. Анализ результатов ранжирования (иерархии терминов) позволяет выявить противоречия и неточности терминологии и дать рекомендации по их устранению (например, необходимость уточнения дефиниций терминов). Имеется в виду, что дефиниции терминов, описывающих понятия более высокого порядка, должны включать базовые термины из глоссария, а не являться произвольным текстом, даже близким по смыслу. Каждому термину присвоен определенный статус, который обеспечивает возможность поиска понятий и терминов в случае нечетко выраженного запроса. База данных будет опубликована в Интернете на MS Windows SharePoint-сайте для ознакомления профессиональным сообществом химиков-аналитиков и в образовательных целях.

## 1 Введение

Запущен проект по созданию онтологии по аналитической химии. Работа над развернутым проектом мотивирована следующими моментами: создание предметной онтологии по аналитической химии в связи с активно обсуждаемой концепцией семантического Интернета, подразумевающему размещение информационных ресурсов, содержащих структурированную и формализованную информацию, «понятную» компьютерам [1]; необходимость гармонизации терминологии по аналитической химии, включая электронное обеспечение научного и образовательного сообщества. Концепция семантического Интернета предполагает наличие в сети предметных онтологий (по сути словарей дефиниций, понятий и терминов той или иной области знания) и увязывание размещаемых в Интернете материалов с этими онтологиями с помощью содержательных дескрипторов XML. Онтологии, в том числе и по аналитической химии, должны создаваться сообществом специалистов.

В качестве первого шага разработана база данных ключевых понятий этой отрасли науки в виде многоязычного глоссария. Имея в виду, что методы аналитической химии часто используется для сертификации различных материалов, продуктов питания, контроля состояния окружающей среды и т.д., то описание аналитических процедур, представление результатов анализа, а значит и соответствующая терминология строго регламентируются уполномоченными органами, как национальными (ГОСТ), так и международными (ISO, IUPAC и др.). Поэтому терминологическое обеспечение аналитической химии заметно более продвинутое по сравнению с другими научными дисциплинами, где часто используется лингвистический анализ реальных текстов для выбора терминологии с последующей экспертизой для снижения информационного шума и формализации данных [3]. В аналитической химии в качестве первичных источников терминов, прошедших высококачественную профессиональную экспертизу служат утвержденные соответствующими официальными органами документы, в том числе и глоссарии терминов. Как правило, такие глоссарии

**Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2009, Петрозаводск, Россия, 2009.**

достаточно полны и содержат как термины, так и их синонимы (в ряде случаев антонимы), комментарии и, обязательно, дефиниции терминов.

Следует отметить, что несмотря на значительные усилия по разработке непротиворечивой системы терминологии существуют разночтения в трактовке терминов, разработанных различными организациями. Отчасти это объясняется и тем, что многие методы аналитической химии возникли на стыке с другими науками, что привело к определенному взаимовлиянию терминов, появлению терминов-синонимов, трактуемых по разному в различных областях аналитической химии. Это относится как к англоязычной, так и русскоязычной терминологии. В этой связи сведение всей официальной терминологии по аналитической химии в единый структурированный электронный источник (что является одной из целей проекта) обеспечивает удобное и мощное средство для ее анализа и, в том числе, для гармонизации терминологии в целом. Естественно, для обеспечения информационно-поисковых задач по аналитической химии глоссарий должен быть многоязычным (на текущем этапе русско-английским).

## 2 Источники информации

Нами использованы следующие официальные источники терминологии:


1. Compendium of Analytical Nomenclature: Definitive rules 1997. Orange Book. 3rd edition Inczedy, J.; Lengyel, T. and Ure, A.M. Blackwell Science, 1998 [ISBN 0-86542-6155]. On-line version: http://www.iupac.org/publications/analytical_compendium (язык: английский),
2. International Vocabulary of Metrology – Basic and General Concepts and Associated Terms. VIM. 3rd edition. Final 2007-05-18 (язык: английский, французский),
3. ГОСТ Р 8.563-96 «Государственная система обеспечения единства измерений. Методики выполнения измерений» - М.: Издательство стандартов (язык: русский),
4. Национальный стандарт Российской Федерации. ГОСТ Р 52361-2005 «Контроль объекта аналитический. Термины и определения». М.: Стандартинформ. 2005 (язык: русский),
5. Государственный стандарт Российской Федерации. ГОСТ Р ИСО 5725.1-5725.6 «Точность (правильность и прецизионность) методов и результатов измерений». Часть 1. Основные положения и определения. М.: Издательство стандартов (язык: русский),
6. Представление результатов химического анализа. Рекомендации ИЮПАК 1994 (IUPAC. 1994. V. 66. P. 595). Перевод с англ. // Журнал аналитической химии. 1998. Т. 53. N 9. С. 999-1008 (язык: русский),
7. РМГ 61-2003 «Показатели точности, правильности, прецизионности методик количественного химического анализа» - М., ИПК Издательство стандартов. 2004 (язык: русский).


## 3 Реляционная база данных

В результате предварительного тестирования различных способов электронной интеграции разнородных источников данных, выбор пал на использование модели реляционной базы данных как удобного средства для хранения данных, поддержания их целостности, представления отношений терминов различного ранга, обеспечения прослеживаемости истории записей, экспорта в XML-формат, публикации в Интернете и т.д. В качестве физической СУБД использован сервер MS SQL 2005.

Разработана структура базы данных, проведено ее наполнение и выполнено ранжирование.

На Рис.1 приведены основные таблицы развернутой базы данных и отношения между ними.

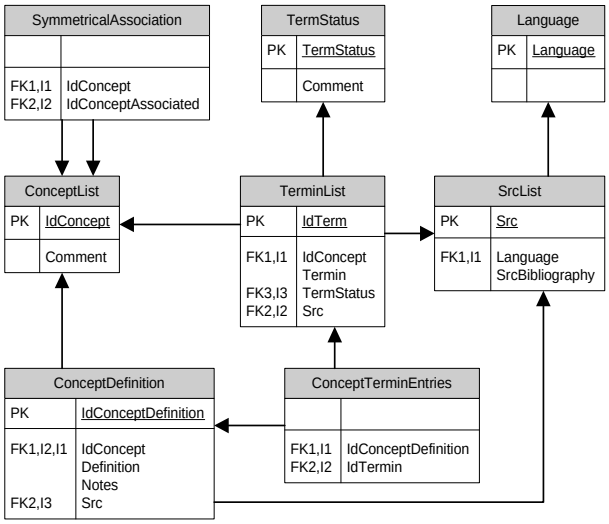



Рис.1. Структура реляционной базы данных глоссария: таблицы и поля, PK и FK(n) первичные и внешние ключи для обеспечения целостности данных, их каскадного изменения/удаления, подстановки, I(n)- индексированные поля. Стрелками показаны реляционные отношения (обычно один - ко многим).

Центральной таблицей является ConceptList (список понятий). Абстрактные понятия попросту представлены целым числом, а поле комментария позволяет дать наиболее частотное (или удобное) их словесное обозначение.

Термины и их лингвистические формы (на любом языке), соответствующие данному понятию, записаны в таблицу TerminList (фрагмент таблицы приведен на Рис. 2). Одно и то же число в поле IdConcept означает семантическую идентичность терминов.

Рис. 2. Фрагмент таблицы TerminList базы данных

| IdTerm | IdConcept | Termin | Src |
|---|---|---|---|
| 107 | 66 | systematic measurement error | JCGM/WG 2 |
| 109 | 66 | systematic error of measurement | JCGM/WG 2 |
| 110 | 66 | systematic error | JCGM/WG 2 |
| 601 | 66 | полная систематическая погрешность | IUPAC_1994 |
| 642 | 66 | bias | ГОСТ Р ИСО 5725-1-2002 |
| 432 | 66 | систематическая погрешность | ГОСТ Р ИСО 5725-1-2002 |
| 672 | 66 | bias | IUPAC_1994 |
| 157 | 66 | систематическая  погрешность | ГОСТ Р 8.563 - 96 ГСИ |

Рис. 3. Фрагмент таблицы ConceptDefinition базы данных.

| IdConcept | Definition | Src |
|---|---|---|
| 66 | Component of measurement error that in replicate measurements remains constant or varies in a predictable manner | JCGM/WG 2 |
| 66 | Разность между математическим ожиданием и истинным значением (со знаком) | IUPAC_1994 |
| 66 | Разность между математическим ожиданием результатов измерений и истинным ( или в его отсутствие - принятым опорным ) значением | ГОСТ Р 8.563 - 96 ГСИ |
| 66 | Разность между математическим ожиданием результатов измерений и истинным (или в его отсутствие - принятым опорным) значением | ГОСТ Р ИСО 5725-1-2002 |

ConceptDefinition (фрагмент таблицы приведен на Рис. 3) содержит определения (дефиниции) того или иного понятия (на языке источника). Количество дефиниций определяется количеством источников данных. Кроме того, имеется ряд вспомогательных таблиц, обеспечивающих подстановку данных, например, источник терминологии и статус термина. О последней таблице есть смысл сказать отдельно. Выделены следующие особенности терминов: основной термин (предпочтительный для данного понятия); его полноценный синоним; синоним употребляемый, но устаревающий и/или не совсем точный (рекомендуется избегать использования); синоним не вошедший в нормативные документы, но, тем не менее, используемый на практике; синоним ошибочный или устаревший (должно избегать использования). Такая иерархия терминов обеспечивает возможность поиска даже нечетко выраженной информации, включая грубую оценку ее соответствия запросу.

Для того, чтобы не перегружать смысловую часть схемы (Рис.1) из нее убрана информация, касающаяся истории редактирования данных (записи при редактировании не удаляются, имеются средства для восстановления истории модификации базы данных).

Для ранжирования данных разработано программное обеспечение (TerminRange), написанное на языке C#, включающее ряд последовательных запросов к базе данных. Результатом работы программного обеспечения являются таблица отношений терминов (ConceptListEntries), построенная на основе анализа вхождений терминов в дефиниции для определенного источника терминов. Анализ этих таблиц позволяет выявить иерархию терминов.

В базе данных в настоящее время представлены следующие отношения:

- наследуемые материнско-дочерние или родовидовые отношения. Эта наиболее многочисленная группа отношений строится автоматически с помощью программы TerminRange (табл. ConceptTerminEntries),

- симметричная ассоциация (очень близкие, но разные понятия). Эта узкая группа отношений задается вручную экспертом при компиляции базы данных, когда приходится учитывать тонкие нюансы различающие понятия (табл. SymmetricalAssociations).

Анализ иерархии терминов позволяет выявить некоторые невязки и неточности терминологии и дать рекомендации по их устранению (например, необходимость уточнения дефиниций терминов). Здесь имеется в виду, что дефиниции терминов, описывающих понятия более высокого порядка, должны включать базовые термины из глоссария, а не являться произвольным текстом, даже близким по смыслу.

База данных, включая отношения терминов, будет опубликована на MS Windows SharePoint-сайте [2] для публичного обсуждения профессиональным сообществом химиков-аналитиков.

## Литература


[1] Berners Lee T. Semantic Web Road Map: http://www.w3.org/DesignIssues/Semantic.html

[2] WSS- сайты Научного совета по аналитической химии Российской академии наук: http://www.rusanalytchem.org/wss

[3] Б.В.Добров, Н.В.Лукашевич, Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска // Web Journal of Formal,Computational & Cognitive Linguistics - Special issue, 2006. http://fccl.ksu.ru/issue_spec/docs/oent-kgu.doc


## Relational database as the structured storage of a multilingual glossary of terms in analytical chemistry. Working out linguistic ontology


V.P.Kolotov, V.I.Shirokova and M.V.Alenina



As the first step to working out ontology in analytical chemistry the database of key concepts in the form of a bilingual Russian-English glossary (its expansion for other languages is meant) is created. The database structure is developed, its filling by the data from various official documents is spent and ranging of the data is executed. The analysis of results of ranging (hierarchy of terms) allows to reveal contradictions and discrepancies of terminology and to make recommendations about their elimination (for example, necessity of specification of definitions of terms). Means that definitions of the terms describing concepts of higher order, should include base terms from a glossary, instead of be any text, even faithful. To each term the certain status which provides possibility of search of concepts and terms in case of indistinctly expressed inquiry is appropriated. The database will be published on the Internet on MS Windows a SharePoint-site for acquaintance by professional community of chemists-analysts and in the educational purposes.



Работа проводится при поддержке Российского фонда фундаментальных исследований (грант N 08-03-00893)

# Related Terms Search Based on WordNet / Wiktionary and its Application in Ontology Matching ♣

© Andrew Krizhanovsky

Institution of the Russian Academy of Sciences St.Petersburg Institute for Informatics and Automation RAS
andrew dot krizhanovsky@gmail.com

© Feiyu Lin

Jönköping University, Sweden
feiyu.lin@jth.hj.se

## Abstract

A set of ontology matching algorithms (for finding correspondences between concepts) is based on a thesaurus that provides the source data for the semantic distance calculations. In this wiki era, new resources may spring up and improve this kind of semantic search. In the paper a solution of this task based on Russian Wiktionary is compared to WordNet based algorithms. Metrics are estimated using the test collection, containing 353 English word pairs with a relatedness score assigned by human evaluators. The experiment shows that the proposed method is capable in principle of calculating a semantic distance between pair of words in any language presented in Russian Wiktionary. The calculation of Wiktionary based metric had required the development of the open-source Wiktionary parser software.

## 1 Introduction

Gruber [6] defined an ontology as, "*an ontology is a formal, explicit specification of a shared conceptualization*". As the main elements of the semantic web, a lot of ontologies are created in different areas and applications. Although these ontologies are developed for various purposes and domains, they always contain overlapping information. To build a collaborative semantic web, it is necessary to find ways to compare, match and integrate various ontologies. Ontology matching which is finding similar entities in the source ontologies or finding translation rules between ontologies is the first step.

There are different strategies in order to find out the similarity between entities in the current ontology matching systems. For example, these strategies can be string similarity, synonyms, structure similarity and based on instances. Synonyms strategy can help to solve the problem of using different terms in the ontologies for the same concepts. For example, an ontology may use "diagram" while the other ontology is using "graph" for the same meaning. Normally synonyms strategy is based on external resources like domain ontology, corpus, thesaurus (e.g., WordNet, Wiktionary).

Goal of this research is to compare WordNet and Wiktionary as the external data sources for a semantic distance calculation and for an ontology matching.

## 2 Ontology matching based on WordNet and Wiktionary

WordNet[1] can be treated as a partially ordered synonym resources. The total of all unique noun, verb, adjective, and adverb strings is actually 147278. WordNet consists of a set of synonyms "synsets" and "gloss" which is the definitions and examples of the concepts. A synset denotes a concept or a sense of a group of terms. Synsets provide different semantic relationships such as synonymy (similar) and antonymy (opposite), hypernymy (superconcept) / hyponymy (subconcept), meronymy (part-of), holonymy (has-a).

Semantic similarity based on WordNet has been widely explored in Natural Language Processing and Information Retrieval. These methods can be classified into three categories [13]:

• Edge-based methods: to measure the semantic similarity between two words is to measure the distance (the path linking) of the words and the position of the word in the taxonomy. For examples see Wu and Palmer [23], Resnik [18].

• Information-based statistics methods: it calculates the probability with concepts in the taxonomy first, then follows information theory. The similarity of two concepts is extent to the specific concept that subsumes them both in the taxonomy. For examples see Resnik [19], Lin [12].

• Hybrid methods: combine the above methods, e.g., X-Similarity [17], Jiang and Conrath [8], Rodriguez [20].

WordNet based semantic similarity methods can be used in two ways in the ontology matching [13]. One way is applying these methods to calculate the entities similarities in two ontologies. If two independent

**Proceedings of the 11th All-Russian Research Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» - RCDL'2009, Petrozavodsk, Russia, 2009.**

ontologies have a common superconcept, some methods like [20] and [17] can be used to measure structure similarity in ontology matching directly.

There are some evaluation works about WordNet based semantic similarity methods, e.g., [2] and [17]. Based on Miller and Charles [14] experiments where the results obtained for 30 pairs nouns were compared with the judgement of each pair on a scale from 0 (not similar) to 10 (total similar) by 38 students, [2] evaluates 5 methods (e.g., Resnik [19], Lin [12], Jiang and Conrath [8], etc.). [17] evaluates 14 methods (e.g., Resnik [19], Lin [12], Jiang and Conrath [8], X-Similarity [17], etc.). Both results show that Jiang and Conrath [8] method gives the best result.

At the time of writing, there are no publications on the use of Wiktionary[2] in ontology matching or related terms search. Nevertheless, one paper [24] describes application programming interfaces for Wikipedia and Wiktionary (English and German Wiktionaries).

## 3 An example of related terms search in ontology matching

To evaluate the increasing number of ontology matching methods and their qualities, OAEI (Ontology Alignment Evaluation Initiative) [3] started arranging evaluation campaigns yearly from 2004. The input of evaluation is two ontologies written in the OWL-DL language. The different elements of ontologies, e.g., concepts, instance and relations can be aligned. The usual output notations are 1:1, 1:m, n:1 or n:m. For example, one entity of one ontology can (e.g., injective, subjective and total or partial) map to entity/entities of the other ontology.

There are a lot of algorithms for semantic similarity which are used for ontology matching. There is following classification of ontology matching algorithms: internal and external [21]. Internal ontology matching algorithm exploit information which comes only with the input ontologies, external ontology matching algorithm exploit external resources such as domain ontology, corpus, thesaurus (e.g., WordNet, Wiktionary).

## 4 A measure of semantic relatedness based on the Russian Wiktionary

An experiments were conducted in order to evaluate the usefulness of the Wiktionary as a resource for related terms search, and consequently for ontology matching. It has been compared with other measures based on WordNet, Wikipedia, Roget's Thesaurus and Google.

### 4.1 Source data

The Russian Wiktionary[4] (the dump of the database as of January 2009) was parsed and the results were stored in a relational database (MySQL). So, the database of the parsed Wiktionary is a source data in the experiment. This database is compared with English and German Wiktionaries in Table 1.

The database of the parsed Wiktionary has a better coverage than WordNet (247,580 words against 150,000). At the same time, WordNet consists of over 115,000 synsets (for a total of 207,000 word-sense pairs) while the total number of semantic relations in the database of the parsed Wiktionary is about 67,000 at this moment (for a total of 177,000 word-sense pairs).

This comparison raises an interesting question: is whether the joint usage of Wiktionary and WordNet can improve the calculation of relatedness measure.

### 4.2 Evaluation based on 353 pairs of English words

WordSimilarity-353 Test Collection (353-TC) consisting of 353 pairs of English words was proposed in [4] in order to evaluate metrics and algorithms which calculates semantic similarity of words.

**Table 1.** The number of entries and selected types of lexical semantic information about three Wiktionaries

| | Wiktionary editions as of September 2007, from [24] | | | | A part of Wiktionary extracted by the parser. Wiktionary edition as of January 2009. | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | English Wiktionary | | German Wiktionary | | Russian Wiktionary | | | | |
| | English | German | English | German | Total[5] | English | German | Russian | Ukrainian |
| Entries | 176,410 | 10,487 | 3,231 | 20,557 | 247,580 | 2,813[6] | 13,072 | 124,301 | 88,575 |
| Part of speech (POS) | | | | | | | | | |
| Nouns | 99,456 | 6,759 | 2,116 | 13,977 | 108,448 | 935 | 336 | 58,843 | 40,607 |
| Verbs | 31,164 | 1,257 | 378 | 1,872 | 26,290 | 342 | 49 | 356[7] | 24,096 |
| Adjectives | 23,041 | 1,117 | 357 | 2,261 | 26,864 | 184 | 18 | 2,168 | 23,536 |
| Unknown | POS which were not recognized by the parser | | | | 80,293 | 1,321 | 12,648 | 57,573 | 331 |
| Semantic relations | | | | | | | | | |
| Synonyms | 29,703 | 1,916 | 2,651 | 34,488 | **28,718** | 1,345 | 665 | 24,338 | 310 |
| Antonyms | 4,305 | 238 | 283 | 10,902 | 10,480 | 238 | 234 | 9,062 | 54 |
| Hypernyms | 42 | 0 | 336 | 17,286 | 18,975 | 444 | 474 | 17,033 | 115 |
| Hyponyms | 94 | 0 | 390 | 17,103 | 8,585 | 176 | 473 | 7,574 | 12 |
| Holonyms | – | – | – | – | 216 | 1 | 0 | 215 | 0 |
| Meronyms | – | – | – | – | 322 | 8 | 2 | 306 | 0 |
| Total | – | – | – | – | **67,296** | 2,212 | 1,848 | 58,528 | 491 |

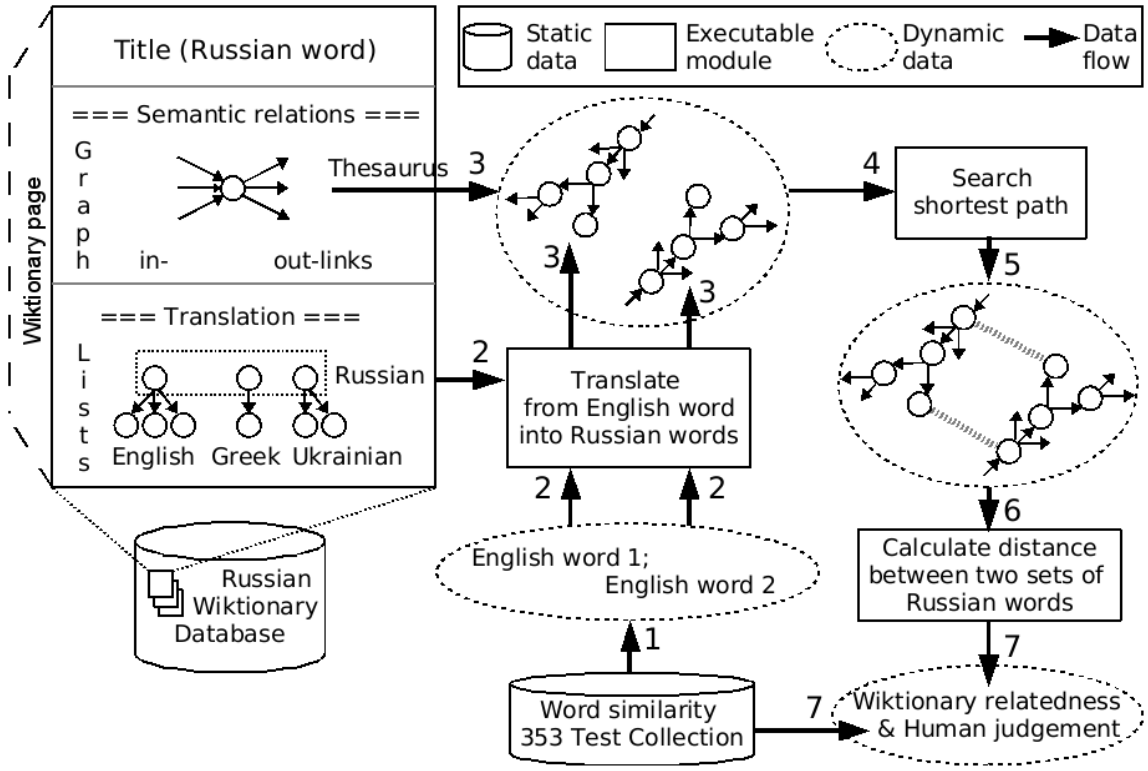



**Fig. 1.** Scheme of the experiment for calculating the semantic relatedness measure based on Russian Wiktionary data

### 4.3 Semantic relatedness measure

The goal is to calculate a relatedness measure between two English words using Russian Wiktionary in order to estimate this measure with the help of 353-TC.

Path based measures (Table 2) are the most attractive candidates because the Wiktionary contains thesaurus for each language (since there are Russian synonyms, English hyponyms, etc.). But the English thesaurus (in Russian Wiktionary) is much more smaller than the Russian one at this moment (2,000 against about 59,000 in Table 1). This sorrowful situation could be remedied simply by using translations from the same Wiktionary.

Thus the Russian thesaurus and translation from English to Russian (as parts of the Russian Wiktionary) will be used (see "Wiktionary page" at Fig. 1).

The process of calculation of relatedness consists of the steps illustrated in Fig. 1:

1. Take a pair of the English words from 353-TC;
2. Translate the English words to the two sets of Russian words using translations from Russian into English in the Russian Wiktionary;
3. Mark the vertices of the subgraph 1 (corresponding to the set 1 of Russian words in the thesaurus) and the subgraph 2;
4. Search the shortest paths between the marked vertices of the subgraphs 1 and 2 (Dijkstra's algorithm);
5. Extract the list of words which link two Russian words in the thesaurus, see Fig. 2 (optional step);
6. Calculate the distance between two sets of Russian words ($path^{max}_{len}$) using the paths lengths (found in step 4);
7. Save the result of the calculation for the comparison with the human judgement.

The preliminary computations include the constructing of the graph corresponding to the thesaurus and calculating the shortest paths between all pairs of nodes (for the steps 3 and 4).

Semantic relatedness measure $path^{max}_{len}$ is a maximum of lengths of shortest paths from each word of a Russian words set 1 to each word of a Russian words set 2 (Fig. 1, step 6). This measure was calculated for 353 English word pairs. The correlation with human judgements from the test collection 353-TC is 0.24 (see column "WT" in Table 2).

Let us stress that we do not make direct comparison between English words only due to the reason of a small English thesaurus in Russian Wiktionary. Notice that a development of an English Wiktionary parser will make the translation step unnecessary (for English words).

The number of cases where the distances between word pairs could not be calculated due to an absent data in the Russian Wiktionary (an absent page, or a translation, or there are no semantic relations) is 115 (32% of 353 word pairs).

### 4.4 Comparison with other metrics

The central place in the paper is occupied by Table 2 with the evaluation of semantic distance calculation algorithms and metrics against the test collection 353-TC.

The substantial part of estimations (metrics *jaccard*, *text*, *res* $_{hypo}$) was taken from [22]. Also *res* $_{hypo}$ metric was estimated in our previous paper [10]. Table 2 includes experimental data found in the following publications: [7] (the metric *jarmasz*), [4] (the search engine *IntelliZap* and *LSA* algorithm), [5] (*ESA* algorithm). The notion about the rest of the metrics could be found in other papers: the metric *wup* [23], *lch* [3] (p. 265-283), *res* [18], and *lesk* [1].

Table 2 shows correlations between the test collection 353-TC and the listed above metrics, algorithms. There are the following metrics and algorithms yielding the best results, which take into account:

I. *taxonomy structure* – 0.48, the metric *lch* [3] (English Wikipedia dataset) and 0.539, the metric *jarmasz* [7] (Roget's Thesaurus);
II. *words frequency in corpus* – 0.75, *ESA* algorithm [5] (English Wikipedia);
III. *text overlapping* – 0.21, the metric *lesk* [1] (WordNet).

Out of the scope is Green [15] algorithm (search in Wikipedia), which was not tested with 353-TC.

Table 2 shows that the Wiktionary based semantic similarity metric yields the worst result (0.24) among the path based measures (I), but it is comparable with values of text overlapping metrics (III). Best WordNet-based metrics (with value 0.34) are *lch* [3] and *res* [18].

**Table 2**. Results on correlation with human judgements of relatedness measures 353-TC to measures based on WordNet (WN), English Wikipedia (WP), Russian Wiktionary (WT)

| Dataset | **WN** | WP | **WT** | Others |
|---|---|---|---|---|
| **Metric or Algorithm** | I. Path based measures (in taxonomy) | | | |
| **wup** | 0.3 | 0.47 | – | – |
| **lch** | **0.34** | **0.48** | – | – |
| **res** $_{hypo}$ | – | 0.25-0.37[8] | – | – |
| **jarmasz** | – | – | – | **0.539** RT[9] |
| **path** $^{max}_{len}$ | – | – | **0.24** | – |
| | II. Words frequency in corpus | | | |
| **jaccard** | – | – | – | Google 0.18 |
| **res** | 0.34 | – | – | – |
| **LSA** | – | – | – | IntelliZap 0.56 |
| **ESA** | – | **0.75** | – | – |
| | III. Text overlapping | | | |
| **lesk** | 0.21 | 0.2 | – | – |
| **text** | – | 0.19 | – | – |

## 5 Implementation

The Wiktionary parser is a part of Wikokit project [10]. The software programming code is based on our previously developed Wikipedia indexing system [11] and the system that searches for related terms by analysing Wikipedia internal links [9].

The database (wikt_parsed) storing data extracted from the Wiktionary was designed using the visual tool MySQL Workbench. A part of lexicographic information from Russian Wiktionary texts has been extracted and stored into this database, namely:

- a word itself (stored into the table *page*);
- a word's language and a part of speech (tables *lang_pos*, *lang*, *part_of_speech*);
- a definition (table *meaning*);
- links for key words in the definition, in the translations, in the semantic relation, i.e. in any wikified text (tables *wiki_text*, *wiki_text_words*, *page_inflection*, and *inflection*);
- semantic relations (tables *relation* and *relation_type*);
- translations (tables *translation* and *translation_entry*), where one record in the table *translation* corresponds to one meaning, and one record in the table *translation_entry* corresponds to the translation of this meaning into one language.

The developed software provides API (application programming interface) that will store and retrieve information from the database of the parsed Wiktionary. This API was used for calculating the relatedness between words in Wiktionary.

A shortest path computation on a graph (Fig. 2) was easily implemented within the Java Universal Network / Graph Framework (JUNG). The JUNG Framework is a free, open-source software library that provides a language for the manipulation, analysis, and visualization of data that can be represented as a graph or network [16].

Fig. 2 shows another problem that arises during a creation of a thesaurus from the Wiktionary data. This is a word sense disambiguation (WSD) task.

See the entry "журнал" in Russian Wiktionary:

1. "дневник" (journal, diary) is a near-synonym of "журнал" (journal) (in accordance with meaning number 2 of "журнал");
2. "издание" (magazine, journal) is a hyperonym of "журнал" (in accordance with meaning number 1 of "журнал").

The entry "дневник" (journal, diary) describes that "журнал" (journal) is a near-synonym of "дневник" with, regrettably, no number of "журнал" meaning number.

Thus, within one Wiktionary page (entry, word) the different meanings (definitions) are presented explicitly, hence the lists of synonyms, antonyms, etc. are also explicitly marked by the number of meaning. But other entries (which "mention" this word by hyperlinks in a definition, or synonym, or translation sections) do not

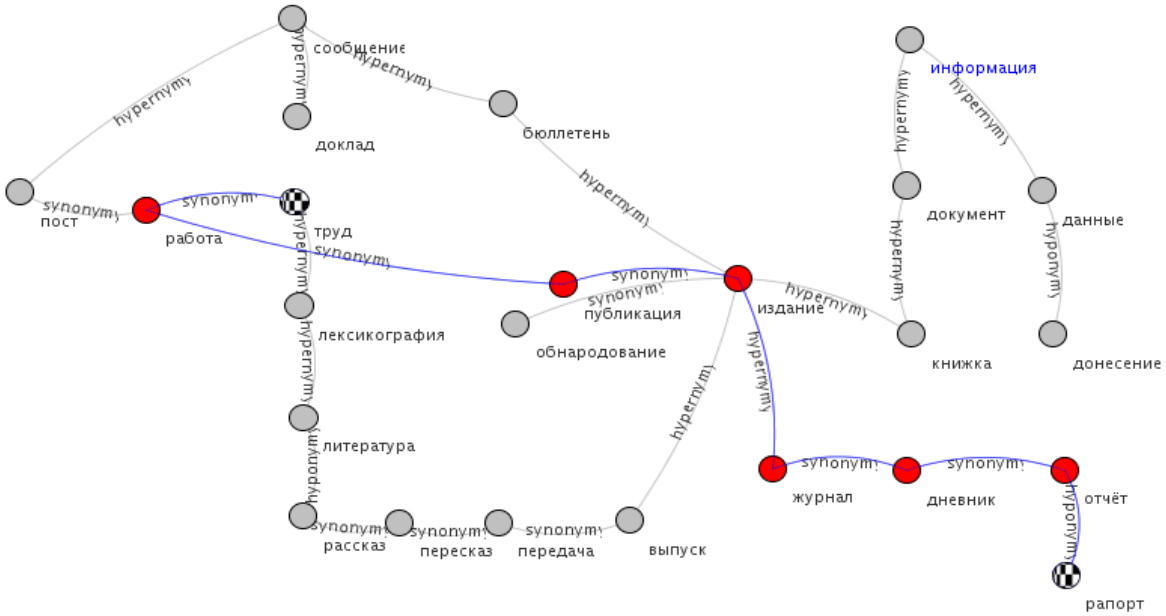



**Fig. 2.** A shortest path from a Russian word “рапорт” (raport) to a word “труд” (work, labour) found in a thesaurus of the Russian Wiktionary (“рапорт”, “отчёт”, “дневник”, “журнал”, “издание”, “публикация”, “работа”, “труд”)

indicate the meaning number of this entry.

It’s not a big problem for the reader, but it requires the additional worthwhile programming work of disambiguating the meanings of the words listed in the “semantic relations” section of a Wiktionary page. Thus, a WSD algorithm should be developed or adapted to the Wiktionary in order to solve this problem.

## 6 Discussion and conclusion

Wiktionary has advantage over WordNet in being of a larger size in number of words, but the database of the parsed Wiktionary has less number of relations (see Table 1). The experiment shows that a similarity calculation between two English words by the joint usage of a thesaurus and translations extracted from Russian Wiktionary could not hope to win a victory over WordNet so far.

But the present experiment shows that the proposed method is capable in principle of calculating a semantic distance between pair of words in any language presented in Wiktionary (more than 200 in Russian Wiktionary).

It should be noted that (1) other language editions of Wiktionary are out of the scope of this paper, (2) only a small part of lexicographic information from Russian Wiktionary texts has been extracted and stored into machine readable dictionary, namely:

- a word’s language,
- a part of speech,
- a definition,
- links in the definition for key words,
- semantic relations,
- and translations.

An extraction from Wiktionary of a pronunciation (phonetic transcription, a sound sample), a hyphenation, an etymology, a quotation (example sentence), a parallel text (examples with translations), a figure (which illustrates a word meaning) were not considered because this is a first step towards the creation of an open-source Wiktionary parser software.

## References

[1] S. Banerjee, T. Pedersen. An Adapted Lesk algorithm for word sense disambiguation using WordNet. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-02). Mexico City, February, 2002. http://www.d.umn.edu/~tpederse/Pubs/cicling2002-b.pdf

[2] A. Budanitsky, G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures, pp. 29–24, 2001.

[3] C. Fellbaum. WordNet: an electronic lexical database. – MIT Press, Cambridge, Massachusetts – 423 pp. – ISBN 0-262-06197-X. 1998.

[4] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin. Placing search in context: the concept revisited. In *ACM Transactions on Information Systems*, volume 20(1), pp. 116-131, 2002. http://www.cs.technion.ac.il/~gabr/papers/tois_context.pdf

[5] E. Gabrilovich, S. Markovitch. Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI). Hyderabad, India, January, 2007. http://www.cs.technion.ac.il/~gabr/papers/ijcai-2007-sim.pdf

[6] T. R. Gruber. A translation approach to portable ontology specifications. In *Knowledge Acquisition*, volume 5(2), pp. 199-220, 1993.
[7] M. Jarmasz, S. Szpakowicz. Roget's Thesaurus and semantic similarity. In Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP). Borovets, Bulgaria, pp. 212-219, 2003. http://www.nzdl.org/ELKB/
[8] J. J. Jiang and D.W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, 1997.
[9] A. A. Krizhanovsky. Synonym search in Wikipedia: Synarcher. In: 11-th International Conference "Speech and Computer" SPECOM'2006. Russia, St. Petersburg, pp. 474-477, 2006. http://arxiv.org/abs/cs/0606097
[10] A. A. Krizhanovsky. Evaluation experiments on related terms search in Wikipedia: Information Content and Adapted HITS (In Russian). 2007. http://arxiv.org/abs/0710.0169
[11] A. A. Krizhanovsky. Index wiki database: design and experiments. In: Corpus Linguistics, 2008. http://arxiv.org/abs/0808.1753
[12] D. Lin. An Information-Theoretic Definition of Similarity. In Proceedings of the Fifteenth International Conference on Machine Learning, 1998.
[13] F. Lin, K. Sandkuhl. A Survey of Exploiting WordNet in Ontology Matching. In Proc. IFIP AI, 2008.
[14] G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. In *Language and Cognitive Processes*, volume 6, pp. 1-28, 1991.
[15] Y. Ollivier, P. Senellart. Finding related pages using Green measures: an illustration with Wikipedia. In Association for the Advancement of Artificial Intelligence. Vancouver, Canada, 2007. http://pierre.senellart.com/publications/ollivier2006finding.pdf
[16] O'Madadhain, D. Fisher, P. Smyth, S. White, and Y.-B. Boey. Analysis and visualization of network data using JUNG (preprint). *Journal of Statistical Software*, pp. 1–35. 2007. http://jung.sourceforge.net/doc/JUNG_journal.pdf
[17] E. G. M. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou. Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Different Ontologies. In Book Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Different Ontologies, pp. 44-52, 2006.
[18] P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Book Using Information Content to Evaluate Semantic Similarity in a Taxonomy, pp. 448—453, 1995.
[19] P. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. In *Journal of Artificial Intelligence Research*, volume 11, pp. 95-130, 1999.
[20] M. Andrea Rodrguez and J. E. Max. Determining Semantic Similarity among Entity Classes from Different Ontologies. In *IEEE Trans. on Knowl. and Data Eng.*, volume 15(2), pp. 442-456, 2003.
[21] P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. In *Journal on Data Semantics*, (4), pp. 146–171, 2005.
[22] M. Strube, S. Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 06). Boston, Mass., July 16-20, 2006. http://www.eml-research.de/english/research/nlp/publications.php
[23] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994.
[24] T. Zesch, C. Mueller, I. Gurevych. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In Proceedings of the Conference on Language Resources and Evaluation (LREC), 2008. http://elara.tk.informatik.tu-darmstadt.de/publications/2008/lrec08_camera_ready.pdf

## Поиск семантически близких слов на основе WordNet / Викисловаря, его применение в задаче сопоставления онтологий

Андрей Крижановский, Фейу Лин

При установлении соответствия между элементами онтологий (Ontology Matching) ряд алгоритмов для вычисления семантического расстояния привлекает данные тезаурусов. Эпоха вики (wiki) знаменательна новыми ресурсами, применение которых даёт шанс улучшить этот тип семантического поиска. В работе сравниваются алгоритмы, решающие задачу вычисления семантического расстояния на основе данных Русского Викисловаря и WordNet. Алгоритмы и метрики оцениваются с помощью тестовой коллекции из 353 пар английских слов, включающей оценку экспертов. Эксперимент показал, что предложенный метод позволяет вычислить семантическое расстояние между парой слов, в принципе, на любом из языков, представленных в Русском Викисловаре. Проведение вычислений на основе данных Викисловаря потребовало разработки программного обеспечения с открытым исходным кодом – парсера Викисловаря.

♣ Part of this work was financed by the Foundation (The Swedish Institute), project CoReLib.
The research is supported partly by the project funded by grant 08-07-00264 of the Russian Foundation for Basic Research, and project 213 of the research program "Intelligent information

technologies, mathematical modelling, system analysis and automation" of the Russian Academy of Sciences.

[1] WordNet, http://wordnet.princeton.edu

[2] Wiktionary, http://wiktionary.org

[3] OAEI, http://oaei.ontologymatching.org/2008

[4] Russian Wiktionary, http://ru.wiktionary.org

[5] Total, i.e. all languages in the Russian Wiktionary.

[6] Different parts of speech are considered as different entries (table *lang_pos* in the database of the parsed Wiktionary).

[7] Russian Wiktionary contains no doubt more than 356 Russian verbs, but only a part of verbs was successfully extracted by the parser.

[8] 0.25, 0.37, where 0.37 was taken from [22] (English Wikipedia, as of Feb. 2006), and 0.25 from [10] (English Wikipedia, as of May 2007).

[9] RT – Roget's Thesaurus [7].

[10] Wikokit. http://code.google.com/p/wikokit

# Роли онтологий в электронной библиотеке КарНЦ РАН*

© В.А. Лебедев

Институт прикладных математических исследований КарНЦ РАН
Vl777@krc.karelia.ru

**Аннотация**

В статье рассмотрены проблемы построения онтологий научных дисциплин по описанию изученности природных объектов и систем, применение онтологий для систематизации и комплектации контента ЭБ КарНЦ РАН. Интеграция информационных ресурсов контента сопровождается их индексацией при помощи онтологий с целью последующего тематического поиска с использованием тех же онтологий для построения запросов.

## 1 Введение

Известный американский ученый Бернерс-Ли (Berners-Lee) в 2001 году предложил концепцию семантического Интернет (Semantic Web) как средство упорядочения контента сети и тем самым существенного уменьшения затрат ручного труда при поиске [12, 13]. Идея состоит в том, чтобы каждый информационный ресурс в сети сопровождала онтология, включаемая как значение специального атрибута в составе метаданных, например, в схеме метаданных Dublin Core это может быть атрибут Subject (тема) или Description (описание). В качестве информационного ресурса может выступать отдельный документ и/или коллекция документов. И каждый документ, включенный в коллекцию, и коллекция в целом сопровождаются онтологиями, причем каждая из них является фрагментом более общей онтологии предметной области, к которой относится публикуемый информационный ресурс. Имея открытую онтологию предметной области, возможно автоматически индексировать публикуемые ресурсы. Индекс при этом будет представлять фрагмент онтологии предметной области, т.е. будет онтологией документа. В этих условиях пользователь может применить для поиска программный агент, содержащий фрагмент онтологии предметной области и алгоритм для сравнения с ним онтологий ресурсов, найденных в сети. Тогда отклик на запрос практически не будет содержать информационного шума[1].

Очевидно, что для реализации Semantic Web в качестве первого шага требуется создание онтологий предметных областей, поэтому группа экспертов, работавшая по заданию Правительства РФ по определению перспективных направлений разработки информационно-коммуникационных технологий в России, определила как одно из приоритетных направление «Общедоступные методы и программные средства построения русскоязычных схем систематизации контента (программирование номенклатур, таксономий и онтологий предметных областей)» [10]. Под онтологией здесь понимается эксплицитная спецификация концептуализации предметной области, которая подразумевает использование некоторой математической модели и языка реализации этой спецификации [6, 14, 15].

Формально онтология определяется как **O = <X,R,F>**, где

- X — конечное множество понятий предметной области,
- R — конечное множество отношений между понятиями,
- F — конечное множество функций интерпретации [7].

Как видно, по форме – это определение графа с помеченными вершинами, где X – множество вершин, R – множество дуг, F- множество помет. Значение пометы интерпретируется некоторой функцией (функциями).

Будем понимать онтологию как иерархический граф связей терминов и названий, принятых в предметных областях, с их толкованиями, пометами и функциями интерпретации помет [11]. Основание: объекты, подлежащие описанию в ЭБ, обладают иерархической структурой, множество их свойств изучает комплекс научных дисциплин, иерархически соподчиненных, что и отображается в графе связей терминов. В составе ЭБ КарНЦ РАН [3-5] онтологии будут выполнять следующие роли:

- моделировать и представлять контент ЭБ,
- способствовать созданию контента, удовлетворяющего требованиям актуальности, достоверности и полноты,
- обеспечивать автоматическую индексацию электронных публикаций и
- построение точных запросов на поиск.

**Труды 11ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

Таким образом, необходимо разработать технологию построения и построить онтологии по биологии и науках о Земле, технологию формирования контента ЭБ с использованием онтологий, отработать технологию поиска релевантных документов при помощи онтологий.

Ряд функций указанных технологий нами был осуществлен ранее и опубликован в серии докладов на российских конференциях [1, 2, 8, 9]. В докладе излагаются решения и схемы технологий новых функций.

## 2 Построение онтологий

Для обеспечения достоверности, актуальности и полноты предметных онтологий необходимо разработать методологию их создания, соответствующую целям ЭБ КарНЦ РАН.

В качестве основы методологии принята следующая парадигма[2].

Природа Карелии состоит из объектов (предметов), подразделяемых на классы в соответствии с классификацией наук и научных дисциплин (например, по рубрикатору ГРНТИ). Каждый класс объектов характеризуется некоторым набором свойств (атрибутов), принимающих значения из соответствующих областей значений (доменов). Некоторые подмножества свойств объявляются признаками и используются непосредственно или их значения для различных классификаций объектов внутри класса. Множество свойств объектов разбивается на группы (темы), изучение которых является предметом соответствующей научной дисциплины.

Каждый объект вступает во взаимодействие с другими объектами, что является основой для выделения различных систем и подсистем. В системах объекты выполняют некоторые роли (функции), которые могут иметь различные оценочные названия (враги, союзники и т. п.), или выражаются соответствующими формулами.

Каждый объект любого класса обладает некоторым строением, то есть состоит из набора частей (тоже объектов), вступающих во взаимодействия и является системой (агрегатом).

Выделяют внешнее строение (морфологию) и внутреннее (анатомию).

Взаимодействие объектов в системах в некотором масштабе времени может быть неизменным (статика) или меняющимся (динамика). Подразделение взаимодействий объектов на классы и виды определяется в соответствующих научных дисциплинах. Статика определяет устойчивость, а динамика (процессы) – внешнее поведение (этологию), внутреннее функционирование (физиологию), происхождение, становление (генетика, генезис). Термины в скобках понимаются расширительно, в предметных областях конкретизируются и детализируются.

Методология построения предметных онтологий, основанная на данной парадигме, определяет структуру графа связей понятий (точнее, терминов и названий) предметной области.

Необходимо установить список классификаций и номенклатуры их классов. При необходимости зафиксировать соответствие (например, в виде табличной функции) значений признаков и классов. Установить номенклатуру (список) свойств (атрибутов) объектов класса, изучаемую данной научной дисциплиной. Затем определить их домены.

Некоторые термины являются многословными сочетаниями (например, сухие сосновые и смешанные леса, сырые засфагненные луга). Такого рода термины будем трактовать как конкатенацию названий классов различных независимых классификаций. В тех же примерах: леса, луга – типы растительности; сухие, сырые – классы по влажности; сосновые, засфагненные – классы по преобладающим видам растений и т.д.

Для обеспечения удобства поиска такого рода классификации в составе таксономии разносятся по уровням иерархии. В запросе они представляются в виде конъюнктивной (например, сырые $\wedge$ засфагненные $\wedge$ луга) или конъюнктивно-дизъюнктивной формы (например, сухие $\wedge$ (сосновые $\vee$ смешанные) $\wedge$ леса .

Далее следует устанавливать термины и названия, относящиеся к морфологии, анатомии, этологии и физиологии, то есть зафиксировать номенклатуры названий частей объектов и систем, их функции и оценки. При этом учитываются следующие типы отношений: классификации, агрегации, синонимии и полисемии. Технология, реализующая указанную методологию, состоит в следующем:

- Корневые понятия (термины) предметных областей принимаются по рубрикатору ГРНТИ.
- Начиная с корневых понятий, организуем поиск их значений (толкований) в Интернет или словарях.
- Используя найденное толкование, выделяем в нем термины более детальных понятий и ищем их толкования.
- Поступаем аналогично с терминами следующего уровня. И так до уровня значений свойств.
- В процессе поиска и нахождения терминов и их толкований фиксируем термин и URL статьи с наиболее полным толкованием его значения в связи с термином предыдущего уровня.


Рис. 1.

| Метка | Предок | Потомок |
|---|---|---|
| Биология А | Экология | Сообщества(экосистемы.биоценозы) |
| | | Связи |
| | | Виды |
| | | Популяции |
| | | Охрана окружающей среды |
| | Сообщества | Суша |
| | | Пресные воды |
| | | Моря |
| | | Атмосфера |
| Сообщества суши К | Типы | Биосфера |
| | | Зона |
| | | Подзона |
| | | Район |
| | | Ландшафт |
| | | Биогеоценоз |
| | | Местообитание(биотоп) |
| Сообщества суши К | Зона | Арктическая пустыня |
| | | Тундра |
| | | Лесотундра |
| | | Лес |
| Сообщества суши К | Подзона | Северная тайга |
| | | Средняя тайга |
| | | Южная тайга |
| | | Широколиственные леса |

Рис. 2.

- Таким образом, определяем как номенклатуру терминов и их связи, так и адреса (URL) толкований.
- После этого материалы передаются на экспертизу специалистам-предметникам, и по результатам экспертизы итерационно выполняется построение таксономий и механизма ссылок на толкования терминов.

В целом таксономия онтологии будет иметь структуру иерархического графа (древовидного или с полуциклами), фрагмент которого представлен на рис. 1.

Вершины графа – термины, дуги – отношения между ними (классификации и агрегации), отношения помечаются в узле разветвления. Отношения синонимии выделены в отдельную структуру (словарь). Полисемия (то есть наличие одинаковых по написанию терминов) разрешается в виду того, что такие термины могут находиться только в разных частях структуры.

Реализация таксономии представляется в виде таблицы (рис. 2), точнее, базы данных реляционной или объектной. Технология загрузки и редактирования таксономии и словаря синонимов отработана. Ведется разработка онтологий первой очереди по геологии, водным ресурсам, ботанике, зоологии, почвоведению, экологии, лесоведению и лесоводству. В настоящее время онтологии содержат около 2000 терминов, не считая видовых названий растений, минералов, химических соединений.

## 3 Технология формирования контента

На первых этапах создания ЭБ контент формировался только с привлечением специалистов КарНЦ РАН. С появлением онтологических моделей предметных областей возникает возможность расширения контента с использованием материалов, опубликованных в Интернете. Для осуществления этой возможности

разработана технология интеграции сторонних материалов в контент ЭБ.

Сущность и назначение интеграции сторонних материалов состоит в том, чтобы обеспечить доступ к ним посредством поисковых сервисов, имеющихся в ЭБ. Это сервисы поиска 1) по названиям коллекций и их документов и 2) с использованием онтологий для формирования тематических запросов.

Каждая коллекция (собственная или внешняя) должна пройти процесс импортирования, который включает формирование: записи в списке коллекций и списка документов коллекции.

Записи в списке коллекций содержат название коллекции и ее URL в виде гиперссылки. Аналогично записи в списке документов также содержат их названия и гиперссылку на текст документа. Для формирования этих списков имеется соответствующая технология [3, 4].

Отличие процесса регистрации интегрируемых сторонних коллекций заключается в том, что их документы могут быть представлены в различных форматах (HTML, PDF и др.) и могут не содержать списка документов в явном виде. Таким образом, необходимо будет разработать дополнительные технологические средства для формирования списков документов привлекаемых коллекций, аналогично тому, как формируются списки терминов онтологий и их толкований.

Для обеспечения тематического поиска документов в коллекциях производится их индексация с использованием соответствующей предметной онтологии. Структура индексного файла – это таблица, которая содержит имя документа, его URL и список встречающихся в его тексте терминов в порядке их иерархии и связей в онтологии.

Документы ЭБ по степени структуризации можно разделить на три категории: базы данных (таблицы), слабо структурированные (XML - документы) и неструктурированные (статьи в форматах PDF, HTML и т.п.). Таблицы и XML-документы структурно соответствуют структуре онтологии, поэтому процесс их индексации сравнительно прост. Документ прочитывается пословно. При этом рубрики документа соответствуют рубрикам онтологий, что обеспечивает сохранение порядка терминов в индексе, принятому в онтологии. Это важно для последующего поиска релевантных.

Неструктурированные документы могут содержать термины не в порядке их подчиненности в онтологии. Тогда, если не принять особых мер при их индексации, индекс документа будет содержать список терминов в порядке их нахождения в тексте, а не в порядке, принятом в онтологии, что впоследствии будет порождать информационный шум.

Чтобы избежать этого, в тексте документа в процессе его чтения сначала ищутся термины, близкие к корню онтологии. И если найден один, то дальше ищутся термины, подчиненные ему вплоть до листьевых терминов, и они помещаются в индекс. Далее ищется следующий термин корневых уровней и подчиненные ему и т. д. В результате список терминов в индексе будет иметь порядок, соответствующий онтологии.

После выполнения указанных операций сторонние коллекции считаются интегрированными в ЭБ и доступ к ним осуществляется при помощи сервисов нашей ЭБ.

Интеграция статей толкований терминов онтологий отличается тем, что списки статей, относящихся к данной предметной онтологии, включаются в соответствующий индексный файл. Причем индексация статьи может не производиться.

Реальные объекты, описываемые в документах коллекций, вступают между собой в различные отношения, которые указываются в виде их ролей в составе системы. Целесообразно использовать эту информацию для создания гиперссылок между документами. В результате получаем не просто наборы документов, а комплексы связанных документов, что полезно при изучении. Для решения этой задачи разработана соответствующая технология [2]. В итоге получим распределенную библиотеку, содержащую описание классов объектов Карелии и толкования терминов онтологии.

## 4 Поиск в ЭБ релевантных документов

Преимущество в использовании онтологий для формирования запросов на поиск заключается в том, что запрос в этом случае представляет собой фрагмент таксономии, в котором термины связаны в иерархию. Тем самым запрос уже не является простым списком терминов, а отражает их зависимость. При этом устраняется возможная полисемия терминов и тем самым отсекается значительная часть информационного шума в отклике на запрос.

Ранее нами была предложена обобщенная схема запросов [1], которая представляет собой редукцию предикатного выражения, а именно, нетерминалы в угловых скобках обозначают предикаты вида X=a, где X — слово в составе индекса, а — термин в запросе. С учетом объединения предметных онтологий на основе рубрикатора ГРНТИ схема запроса принимает следующий вид:

{ <Рубрика ГРНТИ><коллекция>}[<класс> ∧∨ …∧][<агрегат (тема)> ∧∨ … ∧]
[<характеристика> ∧][<список значений> ∧∨]… ∧∨
[<характеристика M > ∧][<список значений> ∧∨][<тема>∧∨ … ∧]
[<характеристика N > ∧][<список значений> ∧∨]… ∧∨

| Рубрика ГРНТИ | Коллекция | Тема | Характеристика 1 | Список значений 1 |
|---|---|---|---|---|
| Биология⋁ Ботаника | Сосудистые растения ⋀ | Экология растений ⋀ | Местообитания ⋀ | Сухие⋀Смешанные⋀ Леса⋁Поляны |

| Характеристика 2 | Список значений 2 |
|---|---|
| Хозяйственное значение ⋀ | Лекарственное ⋀ Противовоспалительное ⋀Пищевое⋀ (Ягоды ⋁Орехи) |

Рис. 3.

[<характеристика N + К > ⋀][<список значений ⋀ /⋁>]
[<класс> ⋀/⋁ … ⋀][<тема> ⋀/⋁ … ⋀][<характеристика> ⋀]
[<список значений> ⋀/⋁] … ,
где
<список значений> : = <значение> ⋀ /⋁<список значений>.

Наличие квадратных скобок указывает на возможность формирования и самых простых запросов из одного термина. При использовании дизъюнкций в составе фрагментов, заключенных в квадратные скобки необходимо правильно расставить круглые скобки, чтобы учитывать приоритеты логических операций.

Нетерминалы «Рубрика ГНТИ» и «коллекция», заключенные в фигурные скобки, определяют как раздел онтологии, так и коллекцию документов, в которой должен выполняться поиск. Очевидно, что поиск осуществляется в одной коллекции. При необходимости поиска в большем их числе запросы должны повторяться.

Нетерминалы «класс», «тема», «характеристика», «значение» отражают иерархическую структуру онтологии. При этом нетерминалы «класс» и «тема» подразумевают возможность иерархических классификаций.

Пример формирования запроса показан на рис. 3.

Очевидно, что построение запроса по указанной схеме непростая задача, поэтому предусмотрены средства оказания помощи пользователю в составлении запроса. Сначала он в процессе поиска в глубину по онтологии формирует список терминов для помещения в запрос, и только далее, обдумав свои потребности, переносит термины в запрос в порядке, определяемом иерархией и советами инструкций. При этом расстановка знаков конъюнкции и дизъюнкции и формирование оператора Select выполняется при помощи программы сервиса, которая контролирует допустимость конъюнктивных связей между терминами, как это показано ниже.

В нашем случае онтология представляет собой множество терминов предметной области, связанных между собой отношениями классификации, агрегации и синонимии.

Классификации разбивают некоторые исходные множества на группу непересекающихся подмножеств (классов) по определенным основаниям, в качестве которых могут использоваться наличие или отсутствие у объекта определенных атрибутов (признаков) и/или определенных значений атрибутов. Классификации могут быть одноуровневыми или многоуровневыми (иерархическими), многоуровневость производит последовательная классификация сначала исходного множества, а затем его подмножеств, подмножеств этих подмножеств и т.д. Одно и то же множество может быть классифицировано несколько раз с использованием различных оснований. По определению, в классификациях допускаются конъюнкции между терминами, лежащими на одном пути в графе онтологии. Все остальные конъюнкции являются пустыми, т. к. связывают непересекающиеся подмножества. В этих условиях, чтобы проверить допустимость любой конъюнкции, заданной в запросе, достаточно проверить, лежат ли входящие в нее термины на одном пути в онтологии и, если не лежат, сообщить об этом пользователю, чтобы он исключил эти конъюнкции из запроса.

Агрегации в отличие от классификаций позволяют представить класс объектов в виде совокупности частей или свойств. Отдельные объекты класса описываются указанием значений свойств или порядковых номеров частей (присваиваемых в процессе производства). При поиске объектов класса в коллекциях в этих случаях допускаются конъюнкции между названиями свойств или частей. При поиске конкретных объектов нужно указывать также значения свойств или номера частей. Запись конъюнкции должна быть аналогичной предыдущей схеме.

В данном случае допустимые конструкции в запросах для классификаций и агрегаций вступают в противоречие. Для его разрешения достаточно пометить в онтологии классификации и агрегации различными знаками (например, классификации метятся буквой К, а агрегации – буквой А) (см. рис. 2). Тогда упомянутый выше контроль допустимости конъюнкций достаточно дополнить анализатором этих пометок.

Синонимические гнезда терминов в онтологиях представляются отдельными словарями синонимов.

Когда пользователь помещает в запрос очередной термин, выполняется поиск в словаре синонимов, и если они есть, то автоматически в запрос помещается дизъюнкция всего синонимического гнезда. Тем самым предотвращается возможность включения в запрос пустых конъюнкций. Доработка и реализация алгоритмов индексирования документов выполнена Старковой В.Г..

## Литература


[1] Вдовицын В. Т., Лебедев В. А. Онтологии для тематического поиска данных в коллекциях электронной библиотеки. // Труды десятой Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции”. Дубна. 2008. С. 63-69.

[2] Вдовицын В. Т., Лебедев В. А., Брагин С. В., Старкова В. Г., Луговая Н. Б. Развитие сервисов электронной библиотеки научных информационных ресурсов //Труды Всероссийской научной конференции Научный сервис в сети Интернет: технологии параллельного программирования, г. Новороссийск, 24 – 29 сентября 2007 г. Издательство Московского университета. 2007. С. 305-310.

[3] Вдовицын В. Т., Лебедев В. А., Луговая Н. Б., Сорокин А. Д., Старкова В. Г.. Развитие и разработка технологии публикации и поиска документов в электронных коллекциях // Труды Восьмой Всероссийской научной конференции по электронным библиотекам, Суздаль, 2006. С. 162-167.

[4] Вдовицын В. Т., Сорокин А. Д., Луговая Н. Б.. Развитие программных сервисов и контента ЭБ КарНЦ РАН. // Труды Седьмой Всероссийской научной конференции по электронным библиотекам, Ярославль, 2005. С. 92-97.

[5] Вдовицын В. Т., Сорокин А. Д., Луговая Н. Б.. Электронная библиотека научных информационных ресурсов КарНЦ РАН. // Труды Шестой Всероссийской научной конференции по электронным библиотекам, Пущино, 2004. С. 41-46.

[6] Добров Б. В., Лукашевич Н. В. и др. Разработка лингвистической онтологии для автоматического индексирования текстов по естественным наукам // Труды Седьмой Всероссийской научной конференции по электронным библиотекам, Ярославль, 2005. С. 70-76.

[7] Загорулько Ю. А. Методы и метологии разработки, сопровождения и реинжиниринга онтологий. Онтологическое моделирование. Труды Симпозиума. Звенигород, май 2008. С. 167-200.

[8] Лебедев В. А., Старкова В. Г., Брагин С. В. Представление онтологии научной коллекции «Водные ресурсы региона» // Труды шестой Всероссийской конференции по электронным библиотекам. Пущино, 2004. С. 86-92.

[9] Лебедев В. А., Старкова В. Г., Брагин С. В. Применение онтологии для ведения и доступа к данным коллекции «Природные ресурсы региона». // Труды седьмой Всероссийской конференции по электронным библиотекам». Ярославль, 2005. С. 87-91.

[10] Перспективные направления развития российской отрасли информационно-телекоммуникационных технологий (Долгосрочный технологический прогноз Российской IT — Foresight) М. , 2007. 223 с.

[11] Фазлиев А. З. Рассуждения о понятии “онтология”. Онтологическое моделирование. Труды Симпозиума. Звенигород, май 2008. С.278-296.

[12] Хорошевский В. Ф. Онтологические модели и Semantic Web: откуда и куда мы идем? Онтологическое моделирование. Труды Симпозиума. Звенигород, май 2008. С. 13-45.

[13] Berners-Lee T., Hendler J., Lassila O. The Semantic Web. Scienific American. 2001.

[14] Gruber T. R. A Translation Approach to Portable Ontology specification // Knowledge Acquition, N 5, 1993.

[15] Uschold M., Gruninger M. Ontologies: Principles, Methods and Applications. // Knowledge Engineering Review, N 11, 1996.


## Roles of ontologies in Karelian Research Centre's digital library


Lebedev V. A.
Institute of Applied Mathematical Research,
Russian Academy of Sciences



The paper considers problems of building ontologies of scientific disciplines by descriptions of the degree of coverage of natural objects and systems by studies, application of the ontologies to systematization and compilation of the contents of the Karelian Research Centre's digital library. Integration of the information resources in the contents is accompanied by their indexing through ontologies to enable further thematic search using the same ontologies to build queries.



[*] Работа поддержана грантом РФФИ № 08-07-00085а.


[1] Информационный шум – документы, ошибочно включенные в состав отклика на запрос из-за содержания в них ряда терминов, обладающих полисемией в различных предметных областях.

[2] Парадигма – это наиболее общая картина устройства природы, в данном случае – описания изученности природных объектов и систем.

# КОЛЛЕКЦИИ НАУЧНЫХ ДАННЫХ

# COLLECTIONS OF SCIENTIFIC DATA

# К инвариантным моделям пульсарных данныхв пространственно-временных координатных системах♣

© А.Е.Авраменко

Пущинская радиоастрономическая обсерватория ФИАН
avr@prao.ru

**Аннотация**

Показана тождественность параметрической модели пульсарных данных в инерциальной – неподвижной и произвольно выбранных – движущихся вместе с Землей координатных системах наблюдений. Определена связь параметрической модели пульсарных данных и уравнений физических процессов импульсного излучения пульсаров в координатных системах. Форматы пульсарных данных приведены к единому виду во всех системах отсчета.

По параметрической модели выявлена численная эквивалентность наблюдаемых параметров вращения пульсара в координатных системах, подтверждающая их беспрецедентную стабильность. На ретроспективных данных показаны новые, перспективные возможности пульсаров в изучении фундаментальных свойств материи и решении актуальных прикладных задач.

## 1 Введение

Наблюдаемая устойчивая периодичность импульсного излучения пульсаров, практически неисчерпаемая энергетика определили уникальную роль пульсаров в изучении фундаментальных закономерностей метрики пространства-времени, включая непосредственные прецизионные измерения времени, а в перспективе – возможность создания внеземных эталонов времени, которые не уступают в точности и стабильности традиционным атомным часам и обладают абсолютной надежностью в пределах любой протяженности наблюдений.

В общем виде, измерение пульсарного времени основано на динамической модели движения небесных тел Солнечной Системы [7]. Модель определяет эфемериды Луны, планет Солнечной Системы, дающие их пространственные координаты как функции независимой переменной – времени, представляющего собой в этой модели координатное время. Поскольку координатное время одного и того же наблюдаемого пульсарного события, зависит от расположения телескопа и в разных пунктах наблюдения оно различно, полученные топоцентрические моменты пересчитываются к барицентру Солнечной системы – неподвижной точке, совпадающей с центром масс [4].

Для топоцентрических и барицентрических моментов наблюдаемых пульсарных событий принят формат хранения и обмена данных Принстонского университета, в соответствии с которым сформирован архив наблюдательных пульсарных данных обсерватории Аресибо [9], и де-факто он стал общепринятым международным форматом. Формат включает модифицированную юлианскую дату MJD наблюдаемого пульсарного события (целая часть – сутки и дробная часть суток от их начала в относительном исчислении), остаточные уклонения – отклонения моментов наблюдаемых импульсов от расчетных значений (resid), статистические оценки отклонений СКО (err) и ряд других характеристик, относящихся к радиотелескопу, системе приема и регистрации радиосигналов пульсара, а также собственные параметры пульсара из каталога.

В инерциальной барицентрической системе расчетные моменты импульсов определяются параметрами вращения пульсара – периодом и его производной Для наблюдаемых моментов используются оценки по остаточным уклонениям и их статистические характеристики – среднеквадратичные отклонения (СКО) [9]. Точность таких оценок оказывается невысокой, и соответствующая им расчетная стабильность пульсарного времени на несколько порядков уступает достижимой стабильности атомных эталонов.

Было показано, что принципиальное, на несколько порядков, улучшение стабильности барицентрического пульсарного времени достигается в результате распространения параметрической расчетной модели также на наблюдательные пульсарные данные [1]. С этой целью интервалы выборочно наблюдаемых пульсарных событий преобразуются с помощью линейного приближения по критерию наименьших квадратов к последовательности интервалов, детерминированных наблюдаемыми параметрами

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

вращения пульсара. Соотношение наблюдаемых интервалов и их приближений выражается:

$$(1+\alpha_i)(1+0{,}5\dot{P}N)P_o^* N = \sum_i dTB_i - R_i \quad (1.1)$$

где $\sum_i dTB_i -$ наблюдаемые в барицентрической системе интервалы от начального до *i*-го пульсарного события, $R_i$ - непараметризуемые вариации наблюдаемых интервалов, $P_o^*$ - начальный период по существующим оценкам на текущую эпоху, $(1+\alpha_i)$ – линейный коэффициент преобразования.

По величине коэффициента $(1+\alpha_i)$ уточняется значение наблюдаемого периода и определяются его вариации в пределах промежутка наблюдений:

$$P_0 = (1+\alpha_i)P_o^* \quad (1.2)$$

В результате соотношение (1.1) приводится к виду:

$$(P_0 N + 0{,}5P_0\dot{P}N^2)_i = \sum_i dTB_i - R_i \quad (1.3)$$

По интервалам пульсарного времени совместно наблюдаемых пульсаров B1937+21 и B1855+09 были получены синхронизирующие поправки атомных часов, с погрешностью, которая не превышает 5-10 нс в 4-летнем промежутке наблюдений [2].

Настоящая работа посвящена интеграции наблюдательных пульсарных данных и их моделей в независимых координатных системах, созданию единой параметрической модели данных для любой произвольно выбранной координатной системы, распространение метрики барицентрического пульсарного времени на универсальную метрику пространства-времени для произвольно выбранных координатных систем отсчета.

## 2 Особенности координатной метрики пульсарного времени

Время, определяемое с помощью эфемерид, относится к выбранной координатной системе отсчета. Международный астрономический союз в 1991 году определил условия реализации в Солнечной системе невращающейся барицентрической системы координат. Пространственные координаты этой системы выбираются таким образом, чтобы в этой системе, центр которой совмещен с барицентром системы масс, собственное время минимально отличалось от рассчитанного в соответствии с метрикой общей теории относительности, с учетом ньютоновского гравитационного потенциала для рассматриваемых систем масс. Кроме барицентрической системы, оси которой располагаются в центре масс Солнечной системы, используются также локальная геоцентрическая система, оси которой располагаются в центре масс Земли. Измерения на радиотелескопе выполняются в координатной системе, вращающейсяся вместе с Землей. В качестве барицентрической системы по международному соглашению используется *международная небесная опорная система координат* (ICRF), которая содержит список прямых восхождений и склонений около 600 квазаров, опубликованных Международной Службой Вращения Земли (IERS). В качестве геоцентрической системы принята *международная земная опорная система координат* (ITRF). Она включает в себя список координат для фиксированной опорной даты и скоростей примерно для 200 пунктов на Земле. ITRF также устанавливается службой IERS [7].

Переход к параметрической модели наблюдаемых интервалов пульсарных событий позволил на несколько порядков снизить уровень случайных вариаций барицентрического пульсарного времени [1]. Выявленные моделью характеристики пульсаров подтверждают возможность воспроизведения высокостабильного периодического процесса, который может быть сопоставлен с традиционным атомным эталоном. Задача заключается в том, чтобы на основе принципа эквивалентности локальных и динамических измерений времени в произвольно выбранной системе пространственно-временных координат, трансформировать удаленный динамический процесс воспроизведения пульсарного времени из инерциальной барицентрической системы в любую другую, совместить его с локальным процессом измерения с использованием физического эталона в выбранной системе, и определить достижимые метрические характеристики пульсарного времени в этой системе.

## 3 Инвариантность координатного пульсарного времени

### 3.1 Параметрические условия инвариантности

Используемый здесь подход основан на неизменности (форминвариантности) уравнений физических процессов в координатных системах отсчета, что имеет следствием неизменность наблюдаемых в них интервалов в 4-мерном пространстве событий, определяемом псевдоевклидовой геометрией пространства-времени, или пространством Минковского. Это свойство метрики пространства-времени, обычно принято относить только к инерциальным системам, однако в более поздних работах [5,6] было показано, что свойство псевдоевклидовой геометрии пространства-времени распространяется и на произвольные, неинерциальные системы отсчета. Выражение

$$J = c^2T^2 - X^2 - Y^2 - Z^2 \quad (3.1)$$

определяет, что в любой системе отсчета, инерциальной или произвольно выбранной, в декартовых координатах заданная абсолютная величина $J$ остается неизменной, тогда как, в зависимости от выбора системы отсчета, проекции $X, Y, Z, T$ являются относительными величинами (здесь $c$ – скорость света). Выражение (3.1) в дифференциальной форме принимает вид:

$$(d\sigma)^2 = c^2(dT)^2 - (dX)^2 - (dY)^2 - (dZ)^2 \quad (3.2)$$

Прямым следствием псевдоевклидовой геометрии четырехмерного пространства событий является одновременность, т.е. единое время для всех точек трехмерного пространства данной координатной системы. В другой координатной системе единое время будет другим. Это означает, что трехмерное пространство любой координатной системы отсчета ортогонально линиям времени, и в совокупности они представляют собой четырехмерное пространство с мерой (3.1) и (3.2).

Необходимо отметить, что уравнение вида (1.1), относящееся к барицентрическим интервалам пульсарного времени, применительно к топоцентрической системе отсчета не определено. Считается, что параметрическая зависимость моментов импульсов, выражаемая через период вращения пульсара и его производную, применима только к барицентрической системе отсчета. Поэтому, для приведения дискретного континуума интервалов пульсарных событий, наблюдаемых в топоцентрической координатной системе телескопа, к виду (1.1), требуется показать правомерность расширения области применения этого уравнения, включив в неё, наряду с инерциальной – барицентрической, также и произвольно выбранные – топоцентрические координатные системы отсчета.

### 3.2 Пульсарное время в координатных системах

При решении этой задачи – приведения к инвариантному виду уравнений интервалов пульсарных событий, наблюдаемых в произвольных координатных системах, будем следовать принципу неизменности уравнений физических процессов во всех системах отсчета, сформулированному А.А.Логуновым в [5,6]. Неизменность (форминвариантность) физических уравнений во всех инерциальных системах отсчета означает, что **физические процессы**, протекающие в этих системах при одинаковых условиях, **тождественны**. Именно поэтому все **естественные эталоны** во всех системах отсчета **одинаковы**. Преобразования координат, которые оставляют метрику (3.1) и (3.2) форминвариантной, приводят к тому, что физические явления, происходящие в таких системах координат при одинаковых условиях, никогда не могут позволить отличить одну систему координат от другой. Отсюда можно дать более общую формулировку **принципа относительности**, которая относится уже не только к **инерциальным системам** координат, но и к **неинерциальным** (выделено А.А.Логуновым, [6]):

*«Какую бы физическую систему отсчета мы ни избрали (инерциальную или неинерциальную), всегда можно указать бесконечную совокупность других систем отсчета, таких, в которых все физические явления протекают одинаково с исходной системой отсчета, так что мы не имеем и не можем иметь никаких экспериментальных возможностей различить, в какой именно системе отсчета из этой бесконечной совокупности мы находимся»*

Из этого факта вытекает несколько основополагающих следствий применительно к наблюдению излучаемых периодических импульсов пульсара:

- инвариантность, т.е. независимость и, следовательно, постоянство *единичного интервала времени*, воспроизводимого измерительным атомным эталоном радиотелескопа, для всех систем отсчета: все естественные эталоны во всех системах отсчета одинаковы [6];
- инвариантность во всех системах отсчета наблюдаемого периода вращения пульсара, отсчитываемого в единицах шкалы измерительного атомного эталона;
- инвариантность во всех системах отсчета наблюдаемых вариаций периода вращения пульсара и обусловленных ими отклонений интервалов пульсарного времени в пределах промежутков, ограниченных одними и теми же наблюдаемыми событиями.

Тогда уравнения для интервалов пульсарного времени $TB_i$ в инерциальной барицентрической системе и интервалов $TT_i$ в топоцентрической системе выглядят так:

$$TB_i = (1+\alpha_i)(P_0 N_B + 0{,}5 P_0 \dot{P} N_B^2)_i \quad (3.3)$$

$$TT_i = (1+\alpha_i)(P_0 N_T + 0{,}5 P_0 \dot{P} N_T^2)_i \quad (3.4)$$

где $N_B$, $N_T$ – порядковый номер излученного импульса пульсара в барицентрической (ICRF) и топоцентрической (ITRF) системах, отсчитываемый от начального события, одного и того же, в барицентрической и топоцентрической системах отсчета. Наблюдаемый период определяется выражением (1.2), как в барицентрической, так и топоцентрической системе.

По отклонениям коэффициента линейного приближения (1+ $\alpha_i$), как было показано в [2] применительно к барицентрической системе отсчета, определяются вариации наблюдаемого периода в промежутке, ограниченном двумя произвольно выбранными наблюдаемыми событиями. Если взять два промежутка, отсчитываемых от общего начала и ограниченных соответственно $i$-м и $j$-м излучаемыми импульсами пульсара $(i>j)$, то относительное отклонение наблюдаемого периода вращения пульсара в промежутке между $i$-м и $j$-м событиями

определяется разностью коэффициентов линейного приближения между ними:

$$\Delta P_i / P_0 = (\alpha_i - \alpha_j) \qquad (3.5)$$

Соотношение (3.5) выполняется также и в топоцентрической системе отсчета.

Отклонение интервалов пульсарного времени вследствие непостоянства наблюдаемого периода вращения пульсара в промежутке между $i$-м и $j$-м событиями совпадает для барицентрической и топоцентрической систем отсчета и определяется выражением:

$$\Delta TB_i = \Delta TT_i = P_0(\alpha_i - \alpha_j)(N_i - N_j), \quad (3.6)$$

где $(N_i - N_j)$ – общее число пульсарных событий, излученных в выделенном промежутке.

Таким образом, благодаря распространению модели интервалов пульсарного времени, установленной для инерциальной барицентрической системы, также и на движущиеся топоцентрические системы, интервалы пульсарных событий, наблюдаемые в произвольных координатных системах, отображаются одними и теми же уравнениями (то есть форминвариантно). Интервалы в этих уравнениях определяются одними и теми же параметрами вращения пульсара – наблюдаемым периодом и его производной. Существенно, что и численные значения наблюдаемых параметров вращения пульсара совпадают, какую бы координатную систему мы ни выбрали. То же относится и к наблюдаемым вариациям периода вращения и отклонения интервалов пульсарного времени: их величины однозначно определяются коэффициентами линейного приближения $(1+\alpha_i)$, также инвариантными для барицентрической и топоцентрической систем отсчета в пределах промежутка наблюдений. Отсюда следует важный вывод: в любых произвольно выбранных координатных системах уравнения наблюдаемых интервалов пульсарного времени одинаковы, а величина наблюдаемых интервалов определяется одними и теми же значениями параметров вращения пульсара, не зависящими от выбора координатных систем. Фундаментальное свойство инвариантности уравнений и наблюдаемых величин, выражаемых через наблюдаемые параметры, позволяют перейти от специфических для каждой системы моделей и форматов данных к единому их виду для любых координатных систем.

## 4 Наблюдательные данные в координатных системах

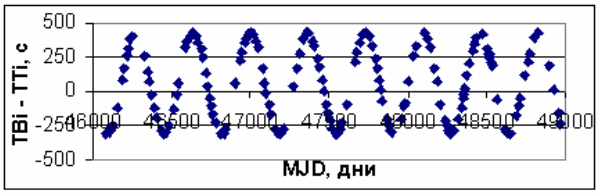
### 4.1 Инвариантность моделей пульсарных данных в координатных системах

Распространение уравнений барицентрических интервалов на произвольно выбранные неинерциальные системы позволяет перейти к единой модели наблюдательных пульсарных данных для всех координатных систем. Существующее различие и несовместимость форматов барицентрических и топоцентрических данных Принстонского университета, отражающие специфику принятых моделей наблюдательных данных в топоцентрической и барицентрической системах, ограничивает функциональные возможности и надежность отождествления свойств пульсаров по результатам наблюдений. Отношения наблюдаемых величин, которые в координатных системах определяются численными методами по уравнениям эфемерид Земли и планет Солнечной системы, не поддаются непосредственному сопоставлению и сравнительным оценкам, даже если исходные данные представлены в виде интервалов, наблюдаемых в этих системах, а тем более в виде остаточных уклонений, которые определяются только в барицентрической системе, тогда как в топоцентрической системе они вообще отсутствуют.

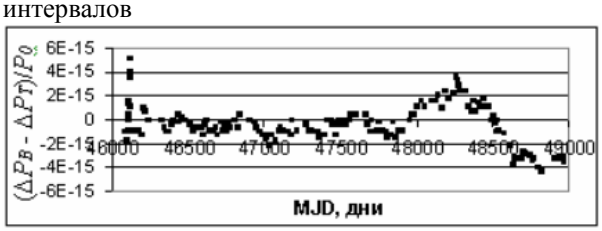
Тождественность уравнений интервалов в координатных системах проиллюстрируем на примере наблюдений миллисекундного пульсара B1937+21. На Рис.4.1(а) приведены результаты сопоставления вариаций периода вращения и отклонений интервалов пульсарного времени в барицентрической и топоцентрической системах отсчета по данным наблюдений этого пульсара в Аресибо [9]. На графике Рис.4.1(а) показана разность интервалов для одних и тех же событий, наблюдаемых в барицентрической и топоцентрической системах отсчета. Интервалы отсчитываются от начального пульсарного события на дату MJD 46053 (19.12.1984г.) в промежутке около 8 лет, ограниченном датой MJD 48970 (14.12.1992г.). Как следует из графика, разность циклически, с периодом обращения Земли вокруг Солнца, изменяется по абсолютной величине в пределах около ±500 с, отображая различие координатного времени в двух рассматриваемых системах.

На графике Рис.4.1(б) приведены разностные вариации периода вращения пульсара, вычисленные по параметрическим приближениям интервалов в барицентрической и топоцентрической системах отсчета. Численно они представляют собой выраженную в относительных единицах разность коэффициентов линейного приближения наблюдаемых интервалов в соответствии с (3.5) в барицентрической и топоцентрической системах. На Рис.4.1(в) показаны отклонения интервалов пульсарного времени, которые определены по наблюдаемым параметрам вращения пульсара в барицентрической (по значениям $TB_i$) и топоцентрической (по значениям $TT_i$) системах отсчета, в соответствии с выражениями (3.3) и (3.4), и их средняя величина. На графике Рис.4.1(г) приведена разность отклонений интервалов

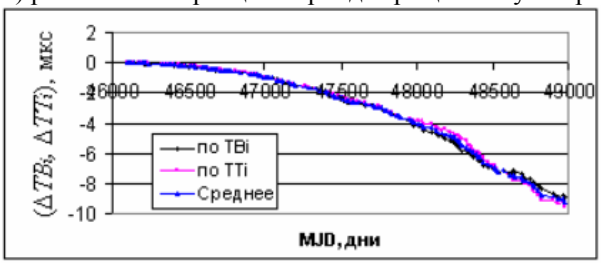
пульсарного времени в барицентрической и топоцентрической системах отсчета.




а) разность барицентрических и топоцентрических интервалов




б) разностные вариации периода вращения пульсара




в) отклонения интервалов пульсарного времени по барицентрическим и топоцентрическим интервалам


г) разность отклонений интервалов координатного пульсарного времени

Рис.4.1. Вариации наблюдаемого периода вращения и интервалов пульсарного времени пульсара B1937+21 в координатных системах отсчета.

Приведенное на Рис.4.1 сопоставление результатов наблюдений экспериментально подтверждают правомерность перехода к единой модели пульсарных данных в координатных системах и тождественность полученных с помощью этой модели характеристик координатного пульсарного времени в топоцентрической и барицентрической системах. Действительно, разностные вариации наблюдаемого в обеих системах периода вращения пульсара (Рис.4.1б) не выходят за пределы статистической погрешности измерений [1], а разность отклонений интервалов координатного пульсарного времени (Рис.4.1г) более чем на порядок меньше абсолютных величин отклонений в каждой системе. Это означает, что сравниваемые величины вариаций наблюдаемого периода и отклонения интервалов пульсарного времени по наблюдениям в обеих координатных системах практически совпадают. Отличия, выраженные разностными вариациями периода и отклонениями интервалов, наблюдаемых в барицентрической и топоцентрической системах отсчета определяются достижимой погрешностью измерительного атомного эталона [2].

Таким образом, по наблюдательным данным экспериментально подтверждена правомерность распространения уравнений барицентрических интервалов пульсарного времени на произвольно выбранные координатные системы отсчета. Благодаря инвариантности уравнений, достигается возможность перехода к единой модели наблюдательных пульсарных данных в любых координатных системах. На основе универсальной параметрической модели данных реализуется надежное сопоставление наблюдаемых величин в различных координатных системах, отождествление по ним физических свойств пульсаров.

### 4.2 Новые возможности пульсаров в метрике пространства-времени

Обобщая рассмотренные здесь результаты, можно отметить, что преобразование моментов наблюдаемых пульсарных событий в интервалы, отсчитываемые от выбранного начального события; распространение барицентрической модели интервалов, определяемых наблюдаемыми параметрами вращения пульсара, также и на неинерциальную топоцентрическую систему; представление интервалов выборочно наблюдаемых пульсарных событий в виде регуляризованного периодического процесса, детерминированного наблюдаемыми параметрами вращения пульсара – всё это вместе позволило реализовать принципиально новый подход к выявлению роли и возможностей пульсаров в решении фундаментальных проблем метрики пространства-времени. Рассматриваемый подход основан на неизменности физических уравнений в любых, произвольно выбранных координатных системах отсчета. Получены экспериментальные подтверждения правомерности обобщения параметрической модели наблюдаемых барицентрических интервалов пульсарного времени на топоцентрическую систему отсчета. Инвариантные свойства интервалов пульсарного времени в координатных системах открывают новые возможности построения шкал пульсарного времени в движущихся системах отсчета с точностью, не уступающей сопоставляемым с ними измерительным атомным эталонам.

Уникальность физического процесса излучения энергии пульсара состоит в том, что, во-первых, одни и те же события, связанные с излучением пульсара, можно наблюдать в любой точке пространства, ограниченного, например, только пределами Солнечной системы. Во-вторых, излучение энергии пульсара имеет выраженный периодический характер, процесс излучения локализуется в любой произвольно выбранной

координатной системе пространства в виде регулярно повторяющихся интервалов. И, наконец, в-третьих, высокостабильная повторяемость импульсов излучения пульсара позволяет производить сличение интервалов пульсарного времени с физическим эталоном атомного времени с целью их синхронизации. Тем самым в любой выбранной координатной системе достигается совмещение локального времени атомных часов и координатного пульсарного времени, основанного на динамической модели движения небесных тел и наблюдениях удаленного физического явления – излучения периодических импульсов пульсара.

Инвариантность уравнений физических процессов излучения импульсов и совпадающие характеристики вариаций параметров вращения пульсара, наблюдаемые в произвольно выбранных координатных системах, определяют эквивалентность применяемых методов и получаемых результатов отождествления свойств пульсаров в этих системах. Так, полученные по результатам совместных наблюдений пульсаров B1937+21 и B1855+09 совпадающие профили изменений наблюдаемых величин периода вращения пульсаров и отклонений интервалов пульсарного времени в барицентрической системе [2] были экспериментально подтверждены аналогичными характеристиками периода вращения и интервалов пульсарного времени, наблюдаемых в движущейся топоцентрической координатной системе. Тем самым, независимо по наблюдениям в любой выбранной координатной системе, реализуются шкалы координатного времени на основе физического эталона, который синхронизирован по пульсарному времени, локализованному в выбранной координатной системе.

## 5 Об инвариантности форматов данных в координатных системах

В основу форматов наблюдательных пульсарных данных положены форматы обмена данными Принстонского университета [3]. В топоцентрической координатной системе исходными являются моменты пульсарных событий, наблюдаемых на указанную модифицированную юлианскую дату MJD. Дробная часть даты (13 десятичных разрядов) представляет собой относительную безразмерную величину, численно равную части суток, отсчитываемой от их начала на дату наблюдения (целая часть). В барицентрической системе координат моменты пульсарных событий представляются в такой же форме, с той лишь разницей, что дробная часть, сильно редуцированная, содержит не 13, а всего 4 десятичных разряда. В формат барицентрических данных, кроме моментов наблюдаемых событий, включены остаточных уклонений (ОУ) – отклонения моментов импульсов от расчетных значений в абсолютном исчислении, а также статистические оценки моментов наблюдаемых событий – среднеквадратичные отклонения, которые приняты одинаковыми для топоцентрической и барицентрической координатных систем.

Принстонский формат недостаточен для преобразования наблюдательных данных к параметрическому виду расчетной модели, так как он не содержит компонентов, отражающих связь наблюдаемых величин с параметрами вращения пульсара и не содержит в явном виде наблюдаемых интервалов, отсчитываемых от выбранного начального события.

Была проведена необходимая модификация формата. Для преобразования топоцентрических данных в интервалы пульсарного времени дробная часть даты пересчитывается из относительной безразмерной величины в измеряемое время, выраженное в его единицах. По этим данным формируется последовательность наблюдаемых интервалов, а по ним – суммарные интервалы каждого наблюдаемого события относительно начального события. Исходные наблюдательные данные, кроме моментов наблюдаемых событий в пределах суток на дату наблюдения, в каждой координатной системе дополняются значениями интервалов между соседними наблюдаемыми событиями, а также величинами суммарных интервалов, отсчитываемых от начального до любого другого события, выбранного в пределах промежутка наблюдений [3]. Интервалы пульсарного времени, полученные по выборочным наблюдениям, трансформируются в регуляризованную последовательность интервалов, которые определяются наблюдаемыми параметрами вращения пульсара. Тем самым, трансформированный и расширенный формат пульсарных данных, наряду с исходными моментами наблюдаемых событий, включает параметры вращения пульсара и коэффициенты линейного приближения наблюдаемых интервалов, входящие в уравнения (3.3) и (3.4), которые определяет топоцентрические и барицентрические интервалы пульсарного времени в пределах рассматриваемого промежутка наблюдений.

Таким образом, в результате выполненных преобразований наблюдательные данные в барицентрической и топоцентрической системах приведены к одному и тому же формату, который по структуре параметров отвечает требованиям уравнений (3.3) и (3.4), описывающих физический процесс излучения импульсов пульсара в одинаковом виде в любой выбранной координатной системе. Тем самым реализованы качественно новые возможности пульсаров в изучении фундаментальных физических основ высокоточной метрики 4-мерного пространства-времени и решении актуальных прикладных задач сопоставления пульсарного времени с физическими атомными эталонами с целью формирования высокостабильных координатных шкал времени.

## Заключение

В работе приведены результаты обобщения параметрической модели пульсарных данных для инерциальной барицентрической системы на произвольно выбранные координатные системы, в том числе связанные с движущейся Землей. Были определены требования к представлению наблюдательных данных, которые соответствуют условиям инвариантности уравнений излучения импульсов пульсаров в любой выбранной координатной системе.

Получен общий вид уравнений для интервалов импульсов, наблюдаемых в инерциальной и движущихся координатных системах. Благодаря совпадению значений параметров вращения пульсара, наблюдаемых в любой системе, достигается метрическая совместимость и прямое сопоставление интервалов физического атомного эталона и интервалов пульсарного времени, локализованных в любой координатной системе.

По результатам наблюдений получено экспериментальное подтверждение правомерности обобщения параметрической модели наблюдаемых барицентрических интервалов пульсарного времени на топоцентрическую систему отсчета. На основе инвариантных уравнений и параметрической модели интервалов пульсарного времени реализован единый формат наблюдательных пульсарных данных в координатных системах, который отвечает требованиям надежного выявления присущих пульсарам свойств, определяющих уникальную стабильность наблюдаемых интервалов, с точностью, не достижимой традиционными методами. Применение рассматриваемых здесь методов к полученным ранее архивным наблюдательным данным позволили на несколько порядков улучшить точность отождествляемых характеристик присущей пульсарам долговременной стабильности. Погрешность отождествления ограничивается по существу только предельно достижимыми характеристиками измерительных эталонов.

Таким образом, преимущества параметрической модели пульсарных данных по отношению к традиционной разностной модели, основанной на остаточных уклонениях, проявляются не только в барицентрической системе, но и, в еще большей степени, при переходе к 4-мерной метрике псевдоевклидовой геометрии пространства-времени. Переход к единой параметрической модели наблюдательных пульсарных данных для любых, произвольно выбранных координатных систем, распространение метрики барицентрического пульсарного времени на универсальную метрику пространства-времени раскрывают исключительно важную роль пульсаров в изучении фундаментальных свойств материи и определяют перспективы создания единой системы эталонов времени на основе физических процессов разной природы.

## Литература


[1] А.Е.Авраменко. Параметрический синтез пульсарного времени. //Измерительная техника, 2006, № 6, 39-44.

[2] А.Е.Авраменко. Параметрическая стабильность пульсарного времени. //Измерительная техника, 2008, № 7, 32-37.

[3] А.Е.Авраменко. К согласованному виртуальному и реальному времени в коллекции астрометрических пульсарных данных. //Труды RCDL 2007. Переславль Залесский, 2007, 103-111.

[4] О.В.Дорошенко. Комплекс программ для фазовых наблюдений пульсаров. Препринт ФИАН. Москва, 1993, № 51.

[5] Логунов А.А. Лекции по теории относительности. М.: Наука. 2002.

[6] Логунов А.А. Анри Пуанкаре и теория относительности. М.: Наука. 2004.

[7] К.Одуан, Б.Гино. Измерение времени. Основы GPS. Пер. с англ. под ред. В.М.Татаренкова. М., ТЕХНОСФЕРА, 2002.

[8] Guinot B. and Petit J. Atomic Time and the Rotation of Pulsars. //Astron.Astrophys., 1991, 248, 292-296.

[9] V.M.Kaspi, J.H.Taylor, and M.F.Ryba. High-precision Timing of Millisecond Pulsars. III. Long-term Monitoring of PSRs B1885+09 and B1937+21. //The Astrophysical Journal, 1994, 428, 713-728.

[10] C.A.Murray. Vectorial Astrometry. Adam Hilger, Bristol, 1983.


## Toward the Invariant Models of Pulsar Data in Spatial-Time Coordinate Systems


A.E.Avramenko



The coincidence of the parametric model of the observed pulsar data in both, inertial barycentric coordinate system, or arbitrary chosen topocentric ones, is shown. The relationship of the parametric pulsar data model and the equations of physical processes of pulsed radiating of the pulsars in coordinate systems, are determined. The formats of the observed pulsar data are modified and transformed into unific, system independent type.

On the parametric model, the numerical equivalence of the observed parameters of rotation of pulsar in any coordinate systems, which confirms unprecedented stability of pulsars, is detected. On the retrospective archive of the observed pulsar data, the new pulsar applications in study of fundamental properties of matter and decision of actual applied problems, are considered.



♣ Работа выполнена при поддержке гранта РФФИ № 06-07-89043

# Новая электронная карта основных параметров гигантского дипольного резонанса атомных ядер♣

© В. В. Варламов, В .В. Вязовский, И. А. Ехлаков, С. Ю. Комаров, Н. Н. Песков,

О. В. Семенов, М. Е. Степанов

Центр данных фотоядерных экспериментов
Научно-исследовательского института ядерной физики имени Д. В. Скобельцына
Московского государственного университета имени М. В. Ломоносова
Varlamov@depni.sinp.msu.ru

**Аннотация**

Описывается новый информационный Интернет-ресурс – реляционная база данных - электронная Карта основных параметров гигантских дипольных резонансов (ГДР) атомных ядер – входящий в систему реляционных баз данных (БД) по физике атомных ядер и ядерных реакций Центра данных фотоядерных экспериментов (ЦДФЭ) НИИЯФ МГУ.
Карта включает большое количество данных об энергетическом положении, амплитуде, ширине, интегральных характеристиках ГДР, широко востребованных в различных областях фундаментальных и прикладных исследований, а также – в разнообразных приложениях.

## 1 Введение

История научного знания представляет собой, по существу, единство двух тенденций: с одной стороны это - получение все новых и новых эмпирических фактов, а с другой - сжатие, концентрация имеющихся массивов данных, сведение их к наименьшим объемам. Наивысшей формой такого представления можно считать аналитические формулы, описывающие многообразие результатов определенного типа. Так, например, Ньютон в одну строчку уравнения закона всемирного тяготения “поместил” громадное количество результатов эмпирических наблюдений, относящихся к взаимному движению масс. По существу это уравнение описывает супер-базу (гипер-базу) данных (БД) - в ответ на определенным образом сформулированный запрос могут быть получены определенные данные об окружающем мире.

Следует обратить особое внимание на то, что такое концентрированное представление информации обладает предсказательными свойствами: содержит в себе не только известные уже факты, но и те, которые станут известными лишь тогда, когда соответствующий запрос будет сделан.

Природа устроена так, что далеко не все ее явления и проявления поддаются, по крайней мере, сразу, аналитическому описанию, и наилучшей формой представления накопленного знания становится именно БД - упорядоченное и концентрированное представление фактов. Безусловно, на такую роль может претендовать лишь достаточно полная, представительная, репрезентативная БД.

Ярким примером такой БД является широко известная и неспециалистам Периодическая таблица химических элементов Д. И. Менделеева. По существу она является ни чем иным, как базой данных химических элементов, последовательно отсортированных сначала по атрибуту - “валентность”, а затем - “заряд”. Эта таблица, как и всякая другая подобная - по существу не что иное, как “склад готовой продукции” - представляет собой эффективный инструмент научных исследований, поскольку обладает мощной предсказательной силой: много новых свойств элементов и самих элементов было открыто на основе простой комбинаторики свойств элементов-соседей “слева - справа” и “сверху - снизу”.

## 2 Деятельность ЦДФЭ по обработке (фото)ядерных данных

### 2.1 Сеть Центров ядерных данных МАГАТЭ

Создание таких инструментов имеет особое значение для исследований в области ядерной физики, в которых количество данных, получаемых и применяемых в современных ядерно-физических экспериментах и востребованных современными технологиями, огромно и с течением времени

**Труды 11[й] Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции” - RCDL’2009, Петрозаводск, Россия, 2009.**

только возрастает. В силу ряда некоторых исторических причин создание таких ядерно-физических БД началось очень давно и носило планомерный, долговременный и глобальный характер. Этому способствовал тот факт, что оно проводилось под эгидой Международного агентства по атомной энергии (МАГАТЭ). Результатом явилось то, что ядерная физика к тому моменту, когда во всем мире начался информационный бум, оказалась едва ли не самой передовой и подготовленной в этом отношении областью знаний.

В настоящее время научному сообществу доступно значительное количество ресурсов, содержащих самые разнообразные данные по атомным ядрам и процессам их превращения друг в друга в ядерных реакциях и радиоактивных распадах. Для поддержания этих ресурсов созданы сети специализированных Центров ядерных данных. Основная задача участников сети - создание репрезентативных БД по свойствам атомных ядер и характеристикам ядерных реакций. Для решения этой благородной, но неблагодарной задачи Центры согласованно:

- организуют поиск и компиляцию данных;
- обеспечивают форматирование данных в согласованных форматах;
- проводят экспертизу точности и надежности данных;
- осуществляют согласование результатов различных экспериментов;
- создают системы доступа к данным (банки и базы данных, Интернет-интерфейсы);
- анализируют и оценивают данные;
- готовят, издают и распространяют аналитические обзоры, указатели, атласы и т.д. и т.п.

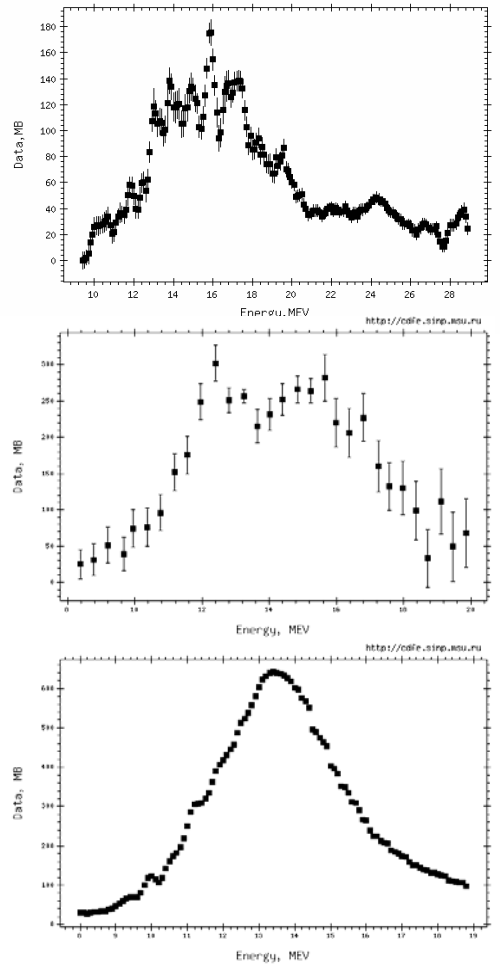
Участником одной из таких сетей Центров ядерных данных МАГАТЭ, специализирующейся на информации по ядерным реакциям, является Центр данных фотоядерных экспериментов (ЦДФЭ), ответственный за данные по ядерным реакциям под действием γ–квантов низких и средних энергий. В ЦДФЭ НИИЯФ МГУ в рамках программы обеспечения российских пользователей к точной и надежной ядерно-физической информацией в течение ряда лет созданы несколько полных реляционных баз ядерных данных [1].

Эти БД широко используются для информационного обеспечения фундаментальных и прикладных ядерно-физических исследований и для проведения новых научных исследований [2]. Мощные и гибкие поисковые системы созданных баз данных предоставляют пользователям широкие и эффективные возможности работы с разнообразной ядерно-физической информацией. При этом могут использоваться либо каждая БД в отдельности, либо объединяющая их фонды и поисковые возможности "Универсальная электронная системы информации по атомным ядрам и ядерным реакциям" (http://cdfe.sinp.msu.ru/services/unifsys/index.html) [3].

Источником информации для указанных баз данных (БД) служат большие фонды числовых данных, создаваемые и поддерживаемые международной сетью Центров ядерных данных МАГАТЭ [4], а также и другие достаточно большие БД, основанные на фондах, подготовленных в ЦДФЭ.

### 2.2 Данные о гигантских дипольных резонансах атомных ядер

Наиболее полной и оригинальной из них является БД "Параметры гигантского дипольного резонанса, сечения фотоядерных реакций" (http://cdfe.sinp.msu.ru/services/gdrsearch.html). Она содержит огромное количество данных об основных параметрах гигантских дипольных резонансов




(ГДР) ядер.

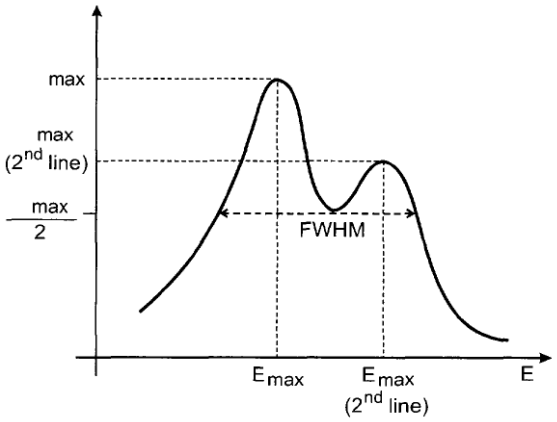
Рис. 1. ГДР в сечениях реакции полного фотопоглощения (γ,abs) – сверху вниз:
- легкое сферическое ядро $^{12}C$;
- средне-тяжелое деформированное ядро $^{165}Ho$;
- тяжелое сферическое ядро $^{208}Pb$.

Гигантские резонансы – мощные и отчетливо выраженные максимумы (в деформированных ядрах

– двойные) проявляются в сечениях всех фотоядерных реакций на практических всех (за исключением легчайших) атомных ядрах и определяют все особенности исследования и использования фотоядерных реакций.

На Рис. 1 приведено несколько типичных примеров ГДР, проявляющихся в сечениях реакции полного фотопоглощения на различных атомных ядрах.

В схематичном виде сечения фотоядерных реакций – гигантские дипольные резонансы – могут быть представлены в виде, приведенном на Рис. 2 и дающем представление о том, какие именно




параметры ГДР находят отражение в БД (Рис. 3).

Рис. 2. Схематичное изображение типичного гигантского дипольного резонанса.

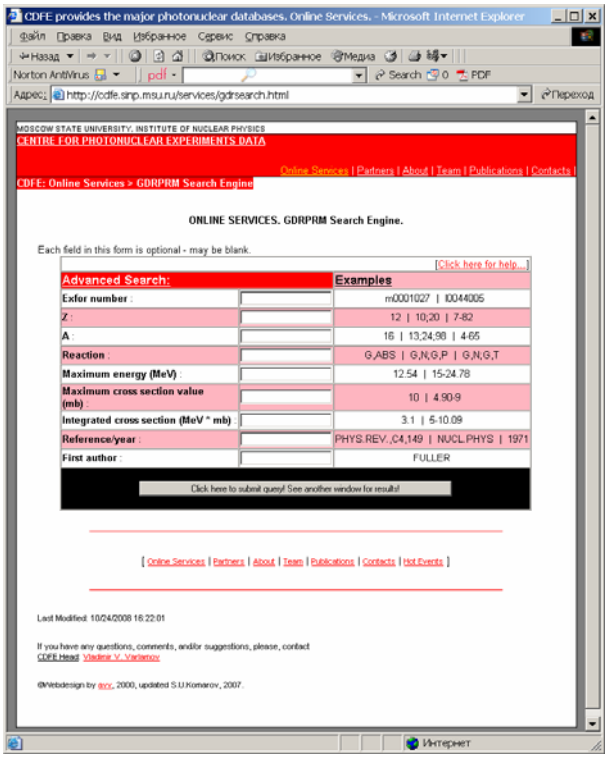



Рис. 3. Web-страница сайта ЦДФЭ НИИЯФ МГУ – поисковая система БД.

Основные параметры (энергия $E_{max}$ резонанса, его амплитуда max и ширина (FWHM), различные интегральные характеристики) ГДР играют важную роль в разнообразных фундаментальных исследованиях электромагнитных взаимодействий ядер, представляют большой интерес с точки зрения изучения структуры и динамики атомных ядер, механизмов ядерных реакций.

Наряду с этим, данные о параметрах ГДР широко используются в разнообразных приложениях в различных областях науки и техники от неразрушающего контроля, гамма-активационного анализа и процессов формирования и исследования наноструктур в реакциях взаимодействия фотонов с материалами до разнообразных медицинских применений.

БД основана, с одной стороны, на фондах международной системы EXFOR, а с другой – на информации из нескольких электронных и печатных коллекций соответствующих данных [5 - 8]. БД содержит информацию из работ, опубликованных в научной литературе в период с 1954 по 2004 годы, в последнее время не обновлялась.

### 2.3 Карта (прототип) данных о форме и размерах ядер

Одной из последних разработок, существенно повышающих эффективность работы с большими массивами данных, на момент своего создания являлась достаточно оригинальная и по существу не имеющая аналогов новая БД - “Карта квадрупольных ядерных деформаций” (http://cdfe.sinp.msu.ru/services/defchart/defmain.html) [9]. Она подготовлена в ЦДФЭ с использованием коллекций данных, либо опубликованных в печатном виде в соответствующих журналах, либо переданных ЦДФЭ в числовом виде авторитетными авторами соответствующих коллекций и содержит полную и систематизированную информацию о таких параметрах, как квадрупольный момент и параметр квадрупольной деформации ядер, с помощью которых описывается их форма.

Впоследствии эта БД была дополнена данными о среднеквадратичных зарядовых радиусах большого количества ядер и преобразована в новую (по существу) БД с названием “Карта данных о форме и размерах атомных ядер” (http://cdfe.sinp.msu.ru/services/radchart/radmain.html) [10]. Она подготовлена в ЦДФЭ с использованием коллекций данных, либо опубликованных в печатном виде в соответствующих журналах, либо переданных ЦДФЭ в числовом виде авторитетными авторами соответствующих коллекций и содержит полную и систематизированную информацию о таких параметрах, как квадрупольный момент и параметр квадрупольной деформации ядер, с помощью которых описывается их форма.

Для представления данных Карты использована удобная графическая форма, в которой традиционно выполняются хорошо известные специалистам карты-плакаты нуклидов, на которых блоки информации, описывающие многообразные спектроскопические свойства отдельных ядер расположены в координатах “число протонов Z” - “число нейтронов N”.

Каждый такой блок информации приводит к данным о квадрупольных моментах Q ядер, параметрах $\beta_2$ их квадрупольной деформации и среднеквадратичных зарядовых радиусах.

С целью облегчения поиска и идентификации данных для искомого ядра специальная расцветка элементов Карты деформаций выполнена по типу той, что используется на традиционных географических картах. Так, коричневым цветом (“горы”) обозначены разделы для ядер с положительной деформацией, синим (“моря”) – с отрицательной, зеленым (“равнины”) – с деформацией неизвестного знака или нулевой. Для идентификации параметров деформации по абсолютной величине использованы пять градаций каждого цвета – наибольшей деформации соответствует более интенсивный цвет. Такой принцип представления обсуждаемой физической информации дает возможность пользователю “на лету” обнаруживать ядра с характерной деформацией – сильно или, напротив, незначительно вытянутые или сплюснутые.

## 3 Новая электронная Карта фотоядерных данных

При создании описанных Карты квадрупольных ядерных деформаций и Карты параметров формы и размеров ядер было разработано новое универсальное программное обеспечение, которое может эффективно использоваться для столь же наглядного представления широкого круга характеристик объектов (и процессов), то есть для БД, содержащих данные различного смыслового содержания. Необходимым условием для такого использования является лишь наличие между соответствующими параметрами объектов и процессов отношений, аналогичных описанным выше - двухпараметрическое расположение на некоторой плоскости и дискретность абсолютных значений.

Очевидно, что содержание представленной выше БД “Параметры гигантского дипольного резонанса, сечения фотоядерных реакций” полностью соответствует сформулированному условию. При этом описанные выше возможности новой Карты по поиску ядер, для которых сечения различных фотоядерных реакций, обладают определенными характеристиками, должны существенно повысить эффективность поиска ядер и процессов их преобразования. Такие ядра и процессы представляют собой кандидатов для проведения огромного количества прикладных исследований, в которых используются фотоядерные реакции на ядрах различных материалов и веществ, сопровождающиеся испусканием различных частиц-продуктов.

### 3.1 Организация доступа к данным

В этой связи в ЦДФЭ начато создание нового электронного ресурса – электронной карты основных параметров гигантского дипольного резонанс ядер, принципы организации и представления информации в которой аналогичны принципам, использованным при создании Карты параметров формы и размеров ядер [9, 10].

Визуальный интерфейс новой Карты представляет собой координатную сетку (Z, N) на которой отображены в виде ячеек все известные на данный момент атомные ядра с соответствующими значениями Z и N (Рис. 4). Координатная сетка с нанесёнными на ней ячейками ядер в дальнейшем именуется картой нуклидов или картой ядер.

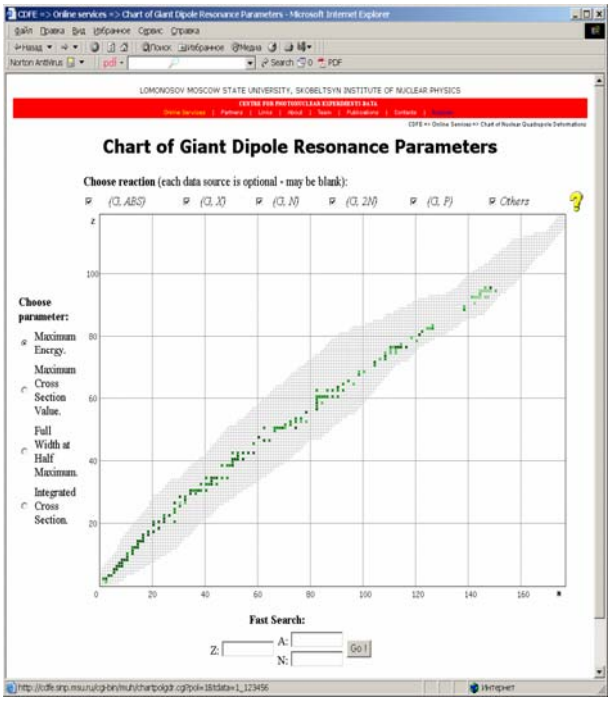



Рис. 4. Входная страница новой электронной карты основных параметров ГДР.

Слева от карты ядер расположена панель выбора одного или нескольких параметров ГДР. Данные параметры выделяют ячейку соответствующего ядра определенным цветом:

- максимальная энергия ГДР – зеленым;
- абсолютная величина (амплитуда) – коричневым;
- ширина (на половине высоты) – красным;
- интегральное сечение – синим.

Интенсивность цвета определяет диапазон количественных значений параметров ГДР. По умолчанию выдаются все доступные данные, однако, может быть выделен определенный канал.

В верхней части страницы Карты представлена панель выбора типа фотоядерных реакций:

- полного фотопоглощения (γ,abs)
- с образованием одного нейтрона (γ,n),
- с образованием двух нейтронов (γ,2n),
- с образованием нескольких нейтронов (γ,xn),
- с образованием протона (γ,p)
- все остальные типы реакций (с образованием дейтрона (γ,d), тритона (γ,t) и α-частицы (γ,α));

Каждая панель представляет собой набор кнопок-переключателей (html-элементов типа "checkbox"), которые позволяют выбрать как одно, так сразу несколько значений.

Элемент "Others" (все остальные типы реакций) сделан в виде выпадающего списка, в котором перечислены все другие реакции, которые не были выделены в элементы панели поиска. Внизу карты нуклидов находится панель поиска по Z, N, A параметров ГДР для конкретного ядра или группы ядер (панель "Fast Search").

Карта имеет 4 уровня масштаба, что облегчает поиск необходимых данных. При выборе самого подробного масштаба нажатие на элемент карты для определенного нуклида открывает окно (Рис. 5), в котором представляются все (соответствующие выбранным каналам и параметрам) характеристики ГДР.

Дополнительно имеется прямая ссылка на исходный документ БД по ядерным реакциям системы EXFOR [11].

Кроме того, приводятся все библиографические сведения об исходных данных. Некоторые из них аналогично тому, как это было сделано в Карте параметров формы и размеров ядер [9, 10], являются ссылками (Рис. 6) на соответствующие документы еще одной полной БД ЦДФЭ по публикациям (http://cdfe.sinp.msu.ru /services/nsr/Search_form.shtml) [12, 13].

### 3.2 Принципы организации программного обеспечения

Программный комплекс новой электронной Карты – реляционной БД "Карта параметров ГДР" является дальнейшей модернизацией информационной системы созданных ранее "Карты квадрупольных деформаций и атомных радиусов" и "Карты параметров формы и размеров ядер". При создании новой электронной карты используются те же средства, с помощью которых создавались и другие БД ЦДФЭ - операционная система Linux, система управления базами данных MySQL, Web-сервер "Apache", технологии обработки запросов CGI. Программное обеспечение новой электронной Карты написано с использованием языков Perl и Javascript, каскадных таблиц стилей CSS. Для создания html-страниц использовался программный пакет HomeSite 5.0.

Центральная часть Карты (координатная сетка) является изображением в формате PNG, которое образуется в зависимости от выбранных параметров на верхней и левой панелях. При этом информация о параметрах ГДР для каждого ядра и типа реакции извлекается из СУБД MySQL. В html-коде главной страницы PNG-картинка разбивается на несколько невидимых пользователю областей, которым соответствуют различные гиперссылки. При нажатии посетителем сайта на какую-либо область картинки курсором мыши происходит переход по соответствующей гиперсылке, привязанной к выбранной области. Данный механизм отображения карты повышает гибкость работы и удобство обновления с массива ядерных данных параметров ГДР.

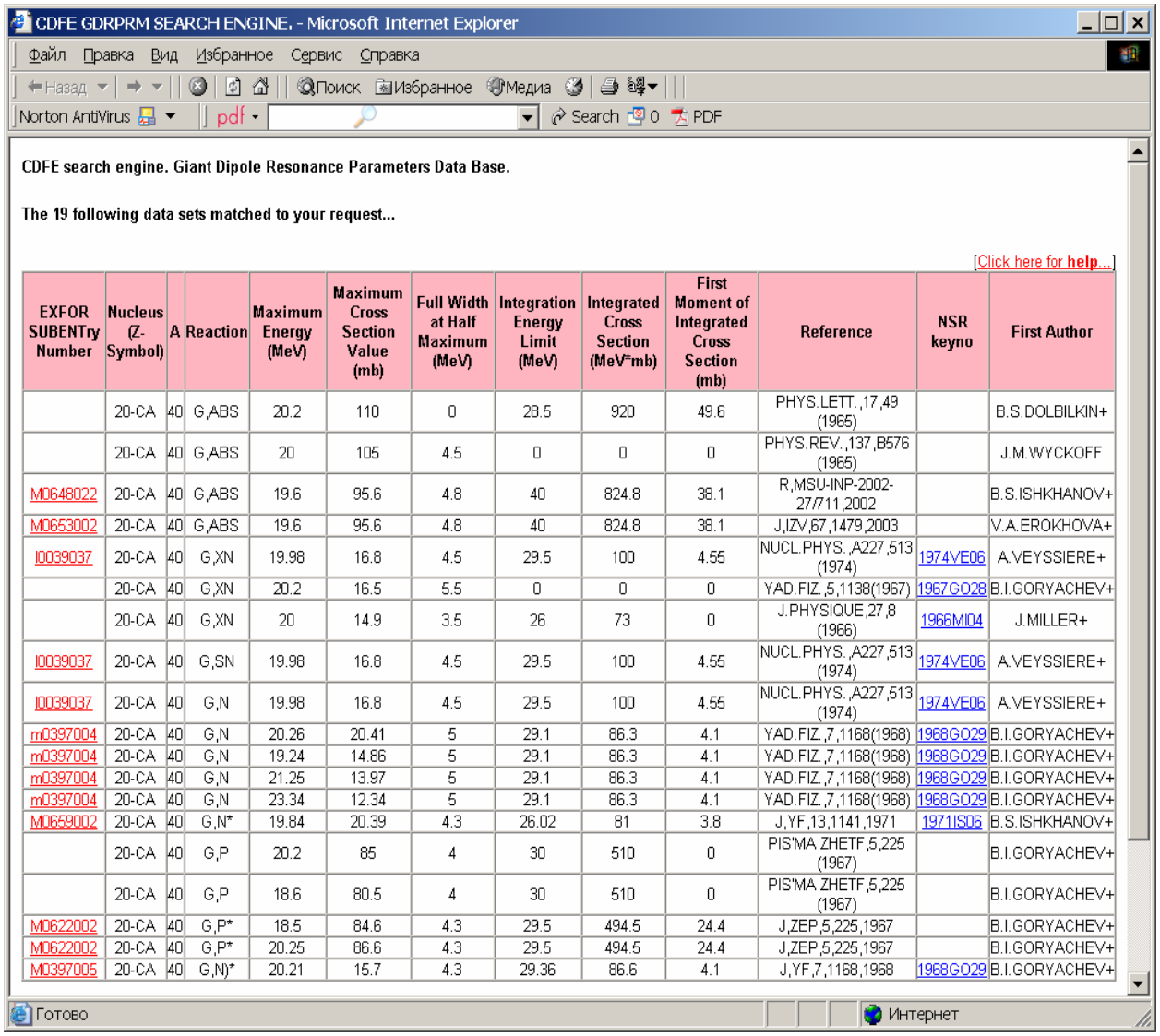
В дополнение к приведенному механизму, для более удобного поиска группы ядер в правом нижнем углу карты расположено вспомогательное окно размером 10x10 ячеек атомных ядер. В результате работы дополнительного js-скрипта, по событию OnMouseMove, в данном окне отображается выделенная пользователем область ядер в увеличенном масштабе. Область ядер является html-таблицой динамически создаваемой js-скриптом по технологии AJAX.

4 уровня масштаба Карты достигаются изменением размера ячейки ядра и диапазона вводимых значений ядер в механизме генерации изображения карты. При этом 2 верхних (мелких) уровня масштаба представляют собой изображения, а 2 нижних (крупных) являются html-таблицами, созданными соответствующими Perl-скриптами. В крупном масштабе, за счет использования средств языка javascript, допускается выделение сразу нескольких ядер.

При выборе самого подробного масштаба нажатие на элемент карты, соответствующий определенному нуклиду скриптом "gdrcart.cgi" открывает всплывающее окно, в котором представляются соответствующие характеристики ГДР. Данные представлены в виде развернутой html-таблицы с гиперссылками в информационные системы NSR и EXFOR (Рис. 5). Для получения информации используется соответствующая БД СУБД MySQL.

Специальная кнопка (вопросительный знак желтого цвета) в правом верхнем углу главной страницы Карты (Рис. 4) является ссылкой на документ справки ("gdrhelp.html"), представленный в русском и английском вариантах (русскоязычный вариант – "gdrhelp_ru.html").

Библиотечный модуль "parcheck.pm" выполняет функции безопасности, отсеивая неприемлемые варианты запросов к программным модулям системы.




| EXFOR SUBENTry Number | Nucleus (Z-Symbol) | A | Reaction | Maximum Energy (MeV) | Maximum Cross Section Value (mb) | Full Width at Half Maximum (MeV) | Integration Energy Limit (MeV) | Integrated Cross Section (MeV*mb) | First Moment of Integrated Cross Section (mb) | Reference | NSR keyno | First Author |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 20-CA | 40 | G,ABS | 20.2 | 110 | 0 | 28.5 | 920 | 49.6 | PHYS.LETT.,17,49 (1965) | | B.S.DOLBILKIN+ |
| | 20-CA | 40 | G,ABS | 20 | 105 | 4.5 | 0 | 0 | 0 | PHYS.REV.,137,B576 (1965) | | J.M.WYCKOFF |
| M0648022 | 20-CA | 40 | G,ABS | 19.6 | 95.6 | 4.8 | 40 | 824.8 | 38.1 | R,MSU-INP-2002-27/711,2002 | | B.S.ISHKHANOV+ |
| M0653002 | 20-CA | 40 | G,ABS | 19.6 | 95.6 | 4.8 | 40 | 824.8 | 38.1 | J,IZV,67,1479,2003 | | V.A.EROKHOVA+ |
| I0039037 | 20-CA | 40 | G,XN | 19.98 | 16.8 | 4.5 | 29.5 | 100 | 4.55 | NUCL.PHYS.,A227,513 (1974) | 1974VE06 | A.VEYSSIERE+ |
| | 20-CA | 40 | G,XN | 20.2 | 16.5 | 5.5 | 0 | 0 | 0 | YAD.FIZ.,5,1138(1967) | 1967GO28 | B.I.GORYACHEV+ |
| | 20-CA | 40 | G,XN | 20 | 14.9 | 3.5 | 26 | 73 | 0 | J.PHYSIQUE,27,8 (1966) | 1966MI04 | J.MILLER+ |
| I0039037 | 20-CA | 40 | G,SN | 19.98 | 16.8 | 4.5 | 29.5 | 100 | 4.55 | NUCL.PHYS.,A227,513 (1974) | 1974VE06 | A.VEYSSIERE+ |
| I0039037 | 20-CA | 40 | G,N | 19.98 | 16.8 | 4.5 | 29.5 | 100 | 4.55 | NUCL.PHYS.,A227,513 (1974) | 1974VE06 | A.VEYSSIERE+ |
| m0397004 | 20-CA | 40 | G,N | 20.26 | 20.41 | 5 | 29.1 | 86.3 | 4.1 | YAD.FIZ.,7,1168(1968) | 1968GO29 | B.I.GORYACHEV+ |
| m0397004 | 20-CA | 40 | G,N | 19.24 | 14.86 | 5 | 29.1 | 86.3 | 4.1 | YAD.FIZ.,7,1168(1968) | 1968GO29 | B.I.GORYACHEV+ |
| m0397004 | 20-CA | 40 | G,N | 21.25 | 13.97 | 5 | 29.1 | 86.3 | 4.1 | YAD.FIZ.,7,1168(1968) | 1968GO29 | B.I.GORYACHEV+ |
| m0397004 | 20-CA | 40 | G,N | 23.34 | 12.34 | 5 | 29.1 | 86.3 | 4.1 | YAD.FIZ.,7,1168(1968) | 1968GO29 | B.I.GORYACHEV+ |
| M0659002 | 20-CA | 40 | G,N* | 19.84 | 20.39 | 4.3 | 26.02 | 81 | 3.8 | J,YF,13,1141,1971 | 1971IS06 | B.S.ISHKHANOV+ |
| | 20-CA | 40 | G,P | 20.2 | 85 | 4 | 30 | 510 | 0 | PIS'MA ZHETF,5,225 (1967) | | B.I.GORYACHEV+ |
| | 20-CA | 40 | G,P | 18.6 | 80.5 | 4 | 30 | 510 | 0 | PIS'MA ZHETF,5,225 (1967) | | B.I.GORYACHEV+ |
| M0622002 | 20-CA | 40 | G,P* | 18.5 | 84.6 | 4.3 | 29.5 | 494.5 | 24.4 | J,ZEP,5,225,1967 | | B.I.GORYACHEV+ |
| M0622002 | 20-CA | 40 | G,P* | 20.25 | 86.6 | 4.3 | 29.5 | 494.5 | 24.4 | J,ZEP,5,225,1967 | | B.I.GORYACHEV+ |
| M0397005 | 20-CA | 40 | G,N)* | 20.21 | 15.7 | 4.3 | 29.36 | 86.6 | 4.1 | J,YF,7,1168,1968 | 1968GO29 | B.I.GORYACHEV+ |

Рис. 5. Выходная форма - таблица физических данных - БД и новой электронной Карты (приведен пример для ядра $^{40}$Ca):

- EXFOR SUBENT Number - номер раздела международной БД фактографической информации системы EXFOR [11] (нажатие на номер позволяет получить графическое изображение соответствующего сечения, аналогичное приведенным на Рис. 1);
- Nucleus (Z-Symbol) - заряд и химический символ элемента;
- A - массовое число элемента;
- Maximum Energy - энергетическое положение максимума ГДР;
- Maximum Cross Section Value - амплитуда (значение в максимуме) ГДР;
- Full Width at Half Maximum (FWHM); - полная ширина ГДР на половине его амплитуды
- Intergration Energy Limit - предел интегрирования по энергии;
- Integrated Cross Section - интегральное сечение;
- First moment of Integrated Cross Section - первый момент интегрального сечения;
- Reference - библиографическая ссылка на исходный документ;
- NSR keyno - код ссылки на документ международной справочно-библиографической БД NSR [13] (нажатие на код позволяет получить соответствующий документ – Рис. 6);
- First Author - фамилия первого автора оригинальной публикации.

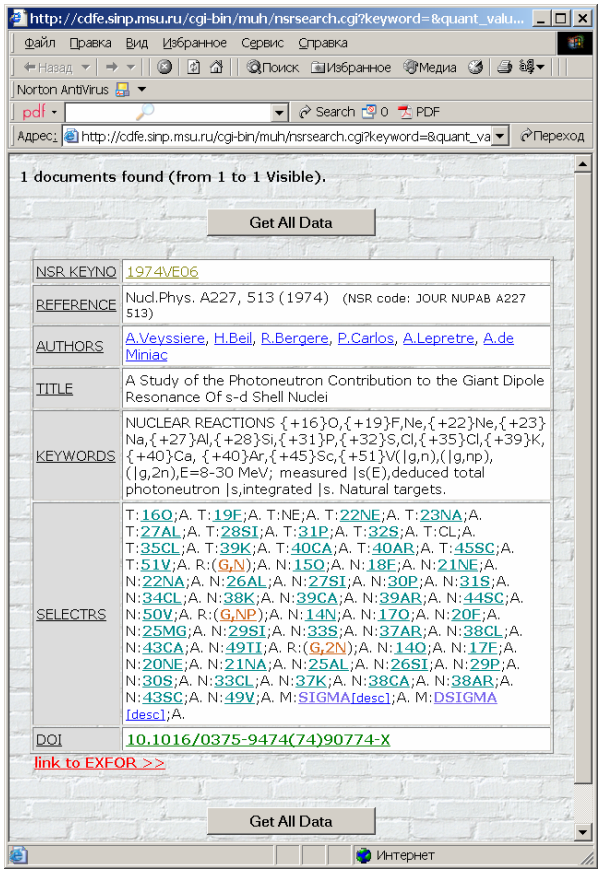



Рис. 6. Документ международной справочно-библиографической БД ЦДФЭ NSR (Nuclear Science References) [13], соответствующий первому из представленных на Рис. 5 номеров “NSR keyno” - 1974VE06.

## Литература


[1] Труды Всероссийской научной конференции “Научный сервис в сети Интернет: многоядерный компьютерный мир. 15 лет РФФИ”. И. Н. Бобошин, В. В. Варламов, С. Ю. Комаров и др., 2007, с. 318.

[2] Труды Десятой Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции”. И. Н. Бобошин, В. В. Варламов, С. Ю. Комаров и др., 2008, с. 259.

[3] Труды Всероссийской научной конференции “Научный сервис в сети Интернет: технологии распределенных вычислений”. И. Н. Бобошин, В. В. Варламов, В. В. Вязовский и др., 2005, с. 156.

[4] Ed. by V. G. Pronyaev, The Nuclear Data Centres Network. IAEA Nuclear Data Section, INDC(NDS)-401, IAEA, Vienna, Austria, 1999.

[5] E. G. Fuller, H. Gerstenberg. Photonuclear Data – Abstracts Sheets 1955 – 1982. NBSIR 83-2742. U.S.A. National Bureau of Standards, 1986.

[6] T. Asami, T. Nakagawa. Bibliographic Index to Photonuclear Reaction Data (1955 – 1992). JAERI-M-93-195, INDC(JPN)-167L,JAERI, Japan, 1993.

[7] V. V. Varlamov, V. V. Sapunenko, M. E. Stepanov. Photonuclear Data Index 1976 – 1995. Izdatel'stvo Moskovskogo Universiteta, Moscow, 1996.

[8] A. V. Varlamov, V. V. Varlamov, D. S. Rudenko, M. E. Stepanov. Atlas of Giant Dipole Resonances. Parameters and Graphs of Photonuclear Reaction Cross Sections. IAEA NDS, INDC(NDS)-394, Vienna, 1999.

[9] Труды Восьмой Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции”. И. Н. Бобошин, В. В. Варламов, С. Ю. Комаров и др., 2006, с. 56.

[10] Труды Десятой Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции”. И. Н. Бобошин, В. В. Варламов, Ю. П. Гангрский и др., 2008, с. 268.

[11] База данных по ядерным реакциям (EXFOR), URL: http://cdfe.sinp.msu.ru/exfor/index.php.

[12] Труды Всероссийской научной конференции “Научный сервис в сети Интернет”. В. В. Варламов, С. Ю. Комаров, С.Б.Семин, В.В.Чесноков. 2003, с. 52.

[13] БД Публикации по ядерной физике (база данных “NSR”), URL: http://cdfe.sinp.msu.ru/services/nsr/Search_form.shtml


## New Digital Chart of Main Parameters of Giant Dipole Resonances of Atomic Nuclei


V. V. Varlamov, V. V. Vyazovsky, I. A. Ekhlakov, S. Yu. Komarov, N. N. Peskov, O. V. Semenov, M. E. Stepanov



The new Internet-resource – relational database – digital Chart of main parameters of Giant Dipole Resonances (GDR) of atomic nuclei – including into the system of relational nuclear databases for physics of atomic nuclei and nuclear reactions of the MSU SINP Centre for Photonuclear Experiments Data (Centr Dannykh Fotoyadernykh Eksperimentov – CDFE) is described. The Chart contains many data on GDR energy positions, amplitudes, widths, integrated characteristics widely used in various fields of basic and applied research and in variety of applications.



♣ Работа выполняется в Лаборатории анализа ядерных данных (Центр данных фотоядерных экспериментов) Отдела электромагнитных процессов и взаимодействий атомных ядер НИИЯФ МГУ. Она использует некоторые результаты и опыт работ по грантам РФФИ №№ 03-07-90431 и 04-02-16275, была частично поддержана грантом Президента РФ для научных школ №НШ-485.2008.2, Госконтрактом № 02.513.12.004, в настоящее время поддерживается грантом РФФИ № 09-02-00368.

# Распределенная информационно-вычислительная система «Атмосферная радиация» ♣

© К.М. Фирсов[1], А.З. Фазлиев[2], Т.Ю. Чеснокова[2], Е.М. Козодоева[2]

1) Волгоградский государственный университет
fkm@iao.ru

2) ИОА СО РАН

**Аннотация**

Представлено описание доступной по сети Интернет распределенной информационно-вычислительной системы «Атмосферная радиация»(ИВС), серверы которой расположены в Институте оптики атмосферы СО РАН (Томск), Волгоградском государственном университете и Уральском государственном университете (Екатеринбург). Создаваемая система ориентирована на исследования проблем переноса радиации в атмосфере Земли, включая прямые и обратные задачи спутникового и наземного зондирования газового состава атмосферы. Адреса ресурсов в Интернете, подготовленных авторами по данному проекту: http://atrad.atmos.iao.ru, http://atmos.physics.usu.ru,. http://atmos.volsu.ru и http://remotesensing.ru.

## 1 Введение

Доступная в сети Интернет информация по атмосферной радиации существует в нескольких формах. Это – многочисленные архивы данных, метеорологические и оптические модели, программы для расчета радиационных характеристик.

Архивы данных. Например, на сайте ARM [9] предоставлен доступ к полевым измерениям радиации и оптического состояния атмосферы для условий Северной Америки, которые позволяют верифицировать современные модели переноса радиации и общей циркуляции. На сайте [6] содержатся архивы спутниковых данных AIRS/AQUA уровня L1 (регистрируемая спектральная яркость излучения атмосферы) и уровня L2 (определенные из L1 в результате решения обратной задачи вертикальные профили температуры и концентрации водяного пара в атмосфере).

В рамках программы ARM проводилось сопоставление радиационных кодов и опубликованы результаты тестовых расчетов, например, [3], доступны также результаты тестовых расчетов для спутниковых радиометров HIRS, AMSU [5] и др. Однако, исходная спектроскопическая информация быстро устаревает, что требует проведения новых расчетов.

В настоящее время созданы пакеты программ для расчета радиационных характеристик (профили поля яркости, потоков излучения), такие как DISORT [1], FASCODE, LBLRTM [7] и др., часть из которых в настоящее время доступна через сеть Интернет.

Следующим шагом в развитии программного обеспечения для радиационных расчетов явилось создание библиотек программ LIBRADTRAN [8], которые на основе отдельных программ позволяют создавать новые компьютерные коды по радиации.

Повышение качества измеряемых данных требует все более сложных радиационных моделей. С другой стороны, постоянно возрастающие объемы информации, требуют повышения скорости обработки информации. Так, например, лежащая в основе многих радиационных расчетов современная спектроскопическая база данных *HITRAN* содержит информацию о более чем 2,5 млн. спектральных линий, и в ближайшее десятилетие ожидается ее увеличение на два порядка в связи с проведением многочисленных *ab initio* расчетов молекулярных спектров. Исследования оптических характеристик аэрозолей и облаков также приводят к увеличению объемов данных, используемых в радиационных расчетах. Помимо этого, различные радиационные модели не только основываются на различных способах расчета, но требуют различных способов представления входных данных.

Поэтому к радиационным моделям предъявляются по сути дела противоречивые требования: высокая скорость и высокая точность. Это требование приводит к привлечению различного типа параметризаций, и как следствие, значительному усложнению моделей. Таким образом, увеличивается сложность программного обеспечения, его структура и связи с программным обеспечением смежных научных дисциплин. Такая

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

ситуация создает ряд трудностей для конечного пользователя как в использовании информационных, так и вычислительных ресурсов в его непосредственной работе.

Следует также обратить внимание на то, что в настоящее время имеется огромное число публикаций по рассматриваемой тематике. Так, например, число публикаций только по исследованию аэрозоля за год достигает несколько десятков тысяч. Все это приводит к тому, что даже специалисты в области атмосферной оптики не всегда могут ориентироваться в этой информации.

В 2004 году в ИОА СО РАН была начата работа по созданию ИВС «Атмосферная радиация». Однако, мы достаточно быстро поняли, что в одиночку такой проект создать очень сложно и с 2007 г. начался новый проект, который объединил уже несколько организаций: ИОА СО РАН, РНЦ «Курчатовский институт», Уральский госуниверситет и Волгоградский госуниверситет. распределенная информационно-вычислительная система «Атмосферная радиация».

## 2 Радиационная модель

Выше отмечалось, что для решения прямых и обратных задач оптики атмосферы к разрабатываемому программному обеспечению по расчету радиационных характеристик предъявляются жесткие требования: высокая точность при высокой скорости счета. В настоящее время высокую точность обеспечивают только прямые методы счета line-by-line, на которые мы и ориентировались. Однако, эти методы очень трудоемки даже для современных ЭВМ. По этой причине необходимы эффективные методы повышения скорости вычислений. Для проведения радиационных расчетов необходима информация о газово-аэрозольном составе атмосферы, облачности, подстилающей поверхности, метеорологическом состоянии и т.п. Задача унификации данных и моделей стояла очень остро. После тщательного анализа выбор был остановлен на моделях, которые основываются на локальных оптических характеристиках среды, таких как коэффициенты поглощения и рассеяния. Однако прямые методы расчета имеют ограниченное применение, поэтому для задач требующих высокой скорости было решено использовать look up table архивы, а для широкополосных радиометров метод рядов экспонент.

Разложение функции пропускания, обусловленной молекулярным поглощением, в ряд экспонент позволяет разрешить целый ряд проблем, возникающих при численных расчетах: оно описывает функцию пропускания с высокой точностью (отличие от прямого счета в среднем составляет величину ~ 1%), является малопараметрическим (число членов ряда не превышает 5-7), имеет экспоненциальную функциональную зависимость как и в случае line-by-line расчетов, что обеспечивает удобство ее применения в различных вычислительных схемах, когда существенно многократное рассеяние [11].

Теоретической основой разложения в ряд экспонент является преобразование Лапласа. Оно позволяет перейти от быстроменяющегося коэффициента молекулярного поглощения к эффективному коэффициенту поглощения, который уже является гладкой, монотонно-возрастающей функцией. Это позволяет радикально на несколько порядков уменьшить число точек при численном интегрировании данной функции. На рис. 1 приведен пример такого преобразования.

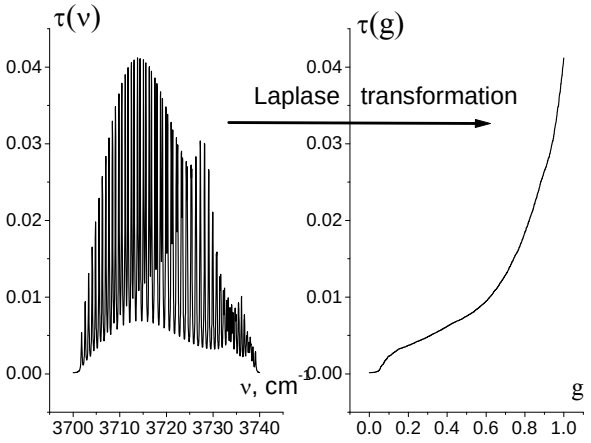



Рис.1. Преобразование Лапласа спектра поглощения.

Развиваемый нами подход для расчета потоков длинноволновой радиации основывается на такой модификации метода k-распределения, когда одновременно учитываются все газы. Термин k-распределение вместо рядов экспонент появился в западной литературе. Смысл его состоит в том, что в заданном спектральном интервале интегральное поглощение определяется функцией распределения коэффициента поглощения. Одно из ограничений данного метода состоит в том, что функция распределения оказывается зависящей от термодинамических параметров атмосферы. Это приводит к тому, что спектр поглощения в приземном слое значительно отличается от спектра поглощения на больших высотах и метод k-распределения дает погрешности в функциях пропускания на уровне 5%. Нами разработан метод учета перекрывания полос разных газов, который позволяет свести эту ошибку к минимуму [2].

В наших работах [13] было показано, что интегральные по спектру радиационные характеристики (яркость, поток) могут быть представлены в виде:

$$I_{\Delta\nu} = \sum_{i=1}^{N} C_i Q_i$$

где $Q_i$ – монохроматические радиационные характеристики, число членов ряда определяется числом параметров при разложении функции пропускания в ряд экспонент.

Расчет радиационных характеристик реализуется в несколько этапов:

Методом line-by-line на основе атласа спектральных линий поглощения атмосферных газов HITRAN рассчитываются вертикальные профили коэффициентов молекулярного поглощения с высоким спектральным разрешением;

Определяются N ($N \leq 10$) значений эффективных коэффициентов молекулярного поглощения для заданной высоты с учетом аппаратной функции прибора и спектрального хода солнечного излучения.

Для расчета широкополосных радиационных характеристик решается уравнение переноса излучения для каждой компоненты (их число не превышает 10), причем эффективный коэффициент поглощения здесь можно использовать как обычный монохроматический коэффициент поглощения, который входит в альбедо однократного рассеяния и оптическую толщу.

Для решения стационарного уравнения переноса с учетом многократного рассеяния излучения аэрозолем и облаками могут использоваться различные методы, такие как метод Монте-Карло, метод сферических гармоник, метод дискретных ординат и др. Экспоненциальный вид функций пропускания обеспечивает возможность применения практического любого метода Мы выбрали метод дискретных ординат DISORT[1], т.к. он сочетает в себе как высокую скорость, так и хорошую точность вычислений.

В ИВС «Атмосферная радиация» реализовано два типа расчетов:

1) эффективные коэффициенты молекулярного поглощения и функций пропускания, обусловленные молекулярным поглощением с использованием рядов экспонент и методом line-by-line.

2) Расчет интенсивностей и потоков радиации на основе разложения в ряд экспонент.

Первый тип расчетов позволяет пользователям получить данные о характеристиках молекулярного поглощения в атмосфере и использовать их в своих радиационных кодах. Сопоставление пропускания рассчитанного методом line-by-line и с применением ряда экспонент дает возможность оценить точность расчета и выбрать необходимое число членов ряда экспонент**.**

Применяемый нами подход к учету молекулярного поглощения обладает рядом достоинств: короткий ряд обеспечивает высокую скорость счета; разделение во времени расчета коэффициентов молекулярного поглощения и решения уравнения переноса излучения позволяет исследователю с минимальной коррекцией использовать алгоритмы и программы, разработанные для радиационных расчетов без учета молекулярного поглощения.

Данный способ расчета радиационных характеристик позволяет унифицифировать входные данные и модели, используемые для расчета радиации.

Входные данные структурированы следующим образом:

- Спектроскопические параметры линий;
- Метеорологические модели;
- Аэрозольные оптические модели;
- Оптические модели облаков.
- Справочная информация о характеристиках облачности и аэрозоля для Северного полушария Земли.

В настоящее время для каждого типа вышеперечисленных данных на портале ATMOS созданы либо информационные системы, либо отдельные интерфейсы.

Поскольку для расчета эффективных коэффициентов поглощения используется прямой метод счета, то для повышения скорости вычислений используется многосеточный метод [12], разработанный Фоминым Б.А., что позволяет на два – три порядка уменьшить число расчетных узлов без потери точности. В настоящее время этот подход получает все более широкое распространение, так как он, в отличие от подхода, используемого в широко известной программе FASCODE (USA), позволяет добиться более высокой скорости счета и не имеет ограничений на форму контура спектральной линии.

Метод, использованный в нашем подходе для расчета полей интенсивностей, выгодно отличается от общепринятых. Это обусловлено не только тем, что расчеты проводятся непосредственно на нашем сайте, где есть дружественный интерфейс, но также и тем, что применение современных подходов к параметризации молекулярного поглощения позволяет использовать разные версии спектроскопических баз данных.

## 3 Общая характеристика РИВС «Атмосферная радиация»

Для того, чтобы радиационные модели стали более доступными создается распределенная информационно-вычислительная система «Атмосферная радиация», которая не только обеспечивает доступ к данным, но и позволяет проводить расчеты радиационных характеристик атмосферы Земли.

ИВС "Атмосферная радиация" является частью научного портала для атмосферных наук ATMOS (http://atmos.iao.ru), который представляет собой интегрированный набор множества распределенных, но координируемых предметных сайтов, содержащих типовую информацию с исследовательскими базами данных, моделями и аналитическим инструментарием для прямого использования и визуализации данных.

Работы по созданию промежуточного программного обеспечения (ППО) и средствам создания и поддержки информационно-вычислительных систем с доступом по сети Интернет ведутся в ИОА СО РАН последние 8 лет. За этот период на основе оригинального ППО создан портал по атмосферным наукам, включающий в себя 6 информационных систем по таким предметным областям как атмосферная спектроскопия, радиация, химия, климат и погода и др. [10]. Ключевым разделом ППО является система управления рабочими потоками, обеспечивающая формирование статического меню, механизм его динамической визуализации и сохранение целостности данных пользователя на стороне сервера. Вспомогательными приложениями, интегрированными в ППО, являются тезаурус предметных областей, с которыми связаны каталогизаторы информационных ресурсов ИВС, словари ссылок на информационные ресурсы, система управления контентом и т.д. Программное обеспечение для формирования распределенной ИВС (промежуточное программное обеспечение, система управления рабочими потоками, система формирования интерфейсов и т.д.), созданное в проекте РФФИ №06-07-89201, используется в нашем проекте. Открытые стандарты W3C (XML, RDF, OWL, и другие) положены в основу его создания.

Комплексный подход, в рамках которого одна часть вычислений проводится на стороне клиента, а другая — на стороне сервера, практически не распространен в Интернете. Тем не менее, его использование перспективно в задачах графической обработки предметных данных. Наиболее значительное ограничение на технологию вычислений в ИВС накладывает архитектура клиент-сервер. В рамках этой архитектуры важным является то, где проводятся вычисления: на стороне клиента или сервера. В Интернет-доступных системах для проведения вычислений только на стороне клиента, как правило, используются Java-script или Java-applet. Однако возможности такого подхода ограничены задачами, не требующими высокопроизводительных вычислительных ресурсов или использующими небольшой по объему пересылаемый клиенту код. Инструментальные средства для организации вычислений на серверной стороне значительно богаче. С точки зрения гибкости интерфейса пользователя при организации вычислений на стороне сервера для решения задачи ключевым моментом является то, как организована работа с данными пользователя. Наиболее распространен подход, при котором для решения задач данные пользователя не хранятся на сервере, а только доставляются клиенту после проведения вычислений. Связано это с тем фактом, что такая организация вычислений не требует создания системы управления данными пользователя (СУДП) на стороне сервера. В РИВС «Атмосферная радиация» используется СУДП, которая позволяет организовать хранение данных пользователя, обеспечить их целостность, переименовывать и удалять структуры данных и проводить сравнение результатов решения однотипных задач.

В настоящее время РИВС «Атмосферная радиация» включает следующие данные:

- Сформирована база данных полей яркости нисходящего диффузного солнечного излучения в безоблачной атмосфере в спектральном диапазоне 0.34-4 мкм для типичных метеорологических и оптических ситуаций (на основе расчетов методом Монте-Карло).
- Разработана база данных аэрозольных статистических моделей на основе моделей Крекова-Рахимова. Реализована возможность использования среднециклической модели Крекова-Рахимова при расчетах переноса радиации «радиационным блоком ИОА»
- Создана и регулярно пополняется база данных результатов экспериментальных измерений на радиационном комплексе ИОА СО РАН. Результаты измерения интегральной оптической толщи и радиационных характеристик атмосферы представляются на сайте в табличном и графическом виде.
- Создана база данных по самолетным измерениям спектральных радиационных характеристик атмосферы и поверхности.
- Созданы архивы оптических характеристик облаков на основе параметризаций Y.X. Hu и K. Stamnes и A. Slingo и H.M. Scherker и создан интерфейс, обеспечивающий вычисление коэффициентов ослабления, альбедо однократного рассеяния и среднего косинуса рассеяния для жидкокапельных облаков в диапазоне длин волн 0.25-4 мкм с разбиением на 24 интервала для эффективного радиуса капель 2.5-60 мкм.

ИВС атмосферная радиация включает следующие модели:

Разработан интерфейс для расчета восходящих и нисходящих спектральных потоков и интенсивности коротковолнового излучения на верхней и нижней границе атмосферы. Для расчета уравнения переноса излучения используется пакет программ DISORT. Молекулярное поглощение учитывается на основе рядов экспонент, параметры которых рассчитываются прямым методом line-by-line. Пользователю предоставляется выбор параметров линий атмосферных газов из различных спектроскопических банков данных. Для этого организована связь с распределенной ИВС «Молекулярная спектроскопия» (http://saga.atmos.iao.ru.

Разработан интерфейс для расчета интегральных потоков радиации для моделей ИВМ РАН и модели В.А.Фролькиса. Данные радиационные блоки дают возможность рассчитать восходящие и нисходящие интегральные потоки радиации, а также скорость радиационного выхолаживания атмосферы.

Информационно-вычислительная система «Атмосферная радиация», доступна для коллективного использования. Доступ к ресурсам организован с помощью веб-интерфейса (http://atrad.atmos.iao.ru) в ИОА СО РАН. В настоящее время созданы зеркала в Волгоградском и Уральском (Екатеринбург) госуниверситетах.

Обмен данными в распределенной сети, предполагает, что решения задач, загруженные в данном узле, автоматически не реплицируются на остальные узлы распределенной системы. Связано это с тем обстоятельством, что на каждом узле РИС существует собственная система регистрации пользователей. Данные, заносимые пользователем в БД РИВС, являются приватными и открываются для свободного доступа после обращения пользователя к экспертам и решения экспертов об их открытой публикации. Публичные данные могут быть реплицированы по решению администратора узла с других узлов системы. Семантические метаданные для опубликованных источников информации должны быть идентичными на всех узлах распределенной системы.

Поскольку разработчики ИВС находятся в разных городах, то для удобства общения создан сервис с функциями передачи и запроса новостей размещенный на всех узлах распределенной информационной системы «Атмосферная радиация» и «Молекулярная спектроскопия». Раздел "Новости сайта" позволяет администраторам узлов информировать пользователей системы о происходящих изменениях ресурса, добавлении новых разделов и их назначении, а также публиковать любую другую информацию соответствующую тематике сайта. Обмен новостями между узлами распределенной системы основан на использовании технологии веб-сервисов. Особенности технической реализации системы обмена подробно описаны в работе [10].

## 4 Интерфейс для расчета переноса радиации в атмосфере Земли

Все входные параметры описываемой нами модели делятся на две группы. Первая группа параметров может задаваться и изменяться пользователем, а вторая представлена в системе в виде встроенных моделей, например, возможность использования среднециклической модели Крекова-Рахимова при расчетах переноса радиации.

Входные параметры, доступные для изменения пользователем, объединены в логические группы, такие как:

"начальные условия", в которых задаются как общие параметры расчета (рис.2) (число слоев, на которые разбита атмосфера, число гауссовских квадратур для расчета эффективных коэффициентов поглощения, альбедо поверхности, зенитный угол Солнца, азимутальный угол Солнца, зенитный угол трассы, число азимутальных углов приемника), так и полярные углы приемника, тип аппаратной функции и ее параметры, а также учитывать ли при расчетах облачность;


Рис 2. Интерфейсы задания "начальных условий" и "параметров атмосферы".

"параметры атмосферы", где пользователь может выбрать метеомодель и задать ее настройки, в которые входит список газов учитываемых в расчетах;

"спектроскопические данные" - в этой части интерфейса происходит взаимодействие с ИВС "Атмосферная спектроскопия", в результате которого пользователь может выбрать интересующий его банк данных, содержащий параметры спектральных линий для выбранных им газов при настройке параметров атмосферы;

"параметры аэрозоля", позволяющие выбрать аэрозольную модель (Крекова или Произвольную) и задать высотные профили коэффициента аэрозольного ослабления и альбедо однократного рассеяния для аэрозоля, а также задать параметры индикатриссы рассеяния, в случае выбора произвольной аэрозольной модели;

"модель облачности" — эта часть интерфейса становится доступной только в случае принятия решения учитывать облачность при задании начальных условий, и позволяет выбрать одну из встроенных моделей облачности, либо задать параметры для произвольной модели.

В 2008 на основе спутниковых данных созданы интерфейсы для базы данных по оптическим характеристикам аэрозоля и облаков для северного полушария Земли. В этом разделе сайта пользователю предоставляется возможность просмотра данных в табличном и графическом представлении. В интерфейсе для входных параметров задаются интервалы по широте и долготе, и месяц года. Выборка данных из базы данных производится с помощью запроса на языке

MySQL по всем девяти величинам. У программного модуля используемого для табличного отображения есть возможность скрывать/включать колонки, поэтому пользователь может выбирать необходимые ему величины непосредственно при просмотре таблицы с результатами запроса. Графическое представление данных реализовано с помощью графического пакета GnuPlot. С помощью цветной поверхности на графике представляется пространственное распределение выбранной пользователем величины в заданных им временном (месяц) и координатном интервалах.(см. рис

Теперь, пользователь, имея в своем распоряжении такую базу данных, имеет возможность при формировании своих наборов данных опираться на справочную информацию и учитывать региональные особенности.

Еще одним новым разделом сайта «Атмосферная радиация» является раздел для расчета переноса радиации в атмосфере Земли в спектральных каналах спутниковых радиометров среднего спектрального разрешения. Этот интерфейс позволяет задавать пользователю такие начальные параметры как: спектральный интервал (0-3000 см-1), учитывать ли континуум H2O и если да, то модель континуума и интервал учета континуума для моделей RSB и ARF. Также пользователь имеет возможность выбрать метеомодель ИОА, AFGL или задать собственную на базе модели AFGL.

Все входные параметры и результаты расчета сохраняются в "портфеле пользователя", что позволяет сохранять связь между входными параметрами модели и ее результатами. После окончания расчета система перенаправляет пользователя на страницу, где отображаются результаты данной задачи. (см. рис.3)

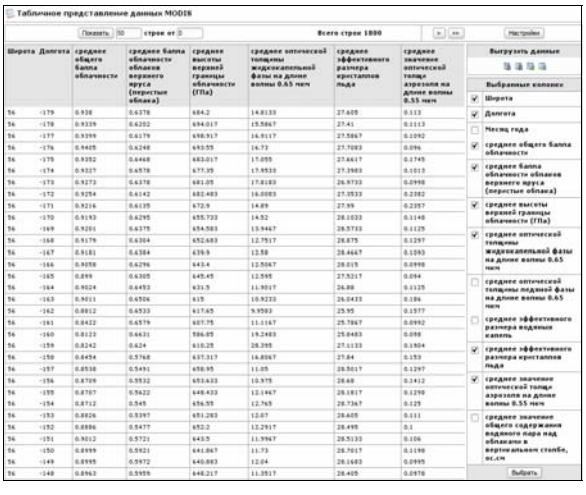



Рис.3а Интерфейс табличного отображения спутниковых данных MODIS.

Все введенные пользователем параметры сохраняются в системе с помощью СУДП. Далее производится последовательный запуск расчетных модулей и результат вычислений снова сохраняется в системе и отображается пользователю. (рис.3) Такой подход позволяет внося незначительные изменения в значения входных параметров сравнивать результаты полученные в ходе расчетов, и не терять из при выходе из системы.

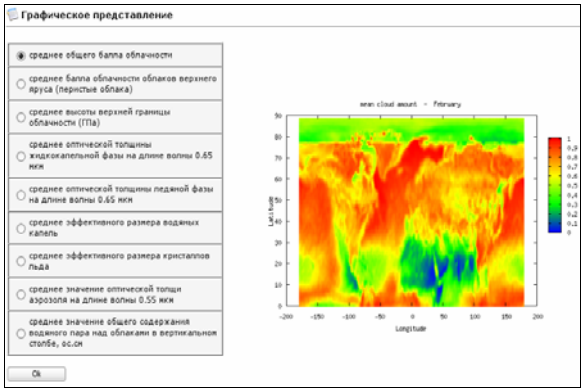



Рис.3б Интерфейс графического отображения спутниковых данных MODIS.

## 5 Расчеты переноса длинноволновой радиации

При обсуждении проблемы повышения температуры Земли вследствие парникового эффекта рассматривают, как правило, радиационный баланс между нисходящим потоком от солнца и восходящим собственным излучением Земли на верхней границе атмосферы [4]. Следует также отметить, что рассматривают метеорологические модели, которые соответствуют средним значениям. Не преуменьшая роли $CO_2$, по нашему мнению следует особое внимание уделить водяному пару, который определяет радиационный баланс нижней атмосферы не только в виде явных потоков, но и в виде скрытых потоков тепла, обусловленных процессами испарения и конденсации водяного пара. Причем, теплота фазовых переходов является одним из основных механизмов непосредственного нагревания атмосферы. Роль паров воды при их высоком содержании в приземном слое резко возрастает. Разрабатываемая ИВС позволяет рассчитывать явные потоки нисходящей радиации, которые определяют приземную температуру. Так, например, было проведено моделирование этих потоков для типичных условий Нижнего Поволжья. Для того, чтобы выбрать экстремально низкое, среднее и экстремально высокое содержание паров воды в атмосфере был проанализирован десятилетний ряд аэрологических наблюдений за вертикальными профилями температуры и влажности вблизи г. Волгоград. Основным исходным материалом для исследования особенностей полей температуры и влажности в свободной атмосфере послужили данные реанализа с сайта www.noaa.gov, рассчитанные по многолетним радиозондовым измерениям мировой

сетью аэрологических станций. Результаты моделирования приведены в Табл. 1.

Табл.1. Вклад паров воды и углекислого газа в нисходящие потоки радиации

| Приземная температура, К | 312 | 286 | 296 | 296 |
|---|---|---|---|---|
| Общее содержание $H_2O$ в вертикальном столбе атмосферы, ос.см | 2.12 | 0.976 | 4.71 | 2.81 |
| Нисходящий поток $F^{\downarrow}$, Вт/мI | 404.3 | 287.6 | 381.4 | 410.3 |
| $\Delta F^{\downarrow}$,Вт/мI Вклад $H_2O$ | 291.4 | 204.2 | 286.9 | 299.6 |
| $\Delta F^{\downarrow}$, Вт/мI Вклад $CO_2$ | 20.9 | 18.7 | 4 | 16 |

Согласно оценкам, полученным в данной работе радиационные процессы в приземном слое атмосферы главным образом обусловлены парами воды, причем вклад $CO_2$ уменьшается по мере возрастания концентрации $H_2O$ и для некоторых ситуаций, характерных для летних условий Нижнего Поволжья его доля в нисходящих потоках снижается до 1%.

## Заключение

В работе дано описание некоторых Web-интерфейсов информационно-вычислительной системы “Атмосферная радиация”. Информационная система построена средствами Интернет-технологий. Для разработки ИВС использовано программное обеспечение: Web сервер Apache, скриптовый язык PHP4, пакет программ для построения графиков GNUPlot.

Модели данной ИВС могут представлять интерес для специалистов в области атмосферной радиации и климата, а также аспирантов и студентов.

Авторы благодарят за финансирование РФФИ (проект 07-07-00269).

## Литература

[1] DISORT (ftp://climate.gsfc.nasa.gov/pub/wiscombe/Multiple_Scatt)

[2] Firsov K/M, Mitsel A.A., Ponomarev Yu.N., Ptashnik I.V. Parametrization of transmittanse for application in atmospheric Optics// Journ.Quant.Spectr. and Radiat.Trasf. –1998. - V.59, No3-5. - pp.203-213.

[3] Haltore R.N., Grisp D., Shwartz S.E. at all. Intercomparison of shortwave radiatiative codes and measuremants// Journ. Geoph. Research. – 2005. – V. 110, pp.D11206, pp.1-18.

[4] Huang Yi, Ramaswamy V., Soden B. An investigation of the sensitivity of the clear-sky outgoing longwave radiation to atmospheric temperature and water vapor// Journ. Geoph. Research. – 2007. – V. 112, D05104, pp.1-13.

[5] Intercomparison of forward and Jacobian radiative transfer models for HIRS and AMSU channels http://collaboration.cmc.ec.gc.ca/science/arma/intercomparison/

[6] Jet propulsion laboratory http://airs.jpl.nasa.gov

[7] LBLRTM (http://rtweb.aer.com/)

[8] Mayer B. and Kylling A. The libRadtran software package for radiative transfer calculations - description and examples of use//Atmos. Chem. Phys., 5, pp. 1855-1877, 2005

[9] The Atmospheric Radiation Measurement Program, http://www.archive.arm.gov/about.html

[10] Козодоев А.В., Козодоева Е.М., Привезенцев А.И., Фазлиев А.З. Труды 13-й Байкальской Всероссийской конференции "Информационные и математические технологии в науке и управлении", 2008.

[11] Мицель А.А., Фирсов К.М. Развитие моделей молекулярного поглощения в задачах переноса излучения в атмосфере Земли //Оптика атмосферы и океана. – 2000. - №2. - С.179-197.

[12] Мицель А.А., Фирсов К.М., Фомин Б.А. Перенос оптического излучения в молекулярной атмосфере. Томск: STT, 2001. 444 С.

[13] Фирсов К.М., Чеснокова Т.Ю., Белов В.В., Серебренников А.Б., Пономарев Ю.Н. Ряды экспонент в расчетах переноса излучения методом Монте-Карло в пространственно неоднородных аэрозольно-газовых средах // Вычислительные технологии 2002. Т. 7. № 5. С. 77-87.

## Distributed information-computational system “Atmospheric radiation”

K.M. Firsov, A.Z. Fazliev, T.Yu. Chesnokova,

E.M. Kozodoeva

The internet-accessible distributed information-computational system “Atmospheric radiation” is described. The servers of the system are situated in the Institute of Atmospheric Optics SB RAS (Tomsk), Volgograd State University, and Ural State University (Ekaterinburg). The created system is aimed at investigations of radiative transfer in the Earth atmosphere, including direct and inverse problems of satellite and ground-based sounding of gaseous composition of the atmosphere.

The Internet resources addresses, prepared by the authors at the project: http://atrad.atmos.iao.ru, http://atmos.physics. usu.ru,. http://atmos.volsu.ru и http:// remotesensing.ru.

♣ Данная работа выполнена при финансовой поддержке РФФИ (грант РФФИ № 07-07-00269).

# Виртуальные коллекции животных и интерактивные определители биологических объектов♣

© А.Г.Кирейчук, А.Л.Лобанов, И.С.Смирнов, А.Т.Вахитов, Е.П.Воронина, О.Н.Пугачев

Учреждение Российской Академии наук Зоологический институт РАН, Санкт-Петербург
smiris@zin.ru

**Аннотация**

Разработка виртуальных коллекций, определительных ключей и систем экологического мониторинга получает за последние годы более интенсивное развитие, связанное с совершенствованием компьютерных технологий и возрастающей ролью оценки и сохранения биоразнообразия под действием антропогенной нагрузки. В связи с распространением операционной системы Windows и Интернета, становится актуальным перевод программ ввода и работы с ключами в эту среду, а также представление данных систем в Интернете для более оперативного использования в целях управления ресурсами и в образовании.

## 1 Введение

В своей книге «Электронные библиотеки» Вильям Армс дает: «Информационное определение ЭБ: управляемая коллекция информации в совокупности с соответствующими сервисами, причем информация хранится в цифровых форматах и доступна по сети» [2, стр. 10]. Большинство компьютерных разработок в Зоологическом институте РАН тесно связано с идеологией электронных библиотек или коллекций [34].

Важным аспектом зоологических исследований является использование реальных коллекций, наличие возможности перепроверить то или иное описание вида, выявить новый признак, рассмотреть детали строения под микроскопом или даже просто удостовериться в наличии именно данного вида в данной точке планеты, что особенно становится актуальным в свете интенсивного распространения чужеродных и, особенно, вредоносных видов (классические примеры: колорадский жук и двустворчатый моллюск-дрейссена). Можно еще упомянуть недавние катастрофические последствия заносов в США из Китая жука-усача Anoplophora glabripennis и жука-златки Agrilus planipennis. Несмотря на усилия множества служб и специалистов, эти два вида принесли многомиллионный ущерб и продолжают расселяться. Виртуальные коллекции позволяют ускорить привлечение настоящих коллекций для анализа материала в данных ситуациях, а в ряде случаев и заменить их.

Во всех научно-исследовательских зоологических музеях, где хранятся собрания образцов животных, как ныне живущих, так и вымерших, которые собраны и поддерживаются учеными для более полного изучения и документирования биологического разнообразия, ведутся традиционные каталоги и картотеки и во многих начинают приступать к созданию или электронных каталогов, или специализированных баз данных, или даже информационно-поисковых систем [29, 31, 32]. Появляются компьютерные каталоги, которые выставляются в глобальной информационной сети. Перевод списков экспонатов, а затем и музейных каталогов, в цифровую форму и создание коллекционных баз данных служит первым шагом на пути создания виртуальных коллекций. Вторым шагом в создании электронных или цифровых коллекций является накопление изображений экспонатов и создание электронных фотогалерей и фотоальбомов для зоологических образцов [19, 25]. Соединение цифровых коллекций с информацией о музее, его истории, кураторах, специалистах (которые определяли материал), постепенно приводит к созданию виртуального музея [8, 21, 26-29].

Интернет-технологии ускорили процессы представления разнообразной музейной информации.

## 2 Электронные публикации

Среди компьютерных технологий определенное место занимают электронные публикации на компакт-дисках CD-ROM и в виде файлов на серверах сети Интернет [1, 22]. Первая серьезная электронная публикация на веб-портале ЗИН - интерактивный каталог коловраток пресных вод Северо-запада России [9]. Исследованиями авторов каталога охвачены около 100 озер, более 70 рек, свыше 10 водохранилищ и другие водоемы.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

В ЗИНе разработана концепция построения компьютерных зоологических руководств типа «Фауна России» и «Определители по фауне России» [3, 12-14, 17, 18, 24]. Эта концепция реализована в пакете программ, получившем название DIALOBIS (DIALOgue Biological Identification Systems). В полном объеме эта идеология и оригинальный пакет программ, созданный сотрудниками ЗИНа, впервые на практике использованы немецкой фирмой «dialobis edition» для подготовки серии биологических изданий на лазерных дисках [16 и др.].

Функциональная основа DIALOBIS - многоаспектное представление об исходном наборе таксонов, который можно исследовать и редуцировать разными способами. Для этого используются специализированные прикладные программы, одновременно являющиеся инструментами исследования какого-то аспекта и фильтрами. Эти программы получают на входе набор таксонов (либо в виде копии исходного полного набора, либо как результат одной из предыдущих редукций, сам текущий набор или то его подмножество, которое проходит через все фильтры), а на выходе могут редуцировать его в соответствии с желанием пользователя. Такие действия отдельных программ не сразу влияют на состав текущего набора, а накапливаются в виде совокупности фильтров, информация о которых постоянно выводится на экран главной управляющей программой пакета. Такой пакет программ и управляемая им информация получили название - гипербаза данных. Гипербаза дает возможность пользователю с помощью отдельных шагов многоаспектного поиска активно конструировать требуемый для детального изучения массив данных. Широкое применение такой технологии позволяет создавать очень эффективные зоологические электронные руководства. К сожалению, на практике бумажные технологии пока превалируют в сознании над электронными. С использованием базы данных по рыбам, разработанной Е.П.Ворониной, было подготовлено и опубликовано несколько каталогов фондовой коллекции Зоологического института [4, 5, 6, 7].

Создание биологических (зоологических, ботанических) информационно-поисковых систем - способствует быстрой и качественной экологической экспертизе, как в интересах рационального природопользования, так и при обеспечении экологической безопасности, например, при освоении различных месторождений полезных ископаемых. Прошедшее в 2004 году международное совещание по информационным системам, применяемым в изучении морского биоразнообразия, показало значительную востребованность ИПС в биологии и, в частности, в морской экологии [30, 33, 35, 36].

Видовой состав организмов той или иной территории или акватории необходимо знать не только из чисто академического интереса. Очень остро последнее время стоит проблема видов-вселенцев, которые способны серьезно нарушать экологическую обстановку и вести к колоссальным экономическим потерям. Эта проблема может быть решена, только тогда, когда хорошо известны местные фауна и флора, т.е. состав населения водоемов, и тогда, когда на протяжении многих лет можно проследить динамику появления чужеродного вида и захват им новых участков и мест обитания. Без современных информационных технологий механизм подобного мониторинга создать невозможно, учитывая современные средства передвижения и пассивного переноса нежелательных чужеродных видов.

### 2.1 Электронные справочные пособия и определители. Краткая предистория

Важным моментом развития коллекционных баз данных и ИПС является создание на их основе полноценных справочных пособий и определителей. Своевременная идентификация тех же чужеродных вселенцев и принятие соответствующих мер, предотвращающих их нежелательное распространение, зачастую может сэкономить колоссальные финансовые и человеческие ресурсы.

Первые опыты применения ЭВМ для диагностики таксономической принадлежности биологических объектов были осуществлены в середине 60-х годов, когда ЭВМ еще были для биологов экзотической техникой. Интересно, что на заре компьютерной биологической идентификации отставания отечественных исследований от зарубежных как по качеству, так и по количеству диагностических систем практически не было. Позднее за рубежом, благодаря развитию компьютерной индустрии, в этой области произошел всплеск активности американских, английских и австралийских биологов, который привел к созданию сразу нескольких машинных систем, и таким образом у зарубежных коллег появился количественный перевес. Пиком этой активности можно считать выход в 1975 году сборника "Biological Identification With Computers".

К началу 80-х годов сложилось представление о специфических функциях компьютерных диагностических систем. Их полный набор включал:

1. Предварительную обработку диагностической информации о таксонах.
2. Накопление, хранение и анализ информации на машинных носителях.
3. Диалоговый диагноз с оптимизацией со стороны ЭВМ.
4. Автоматическое составление оптимизированных текстовых определителей.
5. Анализ параметров готовых определителей.

Первая в СССР действующая полная машинная диагностическая система "Диагностика-1" (т.е. выполняющая все перечисленные выше основные специфические функции компьютерных диагностических программ) была создана

А.Л.Лобановым в 1974 году на ЭВМ "Наири-2". Затем эта система расширялась, модернизировалась в соответствии с растущими возможностями доступных биологам ЭВМ, - сначала ЕС "Ряд-1", затем БЭСМ-6, СМ-4, СМ-1420 и, наконец, IBM PC. Начиная с пятой версии системы ("Диагностика-5") программы разрабатывались только для IBM PC, они использовали базы данных формата DBF и состояли из модулей, написанных на языке Фортран-88 и на внутреннем языке СУБД FoxPro. С 1992 г. к работе над диагностическими компьютерными системами подключился М.Б.Дианов. Усовершенствованная версия системы "Диагностика-5" получила новое название "BiKey5", а входящая в нее специализированная диалоговая программа была названа "PicKey". Последний вариант комплексной системы "BiKey7" (1996-1998 гг., языки программирования FoxPro и Fortran) был создан совместно с М.Б. Дианов [21].

Разработка последней версии системы (Bikey8b/PicKey8b) закончена в 2005 г. Все блоки системы программировались на языке C++ для Windows и использовали базы данных формата DBF. Программа PicKey8b является одной из лучших в мире в своем классе (интерактивные диагностические программы без специальной ориентации на использование в сети Интернет). С помощью перечисленных программ были созданы определители, включающие до 400 таксонов самых разнообразных организмов (жуки - А.Л.Лобанов, деревья - Б. и В.Шиловы, медузы - С.Д.Степаньянц, офиуры - И.С.Смирнов, циклопы - В.Р.Алексеев, нематоды - А.Ю.Рысс и др.). О программе PicKey можно узнать на специальном веб-сайте: http://www.zin.ru/projects/pickey. Этой программе и созданным с ее помощью определителям посвящены десятки публикаций. Ссылки на них имеются на сайте PicKey и веб-странице: http://www.zin.ru/Animalia/Coleoptera/rus/all_ref.htm.

Эти определители охватывали лишь небольшое число таксонов и работали на отдельных персональных компьютерах. Для более совершенного механизма определения необходимо было создать алгоритм, позволяющий производить диагностику значительно большего числа предполагаемых объектов идентификации.

### 2.2 Интернет-определители

С появлением и широким распространением Интернета появилась идея написания программы, которая бы давала возможность дистанционного определения [23].

В ходе подготовки к реализации задуманного, при поддержке РФФИ (грант N 02-07-90105), в течение 2002-2004 гг. был создан электронный Атлас жуков России и сопредельных стран, который размещен на сайте Зоологического института РАН: http://www.zin.ru/Animalia/Coleoptera/index.htm. В 2005 г. была получена поддержка РФФИ специально на разработку программного обеспечения к многовходовым политомическим определителям с использованием сети Интернет. Работа над этим проектом стала возможной благодаря участию специалистов в области сетевых компьютерных систем на основе баз данных - О.Н.Граничина и А.Т.Вахитова. Так родился проект WebKey-X: http://www.zin.ru/projects/webkey-x/index.html, в котором участвовали специалисты по разным группам насекомых и офиурам (грант РФФИ N 05-07-90179а, руководитель А.Г. Кирейчук). Связь определителей с атласами, а также необходимыми ссылками на литературные источники по каждому таксону позволяет достичь высокой эффективности на всех этапах определения и проверки (уточнения) его надежности. В настоящее время совершенствуются методы диагностики по медузообразным (С.Д. Степаньянц), иглокожим (А.В. Смирнов, И.С. Смирнов), насекомым (А.Г. Кирейчук, А.Л. Лобанов, А.И. Халаим) (грант РФФИ N 09-04-00789а).

Проведенный в 2005 г. обзор существующих диагностических систем, показал наличие довольно развитых программных продуктов, обеспечивающих определение различных групп животных и растений [23]. На основе анализа компьютерных систем создана таблица, идея и часть содержания которой заимствованы у M. Dallwitz'a: http://delta-intkey.com/www/comparison.htm. Таблица сильно модифицирована: в нее добавлены новые характеристики (они предложены А.Лобановым и Д.Дмитриевым и выделены в раздел "New features") и ряд отсутствовавших в ней программ (в том числе наши разработки - PICKEY и WebKey-X, а также программа Д.Дмитриева - 3I): http://www.zin.ru/projects/webkey-x/index.html.
Таблица активно дополняется оценками для новых характеристик и для добавленных программ. В таблицу включены следующие программы: WebKey-X, PICKEY, 3I (by D.Dmitriev), Flora Search, Intkey, LucID, MEKA, PollyClave, IdentifyIt, Visual Key, NaviKey, XID, DAP, DAWI.
На основе полученных данных шла разработка структуры типового интерактивного определителя (компьютерного ключа) и создание пилотных вариантов различных определителей. Принципиальные свойства интерактивного определителя или компьютерного ключа могут быть условно разделены на 2 группы - структурные и динамические. Структурные свойства - особенности структуры базы данных ключа. Динамические свойства - специфические особенности шага идентификации, представляющего собой итеративно повторяющийся цикл диагноза (под шагом, более точно, понимается набор процедур, обычно включающий выбор признака, ввод в компьютер информации о состоянии признака и получение ответа компьютера с текущим набором таксонов, обладающих данным состоянием признака.). Главные структурные свойства ключа: число входов и число состояний в одном признаке. Важнейшее свойство - число входов для начала каждого шага диагноза.

Таблица.

| ХАРАКТЕРИСТИКИ \ ОПРЕДЕЛИТЕЛИ | MAX_VALUE | WEBKEY_X | PICKEY | DMITR_3I | FLORA_SRCH | INTKEY | LUCID | MEKA | POLLYCLAVE | IDENTIFYTT | VISUAL_KEY | NAVIKEY | XID | DAP | DAWI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| НОВЫЕ ХАРАКТЕРИСТИКИ | | | | | | | | | | | | | | | |
| Использование традиционных баз данных | 16 | 16 | 16 | 16 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Указание на недопустимые состояния вместо удаления | 4 | 0 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Демонстрация числа таксонов с определенным состоянием признака | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Возможность различных путей определения на основе одной матрицы данных | 2 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Многоязычные ключи (весь текст на нескольких языках) | 8 | 0 | 8 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Хранение, использование и поиск синонимических названий | 2 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Использование на компьютерах MAC и IBM | 4 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Поиск в Интернете дополнительной информации по каждому таксону | 4 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Хранение и поиск библиографических источников по каждому из таксонов | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Хранение и поиск информации по кормовым растениям | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Хранение и поиск данных по распространению с использованием карт | 4 | 0 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ПРЕИМУЩЕСТВА ПЕРЕД ТРАДИЦИОННЫМИ КЛЮЧАМИ | | | | | | | | | | | | | | | |
| Использование признаков с нефиксированными значениями | 8 | 0 | 0 | 8 | 0 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 0 | 0 |
| Изменение и удаление признаков | 4 | 4 | 4 | 4 | 0 | 4 | 4 | 4 | 4 | 4 | 0 | 4 | 4 | 0 | 0 |
| Допустимость ошибок определяющего | 8 | 0 | 8 | 8 | 0 | 8 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Указание программы на ошибки пользователя | 4 | 0 | 0 | 4 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Оценка неопределенности | 8 | 0 | 0 | 8 | 0 | 8 | 8 | 4 | 8 | 0 | 8 | 0 | 4 | 0 | 0 |
| Использование непрерывных числовых признаков | 8 | 0 | 0 | 8 | 0 | 8 | 8 | 0 | 8 | 0 | 0 | 8 | 0 | 0 | 0 |
| ПОМОЩЬ В ВЫБОРЕ ПРИЗНАКОВ | | | | | | | | | | | | | | | |
| Отображения иерархии признаков | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 |
| Указание оптимальных признаков | 8 | 8 | 8 | 8 | 0 | 8 | 4 | 0 | 4 | 8 | 8 | 0 | 8 | 0 | 0 |
| Указание признаков для выделения таксона | 4 | 4 | 4 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| Оптимальные пути определения | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Удаление избыточных признаков | 2 | 0 | 0 | 2 | 0 | 2 | 2 | 2 | 0 | 2 | 2 | 2 | 2 | 0 | 0 |
| Удаление избыточных состояний признаков | | | | | | | | | | | | | | | |
| Оценка надежности признаков | 4 | 0 | 0 | 4 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 0 | 4 | 0 | 0 |
| Оценка надежности состояний признаков | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Составление списка признаков | 4 | 4 | 4 | 4 | 0 | 4 | 4 | 0 | 2 | 0 | 4 | 0 | 0 | 0 | 0 |
| ЗАПИСЬ И ОПРЕДЕЛЕНИЕ ЦЕННОСТИ ПРИЗНАКОВ | | | | | | | | | | | | | | | |
| Сохранение невыясненных признаков | 16 | 0 | 0 | 16 | 0 | 16 | 16 | 16 | 16 | 0 | 16 | 16 | 8 | 0 | 0 |
| Использование взаимосвязи признаков | 4 | 0 | 0 | 0 | 0 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Автоматический контроль признаков | 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Использование интервалов для числовых признаков | 4 | 0 | 0 | 0 | 0 | 4 | 4 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 0 |
| Возможность текстового заполнения признаков | 2 | 2 | 2 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| Специальные оценки ключей | 4 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Вероятностное определение | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Указание неприемлемых комбинаций состояний признаков | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 0 |
| Расширенные диапазоны числовых признаков | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Допустимость неизвестных состояний признаков | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Точные признаки | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Определение ценности признака | 2 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| ИСПОЛЬЗОВАНИЕ ГРУПП ПРИЗНАКОВ | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Поименованные наборы признаков | 4 | 0 | 4 | 4 | 0 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 |
| Общие наборы признаков | 4 | 0 | 4 | 4 | 0 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Конкретные наборы признаков | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Поименованные наборы (группы) таксонов | 4 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 |
| Общие группы таксонов | 4 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Конкретные группы таксонов | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ИНТЕРПРЕТАЦИЯ ПРИЗНАКОВ | | | | | | | | | | | | | | | |
| Комментарии к признакам | 4 | 0 | 4 | 4 | 0 | 4 | 4 | 0 | 0 | 2 | 0 | 0 | 4 | 0 | 0 |
| Глоссарии (словари признаков) | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Иллюстрации признаков | 4 | 4 | 4 | 4 | 0 | 4 | 2 | 0 | 4 | 2 | 4 | 0 | 4 | 0 | 0 |
| Выбор состояния признака по иллюстрации | 4 | 4 | 4 | 4 | 0 | 4 | 2 | 0 | 0 | 2 | 4 | 0 | 0 | 0 | 0 |
| ИЗОБРАЖЕНИЯ И ЗВУКИ | | | | | | | | | | | | | | | |
| Иллюстрации таксонов | 4 | 4 | 4 | 4 | 0 | 4 | 4 | 4 | 2 | 4 | 4 | 4 | 2 | 0 | 0 |
| Иллюстрации таксонов по группам | 4 | 0 | 4 | 4 | 0 | 4 | 4 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 |
| Произвольный показ иллюстраций | 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| Текст с иллюстрациями | 2 | 0 | 0 | 2 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 0 | 2 | 0 | 0 |
| Звуки | 2 | 0 | 2 | 2 | 0 | 2 | 2 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 |
| Видеоизображения | 2 | 0 | 2 | 2 | 0 | 2 | 2 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 |
| Определение без рисунков | 2 | 0 | 0 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| СОПРЯЖЕННЫЕ КЛЮЧИ | | | | | | | | | | | | | | | |
| Интегральные иерархические ключи | 2 | 0 | 0 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Самостоятельные иерархические ключи | 2 | 0 | 0 | 2 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ПОИСК ИНФОРМАЦИИ | | | | | | | | | | | | | | | |
| Поиск по названиям таксонов | 2 | 0 | 0 | 2 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Составление описаний таксонов | 2 | 0 | 2 | 2 | 0 | 2 | 2 | 0 | 2 | 2 | 2 | 2 | 2 | 0 | 0 |
| Различия между таксонами | 4 | 0 | 0 | 0 | 0 | 4 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| Сходство таксонов | 2 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Составление дифференциальных описаний | 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Поиск таксонов по сочетаниям признаков | 2 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Контроль правильности значений состояний признаков | 2 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Распределение признаков по ценности | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Наиболее сходные таксоны | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Текстовые файлы, прикрепленные к таксонам | 2 | 0 | 0 | 2 | 0 | 2 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| СОВМЕСТНОЕ ИСПОЛЬЗОВАНИЕ ДАННЫХ | | | | | | | | | | | | | | | |
| Импорт данных в формате DELTA | 2 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 |
| Экспорт данных в формате DELTA | 2 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 |
| Вывод данных | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Связь с составлением описаний таксонов | 2 | 0 | 0 | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 |
| Связь с генерацией одновходовых ключей | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 |
| Связь с классификацией | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 |
| УДОБСТВА ДЛЯ ПОЛЬЗОВАТЕЛЯ | | | | | | | | | | | | | | | |
| Интерактивная справка | 2 | 0 | 0 | 1 | 0 | 2 | 1 | 0 | 1 | 1 | 0 | 1 | 2 | 0 | 0 |
| Использование командных файлов или макросов | 2 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Использование инструментальной панели по выбору | 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Доступ к тексту программы | 2 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Папки с документацией | 2 | 2 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Неограниченный объем данных | 4 | 4 | 4 | 4 | 0 | 4 | 4 | 4 | 4 | 2 | 4 | 4 | 4 | 0 | 0 |
| Неограниченная длина полей | 2 | 2 | 2 | 2 | 0 | 2 | 1 | 0 | 2 | 0 | 2 | 2 | 2 | 0 | 0 |
| Отсутствие специальных требований к памяти компьютера | 2 | 0 | 0 | 2 | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 |
| Скорость ответа программы на действия пользователя | 4 | 0 | 0 | 4 | 0 | 4 | 0 | 4 | 0 | 4 | 4 | 0 | 4 | 0 | 0 |
| Возможность работы в Интернете | 8 | 8 | 4 | 8 | 0 | 4 | 0 | 0 | 4 | 0 | 4 | 2 | 0 | 0 | 0 |
| Отсутствие необходимости инсталляции программ | 2 | 2 | 0 | 2 | 0 | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 2 | 0 | 0 |

Существуют одновходовые (пользователю предлагается один признак) и многовходовые (предлагается несколько признаков для выбора наиболее надежного и удобного) ключи. Число состояний признака принципиально менее важное свойство. Существуют дихотомические (каждый признак имеет только 2 состояния) и политомические (признак включает 3 и более состояний) ключи. Практически важна способность оперировать изображениями признаков и состояний признаков; существуют ключи, управляемые образами (изображения признаков и их состояний - экранные кнопки управления диагнозом) и ключи, управляемые словесными формулировками (используются альтернативные текстовые формулировки). Изображения дают возможность мгновенно уяснить признак и верно выбрать его состояние; текстовая формулировка требует времени для понимания, сравнения, выбора соответствующего признака и состояния. Повышает возможности определительной программы способность оперировать количественными признаками; возможность фильтровать набор таксонов по диапазону значений количественного признака может существенно сократить путь диагноза [20].

В дальнейшем была предпринята попытка представить себе идеальный компьютерный определитель или «определитель-идеал», к которому нужно стремиться при разработке подобных систем. На основе сравнения всех свойств компьютерных определителей, созданных к настоящему времени как за рубежом, так и в СНГ, можно смоделировать оптимальный идентификационный ключ. Это интерактивный, многовходовый, политомический, управляемый изображениями ключ со следующими динамическими свойствами: ранжированием признаков на каждом шаге в зависимости от их диагностической ценности; возможностью видеть названия и изображения таксонов текущего набора, а также значения вероятности их идентичности определяемому объекту; возможностью возврата на один или несколько шагов диагноза для коррекции ошибок определяющего; возможностью выбрать несколько признаков в каждом шаге и отметить "невозможные" состояния признаков; активным оперированием как качественными, так и количественными признаками с возможностью использования для последних математических функций (диапазон минимум-максимум, среднее, формулы дискриминантного анализа). Позже прибавилась возможность интерактивного представления определителей или ключей в глобальной сети Интернет. По этим направлениям и ведутся поисковые работы, так как ни одна из существующих систем пока не обладает полным набором описанных свойств [10, 11, 15, 20, 23].

## Литература


[1] Алимов А.Ф., Смирнов И.С., Рысс А.Ю., Дианов М.Б., Лобанов А.Л., Голиков А.А. Современные биологические электронные публикации: коллекции, идентификационные системы и базы данных // Информационные и телекоммуникационные ресурсы в зоологии и ботанике. Труды 2-го международного симпозиума. 2001. С. 13-19.

[2] Армс В. Электронные библиотеки. (Перевод С.А.Арнаутова). ПИК ВИНИТИ, 2001. - 275 с.

[3] Дианов М.Б., Лобанов А.Л. PICKEY - Программа для определения организмов с интерактивным использованием изображений // В: С.Степаньянц, А.Лобанов, М.Дианов, ред., Базы данных и компьютерная графика в зоологических исследованиях // Труды Зоологического института РАН, 1997. Т. 269. С. 35-39.

[4] Каталог фондовой коллекции Зоологического института РАН. Класс Костистые рыбы (Osteichthyes). Отряд Камбалообразные (Pleuronectiformes). Сост. Воронина Е.П., Волкова Г.А. Исследования фауны морей. Т. 55(63). – СПб., 2003. 198 с.

[5] Каталог фондовой коллекции Зоологического института РАН. Класс Костистые рыбы (Osteichthyes). Отряд Скорпенообразные (Scorpaeniformes). Подотряд Cottoidei. Часть I. Сиделева В.Г., Неелов А.В., Воронина Е.М., Волкова ГА. Исследования фауны морей. Т. 57(65). - СПб., 2006. С. 1-223.

[6] Каталог фондовой коллекции Зоологического института РАН. Класс Костистые рыбы (Osteichthyes). Отряд Скорпенообразные (Scorpaeniformes). Подотряд Cottoidei. Часть II. Сиделева В.Т., Неелов А.В., Воронина Е.П., Волкова Г.А. Исследования фауны морей. Т. 57(65). - СПб., 2006. С. 225-349.

[7] Каталог фондовой коллекции Зоологического института РАН. Класс Костистые рыбы (Osteichtbyes). Отряд скорпенообразные (Scorpaeniformes). Подотряды Scorpaenoidei, Congiopodoidei, Platycephaloidei, Anoplopomatoidei, Hexagrammoidei, Scorpaenoidei. Воронина Е.П., Волкова Г.А. Исследования фауны морей. Т. 58(66). - СПб., 2007.189 с.

[8] Кривохатский В.А., Лобанов А.Л., Медведев Г.С., Белокобыльский С.А., Дианов М.Б., Смирнов И.С., Халиков Р.Г. Информационная система по энтомологическим коллекциям в Интернете // Труды Русского энтомологического общества, Т. 74, СПб., 2003: С. 59-70.

[9] Кутикова Л.А., Николаева И.П. Каталог видов коловраток (Rotifera) пресных вод Севера-Запада России // [Электрон. ресурс]. СПб, ЗИН РАН, 2002. (Рус.).

http://www.zin.ru/books/rotcatalog/default.asp [22 ноября 2006]

[10] Лобанов А.Л. Проблемы создания единой системы диагностической информации в биологии // Единая система и информационно-поисковых языков. Тезисы докладов Всесоюзной научной конференции. Юрмала, 6-8 сентября 1977 г. 1977. С. 84-87.

[11] Лобанов А.Л. Принципы построения определителей насекомых с использованием электронных вычислительных машин. - Автореферат диссертации на соискание ученой степени канд. биол. наук. Л.: ЗИН АН СССР, 1983. С. 1-19.

[12] Лобанов А.Л. Диалоговые компьютерные биологические диагностические системы BIKEY5 и BIKEY6 // В: Степаньянц, Лобанов, Дианов, ред., Базы данных и компьютерная графика в зоологических исследованиях // Труды Зоологического института РАН, Т. 269. 1997а. С. 61-65.

[13] Лобанов А.Л. Компьютерные определители в биологии: результаты 30-летней эволюции // Компьютерные базы данных в ботанических исследованиях. Сборник научных трудов. 1997б. С. 51-55.

[14] Лобанов А.Л. Компьютерные определители животных и растений: современное состояние и перспективы // В: Рысс, Смирнов, ред., Информационно-поисковые системы в зоологии и ботанике. Тезисы международного симпозиума, май 1999 // Труды Зоологического института РАН. Т. 278. 1999. С. 79-80.

[15] Лобанов А.Л., Дианов М.Б. Компьютерная диагностическая система BIKEY и возможности ее использования в защите растений // Защита растений в условиях реформирования агропромышленного комплекса: экономика, эффективность, экологичность. Всеросс. съезд по защите растений. Тезисы докл. 1995. С. 548-549.

[16] Лобанов А.Л., Дианов М.Б. Мир жуков ("Wir bestimmen Kaefer") - CD-ROM и краткое руководство. 1996. - Berlin: dialobis edition.

[17] Лобанов А.Л., Дианов М.Б. CD-ROM: новый инструмент изучения биологического разнообразия // Компьютерные базы данных в ботанических исследованиях. Сборник научных трудов. 1997. С. 55-57.

[18] Лобанов А.Л., Дианов М.Б. Комплекс программ для создания компьютерных зоологических монографий на компакт-дисках // Отчетная научная сессия по итогам работ 1997 г. Тезисы докладов. 1998. С. 27-28.

[19] Лобанов А.Л., Дианов М.Б. Средства мультимедиа в электронных зоологических и ботанических публикациях // Информационно-поисковые системы в зоологии и ботанике (Тезисы международного симпозиума, май 1999). Труды Зоологического института РАН. 1999. Vol. 278. P. 100.

[20] Лобанов А.Л., Рысс А.Ю. Компьютерные идентификационные системы в зоологии и ботанике: современное состояние и перспективы // Информационно-поисковые системы в зоологии и ботанике (Тезисы международного симпозиума, май 1999). Труды Зоологического института РАН. 1999. Vol. 278. P. 17-29.

[21] Лобанов А.Л., Смирнов И.С. Место и роль информационных технологий в исследованиях Зоологического института РАН // Фундаментальные зоологические исследования: Теория и методы. (По материалам Международной конференции «Юбилейные чтения, посвященные 170-летию Зоологического института РАН», 23-25 октября 2002 г.), М.-СПб.: Товарищество научных изданий КМК. 2004: 283-318 (резюме на англ. яз.).

[22] Лобанов А.Л., Дианов М.Б., Рысс А.Ю. Современные типы биологических электронных публикаций: CD-ROM и Internet // Информационно-поисковые системы в зоологии и ботанике (Тезисы международного симпозиума, май 1999). Труды Зоологического института РАН. 1999а. 278. С. 39-44.

[23] Лобанов А.Л., Кирейчук А.Г., Смирнов И.С., Дианов М.Б., Граничин О.Н.. Интернет и интерактивные определители биологических объектов // Научный сервис в сети ИНТЕРНЕТ: Труды Всероссийской научной конференции (19-24 сентября 2005 г., г. Новороссийск). – М.: Изд-во МГУ, 2005. с. 132-134.

[24] Смирнов И.С., Лобанов А.Л. Компьютерный определитель по офиурам как база данных для хранения таксономической информации // Бюллетень Московского общества испытателей природы (МОИП). Отд. геологии. Т. 72, Вып. 1. 1999. С. 87-88.

[25] Смирнов И.С., Рысс А.Ю. Биологические коллекции и базы данных // Рысс А.Ю., Смирнов И.С. (ред.). Информационно-поисковые системы в зоологии и ботанике // Труды Зоологического института РАН, Т. 278, Санкт-Петербург. 1999. С. 30-38.

[26] Смирнов И.С., Лобанов А.Л., Дианов М.Б. Зоологические виртуальные музеи // Научный сервис в сети Интернет. Тезисы докладов Всероссийской научной конференции (20-25 сентября 1999 г., г. Новороссийск), Изд-во Моск. ун-та, 1999а. С. 185-187.

[27] Смирнов И.С., Лобанов А.Л., Голиков А.А.., Дианов М.Б. Электронные зоологические коллекции // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Первой Всероссийской научной конференции (19-22 октября 1999 г., г. С.-Петербург), Изд-во Санкт-Петербургского ун-та, 1999б. С. 236-240.

[28] Смирнов И.С., Лобанов А.Л., Алимов А.Ф., Голиков А.А.. От электронных коллекций к

виртуальным коллективам зоологов в сети Интернет // Интернет и современное общество. Тезисы докладов II Всероссийской научно-методической конференции, (29 ноября-3 декабря 1999 г., г. Санкт-Петербург), Санкт-Петербург, 1999в. С 61-62.

[29] Смирнов И.С., Лобанов А.Л., Дианов М.Б., Голиков А.А., Алимов А.Ф. Зоологические виртуальные музеи: настоящее и будущее // Научный сервис в сети Интернет: Труды Всероссийской научной конференции (24-29 сентября 2001 г., г. Новороссийск). – М.: Изд-во МГУ, 2001. С. 22-24.

[30] Смирнов И.С., Лобанов А.Л., Дианов М.Б., Голиков А.А., Алимов А.Ф., Неелов А.В., Гаврило М.В. Создание информационно-поисковой системы по экологии бентоса и птиц Антарктики (ECOANT) на основе электронной коллекции беспозвоночных, рыб и птиц. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Сборник докладов Третьей Всероссийской конференции RCDL'2001. Петрозаводск, 11-13 сентября 2001 г. – Карельский научный центр РАН, 2001. С. 197-198.

[31] Смирнов И.С., Лобанов А.Л., Алимов А.Ф., Кривохатский В.А. Электронные коллекции Зоологического института РАН. Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Пятой Всероссийской научной конференции RCDL'2003, (Санкт-Петербург, 29-31 октября 2003 г.): – Санкт-Петербург: НИИ Химии СПбГУ, 2003: 275-278.

[32] Смирнов И.С., Лобанов А.Л., Алимов А.Ф., Пугачев О.Н., Кривохатский В.А.. Информационная система по биологическому разнообразию России // Научный сервис в сети ИНТЕРНЕТ: Труды Всероссийской научной конференции (22-27 сентября 2003 г., г. Новороссийск). – М.: Изд-во МГУ, 2003. С. 12-14.

[33] Смирнов И.С., Воронина Е.П., Лобанов А.Л., Голиков А.А., Неелов А.В. Создание информационно-поисковых систем по коллекциям морских животных (рыб и беспозвоночных) в Зоологическом институте РАН // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Шестой Всероссийской научной конференции RCDL'2004, (Пущино, 29 сентября - 1 октября 2004 г.): – Москва, типография ООО «Мультипринт», 2004: 30-33.

[34] Смирнов И.С., Лобанов А.Л., Пугачев О.Н., Алимов А.Ф., Воронина Е.П. Электронные коллекции в зоологии и электронные библиотеки // Электронные библиотеки, 9 (4). 2006 http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2006/part4/SLPAV

[35] Smirnov I.S., Voronina E.P., Lobanov A.L., Neyelov A.V. The information system of the marine animals collection (fish and invertebrates) in the Zoological Institute Russian Academy of Sciences // Ocean Biodiversity Informatics. International Conference on Marine Biodiversity Data Management. Hamburg, Germany: 29/11-1/12/2004, 2004. p. 27.

[36] Ocean Biodiversity Informatics International Conference on Marine Biodiversity Data Management Hamburg, Germany: 29 November to 1 December 2004. (Engl.). http://www.vliz.be/obi/ [22 November 2006]

## Digital animal collection and interactive keys of biological objects

A.G.Kireitchuk, A.L.Lobanov, I.S.Smirnov, A.T.Vakhitov, E.P.Voronina, O.N.Pugachev

Elaboration of digital collections, identification keys and ways of ecological monitoring requires more intensive usage of more complex computer technologies. Increasing necessity of estimation and preservation of a biodiversity under anthropogenic action for last years makes this problem particularly important. Wide expansion of the Windows operational system and Internet makes also actual a translation of programs of input and work with keys on this environment, and also data presentation of systems on the Internet for more efficient use of available resources of management and information. Identification keys allow to structure diverse information and to fulfil a fast-access retrieval of it in different sources.

♣ Работа по теме осуществляется частично при поддержке грантов РФФИ N 09-04-00789а и 07-04-00514а, проекта N11 «Исследование Антарктики. Проведение комплексного изучения антарктической биоты», Федеральной Программы «Мировой Океан» и программы «Биоразнообразие».

# Информационное обеспечение вибросейсмического мониторинга ♣

А.П.Григорюк, Л.П.Брагинская

Институт вычислительной математики и математической геофизики СО РАН
ludmila@opg.sscc.ru

**Аннотация**

В работе предложены концептуальные подходы к методике построения информационно-вычислительных систем экспериментальных данных, содержащих неструктурированные данные в виде длинных числовых массивов. Далее рассмотрены функции веб-ориентированной информационно-вычислительная системы (ИВС) «Вибросейсмическое просвечивание Земли» (*http://opg.sscc.ru/db)*, разработанной авторами на основе предложенной ими концепции. Система включает файловый архив данных, полученных в ходе уникальных экспериментов по вибросейсмическому просвечиванию Земли с помощью мощных 40- и 100-тонных сейсмических вибраторов. Пользователи ИВС имеют возможность с помощью стандартного веб-браузера проводить поиск в базе данных по различным параметрам вибросейсмического просвечивания (около 20 параметров), просматривать найденные файлы волновых форм, осуществлять интерактивный анализ данных по классическим и специальным алгоритмам с выводом результатов непосредственно в веб-браузер.

Картографическая подсистема, выполненная на базе сервиса Google Maps, позволяет пользователям работать с интерактивными картами и спутниковыми снимками районов проведения экспериментов.

## 1 Активный вибросейсмический мониторинг

На протяжении 1995–2008 годов институтами Сибирского отделения Российской академии наук (СО РАН) совместно с другими отечественными и зарубежными научными учреждениям проводились экспериментальные работы по активному вибросейсмическому мониторингу литосферы в сейсмоопасных районах России. Работы проводились в Алтае-Саянской сейсмоактивной зоне, Байкальской рифтовой зоне, Таманской грязе-вулканической провинции [1].

Активный вибросейсмический мониторинг относится к числу новых геофизических технологий, включающих методы наблюдения за состоянием земной коры по изменению характеристик распространения сейсмических волн.

В сейсмологии основным источником волн является землетрясение — природный процесс, не управляемый ни по времени, ни по месту возникновения, ни по энергетике. С применением мощных сейсмических вибраторов были разработаны новые геотехнологии, которые позволяют избежать ряда ограничительных обстоятельств сейсмологии землетрясений и больших взрывов. В то же время, несколько десятков минут работы 100-тонного сейсмического вибратора по энергетической эффективности эквивалентны среднему землетрясению. Вибрационные геотехнологии имеют следующие преимущества [2]:

- точно определенные координаты источника и времени его работы;
- повторяемость эксперимента;
- возможность возбуждения колебаний с заданными параметрами;
- возможность автоматизации управления экспериментом;
- повсеместность применения;
- экологическая безопасность, т.к. регистрируемый сигнал находится под микросейсмами, а необходимые соотношения сигнал/шум обеспечиваются накоплением.

В экспериментальных работах по вибросейсмическому мониторингу использовались мощные управляемые вибраторы, расположенные на Быстровском (Новосибирская обл.), Байкальском

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

и Краснодарском сейсмологических полигонах. Регистрация осуществлялась специальными мобильными регистраторами и региональными сейсмостанциями в радиусе до 500 км от источников. В ходе некоторых экспериментов регистрировались также промышленные взрывы на шахтах и разрезах Кузбасса и на Семипалатинском ядерном полигоне.

Всего в ходе полевых работ было проведено более 33 экспериментов, в результате которых зарегистрировано около 30000 сейсмотрасс. Общий объем накопленного архива файлов волновых форм и сопутствующей табличной информации (тип сейсмического источника, параметры излучаемого им сигнала, параметры регистратора, географические координаты источника и регистратора и т.д.) составляет около 20 ГБ

Полученный экспериментальный материал может быть использован при решении следующих задач геофизики [1]:

- развитие нового метода активной сейсмологии и геофизических технологий с использованием мощных вибрационных источников сейсмических волн;
- экспериментальные и теоретические исследования по сейсмическому зондированию Земли с целью изучения деформационных процессов в коре и верхней мантии;
- создание методики вибросейсмического мониторинга сейсмоопасных зон с целью прогноза землетрясений;
- развитие методов обработки и интерпретации вибросейсмических данных;
- математическое моделирование в задачах разведочной геофизики и прогноза землетрясений;
- разработка численных методов решения прямых и обратных задач геофизики, включая комбинированные обратные задачи;
- разработка новых численно-аналитических методов решения многомерных прямых задач и их приложений в исследованиях различных аспектов сейсмологии и сейсморазведки.

## 2 Модель данных

Для обеспечения доступа широкого круга исследователей к накопленному экспериментальному материалу было решено создать базу данных и информационно-вычислительную систему (ИВС) с возможностью эффективного поиска и интерактивного анализа данных. Организация структурированных табличных (реляционных) данных обычно затруднений не вызывает. Для этого идеально подходят реляционная модель и язык SQL, лежащие в основе современных СУБД. Сложнее обстоит дело с файлами волновых форм, представляющими собой n-мерные, в общем случае, числовые массивы, которые не могут быть структурированы и поэтому не поддерживаются реляционными СУБД непосредственно.

В настоящее время для работы одновременно с реляционными и нереляционными данными, в основном используют одну из двух архитектур:

- как реляционные, так и нереляционные данные находятся в базе данных;
- реляционные данные находятся в базе данных, а нереляционные данные – в файловых системах или на файловых серверах.

Каждый из этих двух подходов имеет свои преимущества и недостатки [8]. В первом случае одна база данных становится удобным централизованным хранилищем для обоих типов данных. Однако нереляционные данные хранятся в формате больших двоичных объектов (BLOB), скорость доступа к этим объектам существенно уступает скорости доступа к файлам. Во втором случае обеспечивается высокая скорость доступа, но усложняется разработка приложений и управление ими, так как приложения должны поддерживать согласованность между записями в базе данных и файлами, связанными с этими записями. Данную проблему можно частично или полностью решить за счет модели данных, обеспечивающей эффективную индексацию файловой системы из базы данных.

При построении концептуальной модели предметной области мы исходили из того, что экспериментально изучаемому объекту может быть приписан определенный набор параметров, соответствующий представлениям исследователей о состоянии и поведении этого объекта. Параметры это то, что можно измерять, наблюдать и изменять в процессе исследований. В процессе экспериментов на изучаемый объект воздействуют некоторые факторы с контролируемыми параметрами и, с помощью сенсоров, регистрируется ряд параметров объекта при фиксированных других параметрах. Тогда концептуальную модель эксперимента можно представить с помощью приведенной на Рис.1 ER-диаграммы (диаграмма *«сущность-связь»*) [6].


Рис. 1. Концептуальная модель эксперимента

Диаграмма содержит три класса сущностей: ОБЪЕКТ, ФАКТОР и СЕНСОР. Каждый класс с набором атрибутов, определяемых конкретной областью исследований. Взаимоотношения сущностей выражаются двумя классами связей: ВОЗДЕЙСТВИЕ и ОТКЛИК. В случае пассивного эксперимента или наблюдения класс сущностей ФАКТОР может отсутствовать.

Для перехода к реляционной модели данных заменим сущности и связи ER-диаграммы на соответствующие отношения $R$ с первичными ключами $K$ и атрибутами $A$:

ОБЪЕКТ – $R1(\underline{K_1}, A_{11}, A_{12}, \ldots, A_{1n})$;

ФАКТОР – $R2(\underline{K_2}, A_{21}, A_{22}, \ldots, A_{2n})$;

СЕНСОР – $R3(\underline{K_3}, A_{31}, A_{32}, \ldots, A_{3n})$;

ВОЗДЕЙСТВИЕ – $R4(K_1, K_2)$;

ОТКЛИК – $R5(K_1, K_3)$.

Вспомогательные отношения *R4* и *R5* служат для организации связи типа M:N (многие-ко-многим) между отношениями *R1, R2* и *R1, R3* соответственно. Первичными ключами *K* могут служить, например, порядковые номера кортежей соответствующих отношений. В общем случае ключевые атрибуты должны содержать значения из конечных множеств *P:*

$P_1=\{ K_{11}, K_{12}, \ldots, K_{1m} \}$;
$P_2=\{ K_{21}, K_{22}, \ldots, K_{2m} \}$;
$P_3=\{ K_{31}, K_{32}, \ldots, K_{3m} \}$.

Такая модель позволяет организовать адресацию файлового архива, имеющего следующую иерархическая структуру:

$$/ data / P_1 / P_2 / P_3 / P_1 P_2 P_3 N \qquad (1).$$

где строка «$P_1 P_2 P_3 N$» — имя файла данных, образованное конкатенацией атрибутов $P_1, P_2, P_3$ и номера канала сенсора *N* для многоканальных сенсоров. Данная структура соответствует естественной древовидной структуре файловой системы.

Предложенная модель данных в сочетании со способом адресации неструктурированных данных обеспечивает естественную однозначную связь между записями в базе данных и соответствующими файлами. В то же время пользователи могут полностью абстрагироваться от имен или шаблонов имен файлов и каталогов, работая только с атрибутами, каталогизирующими свойства и происхождение каждого файла.

Как и модель эксперимента, модель данных является обобщенной, ее необходимо адаптировать для каждой конкретной научной области и вида экспериментов. В большинстве случаев может потребоваться декомпозиция отношений ОБЪЕКТ, ФАКТОР и СЕНСОР с учетом функциональных зависимостей между атрибутами.

Каждый из экспериментов по вибросейсмическому просвечиванию Земли проводится при некоторой фиксированной расстановке задействованных в эксперименте сейсмических вибраторов и регистраторов. В ходе эксперимента проводятся сеансы работы виброисточников с различными видами и параметрами излучаемых сигналов. Каждому сеансу соответствует один кортеж отношения ФАКТОР и несколько кортежей (по числу регистраторов) отношения СЕНСОР. Параметры изучаемого объекта, которым является некоторая область литосферы и верхней мантии, определяются географическими координатами сейсмических источников и регистраторов. Эти координаты представлены соответствующими атрибутами отношений ФАКТОР и СЕНСОР. Поэтому отношение ОБЪЕКТ в данной системе содержит только порядковый номер эксперимента в качестве ключевого атрибута, а также несколько атрибутов, представляющих вспомогательные метаданные.

## 3 Основные функции ИВС «Вибросейсмическое просвечивание Земли».

На основе предложенной модели данных нами была разработана база данных и веб-ориентированная информационно-вычислительная система «Вибросейсмическое просвечивание Земли» (*http://opg.sscc.ru/db*) [4]. ИВС реализует следующие основные функции:

- получение подробной информации по любому из проведенных экспериментов (метаданные);
- поиск в базе данных одновременно по 18 параметрам вибропросвечивания (типы источников, вид и параметры сигналов, географические координаты и др.) поисковая форма системы показана на Рис. 2;
- интерактивный on-line анализ (корреляционный, спектральный, спектрально-временной и т.д.) найденных сейсмотрасс с отображением результатов непосредственно в веб-браузере пользователя;
- построение по результатам поиска интерактивных карт и спутниковых снимков с обозначенными источниками и регистраторами сейсмических волн.

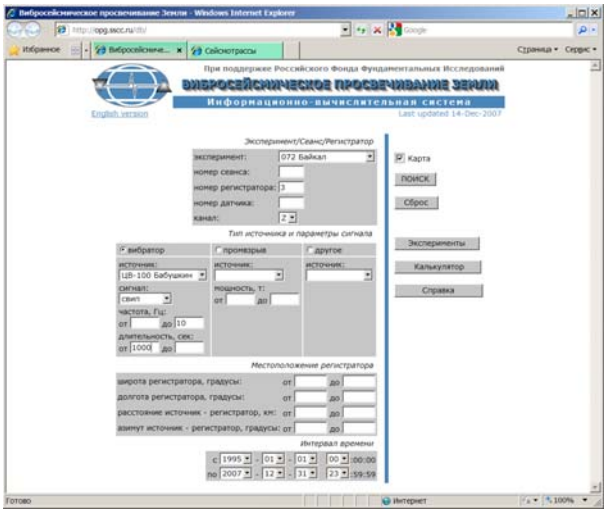



Рис. 2. Поисковая форма ИВС

## 4 Структурная схема ИВС

Структурная схема информационно-вычислительной системы приведена на Рис. 3. Пользователи взаимодействуют с системой с помощью стандартного веб-браузера, посылая запросы на поиск и анализ данных. В запросе на поиск указываются интересующие пользователя параметры объекта, параметры воздействующих на объект факторов и параметры сенсоров, регистрирующих данные. Запрос на анализ должен содержать перечень процедур анализа, которые будут применены к найденным данным и параметры этих процедур.

В результате выполнения запроса на поиск из базы данных извлекаются необходимые для обращения к файловому архиву атрибутивные данные. На основе этих данных веб-приложение формирует адреса файлов в соответствии с (1) и передает их модулю анализа.

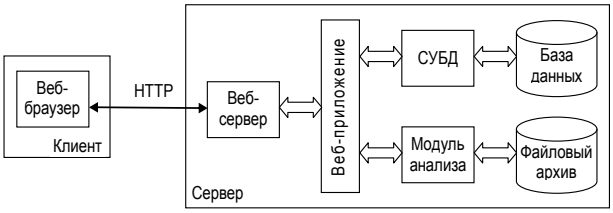



Рис. 3. Структурная схема ИВС

Модуль анализа представляет собой приложение, выполняющее анализ данных в соответствии с алгоритмами, применяемыми в конкретной области экспериментальных исследований. В большинстве случаев это классические математико-статистические процедуры анализа числовых последовательностей. Соответствующие вычисления выполняются современными многоядерными и мультипроцессорными системами с быстродействием, достаточным для обеспечения online режима при работе с массивами данных практически любых размеров. Полученные в результате анализа числовые массивы возвращаются веб-приложению, которое «на лету» формирует графики, таблицы, текст и отправляет всё это пользователю в виде готовой веб-страницы. Конечное представление информации реализуется при помощи клиентских технологий JavaScript, HTML и стилевых таблиц CSS.

Конкретная аппаратно-программная реализация структурной схемы Рис. 2 определяется масштабами системы, сложностью применяемых алгоритмов анализа, количеством пользователей и т.д. Так, ИВС «Вибросейсмическое просвечивание Земли» в настоящее время базируется на единой аппаратной платформе на основе двух процессоров Intel Xeon 3.0GHz, оперативная память DDR2 2GB. Использовалось только свободно распространяемое программное обеспечение: операционная система Linux, веб-сервер Apache, СУБД MySQL. Веб-приложение написано на языке PHP, этот выбор обусловлен широким набором графических функций и высокой скоростью генерации изображений. Для повышения быстродействия системы модуль анализа был реализован на языке C++ с использованием программных библиотек Intel Performance Libraries [3].

Пример веб-страницы с результатами поиска и анализа данных приведен на Рис. 4.

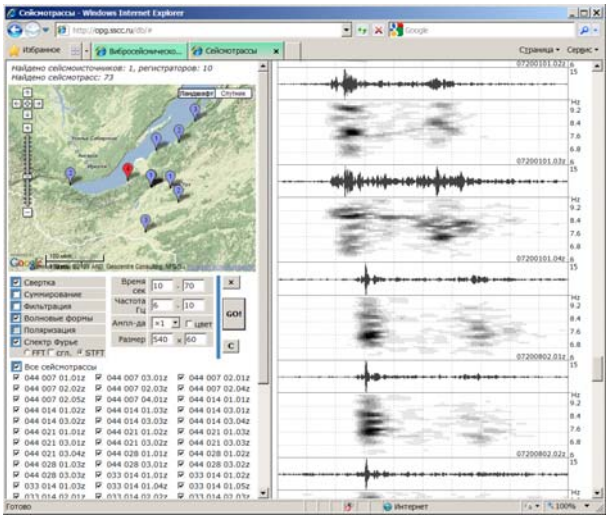


Рис. 4. Отображение результирующей страницы в веб-браузере

## 5 Использование сервиса Google Maps для построения карт эксперимента

В геофизике, как и во многих других областях, исследователи имеют дело с пространственно обусловленными данными или геопространственными данными. Поэтому архитектура ИВС должна предусматривать подсистему управления геоданными и картографическую подсистему.

Большинство современных СУБД, как коммерческих, так и свободно распространяемых, поддерживают класс пространственных данных непосредственно или с помощью специальных расширений. Картографические сервисы, в частности веб-сервисы, до недавнего времени строились преимущественно на основе специализированного серверного ПО, позволяющего публиковать в сети Интернет карты, сопровождаемые базовым ГИС-инструментарием.

Однако в последние годы в Интернете все большее распространение получают гибридные ГИС. В таких системах геоданные из прикладной базы данных интегрируются с картографическим сервисом, предоставляемым специализированным веб-сервером. На сегодняшний день наиболее развитым картографическим веб-сервисом является Google Maps компании Google [5, 9]. Сервис базируется на данных дистанционного зондирования (спектрозональные снимки со

спутников Landsat, SPOT, Quickbird с разрешением до 0.68м) совмещенных с топографическими картами в проекции Меркатора. Компания Google предоставляет пользователям интерфейс Google Maps API в виде классов объектов JavaScript для генерации карт и нанесения на них собственных маркеров, контуров, а также готовых слоев в формате KML. Данные для отображения могут находиться как непосредственно в коде веб-страниц, так и во внешних XML и KML файлах. Схема взаимодействия сервера ИВС, сервера Google Maps и клиентского браузера показана на Рис. 5.

Использование технологии AJAX обеспечивает обновление содержимого результирующей веб-страницы без ее перезагрузки, таким образом, изменение масштаба и перемещение по карте осуществляется без каких-либо задержек. Пример карты можно видеть на Рис. 4.

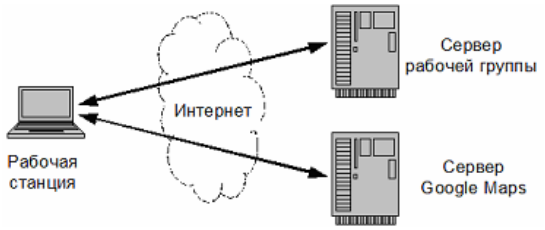



Рис. 5. Структура гибридной ГИС

## 6. Сравнение с зарубежными аналогами

Наиболее близким аналогом созданной ИВС, который нам удалось найти в Интернете, является система «Сейсмический монитор», доступная через портал IRIS (Incorporated Research Institutions for Seismology) [10]. Система позволяет в режиме on-line осуществлять поиск и просмотр сейсмограмм и метаданных землетрясений, зарегистрированных любой из 150 сейсмостанций Глобальной сети (Global Seismographic Network - GSN). Пользователи могут работать с системой через специальный интерфейс WILBER с помощью стандартного веб-браузера.

«Сейсмический монитор» обеспечивает доступ к громадному объему данных, которые постоянно пополняются. Однако по функциональности система значительно уступает разработанной нами ИВС. Так, в системе отсутствуют функции анализа волновых форм (фильтрация, спектральный анализ и т.д.), возможен только их непосредственный просмотр. Картографическая подсистема представлена только мелкомасштабной схематической картой мира. Не представлены данные дистанционного зондирования, которые позволили бы наблюдать особенности геологического строения районов землетрясений.

## 7. Заключение

В работе представлены концептуальные основы, архитектура и программное обеспечение информационно-вычислительной системы для поддержки экспериментов, проводимых научными коллективами, состоящими из специалистов, территориально удаленных друг от друга. Разработанное инвариантное ядро системы обеспечивает эффективное управление неструктурированными данными, получаемыми в ходе экспериментов от приборов или компьютерных моделей. Практическая реализация показана на примере ИВС «Вибросейсмическое просвечивание Земли».

На основе представленного инвариантного ядра была также разработана информационно-вычислительная система «Землетрясения Камчатки» (*http://opg.sscc.ru/kg/*) [7]. Система предназначена для информационной поддержки теоретических и прикладных исследований в области сейсмологии, вулканологии, физики землетрясений.

## Литература

[1] Активная сейсмология с мощными вибрационными источниками / Отв.ред. Г.М. Цибульчик. –Новосибирск: ИВМиМГ СО РАН, Филиал «Гео» Издательства СО РАН, 2004.

[2] А.С. Алексеев и др. Вибрационные геотехнологии на пороге 21 века.// Сейсмология в Сибири на рубеже тысячелетий. Материалы международной геофизической конференции, Новосибирск, ОИГГиМ, 2000 г.

[3] Библиотеки Intel Performance Libraries. http://www.intel.com/cd/software/products/emea/rus/perflib/358868.htm

[4] Григорюк А.П., Брагинская Л.П. Управление данными вибросейсмического мониторинга. // Мониторинг окружающей среды, геоэкология, дистанционные методы зондирования Земли и фотограмметрия. Сб. материалов междунар. науч. конгресса «ГЕО-Сибирь-2007» Т.3. – Новосибирск: СГГА, 2007.

[5] Григорюк А.П., Брагинская Л.П. Опыт веб-картографирования на основе сервиса Google Maps // Мониторинг окружающей среды, геоэкология, дистанционные методы зондирования Земли и фотограмметрия. Сб. междунар. Науч. Конгресса ГЕО-Сибирь-2008. Т.3. –Новосибирск: СГГА,2008.

[6] Дейт К. Д. Введение в системы баз данных. 7-е изд. – М.: Вильямс, 2001. – 702 с.

[7] Чебров В.Н., Григорюк А.П., Пантюхин Е.А., Брагинская Л.П. и др. Информационно-вычислительная система «Землетрясения Камчатки», доступная в сети Интернет. // Геофизический мониторинг и проблемы сейсмической безопасности Дальнего Востока России Труды региональной научно-технической конференции. Петропавловск-

Камчатский. 11-17 ноября 2007 г. Петропавловск-Камчатский: ГС РАН, 2008. 238 с.
[8] Электронная документация по SQL Server 2008. Управление неструктурированными данными. http://msdb.ru/Downloads/SQL2008/white_papers/UnstructuredData%20RU.docx
[9] Google Maps API Documentation. http://www.google.com/apis/maps/documentation
[10] Seismic Monitor Documentation. http://www.iris.edu/seismon/html/Help.html

## Information support of vibroseismic monitoring

A.P. Grigoruk, L.P. Braginskaya

In this paper we propose the conceptual approach to constructing informational computational systems of experimental geodata. The web-oriented experiment information system has been created. The system based on a data obtained in unique vibroseismic Earth sounding experiments. These experiments were conducted during 1995-2008 at Altay and Sayan regions, Baikal rift zone and Taman mud volcano province. Most of data is obtained using powerful eccentric 40- and 100-ton seismic vibrators.

Structured information about 33 experiments has been brought into the database. The system file archive contains over 30000 seismic traces obtained in these experiments. Users can search the database by various parameters (about 20) of the vibroseismic sounding, retrieve, view, and analyze the traces have been found. The cartographic system constructed on Google Maps service allows users to work with interactive maps and satellite images of areas of carrying out of experiments.

♣ Работа выполнена при поддержке грантов РФФИ № 05-07-90081, № 07-07-00106, 09-07-00515

# Наукометрическое исследование развития работ по малоактивируемым материалам для термоядерного реактора

© Аленина М.В., Колотов В.П.

Институт геохимии и аналитической химии им. В.И. Вернадского РАН
alenina@geokhi.ru

**Аннотация**

Проведен наукометрический анализ информационного потока в области разработки малоактивируемых материалов (МАМ). Анализ показал, что наиболее важные направления в области МАМ, дающие основной вклад в увеличение информационного потока, связаны с изучением свойств МАМ, оценками активации, технологиями МАМ, микроструктурными исследованиями материалов и конструкционными вопросами разработки термоядерного реактора. Не менее 80% публикаций в ведущих журналах по этой теме цитируется. Период удвоения информации составляет 11,5 лет. Сделан прогноз о дальнейшем развитии данной области материаловедения.

## 1 Введение

Одной из главных задач наукометрии является изучение состояния развития науки и техники в той или иной области путем анализа потоков научной информации. Некоторые авторы [2-3, 6] предлагают для оценки развития какого-нибудь направления в науке (а также прогнозирования трендов его дальнейшего развития) использовать многоаспектный анализ информации, т.к. считают, что одного-двух количественных параметров возможно не достаточно. В данной работе представлена оценка одного из направлений современного материаловедения, связанного с разработкой экологически безопасных конструкционных материалов, отличающихся ускоренным спадом наведенной радиоактивности или малоактивируемых материалов (МАМ) для термоядерного реактора, международный проект по строительству которого сейчас активно развивается.

## 2 Результаты и обсуждение

Наукометрический анализ информации по МАМ проводили в два этапа. Первый этап включал сбор тематической информации путем запроса по ключевым словам к реферативной базе научных публикаций SCOPUS [5] и дальнейший её анализ. Хронологический охват с 1960г. по настоящее время. Учитывая, что наиболее весомые источники подробной информации — это статьи в периодических журналах [7], поиск был ограничен тремя ведущими журналами, публикующими основные работы по исследуемой теме:

- ✓ Journal of Nuclear Materials;
- ✓ Fusion Engineering and Design;
- ✓ Fusion Technology.

Поисковая система по запросу «(low activation materials) OR (reduced activation materials)» дала 525 публикаций. Экспертная оценка рефератов публикаций показала, что 2% полученной информации не относится к теме по МАМ. Столь малая величина случайно попавшей непрофильной информации объясняется корректным выбором тематических журналов. Полученные данные были загружены в таблицу базы данных MS Access для последующей обработки.

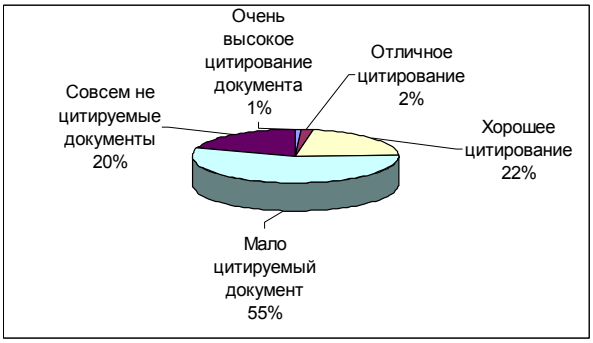



Рис.1 Распределение публикаций по категориям цитируемости.

Одно из достоинств базы данных SCOPUS – это наличие информации по цитированию статей. Согласно работе [2] все публикации могут быть разделены по категориям цитируемости. На Рис.1

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

представлено распределение полученных публикаций по категориям цитируемости (где $k$ – количество цитирований, приходящихся на статью):

- совсем не цитируемые статьи ($k$=0);
- мало цитируемые статьи ($1 \le k < 10$);
- хорошее цитирование ($10 \le k < 40$);
- отличное цитирование ($40 \le k < 90$);
- очень высокое цитирование статьи ($k \ge 90$).

Из Рис.1 видно, что 80% опубликованной информации данного направления материаловедения востребовано и, следовательно, наблюдается заметный обмен идеями.

Из общего информационного потока по МАМ также исключены статьи 2009 г. (т.к. год ещё не завершен и полной информации о нем нет). В Табл.1 приведена суммарная статистика по статьям, авторам и авторству в области разработки МАМ за период 1981-2008гг. В данном случае под авторством понимают общую сумму имён авторов за рассматриваемый период времени. Например, в 2001 году Иванов опубликовал 2 статьи, Петров – 3 статьи. В этом примере количество авторов равно 2, а авторство – 5.

Таблица 1 Суммарная статистика по статьям, авторам и авторству в области разработки МАМ.

| Годы | Кол-во статей | Кол-во авторов | Авторство |
|---|---|---|---|
| 1981 | 2 | 9 | 9 |
| 1982 | 1 | 2 | 2 |
| 1983 | 1 | 8 | 8 |
| 1984 | 10 | 47 | 47 |
| 1985 | 15 | 48 | 51 |
| 1986 | 17 | 52 | 53 |
| 1987 | 2 | 8 | 8 |
| 1988 | 14 | 39 | 44 |
| 1989 | 13 | 37 | 45 |
| 1990 | 5 | 11 | 11 |
| 1991 | 27 | 78 | 82 |
| 1992 | 32 | 105 | 117 |
| 1993 | 11 | 41 | 62 |
| 1994 | 34 | 116 | 134 |
| 1995 | 14 | 43 | 45 |
| 1996 | 35 | 118 | 128 |
| 1997 | 7 | 34 | 34 |
| 1998 | 52 | 176 | 232 |
| 1999 | 9 | 30 | 30 |
| 2000 | 53 | 220 | 272 |
| 2001 | 22 | 96 | 106 |
| 2002 | 34 | 137 | 164 |
| 2003 | 11 | 48 | 55 |
| 2004 | 4 | 16 | 16 |
| 2005 | 16 | 91 | 97 |
| 2006 | 26 | 130 | 170 |
| 2007 | 28 | 127 | 150 |
| 2008 | 18 | 82 | 91 |

Проведена оценка полученных результатов с точки зрения производительности авторов и их совместной деятельности (кооперативности), как предложено в статье [3]. На Рис.2 приведен ряд трендов:

- производительности: как по полной сумме (full count) – отношению авторства к количеству авторов за рассматриваемый год, так и по частичной сумме (fractional count) – отношению количества публикаций за год к количеству авторов, опубликовавших эти статьи;
- кооперативности – отношению авторства к количеству публикаций за рассматриваемый год.

Авторы работы [1] считают, что частичная сумма отражает степень сотрудничества учёных в научных исследованиях, тем самым объединяя оба аспекта: производительность и кооперативность. График кооперативности (Рис.2) носит явно колебательный характер. На некоторых временных отрезках, на графиках наблюдаются синхронные изменения. Так, например, рост совместной деятельности специалистов в области МАМ с 1990г. по 1993г. (а также с 2004г. по 2006г.) сопровождается одновременным увеличением производительности (по полной сумме) с последующим снижением обоих параметров. В последнее время наблюдается снижение кооперативности, в то время как производительность по частичной сумме не изменяется.


Рисунок 2 Тренды производительности и кооперативности авторов в области разработки МАМ.

Очевидно, что приведенного выше материала явно не достаточно ни для составления картины развития выбранной нами области материаловедения, ни для прогнозирования дальнейшего её существования. Необходим более детальный анализ. Поэтому второй этап исследований представляет собой наукометрический анализ информации, хранящейся в специализированной библиографической базе данных по МАМ, разработанной авторами данной статьи.

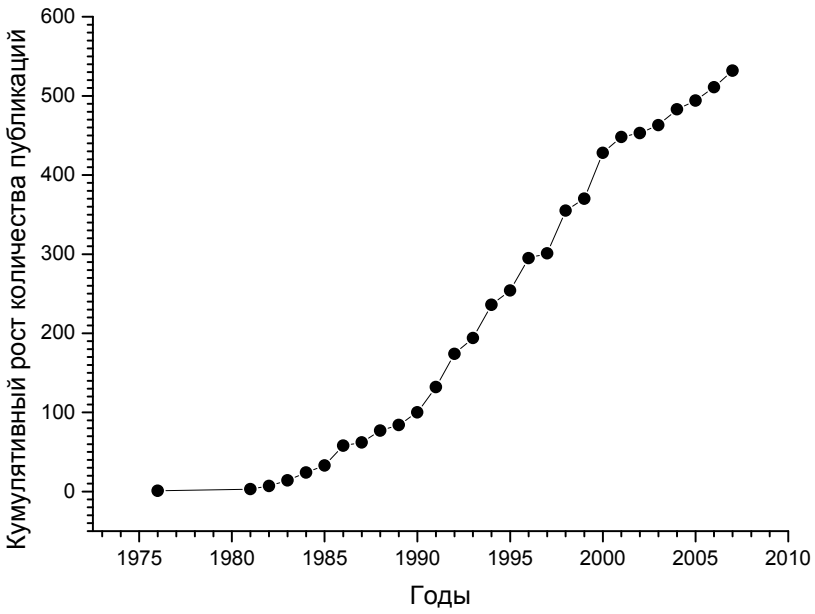



Рисунок 3 Кумулятивный рост публикаций в журналах по МАМ.

Наполнение базы данных проходило путем анализа и последующей индексацией текстовой библиографической информации, полученной в результате запроса к таким известным базам данных, как INIS[1], COMPENDEX[2], SCI[3]. Хронологический охват с 1976г. по 2008г. База данных содержит около 800 записей, охватывающих различные типы публикаций – книги, труды и тезисы конференций, лабораторные отчёты, статьи в журналах (46 журналов), а также рефераты публикаций.

На Рис.3 приведена динамика роста публикаций в журналах во времени (1976-2008гг.). Как видно, логистическая кривая не выходит на «плато», что свидетельствует о продолжающемся сравнительно быстром развитии этой специфической области материаловедения. Однако, период удвоения информации увеличился по сравнению с ранее полученными данными [7] с 4,5 года до 11,5 лет.

В Табл.2 представлено распределение информации, опубликованной в различных журналах по МАМ между странами. Очень сильный рывок за последние 10 лет сделали ученые из Германии и Франции. Такой скачок можно объяснить началом строительства международного термоядерного реактора, на территории Франции в Кадараше (Cadarache). Также появились новые страны-участники: Китай и Испания, вносящие свой вклад (по 3% каждая) в развитие МАМ. Расширение круга заинтересованных организаций по данной проблеме может быть связано и с подписанием в 2006 году Соглашения[4] о реализации проекта ИТЭР между Россией, Китаем, Европейским Союзом, Индией, Кореей, Японией и США.

Таблица 2 Распределение информационного потока по МАМ между странами, в %.

| № | 1976-1997гг. | | 1998-2008гг. | |
|---|---|---|---|---|
| п/п | Страна | % | Страна | % |
| **1** | США | 39,8 | Япония | 29,7 |
| **2** | Япония | 27,7 | Германия | 14,5 |
| **3** | Италия | 8,8 | США | 13 |
| **4** | Германия | 7 | Италия | 9,3 |
| **5** | Англия | 5,5 | Франция | 7,4 |
| **6** | Россия | 4 | Россия | 6,3 |
| **7** | Франция | 0,9 | Англия | 5,2 |
| **8** | - | - | Китай | 3,3 |
| **9** | - | - | Испания | 3 |

Учитывая, что действительная модель потока информации в специфической области науки может быть представлена как мульти-логистическая кривая [4], ранее [7] было выдвинуто предположение, что следующий подъем кумулятивного роста публикаций будет связан с развитием технологии МАМ. Из Рис.3 видно, что начиная с 1997 года действительно можно наблюдать дальнейший кумулятивный рост количества публикаций. Анализ показал, что наиболее важные направления в области МАМ, дающие основной вклад в ее развитие, связаны с изучением свойств МАМ, оценкой активации, технологией МАМ, микроструктурными исследованиями материалов и конструкционными вопросами ТЯР (Рис.4). Правда, в последние годы кумулятивный рост публикаций вызван не

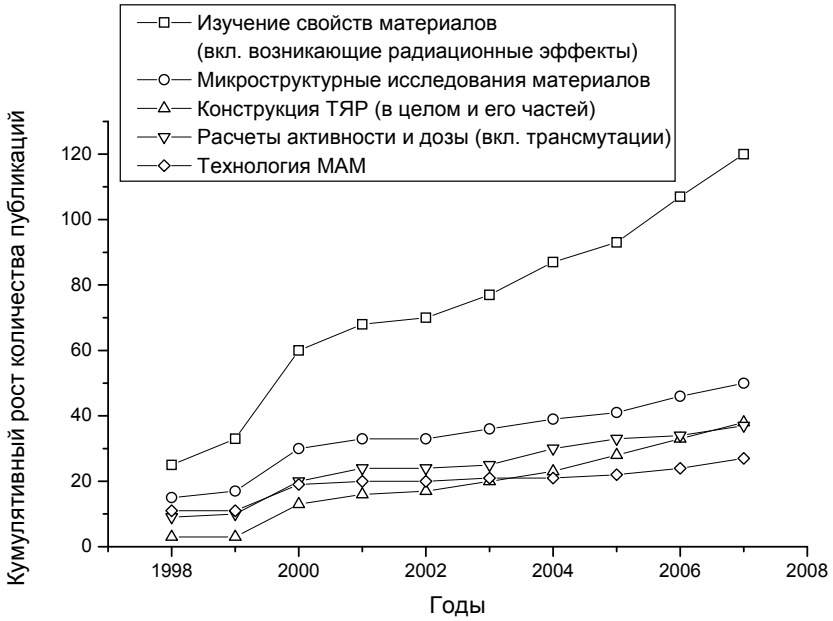



Рисунок 4 Вклад наиболее важных направлений в области развития МАМ.

развитием технологии МАМ, а повышенным интересом к конструкционным особенностям реактора, исследованиям свойств и микроструктуры материалов, моделирования и роли дефектообразующих факторов при облучении.

В специализированной базе данных по МАМ при классификации публикаций по свойствам материалов были выделены следующие индексированные кластеры:

- изучение свойств материалов (механических, физических и т.п.);
- эффекты, возникающие при нейтронном облучении материалов (радиационные дефекты, радиационное охрупчивание, распухание и т.п.).

Таблица 3 Распределение публикаций по разным темам.

| Количество публикаций, в % | Название темы |
|---|---|
| 55,1 | механические (включая термомеханические) свойства |
| 10,2 | охрупчивание материала |
| 6,1 | коррозия материала |
| 4,8 | радиационное упрочнение |
| 3,4 | распухание |
| 2,7 | термические свойства |
| 2 | эрозия материала |
| 5,4 | радиационные дефекты/повреждения |
| 2 | физические свойства |
| 8,1 | другие |

Эти кластеры, в свою очередь, включают ещё более детальную разбивку на темы. В Табл.3 приведено распределение публикаций за 1998-2007гг по темам. Как известно, механические свойства конструкционных материалов (металлов, сплавов, керамики и др.) устанавливают механическими испытаниями, целью которых чаще всего является нахождение связи между приложенными механическими напряжениями к материалу и его деформацией. Не удивительно, что 55% информации раздела связано с исследованиями механических и термомеханических свойств материалов для ТЯР.

## 3. Заключение

Наукометрический анализ тематических баз данных или коллекций более продуктивен для описания состояния и перспектив дальнейшего развития выбранной области науки. В тоже время исследования связанные с цитируемостью работ, возможны только, благодаря таким крупным базам данных, как SCOPUS, SCI или «Web of Sciences». Двухэтапный наукометрический анализ информационного потока по МАМ показал:

1. Несмотря на увеличение периода удвоения информации до 11,5 лет кумулятивный рост количества публикаций позволяет сделать оптимистичный прогноз по перспективам развития этой специфической области материаловедения.
2. 80% публикаций в ведущих журналах цитируется, что свидетельствует об их эффективности и дает возможность не только для поддержания развития данного

направления материаловедения, но и для обмена идеями.

3. Снижение кооперативности в последние годы не повлияло на производительность авторов.

## Литература


[1] Ajiferuke I, Burell Q., Tague J. Collaborative coefficient: A single measure of the degree of collaboration in research. Scientometrics, 1988, vol.14, No.5-6, p.421-433.

[2] Braun T., Glanzel W., Schubert A. Evaluation of citedness in analytical chemistry: How much is much? Analytical Proceedings, 1990, vol.27, p.38-41.

[3] Braun T., Glanzel W., Schubert A. Publication and cooperation patterns of the authors of neuroscience journals. Scientometrics, 2001, vol.51, No.3, p.499-510.

[4] Stromberg A.G., Orient I.M., Svishhenko N.M. Development of analytical chemistry in 1955-1981. A scientometric study. Zh. Analyt. Khim., 1984, vol.39, No.9, p.1704-1707.

[5] База данных SCOPUS. http://www.scopus.com/scopus/home.url

[6] Гарфилд Ю. Можно ли выявлять и оценивать научные достижения и научную продуктивность? Вестник Академии наук СССР, 1982, № 7, с.42-50.

[7] Колотов В.П., Аленина М.В., Иванов Л.И. Наукометрический анализ информационного потока в области развития малоактивируемых материалов для атомных и термоядерных установок. Перспективные материалы, 1998, №5, с.50-53.


## Scientometric investigation on development of works on development of low activation materials for fusion reactor


Alenina M.V., Kolotov V.P.



Scientometric analysis of the information flux in the field of development of low activation structural materials (LAM) for fusion reactor has been done. The analysis showed that the directions dealing with investigation of materials properties, estimation of activation, technology of LAM, investigation of microstructure evolution of materials and questions of fusion reactor construction display the most important contribution in the information flux in the field of LAM. Not less than 80% publications in the leading journals on the development of low activation materials are citing. The period of the information doubling is 11,5 years. Prognostication of some trends of this field of material science has been done.


---

[1]INIS - Международная информационная система по ядерной энергетике, которая охватывает мировую научно-техническую литературу по ядерным исследованиям и технологиям. Её генератором является Международное агентство по атомной энергии (МАГАТЭ) в Австрии.

[2]COMPENDEX - Компьютеризованный указатель технической литературы, который содержит библиографию технических журналов и трудов конференций из всех стран мира. Его генератором является фирма "Техническая информация" в США.

[3]SCI – Science Citation Index политематическая база данных по естественным наукам, медицине и с/х. Охватывает 3500 периодических изданий (мировой поток).

[4]Соглашение о создании Международной организации ИТЭР по термоядерной энергии для совместной реализации проекта ИТЭР и других международных договоров, направленных на реализацию указанного Соглашения (Распоряжение Правительства РФ от 04.09.06г. №1234-р).

# Сервисы геоинформационной системы сбора, хранения и обработки данных натурных наблюдений ♣

© Молородов Ю.И., Смирнов В.В., Федотов А.М.

Институт вычислительных технологий СО РАН
yumo@ict.nsc.ru

**Аннотация**

Описаны сервисы, расширяющие возможности геоинформационного сервера построенного на основе стандартов Open Geospatial Consortium. Предлагается модель распределенной информационно-аналитической системы с единой точкой доступа к геоданным (через разнородные пространственно распределенные базы метаданных) и к инструментарию для их обработки и визуализации.

## 1 Введение

Создание информационных ресурсов и интеграция их в единую информационную среду являются приоритетными направлениями развития современного общества. Разработка механизмов, обеспечивающих как функционирование общей информационно-аналитической среды, так и доступ к научным ресурсам, и их сохранность, имеет первостепенное значение в задачах информационной поддержки научных исследований. Эти вопросы особенно важны при исследованиях в области экологии и, в частности, биологического разнообразия, выполняемых в рамках междисциплинарных проектов СО РАН. Исследования проводятся различными группами ученых, разделенными географически, связанных необходимостью проведения совместных работ, обмену данными и координации своих действий. Применение информационных технологий в науках о Земле способствует пониманию как глобальных, так и региональных природных процессов, формирующих окружающую среду.

Развивающиеся геоинформационные технологии в настоящее время переходят на новый этап развития – создание распределенных ГИС (РГИС). Предпосылками этого перехода стали внедрение передовых технологий (распределенные базы данных, распределенные вычисления, высокоскоростные линии связи) и доступность веб-технологий широкому кругу пользователей. Этому также способствует стремительный прогресс в области создания и развития средств и технологий дистанционного зондирования Земли, обеспечивающий миллионам пользователей доступ к данным, получаемым с помощью спутниковых систем нового поколения (QuickBird, IKONOS, OrbView, GeoEye, SPOT, TerraSarX, Ресурс-ДК и др.) [1]. С другой стороны, в последние годы крупнейшие разработчики программного обеспечения объединяются в консорциумы по стандартизации. В области геоинформатики создан Консорциум открытых ГИС (Open Geospatial Consortium – OGC) [2]. Соответствие стандартам OGC позволяет программным продуктам различных производителей (в том числе коммерческим) взаимодействовать при решении конкретных задач. Переходу к РГИС также способствует внедрение спутниковых систем навигации NAVSTAR GPS и ГЛОНАСС, позволяющих с высокой точностью определять местоположение объекта в любой точке планеты. Распределенные геоинформационные системы имеют неоспоримые преимущества перед настольными, благодаря таким факторам, как:

- распределенный доступ к системе (наличие веб-интерфейса, позволяющего избежать установки дорогостоящего программного обеспечения, простота изменения и обновления программного обеспечения, доступность широкому кругу пользователей);
- распределенное хранение данных (организация доступа к архивам данных, возможность хранения пользовательских данных на сервере);
- распределенная обработка данных (возможность проведения обработки на высокопроизводительных вычислительных системах).

В настоящей работе представлена модель информационно-аналитического портала с единой точкой доступа к широкому диапазону данных и к инструментарию для их визуализации и обработки на высокопроизводительных вычислительных системах. Потребность в системах подобного рода особенно остро ощущается при проведении фундаментальных и прикладных исследований в области экологии и рационального природопользования.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

## 2 Принципы построения и структура Системы

Разрабатываемая геоинформационная система базируется на операционных системах семейства UNIX и наборе специализированных программных продуктов с открытым исходным кодом, распространяемых под лицензией GPL (GNU General Public License). Она полностью удовлетворяет требованиям OGC, предъявляемым к аналогичным системам, и допускает возможность подключения других ГИС. Основная цель ее разработки– создание виртуальной информационно-аналитической среды для поиска, обработки и анализа пространственных данных. Она позволит организовать единую точку доступа к различным геоинформационным системам и распределенным разнородным хранилищам данных и атрибутивной информации.


Рис. 1. Структура распределенной ГИС

На рис. 1 представлена структура системы. Доступ осуществляется через стандартный веб-браузер, что обеспечивает платформенную независимость. Многоуровневая система разграничения прав доступа после авторизации позволяет генерировать “на лету” графический интерфейс пользователя (в соответствии с уровнем доступа). Он представляет собой виртуальный рабочий стол с традиционными элементами управления. Ядро виртуальной среды состоит из набора Perl/PHP/JavaScript-приложений (с поддержкой технологии AJAX) и интерфейсов к внешним компонентам системы, работающих под управлением веб-сервера Apache. Внешние приложения взаимодействуют с системой через интерфейсы, описанные на языке XML.

Apache 2.0 (с расширением Tomcat) используется в качестве *HTTP-сервера* для платформы UNIX. Это веб-сервер разрабатывается и поддерживается открытым сообществом разработчиков под эгидой Apache Software Foundation и включен во многие программные продукты, среди которых СУБД Oracle и IBM WebSphere. К его основным достоинствам относятся надежность и гибкость конфигурации. Для размещения пространственно-координированных данных используется *картографический сервер* UMN MapServer. В настоящее время он приобретает все большую популярность среди разработчиков геоинформационных веб-сервисов. По своим функциональным возможностям он не уступает коммерческим программам, а по ряду параметров превосходит их, в частности по производительности [10]. MapServer включает в себя все необходимое для разработки основных картографических сервисов WMS/WFS, рекомендованных к использованию OGC [11]. Он позволяет пользователю формировать карты с одновременным использованием материалов, хранящихся как в локальных, так и в удаленных архивах. Богатая функциональность, легкость интеграции с различными СУБД и открытость исходных кодов предопределили популярность программы.

В качестве базового *инструментария для обработки и анализа данных дистанционного зондирования* выбран пакет программ GRASS GIS (Geographic Resources Analysis Support System). Его отличительной особенностью является о интеграция в среду UNIX, поддержка основных типов пространственных данных, мощный процессор обработки растровых данных, модульность и открытый инструментарий для быстрой и эффективной разработки модулей расширения. Использование в пакете библиотек GDAL и PROJ4 обеспечивает поддержку всех современных стандартов геоданных и большой набор функций для трансформации и перепроецирования изображений. Пакет GRASS GIS позволяет разрабатывать модули расширения практически на всех языках программирования, для которых есть компилятор под UNIX (Perl, sh, C/C++, Fortran и т.,д.). Он включает библиотеки для работы практически со всеми современными СУБД. Внутреннее представление растровых данных базируется на вокселях (3D-пикселях), что существенно повышает возможности по обработке. Интеграция пакета GRASS GIS в разрабатываемую систему позволяет с минимальными временными затратами обеспечить пользователя доступом к полнофункциональной ГИС, расширенной специализированным математическим аппаратом.

Большое внимание в системе уделено статистической обработке данных и визуализации результатов. Для этих целей в нее интегрирован статистический пакет R., распространяемый на условиях лицензии GPL.

Функциональность системы расширяется за счет использования *сервера приложений*. С обеспечения работы дополнительных сервисов, на языке XML, разработаны интерфейсы для взаимодействия с внешними приложениями.

Функционирование предлагаемой системы в распределенном режиме по протоколам доступа к метаданным основано на *модулях поддержки протокола Z39.50* [3] (и CIP [4] как одного из его профилей). Для этого в систему включены:

- сервер Z39.50 (ZooPARK), обеспечивающий базовую функциональность сервисов Z39.50 в

соответствии с различными прикладными профилями;

- шлюз Z39.50-HTTP, обеспечивающий простые пользовательские интерфейсы для доступа к ресурсам Z39.50 по протоколу HTTP;
- набор динамических провайдеров данных, каждый из которых описывает условия и протокол взаимодействия с конкретной СУБД, в которой хранятся метаданные. Вся логика работы с конкретной СУБД локализована в соответствующем провайдере данных. Взаимодействие базового сервера с провайдерами данных осуществляется через единый интерфейс.

## 2.1 Каталог пространственных данных

Для добавления данных дистанционного зондирования в хранилище системы создан специализированный сервис. Он производит начальную обработку данных, с последующей индексацией. После обработки данные перемещаются для долгосрочного хранения в систему хранения данных, а полученные метаданные приводятся к стандартному виду и размещаются в соответствующих разделах поисковой системы. Только после этого данные становятся общедоступными. Разработанная поисковая система позволяет находить данные по метаданным и выполнять комплексные запросы (содержащие географические координаты, номера трека и кадра, дату и время, параметры облачности и др.).

Для систематизации данных, организации поиска и извлечения из архива необходимой информации разработан каталог пространственных данных. В его основе лежит набор стандартных и специализированных программных продуктов с открытым исходным кодом, распространяемых под лицензией GPL (GNU General Public License). В основе структуры каталога лежит набор Perl/Java/JavaScript-приложений, работающих под управлением веб-сервера Apache.

Доступ к каталогу реализован посредством модуля Central Authentication Service (CAS), разрабатываемого в рамках проектом JA-SIG [5]. Он позволяет организовать многоуровневую систему разграничения прав доступа с централизованной базой пользователей на основе LDAP-каталога Сибирского отделения РАН. Функционально сервис реализован в виде прокси-сервера для аутентификации, возможна тонкая настройка параметров доступа (например, в зависимости от IP-адреса, доменного имени, даты и времени суток, используемого метода аутентификации, расширения файла и т.п.). Это позволяет реализовать практически индивидуальные настройки доступа к любому защищаемому ресурсу. Модуль CAS позволяет легко создавать защищенные ресурсы как на основе Apache/Tomcat, так и при использовании технологий PHP/JavaScript на платформе Apache.

## 2.2 Картографические сервисы

GeoServer предназначен для публикации набора векторных и растровых слоев, в частности слоя покрытия территории данными оперативного спутникового мониторинга за интересующий период. Приложение взаимодействует непосредственно с СУБД PostgreSQL/PostGIS что позволяет построить высокопроизводительный и легкий в настройке сервис.

*Подсистема картографических сервисов* реализована на программных продуктах, распространяемых под лицензией GPL(GeoServer и UMNMapServer).

*Картографический сервер* UMN MapServer предназначен для публикации растровых данных, и для извлечения информации из сопряженных баз данных и публикации законченных проектов. В настоящее время он становится одним из наиболее популярных инструментов для создания геоинформационных веб-сервисов и ресурсов[7]. MapServer содержит все необходимое для разработки картографических сервисов WMS/WFS, в соответствии со спецификациями OGC. Он позволяет формировать карты, одновременно используя информационные слои, размещенные как в локальных, так и в удаленных архивах. Богатая функциональность, легкое взаимодействие с различными СУБД и открытость исходного кода предопределили популярность программы.

К настоящему времени запущен в эксплуатацию картографический сервис, позволяющий организовывать хранение и публикацию в сети Интернет различных картографических материалов. Он обеспечивает доступ к сопряженной атрибутивной информации. Сервис обеспечивает доступ к картографическим материалам посредством стандартизированных протоколов (WMS, WFS), что позволяет использовать данный сервис как для разработки различных веб-приложений, так и для использования данных при работе в различных геоинформационных продуктах (например, ENVI, ArcGIS, MapINFO и др.).

### 2.2.1 Сервис доступа к картографическим данным

Картографический сервис обеспечивает доступ к картографическим данным посредством использования стандартов WMS/WFS и решает задачи:

1. обеспечения централизованного доступа к картографическим сервисам для организаций и сотрудников СО РАН;
2. повышение скорости работы с удаленными картографическими сервисами;
3. позволяет кэшировать поток данных предоставляемых сервисами, для снижения нагрузки на сервера, выполняющие функции генерации данных;
4. снижает загрузку внешних каналов связи.

В основе сервиса - система хранения данных ИВТ СО РАН с общим объемом дисковой памяти около 40 тбайт (рис. 2).

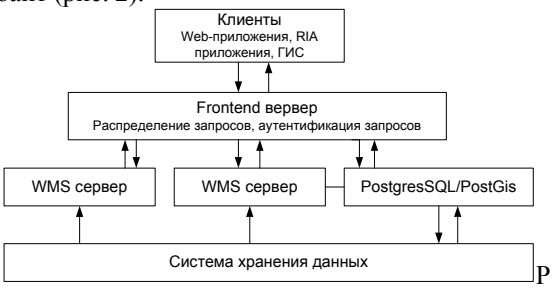



Рис.2. Структурная схема сервиса

Программная реализация серверной части системы выполнена на основе картографического сервера *GEOSERVER*, распространяемого на основе лицензии с открытым исходным кодом (GPL). Сервер обеспечивает доступ пользователей посредством использования протоколов WMS/WFS, с использованием, как программный клиентов реализованных с использование веб-технологий, так и с использованием стандартного, в том числе коммерческого (ArcGIS, ENVI и др.), программного обеспечения. Хранение атрибутивной информации, а также связанной пространственной информации реализовано на платформе PostgreSQL/PostGIS, распространяемой по лицензии GPL. Использование данной СУБД позволяет легко развернуть полномасштабную систему с возможностью работы с пространственной информацией.

Важной составляющей в разрабатываемой системе является подсистема кэширования геопространственных данных. Использование технологий кэширования позволяет существенно увеличить время доступа к данным, а также значительно снизить нагрузку на серверное оборудование. Дополнительно, при использовании технологий обмена содержимым кэша, имеется возможность строить распределенные по регионам системы кэширования, обеспечивающие оптимальный доступ пользователей к данным. Подсистема кэширования пространственных данных предполагает развитие нескольких стратегий работы с пользователями сервиса:

1. Организация кэширования во время выполнения текущих запросов.
2. Кэширование дополнительных областей, вокруг зон повышенного интереса.
3. Постепенное кэширование зон интересов по заказу.

Пространственный поиск по картографическим данным обеспечены протоколами Z39.50 и SOAP.

### 2.2.2 Векторная карта ландшафтов юга Восточной Сибири

На платформе Flash/Flex реализовано клиентское приложение WMS Browser, которое обеспечивает работу с различными картографическими слоями. Через это приложение реализован интерфейс работы с векторной картой ландшафтов растительности юга Восточной Сибири (Иркутская область и Республика Бурятия, масштаб 1:1 500 000). Авторы карты: В.Б. Сочава, В.С. Михеев, О.П. Космакова (цифровая версия карты представлена И.Н. Владимироввым и А.А. Сороковым). Адрес ресурса (http://gis-app.ict.nsc.ru/wmsbrowser). Здесь разными цветами выделены группы и геомы растительности.


Рис.3. Приложение WMSBrowser

### 2.2.3 Электронные библиотеки для ботаники

Одной из важных задач создания геоинформационного портала (http://gis-app.ict.nsc.ru), является создание виртуальной информационно-аналитической среды для поиска, обработки и анализа спутниковых и наземных данных [8].

В Новосибирском научном центре специалисты ботаники Центрального сибирского ботанического сада (ЦСБС) и специалисты по информационным технологиям Института вычислительных технологий (ИВТ) СО РАН, начиная с 1993 г. занимаются разработкой программных продуктов и информационных ресурсов для нужд биологических наук. Основными направлениями разработок были выбраны: 1) создание тематических виртуальных библиотек - публикация обобщенной и систематизированной биологической информации с возможностью оперативного поиска; 2) разработка виртуальных коллекций - программных продуктов для хранения и систематизации первичных биологических данных. Были созданы Электронный атлас "Биоразнообразие животного и растительного мира Сибири", построена база данных "Зеленая книга Сибири", электронная библиотека "Разнообразие растительного и животного мира Сибири" и “Электронный каталог растений Сибири”. Практически они являлись электронными версиями опубликованных в издательстве Сибирского отделения “Наука” книг Зеленая книга Сибири и Флора Сибири.

Они представляют собой виртуальные путеводители (электронные библиотеки) по флоре Сибири и Дальнего Востока России. Здесь можно узнать о распространении отдельных родов и видов в разных регионах Сибири и Дальнего Востока, в каких местообитаниях они чаще всего встречаются, посмотреть фотографии и рисунки, познакомиться с

общим распространением видов. Материалы, представленные в этих разработках, преследовали, по крайней мере, две цели: изучение возможностей современных вычислительных технологий для организации многопользовательской информационной системы по растениям крупного региона Сибири и определение эффективности работы с электронным информационным ресурсом по сравнению с публикацией на бумажных носителях. Главным отличием электронных библиотек стала возможность размещения в них многочисленных цветных иллюстраций и фотографий растений, выполненных в природной обстановке.

Кроме публикации обобщенной информации, насущной проблемой биологических (ботанических в частности) исследований является формализации, интеграция и систематизация первичных описательных данных; в противном случае накопление информации быстро утрачивает смысл. На ГИС-портале ИВТ СО РАН реализована возможность создания многопользовательских виртуальных коллекций. Предоставляемые в виртуальных коллекциях возможности отвечают основным потребностям первичной обработки ботанической информации. Помимо чисто научного значения, построение и развитие виртуальных коллекций важно для природоохранной деятельности – это позволяет оперативно выявлять действительно редкие виды, анализировать их распространение, выявлять наиболее интересные и подлежащие охране территории.

На страницах портала реализован доступ к информационным ресурсам и разработанным ранее базам данных: (http://old.ict.nsc.ru/win/elbib/bio/) "Биоразнообразие животного и растительного мира Сибири", (http://old.ict.nsc.ru/win/elbib/bio/green/) - "Зеленая книга Сибири" электронной библиотеки "Разнообразие растительного и животного мира Сибири" и Электронный каталог растений Сибири (http://old.ict.nsc.ru/win/elbib/atlas/flora/).

### 2.2.4 Атлас «Мхи России»

Создан прототип информационно-поисковой системы для хранения связной и формализованной информации о видах растений (http://gis-app.ict.nsc.ru/bio/). В качестве исходного массива данных используется бриологическая база данных “Мхи России” (www.arctoa.ru), которая была получена путем объединения личных БД специалистов-бриологов России. Информационная система содержит базу данных "Мхи России", которая на текущий момент содержит около 30 000 документов.

Атлас построен по модульному принципу. Существует ядро, включающее в себя несколько программ, и модули, которые вызываются ядром по мере необходимости. Подобный подход позволяет уменьшить расходы, как связанные с модификацией системы в целом, так и отдельных её частей. Следует отметить, что в данном случае *модульный подход* используется как принцип построения системы в целом, а не отдельных её частей. Внутри каждого модуля и ядра может использоваться как модульный принцип программирования, так и объектно-ориентированный.

Среди главных подсистем ядра можно выделить следующие блоки :

- разграничения доступа и управления пользователями;
- изменения фактографической информации;
- поиска информации и вывода результатов;
- взаимодействия с географической картой;
- многоязычного интерфейса.

При этом каждый из них представляет собой один или несколько программных модулей.

База данных состоит из трех основных элементов:

1. *Таблица названий видов – “Species_list”.* Содержит названия видов мхов, авторов названий, согласно сводке (Ignatov, Afonina, Ignatova et al., 2006), синонимы и русские названия. Возможно редактирование названий соответственно современной номенклатуре..
2. *Таблица регионов России – “Regioni_Russia”.* Содержит данные о регионах России: код региона, буквенный код региона, русское и английское название (разработаны М.С.Игнатовым (ГБС РАН, Москва) для «Флоры мхов России» (www.arctoa.ru)).
3. *Таблица коллекции Гербарных Этикеток – “etic_Russia”.*Содержит в себе данные о конкретных образцах, взятых при полевых исследованиях.

Важным этапом в работе являлось обеспечение двустороннего взаимодействия с картой. Работа подсистемы осуществляется следующим образом. После выдачи поискового запроса, пользователь имеет возможность перейти по ссылке «показать все результаты поиска на карте». При активизации ссылки (нажатии на нее) программа находит в базе данных координаты всех найденных образцов и записывает их во временный файл специального вида. После этого открывается картографический интерфейс созданный файл передается программе-скрипту отображения. Далее он добавляет данные из этого временного файла отдельным слоем на карту, происходит центрирование и изменение масштаба для комфортного просмотра результатов (Рис.4).

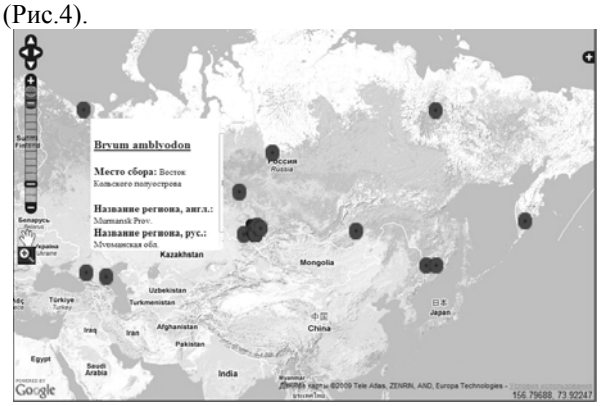


Рис. 3. Страница отображения гербарных этикеток на карте

Взаимодействие с картой происходит в обе стороны, так после просмотра результатов на карте, пользователь может перейти к просмотру конкретной гербарной этикетки. После щелчка по соответствующему маркеру всплывает краткое описание этикетки, на котором присутствует ссылка на полное описание. Либо как уже отмечалось выше выбрать с помощью картографического интерфейса область поиска данных

## 3. Заключение

В работе описаны сервисы распределенной информационно-аналитической системы, обеспечивающей поддержку исследований, связанных с обработкой и анализом пространственно распределенных данных. Создаваемая система взаимодействует с системой хранения данных общим объемом памяти более 40 Тбайт. Функциональные блоки объединяются в единую систему и производятся разработка графического интерфейса пользователя и подключение баз данных, созданных в ходе многолетних исследований специалистами СО РАН. Создан прототип информационно-поисковой системы для хранения связной и формализованной информации о видах растений (http://gis-app.ict.nsc.ru/bio/). В качестве исходного массива данных используется бриологическая база данных “Мхи России”, которая была получена путем объединения личных БД специалистов-бриологов России. Первоначальный вариант базы данных был подготовлен в формате MS Access, что потребовало разработки алгоритмов преобразования в формат базы данных MySQL. Информационная система содержит базу данных "Мхи России", которая на текущий момент содержит около 30 000 документов.

Поиск в базе данных нужного вида мха осуществляется как по метаданным гербарному или полевому номеру образца, по виду и внутривидовому таксону, по точке сбора и местообитанию и др., так и по географическим параметрам и высоте. Результаты поиска могут быть выведены в таблице или в виде маркера на географической карте. При наведении на него курсора мыши, высвечивается гербарная этикетка с указанием названия вида, региона и места сбора.

Информационная система используется отечественными и зарубежными специалистами-бриологами. Она является основой для обобщения и ревизии гербарных материалов к подготовке первого издания «Флоры мхов России».

## Литература


[1] SOVZOND Материалы Междунар. конф. “Космическая съемка - на пике высоких технологий”. Россия, Москва, 18-20 апреля 2007 г. // http://www.sovzondconference.ru/archive-2007/rus/agenda.html..

[2] The Open Geospatial Consortium, 2007. http://www.opengeospatial.org.

[3] ANSI/NISO Z39.50-1995. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. Z39.50 Maintenance Agency Offical Text for Z39.50. 1995.

[4] Жижимов О.Л., Мазов Н.А. Принципы построения распределенных информационных систем на основе протокола Z39.50. Новосибирск: ОИГГМ СО РАН; ИВТ СО РАН, 2004. 361 с.

[5] Catalogue Interoperability Protocol (CIP) Specification - Release B 2.4. CEOS/WGISS/ PTT http://www.dfd.dlr.de/ftp/pub/CIP_documents/cip2.4/S_cover.pdf.

[6] http://www.ja-sig.org/products/cas/index.html.

[7] ANSI/NISO Z39.50-1995. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. Z39.50 Maintenance Agency Offical Text for Z39.50. 1995

[8] Шокин Ю.И., Добрецов Н.Н., Пестунов И.А., Молородов Ю.И., Смирнов В.В., Синявский Ю.Н. Система сбора, хранения и обработки спутниковых и наземных данных Новосибирского научного центра СО РАН // Выч. тех.– 2008.– Т.13.– Вестн. КазНУ им. аль-Фараби. Серия: Математика, механика, информатика.– №4 (59).– Совместный вып. по материалам междунар. конф. «Вычислительные и информационные технологии в науке, технике и образовании».– Ч.III.– С.371-376.


## Services GIS data collection, storage and data processing field observation


Molorodov Yu.I., Smirnov V.V., Fedotov A.M.



We describe the services that enhance the possibility of geo-information server based on the standards of Open Geospatial Consortium. The model of distributed data-processing system with a single point of access to geo-data (in a spatially heterogeneous distributed database of metadata) and tools for processing and visualization.



♣ Работа частично финансировалась Президентской программой НШ 931.2008.9, РФФИ (грант № 09-07-00277), и Интеграционным проектом СО РАН № 50.

# МУЛЬТИМЕДИЙНЫЕ КОЛЛЕКЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

# MULTIMEDIA COLLECTIONS, INFORMATION SECURITY

# Электронная библиотека ВУЗа - как инструмент автоматического формирования учебных мультимедийных коллекций

© Курчинский Д. Н., Палей Д. Э., Смирнов В. Н.

Ярославский государственный университет им. П. Г. Демидова
reno@econom.uniyar.ac.ru, paley@jugra.yar.ru, smirnov@uniyar.ac.ru

**Аннотация**

Основная задача, которая рассматривается в предлагаемой работе - это построение системы эффективного использования данных электронной библиотеки ВУЗа в учебном процессе. Для этого авторы предлагают автоматически формировать 'учебные коллекции' на основе данных системы автоматизации фундаментальной библиотеки и цифрового каталога мультимедийных объектов.

## 1 Введение

Проблемы и задачи создания, развития и использования электронных библиотек (ЭБ) ВУЗа решаются различными способами уже достаточно давно, фактически с момента возникновения таких общепринятых в наше время понятий, как "цифровая библиотека" или "электронный архив". Каждый год появляется достаточно большое количество работ и проектов, посвященных этой или близких к этой тематике (к примеру, только на RCDL'2008 были представлены работы [1, 2]). Они предлагают различные подходы к построению ЭБ, в том числе, ЭБ учебных заведений высшего образования. Характерным для многих проектов является то, что в них решаются задачи систематизации, упрощения доступа и повышения эффективности использования больших объемов данных, накопленных в ВУЗах.

В работе [3] приведены данные по мониторингу электронных библиотек учебных заведений, которые показывают, что большинство библиотек ВУЗов в настоящее время имеют электронные каталоги составляющие основу электронных библиотек. Что касается полных текстов, видео и аудиоматериалов, то тут положение дел существенно хуже. В настоящее время такие коллекции представляют собой, в основном, полные тексты изданий (что само по себе очень даже неплохо). Массовые работы по наполнению, и главное, систематизации таких электронных ресурсов начались значительно позже создания библиографических систем.

Таким образом, можно сделать вывод, что именно системы обработки библиографических данных сегодня составляют основу ЭБ высших учебных заведений.

## 2 Электронная библиотека Ярославского государственного университета им. П. Г. Демидова

Проектированием, разработкой и созданием подобных систем уже долгое время занимаются и авторы доклада. Реально созданы и развиваются системы: "Digital Library" (DL) [8] и система автоматизации библиотеки (АБИС "Буки") [5].

### 2.1 Каталог цифровых объектов "Digital Library"

Система "Digital Library" представляет собой универсальный цифровой каталог, предназначенный для описания произвольных объектов и сущностей. Метаинформация (метаописания сущностей) в виде определений классов объектов хранится в самом каталоге и являются его частью. Для классов реализован механизм "наследования".

Экземпляры классов (объекты) организованы в виде иерархического цифрового каталога. Каждый объект может иметь произвольное количество зависимых объектов и принадлежать какому-либо классу, описанному в метаданных.

На основе этого каталога построены различные сервисы доступа и навигации по объектам. Наряду с предоставлением стандартных возможностей (поиск, модификация, копирование, удаление) на уровне каталога поддерживаются возможности связанные с тем, что метаописания являются частью системы. Особо отметим сервис "Наследования данных" [8].

Упрощенно, этот сервис предоставляет возможности эффективного использования данных объекта некоторого класса Q. Пусть мы унаследовали от класса Q некоторый класс T. Далее, пусть мы хотим создать объект класса T – $O_T$, уточняющий некоторый объект класса $O_A$.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

(Например, класс A – “сотрудник”, класс “T” - преподаватель). В этом случае “наследование данных” позволит использовать данные $O_A$ в объекте $O_T$ автоматически.

Являясь универсальной системой, DL позволяет описывать и сохранять информацию о произвольных документах (объектах). Следует отметить, что эта универсальность одновременной является и достаточно большой проблемой, т. к. требует значительных затрат на внесение, администрирование и поддержку актуальности данных.

На основе DL выполнено несколько проектов по грантам РФФИ РГНФ (98-07-91152, 98-07-03270, 03-07-91152, 04-07-90154) и РГНФ (97-04-12016, 03-04-12019в, 04-06-12016в).

### 2.2 АБИС “Буки”

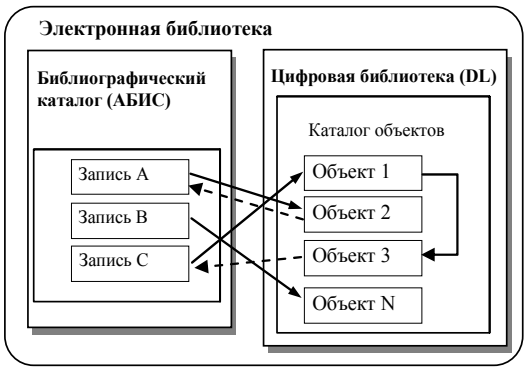
АБИС “Буки” используется на практике в библиотеке ЯрГУ и в других ВУЗах региона. В рамках системы реализованы все основные функции по каталогизации, поиску и обработке библиографических записей. Также поддерживается работа всех основных отделов библиотеки - абонемент, новые поступления, материальный учет и т. д. Особо отметим, что в рамках АБИС реализован специфичный для ВУЗов модуль “Книгообеспеченность” [6, 7]. Именно на основе этого модуля и предполагается реализовать новые сервисы электронной библиотеки, представляемые в данном докладе.

### 2.3 Интегрированная система

Интеграция этих систем в одну общевузовскую электронную библиотеку давно уже является актуальной проблемой. Изначально ставилась задача: получить и упростить доступ из библиографической системы к публикациям в DL (прежде всего, к полным текстам или отсканированным документам) и наоборот сделать возможным доступ к библиографическим записям в качестве объектов цифрового каталога [4]. Решение этой задачи вылилось в организацию системы ссылок объектов DL на библиографические записи АБИС и наоборот. Отметим, что большинство практических подходов, на сегодняшний день, с различными вариациями сводятся именно к такому решению. Функционально ЭБ университета строится следующим образом:

- информационная система состоит из двух основных частей - АБИС и DL;
- каждый модуль выполняет собственные специализированные функции;
- связь осуществляется на уровне ссылок записей АБИС на объекты каталога и объектов (атрибутов объектов) на записи АБИС.

Т. к. работы ведутся давно - уже можно обсуждать итоги и опыт практического внедрения и эксплуатации этих систем, а также степень соответствия результатов и подходов поставленным ранее задачам. На данный момент авторы могут констатировать, что основные идеи, заложенные в основу проектов, оказались практически эффективными.




При относительной независимости использования и администрирования обеих подсистем ЭБ достигнута достаточно тесная интеграция, которая позволяет решать многие задачи.

## 3. Проблемы использования электронной библиотеки и цель проекта

Вместе с тем стали очевидны проблемы использования, которые являются общими для систем такого типа и вытекают как из предложенного подхода, так, очевидно, обусловлены спецификой использования системы в ВУЗе.

В первую очередь это касается информационного наполнения DL. Изначально цифровая библиотека создавалась для выполнения работ по различным грантам. В рамках этих работ было введено большое количество мультимедийных данных (полные тексты, отсканированные копии документов, графика, видео и т. д.). Но на постоянной основе такая информация, к сожалению, практически не вносится. Для этого необходимо помимо отработанных технологических решений решить массу организационных вопросов (обучение персонала, выделение средств, оборудования и т.д.)

Другой важной проблемой является доступность данных для использования. Даже в составе электронной библиотеки информационный ресурс должен быть найден конечным пользователем и представлен ему в доступном виде. Если рассматривать эту ситуацию применительно к учебному процессу, то предполагается, что существуют специально подобранные и отобранные ресурсы. Эти электронные документы соответствуют той или иной задаче обучения - чаще всего обеспечивают поддержку дисциплины, преподаваемой в ВУЗе. Далее будем называть такие ресурсы “**учебными коллекциями”.**

Следует отметить, что задачи формирования таких коллекций могут быть успешно решены в рамках уже применявшихся подходов

извлечения/обработки данных ЭБ. При этом подразумевается формирование непосредственно в ЭБ специализированного каталога, поиск и добавление в него объектов ЭБ и ссылок на библиографические описания.

Для проведения такой работы необходим квалифицированный эксперт (в ВУЗе таковым является преподаватель), который имеет хотя бы минимальные навыки поиска и администрирования данных в библиографических каталогах и цифровой библиотеке. Для неподготовленного пользователя это представляет собой довольно сложную задачу, а объем и количество различных курсов, загруженность преподавателей на практике исключают возможность массовой подготовки таких коллекций вручную. С другой стороны, в силу понятных причин, самостоятельное составление ”учебных коллекций” студентами так же практически невозможно. Более того, часто возникает потребность решать такую задачу без привлечения конкретного эксперта или с привлечением его на заключительных этапах формирования наборов данных.

Потребность в учебном заведении в использовании учебных ресурсов на основе данных электронной библиотеки высока как со стороны преподавателей, так и со стороны студентов. Таким образом, возникает проблема автоматического формирования специализированных “учебных коллекций”, состоящих из объектов цифровой библиотеки на основе данных АБИС в рамках единой электронной библиотеки ВУЗа.

Актуальность такой задачи так же может быть подтверждена объемами литературы по специальностям изучаемым в ВУЗах. В качестве примера можно привести данные по ЯрГУ:

| | |
|---|---|
| Количество специальностей | **68** |
| Количество дисциплин | **>2000** |
| Количество кафедр | **52** |
| Общее количество наименований литературы | **>5000** |
| Максимальное количество наименований по специальной дисциплине | **175** |
| Среднее количество наименований литературы по дисциплине | **12** |

Создать вручную наборы электронных документов по всем специальностям представляется проблематичным и трудоемким.

Из вышесказанного следует, что для решения этой проблемы необходимо решить следующие задачи:

- обеспечить ввод данных по полным текстам, аудио- и видеоданным в DL одновременно с описанием этого контента в АБИС;
- модифицировать АБИС с целью реализации возможностей автоматического составления библиографических списков для задач учебного процесса;
- добавить в ЭБ возможности администрирования готовых “учебных коллекций”.

## 4. Реализация проекта

### 4.1 Задачи использования ЭБ в учебном процессе

Как уже было сказано ранее, эффективно использовать медиаконтент можно только в том случае, если одновременно с его помещением в каталог ЭБ будет составлено его описание. В нашем случае - это библиографическая запись в АБИС. Авторы прекрасно понимают, что это большей частью организационная задача. Вместе с тем, на уровне программных решений можно обеспечить приемлемо корректное и, главное, обязательное описание любого мультимедийного документа. Форматы RUSMARC, USMARC позволяет описывать самые различные типы документов и публикаций, в том числе и медиаконтент [9, 11].

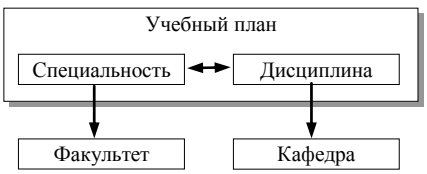
Рассмотрим основные задачи, для которых можно применить электронную библиотеку в учебном процессе. Традиционно выделяют следующие направления:

- формирование списков основной и дополнительной литературы по отдельным темам, курсам, специальностям, с возможностью просмотра полных текстов или других материалов;
- формирование виртуальных обучающих курсов на основе содержимого каталога ЭБ;
- формирование тестовых заданий для контроля и самоконтроля обучающихся.

Все эти задачи в конечном итоге сводятся к формированию некоторого набора объектов, объединенных в именованную “учебную коллекцию”. Далее такая коллекция доступна конечному пользователю в специализированном средстве просмотра. Каким образом сформированные коллекции/курсы будут представлены не критично. Удобнее всего это сделать через web интерфейс, хотя можно это реализовать, например, и через клиента доступа к данным АБИС или ЭБ.

### 4.2 Модуль “Книгообеспеченность” – источник шаблонов “учебных коллекций”

Для формирования коллекций по учебным курсам авторы доклада предлагают использовать данные модуля АБИС "Книгообеспеченность". Схожие идеи применительно к ссылкам на полные тексты уже высказывались в работе [10].

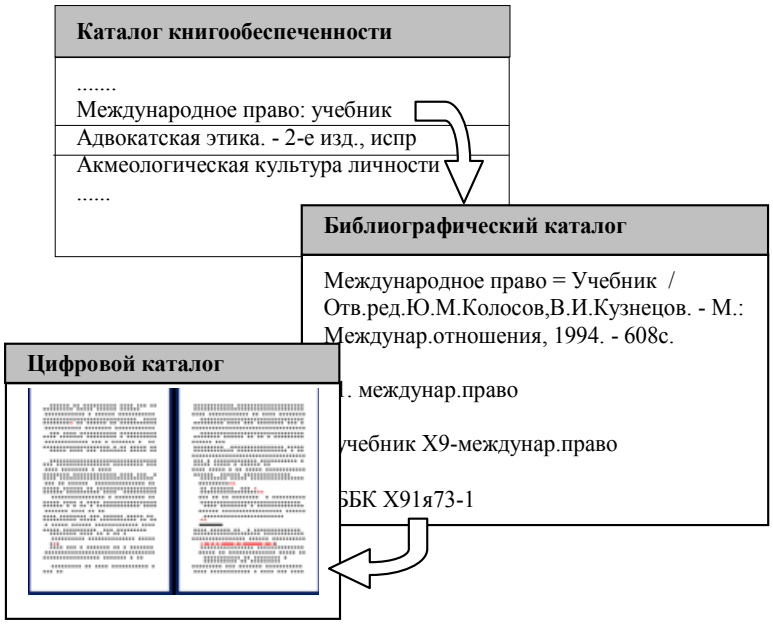



**Структура учебного процесса в подсистеме книгообеспеченности АБИС БУКИ**

Учебный план (по специальностям)

| | Фак./Спец./Дисциплина | Кафедра |
|---|---|---|
| C | экология и природопользование (бакалавриат) | |
| F | **Информатики и вычислительной техники** | |
| C | информационные технолоии (бакалавриат) | |
| C | математическое обеспечение и администрирование ИС | |
| C | прикладная информатика в экономике | |
| D | алгоритмы обработки информации | вычислительных и программных систем |
| D | английский язык для начинающих (факультатив) | иностранных языков |
| D | базы данных | информационных и сетевых технолоий |
| D | Библейская мифолоия и возникновение христианства (по выбору) | всеобщей истории |
| D | бухгалтерский учет | информационных и сетевых технолоий |
| D | веб-дизайн (по выбору) | информационных и сетевых технолоий |
| D | высокоуровневые методы информатики и программирования | теоретической информатики |
| D | вычислительные системы, сети и телекоммуникации | информационных и сетевых технолоий |
| D | глобальные компьютерные сети | информационных и сетевых технолоий |
| D | дипломы | ректорат |

**Фрагмент учебного плана подсистемы книгообеспеченности АБИС БУКИ**




**Связь записи в каталоге книгообеспеченности и полного текста публикации**

Данные этого модуля в АБИС “Буки” фактически представляют собой полный справочник по всей программе обучения ВУЗа. Все это объединено в учебном плане. Каждая строка учебного плана однозначно идентифицируется набором: *факультет <- специальность <-> дисциплина -> кафедра*

Выбор такой структуры хранения данных позволяет легко получить выборки необходимой литературы, как по дисциплинам, так и по кафедрам или специальностям.

Каждому пункту учебного плана соответствует список основной и дополнительной литературы. Издания прикреплены к соответствующим дисциплинам вместе с аналогами использования. Подобный подход хорошо отработан и показал на практике свою эффективность. Важным является то, что актуальность информации в модуле поддерживается сотрудниками фундаментальной библиотеки в силу своих должностных обязанностей. А формируют эти данные преподаватели ВУЗа, т. е. непосредственно те, кто будет в них нуждаться и будет использовать в работе.

Таким образом, эта информация фактически является готовым эффективным шаблоном для генерации “учебных коллекций” мультимедийных материалов по различным курсам. Очевидно, что при этом можно полностью использовать существующие ссылки библиографического каталога на объекты ЭБ. Более того, можно воспользоваться возможностями сервиса “Наследования данных”. Если библиографическое описание ссылается на некоторый объект $O_n$, то можно, получить все объекты “наследующие

данные“ от $O_n$ , автоматически расширив тем самым “учебную коллекцию”.

Схема связи сущностей при этом выглядит так:

*[Учетная запись книгообеспеченности] ->*
*[Запись библиографического каталога] ->*
*[Объекты DL]*.

В соответствии с такой схемой и происходит начальное формирование коллекции в автоматическом режиме:

- стандартная выборка литературы книгообеспеченности по некоторой дисциплине;
- извлечение из АБИС библиографических описаний для каждого отобранного наименования;
- формирование по библиографическим данным массива ссылок на объекты ЭК;
- формирование из отобранных объектов виртуальной коллекции с прикрепленными к ней библиографическими описаниями (если это необходимо).

На заключительных этапах редактирование полученных коллекций может производиться экспертом.

При изменении параметров и данных модуля “Книгообеспеченность” возможно автоматическое перестроение соответствующих коллекций. Таким образом, достигается максимальная эффективность и оперативность использования данных.

### 4.3 Использование дополнительной и преиодической литературы

Следует отметить, что описанный подход применим для формирования “учебных коллекций” различных типов. Основными являются коллекции на основе списка основной литературы по дисциплинам. Обычно это учебники и учебные пособия, хранящиеся в библиотеке и необходимые по стандартам образования для изучения дисциплин. Но не менее важными (а во многих случаях и более важными) являются наборы дополнительной литературы. Помимо учебников это касается изданий, выпускаемых небольшими тиражами, периодических изданий и Интернет-ресурсов.

Небольшими тиражами, часто прямо в переделах ВУЗа, издаются различные методические указания, авторами которых являются преподаватели, непосредственно обучающие студентов. Эти издания обычно каталогизируются в АБИС ВУЗа и изначально существуют в электронном виде, т. е. их исходные тексты легкодоступны. Таким образом, однажды добавив в каталог книгообеспеченности такое издание - можно гарантированно включить его во все “учебные коллекции” по дисциплине. Преподаватель в результате этого получает в распоряжение готовый информационный ресурс.

Аналогичная ситуация с периодическими изданиями:

- все поступления в библиотеку каталогизируются;
- при организации работ по переводу журналов в электронный вид их можно помещать в DL;
- далее, поместив описание периодического издания (или отдельной статьи в нем) в каталог модуля книгообеспеченности, можно автоматически включать их в “учебные коллекции”.

## Литература


[1] Абросимов А. Г., Зуев Д. С. Научно-образовательная электронная библиотека ВУЗа. // Труды десятой всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" RCDL’2008, (Дубна, 7 – 11 октября 2008 г.). - Казань: Казанский государственный университет, 2008. - С. 374 – 379.

[2] Амалиева Г. Г., Елизаров А. М. Создание электронной коллекции личных дел студентов Казанского университета (1917-1925 гг.). // Труды десятой всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" RCDL’2008, (Дубна, 7 – 11 октября 2008 г.). - Казань: Казанский государственный университет, 2008. - С. 370 – 373.

[3] Котляр Э. А., Гужеля Д. Ю., Полихина Н. А. Результаты мониторинга и основные рекомендации по вопросам развития электронных библиотек ВУЗов. // Труды десятой всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" RCDL’2008, (Дубна, 7 – 11 октября 2008 г.). - Центр инновационных технологий и образования в науке, 2008. - С. 380 – 386.

[4] Курчинский Д. Н., Палей Д. Э., Смирнов В. Н. Информационная система учреждения культуры - как система обработки объектов электронного каталога // Труды Седьмой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" RCDL’2005, (Ярославль, 4 - 6 октября 2005 года). - Ярославль: Ярославский государственный университет им. П. Г. Демидова, 2005. С. 92 - 97.

[5] Палей Д. Э., Курчинский Д. Н., Смирнов В. Н. АБИС "Буки". Первые итоги развития, новые возможности, перспективы на будущее // Информационные технологии, компьютерные системы и издательская продукция для библиотек: Доклады и тезисы докладов. Седьмая Международная конференция и Выставка “LIBCOM-2003”, пансионат “Ершово”, Звенигород, Московская область, 17 – 21 ноября 2003 г. – Москва: ГПНТБ России, 2003. – С. 192 – 200.

[6] Палей Д. Э. Учет книгообеспеченности в библиотеке ВУЗа: проблемы и пути их решения / Палей Д. Э., Курчинский Д. Н., Смирнов В. Н.

// Информационные технологии, компьютерные системы и издательская продукция для библиотек: Доклады и тезисы докладов. Восьмая Международная конференция и Выставка “LIBCOM-2004”, пансионат “Ершово”, Звенигород, Московская область, 15 - 19 ноября 2004 г. – М.: ГПНТБ России, 2004. – С. 155 – 160.

[7] Палей Д. Э. Опыт внедрения учета книгообеспеченности в библиотеке ВУЗа / Палей Д. Э., Смирнов В. Н., Курчинский Д. Н. // Библиотеки и образование: Сборник материалов первой Международной конференции, Ярославль, 19 – 22 апреля 2005 г. – Ярославль: МУБиНТ, 2005. – С. 107 – 112.

[8] Палей Д. Э., Курчинский Д. Н., Смирнов В. Н. Цифровая библиотека Ярославского региона. Итоги работы, перспективы развития. // Труды пятой всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" RCDL’2003, (Санкт-Петербург, 29 – 31 октября 2003 г.). - Санкт-Петербург: НИИ химии СПбГУ, 2003. - С. 315 – 319.

[9] Российский коммуникативный формат представления библиографических записей в машиночитаемой форме (Рос. вариант UNIMARC), СПб.: Изд-во РНБ, 1998.

[10] Тимонина Л. С., Шарова Т. С. Применение библиографических ссылок в автоматизированном учете книгообеспеченности и в представлении результатов поисковых запросов. // Труды десятой всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" RCDL’2008, (Дубна, 7 – 11 октября 2008 г.). - Библиотечный комплекс университета "Дубна", 2008. - С. 76 – 82.

[11] Форматы USMARC. Краткое описание. В 3-х ч. / Пер. с англ.; ГПНТБ России. – М., 1996.

# Institute Of Higher Education Digital Library As System For Creating Educational Multimedia Objects Catalogue

Kurchinsky D., Paley D., Smirnov V.

Institute of higher education digital library application for generate educational multimedia collections is considered in this paper.

Authors suggest using the universal library system as the foundation of this work. The positive and negative aspects of this solution are discussed in the article.

# Информационная система для создания и управления электронными коллекциями графических документов ♣

А.А. Рогов, К.А. Рогова, П.В. Кириков, М.Ю. Быстров

Петрозаводский государственный университет
rogov@psu.karelia.ru, ksushar@mail.ru, lispad@gmail.com, maksimkab@yandex.ru

**Аннотация**

Разрабатываемая информационная система позволяет создавать коллекции графических документов, хранить различные наборы графических (основанных на цветовосприятии и текстурных характеристиках) и текстовых параметров для каждого графического документа, выполнять классификацию и поиск по различным комбинациям параметров, а так же на основе сходства изображений. Система разрабатывается под Интернет.

## 1 Введение

В настоящее время все большую популярность получают электронный коллекции графических документов. Обычные пользователи создают свои фотоальбомы (не только на дома, на локальном компьютере), но и в сети Интернет; научные сообщества используют большие объемы графической информации в своих исследованиях. К сожалению, сейчас нет единой системы, которая позволила бы не только хранить изображения в определенном порядке, но и классифицировать графическую информацию по выделенным параметрам и осуществлять поиск.

Целью данной работы является создание информационной системы, позволяющей создавать, управлять и анализировать коллекции графических документов. Особенностью создаваемой информационной системы является возможность хранения и использования иерархии документов и значений наборов признаков, зависящих от типа документа. Создаваемое в коллекции дерево для хранения иерархии графических объектов может иметь неограниченное число уровней и неограниченное число узлов на каждом уровне. Не смотря на то что, создаваемая система предназначена в первую очередь для научных исследований, подобный способ описания графических объектов позволяет применять систему для создания каталога запчастей/товаров с иллюстрациями, произведений художников и т.д.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

## 2 Общие характеристики разрабатываемой системы

Информационная система предназначена для размещения на WEB-сервере и обеспечивает удобный доступ через сеть Интернет с любого подключенного к сети устройства. Она пригодна для решения большого ряда задач, среди которых можно выделить:

- создание и редактирование иерархической структуры коллекций графических документов;
- хранение различных наборов параметров для каждого графического документа;
- классификация и поиск графических документов по различным комбинациям параметров, а так же на основе сходства текстур, цветового восприятия и т.д.;
- описательная статистика коллекции;
- разделение доступа к системе.

Система логически разделена на две части – пользовательскую и административную и предоставляет простой и удобный интерфейс. Для ввода информации будут пригодны графические документы в различных форматах, автоматически осуществляться их приведение к единому стандарту, и в автоматизированном режиме предоставлять возможность выделять объекты на них.

Главное отличие предлагаемой системы от существующих систем создания электронных фотоальбомов состоит в возможности приписывать графическому документу набор индивидуальных признаков и осуществлять поиск по выделенной комбинации признаков. На основе признаков автоматически производится классификация объектов коллекции с целью поиска наиболее близких между собой. Кроме того, пользователю предлагается статистическая информация о наличии в коллекции объектов с выделенным набором признаков и анализ выделенных признаков статистическими методами.

Информационная система реализуется на php с использованием web-сервера apache и сервера баз данных mysql. В данный момент прототип информационной системы апробируется на основе коллекции графических документов петроглифов Северной Фенноскандии.

## 3 Административная и пользовательская составляющие системы

Для зарегистрированных пользователей системы, в административном разделе заложены дополнительные функции работы с графической информацией: пополнение информации, редактирование и удаление; создание набора признаков; присвоение признаков изображениям, создание иерархии рисунков и т.д.

Рассмотрим работу этого модуля на примере создаваемой коллекции петроглифов Северной Фенноскандии. В Северной Фенноскандии выделены несколько крупных местонахождения наскальных рисунков, среди которых мы рассматриваем Норвегию, Мурманскую область и Карелию. В каждом месте выделяют более мелкие группы, потом еще более мелкие и т.д., пока не доходят до сюжетных схем и отдельных петроглифов. Наглядным представлением такого расположения являются иерархические наборы карт, схем и рисунков. Для этого необходимо загрузить карты в систему и выделить соответствующие точки местонахождения. Точно так же можно работать со схемой петроглифов: квадратной областью выделяется петроглифов и в базу данных заносится вся необходимая информация о нем. Таким образом, с этой часть системы может работать специалист в любой области, без необходимости изучения языков работы с базами данных. Основными пользователями этого раздела будут создатели коллекции. Для больших научных коллективов (сетевого научного сообщества) возможна распределенная работа научного коллектива без принудительной синхронизации получающейся базы данных.

Для обычных пользователей система является открытой и доступна через Интернет любому интересующемуся. Доступ к функциям системы в таком случае ограничен лишь «чтением» и «анализом». То есть, пользователь может работать с системой как с обычным сайтом, классифицировать и искать изображения по выделенным признакам, но изменение и дополнение информации для него невозможно.

## 4 Типы признаков изображений

Все признаки можно разделить на 2 части. Первая - значения признаков вводит администратор системы. При этом, для вычисления некоторых признаков возможна частичная автоматизация. Во второй группе признаки получаются в автоматическом или авторизованном виде и касаются параметров цветовосприятия и текстур изображения.

В первой группе каждое изображение может быть описано по параметрам следующего вида:

- количественные характеристики изображений;
- номинальные переменные с отношением порядка и категоризированные переменные, кроме того, отдельные признаки могут иметь более сложную фасетную структуру, которую можно описать с помощью графа;
- текстовые описательные признаки (они не используются при статистическом анализе);
- фрагменты изображений: взаимосвязь и повтор объектов, описываются с помощью ориентированных и неориентированных графов.

Приведем пример фасетной структуры на основе признаков петроглифов Северной Феноскандии. Примерами признаков для лосей/оленей, птиц и лодок являются следующие:

<u>Для лосей и оленей:</u>

1. голова
   - o удлиненная
   - o укороченная
   - o нормальная
2. уши
   - o наличие
   - o отсутствие
3. холка
   - o наличие
   - o отсутствие
4. рога
   - o отсутствие
   - o лося
   - o оленя
   - o ни те, ни другие
5. шея
   - o короткая
   - o длинная
   - o нормальная
6. шея
   - o утолщенная
   - o узкая
   - o нормальная
7. серьга на шее
   - o наличие
   - o отсутствие
8. корпус
   - o массивный
   - o грузный
   - o линейный
9. изгиб спины
   - o внутрь
   - o вне
   - o отсутствует

10. изгиб живота
    - o внутрь
    - o вне
    - o отсутствует
11. изгиб передней пары ног
    - o внутрь
    - o вне
    - o отсутствует
12. изгиб задней пары ног
    - o внутрь
    - o вне
    - o отсутствует
13. хвост
    - o отсутствует
    - o короткий
    - o удлиненный
14. животное обращено
    - o вправо
    - o влево

Для птиц:

1. Фигура
    - o реалистичная
    - o схематизированная
2. степень выбивки
    - o контурное с заполнением
    - o контурное без заполнения
    - o силуэтная
3. ориентация фигуры
    - o правая
    - o левая
4. лапы
    - o одна лапа
    - o две лапы
    - o нет лап
5. хвост
    - o наличие
    - o отсутствие
6. шея
    - o длинная
    - o короткая
7. шея
    - o прямая
    - o изогнутая
8. шея (угол между головой и нижней линией туловища)
    - o прямой
    - o тупой
    - o острый
9. клюв
    - o не выделен
    - o длинный
    - o короткий

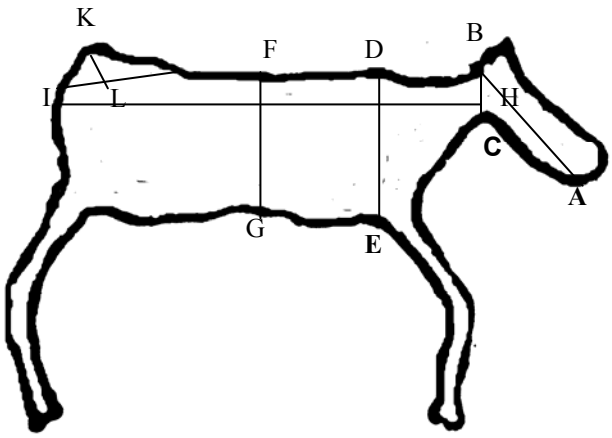
Для лодок:

1. Фигура
    - o силуэтные
    - o контурные
2. длина
    - o длинная
    - o короткая
3. борта
    - o высокие
    - o низкие
4. пассажиры
    - o наличие
    - o отсутствие
5. пассажиры
    - o в виде столбиков
    - o реалистично
6. Носовое украшение
    - o отсутствие
    - o в виде головы лесного животного
    - o в виде головы птицы
7. киль
    - o отсутствует
    - o один
    - o два
8. угол наклона форштвеня к корпусу лодки
    - o прямой
    - o тупой
    - o острый
9. корпус
    - o прямой
    - o изогнутый
    - o расширяющийся
10. ориентация фигуры
    - o правая
    - o левая

Возможно вычисление некоторых параметров изображения в полуавтоматическом режиме с использованием встроенного функционально программируемого калькулятора. Проиллюстрируем его работу на примере изображений лосей и оленей.

Определение некоторых признаков, таких как толщина шеи, тип головы, тип корпуса и т.д. визуально по рисунку не всегда является точным и часто зависит от эксперта. Для того, чтобы избежать двусмысленности восприятия предлагается использовать формулы отношения между длинами соответствующих отрезков.

На изображение петроглифа лося/оленя выделяют следующие отрезки, приведенные на рисунке: BA - длина головы, HI - длина корпуса, KL - длина хвоста, BC - толщина шеи, DE, FG - толщина передней и средней частей корпуса. Для



того, чтобы отметить эти отрезки на рисунке в программе, необходимо мышкой отметить точки начала и конца отрезков. После этого автоматически высчитываются длины этих отрезков. Чтобы определить значение признака, необходимо сравнить длины соответствующих отрезков. В этом случае формулы задаются пользователем. Имеется возможность использовать все арифметические операции, а так же операции сравнения. При выполнении всех действий учитывается вычислительная погрешность.

Используя подобный калькулятор происходит формализация восприятия изображения и упрощается ввод характеристик и они становятся более точными.

Выделение фрагментов изображения и их взаимосвязь происходит с помощью специального графического интерфейса при помощи мыши прямо в окне браузера. Данный признак описываются с помощью ориентированных и неориентированных графов.

Во второй группе признаков выделим:

- характеристики текстуры;
- характеристики цветовосприятия.

Одним из стандартных способов представления цветовой характеристики изображения является цветовые гистограммы. Для ее построения пространство всех цветов разбивается на подмножества так, чтобы схожие цвета попали в один интервал. Для каждого интервала подсчитывается количество пикселей, чей цвет принадлежит данной области. Для анализа гистограмм используются различные метрики, например, сумма модулей разностей значений элементов гистограмм для каждой цветовой области [3]. Вместо гистограммы можно брать вектор цветовой когеренции [2]. Другим вариантом представления является статистическая модель: рассматривается статистическое распределение различных цветовых каналов. Сравнение распределений является оценкой схожести [4]. Кроме того, можно рассматривать не только одномерные распределения, но и трехмерные, учитывая все взаимосвязи между каналами (ковариации). Рассматриваются интервалы наиболее часто встречающихся цветов, размеры одноцветных цветовых фрагментов, перевод цветных изображений в бинарное и их анализ.

Для анализа текстур, одним из применяемых методов является анализ независимых компонент. С его помощью выделяют фильтры, которые признаны отражать основные направления текстур для той базы изображений, на основе которой они строятся [3]. Кроме этого используется спектр фрактальной размерности Реньи [3, 4].

Возможно вычисление некоторых параметров изображения в полуавтоматическом режиме с использованием встроенного функционально программируемого калькулятора. Проиллюстрируем его работу на примере изображений лосей и оленей.

Определение некоторых признаков, таких как толщина шеи, тип головы, тип корпуса и т.д. визуально по рисунку не всегда является точным и часто зависит от эксперта. Для того, чтобы избежать двусмысленности восприятия предлагается использовать формулы отношения между длинами соответствующих отрезков.

## 5. Анализ признаков

С помощью специального модуля корреляционного анализа введенные признаки можно проверить на статистическую независимость с помощью критерия $\chi^2$ Пирсона. Для этого выделяют признаки, группы объектов и задают уровень значимости. Кроме того, для анализа признаков используются методы описательной статистики.

## 6. Алгоритмы классификации и кластеризации

Для различных задач и типов признаков используются различные методы классификации и кластеризации. Для этого создаются модули статистического анализа документов. Рассмотрим первую группу признаков. Признак можно описать как $f : X \to D_f$, где $D_f$ - множество допустимых значений признака, тогда, если заданы признаки $f_1,\ldots, f_n$, то вектор $x = (f_1(x),\ldots, f_n(x))$ называется признаковым описанием графического документа. Ввиду различия множеств допустимых значений для различных признаков для корректной работы алгоритма выполняется нормировка значений признаков: $\tilde{f}_j = \frac{f_j(x) - \min(f_j)}{\max(f_j) - \min(f_j)}$. В этом случае значение каждого признака будет лежать в пределах $[0,1]$. В качестве меры расстояния между документом и эталоном берётся эвклидово расстояние $(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$. Для классификации объектов коллекции на основе признаков применяются методы дискриминантного и кластерного анализов. Например, методом иерархического кластерного анализа (метод ближайшего соседа).

Классификация по цветовому и текстурному анализу может быть осуществлена несколькими методами. Примерами являются поиск по цветовым моментам и цветовым гистограмм [1]. Кроме того, существуют методы поиска нечетких дубликатов 12]. Поиск нечетких дубликатов позволяет предположить, являются ли два объекта частично

одинаковыми или нет. Частично одинаковые изображения могут образовывать один кластер. Кроме того, схожесть изображений по степени цветового восприятия может быть осуществлена на: сравнительной площадь белого, при наличии большого фрагмента определенного цвета и т.д.

## 7. Алгоритмы поиска изображений

Управление типами документов, а также наборами признаков, общих для всех графических документов коллекции и уникальных для каждого типа, позволяет хранить разнообразные данные о каждом графическом документе, организовать поиск документов по типу, признаку или набору признаков, задав точное значение или границы варьирования значения каждого признака и точность поиска (количество совпадений признаков для номинальных и категоризированных и интервалы изменения для количественных переменных).

Поиск по изображениям предназначен для поиска изображений, похожих на данное или на его фрагмент. На вход подается исследуемое изображение, а на выходе должны появиться изображения из базы данных, наиболее похожие на исходное. Для поиска похожих графических объектов на основе текстур изображений – методы нейронных сетей и геометрического программирования.

Рассмотрим его на примере наскальных изображений северной Фенноскандии. На сегодняшний день все новейшие материалы по петроглифам представляют собой набор цветных фотографий. Определенную сложность поиска создает фактическое отсутствие некоторых частей изображения. Поиск также осложняется тем, что часто невозможно определить, где верх, а где низ изображения. При этом, требование, что при поиске необходимо только совпадение контура изображения, позволяет упростить поиск, а значит, изображение петроглифа можно рассматривать, как бинарное (скале соответствует белый цвет, а петроглифу - черный). В зависимости от выбранных параметров поиска (точность поиска, процент совпадений элементов изображений) будет найдено одно или несколько изображений. Для поиска используются сеть адаптивного резонанса и структурный метод поиска.

В результате поиска, пользователю предоставляется доступ к информации о кодовом номере, месторасположении, характерных признаках найденного петроглифа и петроглифах, близких к нему по ранее описанным признакам.

## 8 Технические особенности реализации

Для создания и функционирования информационной системы использовалось свободно распространяемое программное обеспечение, и использование системы не нарушает лицензионных соглашений третьих сторон. В качестве WEB-сервера был использован сервер Apache/1.3.23, интерпретатор PHP 4.4.9, сервер баз данных MySQL 3.23.49. Одним из возникших вопросов был вопрос о средстве хранения графических документов. Было рассмотрено два варианта: хранение изображений в файлах и в базе данных в биде BLOB. Преимущество первого способа - более высокая скорость работы, второго - более простой контроль над целостностью данных и контролем доступа к ним. Были произведены вычислительные эксперименты, которые показали что, скорость чтения файлов сравнима со скоростью извлечения информации из БД (порядок $10^{-2}$ сек). При этом время обработки (масштабирования) изображения СБ2-модулем интерпретатора PHP имеет порядок $10^{-1}$ сек. Таким образом, ввиду незначительности потерь в скорости доступа при большем числе положительных сторон был выбран метод хранения в БД. В данной реализации используется тип MEDIUMBLOB позволяющий хранить до 16 мегабайт данных.

## 9 Преимущества разрабатываемой системы

Первой попыткой создания информационной системы была локальная версия [3]. С ее помощью была создана электронная коллекция петроглифов Карелии. В созданной ранее системе для навигации по каталогу групп петроглифов использовала дерево, жестко прописанное в программном коде системы. Для добавления новой группы необходимо было перекомпилировать продукт, что вызывало определенные трудности. Интерфейс программы содержал графические схемы групп, позволяя пользователю кликом мышки выбирать подгруппу или петроглиф и получать о них информацию. В системе использовались изображения схем в BMP формате, вручную подготовленные в графическом редакторе: необходимо было раскрасить области перехода на следующий уровень в оттенки красного цвета, так чтобы код цвета соответствовал коду подгруппы. Число возможных узлов на уровне было ограничено 255 возможными оттенками красного цвета, при изменении только составляющей красного цвета в палитре RGB.

Использование цветового кодирования представлялось невозможным при создании WEB-реализации информационной системы: размер BMP изображений схем достигал десятков мегабайт, что неприемлемо для загрузки с сервера даже при быстром Интернет-соединении. При конвертации изображения в другие форматы (JPEG/PNG) для сокращения размера происходит частичная потеря цветовой информации, и полученную схему невозможно использовать для навигации теми же средствами, что в локальной системе. Этот метод привносил трудности для добавления новых

разделов в дерево графических документов - возникала необходимость ручной раскраски схем. Решением данного вопроса стал отказ от цветового кодирования. Для навигации используются отмеченные прямоугольные области на изображении, а в административной части администратору системы предлагается графический интерфейс для разметки изображения с помощью мыши прямо в окне браузера. В базе данных сохраняются относительные координаты начала и конца области, что позволяет использовать навигацию при отображении схем в различных масштабах.

## Заключение

В настоящее время, кроме создаваемой с помощью системы коллекции петроглифов Северной Фенноскандии, система апробируется на материалах Карельского государственного краеведческого музея при создании коллекции открыток и коллекции фотографий со строительства Беломорско-Балтийского канала.

## Литература


[1] Васильева Н., Марков И. Синтез цветовых и текстурных признаков при поиске изображений по содержанию. // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. – Санкт-Петербург: НУ ЦСИ, 2008, С. 135-144.

[2] Кисель Я. Алгоритм поиска нечетких дубликатов в коллекции изображений. // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. – Санкт-Петербург: НУ ЦСИ, 2008, С. 170-173.

[3] Рогов А.А., Рогова К.А., Спиридонов К.Н., Быстров М.Ю. Система поиска в электронной коллекции изображений петроглифов Карелии. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 10 Всероссийской научной конференции "RCDL-2008"(Дубна, Россия 7-11 октября 2008г.). - Дубна: ОИЯИ, 2008. С. 246-251.

[4] Рогов А.А., Спиридонов К.Н. Применение спектра фрактальных размерностей Реньи как инварианта графического изображения. // Вестник Санкт-Петербургского университета. Сер. 10. 2008. Вып. 2. С. 30-43.

[5] Sticker M., Dimai A. Color Indexing with Weak Spatial Constraints. In Proceeding of the SPIE Conference, 1996.


## The Information System for Graphic Documents Electronic Collections Creating and Administration


A.A. Rogov, K.A. Rogova, P.V. Kirikov, M.Yu. Bystrov



Creating information system allows to develop collections of graphic documents, to storage different sets of grahic (based on color perception and texture characteristics) and text parameters for each document, to carry out classification and search by different parameters combinations and by depictions similarity. The system is designed for Internet.



♣ Исследования поддержаны грантом РГНФ № 08-01-12116в (руководитель Н.В. Лобанова).

# Повышение достоверности обработки данных на основе избирательного избыточного кодирования семантических единиц текста

© С.В. Минаков, О.А. Финько

Краснодарское высшее военное училище (военный институт)
имени генерала армии С.М.Штеменко
ofinko@yandex.ru

**Аннотация**

Рассматриваются пути повышения достоверности текстовых данных на основе использования гибридной семантико-кодовой избыточности применительно к семантическим единицам текста наиболее подверженных ошибкам.

## 1 Введение

Текстовый тип данных в ближайшем будущем все еще будет составлять основу документированной информации. Для повышения достоверности электронных документов в настоящее время успешно применяются всевозможные избыточные коды [2], а в необходимых случаях и электронная цифровая подпись (ЭЦП). Как известно избыточность текста носит крайне неравномерный характер. В одних случаях морфологическую ошибку легко распознать и исправить, руководствуясь здравым смыслом и грамматикой языка. А в других - распознать ошибку, основываясь только на избыточности естественного языка, невозможно (цифровые последовательности, слова, например, такие как «июль» и «июнь», буквенные литеры, пароли в системах разграничения доступа и т.п.).

Текстовая и файловая обработки информации могут существенно отличаться друг от друга. Текст необходимо создавать и редактировать. В некоторых наиболее ответственных случаях текст даже обрабатывается и передается побуквенно (криптография, шифр гаммирования). Более того, при хранении и передаче особо ценной информации [5] распределенный принцип обработки просто необходим. В этих случаях «Шенноновский» отрыв от смысловой нагрузки на слово не всегда положителен. Поэтому применения традиционных методов повышения достоверности файлов, основанных на ЭЦП и избыточном кодировании недостаточно.

## 2 Неравномерность избыточности текста

Текст неравномерен по смысловой нагрузке, следовательно, требования к повышению достоверности семантических единиц (СЕ) текста, влияющих на смысл, должны быть разными. Целесообразно повысить достоверность СЕ вероятность искажения, которых максимально влияет на смысл текста, и применять к ним методы повышения достоверности.

Одним из известных методов повышения достоверности текстовых данных является внесение семантической и лингвистической избыточностей (повторы текстовых единиц, применение синонимов, перефразировки, повторы цифровых данных словами и т.д.). Другим путем повышения достоверности текстовых данных является помехоустойчивое кодирование [2], достоинством которого является возможность обнаруживать или исправлять ошибки на выходе конечных устройств. Недостаток – кодирование не учитывает смысл (ценность) передаваемой информации.

В основе предлагаемого решения повышения достоверности СЕ лежит автоматическое определение СЕ минимальное искажение которых может привести к трансформации в другое существующее (семантически близкое) слово текста на каком-либо естественном языке. Причем предлагается СЕ сравнивать не с множеством СЕ, составляющих текст, а с множествами образованными онтологическими рядами относительно анализируемой СЕ. Для выявления семантически близких СЕ применим метод динамического программирования [6]. В процессе вычислений значения $d_{i,j}$ (операции удаления, вставки, замены) записываются в массив

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

$(m+1)(n+1)$, вычисляются с помощью следующего рекуррентного соотношения:

$$d_{i,j} = \min\left\{d_{i-1,j} + w(a_i,\varepsilon), d_{i,j-1} + w(\varepsilon,b_j), d_{i-1,j-1} + w(a_i,b_j)\right\},$$

где:

$a_i$ - СЕ текста множества $A$ соответствующего вводимому тексту;

$b_j$ - слова, содержащиеся в базе данных множества $B$ (эти слова объединяются в онтологические ряды путем разбиения единой базы данных по тематическим признакам);

$w(a_i,b_j)$ - цена преобразования символа $a_i$ в символ $b_j$.

Ниже приведен массив (табл. 1), полученный при вычислении расстояния Левенштейна между строками «Моховое» и «Меховое». Из него видно, что расстояние между этими строками, то есть $d_{7,7}$, равно 1.

Таблица 1

Пример вычисления расстояния Левенштейна

| | j | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| i | | | м | е | х | о | в | о | е |
| 0 | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | м | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 2 | о | 2 | 1 | 1 | 2 | 3 | 4 | 5 | 6 |
| 3 | х | 3 | 2 | 2 | 1 | 2 | 3 | 4 | 5 |
| 4 | о | 4 | 3 | 3 | 2 | 1 | 2 | 3 | 4 |
| 5 | в | 5 | 4 | 4 | 3 | 2 | 1 | 2 | 3 |
| 6 | о | 6 | 5 | 5 | 4 | 3 | 2 | 1 | 2 |
| 7 | е | 7 | 6 | 5 | 5 | 4 | 3 | 2 | 1 |

## 3 Избирательное избыточное кодирование семантических единиц текста

Для выбора СЕ текста (рис. 1) возьмем $d = 1$.

Данному условию удовлетворяют СЕ выделенные рамкой (все цифры; название месяцев «июль/ июнь»; фамилии «Рыбалко/ Рыбалка/ Рыбалков», «И.Сталин/ В.Сталин/ И.Салин», название населенных пунктов «Моховое/ Меховое/ Мохово»; «р.Ока/ р.Оса/ р.Ака» и т.п.).

Срочно

ТЕЛЕГРАММА

Командующему Брянским фронтом

Ставка Верховного Главнокомандования приказывает: 3-й танковой армии Рыбалко в 08:00 20 июля нанести удар в направлении Протасово, Отрада, к исходу дня 20 июля перерезать шоссейную и железную дороги Мценск — Орел и, развивая в течение 21 июля наступление на Мценск с юга, совместно с 3-й армией Горбатова завершить уничтожение мценской группировки противника и освободить город Мценск.

После выполнения этой задачи 3-ю танковую армию Рыбалко направить на юг с целью перерезать железную дорогу Моховое — Орел и содействовать 63-й армии Колпакчи в выходе ее также на р. Ока.

И. Сталин

02 июля 1942. 16:05

№ 8/0123

Рис. 1. Пример выделения СЕ в тексте телеграммы

Присвоим выделенным СЕ текста (рис. 1) кодовые обозначения (табл. 2).

Таблица 2

Пример кодирования алфавита

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| А | - | 0 | Т | - | 18 | ( | - | 40 |
| Б | - | 1 | У | - | 19 | ) | - | 41 |
| В | - | 2 | Ф | - | 20 | , | - | 44 |
| Г | - | 3 | Х | - | 21 | — | - | 45 |
| Д | - | 4 | Ц | - | 22 | . | - | 46 |
| Е | - | 5 | Ч | - | 23 | / | - | 47 |
| Ж | - | 6 | Ш | - | 24 | : | - | 58 |
| З | - | 7 | Щ | - | 25 | ; | - | 59 |
| И | - | 8 | Ъ | - | 26 | 0 | - | 48 |
| Й | - | 9 | Ы | - | 27 | 1 | - | 49 |
| К | - | 10 | Ь | - | 28 | 2 | - | 50 |
| Л | - | 11 | Э | - | 29 | 3 | - | 51 |
| М | - | 12 | Ю | - | 30 | 4 | - | 52 |
| Н | - | 13 | Я | - | 31 | 5 | - | 53 |
| О | - | 14 | ! | - | 33 | 6 | - | 54 |
| П | - | 15 | ? | - | 34 | 7 | - | 55 |
| Р | - | 16 | № | - | 36 | 8 | - | 56 |
| С | - | 17 | ¶ | - | 37 | 9 | - | 57 |
| | | | | | | пробел | - | 32 |

Вычислим для каждой полученной, таким образом, числовой последовательности проверочное число. Например, суммируем каждую последовательность по модулю 64:

$$M = \left(\sum_{i=1}^{n} a_i\right) \bmod 64,$$

где $n$ – количество символов в СЕ.

Таблица 3

Пример базы онтологических рядов для анализируемого текста

| Тематические признаки | Элементы множества $B$ |
|---|---|
| годы | …, 1939, 1940, 1941, 1942, 1943, 1944, 1945 …, 2008, 2009, … |
| месяцы года | январь, февраль, март, апрель, май, июнь, июль, август, сентябрь, октябрь, ноябрь, декабрь |
| даты | 01, 02, 03, 04, 05, … , 29, 30, 31 |
| время | 00:00, 00:01, … 01:00, 01:01, 01:02, … , 23:58, 23:59 |
| военачальники ВОВ (маршалы) | Василевский, Говоров, Жуков, Конев, Малиновский, Мерецков, Рокоссовский, Сталин, Тимошенко, Толбухин |
| военачальники ВОВ (генералы арий) | Антонов, Апанасенко, Василевский, Еременко, Жуков, Конев, Малиновский, Мерецков, Павлов, Попов, Рокоссовский, Соколовский, Тюленев, … |
| орган управления | отделение, взвод, …, дивизия, корпус, армия, …, Генеральный Штаб. |
| виды и рода ВС | авиация, ВВС, сухопутные, …, морские, ВМФ, связь |
| воинские звания | рядовой, ефрейтор, младший сержант, …, генерал-армии, маршал |
| вооружение (танки) | Т-34, Т-34-57, ОТ-34, ТО-34, КВ-1с, КВ-85, ИС-1, ИС-3, ИСУ-152, ИСУ-122, ИСУ-122С, … |
| … | … |
| географические наименования (реки) | Большая Чернава, Быстрая Сосна, Кшень, Нерусса, Общерица, Ока, Олым, Орлик, Семенек |
| географические наименования (города) | …, Мценск, Мымрино, Мыцкое, Навесное, Навля, …, Шатилово, Шахово, Шашкино, Щербово, Юшково, Яковлево, … |
| … | … |

Полученные проверочные числа добавим к исходным выделенным СЕ. Для отличия их от текста введем маркер – ## (редко встречающееся сочетание символов) и двухзначные проверочные цифры, соответствующие числу $M$ избыточного кода, скрытые от пользователя (рис. 2).

Принимающая сторона вычисляет для соответствующих СЕ проверочные символы и сравнивает их значение с прикрепленными. При несовпадении результатов адресат делает вывод об искажении данной СЕ текста.

Элементами $b_j$ единой базы данных $B$ являются семантические единицы передаваемых сообщений. Пример таких СЕ представлен в таблице 3.

## 4 Заключение

Введение семантико-кодовой избыточности позволит повысить достоверность данных при соблюдении принципа *распределенной* (посимвольной) обработки, хранения и передачи. Причем предложенный метод может использоваться и для обработки документированной информации на «твердых» (бумажных) носителях.

В рассмотренном случае был применен достаточно простой избыточный код с контролем по модулю. Однако предполагается использовать более подходящие для данной задачи многозначные коды [1, 3, 4, 7, 8], избыточность которых будет устанавливаться адаптивно в соответствии с избыточностью и ценностью СЕ текста.

Срочно

ТЕЛЕГРАММА

Командующему Брянским фронтом

Ставка Верховного Главнокомандования приказывает:
3**##51**-й танковой армии Рыбалко**##15** в 08:00**##18** 20**##34** июля**##18** нанести удар в направлении Протасово**##46**, Отрада**##52**, к исходу дня 20**##34** июля**##16** перерезать шоссейную и железную дороги Мценск**##15** — Орел**##46** и, развивая в течение 21**##35** июля**##16** наступление на Мценск**##15** с юга, совместно с 3**##51**-й армией Горбатова**##04** завершить уничтожение мценской**##38** группировки противника и освободить город Мценск**##15.**
После выполнения этой задачи 3**##51**-ю танковую армию Рыбалко**##15** направить на юг с целью перерезать железную дорогу Моховое**##18** — Орел**##46** и содействовать 63**##41**-й армии Колпакчи**##27** в выходе ее также на р. Ока**##54.**
И. Сталин**##25**
02**##34** июля**##16** 1942**##16.** 16:05**##06**
№ 8/0123**##45**

Рис. 2. Текст документа с введенной кодовой избыточностью

## Литература


[1]. Кирилов, А.А. Коды с произвольным основанием, исправляющие одиночные ошибки/ А.А. Кирилов. – В кн.: Проблемы кибернетики. - М.: Наука, 1970, Вып. 22, С. 282-287.

[2]. Питерсон, У. Коды, исправляющие ошибки/ У. Питерсон, Э. Уэлдон. – М.: Мир, 1976. – с. 596.

[3]. Тимофеев, Б.Б. Методы обнаружения ошибок в алфавитно-цифровых последовательностях на этапе подготовки и ввода данных в ЭВМ/Б.Б. Тимофеев, В.А. Литвинов. – УСиМ, 1977. №4, С. 20-27.

[4]. Четвериков, В.Н. Подготовка и телеобработка данных в АСУ/ В.В. Четвериков. – М.: Высшая школа, 1981.

[5]. Шанкин, Г.П. Ценность информации. Вопросы теории и приложений/ Г.П. Шанкин. – М.: Филоматис, 2004. – с. 128.

[6]. Graham, Stephen. String Search/, Stephen A. Graham. – UK. School of Electronic Engineering Science University College of North Wales, 1992. – p. 103.

[7]. Herr, J.R. Self-checking number system/ J.R. Herr. – Comput. Des., 1974, vol. 13, N 1, p. 85-91.

[8]. Sethi, A.S. An error-correcting coding scheme for alphanumeric data/ A.S. Sethi, V. Rajaraman, P.S. Kenjale. – Inform. Process. Lett., 1978, N 2, p. 72-77.


## Increasing reliable processing data on base of the electoral surplus coding semantics text units


S.V. Minakov, O.A. Finko



Considering ways of increasing the validity of textual data on base the use of hybrid semantics-code redundancy with reference to semantics text units most subject to errors.

# Опыт построения системы защиты электронных библиотек от несанкционированного копирования документов♣

© Ивашко Е. Е., Никитина Н. Н.

Карельский научный центр РАН
Институт прикладных математических исследований
{ivashko, nikitina}@krc.karelia.ru

**Аннотация**

В работе описаны результаты экспериментов, проведенных в рамках разработки системы защиты электронных библиотек от несанкционированного копирования документов. Статья представляет выводы, сделанные на основании практической проверки теоретических разработок, представленных на конференции RCDL-2007.

## 1 Введение

За последнее десятилетие электронные библиотеки (ЭБ) стали важной составляющей системы формирования и распространения научного знания. Свободный доступ к результатам исследований в различных областях является залогом дальнейшего развития науки. На создание и сопровождение коллекций электронных документов тратятся большие материальные и нематериальные ресурсы. При этом, зачастую дальнейшее развитие ЭБ ставится в прямую зависимость от посещаемости (популярности) ресурса.

Однако нередко полученные из ЭБ документы используются третьими лицами для получения прибыли в обход интересов правообладателей или для создания клона исходной ЭБ. Это делает актуальной задачу защиты ЭБ от полного несанкционированного копирования документов. Под полным несанкционированным копированием здесь и далее подразумевается получение электронных копий всех или большой части цифровых документов ЭБ без разрешения ее владельцев (правообладателей).

Основная идея, лежащая в основе исследования, представленного в данной работе и статье [1], заключается в следующем.

При использовании сервисов ЭБ пользователь решает актуальные для него задачи. При этом, обращаясь к различным электронным документам, он предполагает в рамках своих задач некоторую (возможно и несуществующую в действительности) субъективную семантическую связь между интересующими его документами. Например, студент при поиске материала для реферата по истории математики, может использовать биографии А. Пуанкаре и Г. Минковского, однако вряд ли будет обращаться к их математическим статьям и монографиям. Очевидно также, что ситуация, когда один и тот же пользователь интересуется одновременно узкоспециализированными темами из области физики, искусствоведения, генетики и др., является аномальной. Мы полагаем, что интерес к разнородным (семантически не связанным) документам является аномальным и может свидетельствовать о попытке копирования большой части разнородных документов в целях, связанных с нарушениями авторских прав (например, для создания «клона» исходной ЭБ).

Для обнаружения такого аномального поведения мы используем аномальный подход в обнаружении вторжений, основанный на предположении, что вторжение проявляется как отклонение от обычного («нормального») или ожидаемого поведения пользователя, и может быть обнаружено путем сравнения последовательности действий пользователя с некоторым заданным «шаблонным» поведением.

Здесь и далее полное несанкционированное копирование документов и вторжение трактуются в одном и том же смысле.

В данной работе представлены результаты экспериментов, проведенных в рамках разработки системы защиты электронных библиотек от полного несанкционированного копирования документов. Работа, описанная в статье, является продолжением исследований, представленных в рамках конференции RCDL в 2007 г. [1].

## 2 Описание модели

В этом разделе будут кратко описаны

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

аномальный подход в обнаружении вторжений и адаптированная технология обнаружения полного несанкционированного копирования документов. Более подробное формальное описание модели и связанных с ней вопросов можно найти в работе [1].

Аномальный подход в обнаружении вторжений основан на предположении, что вторжение проявляется как отклонение от обычного («нормального») или ожидаемого поведения пользователя, и может быть обнаружено путем сравнения последовательности действий пользователя с некоторым заданным «шаблонным» поведением.

При разработке системы, реализующей аномальный подход в обнаружении вторжений, возникают следующие основные задачи:

1. построение «нормального» профиля поведения пользователя;
2. разработка классификатора, позволяющего отличить «нормальную» последовательность действий от аномальной;
3. определение граничных значений характеристик классификатора для снижения вероятности появления ошибок классификации;
4. обновление шаблонов «нормального» поведения.

Основой шаблона «нормального» поведения пользователя является Марковская цепь, построенная по записям поведения обычных пользователей ЭБ. Опишем кратко метод построения Марковской цепи.

Пусть имеется алфавит атомарных действий (например, список доступных документов ЭБ) ∑, множество всех конечных следов T* и тренировочный набор, составленный из заведомо нормальных следов $T_{tr} \in T^*$.

Расширим алфавит ∑специальным символом Ø. При построении МЦ задается параметр – «окно» размера w. Состояние в МЦ связано со следом длины w через алфавит ∑UØ, т. е. каждое состояние – набор из wсимволов алфавита ∑UØ. Переход – это пара (s, s'), определяющая в МЦ переход из состояния s в s'. Каждое состояние и переход также связаны со счетчиком количества переходов.

Операция shift(σ, x) сдвигает след σвлево и добавляет символ x в конец следа, т. е. shift(«aba», c) = «bac».

Начальное состояние МЦ определяется как след длины w, состоящий из нулевых символов, т. е. если w=3, то начальное состояние будет следом [Ø, Ø, Ø].

Операция next(σ) возвращает первый символ следа σи сдвигает σ на одну позицию влево, т.е. next(«abcd») возвращает a и обновляет след до «bcd».

Для каждого следа $\sigma \epsilon T_{tr}$, пока не обработаны все символы, входящие в алфавит, выполняются следующие шаги:

1. полагаем c = next(σ).
2. устанавливаем <следующее состояние> = shift(<текущее состояние>,c).
3. увеличиваем счетчики для состояния <текущее состояние> и перехода (<текущее состояние>, <следующее состояние>).
4. обновляем <текущее состояние> до значения <следующее состояние>.

После того, как все следы из набора $T_{tr}$обработаны, каждое состояние и переход имеют связанные с ними целые положительные числа – счетчики. Вероятность перехода из состояния sв состояние s' (P(s,s')) полагается равной N(s,s')/N(s), где N(s,s') и N(s) счетчики, связанные с переходом (s,s') и s соответственно.

По построению P является корректной мерой, т.е. выполняется следующее соотношение для всех состояний s:

$$\sum_{s \in SUCC(s)} P(s,s') = 1$$

Здесь SUCC(s)={s′: в построенной МЦ существует переход (s,s’)} определяет набор преемников s.

На рис. 1 показан пример МЦ, построенной по набору $T_{tr}$={aabc, abcbc}.

Построенная по такому алгоритму МЦ представляет собой шаблон «нормального» поведения, который создается для каждого зарегистрированного в системе пользователя.

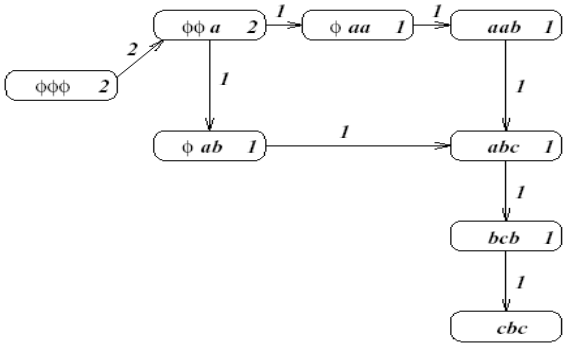



Рис. 1: Структура Марковской цепи

Целью работы, описываемой в данной статье, является проведение экспериментов и анализ их результатов, позволяющих на практике проверить эффективность модифицированного аномального метода обнаружения вторжений (представленного в [1]) применительно к обнаружению полного несанкционированного копирования документов ЭБ.

## 3 Описание экспериментов

Для проведения экспериментов была разработана программная система, выполняющая предварительную обработку файла исходных данных, построение профиля «нормального» поведения и проверку сессий работы пользователей на аномальность.

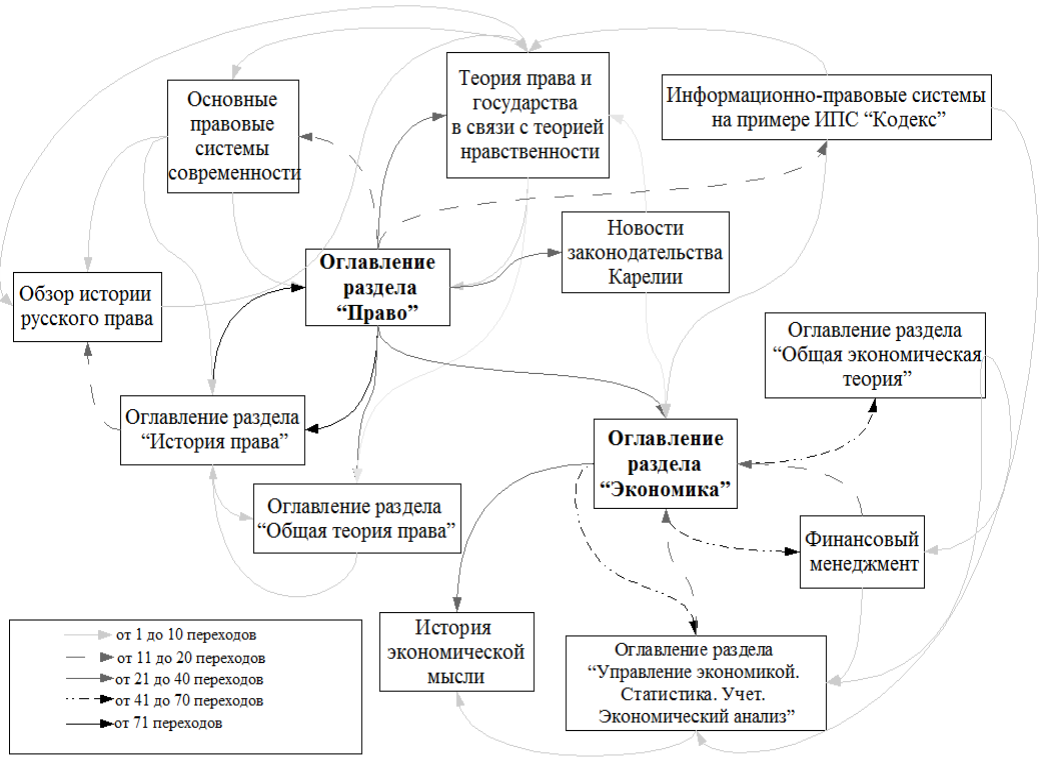



Рис. 2: Пример (фрагмент) профиля «нормального» поведения

### 3.1 Исходные данные

Исходными данными для проведения экспериментов послужили лог-файлы доступа к Электронной библиотеке Республики Карелия [2] за период с июня 2007 г. по февраль 2009 г. включительно. Всего в ЭБ содержится порядка 1000 документов, из них за рассматриваемый период были зафиксированы обращения к примерно 700 документам.

Лог-файл обращений к цифровым документам записан в формате Common Log Format (CLF, [3]). Каждый запрос к серверу записан в отдельной строке, состоящей из полей, разделенных пробелами. При проведении экспериментов использовалась следующая информация, зафиксированная в лог-файле:

I— IP-адрес компьютера пользователя;

D — отметка времени (в CLF-формате);

R — строка запроса (содержит идентификатор запрашиваемого документа);

C — статус ответа сервера.

### 3.2 Преобразование исходных данных

Для построения шаблона «нормального» поведения была проведена предварительная обработка исходных данных. Несмотря на то, что в системе предусмотрена идентификация/аутентификация по паре имя пользователя/пароль, в лог-файле фиксируется лишь IP-адрес пользователя. Всего зафиксированы обращения с более 10000 различных IP-адресов. К сожалению, отсутствие в лог-файле записей об имени пользователя накладывает ограничения на возможности построения профиля «нормального» поведения, так как IP-адрес может являться адресом proxy-сервера, через который к документам ЭБ обращается одновременно несколько пользователей с различными интересами. Для того, чтобы такие обращения не влияли на итоговый результат, были отброшены записи, IP-адрес которых встречался в лог-файле более, чем в 4% случаев. Кроме того, были отброшены неинформативные сессии, содержащие менее 10 запросов к ЭБ.

Сессией работы пользователя считалась последовательность всех обращений к документам с конкретного IP-адреса. Всего в лог-файле содержится 4718 сессий работы пользователей.

### 3.3 Профиль «нормального» поведения

Согласно модели, для построения «нормального» профиля необходимы два набора данных: тренировочный (заведомо нормальные данные) и тестовый (для подбора оптимальных параметров). МЦ, являющаяся «нормальным» профилем поведения, была построена на основе обращений к ЭБ, зафиксированных в период с июня 2007 г. по май 2008 г. включительно. Остальная часть лог-файла служила в качестве тестового набора данных.

На рис. 2 представлен пример (фрагмент) «нормального» профиля, показывающий семантические связи между документами, выявленные на основе поведения пользователей ЭБ.

Естественно, что наиболее сильно оказываются связаны отдельные электронные документы и оглавления разделов. Однако наряду с этим связанными в профиле являются, например, такие

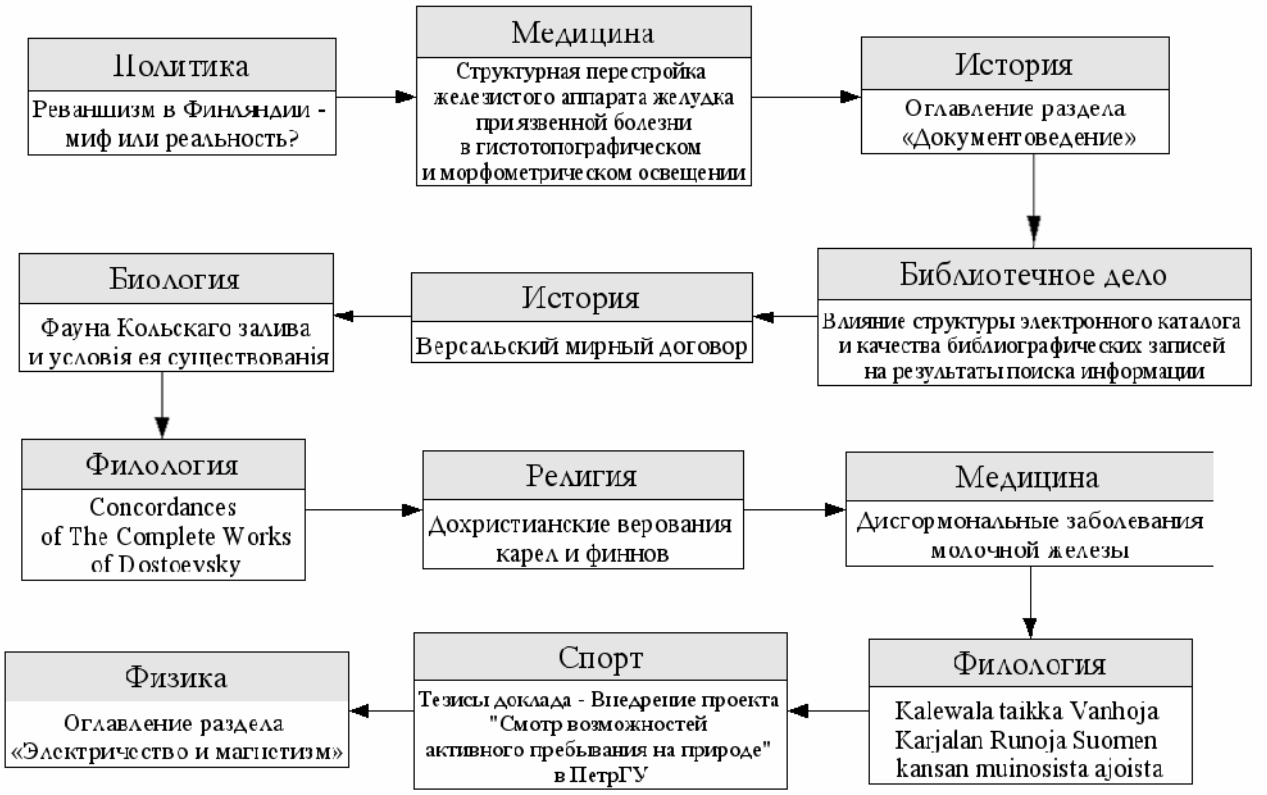



Рис. 3: Пример аномальной сессии работы пользователя

документы как «Основные правовые системы современности» и «Обзор истории русского права». Из названий этих документов понятна их семантическая близость, что подтверждает исходный тезис о возможности выявления семантических связей между документами на основе анализа поведения пользователей ЭБ.

### 3.4 Классификатор и аномальное поведение

Согласно модели аномального обнаружения вторжений [1], классификатор предназначен для определения значимых отличий проверяемой сессии работы пользователя от «нормального» поведения, представленного МЦ.

Одна из наиболее характерных аномальных сессий показана на рис. 3. В заголовке каждого запрошенного пользователем документа, указан раздел, в котором этот документ располагается в ЭБ.

Не вызывает сомнения, что такое разнообразие в выборе документов и разделов не является стандартным поведением пользователя. Выявление подобных аномальных сессий работы и является целью рассматриваемого в данной статье подхода.

## 4 Заключение

В работе представлены первые результаты ряда экспериментов, проведенных для проверки применимости и определения характеристик аномального подхода к защите ЭБ от полного несанкционированного копирования. Для проведения экспериментов была разработана программная система, выполняющая предварительную обработку файла исходных данных, построение профиля «нормального» поведения и проверку сессий работы пользователей на аномальность.

Несмотря на некоторые ограничения, (связанные, в частности, с отсутствием в лог-файле информации о пользователях), можно сделать вывод о применимости аномального подхода в обнаружении вторжений к защите от полного несанкционированного копирования документов ЭБ:

- на основе анализа поведения пользователей возможно автоматически выявлять семантические связи между электронными документами;
- возможно автоматическое выявление последовательностей обращений, противоречащих семантическим связям между документами.

При этом, однако, остается открытым вопрос, связанный с объемом данных по обращениям к документам, достаточным для построения полезных шаблонов нормального поведения. В работах, связанных с обнаружением вторжений на основе аномального подхода, как правило, указывается, что таких данных должно быть «достаточно много», однако какие-либо убедительные оценки (аналитические или эмпирические) отсутствуют.

В дальнейшем планируется сосредоточиться на подборе оптимальных параметров и определении следующих характеристик (согласно модели, представленной в [1]) подхода:

- размер окна при построении шаблона «нормального» поведения;
- количество ошибок классификации и среднее время до первого сообщения об аномальности сессии работы.

Итоговой целью работы является разработка

системы защиты от несанкционированного полного копирования документов, основанной на подходе, представленном в данной работе и статье [1], которая сможет дополнить имеющиеся в ЭБ средства защиты от копирования (например, ограничение числа документов, к которым может обратиться пользователь в единицу времени, и заключение договоров, гарантирующих права владельцев ЭБ). При этом, обнаружение аномального поведения в действиях пользователя может являться основанием для временного блокирования доступа пользователя к ресурсам ЭБ и проведения экспертизы.

## Литература


[1] Ивашко Е. Е. Построение системы защиты электронных библиотек от несанкционированного копирования документов. //Труды Девятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Переславль, Россия, 15-18 октября 2008 г. - Переславль-Залесский: изд-во «Университет города Переславля», 2007. С. 300-306.

[2] Электронная библиотека Республики Карелия. www.elibrary.karelia.ru.

[3] Описание формата Common Log Format. http://httpd.apache.org/docs/1.3/logs.html#common.


## Some results of developing the unauthorized documents-copying protection system for digital libraries


E. Ivashko, N. Nikitina



In this article we consider results of the experiments made to check the workability of statistical anomaly detection algorithm to preserve digital libraries from unauthorized large-scale copying of documents. This work aims to check the theoretical model, introduced in RCDL-2007.



♣ Работа поддержана грантом РФФИ №08-07-00085а «Исследование технологических проблем создания и использования электронных коллекций научных информационных ресурсов»

# ЦИФРОВЫЕ АРХИВЫ

# DIGITAL ARCHIVES

# Организация открытого архива научных публикаций сотрудников ОИЯИ

© В.Ф. Борисовский
vborisov@jinr.ru

В.В. Кореньков
korenkov@cv.jinr.ru

С.В. Куняев
kouniaev@jinr.ru

Ж.Ж. Мусульманбеков
genis@jinr.ru

Э.Г. Никонов
e.nikonov@jinr.ru

И.А. Филозова
Irina.Filozova@jinr.ru

Объединенный институт ядерных исследований

**Аннотация**

Статья посвящена вопросам создания в рамках Open Access Initiative репозитария публикаций научных работ в библиотечной среде Объединенного института ядерных исследований Jinr Document Server (JDS). Рассмотрены требования к выбору программного обеспечения.

## 1 Введение

Традиционные каналы распространения научных результатов посредством публикации в научных журналах претерпевают глубокие изменения в связи с изобретением Интернета и широким доступом к электронным ресурсам. Эти изменения связаны с переходом от парадигмы традиционной публикации к созданию открытых архивов (репозитариев) научной продукции. Эта парадигма, выдвинутая в Декларациях Будапештской и Берлинской Инициатив, получила название "Инициативы Открытого Доступа" (OAI – Open Access Initiative). Под «открытым доступом» подразумевается доступность для любого читателя публикаций в Интернете, которые можно читать, загружать, копировать, распространять, распечатывать или использовать для других законных целей при отсутствии финансовых, правовых и технических преград. Единственным ограничением на воспроизводство и распространение публикаций и условием копирайта в этой области должно быть право автора контролировать целостность своей работы и обязательные ссылки на его имя при использовании публикации и ее цитировании. В настоящее время растущее число академических институтов во всем мире создают собственные репозитарии, накапливая, организуя их в форме открытого доступа для мирового сообщества. Эта тенденция является следствием растущей необходимости перехода к открытому доступу к научной литературе, определяемому принципами OAI: электронный доступ, бесплатный для любого пользователя, снижение ограничений по лицензионному доступу и авторским правам. В перспективе предполагается, что этот переход будет развиваться по двум направлениям: репозитарии с Открытым Доступом (ОД), создаваемые библиотеками университетов и научно-исследовательских институтов и бесплатные, реферируемые журналы с ОД. С другой стороны, издательства, предоставляющие доступ к своим ресурсам за плату, обеспокоены этой тенденцией, поскольку они будут вынуждены перестраивать свою «бизнес-модель» доступа на более демократичную модель ОД. Таким образом, существует определенные проблемы между необходимостью перехода к ОД и монетарной политикой издательств, которые могут быть решены новой моделью «открытых публикаций». Эта модель будет объединять деятельность издательств, библиотек академических институтов и университетов и различных фондов для продвижения и распространения OAI.

## 2 Современное состояние архивов с открытым доступом

Все большее число академических институтов и университетов создают свои собственные репозитарии научной литературы с открытым доступом**.** Созданные репозитарии регистрируются в одном из международных реестров репозитариев открытого доступа, например, Registry of Open Access Repositories — ROAR, что делает их доступными для всего мирового сообщества [8].

В Объединенном институте ядерных исследований (ОИЯИ) одним из основных направлений научных исследований является физика высоких энергий. Наиболее активно используемым в физике высоких энергий является архив электронных препринтов научных статей arXiv.org, где авторы еще до принятия статьи в рецензируемый журнал депонируют свои работы. arXiv.org предоставляет ограниченный доступ к публикациям, а именно, только к препринтам и не является в полном смысле репозитарием открытого доступа. В настоящее время тематика научных направлений, охватываемая в arXiv, значительно расширена и включает математику, компьютерные науки, химию и биологию. Хотя arXiv.org является репозитарием коллективного использования для всего физического сообще-

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

ства и, поэтому служит оперативным и эффективным источником научной информации для любого исследователя, он имеет и ряд недостатков. У него отсутствует пользовательский интерфейс, необходимые библиотечные службы, возможность открытой дискуссии и реферирования загруженных статей. Первым репозитарием ОД, созданным в физике высоких энергий и удовлетворяющий этим требованиям, является архив публикаций на сервере CERN Document Server (CDS) Европейского центра ядерных исследований (ЦЕРН). В качестве программного обеспечения архива используется разработанный в ЦЕРН пакет CDS Invenio, обеспечивающий протокол обмена данных, совместимый с принятым в OAI [2]. ОИЯИ входит в сообщество физических центров (HEP community), многие из которых используют пакет CDS Invenio для организации и поддержки собственных репозитариев ОД. В России и странах ближнего зарубежья известно несколько проектов, направленных на решение задачи поддержки научных публикаций. Среди них можно назвать систему Соционет [5], специализированный портал информационной поддержки научной деятельности в Тульской области [2], открытую научную электронную библиотеку периодических изданий Национальной академии наук Украины НАНУ [9] и многие другие. Названные проекты и созданные системы преследуют общую цель — предоставление неограниченного открытого доступа к научным публикациям и, что очень важно, долгосрочное хранение. Методы в самых общих чертах совпадают, проекты ориентированы на создание портала публикаций, который предоставляет свободный доступ к контенту. Инструментарий для создания архивов в некоторых проектах (например, Соционет) — это самостоятельно разработанное программное обеспечение, что требует немалых затрат. Многие проекты используют хорошо зарекомендовавшие себя свободно распространяемые бесплатные программные продукты. В большинстве случаев это пакеты DSpace и EPrints — системы, которые не требуют больших затрат на создание архива и его поддержку.

Таким образом, в ОИЯИ созрела необходимость создания собственного репозитария и включения его в международную систему архивов открытого доступа в рамках OAI. При этом выбор программного обеспечения для создания и поддержки архива диктуется следующими соображениями: во-перых, архив должен быть интегрирован в международную сеть репозитариев открытого доступа в рамках OAI и, во-вторых, вследствие тесных научных связей между ОИЯИ и ЦЕРН придерживаться существующей тенденции последовательной унификации информационных ресурсов обоих институтов. В следующем разделе приводятся цели и требования к создаваемому репозитарию. В разделе 3 проводится сравнительный анализ уже упомянутых пакетов EPrints, DSpace и CDS Invenio, которые рассматриваются в качестве кандидатов для программного обеспечения создаваемого репозитария ОИЯИ.

## 3 Цели и требования к создаваемому репозитарию ОД

Создание репозитария ОД публикаций сотрудников ОИЯИ преследует следующие цели:

- Сделать доступными для международной научной общественности научные результаты и разработки сотрудников ОИЯИ в существенно короткие сроки.
- Повышение эффективности использования информационных ресурсов Издательский отдела и Научно-технической библиотеки ОИЯИ.
- Повышение уровня информационного обеспечения сотрудников ОИЯИ, в том числе благодаря предоставлению доступа к репозитариям ОД других научных центров.
- Оценка эффективности результатов научной деятельности сотрудников института.

Для реализации этих целей создаваемый репозитарий должен быть интегрирован с международными каталогами, реестрами и репозитариями. Кроме того, необходима интеграция с внутренними информационными ресурсами ОИЯИ, как фактографического так библиографического содержания.

Система сопровождения репозитария должна удовлетворять следующим требованиям:

- Возможность пополнять архив полнотекстовыми документами из других в том числе международных архивов.
- Возможность классификации пополняемых материалов по тематике.
- Централизованное администрирование репозитария и предоставление различных услуг авторам статей и пользователям.
- Предоставлять авторам и библиотечным работникам возможность загружать и передавать статьи в систему используя Интернет – браузер.
- Предоставлять заинтересованным пользователям средства открытого рецензирования и дискуссии еще до принятия статьи в рецензируемый журнал.
- Система должна иметь развитые средства поиска, уведомлений о новых поступлениях и последних изменениях (alerts, RSS feeds).

Таким образом, для создания репозитария в информационной среде ОИЯИ и встраивания его в международную сеть репозитариев ОД необходимо выбрать программное обеспечение, возможности которого адекватны предъявляемым требованиям.

## 4 Анализ программного обеспечения, применяемого для организации электронных архивов

На данный момент в мире существует немало систем для создания электронных архивов как платных, так и бесплатных: EPrints, DSpace, Bepress, OPUS, CDS Invenio и др.

По данным реестра репозитариев открытого доступа (ROAR) лидирующие позиции в этой области

занимают два бесплатных программных продукта EPrints [6] и DSpace [7]. В мире функционирует около 400 электронных репозитариев, построенных с помощью ПО DSpace и более 300 — EPrints.

При выборе программного обеспечения для организации архивов открытого доступа учитываются следующие характеристики:

1. Функциональность;
2. Модель данных;
3. Форматы файлов;
4. Метаданные;
5. Поддержка протокола OAI-PMH;
6. Экспорт/Импорт;
7. Разграничение прав доступа;
8. Депонирование;
9. Внешнее использование;
10. Установка и необходимое программное обеспечение;
11. Поддержка многоязычности;

Далее анализируются и сравниваются функциональные возможности пакетов EPrints, DSpace и CDS Invenio как кандидатов для программного обеспечения создаваемого репозитария.

### 4.1 EPrints и DSpace

Основные возможности и особенности популярных свободно распространяемых систем EPrints и DSpace, а также история их создания и развития, архитектурные отличия и требования к установке были подробно рассмотрены в работах [3].

Системы EPrints и DSpace являются представителями одного класса и имеют много общего, эквивалентны по функциональности.

Общие функциональные возможности систем:

✓ хранение и индексация метаданных разнообразных форматах в БД;

✓ хранение информации о пользователях системы;

✓ авторизация пользователя;

✓ поиск и просмотра (навигация) коллекций;

✓ поддержка протокола сбора метаданных OAI-PMH;

✓ автоматическая рассылка уведомлений по электронной почте через службу подписки;

✓ обработка данные произвольных форматов;

✓ доступ к перечисленным функциональным возможностям посредством веб-интерфейса.

Имеются некоторые отличия в архитектурном, техническом аспекте, а также в использовании.

✓ DSpace использует строгую иерархическую систему организации данных, которая позволяет отразить структуру организации. Модель данных EPrints заключается в том, что все записи эквивалентны и являются одноуровневыми.

✓ DSpace сохраняет метаданные в формате квалифицированного Дублинского Ядра, а в EPrints каждому типу поставлено в соответствие внутренний набор метаданных. Отличительной чертой EPrints является возможность динамически генерировать метаданные в различных форматах из внутреннего представления.

✓ Также существуют отличия, касающиеся использования систем, а именно роли пользователей, процесса внесения данных, просмотра и поиска. С точки зрения ролей и прав доступа EPrints лучше подходит для однородных репозитариев, где не имеют значения права пользователя, выходящие за рамки обычного. Такая система не требует настройки прав пользователей. DSpace обладает более гибкой системой прав доступа, позволяющей ограничивать доступ к различным частям архива. Администрирование и управление процессом внесения осуществляется посредством web-интерфейса. В EPrints конфигурация прав доступа для каждой категории пользователей осуществляется посредством редактирования соответствующих файлов.

✓ EPrints обладает чуть более дружественным интерфейсом для депонирования. Однако, DSpace более производительна, поэтому процесс депонирования в целом проходит быстрее.

### 4.2 CDS Invenio

CDS Invenio (ЦЕРН / CERN, Швейцария) — интегрированная электронная библиотечная система. Представляет собой набор приложений для построения и управления автономным сервером электронной библиотеки [1]. Программное обеспечение бесплатное, распространяемое под лицензией GNU General Public License. Технология, предлагаемая данным продуктом, покрывает все аспекты поддержки электронной библиотеки, совместима по протоколу OAI-PMH, использует формат MARC21 как основной библиографический стандарт. Система CDS Invenio является комплексным решением управления репозитариями документов средних и больших объемов.

Посредством CDS Invenio создан и поддерживается архив публикаций сервера документов CERN (CERN Document Server). В CERN CDS Invenio управляет более чем 500 коллекциями данных, состоящих из более чем 800 000 библиографических записей и 350 000 полнотекстовых документов, покрывая препринты, статьи, книги, журналы, фотографии, видеоматериалы и др. Помимо CERN, CDS Invenio в настоящее время инсталлирована и используется в 14-ти научных и образовательных учреждениях мира.

#### 4.2.1 История создания и развития CDS Invenio

Система разработана в Европейском центре физики высоких энергий в Женеве и прошла несколько этапов развития.

1993 – Web-сервер препринтов CERN. Институциональный репозитарий.

1996 – Библиотечный Web-сервер CERN (weblib): добавлены книги и периодические издания.

2002 – Сервер документов CERN (CDSware) / CERN Document Server Software: поддержка мультимедийных данных и OAI Protocol.

2006 – развитие CDSware, изменение названия на CDS Invenio.

### 4.2.2 Описание CDS Invenio

***Архитектура CDS Invenio***

Архитектура CDS Invenio построена на технологиях Open Source (Python, RDBMS MySQL, PHP, Apache), использовании открытых стандартов (MARCXML, MARC21,OAI-PMH, OpenURL и др.) и принципах модульности. Движение документов от их загрузки в репозитарий до выполнения запросов в соответствие с архитектурой Invenio схематично представлено Рис.1. Пополнение данных осуществляется из трех источников: внесение статьи непосредственно автором (с помощью электронной почты или web интерфейса), сборка данных из OAI и non–OAI репозитариев. Собранные метаданные конвертируются во внутреннее представление метаданных (MARCXML) и поступают на библиографический сервер, а полнотекстовые документы конвертируются при необходимости в формат PDF и загружаются на Document Server. До загрузки в библиографический сервер метаданные могут быть подвергнуты процедуре качественного анализа библиотечным работником. Дополнительно метаданные обогащаются ссылками, извлеченными из соответствующих полных текстов. В результате библиографический сервер сможет генерировать индексы и форматы библиографических данных, необходимых для быстрого поиска. И, окончательно, информация поступает к пользователям и провайдерам сервисов OAI в ответ на OAI-PMH, e-mail и web запросы. Рассмотрим далее более детально структуру и функциональные возможности CDS Invenio.

***Функциональность***

CDS Invenio имеет всю необходимую функциональность для обеспечения поддержки электронных публикаций

- ✓ удобный навигационный механизм в коллекциях, предусматривающий настройки для каждой коллекции;
- ✓ мощная поисковая машина (специально разработанные индексы обеспечивают Google-подобные скорости для архивов до 1 500 000 записей); поисковый интерфейс доступен на 20-ти языках.
- ✓ одновременный поиск по метаданным, полным текстам и цитатам; результаты группируются по коллекциям;
- ✓ настраиваемые пользовательские интерфейсы;
- ✓ развитый информационный сервис, включая пользовательские корзины документов, автоматическое уведомление пользователя по электронной почте;

Типы документов, поддерживаемых CD Invenio:

- ✓ опубликованные статьи;
- ✓ препринты;
- ✓ книги;
- ✓ тезисы;
- ✓ труды конференций;
- ✓ презентации и доклады;
- ✓ отчеты;
- ✓ фотографии;
- ✓ видео-материалы;
- ✓ музейные экспонаты;

***Модель данных***

Дерево коллекции строится на основе классификации наполнения и имеет иерархическую структуру: статьи и препринты, книги и труды конференций, периодические издания и отчеты, презентации и доклады, мультимедиа, архивы. Каждый из этих разделов состоит из подразделов. Так, раздел «Статьи и препринты» включает опубликованные статьи, препринты, тезисы, протоколы комитетов, служебные инструкции.

***Форматы файлов***

Все библиографические данные представлены в формате MARC21. CDS Invenio поддерживает широкий набор форматов для хранения объектов: PDF, PS, HTML, XML, JPEG, GIFF, TIFF, PNG, MPEG, AVI, PPT, RTF, DOC и др.

***Метаданные***

Используется гибкий стандартный формат метаданных MARC XML. Структура метаданных в любой момент может быть расширена и адаптирована. Текущая MARCXML схема в CERN включает более 150 полей метаданных.

Внесение метаданных выполняется автоматизированными и полуавтоматическими процедурами модуля BibHarvest. Документ может быть добавлен непосредственно авторами по Сети или электронной почте через модули WebSubmit и ElmSubmit. В обоих случаях метаданные преобразуются в родное представление метаданных CDSware и загружаются на сервер. Чтобы установить подлинность метаданных, каталогизатор может выполнить качественную оценку через модуль BibCheck.

***Поддержка протокола OAI-PMH***

CDS Invenio поддерживает протокол сбора метаданных Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

***Экпорт/Импорт***

Модуль BibConvert реализует преобразования между различным форматами — последовательными (например, ISO2709) и слабоструктурированными (например, XML), форматами метаданных (MARC21, DublinCore, RFC1807, и т.д.) — и учитывает особенности форматирования текста. Преобразование метаданных через BibConvert обеспечивает высокую степень автоматизации: отчеты метаданных из различных источников могут быть легко импортированы в MARCXML и немедленно введены в систему стандартных файлов конфигурации. Есть возможность экспорта в EndNote.

***Разграничение прав доступа***

Выполнять поиск и/или просмотр можно анонимно, для загрузки публикации нужно пройти процедуру аутентификации. Кроме того, можно ограничивать доступ к коллекциям. Для этого используется механизм ролей, где пользователи принадлежат нескольким группам согласно их роли в системе. Это может

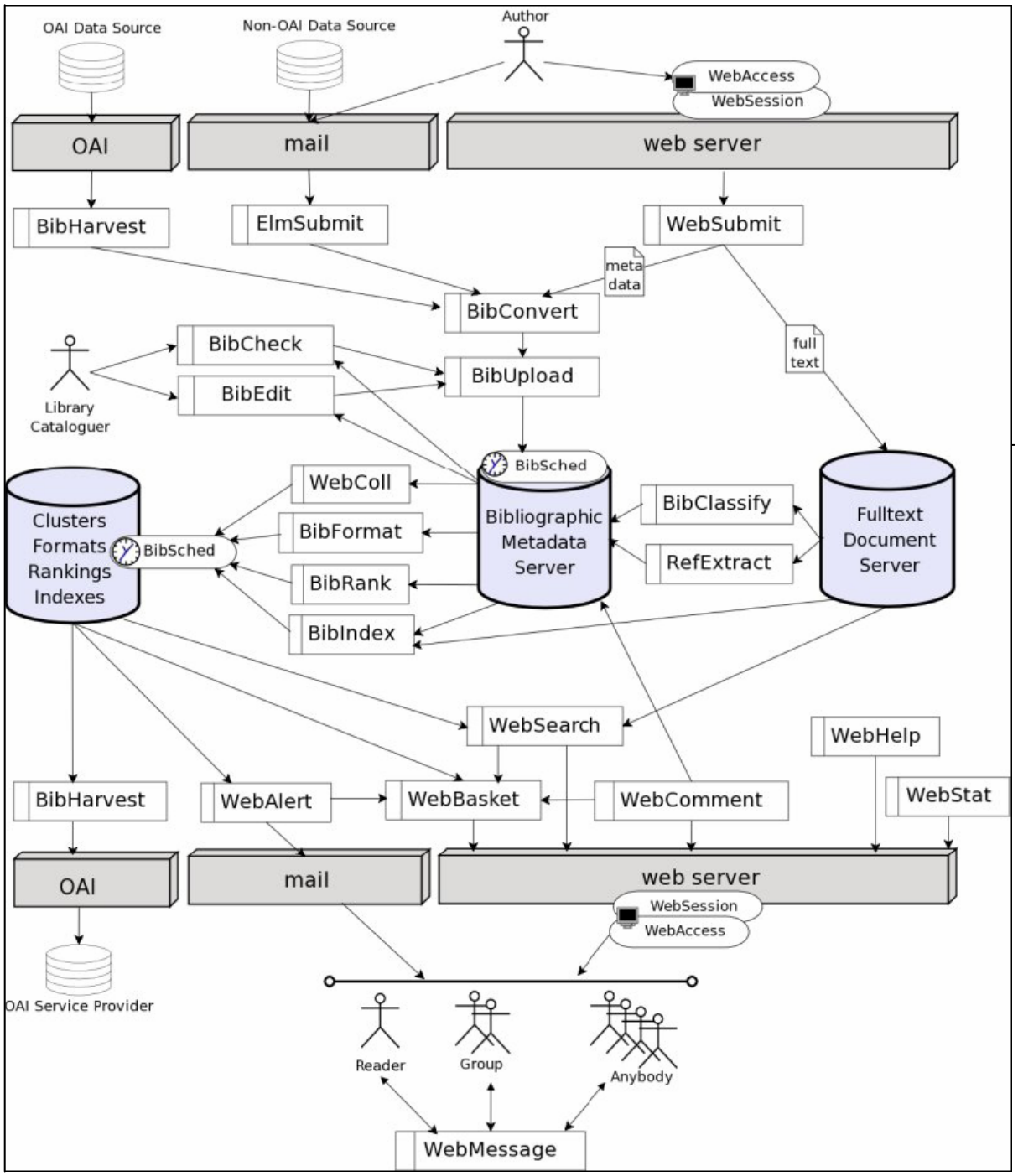



**Рис. 1.** Архитектура Invenio.

.

быть редактор, каталогизатор, менеджер данных и проч. Каждой пользовательской группе можно предоставить права на выполнение определенных действий. Администрирование осуществляется через набор модулей, обеспечивающих различные задачи администрирования — конфигурирование портала, настройка поисковой машины, сбор метаданных, разграничение прав доступа и т.д.

***Депонирование***

Документ может быть добавлен непосредственно авторами по Сети или электронной почте через модули WebSubmit и ElmSubmit.

***Внешнее использование***

Имеет развитый настраиваемый интерфейс. Обладает следующими возможностями:

✓ Простой поиск;

✓ Расширенный многокритериальный поиск с сортировкой по этим критериям;

✓ Возможность ввести указания для поиска;

✓ Навигация по репозитарию возможна по типу или тематике документа;

✓ Автоматическая рассылка различных уведомлений пользователя, например, при поступлении нового документа, соответствующего заданным критериям. Реализуется модулем WebAlert.

✓ Читательская корзина или виртуальная книжная полка. Модуль WebBasket позволяет конечному пользователю системы хранить отобранные документы в личной корзине. Одному пользователю могут принадлежать несколько корзин. Корзина может быть личной или коллективной в пределах группы.
✓ Оценка документов читателями. Модуль WebComment совместно с другим инструментарием — модули WebBasket, WebGroup, WebMessage, WebComment — позволяет учитывать социальные особенности сети.
✓ Информационно-справочная система реализуется модулем WebHelp с учетом прав доступа.
✓ Коммуникация пользователей. Модуль WebMessage обеспечивает коммуникацию между (включая анонимных) конечными пользователями через доски объявлений.
✓ Сбор статистики реализуется конфигурируемым модулем WebStat.

***Установка и необходимое программное обеспечение***

Таблица 1. Требования к установке CDS Invenio

| | |
|---|---|
| *Операционная система* | Unix-подобная |
| *Web-сервер* | Apache 2<br>mod_python<br>PHP |
| *Сервер баз данных* | MySQL<br>MySQLdb |
| *Библиотеки языков программирования* | Python |

***Поддержка многоязычности***

CDS Invenio поддерживает многоязычный интерфейс, доступный на 20 языках.

### 4.3 Выбор программного обеспечения

Проведенный анализ показал, что исследуемые системы практически равноценны по предоставляемым возможностям и полностью удовлетворяют выдвинутым критериям.

Учитывая международный статус ОИЯИ, а также тесное сотрудничество ОИЯИ и CERN, был сделан выбор в пользу CDS Invenio. ОИЯИ и CERN выполняют много совместных программ в области теоретической и экспериментальной физики, особенно в физике высоких энергий, и поэтому интеграция, более того унификация в сфере информационного обеспечения научных исследований является требованием сегодняшнего дня. В ОИЯИ уже эксплуатируется система CDS Agenda, позволяющая управлять организацией семинаров, рабочих встреч, конференций и пр. В ближайшее время планируется переход на web-приложение CDS Indico, которое является развитием CDS Agenda и предоставляет широкий спектр возможностей для информационного обеспечения различных событий в научном сообществе — встреч, семинаров, конференций в электронном и традиционном форматах.

## 5 JINR Document Server

В целях унификации информационных ресурсов с CERN в ОИЯИ по подобию CERN Document Server организуется сервер научных документов JINR Document Server (JDS), на котором планируется разместить архив–репозитарий публикаций сотрудников ОИЯИ и архив материалов конференций, проводимых в ОИЯИ. В качестве программного обеспечения для создания и сопровождения архивов будут использованы соответственно CDS Invenio и CDS Indico. В ОИЯИ имеется богатый опыт создания современных web-приложений. Примером является разработанная в ОИЯИ система ПИН (Персональная ИНформация), предназначенная для распределенного сбора и анализа информации о результатах научной деятельности сотрудников [9]. Создана библиографическая база публикаций сотрудников ОИЯИ. Разработаны web-приложения автоматизации документооборота ОИЯИ (в эксплуатации с 2006 г), управленческого учета ADB2 (в эксплуатации с 2004 г) и бухгалтерского учета BHT (в эксплуатации с 1997 г.).

### 5.1. Создание архива JINR Document Server

В ходе выполнения работ по созданию и эксплуатации в начальный период архива–репозитария ОД на сервере JDS планируется:
✓ на основе CDS Invenio обеспечить базовый вариант открытого архива публикаций сотрудников ОИЯИ;
✓ разработать методику наполнения и функционирования архива — регламент подачи публикаций, процедуру депонирования публикаций в режимах “самоархивирования” и “депонирования по доверенности”, соблюдение авторских прав;
✓ обеспечить программно-методическое сопровождение, включая разработку инструкций, руководств для всех категорий пользователей, консультирование;
✓ обеспечить сбор статистики по следующим показателям:

- количество посещений в день/месяц;
- количество уникальных посетителей;
- количество посещенных страниц в день/месяц;
- наиболее популярные страницы;
- количество посещений страницы, раздела, публикации;

✓ обеспечить дальнейшую эксплуатацию архива — техническое оснащение, конфигурирование и настройку, соблюдение политики наполнения;
✓ разработать методику интеграции архива с приложением ПИН, что позволит использовать данные архива при расчете индекса ПРНД в системе ПИН [4].

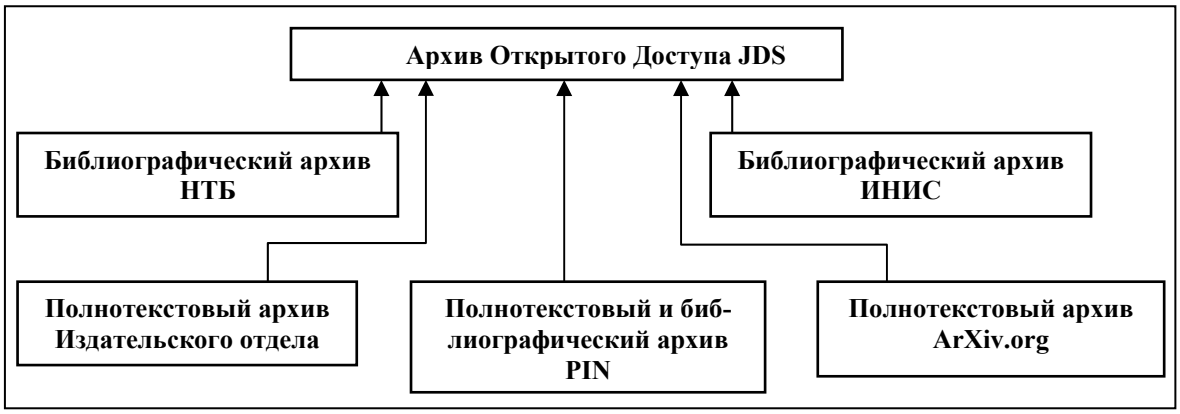



**Рис. 2.** Источники библио- и фактографической информации для наполнения архива публикаций в библиотечной среде ОИЯИ

Для этого необходима разработка модуля интеграции в виде web-приложения. Модуль должен обеспечивать:

- ✓ единую регистрацию пользователя;
- ✓ интерфейс для депонирования публикаций в архиве, обеспечивающий возможность автоматического импорта метаданных о публикации из архива в ПИН с привязкой к авторам-сотрудникам ОИЯИ. Эти данные впоследствии могут быть использованы для автоматического расчета ПРНД сотрудников.

Особую проблему представляет задача наполнения архива ранее изданными публикациями. Прежде всего необходим библиографический список авторов–сотрудников ОИЯИ, который может быть получен из библиографической базы Научно-технической библиотеки, НТБ ОИЯИ (Рис.2). Дополнительным источником библиографической информации может служить база данных ИНИС, поддерживаемая МАГАТЭ. Руководствуясь полученной библиографической информацией, можно извлекать полные тексты публикаций из баз данных Издательского отдела ОИЯИ, архивной базы ArXiv, архивов различных издательств. Кроме того, предполагается, что сами авторы будут заинтересованы депонировать в институтский репозитарий не только текущие, но и ранее изданные работы.

### 5.2. Процедура депонирования публикаций

Внесение публикаций могут выполнять только зарегистрированные пользователи системы, наделенные соответствующими правами. Документ может быть добавлен или отредактирован автором или доверенным лицом интерактивно или по электронной почте. Процедура интерактивного депонирования более предпочтительна. Точная последовательность шагов данной процедуры зависит от типа публикации.

Интерактивное депонирование:

1. Выбрать тип публикации в нужном разделе архива.
2. Выбрать категорию из списка доступных для данного типа документа. Если тип документа использует такие категории, появится список радио-кнопок с названием каждой категории на странице.
3. Выбрать вид действия. В общем случае определены следующие возможные операции: внесение новой записи, редактирование существующей записи, управление прикрепленными к документу файлами.

**В режиме *внесения:***

4. Ввести необходимую информацию в форму: название документа, авторы, дата создания, аннотация и т.п.
5. Загрузить файл полного текста.
6. Подтвердить внесение документа.
7. Дождаться сообщения о том, что документ успешно внесен в репозитарий.

**В режиме *изменения:***

Редактирование документа — двухступенчатый процесс.

*Первый этап:*

4. Ввести идентификационный номер документа, который необходимо отредактировать.
5. Отметить поля, подлежащие модификации.
6. Нажать кнопку «Продолжить».

*Второй этап:*

7. Заполнить сгенерированную системой форму.
8. Подтвердить запрос на изменение документа.
9. Дождаться сообщения о том, что изменения внесены.

**В режиме *управления прикрепленными файлами:***

В данном случае доступны операции — добавление / изменение / удаление файла(ов), прикрепленных к существующей записи. Данное действие доступно не всегда

4. Выбрать нужное действие, нажав одну из кнопок: «Изменить файлы», «Загрузить файлы», «Удалить файлы»

5. Заполнить сгенерированную системой форму.
6. Подтвердить запрос.
7. Дождаться сообщения о том, что изменения приняты.

### 5.3 Перспективы расширения структуры JDS

По мере наполнения архива структура сервера JDS будет расширяться. Помимо публикаций сотрудников ОИЯИ на сервере JDS будут размещаться следующие материалы:

- Книги
- Препринты
- Статьи
- Периодические издания
  - ✓ ЭЧАЯ
  - ✓ Письма в ЭЧАЯ
  - ✓ Новости ОИЯИ
- Диссертации и авторефераты
- Труды конференций
- Отчеты
  Годовые отчеты ОИЯИ
- Публикации об ОИЯИ
- Презентации и доклады
  - ✓ Конференции
  - ✓ Курсы лекций
  - ✓ Руководства
- Мультимедиа
  - ✓ Фото
  - ✓ Видео
  - ✓ Аудио
  - ✓ Постеры

## 6 Заключение

Создание архива–репозитария ОД публикаций сотрудников ОИЯИ позволит интегрироваться в мировую систему репозитариев в рамках OAI. Это дает возможность заинтересованным ученым независимо от их местонахождения иметь быстрый доступ к научным результатам сотрудников ОИЯИ. Выбор соответствующего инструментария для создания архива облегчает дальнейшую интеграцию информационных сред ОИЯИ и CERN. Создаваемая система позволит существенно повысить эффективность использования информационных ресурсов ОИЯИ как для внутренних, так и внешних пользователей.

## Литература


[1] CDSware Overview. http://cdsware.cern.ch/invenio/index.html.

[2] В.В. Глаголев, А.Н. Мерцалов. Система информационной поддержки научных конференций на специализированном портале http://www.ict.edu.ru/vconf/files/9246.pdf

[3] К.А. Кудим, Г.Ю. Проскудина, В.А. Резниченко. Сравнение систем электронных библиотек EPrints 3.0 и DSpace 1.4.1.// Труды Девятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2007. (Переславль-Залесский, Россия, 15 - 18 октября 2007 г.). http://www.rcdl.ru/papers/2007/paper_66_v2.pdf.

[4] Научные и научно-организационные электронные ресурсы Объединенного института ядерных исследований / В.Ф. Борисовский, В.В. Кореньков, С.В. Куняев, Н.А. Ленская, Ж.Ж. Мусульманбеков, Э.Г. Никонов, И.А. Филозова. // Труды Десятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL' 2008. — ISSN 5-9530-0193-2 — ОИЯИ, 2008 — с. 277-283.

[5] Онлайновая научная инфраструктура «Соционет» — http://socionet.ru/idea.htm

[6] Официальный сайт EPrints — http://www.eprints.org/

[7] Официальный сайт DSpace — http://www.dspace.org/

[8] Реестр репозитариев открытого доступа / Registry of Open Access Repositories — http://roar.eprints.org/

[9] В.А. Резниченко, Г.Ю. Проскудина, К.А. Кудим. О создании открытой научной электронной библиотеки периодических изданий НАНУ // Труды Десятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL' 2008. — ISSN 5-9530-0193-2 — ОИЯИ, 2008 — с. 347-356.


## On Open Access Archive for publications of JINR staff members


V.F. Borisovsky V.V. Korenkov S.V. Kuniaev
G. Musulmanbekov E.G. Nikonov I.A. Filozova



The paper concerns with the problems of building the OAI-compliant archive (repository) of the JINR staff members publications. The analysis of the software packages for building and management of such repositories has been performed.

# Разработка информационной системы для медицинских учреждений

© Гордеев Д.А.
ИСИ СОРАН
GordeevDima@mail.ru

**Аннотация**

В статье рассматриваются структура и интерфейс медицинской информационной системы. Так же показаны пути дальнейшего развития программы. Данная система предназначена для автоматизации движения потоков информации и для взаимодействия отделений медицинского учреждения.

## 1. Введение

В России, в сфере медицинского обслуживания, практически не применяются информационные технологии. Считанные медицинские учреждения пытаются внедрить у себя информационные системы. В большинстве поликлиник и больниц основным носителем информации является папка с бумагами. Но в нынешних условиях, когда человек больше не привязан к конкретной поликлинике по месту жительства, такой подход не позволяет быстро получить все данные о пациенте, о его болезнях, о непереносимости лекарств, так как вся эта информация разбросана по больницам, которые он когда-то посещал. В экстренных случаях возможность быстрого доступа к этой информации может сыграть решающую роль в борьбе за жизнь больного. Поэтому создание сети медицинских информационных систем является необходимым условием для функционирования всех областей здравоохранения.

Внедрение информационных систем в медицинские учреждения позволит создать сеть с распределенным хранением данных о пациенте, что решит проблему получения информации о больном, упростит взаимодействие больниц и даже подтолкнет медицинские учреждения к сотрудничеству.

Информационная система в медицинском учреждении позволяет снизить затраты времени при обследовании и лечении больного, помогает избежать бумажной волокиты. А настрой пациента, его душевное спокойствие играет не последнюю роль в процессе выздоровления. Удобство в работе врачей, которое предоставляет информационная система, позволяет им сконцентрироваться на пациенте и не “забивать голову” административными требованиями.

## 2. Особенности предметной области

Проведенные исследования показали, что существующие технологии создания корпоративных информационных систем для здравоохранения не совсем подходят. Говоря обобщенно, в здравоохранении пользователь информационной системы имеет дело и с данными, и с документами, которые в себя эти данные включают, и с «рабочими потоками», определяющими бизнес--процессы лечебного учреждения, в том числе регламент работы с документами. А существующие технологии в основном ориентированы либо на обработку данных, либо на работу с документами, либо на поддержку жестких бизнес-процессов учреждения.

Важное отличие здравоохранения от других областей заключается в специфике финансово-экономических взаимоотношений между сторонами, оказывающими медицинские услуги (ЛПУ), оплачивающими эти услуги (страховые организации) и потребляющими медицинские услуги (пациенты). То есть, в отличие от других областей, в здравоохранении в системе взаиморасчетов участвуют не две стороны (поставщик и потребитель услуги), а три стороны. При этом, в каких услугах нуждается пациент, решает поставщик этих услуг, то есть ЛПУ. Отсюда возникают проблемы оценки качества и адекватности объемов оказываемых услуг. Это обусловливает определенные проблемы для каждой из сторон и необходимость разработки медицинских стандартов качества, медико-экономических стандартов и т.д. Сюда еще добавились проблемы, связанные с так называемыми источниками финансирования. Эти же проблемы отражаются в требованиях к медицинским информационным системам. Иными словами, при разработке медицинских информационных систем исключительно сложны поддержка медицинских технологических процессов и оценка медицинской помощи в терминах услуг.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2009, Петрозаводск, Россия, 2009.**

Медицина – очень сложная область человеческой деятельности, которая позволяет использовать самые разнообразные информационные технологии и информационные системы.

В нашей стране развитие медицинской информатики совпало с периодом серьезных преобразований в системе здравоохранения, связанных с изменениями в общественной и политической жизни страны, что конечно создает для разработчиков новые и серьезные проблемы.

## 3. Постановка задачи

Медицинская информационная система предназначена для автоматизации работы медицинских учреждений независимо от принадлежности (государственное или ведомственное) и специализации. При этом она должна быть разработана с учетом возможности ее использования в поликлинике, многопрофильном стационаре с различными клиническими и диагностическими отделениями. Медицинская информационная система должна удовлетворять все потребности медицинского персонала и пациентов, ради которых она создавалась.

Информационная система для медицинских учреждений должна:

1. уменьшить время прохождения пациентом регистратуры, оформления документов при посещении, при осмотре
2. повысить удобство работы врача с пациентом, эффективность работы докторов
3. перевести весь документооборот на электронный вариант
4. автоматизировать взаимодействие между отделениями
5. предоставлять быстрый доступ ко всей необходимой информации по пациенту
6. предоставлять полную и гибкую систему отчетности

Для выполнения этих требований в медицинскую информационную систему должны входить следующие подсистемы:

1. Учет пациентов
2. Учет движения пациентов (амбулаторный и стационарный этапы)
3. Ведение истории болезней
4. Составление расписаний работы врачей и отделений
5. Подготовка отчетов, результатов обследований, другой документации
6. Модули работы отделений и других подразделений
7. Взаимодействие отделений, диагностических подразделений, лабораторий

## 4. Реализация

Информационная система состоит из трех больших модулей: поликлиника, стационар и диагностические подразделения (рис. 1). Данные модули выделены по принципу прохождения лечения пациентом. Больной может приходить на амбулаторное лечение в поликлинику, а при необходимости лечь на стационарное лечение. В итоге мы получаем два связанных, но непересекающихся этапа лечения (так как пациент не может одновременно находиться на амбулаторном и стационарном этапе лечения). Что удобно для моделирования двух подсистем. При переносе системы на поликлиники или же больницы без амбулаторного этапа лечения подобное разделение поможет легко отделить ненужные модули и масштабировать систему.

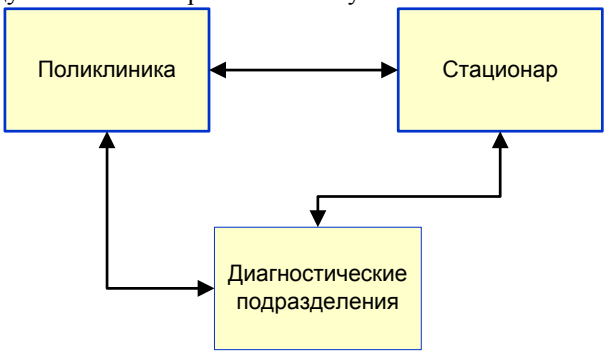



Рис. 1. Основные модули системы

Некоторые диагностические подразделения могут работать без прямого контакта с пациентами (например, клинико-биохимическая лаборатория может работать с анализами, пришедшими из других медицинских учреждений). Данная возможность реализована с помощью отдельных интерфейсной части и набора сущностей, взаимодействующих с остальной системой как внешний блок. Эта заявочная система может размещаться как на внутреннем, так и на внешнем сервере, что позволяет дать доступ к ней максимальному количеству медицинских учреждений. К тому же немногие медицинские центры и больницы имею диагностические блоки в своем составе. Так что выделения этих подразделений в отдельный модуль является естественным.

При рассмотрении взаимодействия модулей можно выделить так же приемное отделение (рис. 2), которое не имеет прямого отношения к лечебному процессу, а осуществляет прием населения и регистрирует все перемещения пациентов в ходе лечебного процесса. Например, на амбулаторном этапе пациента могут направить на госпитализацию. А после госпитализации порекомендовать процедуры для скорейшего восстановления здоровья, которые нужно проходить в поликлинике. Все переходы назначения и действия врачей регистрируются системой.

Конечно, модульность не должна означать разделение информации по пациенту. Разделения на модули подразумевают лишь минимизацию связей между ними, для возможности оперировать ими отдельно.

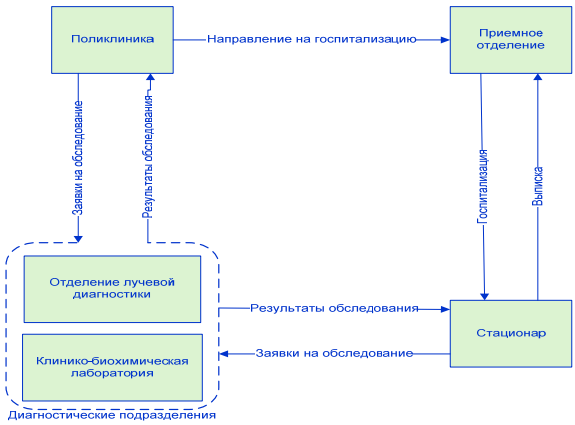



Рис. 2 Взаимодействие модулей

### 4.1. Приемное отделение

Основной сложностью реализации регистратуры является то, что неизвестно какое количество данных будет доступно о пациенте, так как больной может прийти самостоятельно, перевестись из другого лечебного учреждения или же поступить экстренно. А так как приемное отделение является отправной точкой для дальнейшего ведения пациента по информационной системе – необходимо знать, на какие данные мы можем рассчитывать. Для решения этой проблемы информация о пациенте была разведена на две сущности: Пациент и Амбулаторная карта. Первая сущность содержит в себе информацию необязательную для дальнейшей работы с пациентом, такую как место работы, родственники и т.д. Амбулаторная карта же содержит в себе данные медицинского характера, необходимые для дальнейшей работы. Так, например, анамнез жизни является обязательным полем и его заполняет врач либо со слов больного, либо при его осмотре, если пациент прибыл в бессознательном состоянии и не может рассказать о предыдущих заболеваниях.

### 4.2. Амбулаторный этап

Этот этап предназначен для обследования пациента, назначения лечения или принятия решения о госпитализации пациента. Данная схема подразумевает так называемое ведение пациента «до конца», то есть: от первичного приема, на котором определяется диагностика и лечение - через дополнительное обследование - к решению о необходимости госпитализации больного. В ходе амбулаторного лечения при каждом обследовании дополняется карта диагнозов, которая заводится при первичном приеме с целью слежения за состоянием пациента.

Основными сущностями этого этапа являются Клинический случай и Амбулаторная карта. Клинический случай содержит в себе Посещения, по которым прослеживается общая динамика обследования и лечения пациента. Посещение представляет собой обследование врачом пациента, в котором по заключениям, пришедшим из диагностического модуля, и по результатам осмотра выставляется диагноз и делаются назначения. Изменения и дополнения вносятся и в амбулаторную карту.

Так же в амбулаторном этапе находится дневной стационар, работа которого основана на назначениях, сделанных во время посещения. При согласии пациента пройти лечение в поликлинике НИИТО назначения передаются из модуля поликлиники в регистратуру модуля дневного стационара. Это позволяет предоставить больному удобные варианты графиков лечения, как только он решит обратиться в поликлинику.

### 4.3. Стационарный этап.

На этом этапе пациент проходит обследования и лечение. Существуют обязательные этапы обследования, такие как первичный осмотр, этапный и выписной эпикризы. В стационарном модуле собирается и обрабатывается информация со всех блоков информационной системы, в том числе операционного и реанимационного. При выписке больного формируется и предоставляется полный набор документов по его лечению, в том числе и статистическая карта, форма которой утверждена Министерством здравоохранения Российской Федерации. Карта заполняется автоматически на основе данных, полученных на всех этапах диагностики и лечения пациента.

Основной сущностью стационара является Врачебное заключение. Были выделены несколько типов врачебных заключений: первичный осмотр, дневник, совместный осмотр с заведующим, этапный эпикриз, предоперационный эпикриз, протокол операции, выписной эпикриз. Были проанализированы типы заключений различных отделений стационара (травматологического, нейрохирургического и др.) и предложена единая схема ведения документов по пациенту во время периода госпитализации. До введения информационной системы в стационар врачи отделений вели индивидуальный набор документов по лечению больного, что не позволяло ввести единую форму истории болезни для всего стационара. Данная схема была принята как обязательная для всех отделений, она реализована и теперь используется.В системе сделана проверка структуры документов стационарного этапа. Например, не может быть создано два выписных эпикриза или не может быть создан любой другой документ, если нет первичного осмотра.

Так же в системе реализован «ввод на основании». Для сильно связных типов заключений

(например, предоперационный эпикриз и протокол операции, которые соответствуют одной и той же операции) вводится соответствие между ними. Так при добавлении нового протокола операции нужно выбрать предоперационный эпикриз, на основании которого он будет создан. Таким образом, у нас есть возможность максимально заполнить документ данными. А так же избавиться от путаницы в случае, когда пациенту было сделано несколько операций. Автоматическое заполнение полей уже имеющейся информацией при создании новых документов реализовано во всех модулях системы.

В стационарном этапе задействовано множество отделений. Информационная система тесно связывает их в единый модуль. Среди таких составных частей можно выделить операционное отделение, отделение реанимации, отдел госпитализации. Операционный блок в системе позволяет автоматизировать работу операционных, хирургов, операционных бригад и анестезиологов. При написании предоперационного эпикриза автоматически создается заявка на операцию, которая попадает ко всем сотрудникам, задействованным в операции (заведующий операционным блоком, отделение анестезиологии, отделение переливания крови, а так же отделение, в котором состоит больной). После внесения необходимой информации отделениями заведующий операционным блоком выставляет номер операционной и время, и заявка переходит в состояние подтвержденной.

Блок отдела госпитализации позволяет вести учет возможностей медицинского учреждения по госпитализации пациентов. В данном блоке учитываются данные стационара о предполагаемой либо точной дате выписки больного, на основании этих данных составляется предварительное расписание госпитализации новых больных.

### 4.4. Диагностические подразделения

В подразделение приходит заявка, после обработки которой, перед проведением исследования пациент подписывает соглашение на исследование. После исследования на результаты пишутся заключения, которые идут в карту пациента. Диагностический модуль состоит из нескольких частей, соответствующих видам проводимых исследований (лучевая диагностика, биохимическая лаборатория). В подсистеме биохимической лаборатории программно реализован подсчет некоторых параметров аналитов, ранее проводивший вручную.

В системе реализовано взаимодействие с диагностическим оборудованием, как прямое, через прямой обмен данными, так и через сохранение промежуточной информации. Создан отдельный блок работы с электронными данными, поступающими от исследовательского оборудования, за счет чего стало возможным выдавать пациентам полную информацию о проведенном обследовании, включая объемные изображения и видео. Так же информация об исследовании сразу попадает в историю болезни и становится доступна как врачам, так и специалистам для последующего описания. Что позволяет пациенту не ходить за результатами обследования, а сразу идти к врачу. А специалисты по описанию могут более оперативно отреагировать на экстренные случаи проведения исследований.

### 4.5. Роли

Врачебная тайна, в силу особенностей самой врачебной деятельности - важнейшее понятие деонтологии (от греч. deon - должное и logos - учение) как учения о принципах поведения медицинского персонала в общении с больным и его родственниками.

Врачебная тайна – это не подлежащие разглашению сведения о пациенте, факте обращения за медицинской помощью, диагнозе и иные сведения о состоянии здоровья и частной жизни, полученные в результате обследования и лечения. Сохранность врачебной тайны гарантируется законодательно Федеральным законом от 22.07.1993 г. № 5487-1 «Основы законодательства Российской Федерации об охране здоровья граждан» (с изменениями на 27 февраля 2003 года).

Так же общий объем данных по пациенту довольно велик. И медсестре выполняющей назначения врача совершенно не нужно видеть полную историю болезни пациента.

Исходя из этих фактов в системе введено понятие роли. Список и группировка ролей показана на рисунке 3.

Для каждой роли введены:

- индивидуальный интерфейс рабочего меню, интерфейс которого построен учитывая специфику работы
- ограничения на просмотр документов и информации по пациенту
- ограничения на возможность введения и редактирования документов (это обусловлено тем, что, например заведующему отделением не нужно редактировать заключения, это не входит в его обязанности, он имеет только возможность просмотра)

При входе в систему пользователю по идентификатору роли предоставляется доступ к функционалу, предназначенному для его обязанностей. То есть для каждой роли существует своя интерфейсная часть любого документа.


Рис. 3 Роли

### 4.6. Интерфейс

Каждый документ, каждая форма может находиться в трех состояниях: добавление (заносится новая запись в базу данных), редактирование (изменение существующего документа) и просмотр (в этом состоянии нет возможности редактирования). Первоначально вид формы в разных состояниях был одинаковым, различия были лишь в функциональности и возможности редактирования. Но при разработке стационарного модуля возникла проблема отображения большого количества текстовой информации в режиме просмотра. Существующий интерфейс не позволял удобно отображать эти данные, так как пользователям приходилось пользоваться прокруткой в каждом поле ввода. В связи с этим был введен новый режим просмотра, который представляет собой набор текстовой информации, без использования элементов управления.

Данный способ позволил избавиться от незаполненных элементов управления на форме и увеличить пространство для отображения информации. Использование этих состояний форм позволяет использовать одну и ту же форму и при этом разграничивать пользователей по возможности добавления, редактирования, просмотра. В системе используются модальные окна, что позволяет увеличить рабочую площадь для пользователя. При перемещении окна можно посмотреть содержание предыдущих форм, но при этом доступ к информации строго структурирован, так как нельзя перейти на предыдущее окно, не закончив работу с текущим.

Для увеличения скорости работы с формами создана система «быстрый ввод». Врачи в специальных справочниках вводят наиболее часто встречающиеся формулировки диагнозов, рекомендаций, общих и неврологических статусов и т.п. На форме с элементами управления быстрого ввода находятся кнопки, по нажатию которых появляется структурированный список, из него можно выбрать нужные формулировки. В данный момент используется трех уровневое дерево справочников быстрого ввода, что позволяет объединять данные в группы и подгруппы. Так же возможно создать более сложные справочники быстрого ввода. Данный метод очень сильно повысил скорость заполнения врачами форм. Для добавления и редактирования форм с большим количеством информации введено поэтапное редактирование. Документ разделяется на блоки. При его редактировании пользователь перемещается между блоками. Во время смены страницы происходит промежуточное сохранение введенной информации во временную таблицу.

### 4.7. Работа с филиалами

В данный момент с информационной системой работают три филиала. Они работают через прямое подключение к главному серверу системы. Это временное решение, так как при увеличении количества удаленно работающих пользователей один сервер не сможет справиться с их обслуживанием. В ближайшее время планируется создание собственных серверов в филиалах, на которых будет работать серверная часть системы, при этом работа с базой данных будет вестись с главным сервером в НИИТО. Это позволит снизить трафик, а так же нагрузку на сервер. В данный момент тестируются временные потери при удаленной работе с базой данных.

В конечном результате планируется введение распределенного хранения информации. Проведенный анализ структуры медицинских данных и медицинской специфики показал, что она хорошо распараллеливается по специализациям медицинских учреждений. Таким образом, централизованно хранить можно только данные фактически не участвующие в лечебном процессе, такие как адрес проживания больного, место работы и т.д. (то есть данные, которые может изменить регистратура любого медицинского учреждения, в которое обратился пациент), а так же историю посещений пациентом тех или иных медицинских учреждений. Последнее нужно для доступа ко всем данным по пациенту (документы стационара, результаты обследований, диагнозы, рекомендации). Но этот доступ необходим только для чтения, которое ведется из системы конкретного медицинского учреждения, так как все эти

документы могут изменяться только их авторами. Это избавляет нас от необходимости следить за корректностью изменения информации.

**4.8. Отчетность**

В системе реализовано предоставление как внутренней, так и внешней отчетности. Отчеты разрабатываются в системе Crystal Report for VisualStudio.NET. После формирования документы экспортируются в распространенные форматы rtf и pdf.

Во внешней отчетности реализованы как печатные формы всех необходимых документов, так и статистические данные необходимые для Министерства здравоохранения Российской Федерации.

Внутренняя система отчетности позволяет генерировать любые статистические выборки, необходимые для работы врачей и администрации. Реализация данной возможности говорит о правильной разработке структуры и сущностей системы.

## 5. Заключение

Информационная система реализована на платформе .NET с использованием технологий C#, ASP.NET, MSSQL, XML.

Разработка информационной системы еще не завершена. В данный момент реализовывается связь системы с финансовым отделом. Пока не до конца реализованы системы для администрации, за исключением формирования статистики и отчетов. В планах на реализацию стоят: создание функционала для распределенного доступа к данным, для создания городской медицинской информационной сети; реализация модели онтологического построения системы; множество небольших улучшений, например, проверка орфографии.

Хотя система еще не закончена, ее реализованный функционал уже используется в Новосибирском НИИТО. Весной 2007 года был проведен первый пациент от регистрации, через стационар, до выписки (включая все документы диагностических подразделений). Пациент получил весь список необходимых документов, сформированных системой. Вся информация по больному находится в базе данных. История болезни закрыта.

Работа напрямую с врачами позволяет утверждать, что информационная система реально улучшила их работу, как в плане уменьшения затрат времени на больного, так и со стороны удобства для работы врачей и более легкого прохождения организационных этапов для пациента.

## Information system for medical institutions

Gordeev Dmitriy

The article tells about structure and interface of the medical information system. Also it shows the ways for future development. This system is intended for automation of the information flows movement and for departs cooperation of medical institution.

# Решение задач визуализации и поиска мотивов в электронной библиотеке фольклорных текстов

© Н. Д. Москин

Петрозаводский государственный университет
moskin@karelia.ru

**Аннотация**

В данной статье описаны методы исследования электронной коллекции фольклорных текстов с представлением их семантической структуры в виде теоретико-графовых моделей. В частности рассмотрены вопросы хранения и визуализации моделей, а также методика обнаружения схожих мотивов, основанная на структурном соответствии графов и подграфов.

## 1 Введение

В предыдущих докладах на конференцию RCDL [1, 3] автором была рассмотрена электронная коллекция фольклорных песен Заонежья конца XIX – начала XX века, основанная на представлении семантической структуры текстов в виде теоретико-графовых моделей. В дальнейшем эту коллекцию дополнил корпус Лужских песен Городенского хора, тексты духовных стихов о Голубиной книге, записи о народных святых Нижегородского края. В [2] был предложен проект специализированного Интернет-ресурса для представления и анализа фольклорных коллекций. Результаты анализа текстов, а также методику получения этих результатов можно записать в специальном формате, например, в формате RuleML, основанном на технологии XML [4]. Это позволит применить результаты исследования другими специалистами в рамках деятельности сетевых научных сообществ.

В данной работе рассмотрены вопросы хранения и визуализации теоретико-графовых моделей, а также методика обнаружения схожих мотивов, основанная на структурном соответствии графов и подграфов. Для хранения исходных фольклорных текстов коллекции и их теоретико-графовых моделей предусмотрен специальный формат TextGML (Textual Graph Modelling Language), разработанный на основе XML. Этот формат позволяет сохранить исходный текст и его характеристики, объекты и отношения в тексте, описать упорядоченность элементов семантической структуры, ее иерархическую организацию, выделить фольклорные мотивы и предоставить возможности для дальнейшего анализа.

При создании, редактировании и исследовании теоретико-графовых моделей необходимо использовать инструменты их визуализации. В третьей части описывается алгоритм отображения теоретико-графовых моделей фольклорных текстов, представляющий собой модификацию метода визуализации графов на основе физических аналогий. В четвертой части предложено решение проблемы поиска схожих мотивов на основе алгоритма поиска изоморфизма подграфу [17].

## 2 Описание теоретико-графовых моделей фольклорных текстов ня языке TextGML

В настоящее время разработано несколько общепризнанных стандартов описания графов и графовых моделей на основе технологии XML. Одним из предшественников таких форматов является язык GML (Graph Modelling Language) [16], который появился в результате работы, начатой на конференции «Graph Drawing-1995» в Пассау и завершенной на «Graph Drawing-1996» в Беркли. GML до сих пор поддерживается многими прикладными программами и библиотеками для работы с графами.

В 2000 году на 8-ом симпозиуме «Graph Drawing» в Вильямсбурге был предложен язык описания графов GraphXML [15]. На этом языке могут быть описаны как абстрактные графы, так и более сложные структуры: иерархии графов, динамические графы и т. д. В это же время на симпозиуме «Graph Drawing-2000» комитетом «Graph Drawing Steering Committee» был начат проект GraphML (Graph Markup Language) [11]. Рабочая встреча относительно формата файла была проведена накануне симпозиума, и на ней было согласовано создание группы, которая определила новый, основанный на языке XML, формат файла, который должен, в конечном счете, лечь в основу стандарта описания графов.

На базе GML был также создан другой язык XGMML (eXtensible Graph Markup and Modeling Language). XGMML использует все основные теги GML, а также несколько дополнительных тегов, поэтому перевод графов из одного формата в другой

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

осуществляется очень просто. Описание этого языка можно найти на сайте [13]. Другой язык GXL (Graph eXchange Language) также задумывался как стандартный формат для обмена графами [12]. Формально GXL описывает помеченные, ориентированные и упорядоченные графы, которые могут быть расширены до гиперграфов и иерархических графов. GXL поглотил такие спецификации как GraX [14], GRAph eXchange, Tuple Attribute Language и др.

Однако данные форматы предназначены для описания произвольных графов и графовых моделей, не привязанных к тексту. Для формального описания, хранения и изучения теоретико-графовых моделей текстов мы предлагаем использовать язык разметки TextGML (Textual Graph Modelling Language), разработанный на основе XML. DTD-описание данного языка изложено в работе [8]. В его основе лежат следующие элементы (теги):

- *tgml* – корневой элемент.
- *text* – элемент, определяющий границы текста. Элемент *text* имеет два атрибута: *name* – название текста и *type* – тип текста (например, «стихотворение», «басня», «статья», «эссе» и т. д.).
- *text_parameter* – характеристики текста (например, автор, год и место издания), которые определяются в виде элементов *parameter*. Каждому параметру соответствует два атрибута: *id* – идентификатор параметра и *name* – название параметра.
- *graph* – граф, соответствующий тексту. Каждый граф задается набором вершин (*node*) и ребер (*link*), соединяющих эти вершины. У элемента *graph* четыре атрибута: *id* – идентификатор графа, *name* – название графа (например, «дерево зависимостей первого предложения»), *type* – тип графа и *directed* – индикатор, указывающий, является ли граф ориентированным.
- *node* – структурные единицы текста. У этого элемента пять атрибутов: *id* – идентификатор вершины, *name* – название вершины (например, «основная форма слова»), *type* – тип вершины, *order* – порядок вершины в графе и *id_graph* – ссылка на идентификатор графа-потомка. Последний параметр позволяет организовать в тексте иерархию уровней графа, где граф низшего уровня является вершиной графа более высокого уровня.
- *link* – отношения между единицами текста. У данного элемента семь параметров: *id* – идентификатор ребра, *name* – название ребра, *source* и *target* – ссылки на идентификаторы вершины-источника и вершины-приемника, *type* – тип ребра (например, «однородность слов»), *cost* – сила связи и *order* - порядок ребра в графе.

Рассмотрим фрагмент бесёдной песни «Как назябло, навеяло лицо» из сборника В. Д. Лысанова «Досюльная свадьба, песни, игры и танцы в Заонежье Олонецкой губернии» (запись 1916 года, г. Петрозаводск) [6]:

Красна девица во тереме сидит, да
Жемчужное ожерельицо садит; да
Разсыпалось ожерельицо, да
По всему высоку терему. Да
Не собрать, не собрать жемчуга, да
Что ль ни батюшку, ни матушки, да
Что ль ни братцам, ни ясным соколам, да
Ни сестрицам, белым лебедям, да
А собрать соберет жемчужок, да
Разудалый, добрый молодец.

Теоретико-графовая модель этого текста выглядит следующим образом:

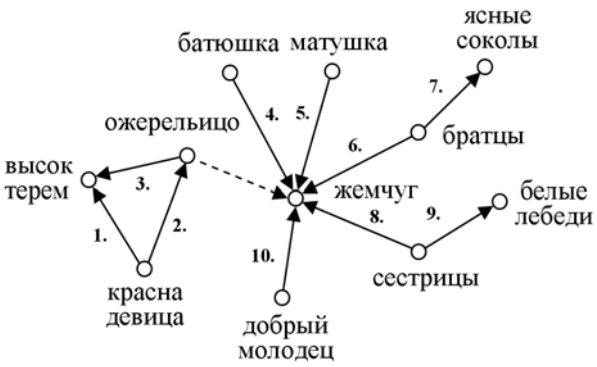



Рисунок 1. Теоретико-графовая модель фрагмента бесёдной песни «Как назябло, навеяло лицо»

Эту модель образуют шесть объектов группы «люди» (H), два объекта - «животный мир» (A), два объекта - «одежда, украшения» (CL), один объект - «постройки» (B). Единственная глобальная связь – отношение принадлежности – существует между объектами «ожерельицо» и «жемчуг», остальные связи локальные. Формат TextGML позволяет хранить сам текст, его характеристики, объекты и связи теоретико-графовой модели, их свойства: тип, упорядоченность, внутреннюю иерархию:

```
<tgml>
<text name="Как назябло, навеяло лицо" type=
"фольклорная песня">
<text_parameter>
<parameter id="p1" name="gubernia"> Олонецкая
</parameter>
<parameter id="p2" name="place_zap">1880-е годы
</parameter>
<parameter id="p3" name="sobiratel">В. Д. Лысанов
</parameter>
……
<parameter id="p8" name="movies">при игре парами
</parameter>
</text_parameter>
……
<!-- Мотив 1 -->
<graph id="g1" name="мотив 1" directed="true">
<node id="n1" type="H">Красна девица</node> во
<node id="n2" type="B">тереме</node> сидит, да
<node id="n3" type="CL">Жемчужное ожерельицо
</node> садит; да
```

```
<link id="l1" source="n1" target="n2" type="local"
order="1"/>
<link id="l2" source="n1" target="n3" type="local"
order="2"/>
</graph>

<!-- Мотив 2 -->
<graph id="g2" name="мотив 2" directed="true">
Разсыпалось <node id="n3" type="CL">
ожерельицо</node>, да
По всему <node id="n2" type="B">высоку тере-
му</node>. Да
Не собрать, не собрать <node id="n4" type="CL">
жемчуга</node>, да
Что ль ни <node id="n5" type="H">батюшку
</node>, ни <node id="n6" type="H">матушки
</node>, да
Что ль ни <node id="n7" type="H">братцам
</node>, ни <node id="n8" type="A">ясным соко-
лам</node>, да
Ни <node id="n9" type="H">сестрицам</node>,
<node id="n10" type="A">белым лебедям</node>,
да
<link id="l3" source="n3" target="n2" type="local"
order="3"/>
<link id="l4" source="n5" target="n4" type="local"
order="4"/>
<link id="l5" source="n6" target="n4" type="local"
order="5"/>
<link id="l6" source="n7" target="n4" type="local"
order="6"/>
<link id="l7" source="n7" target="n8" type="local"
order="7"/>
<link id="l8" source="n9" target="n4" type="local"
order="8"/>
<link id="l9" source="n9" target="n10" type="local"
order="9"/>
</graph>

<!-- Мотив 3 -->
<graph id="g3" name="мотив 3" directed="true">
А собрать соберет <node id="n4" type="CL"> жем-
чужок</node>, да
<node id="n11" type="H">Разудалый, добрый моло-
дец </node>.
<link id="l10" source="n11" target="n4" type="local"
order="10"/>
</graph>

<!-- Структура мотивов песни -->
<graph id="g4" name="граф мотивов" directed=
"true">
<!-- Мотивы песни -->
<node id="n12" name="мотив 1" type="motive"
id_graph="g1"/>
<node id="n13" name="мотив 2" type="motive"
id_graph="g2"/>
<node id="n14" name="мотив 3" type="motive"
id_graph="g3"/>
<node id="n15" name="песня" type="song"/>
<!-- Глобальные связи между вершинами -->
<link id="l11" source="n3" target="n4" type=
"global"/>
<!-- Глобальные связи между мотивами -->
<link id="l12" source="n15" target="n12" type=
"global"/>
<link id="l13" source="n15" target="n13" type=
"global"/>
<link id="l14" source="n15" target="n14" type=
"global"/>
</graph>
</text>
</tgml>
```

## 3. Визуализация теоретико-графовых моделей фольклорных текстов

При создании, редактировании и исследовании теоретико-графовых моделей фольклорных текстов необходимо иметь инструменты их визуализации. При этом алгоритм визуализации должен учитывать особенности построения фольклорного текста, и, следовательно, особенности теоретико-графовой модели. В методе, основанном на физических аналогиях, [5] граф рассматривается как система объектов с силами, взаимодействующими между этими объектами, где, например, вершины графа считаются телами, а ребра – пружинами. Рассмотрим модификацию этого метода для нашего случая [8]. Определим силу, приложенную к вершине $u$, по следующей формуле:

$$F(u) = \sum_{e=(u,v)\in E} f_e + \sum_{(u,v)\in V^2} g_{(u,v)} + \sum_{e=(u,v)\in E} h_e$$

где $f_e$ – сила растяжения, действующая на вершину $u$ из-за пружины $(u,v)$. Здесь $x$-я координата силы $f_e$ вычисляется по формуле:

$$f_e^x = k_e^{(1)}(d(u,v) - \lambda_1\left|p(u) - p(v)\right| - l_{\min})\frac{x_u - x_v}{d(u,v)},$$

где $d_{(u,v)}$ обозначает расстояние между $u$ и $v$, $k_e^{(1)}$ – коэффициент жесткости (упругости) пружины между $u$ и $v$. Чем он больше, тем сильнее пружина стремится установить расстояние между $u$ и $v$, равным $l_e = l_{\min} + \lambda_1 \cdot \left|p(u) - p(v)\right|$. Подобное выражение для естественной длины пружины $l_e$, где $p(u)$ и $p(v)$ – номера слов в тексте, соответствующих объектам $u$ и $v$, $l_{\min}$ – минимальная длина пружины, позволяет ближе расположить те объекты, которые в тексте находятся недалеко друг от друга. Коэффициент $\lambda_1 \geq 0$ характеризует значимость данного критерия.

Силу отталкивания $g_{(u,v)}$, существующую между вершинами $u$ и $v$, определим по следующей формуле:

$$g_{(u,v)}^x = \frac{\lambda_2}{\Delta(u) + \Delta(v)} \frac{x_u - x_v}{(d(u,v))^3},$$

где $\Delta(u)$ – число ребер, инцидентных вершине $u$, $\lambda_2$ – коэффициент отталкивания, постоянный для всех вершин. Чем больше будет у вершины инцидентных ребер, тем меньше будет коэффициент отталкивания, а, следовательно, и сила $g_{(u,v)}$ для всех вершин $v$. Это позволит расположить вершины с большой степенью (основные персонажи фольклорного текста) в центре экрана, а вершины с меньшей степенью ближе к его границам.

Чтобы учитывать порядок появления связей в сюжете фольклорного текста, для каждого ребра $e=(u,v)$ введем дополнительную силу $h_e$. Эта сила будет стремиться расположить ребра графа как можно ближе к установленным заранее упорядоченным точкам $q_{(u,v)}$. Точки $q_{(u,v)}$ следует расположить последовательно на одинаковом расстоянии друг от друга по окружности (или полуокружности) с центром в середине экрана:

$$h^x_{e=(u,v)} = k^{(3)}_{(u,v)} \cdot d(q_{(u,v)}, c_{(u,v)}) \cdot (x_q - x_c),$$

где $c_{(u,v)}$ – центральная точка ребра (центр ребра), координаты которой вычисляются как среднее арифметическое координат вершин $u$ и $v$, а $k^{(3)}_{(u,v)}$ – коэффициент силы притяжения между $q_{(u,v)}$ и $c_{(u,v)}$. Чем он больше, тем сильнее ребро $e=(u,v)$ стремится к точке $q_{(u,v)}$.

Алгоритм визуализации работает в два этапа: вначале вершины размещаются на плоскости случайным образом, затем выполняется последовательность итераций до стабилизации, на каждой из которых для всех вершин $u$ вычисляется сила $F(u)$ и происходит перемещение вершины в направлении этой силы на расстояние, пропорциональное модулю силы [5]. Применение данного алгоритма позволит упорядочить элементы графа по мере их появления в сюжете фольклорного текста, сгруппировать вершины и ребра согласно структуре мотивов и их функциональному весу.

Обычно при анализе результатов визуализации графов используют ряд эстетических критериев, среди которых можно выделить следующие: минимизация пересечений ребер, минимизация области размещения, минимизация и унификация длин ребер, минимизация и унификация сгибов, максимальная симметричность и др. [5]. В нашем случае дополнительным критерием является упорядоченность ребер, отражающая развитие сюжета по времени.

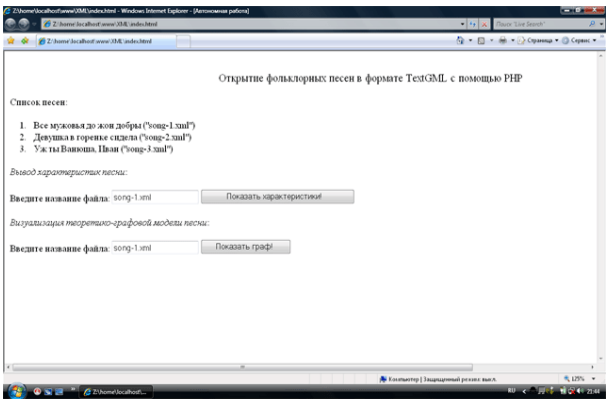



Рисунок 2. Процедура выбора фольклорного текста

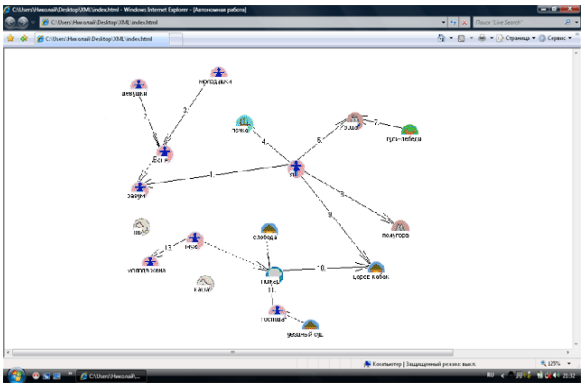


Рисунок 3. Отображение теоретико-графовой модели в окне браузера

С помощью рассмотренных выше коэффициентов можно задать значимость того или иного критерия, например:

- $\lambda_1$ и $k^{(3)}_{(u,v)}$ определяют упорядоченность вершин и ребер графа;
- $k^{(1)}_e$ минимизирует и унифицирует длины ребер;
- $\lambda_2$ влияет на минимизацию области размещения и число пересечений ребер.

В результате настройки данных коэффициентов алгоритм показал достаточно хорошие результаты: пересечения ребер появляются тогда, когда их нельзя избежать, ребра графа упорядочены по мере появления связей в сюжете, длины ребер соответствуют расстоянию между объектами в фольклорном тексте. Недостатком алгоритма является то, что результат визуализации достаточно сильно зависит от начального (случайного) распределения вершин на плоскости. Поэтому для получения более хорошего изображения следует либо использовать несколько (до десяти) случайных разбиений, а затем выбрать лучший вариант, либо вершины графа изначально расположить на плоскости в соответствии с некоторым порядком (например, по мере появления объектов в тексте).

На рисунках 2 и 3 представлены процедура выбора фольклорного текста, который хранится в формате TextGML, и визуализация его теоретико-графовой модели, выполненные на языке PHP 5.

## 4. Поиск мотивов в электронной коллекции фольклорных текстов

Другой важной задачей, возникающей при работе с фольклорной коллекцией, является проблема обнаружении в текстах схожих мотивов и их сравнительный анализ. Мотивы – это композиционные фрагменты, которые повторяются в других песнях (не всегда в одной и той же последовательности) и служат исходными элементами для построения новых текстов. По выражению известного фольклориста Б. Н. Путилова мотив является «узловой категорией художественной организации произведения фольклора».

Рассмотрим фрагмент беседной песни, записанной Ф. Студитским в 1841 году [8]:

На матушке на Неве
Гуси, лебеди сидели,
Гуси, лебеди сидели,
Серы утки налетели,
Серы утки налетели,
Свежу воду помутили.

Допустим, перед исследователем стоит задача: обнаружить подобный мотив в других песнях (возможно в скрытой форме). Самый простой способ – это поиск по ключевым словам: *гуси, лебеди, Нева, серы утки, вода*. Однако это решение будет недостаточным по следующим причинам. Во-первых, автор, исполняя произведение, мог заменить существительные, прилагательные и глаголы синонимами или близкими по звучанию словами. Во-вторых, наличие в тексте ключевых слов еще не говорит об их семантической связности, тем более о наличие схожего мотива. Например, объекты «девушка» или «парень» встречаются почти во всех текстах, образуя совершенно разные сюжеты.

Другое решение основано на использовании графов. В этом случае задача поиска мотива в коллекции сводится к задаче поиска изоморфного подграфа (например, для данного фрагмента текста искомый граф изображен на рисунке 4).

1) гуси, лебеди *сидели* на Неве;
2) серы утки *налетели* к Неве;
3) серы утки *помутили* свежу воду;
4) вода *есть* (*принадлежит*) в Неве.

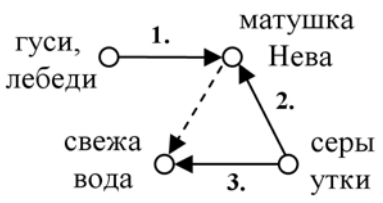
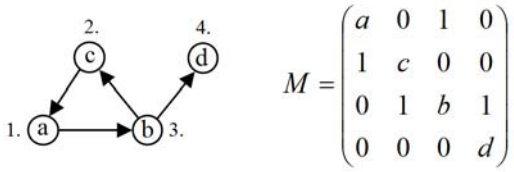



Рисунок 4. Граф мотива песни

Для решения данной задачи применим классический алгоритм поиска изоморфизма подграфу, предложенный Ульманом [17]. Пусть дан граф $G = (V, E, \alpha, \beta, L_v, L_e)$ с множеством вершин $V = \{v_1, v_2, \ldots, v_n\}$ и множеством ребер $E \subseteq V \times V$. Функции $\alpha : V \to L_V$ и $\beta : E \to L_E$ задают метки вершинам и ребрам $G$ соответственно. Граф можно представить с помощью матрицы смежности следующего вида: $M = \{m_{ij}\}_{i,j=1}^{n}$, где $m_{ii} = \alpha(v_i)$ и $m_{ij} = \beta((v_i, v_j))$ для $i \neq j$ (см. пример на рис. 5).




$$M = \begin{pmatrix} a & 0 & 1 & 0 \\ 1 & c & 0 & 0 \\ 0 & 1 & b & 1 \\ 0 & 0 & 0 & d \end{pmatrix}$$

Рисунок 5. Граф G и его матрица смежности

Очевидно, что матрица $M$ является не единственным представлением графа $G$. Если в ней определенным образом поменять строки и столбцы, то полученная матрица тоже будет однозначно определять $G$.

Определение 1. Матрица $P = \{p_{ij}\}_{i,j=1}^{n}$ называется матрицей перестановок, если выполняются следующие условия:

1. $p_{ij} \in \{0,1\}$ для $i, j = 1, \ldots, n$;
2. $\sum_{i=1}^{n} p_{ij} = 1$ для $j = 1, \ldots, n$;
3. $\sum_{j=1}^{n} p_{ij} = 1$ для $i = 1, \ldots, n$.

Если граф $G$ представлен с помощью матрицы смежности $M$ размерности $n \times n$ и $P$ - это матрица перестановок размерности $n \times n$, то тогда матрица

$$M' = PMP^T,$$

где $P^T$ - транспонированная матрица $P$, также является матрицей смежности для графа $G$. При этом если $p_{ij} = 1$, то $j$-я вершина в $M$ становится $i$-й вершиной в $M'$. Например, на рисунке 6 приведена матрица перестановок, которая первую вершину делает второй ($p_{21} = 1$), вторую вершину – четвертой ($p_{42} = 1$), четвертую вершину – первой ($p_{14} = 1$), а третью оставляет без изменения ($p_{33} = 1$).

$$P = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad M' = \begin{pmatrix} d & 0 & 0 & 0 \\ 0 & a & 1 & 0 \\ 1 & 0 & b & 1 \\ 0 & 1 & 0 & c \end{pmatrix}$$

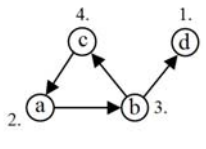



Рисунок 6. Матрица перестановки P и новая матрица смежности M'

Определение 2. Обозначим матрицу $S_{k,m}(M)$ размерности $k \times m$, которая получена из $M$ путем удаления строк с номерами $k+1, \ldots, n$ и столбцов $m+1, \ldots, n$, где $k, m \leq n$. Таким образом, $S_{k,m}(M)$

совпадает с $M$ только при $i=1,\ldots,k$ и $j=1,\ldots,m$. Например, для матрицы смежности $M$ на рисунке 5, $S_{2,3}(M)$ принимает следующий вид:

$$S_{2,3}(M)=\begin{pmatrix} a & 0 & 1 \\ 1 & c & 0 \end{pmatrix}$$

Теперь сформулируем понятие изоморфизма подграфу в терминах матриц смежности и перестановок. Пусть $G_1$ и $G_2$ - графы с матрицами смежности $M_1$ и $M_2$ размерности $m\times m$ и $n\times n$ соответственно, где $m\le n$. Тогда изоморфизм графа $G_1$ подграфу в $G_2$ существует тогда, когда существует матрица перестановок $P$ размерности $n\times n$, такая что

$$M_1=S_{m,m}(PM_2P^T).$$

Рассмотрим рекурсивную процедуру Backtrack (на входе счетчик $k=1$, $G=(V,E,\alpha,\beta,L_v,L_e)$, $n=|V|$, $G_I=(V_I,E_I,\alpha_I,\beta_I,L_v,L_e)$, $m=|V_I|$, $M$ и $M_I$ - матрицы смежности для $G$ и $G_I$ соответственно; $P=(p_{ij})$ - матрица перестановок $n\times n$;).

**Procedure Backtrack**:
1. **if** $k>m$ **then**
2. // $P$ определяет изоморфизм $G_I$ подграфу $G$
3. Print ($P$);
4. **return**;
5. **end if**;
6. **for** $i$ :=1 to $n$ **do**
7. $p_{ki}:=1$;
8. **for** $j\ne i$ **do**
9. $p_{kj}:=0$;
10. **end for**;
11. **if** $S_{k,k}(M_I)=S_{k,n}(P)M(S_{k,n}(P))^T$ **then**
12. $Backtrack(M,M_I,P,k+1)$;
13. **end if**;
14. **end for**;
15. **return**.

Алгоритм основан на идее поиска всех изоморфизмов подграфу с помощью последовательного определения строка за строкой матрицы перестановок $P$. Из определения 1 следует, что каждая строка $k$ матрицы $P$ содержит в точности одно ненулевое число $p_{ki}=1$ (остальные элементы $p_{kj}$ строки $k$ при $j\ne i$ равны 0). Рекурсивная процедура *Backtrack* начинается установкой первого элемента $p_{11}=1$, а всех остальных элементов первой строки 0. Если $S_{1,n}(P)$ - неполное соответствие, которое представляет изоморфизм подграфу, тогда процедура *Backtrack* рекурсивно вызывается еще раз и вторая строка предварительно устанавливается. Процесс продолжается до тех пор, пока $m$ строк $P$ не будут успешно установлены и найден изоморфизм подграфу или условие на шаге 11 не будет выполнено. В любом случае процедура возвращается на предыдущий уровень и проверяет другие значения $p_{ki}$. В данный алгоритм можно ввести дополнительные ограничения на матрицу перехода $P$, например, учитывающие порядок появления связей в графе. Это позволит сократить количество переборных вариантов и приведет к ускорению работы алгоритма.

В настоящее время процедура поиска схожих мотивов реализована в локальной версии системы, написанной на языке Delphi 7.0. Искомый мотив можно задать двумя способами: либо пользователь самостоятельно определяет объекты и связи, либо выделяет в фольклорном тексте границы мотива и программа автоматически строит граф (см. рис. 7).

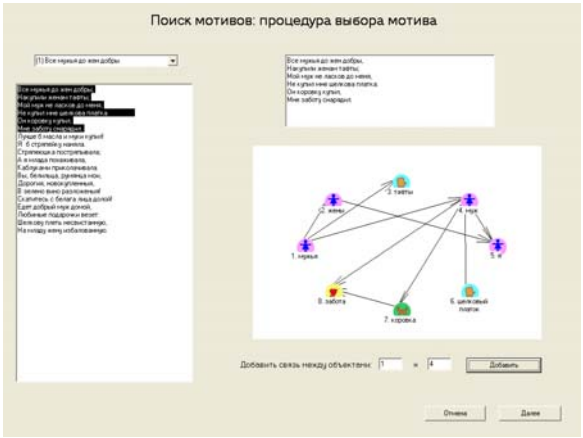



Рисунок 7. Процедура поиска мотивов

Поскольку алгоритм ищет точное соответствие графов и подграфов, наилучшие результаты получаются при поиске мотивов с 4-6 объектами. Особенно хорошие результаты программа выдает при сравнении фольклорных текстов с одинаковыми сюжетами, записанными разными собирателями (например, «Как назябло, навеяло лицо», «Позябло, позябло лицо» и «Разсыпалось ожерельице»). В остальных случаях получается либо много лишних вариантов (при небольшом числе объектов), либо их отсутствие (при числе объектов больших 10).

При проведении подобного анализа программа способна обнаружить «скрытые» мотивы, которые сложно обнаружить традиционными методами. Например, в бесёдной песне «Девушка в горенке сидела» объект «коса» встречается как в начале текста (девушка *плела* русу косу), так и в конце (коса *повысушила* парня, *с ног сронила*), образую два мотива. С другой стороны, эти мотивы взаимосвязаны и их можно рассматривать как единый мотив, разбитый в тексте на несколько частей. При традиционном поиске обнаружить данный фрагмент будет достаточно сложно. При использовании теоретико-графовых моделей мотив представлен как связанный подграф, который структурно не отличается от остальных подграфов. Этот эффект особенно проявляется при исследовании больших текстов.

Для того чтобы данный метод давал лучшие результаты, его необходимо дополнить поиском по ключевым словам. Также можно ввести ограничения на принадлежность объектов к определенной группе, на тип и порядок появления связей в тексте.

## 5. Заключение

В данной работе были рассмотрены метод визуализации теоретико-графовых моделей и алгоритм поиска схожих мотивов, основанный на структурном соответствии графов и подграфов. Кроме того, в статье описывается способ хранения теоретико-графовых моделей на языке TextGML.

Предложенные методы могут найти свое применение при составлении тематических указателей, указателей фольклорных мотивов и формул. Также их можно использовать при исследовании других видов семантических сетей, построенных на основе текстов.

## Литература


[1] Варфоломеев А. Г., Кравцов И. В., Москин Н.Д. Информационная система по фольклорным песням Заонежья как инструмент формализации и классификации песен // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды IV Всероссийской научной конференции RCDL'2002 – Дубна, 2002. – Т. 2. – С. 143-147.

[2] Варфоломеев А. Г., Кравцов И. В., Москин Н.Д. Проект специализированного Интернет-ресурса для представления и анализа фольклорных песен // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды V Всероссийской научной конференции RCDL'2003. – Санкт-Петербург, 2003. – С. 339-343.

[3] Варфоломеев А. Г., Москин Н. Д. Об электронной коллекции фольклорных песен с теоретико-графовой формализацией текстов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Сборник аннотаций стендовых докладов III Всероссийской научной конференции RCDL'2001. – Петрозаводск, 2001. – С. 20.

[4] Каргинова Н. В., Кравцов И. В., Москин Н. Д., Варфоломеев А. Г. Проект электронной библиотеки методик и результатов исследований текстовых коллекций для системы «Источник» // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды X Всероссийской конференции RCDL'2008. – Дубна: ОИЯИ, 2008. – С. 239-245.

[5] Касьянов В. Н., Евстигнеев В. А. Графы в программировании: обработка, визуализация и применение. – СПб.: БХВ-Петербург, 2003. – 1104 с.

[6] Лысанов В. Д. Досюльная свадьба, песни, игры и танцы в Заонежье Олонецкой губернии. – Петрозаводск: Северная скоропечатня Р. Г. Каца, 1916. – 119, 30 с.

[7] Москин Н. Д. Теоретико-графовые модели структуры фольклорных песен и методы их анализа // Круг идей: Междисциплинарные подходы в исторической информатике. Труды X конференции Ассоциации «История и компьютер» / Под ред. Л. И. Бородкина, И. М. Гарсковой. – М.: Изд-во МГУ, 2008. – С. 280-300.

[8] Москин Н. Д. Теоретико-графовые модели структуры фольклорных текстов, алгоритмы поиска закономерностей и их программная реализация // Дисс. на соиск. уч. ст. к.т.н. – Петрозаводск, 2006.

[9] Новиков А. И. Семантика текста и ее формализация. М: Наука, 1983. – 215 с.

[10] Скороходько Э. Ф. Семантические сети и автоматическая обработка текста. – Киев: Наукова думка, 1983. – 218 с.

[11] Спецификация GraphML. http://graphml.graphdrawing.org/index.html

[12] Спецификация GXL. http://www.gupro.de/GXL/

[13] Спецификация XGMML. http://www.cs.rpi.edu/~puninj/XGMML/

[14] Ebert J., Kullbach B., Winter A. GraX: Graph Exchange Format // In Proceedings of the Workshop on Standard Exchange Formats (WoSEF) at ICSE'00, 2000. http://www.gupro.de/winter/Papers/ebert+2000.pdf

[15] Herman I., Marshall M. S. GraphXML – An XML-based graph description format // Proceedings of Graph Drawing 2000 (Lecture Notes in Computer Science, vol. 1984). – Springer: Berlin, 2001. – P. 52-62. http://www.euronet.nl/users/scott/public_html/publications/GraphXMLShort.pdf

[16] Himsolt M. GML: A portable Graph File Format // Technical report, University at Passau, 1997. http://www.infosun.fim.uni-passau.de/Graphlet/GML/gml-tr.html

[17] Ullmann J. R. An algorithm for subgraph isomorphism // Journal of the Association for Computing Machinery. – 1976. – Vol. 23, No. 1. – P. 31-42.


## The solution of visualization and motive search problems in the digital library of folklore texts


N. D. Moskin



This paper describes research methods for the digital library of folklore texts with graph representation of their semantic structures. In particular the questions of storing and visualization graph models, motive search algorithm based on graph and subgraph matching are considered.

# СТЕНДОВЫЕ ДОКЛАДЫ

# POSTER PAPERS

# Технологический процесс подготовки изданий на примере Фундаментальной электронной библиотеки «Русская литература и фольклор». Текущее состояние и принципы модернизации

© С. И. Трифонов, А.Е. Поляков

ФУГП НТЦ «ИНФОРМРЕГИСТР»
trf@ya.ru, pollex@mail.ru

**Аннотация**

В докладе описывается технология перевода изданий из печатного вида в электронную форму, разработанная и внедрённая в отделе электронных библиотек ФУГП НТЦ «ИНФОРМРЕГИСТР», в частности, в Фундаментальной электронной библиотеке «Русская литература и фольклор» (http://www.feb-web.ru)
Обсуждаются как текущее состояние технологии, так и принципы проводящейся в настоящее время модернизации
Перевод изданий из печатного вида в электронную форму представляет собой трудоёмкую задачу, крайне актуальную для библиотечного сообщества. Далеко не все стадии подготовки электронного массива поддаются полной автоматизации, поэтому качественно подготовленный документ оказывается во многом результатом ручного труда. Цель наших разработок - создание программного окружения, оптимизирующего ручные процедуры, в котором трудоёмкий полуавтоматический процесс подготовки документов выполнялся бы максимально удобно, просто и быстро.

## 1 Структура электронного издания

Существуют по крайней мере несколько способов оформления документа в электронном виде. Технология, обсуждаемая в докладе, предназначена для подготовки документов в формате HTML, с полноценной отработкой структурных элементов, стандартных для традиционных печатных изданий: иллюстрации, сноски, разбиение на страницы. Результирующий электронный документ включает также важную информацию, содержащуюся в исходном документе условно, но не достаточно формально. Это - логическая структура документа, отображающаяся обычно в виде оглавления, а также ссылки между документами и фрагментами.

Все электронные издания и произведения, которые готовятся по технологии, в обязательном порядке снабжаются библиографическими описаниями в формате, соответствующем стандарту ГОСТ 7.1-2003.

## 2 Процесс подготовки электронных изданий

По технологическим причинам, процесс производится раздельно для каждого издания. Как правило, под изданием подразумевается книга целиком.

В общем виде процесс подготовки разбивается на три фазы, хотя в зависимости от формата издания и целей работы возможны вариации.

На фазе оцифровки производятся серийные процедуры: постраничное сканирование издания, программное распознавание текста, первичное вычитывание текста. Основной результат фазы — единый файл в формате Word, содержащий текстовую информацию издания. По необходимости, на этой же фазе готовятся иллюстрации и рисунки в адекватном качестве

Подобная процедура весьма стереотипна, она подразумевается в любом варианте подготовки электронного издания, не только в обсуждаемом. Выполнение её, безусловно, трудоёмко, но требует от исполнителей достаточно стереотипных навыков. Качество результата этой фазы невысокое, что осложняет технологию подготовки на двух следующих фазах.

Фаза разметки на данный момент производится в тестовом процессоре Word, с использованием специально разработанных средств. Здесь происходит оформление логической структуры издания и включённых в него произведений, а также всех структурных элементов, связанных с

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

текстом: ссылки, сноски, иллюстрации, страницы и их нумерация. Одновременно производится окончательное вычитывание.

Синхронно с разметкой, для всего издания и для произведений оформляются библиографические описания.

На фазе окончательной подготовки издание разделяются на отдельные произведения, оформленные в формате HTML. Здесь производятся проверки целостности информации и комплексной оценки качества результата.

Две завершающие фазы подразумевают от исполнителей значительно более высокий уровень ответственности и специализированной подготовки. Существующие на данный момент программные средства явно недостаточны, чтобы заменить ручной труд этих специалистов без потери качества результата. В то же время, вполне возможно радикально облегчить их труд, автоматизировав наиболее трудоемкие части процедур.

В настоящее время мы начали модернизацию наших технологий, исходя именно из этого принципа: ***создать среду, в которой процессы фаз разметки и окончательной подготовки были бы поддержаны программно — причём максимально удобным и современным способом***. Необходимо заметить: добиваясь эстетических качеств у программного продукта, мы в конечном итоге сокращаем издержки технологии и существенно оптимизируем затраты.

## 3 Роль среды Word в текущей технологии и в условиях модернизации

Дополнительная цель, преследуемая при создании среды: избавиться от издержек в использовании промежуточного формата Word, и в частности слить фазы разметки и окончательной подготовки. В связи этим уместно сделать следующий комментарий.

В период формирования технологии во второй половине 90-х годов, выбор текстового процессора Word был обоснованным решением. Он позволил автоматизировать многие ручные процедуры, характерные для процесса подготовки. В то же время, концепция электронного документа, поддерживаемая редактором Word, отличается от необходимой для наших целей концепции электронного издания, ориентируемой на возможности формата HTML. Различия эти на первый взгляд незначительны, но в рамках технологии, при подготовке больших текстовых массивов, приводят к очень серьёзным трудностям. Учёт этих трудностей привёл к необходимости поддерживать в среде Word не только систему "скриптов", но и специальный язык разметки документов. Выстроенная таким образом и существующая на данный момент технология оказалась вполне работоспособной, но весьма тяжеловесной.

С момента создания нашей технологии, средства Word по поддержке формата HTML были развиты радикально. Однако с того же времени столь же радикально были развиты и другие средства текстового редактирования, в частности и средства непосредственного редактирования файлов в формате HTML. Поэтому в контексте задач по разметке текста принципиальной необходимости в промежуточном формате Word на нынешний день нет.

В то же время, формат Word остаётся базовым для первого этапа подготовки - оцифровки. Наиболее удобные из наработанных средств для формата Word планируется перенести сюда, частично разгрузив тем самым более ответственную последующую фазу разметки.

## 4 Анализ доступных решений

В принципе, подготовку изданий в желаемом формате можно выстроить вокруг любого текстового редактора. Однако процесс подготовки включает много различных операций, и для достижения качества результата все эти операции необходимы. Существенная особенность нашей задачи: операции должны проводиться над большими массивами текста. Таким образом, формальный список требований (вернее, пожеланий) к создаваемой среде состоит из многих пунктов, каждый из которых в отдельности не имеет принципиального значения. Общее требование: оптимизировать по возможности большую часть разнообразных операций.

Приступая к построению среды, мы ориентировались на максимальное использование уже готовых программных решений, соответственно, к минимизации собственных затрат на разработку. Анализ доступных решений выявил довольно странную картину.

Оказалось, что существует множество вариантов, которые реализуют требуемую нами функциональность, но только по частям, фрагментарно. В целом, использование каждого из готовых решений для наших целей оказалось не удобно. При этом существенно, что открытое программное обеспечение по функциональности показало себя ничем особенным не хуже и не лучше, чем коммерческие варианты.

В результате мы отказались от варианта взять определённый программный продукт, и выстраивать наше решение на нём одном, пользуясь возможностями «плагинов» и настроек. Вместо этого было принято решение создавать самостоятельную программу, используя доступные решения с открытым кодом (Open Source).

## 5 Анализ концепции интерфейса

Дополнительный анализ показал, что ключевой частью задачи по построению нашей среды оказывается концепция интерфейса. Технические же средства — поддержка форматов, алгоритмов —

доступны во многих качественно оформленных вариантах, и их использование не составляет тяжёлой проблемы.

К решению проблемы интерфейса нас подтолкнуло следующее обстоятельство, которое можно было бы назвать недоразумением. Общеизвестный стереотип интерфейса под названием «редактор» на самом деле предоставляет средства не для *редактирования* документов, а для их *создания*. Соответственно, при построении нашей среды нам важно реализовывать как раз те части функциональности, которые в стереотипе «редактора» делаются неудобно. И наоборот: важнейшие для стереотипа задачи в нашем случае оказываются второстепенными:

- «в редакторе» главное время пользователя тратится на создание текста; «в среде» - это происходит довольно редко;
- «в редакторе» документ, как правило, маленький — его создают; «в среде» он всегда большой
- «в редакторе» средства навигации не слишком важны, а структуры, по которым производится навигация (ссылки, сноски, разделы) очень динамично изменяются; «в среде» навигация крайне важна, а навигационные структуры требуют постоянной проверки на корректность, а также внятной маркировки проблемных ситуаций
- «в редакторе» запуск «скриптов», делающих автоматические проверки и исправления, производятся в свободном порядке; «среда» должна действительно быть средой, в которой автоматические процедуры могут запускаться на регулярной основе, а отработка ошибок и сообщений этих процедур должна быть удобной и очевидной

## 6 Программная среда для подготовки электронных документов

Отталкиваясь от этих наблюдений, у нас сложилось следующее интерфейсное решение.

Среда выстраивается вокруг режима просмотра документа и его оглавления. При этом всевозможные ошибки и конфликты, неизбежные при подготовке, отображаются пометками на полях.

Операции редактирования при подготовке разделяются на две группы. Если автоматические процедуры («скрипты») могут выполняться как на всём массиве, так и на его частях, пользовательские операции выполняются только на небольших фрагментах, как правило, размером в один, максимум несколько абзацев текста.

Автоматические процедуры, используемые нами в технологии, очень сильно зависят от контекста конкретной задачи. Как правило, конкретная процедура не может быть рассчитана на «все случаи жизни», и возможность отследить «неправильную ситуацию» при запуске процедур — крайне важная техническая задача. Предлагаемая схема интерфейса дает такую возможность: процедура получает возможность оставить после своего выполнения «заметку на полях», в наглядном и удобном виде.

Пользовательские задачи, требующие повышенного внимания, в данном решении реализуются, как операции над ясно определёнными фрагментами текста. Таким образом, среда гарантирует, что при исправлении фрагмента не испортятся другие фрагменты, и процесс редактирования выстраивается, как понятная последовательность атомарных процедур.

## 7 Платформа разработки

Рассмотрев различные варианты, в качестве платформы для своего проекта мы выбрали решение из проекта Mozilla. Конкретно, это версия продукта XUL Runner, с включённой поддержкой всех возможностей языка Python. Это решение было стабилизировано разработчиками Mozilla сравнительно недавно, и обещает стать популярным в разнообразных мировых разработках открытого программного обеспечения.

Привлекательность решения, с нашей позиции, заключается в следующем свойстве комбинированного продукта. XUL Runner представляет собой не готовое приложение, а среду разработки приложений, обеспечивающую полноценный доступ к ядру, общему для всех продуктов проекта Mozilla. В рамках самой среды можно создавать различные автономные ("клиентские") приложения с определённым креном в сторону сетевых решений. Это очень удобно для наших задач, поскольку результат нашей технологии — файлы в формате HTML — предназначены для сетевого использования. В то же время, использование только XUL Runner порождает определённые проблемы: встроенные средства языка Java Script вполне достаточны для клиентской части сетевого решения, но представляются недостаточно удобными для полновесного локального приложения, каким является наша среда.

Подключение к конфигурации языка Python со всеми его стандартизированными пакетами полностью снимает этот недостаток, открывая доступ к всевозможным общедоступным техническим средствам.

## 8 Заключение

Предыдущий вариант нашей технологии разрабатывался с середины 90-х годов прошлого века. За прошедшее время он, безусловно, морально устарел. В то же время, до последнего времени сохранялась ситуация, когда состояние общих программных технологий не позволяло существенно облегчать процесс подготовки текстов. По нашим предположениям, сейчас ситуация изменилась, и доступные в настоящее время

средства достаточны, чтобы эффективно провести модернизацию процесса подготовки.

На момент написания этого доклада обсуждаемая программа (техническое название "H4") находится в завершающей стадии разработки, и готовится к внедрению в технологический процесс.

## Литература


[1] Вигурский К.В. К проблеме оценки качества электронных библиотек. *Электронные библиотеки России: управление и координация: Всероссийская научно-практическая конференция*, Москва, РГБ, 21–22 февраля 2007 г.

[2] Вигурский К. В. Представление текстовых произведений в электронной форме (Опыт Фундаментальной электронной библиотеки "Русская литература и фольклор"). *Книга и мировая цивилизация: Материалы XI Международной научной конференции по проблемам книговедения* (Москва, 20–21 апреля 2004 г.): В 4 т. – М.: Наука, 2004. – Т. 1. —С. 346 – 386

[3] Вигурский К. В., Пильщиков И.А. Филология и информатика. *Известия АН*. Серия литературы и языка, 2003, Т. 62, № 2. – С. 9–16

[4] Кузнецов Ф.Ф., Вигурский К. В. Фундаментальная электронная библиотека «Русская литература и фольклор»**.** *Вестник Российского гуманитарного научного фонда*, 2004, № 3 (36). – С. 54 – 63

[5] Creating Python GUI Applications using XULRunner. http://pyxpcomext.mozdev.org/no_wrap/tutorials/pyxulrunner/python_xulrunner_about.html


## Technological process for preparation of publications, on example of Fundamental electronic library «Russian literature and Folklore». Current state and the modernization principles.


S.I.Trifonov, A.E.Polyakov



We discuss the technology used for transformation of publications from their printed form to electronic one. This technology was developed and adapted in the Electronic library dept. of the SRC INFORMREGISTR, in particular for the Fundamental electronic library «Russian literature and Folklore» (http://www.feb-web.ru). The current state and the principles of modernization are considered.

# Информационно-поисковая система «Справочные издания о населении и природе Тамбовской области XIX-XX вв.» в образовательной среде вуза♣

Л.А. Пронина, Н.Е. Копытова, Н.В. Шаталова, А.Н. Евстигнеев

Тамбовский государственный университет имени Г.Р. Державина
ralk@tsu.tmb.ru

## Аннотация

Рассмотрены вопросы создания информационно-поисковой системы «Справочные издания о населении и природе Тамбовской области XIX-XX вв.» и ее использование в образовательной среде вуза и региона.

Эффективность процессов обучения зависит от качества образовательной среды. В современных условиях она динамично изменяется за счет увеличения количества электронных ресурсов, которые занимают все более важное место в системе информационного обеспечения науки, образования и практики. Выбор типа продукта и его содержания зависит от информационных потребностей пользователей.

Традиционно одним из достоверных источников информации считаются справочные издания: содержащаяся в них фактографическая информация является доказательной базой, нормой использования терминов и понятий и т.п. Поэтому велика популярность и справочных баз данных, которые востребованы на информационном рынке. Современная тенденция в развитии электронных ресурсов, на наш взгляд, – это создание мощных интегрированных информационно-поисковых систем с разнообразным и разнородным контентом, обладающих мощными поисковыми возможностями. Поэтому было принято решение – создать информационно-поисковую систему краеведческой направленности (население и природа края), полнотекстовую (тексты справочных изданий), с достаточной хронологической глубиной (200 лет), с возможностями использования в научно-образовательной, практико-производственной деятельности. Изучение состояния регионального сегмента электронного пространства показало, что подобных электронных ресурсов нет, в то же время потребность в краеведческом электронном ресурсе – огромна.

Ориентируясь на региональные информационные потребности, нами создается информационно-поисковая система (ИПС) «Справочные издания о населении и природе Тамбовской области XIX-XX вв.». Данная ИПС задумывалась как электронный ресурс с широкими поисковыми возможностями и дидактической направленностью. Впервые осуществляется оцифровка незащищенных авторским правом постоянно востребованных справочных изданий о Тамбовской области (географические словари, административно-территориальные справочники, сборники статистических сведений, обзоры Тамбовской губернии и т.п.) и создание ИПС с дидактическим эффектом. Предполагается перевести в электронный вид соответствующие разделы универсальных, отраслевых и специализированных справочных изданий федеральных и местных издательств с большим хронологическим охватом.

На первом этапе максимально полно выявлялись справочные издания о населении и природе Тамбовской области. Были обследованы: сводный систематический краеведческий каталог Тамбовской областной универсальной научной библиотеки имени А.С. Пушкина, каталоги и картотеки научной библиотеки Тамбовского государственного университета имени Г.Р. Державина, Государственного архива Тамбовской области, научной библиотеки Тамбовского областного краеведческого музея, каталоги Российской государственной библиотеки, Государственной публичной научно-технической библиотеки, Государственной исторической библиотеки, Российской национальной библиотеки. В качестве дополнительных источников поиска использованы государственные библиографические указатели Российской книжной палаты и библиографические издания Института научной информации по общественным наукам. При обследовании библиографических источников рассматривались предметный, географический, именной вспомогательные указатели или разделы изучались по принципу просмотра каждой библиографической записи. В некоторых моментах нами использовался

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

метод «снежного кома». В результате создана рабочая картотека справочных изданий, которые в настоящее время подвергаются вторичному отбору по качественным параметрам с использованием и формальных признаков.

Списки электронных копий отдельных справочных изданий систематизируются в соответствии с ББК и УДК, что обеспечит их удобный поиск в любых целях: научных, производственных, образовательных, самообразовательных и т.д. Каждое издание сопровождается справочной аннотацией, предметным, географическим и именным указателями. Таким образом, на данном этапе документ оцифровывается, на него создаются полное библиографическое описание, он индексируется, к каждому источнику создаются три вспомогательных указателя – предметный, именной, географический. Территориальные границы области неоднократно менялись. Поэтому с учетом различных временных изменений был составлен географический указатель. В именной указатель решено включать имена собственные, которые есть в библиографическом описании документа, имена выдающихся жителей нашей области, имена из текста издания, если оно представляет исторический интерес. Предметный указатель создается к каждому конкретному изданию на базе выявления основных ключевых слов и словосочетаний. Нами принято решение создать примерную схему из таких ключевых слов и словосочетаний, используя Библиотечно-библиографическую классификацию, предметные вспомогательные указатели к выпускам библиографического указателя Института научной информации по общественным наукам «История, Археология. Этнография» и собственные предметные «ключи» к нескольким изданиям. В результате создан предметный указатель, имеющий двухуровневую, в некоторых случаях трехуровневую глубину классификации.

Дидактический эффект будет достигаться за счет включения в систему различных вузовских и школьных учебных программ регионального компонента образования (географического, исторического, литературного, библиотечного краеведения, регионоведения, региональной экономики и др.), которые станут составной частью системы и будут связаны с электронными текстами справочников. Предполагается введение специальной информации – оригинальных методических указаний по использованию справочных изданий в учебном процессе, что обеспечит дидактическую направленность ресурса.

«Переходным звеном» от содержательной к технологической составляющей проекта станет построение логической структуры гиперсвязей электронных текстов справочников с аннотациями, указателями, учебными программами, методическими указаниями. Важнейшими элементами технологии реализации проекта являются создание программной «оболочки» и дизайна информационной системы, заполнение информационно-справочной системы, которые могут осуществляться по ходу оцифровки справочников и их библиографического описания.

При создании оболочки использовалась программная среда Delphi и СУБД Paradox. БД информационного ресурса состоит из нескольких ключевых таблиц: таблица авторов, таблица ББК, таблица указателей, основной таблицы, где хранится название и идентификационный индекс элемента, по которому осуществляется привязка составляющей информации и др. Для отображения оцифрованного материала используется формат *.pdf.

Последним из этапов является разработка пользовательского интерфейса. Его примерная структура: руководство пользователя, библиографический список оцифрованных источников, указатели. Дополнительно приводится библиографический список дополнительных источников, куда входят документы, содержащие аналогичную информацию или информацию, касающуюся одного населенного пункта. Все источники доступны через библиотечную сеть. Отдельный блок в системе занимают программы и методические рекомендации по работе с ними.

Конкретным итогом работы станет тиражирование информационной системы на CD для библиотек, архивов, музеев, административных учреждений, учебных заведений области. В целом, мы надеемся, что создание информационно-поисковой системы позволит, во-первых, обеспечить широкий доступ к ценным изданиям ученых, сотрудников библиотек, архивов, музеев, преподавателей, студентов, учащихся учебных заведений и т.д.; во-вторых, сохранить основные экземпляры редких изданий, представляющих собой культурное наследие региона; в-третьих, создать качественный сегмент в региональном информационном пространстве.

## Information-retrieval system "Reference editions on population and natute of the Tambov region of the 19-20th centuries" in the educational environment of the university

L.A. Pronina, N.E. Kopytova, N.V. Shatalova, A.N. Evstigneev

The article is devoted to the ussues of development of the tnformation-search system "Reference editions on population and natute of the Tambov region of the 19-20th centuries" and its use in the educational environment of the university and of the region.

♣ Работа выполнена при поддержке РГНФ, проект № 08-01-12107в.

# Электронная библиотека как способ повышения эффективности работы аспирантуры

© Кирилл Теймуразов

ВЦ РАН
Kbt@ccas.ru

## Аннотация

Подготовка квалифицированных научных кадров одна из важнейших задач, стоящих перед государством. Аспекты использований электронной библиотеки в процессе подготовки и защиты диссертаций с целью повышения эффективности работы отдела аспирантуры рассматриваются в статье.

Для развития современного общества необходимо увеличение научного потенциала. В связи с этим возникает вопрос о необходимости подготовки квалифицированных кадров, занимающихся формированием новых технологий. Возможности по раскрытию научного творческого потенциала создают аспирантуры, которые являются важнейшей формой планомерной подготовки высококвалифицированных кадров.

Результатом работы аспирантуры является количество подготовленных и защищенных квалификационных работ (диссертаций). Поэтому эффективность работы аспирантуры можно оценивать процентном отношении числа защитившихся от общего числа поступивших. По разным данным это число составляет от 15 до 25% [1]. Среди причин, по которым это происходит, можно выделить следующие: низкая мотивированность аспирантов и организационная сложность процесса защиты диссертации.

Цифровая библиотека, в первую очередь, является электронным каталогом материалов, определенным образом накопленных и систематизированных. Однако цифровая библиотека не только позволяет каталогизировать данные, но и предоставляет различные сервисы для работы с данными.

Рассмотрим библиотеку диссертаций [2] – каталог диссертаций, содержащий метаданные диссертаций и их полные тексты. Первая возникающая проблема — пополнение данных. И наиболее простой способ решения данной проблемы — предоставить возможность ввода данных непосредственно автору диссертации, стимулировав его дополнительными сервисами.

Анализируя предметную область, можно заметить, что процесс подготовки и защиты диссертации состоит из следующих этапов: изучение материалов по теме — сбор требований; написание первой главы диссертации — анализ требований, регистрация понятий их взаимосвязей; написание второй главы (метод решения и его обоснование) — проектирование; написание третьей главы (применение метода решения) — реализация; написание четвертой главы (оценка результатов, доработка, выводы) — опытная эксплуатация; защита диссертации - ввод в эксплуатацию.

Таким образом, мы наблюдаем аналогию между процессом подготовки диссертации и процессом создания программного обеспечения. Аналогию можно провести и между участниками процесса: аспирант (соискатель) — исполнитель; диссертационный совет — экспертный совет со стороны заказчика; аспирантура — руководство проектом.

Отрасль управления проектами достаточно хорошо изучена, и существуют различные системы управления проектами, способствующие эффективной работе по управлению проектами.

Следовательно, одним из способов уменьшения «организационной составляющей» является внедрение системы управления проектами, и рассмотрение самого процесса подготовки защиты диссертации в качестве реализуемого проекта.

Системы управления проектами позволяют провести планирование хода выполнения проекта и контролировать (как самостоятельно, так и внешними пользователями) ход его исполнения.

Таким образом, основной задачей является создание многофункциональной среды за счет расширения электронной библиотеки диссертаций сервисами, открывающими возможности систем управления проектами и электронным документооборотом.

Создаваемая среда должна обеспечивать поддержку деятельности аспиранта (соискателя) (исполнителя) средства автоматизации «СУП», предоставляя доступ к информационному каталогу с нормативными материалами, а также с

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

материалами, содержащими знания о процессе, которые накоплены предыдущими поколениями соискателей, опыт применения (базы знаний/результатов/методов) и обеспечивая ввод данных о ходе процесса защиты: календарное планирование, взаимосвязь этапов процесса, получение информационных сообщений и напоминаний со стороны остальных участников процесса, контроль исполнения, управление конфигурациями-изменениями, «бактрекинг», автоматизированную подготовку документов, подготовку артефактов проекта.

Информационное наполнение среды осуществляется ученым секретарем диссертационного совета (руководитель экспертов), который также получает возможность создавать и модифицировать «заседания совета» (реализации бизнес-процесса «приемки»), планировать их — место и время, повестка дня, материалы к заседанию, результаты заседания, отправлять информационные сообщения в ручном и автоматическом режиме другим пользователям системы, вводить данные о ходе процесса защиты, создавать документы, получать статистические отчеты.

Взаимодействие с системой можно разбить на два основных этапа.

На любом из этапов пользователи системы имеют доступ к электронной библиотеке, включающей в себя и актуальную нормативно-правовую документацию, и открытые результаты работы других пользователей системы, и шаблоны различных артефактов проекта. С другой стороны, аспирант, использующий систему при подготовке диссертации, пополняет библиотеку своими материалами.

На первом этапе (этапе написания диссертации) происходит ввод в систему метаданных о диссертации и ходе процесса ее подготовки, со стороны аспирантуры осуществляется контроль над ходом выполнения проекта, а секретарь совета имеет возможность спланировать заседания совета в зависимости от хода процесса подготовки защит.

С момента принятия решения о том, что диссертация готова к защите, начинается второй этап — этап подготовки к защите диссертации, защита диссертации и действия после защиты диссертации. Действия этого этапа достаточно сильно регламентированы. С одной стороны, это является преградой на пути к успешной защите, а с другой стороны позволяет, воспользовавшись возможностями системы, автоматизировать большинство действий: взаимодействие с диссертационным советом, получение необходимых документов, размещение автореферата на сайте совета, планирование и подготовка заседания совета, взаимодействие с ведущей организацией, оппонентами. Более того, система предоставляет возможность получения отзывов на опубликованные авторефераты со стороны третьих пользователей системы.

С 2003 года идут разработки единого научного информационного пространства (ЕНИП) РАН [3]. В рамках этих разработок создана концепция, взяв которую за основу, можно будет получить набор решений, во многом перекрывающих наших потребности, и сосредоточится на специфике нашей задачи.

В рамках ЕНИП уже разработана модель данных и интерфейсы для поддержки цифровых библиотек и библиотеки диссертаций. Система ЕНИП также поддерживает работу с «рабочими потоками».

Разработанная среда станет в свою очередь имитационной моделью, которая позволит оценить имеющуюся эффективность и, по результатам эксплуатации, выявить узкие места, требующие расширения для дальнейшего улучшения эффективности работы аспирантуры.

## Литература


[1] Василий Ушаков «Об эффективности работы аспирантуры и докторантуры», http://za-kadry.tpu.ru/article/2975/1625.htm

[2] К. Б. Теймуразов «Подсистема «Электронная библиотека диссертаций» в рамках Информационной системы «Научный институт РАН» // Труды XLVIII научной коференции МФТИ часть VII.

[3] А. А. Бездушный, А. Н. Бездушный, А. К. Нестеренко, В. А. Серебряков, Т. М. Сысоев, К. Б. Теймуразов, В. И. Филиппов «Информационная Web-система «Научный институт» на платформе ЕНИП, ВЦ РАН, Москва, 2007


## Digital library as a tool of increasing efficiency of a postgraduate course


Kirill Teymurazov



Training of skilled scientific personnel is one of the important problems of our state. This article concerns some aspects of digital library using in the process of preparing and defense of a thesis to increase efficiency of postgraduate courses.

# АЛФАВИТНЫЙ УКАЗАТЕЛЬ АВТОРОВ

# AUTHOR INDEX

Научное издание

**ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:**
**ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,**
**ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ**

**XI Всероссийская научная конференция RCDL'2009**

*Труды конференции*

**DIGITAL LIBRARIES: ADVANCED METHODS AND TECHNOLOGIES,**
**DIGITAL COLLECTIONS**

**XI All-Russian Research Conference RCDL'2009**

*Proceedings of the Conference*

Составители Н.Б. Луговая, Н.Н. Никитина.

Сборник отпечатан методом прямого репродуцирования
с оригиналов, предоставленных оргкомитетом.

D10,11-2008-124

Сдано в печать **.**.2009.
Формат **×**/. Бумага офсетная. Печать офсетная.
Усл. печ. л.**. Уч.-изд. Л**.. Тираж 180. Заказ №.

Карельский научный центр РАН
Редакционно-издательский отдел
185003, Петрозаводск, пр. А. Невского, 50