

Оглавление

Д.Ю. Ахметов, А.М. Елизаров, Е.К. Липачёв

**СЕРВИС-ОРИЕНТИРОВАННАЯ ИНФОРМАЦИОННАЯ СИСТЕМА
НАУЧНОГО ЖУРНАЛА «ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ»**

А.М. Гусенков

**ИНТЕЛЛЕКТУАЛЬНЫЙ ПОИСК СЛОЖНЫХ ОБЪЕКТОВ В МАССИВАХ
БОЛЬШИХ ДАННЫХ**

УДК 004.82

СЕРВИС-ОРИЕНТИРОВАННАЯ ИНФОРМАЦИОННАЯ СИСТЕМА НАУЧНОГО ЖУРНАЛА «ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ»

Д.Ю. Ахметов¹, А.М. Елизаров², Е.К. Липачёв³

*Институт математики и механики им. Н.И. Лобачевского
Казанского (Приволжского) федерального университета*

¹ akhmetov.dy@gmail.com, ² amelizarov@gmail.com, ³ elipachev@gmail.com

Аннотация

Описаны современные модели публикации и распространения научных знаний, а также различные подходы к организации персонального информационного пространства ученого. Охарактеризованы современные информационные системы управления научными изданиями и сервисы, определяющие их функциональность. На основе технологии расширения функционала открытой системы Open Journal Systems созданы программные модули, обеспечивающие автоматизацию ряда редакционных процессов электронного научного журнала. Представлена архитектура универсальной платформы управления электронными научными журналами и подробно описана платформа управления электронным научным журналом «Электронные библиотеки».

Ключевые слова: издательские системы, современные модели публикации и распространения научных знаний, информационное общество, персональное информационное пространство ученого, электронный научный журнал, сервис-ориентированная информационная система, информационные системы управления научными изданиями и публикациями, интеграция электронных ресурсов, извлечение метаданных, инфраструктура научного журнала «Электронные библиотеки», платформа Open Journal Systems.

ВВЕДЕНИЕ

Как известно (см., например, [1]), в настоящее время во всем мире происходит активное формирование информационного общества (ИО) (общества, основанного на информации и знаниях) как новой ступени развития человеческой цивилизации. Как отмечено в 2003 году в Женевской Декларации принципов «Построение информационного общества – глобальная задача в новом тысячелетии» (см., например, http://www.itu.int/net/wsis/outcome/booklet/declaration_Aru.html), всеми странами проявлено общее стремление построить общество, ориентированное на интересы людей, открытое для всех и направленное на развитие. Открытость для всех предполагает, что в ИО каждый сможет создавать информацию и знания, иметь к ним доступ, пользоваться и обмениваться ими, чтобы дать отдельным лицам, сообществам и народам возможность в полной мере реализовать свой потенциал для собственного устойчивого развития и повышения качества жизни.

Характерной чертой формируемого информационного общества является повсеместное использование информационно-коммуникационных технологий (ИКТ): происходит приобщение людей не только к ИКТ, но и к новым их приложениям и «электронным» услугам (электронному правительству, электронной библиотеке, интернет-магазину, умному дому и многому другому); расширяется активное использование ИКТ как в традиционных сферах человеческой деятельности (включая образование и науку), так и в развитии информационной индустрии (ИКТ-сектора экономики, секторов контента и медиа и др.). Результатом этих процессов должны стать экономия времени, более высокий уровень культуры потребления и лучшее качество жизни. Вместе с тем, информационное общество сегодня характеризуется существенным цифровым неравенством: преимущества революции в области ИКТ неравномерно распределены между развитыми и развивающимися странами, а также внутри стран между различными социальными группами. Поэтому необходимо превратить разрыв в цифровых технологиях в цифровые возможности для всех.

Интернет является ключевой инфраструктурой информационного общества, а основополагающая роль в технологическом переоснащении глобальной

сети принадлежит технологиям Семантического Веба, разработку которых координирует консорциум W3C (<http://www.w3c.org>). В процессе построения ИО формируется сетевое информационное пространство, составной частью которого является научное информационное пространство. Последнее характеризуется использованием специализированных технологий, связанных, в частности, с особенностями жизненного цикла электронного документа и научной публикации.

Как известно, существенным импульсом в процессах развития ИКТ и создания информационного общества стала инициатива администрации президента США Б. Клинтона по развитию интернета (The National Information Infrastructure: Agenda for Action. 15 September 1993. Report, Department of Commerce, USA; <https://www.ibiblio.org/nii/NII-Agenda-for-Action.html>; Next Generation Internet Initiative, <http://web.archive.org/web/19980209111241/>; <http://www.ccic.gov/ngi/concept-Jul97/>). В [2] выделены следующие пять важных результатов названной инициативы:

- заложены юридические основы свободного распространения информации в мультимедийных средах, что привело к ускоренному обмену результатами интеллектуальной деятельности;
- созданы предпосылки для финансирования разработки свободно распространяемого программного обеспечения и, как следствие, построения на этой основе эффективных информационных систем;
 - создан новый сектор экономики по продаже услуг через интернет;
 - появились качественно новые услуги в сфере научной деятельности;
 - создана ИКТ-инфраструктура для науки и образования с целью проведения на новом уровне научных исследований и разработки новых методов, подходов и программно-аппаратных устройств с дальнейшим внедрением в различные сферы экономики и общественной деятельности.

В [2] подробно проанализирован, прежде всего, пятый из названных результатов и выделены следующие черты, характерные для современной научной деятельности:

- глобализация исследовательского процесса, в который вовлекаются ученые разных стран и исследовательских структур (институтов, лабораторий и др.);

- необходимость обучения и воспроизводства специалистов для быстро возникающих новых областей знаний;
- высокая конкуренция исследований, особенно по хорошо финансируемым тематикам, в том числе, в области создания новых ИКТ;
- жесткие временные рамки проведения исследований вследствие высокой конкуренции исследовательских групп, высокие требования к самим группам;
- большой объем данных и информации, на которых базируются поисковые исследования.

Эти черты в полной мере характеризуют глобальность современной научной деятельности и, как следствие, глобальность современного научного сообщества, заинтересованного в максимально быстром распространении полученного нового знания среди своих членов. Последнее невозможно без применения современных ИКТ, в частности, без развития инфраструктуры, построенной на новых информационных технологиях. Кроме того, современная высокая конкуренция в сфере исследований по актуальным научным направлениям требует быстрого распространения научного знания как с целью популяризации результатов, полученных различными группами исследователей, так и для максимально оперативного доступа к новым научным результатам других исследователей. В современном мире только ИКТ способны обеспечить выполнение этих условий. Поэтому в настоящее время даже традиционные задачи, связанные с подготовкой, последующими процессами публикации статьи в научном журнале или материалах научной конференции и информирования заинтересованных читателей, решаются на новых организационном и информационном уровнях и базируются на широком использовании интернета и ИКТ. Кроме того, существенно изменилась инфраструктура современных научных изданий – речь уже идет не столько о формах и средствах использования ИКТ, сколько о создании программной платформы, реализующей развитую систему сервисов для работы с электронным контентом: без информационных сервисов и специализированных программ становится невозможным подготовить научную работу к публикации, а без сетевых коммуникаций – опубликовать ее. При этом стираются различия между электронным и печатным изданиями, являющимися лишь различными формами представления соответ-

ствующей информации, и можно говорить об электронных научных изданиях, отдельно не выделяя печатные их формы, а электронные научные публикации можно рассматривать как разновидность электронных документов, особенности которых отражаются в их жизненном цикле. Отметим, что базовые понятия «электронный документ» и «жизненный цикл электронного документа», свойства электронных научных документов и их преимущества перед традиционными бумажными носителями обсуждены в [3]. Важно подчеркнуть, что жизненный цикл электронной научной публикации, в отличие от «бумажной», предполагает дальнейшее улучшение её качества и расширение свойств. Отметим в этой связи новую форму научных публикаций, связанную с подготовкой «живых» документов (см., например, [4, 5]). Как отмечено в [6], онлайн-представление становится центральным звеном публикационной деятельности ученого, и с него начинается жизненный цикл публикации.

Все, сказанное выше, полностью подтверждает заключение о том, что современный электронный научный журнал должен базироваться на программной платформе, поддерживающей жизненный цикл научной публикации и реализующей развитую систему сервисов для работы с электронным контентом.

В настоящей работе обсуждаются современные модели публикации и распространения научных знаний, а также различные подходы к организации персонального информационного пространства ученого; характеризуются современные информационные системы управления научными изданиями и публикациями и сервисы, определяющие их функциональность. Представлена архитектура универсальной платформы управления электронными научными журналами и подробно описана платформа управления электронным научным журналом «Электронные библиотеки».

СОВРЕМЕННЫЕ МОДЕЛИ ПУБЛИКАЦИИ И РАСПРОСТРАНЕНИЯ НАУЧНЫХ ЗНАНИЙ

Сегодня электронные издания являются основным источником достоверной научной информации, а подавляющее большинство современных печатных изданий имеет электронные варианты, доступные по подписке или в свободном доступе.

В настоящее время научные электронные издания и ресурсы растворены в потоке электронной информации, объем которого лавинообразно растет, а использование стандартных сервисов и поисковых средств интернета применительно к научной электронной информации становится все менее эффективным. Кроме того, стандартная технология полнотекстового поиска не содержит инструментов для распознавания семантических отношений между объектами знания. Реализовать сервисы с расширенной поисковой функциональностью позволяют технологии Семантического Веба, нацеленные, в частности, на разработку средств создания и поддержки больших интегрированных баз знаний на основе унифицированных семантических моделей.

В научном информационном пространстве все более актуальным становится применение новых ИКТ для организации совместной работы распределенных и виртуальных научных коллективов. Результатами такого применения должны стать лучшая синхронизация деятельности отдельных ученых, новые формы объединения исследователей в эффективно действующие группы и лучшие координация деятельности таких групп и управление ими. Сказанное характеризует роль ИКТ в развитии процесса глобализации научных исследований (см. также [2]).

В новых условиях и издатели, и библиотеки стараются адаптироваться к быстро развивающемуся цифровому миру. Сегодня организационные и экономические модели информационной деятельности, апробированные ранее и используемые до сих пор, устарели и уже не в состоянии в рамках действующих новых моделей подписки сохранить баланс между снижением объема имеющихся средств на приобретение информационных ресурсов и растущей стоимостью последних. При этом профессиональные научные сообщества, в том числе математическое, чрезвычайно заинтересованы во всестороннем обсуждении будущей роли научных журналов в соответствующих профессиях (см., например, [7]).

Как отмечено П. Олвером в [7, 8], традиционная модель издания и распространения научных журналов опирается на четыре основных процесса, признанных профессиональным сообществом ученых, – enhancement, dissemination, archiving, validation, что с успехом позволяло в прошлом привлекать к работе на безвозмездной основе (а иногда и оплачивать часть расходов) исследователей и

ученых в качестве авторов, рецензентов и редакторов. При этом библиотеки и читатели всегда оплачивали доступ к опубликованным материалам.

Под *enhancement* П. Олвер понимает процесс превращения рукописи в профессиональный текст, готовый к тиражированию. В области математики в связи с широким распространением издательской системы TEX основной акцент в этом процессе сместился к автору: большинство физико-математических журналов ожидает от авторов исходные LATEX-файлы их материалов, адаптированные к конкретным особенностям и правилам подготовки рукописей в этих журналах. Процесс такой адаптации проходит практически без участия редакций. В результате издатели сумели расширить модель свободного труда, передав авторам большую часть своих традиционных функций, связанных с *enhancement*.

Dissemination – одна из важнейших функций журналов, предполагающая распространение результатов исследований в более широких сообществах. В настоящее время часто сначала появляются препринты на серверах различных архивов, высших учебных заведений, научных организаций или на личных сайтах авторов, свободно доступные любому читателю с достаточно быстрым и бесцензурным доступом через интернет. При этом во многих областях документы появляются в журналах (в электронном или печатном виде) уже после того, как их содержимое стало доступно заинтересованному сообществу. По этой причине многие издатели позволяют авторам до публикации в журнале размещать в интернете препринтные версии своих работ.

Archiving – функция архивации имеющегося научного материала – характеризует собой иную задачу, так как многие коммерческие издатели стремятся извлечь выгоду, ограничивая абонентам доступ к своим архивам и предлагая лишь некоторый уровень краткосрочной архивной поддержки опубликованных материалов. В результате может быть потерян платный или бесплатный доступ к ранее опубликованным материалам, а быстрое развитие новых технологий, а также устаревание оборудования, программного обеспечения, форматов представления файлов привели к еще нерешенным проблемам с долгосрочной электронной сохранностью больших объемов материала и предотвращением потери данных.

Под *validation* П. Олвер понимает процесс подтверждения рецензируемым журналом правильности и оригинальности содержания публикуемого документа,

что, в свою очередь, помогает автору этого документа подтвердить свой статус при приеме на работу, продвижении по службе, получении грантового финансирования и в других ситуациях, когда существенно учитываются научные достижения претендента.

Таким образом, из перечисленных четырех процессов, лежащих в основе традиционной модели системы издания и распространения научного журнала, в значительной степени последний должен быть сохранен в модели журнала, соответствующей новым экономическим и информационным условиям.

В информационно-издательской деятельности использование ИКТ позволило передовым современным издательствам не только наладить опережающий выпуск электронных версий научных изданий, но и предоставить авторам, читателям, редакционным коллегиям и редакциям новые сервисы для работы с информацией (см. [9]). Например, практически все современные информационные системы управления научной информацией предоставляют сервисы получения наукометрических данных, на основе которых проводится анализ публикационной активности сотрудников научных учреждений и университетов и определяются наиболее перспективные направления развития научных исследований.

В настоящее время одним из самых дискуссионных является вопрос ранжирования научных журналов. Традиционно престиж журнала характеризуется ценностью информации, публикуемой в нем. Такие же показатели, как различные варианты импакт-факторов журналов, могут управляться в результате целенаправленной деятельности издателей и редакторов, имеющих возможности для искусственного увеличения цитируемости материалов, публикуемых в журнале, и, тем самым, искусственного манипулирования импакт-факторами и повышения рейтингов. Установлено, что профессиональные исследователи, как правило, имеют достаточно состоятельные оценки рейтингов научных журналах в своей предметной области, в то время как непрофессионалы в этой области легко могут быть введены в заблуждение, не говоря уже о чиновниках, которые часто совсем не ориентируются в достаточно узких предметных областях, но бывают уполномочены принимать административные решения, приводящие к непредсказуемым результатам. Кроме того, ранжирование журналов не может однозначно определять ранжирование отдельных материалов, опубликованных в них. Вместе с тем,

репутация журнала и его редколлегии, а также строгость процесса рецензирования и отбора материалов не могут не отражаться на качестве и рейтинге публикаций, выпускаемых журналом. Подтверждением сказанному служит документ [10], одобренный в 2010 году Генеральной Ассамблеей Международного математического союза, в котором описана практика управления научными математическими журналами путем строгого соблюдения основополагающих принципов прозрачности, честности и профессионализма, а также подробно охарактеризованы права и обязанности авторов, рецензентов, редакторов и издателей. По мнению П. Олвера [7], сегодня задача состоит в том, чтобы разработать прочную экономическую модель для оптимальной реализации этих основополагающих принципов, и настало время для радикального переосмысления традиционной академической модели для научной коммуникации, в том числе, в области математики.

ПЕРСОНАЛЬНОЕ ИНФОРМАЦИОННОЕ ПРОСТРАНСТВО УЧЕНОГО: ПОДХОД, ОСНОВАННЫЙ НА ИСПОЛЬЗОВАНИИ ОНТОЛОГИЙ

В настоящее время для научного сообщества одной из важных проблем, связанных с информатизацией исследовательской деятельности, является организация персонального информационного пространства ученого (см., например, [11, 12]), а наиболее эффективный подход к решению проблем управления научной информацией обеспечивают семантические технологии: в результате автоматизации процессов обработки семантических связей и формируется персональное информационное пространство ученого.

Одна из центральных идей Семантического Веба базируется на понятии «связанные данные» (Linked Data) [13], в основе которого лежит признание того, что первичными объектами Веба являются описания сущностей с явным указанием их семантики и семантики ссылок (отношений) между ними. Семантические связи между данными повышают их ценность и предоставляют дополнительное удобство в их использовании. В результате данные, интегрированные в единую Семантическую паутину, представляют собой пространство знаний о некоторой предметной области. Технологически этот подход обеспечивается представлением данных в виде триплетов «субъект – предикат – объект» формата RDF

(Resource Description Framework), идентификацией данных с помощью универсальных идентификаторов ресурсов URI, механизмом доступа по протоколу HTTP и спецификацией контролируемых словарей на различных языках (в частности, RDFS и OWL).

Одним из первых проектов, непосредственно реализуемых с 2007 года в рамках названного подхода, является Linking Open Data (LOD) (см., например, <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>). Этот проект выявил важные преимущества представления в виде единого связанного «облака» гетерогенных данных, полученных от различных контент-провайдеров. Сегодня объем данных в LOD составляет более 40 млрд. триплетов. Научные коллекции представлены в LOD в виде неофициальных наборов данных, не поддерживаемых на постоянной основе. Содержимым этих наборов являются стандартные метаданные статей (название, год публикации, информация об авторах и др.), а главное преимущество заключается в стандартизованном подходе к структурированию и хранению интегрированных данных, которые, как правило, загружаются и представляются в RDF-формате из таких традиционных хранилищ, как реляционные базы данных или (реже) веб-страницы и неструктурированные текстовые документы.

Подход, развиваемый нами, основан на использовании онтологий естественнонаучных областей, прежде всего, онтологии математических знаний OntoMathPRO [14–16]. При формировании этой онтологии использовались методы онтологического моделирования с применением алгоритмов автоматического выделения понятий и связей [17]. Существенное значение для внедрения подобной системы имеет полнота «покрытия» онтологией терминологического массива электронной коллекции. Онтологические модели предметных областей также используются нами для решения задачи извлечения знаний в терминах онтологий и как технологическая основа построения рекомендательного сервиса, позволяющего выполнить персонализированный отбор научных документов в соответствии с семантическим профилем учёного (см. [18]). Кроме того, создается система аннотирования, основанная на онтологиях предметных областей: в исходный текст научной работы система автоматически добавляет ссылки на опре-

деления терминов – результатом является документ, содержащий помимо исходного текста научной работы блоки аннотаций, доступные через интерфейс пользователя. Отметим в связи с этим платформу ScienceWISE (<http://sciencewise.info>), с помощью которой пользователи получают научную информацию, упорядоченную в соответствии с их личными интересами. При этом научные статьи ранжируются в соответствии с интересами пользователей, а на основе онтологий создается система закладок и аннотаций (см. [19]).

ИНФОРМАЦИОННЫЕ СИСТЕМЫ УПРАВЛЕНИЯ НАУЧНЫМИ ИЗДАНИЯМИ И ПУБЛИКАЦИЯМИ

Большинство современных научных изданий представлено в интернете; периодические научные издания имеют сайты с автономной системой управления либо являются частью некоторой объемлющей информационной системы (например, университета или издательства). Как правило, такие автономные системы ограничены функционально и не обеспечивают автоматизации всех бизнес-процессов научного журнала.

Практически все ведущие научные издательства внедряют системы автоматического управления рабочими процессами, в числе которых – наиболее сложные и длительные по времени редакционные процессы, включая организацию независимого рецензирования. В настоящее время программные платформы управления электронными научными журналами обеспечивают компьютерную поддержку традиционных процессов издания журналов в соответствии со сложившимися стандартными алгоритмами их реализации. Одновременно ведется работа в направлении автоматизации рабочих процессов, реализуемых редколлегией научных журналов, в частности, на основе технологий Cloud Computing. К настоящему времени разработано достаточно много информационных систем автоматизации деятельности издательств, которые можно условно разделить на системы с открытым и закрытым исходными кодами.

К электронным ресурсам, формируемым в рамках научных изданий, должен быть обеспечен эффективный доступ. Одной из распространенных технологий обеспечения такого доступа, а также повышения степени релевантности поисковых запросов, предоставления сервисов, связанных с размещением и получением информации, являются специализированные порталы, примерами которых

служат eLibrary.ru и Math-Net.ru. Отметим, что применяемые здесь технологии базируются на развитых и широко распространенных веб-стандартах [20, 21], используемых, в частности, для предоставления информационных услуг в системах электронного правительства.

В [22] проведен сравнительный анализ технологий электронных библиотек (ЭБ), позволивший выделить основной набор сервисов, реализуемых в научных электронных коллекциях и электронных журналах, и определить ключевые методы интеграции электронных коллекций в научное информационное пространство. Рассмотрены также особенности использования для управления электронными научными коллекциями технологий, примененных при реализации государственных информационных систем, в частности, сервисов электронного правительства. В качестве положительных примеров описаны сервисы, функционирующие на основе технологий Семантического Веба и внедренные в практику работы ряда электронных научных изданий.

Несмотря на использование ИКТ практически на всех этапах подготовки и рассмотрения научных работ в редакциях научных журналов, организация рабочих процессов самих редакций пока по-прежнему ориентирована на бумажную форму как завершающую. Причины этого – в сложившейся системе рассмотрения материалов, поступающих в редакции журналов, которые в первую очередь регламентируют правила их оформления и рецензирования. Одновременно остро стоит проблема оперативного доступа к результатам научных исследований, поэтому для обработки постоянно растущего объема информации стали создаваться и внедряться информационные системы управления научными изданиями и публикациями – сегодня в мире внедрены десятки таких систем, построенных на различных программных платформах и позволяющих управлять редакционно-издательским процессом (см., например, [9, 23, 24]). Отметим, что к настоящему времени сформировано несколько подходов к построению таких систем, причем часть из них разработана научными сообществами. В частности, создана семантическая сеть систем, где на одной платформе могут быть объединены в единое информационное пространство несколько журналов (см. [25]).

Программную поддержку процессов подготовки научной статьи в соответствии с редакционными требованиями журнала, а также взаимодействие с редакцией и сбор наукометрических данных целесообразно обеспечить в рамках единой программной платформы. Например, полный цикл управления редакционно-издательскими процессами поддерживается платформой Open Journal Systems (OJS) (<https://pkp.sfu.ca/ojs/>).

Основные информационные системы управления научными журналами и публикациями были созданы в 2004 – 2008 гг. Наибольший практический интерес вызывают те из них, которые являются свободно распространяемыми («open source»): благодаря открытому коду появляется возможность доработки таких систем и придания им требуемой функциональности. Существенное обстоятельство – наличие у многих систем групп разработчиков, выкладывающих на соответствующие сайты новые модули, часто выполненные инновационными методами с применением передовых ИКТ.

Важным компонентом современных информационных систем управления научными журналами являются сервисы, регулирующие процесс рецензирования и обеспечивающие коллективное редактирование электронных документов. Системы такого типа должны предоставлять такие редакционные сервисы, как классификация, аннотирование, выделение метаданных, публикация, долгосрочное хранение, конвертирование, распространение, синдикация, статистика использования, харвестинг, объединение в коллекцию, взаимодействие с институциональными репозиториями, контроль доступа, подписка, рассылка уведомлений, новые поступления. Помимо удаленного представления статей в научный журнал и их дальнейшей обработки для окончательной публикации системы управления научными журналами и публикациями обеспечивают доступ к сформированному контенту и расширенный поиск (по автору, названию статьи, ключевым словам и др.) в соответствующих электронных коллекциях, т. е. в полном объеме реализуют функциональные возможности, присущие электронным библиотекам. С этой точки зрения электронный научный журнал можно рассматривать как научную ЭБ, оперирующую статьями журнала как информационными объектами. Следовательно, при создании информационных систем управления электронными научными публикациями могут быть использованы хорошо развитые технологии

ЭБ, а при анализе существующих систем такого типа – подходы, разработанные при формировании концептуальных моделей, обобщающих накопленный опыт в сфере создания и использования ЭБ, в частности, эталонной модели ЭБ (Digital Library Reference Model, DLRM) [26], построенной в рамках европейского проекта DELOS (<http://www.delos.info>).

Сравнительный анализ существующих информационных систем управления электронными научными журналами, выполненный с использованием критериев оценки, заложенных в модели DELOS DLRM, проведен в [9, 27]. При анализе использовались результаты работ [28–31], а сравнение проведено по набору таких параметров, как базовое программное обеспечение, количество успешных инсталляций, полнота сопровождающей технической документации.

В соответствии с моделью DELOS DLRM электронный научный журнал – это хранилище статей с набором программного обеспечения, реализующего функции хранения, сбора и предоставления доступа к информационным объектам, а система управления научными журналами является разновидностью системы управления ЭБ, также имеет ролевую модель пользователей и использует метаданные при формировании выпусков журнала и описании статей. Особенность состоит в том, что в системе управления электронным журналом должен быть предусмотрен более сложный процесс публикации информационных объектов, отражающий фактическую работу любого издательства, публикующего научные журналы.

Важным аспектом организации системы управления электронными научными журналами является формирование эффективной поисковой системы. Эффективность поиска в значительной степени связана с используемыми системами ключевых слов и метаданных. Как известно, метаданные являются особым видом информационных ресурсов и содержат обобщенную информацию о структуре и содержании информационного источника, в том числе, данные об авторах публикации, ее названии и дате публикации, кодах предметной области и т. д. (см. [32]). В электронных научных журналах комплексное использование метаданных, записанных по правилам Dublin Core (DC) и RDF, позволяет существенно повысить эффективность поиска (см., например, [33, 34]). Поэтому выделение метаданных

можно использовать как технологию, позволяющую включать научные издания в систему глобального поиска.

Далее, для научных журналов комплекс мероприятий по использованию методов поисковой оптимизации на основе Search Engine Optimization (см. [35]) необходимо дополнить процедурами, учитывающими специфику предметной области. Например, для математических журналов важной является задача поиска по фрагментам формул. Система поиска по формулам реализована нами в электронном хранилище журнала Lobachevskii Journal of Mathematics. В основе применяемого алгоритма поиска лежит технология преобразования математических документов в формат XML и формульных конструкций – в MathML-нотацию (см. [34, 36]).

Проведенный нами анализ показал, что для поддержки жизненного цикла как отдельных научных статей, так и научных журналов в целом целесообразно использовать в качестве ядра системы управления электронными научными журналами программную платформу Open Journal Systems (см. <https://pkp.sfu.ca/ojs/ojs-usage/>). Ее внедрение позволяет, прежде всего, автоматизировать наиболее трудоемкие рабочие процессы, а порталное решение дает возможность интегрировать журналы в мировое информационное научное пространство: согласно статистике OJS, в 2015 году система использовалась почти в 10000 активных журналах (публикующих не менее 10 статей), и это количество постоянно росло с момента начала ее внедрения (см. <https://pkp.sfu.ca/ojs/ojs-usage/ojs-map/>).

Вместе с тем, для развития электронного научного журнала необходимо расширение функциональности его базовой информационной системы, что связано как с особенностями предметной области этого журнала, так и со сложившимися традициями работы его редакции и редколлегии, что достигается разработкой специализированных модулей, например, для математических журналов – сервисов конвертации в специализированные форматы (TEX, MathML и др.). Для системы OJS нами разработан соответствующий плагин и предложен метод автоматизации процесса первичной обработки научной статьи, использующей TEX-нотацию, в журнальную информационную систему. Первичная обработка предполагает валидацию электронного документа, включая его анализ на соответствие тре-

бованиям редакции по стилевому оформлению публикации, а также TEX-компиляцию документа. Система позволяет на этапе загрузки статьи исключить возможные отклонения от редакционных правил и сообщить автору об обнаруженных ошибках при подготовке электронной публикации. Автоматизированная обработка электронных публикаций на платформе OJS реализована нами по схеме, разработанной ранее в электронном математическом журнале Lobachevskii Journal of Mathematics (см. [37]).

Большая часть электронных математических коллекций состоит из неструктурированных текстов в различных форматах, как правило, pdf и TEX. Актуальной является разработка методов автоматического извлечения из электронного источника идентифицирующих метаданных (в терминологии работы [32]), ключевых слов и библиографических ссылок (см., например, [38, 39]). Подходы к решению этой проблемы, реализованные нами в электронных коллекциях математических изданий Казанского университета, описаны в [40]. В частности, особые требования к оформлению научных статей, предполагающие использование стилевых правил (обозначенных каждым научным журналом), были использованы как основа определения признаков, по которым выполнены автоматический разбор статей и выделение из их текста основных метаданных (например, названия работы, ее авторов) (см. [40]). Теговая структура документа позволяет в автоматическом режиме однозначно выделить ряд текстовых элементов. Наличие структурных TEX-конструкций, таких, как `\author`, `\title` и др., используемых в журнальных статьях, значительно упростило алгоритм разбора существенной части электронной коллекции.

Следует также отметить, что использование структурных конструкций системы TEX не является общепринятой практикой, и большие объемы электронных документов, в частности, материалы научных конференций, не имеют структурной разметки, а блок библиографии также не всегда выделен структурно. Для этой части электронных документов в алгоритме автоматического разбора использованы особенности шрифтового оформления и порядок размещения текстовых элементов в документе.

Помимо сервисов извлечения метаданных в информационной системе управления электронными научными коллекциями необходимы сервисы контроля доступа к данным и метаданным, поиска в массивах разнородных метаданных, импорта метаданных из других информационных систем. Методы организации сервисов управления метаданными в распределенных информационных системах, интегрирующих разнородную научную информацию, рассмотрены, например, в [41].

Таким образом, для обеспечения хранения и поддержки жизненного цикла как отдельных статей, так и журналов в целом, целесообразно использовать платформу OJS в качестве ядра инфраструктуры электронного научного журнала. Под термином «инфраструктура научного журнала», как правило, подразумевают систему поддержки определенной части жизненного цикла научной публикации (см., например, [42–44]). Такая система может быть реализована в виде сервисов и программных инструментов, связанных, например, с подготовкой публикации в соответствии с требованиями журнала, включая оформление текста, графических, мультимедийных и других необходимых материалов.

Трехуровневая архитектура платформы управления электронными научными журналами, использующая в качестве ядра систему OJS, предложена в [9], ее уровни – физический, базовый и уровень сервисов.

Физический уровень характеризует аппаратную составляющую системы, обеспечивающую функционирование верхних уровней, и содержит системное и прикладное программное обеспечение. Все эти компоненты должны сопровождаться технической поддержкой с использованием технологий виртуализации и облачных вычислений, хотя возможна реализация и без использования виртуальных машин.

Базовый уровень отвечает за предоставление основных сервисов управления электронными научными журналами, обеспечивающих, в частности, регистрацию авторов и пользователей, прием и первоначальную обработку статей, включая автоматическую проверку соблюдения правил редакции и рецензирования. Должен также осуществляться контроль соблюдения сроков рассмотрения статей, назначения рецензентов и рассылки различных уведомлений; должны

предоставляться сервисы удаленного взаимодействия и совместной работы, поиска в электронном хранилище и автоматического извлечения метаданных, структурирования входящей информации. Также необходимы поддержка управления пользователями, разграничения прав доступа и возможность организация платного доступа к контенту. На базовом уровне развернута система OJS, с помощью которой реализуются все бизнес-процессы электронного издательства, а также хранение контента.

На уровне сервисов размещены дополнительные надстройки и функции, учитывающие специфику предметной области научного журнала. Например, для математических журналов востребованы сервисы конвертации в специализированные форматы (TeX, MathML и др.). Здесь реализуется front-end системы и происходит взаимодействие с конечным пользователем. Взаимодействие с системой управления электронными научными журналами может быть организовано либо через собственный веб-портал, либо через специальные программные адаптеры с сайта конкретного журнала, размещающего свой контент в хранилище системы. При первом способе взаимодействия зарегистрированный пользователь получает доступ ко всем журналам, размещенным в системе, а веб-портал служит единой точкой входа. Такой способ наиболее удобен для новых журналов, не имевших собственных сайтов в Сети. Для журналов, уже длительное время поддерживающих собственные сайты, оптимальным, на наш взгляд, является второй способ взаимодействия. В этом случае удастся сохранить привычный адрес сайта журнала и его «историю» в интернете, максимально автоматизировав редакционные процессы.

С 2013 г. в Республике Татарстан на основе разработанной архитектуры создана и развивается система управления электронными научными журналами. На текущий момент времени система реализована технически, создан ее веб-портал (<http://science.tatarstan.ru/services>), и ряд научных журналов переведен под ее управление. Система проходит тестирование с целью ее дальнейшей интеграции в единую научно-образовательную среду. Выделен набор дополнительных модулей, функций и сервисов, который реализован на портале. По окончании тестирования будут сделаны выводы о достаточности сформированного перечня функций, их полезности и применимости.

В заключение настоящего раздела остановимся на технологиях расширения функционала OJS (см. также [16]).

Система OJS имеет модульную архитектуру, что позволяет разрабатывать собственные классы и модули, расширяя ее функциональные возможности. OJS основана на паттерне Model-View-Controller (MVC), как следствие, хранилище данных, пользовательские интерфейсы и управляющие функции разделены на разные уровни взаимодействия. Система OJS платформонезависима и может быть установлена как под ОС Windows, так и на Unix-ориентированных операционных системах. В OJS используются свободно распространяемые язык программирования PHP, веб-сервер Apache, а также СУБД (MySQL, PostgreSQL).

Расширить функционал системы можно с помощью модулей, имеющих специальный формат (см., например, [45]). Большое количество инструментов, расширяющих возможности пользовательского интерфейса системы, а также обеспечивающих поддержку различных форматов данных, содержится в галерее модулей (плагинов) OJS (<http://pkp.sfu.ca/support/forum/viewforum.php?f=28>).

ИНФРАСТРУКТУРА ЖУРНАЛА «ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ»

Журнал «Электронные библиотеки» представлен в нескольких вариантах: на платформе OJS (<http://ojs.kpfu.ru/index.php/elbib>) и в виде веб-портала (<http://elbib.kpfu.ru/>), на котором реализованы разработанные сервисы управления контентом. Кроме того, на портале организовано электронное хранилище статей журнала за 1998–2014 гг. Исходная архивная коллекция (<http://elbib.ru/index.phtml?page=elbib/rus/journal>) представляет собой набор html-файлов и не имеет метаданных. В автоматизированном режиме проведена обработка этой коллекции с целью преобразования статей к формату .pdf, формирования метаданных и организации поиска. Загрузка новых статей происходит в онлайн-режиме через систему OJS с выполнением стилевой валидации [16].

Веб-портал организован в виде набора тематических разделов:

- информация о журнале: история, редакционная коллегия, главный редактор, тематика, статистика журнала;
- редакционная политика: цели и задачи, правовые основы, рекомендации авторам, рецензентам и редакторам;
- информация для авторов: правовые нормы (вопросы соблюдения

правовых и этических норм научных публикаций), правила представления материалов (шаблоны оформления статьи и договора о передаче авторских прав);

- блок загрузка материалов через OJS;
- «Архивы» – содержит три связанные страницы (архив за весь период, архивы по годам и содержание отдельных выпусков);
- информация для читателей: последний выпуск, архивы, авторский указатель, блок ключевых слов; страница «Последний выпуск» формируется при входе на портал, на ней отображается содержание текущего опубликованного номера (см. рис. 1); при выборе гипертекстовой ссылки с названием статьи происходит переход на страницу с аннотацией статьи, списком ключевых слов и ее полным текстом (см. рис. 2);
- «Авторский указатель» – список всех авторов, публиковавшихся в журнале; при нажатии на фамилию автора происходит переход на страницу со списком его опубликованных работ (рис. 3);
- «Ключевые слова» – список ключевых слов, использованных во всех статьях, опубликованных в журнале (см. рис. 4); выбор ключевого слова выводит на страницы тех статей, в которых оно было использовано.

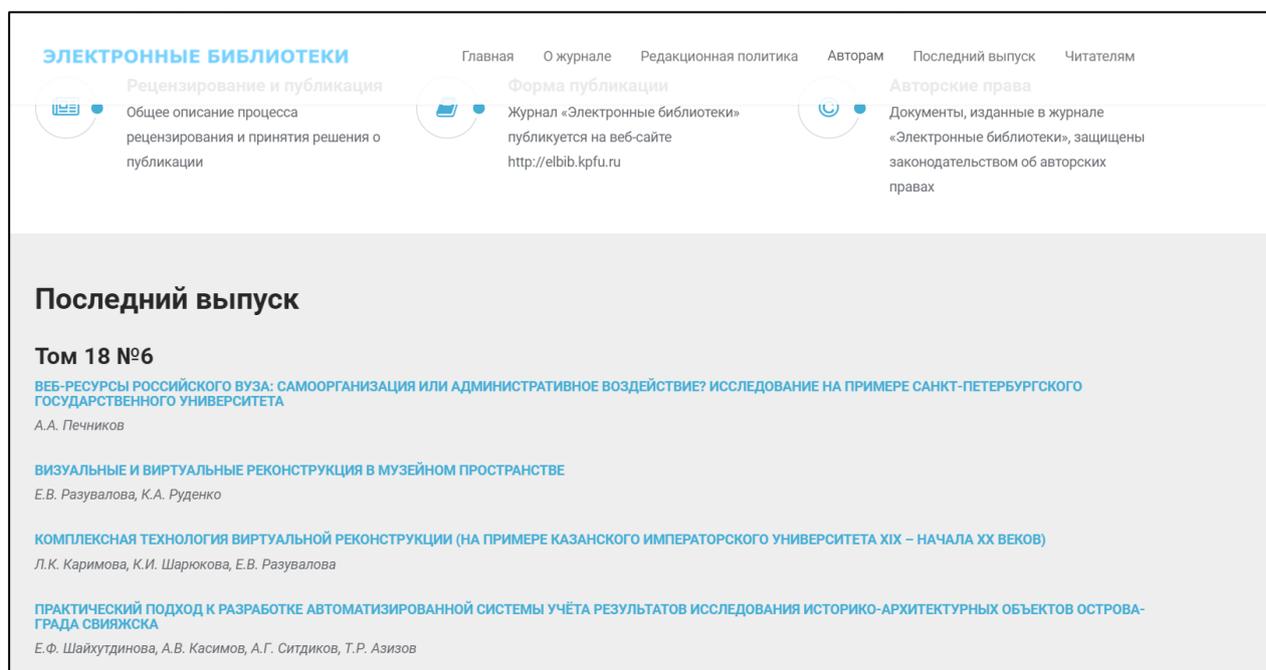


Рис. 1. Страница портала с содержанием текущего номера журнала

Рус Eng

Российский научный электронный журнал
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ

ISSN 1562-5419

Поиск

ВЕБ-РЕСУРСЫ РОССИЙСКОГО ВУЗА: САМООРГАНИЗАЦИЯ ИЛИ АДМИНИСТРАТИВНОЕ ВОЗДЕЙСТВИЕ? ИССЛЕДОВАНИЕ НА ПРИМЕРЕ САНКТ-ПЕТЕРБУРГСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

А.А. Печников

Аннотация:

Веб-пространство организации – это множество веб-сайтов организации, связанных посредством гиперссылок. Основной вопрос, рассматриваемый в статье, заключается в том, является ли такое веб-пространство самоорганизующимся, то есть происходит ли упорядочение его элементов за счёт их внутреннего взаимодействия без внешних воздействий или же внешние (так называемые «административные») воздействия настолько сильны, что их влияние можно обнаружить. В статье предлагается общий подход, демонстрируемый на примере веб-пространства Санкт-Петербургского государственного университета. На основе проведенного анализа на вопрос, являются ли значимыми административные воздействия на веб-пространство университета, в данном случае дается положительный ответ.

Ключевые слова:

гиперссылка, веб-сайт, веб-граф, динамическая модель, административное воздействие

Полный текст:

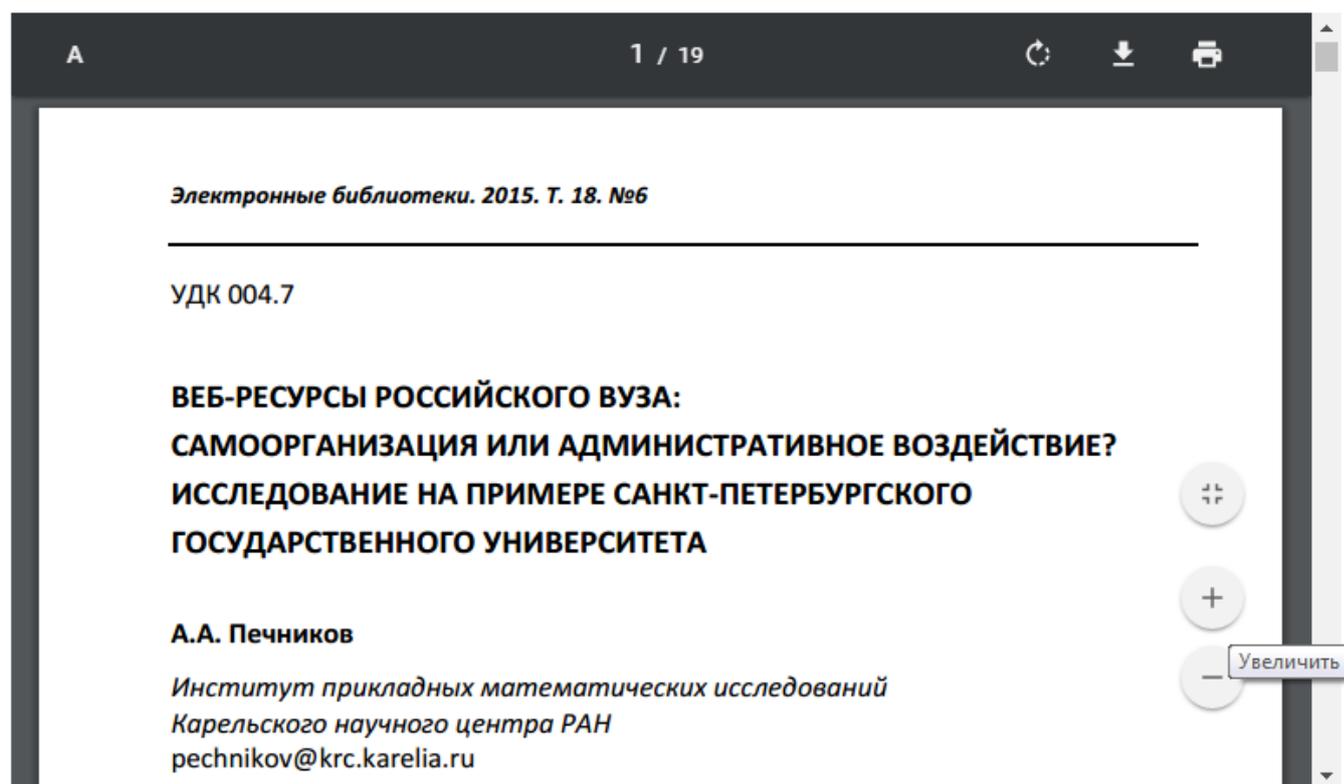


Рис. 2. Страница статьи на портале журнала

ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ

Главная О журнале Редакционная политика Авторам Последний выпуск Читателям

Издается с 1998 года ISSN 1562-5419 English

СПИСОК АВТОРОВ

В журнале опубликованы 573 автор(-а)(-ов)

А Б В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Ь Э Ю Я

А

Аббакумова Н.П.	Абросимов А.Г.	Абызгильдин А.Ю.	Абызгильдина С.Ш.	Аветисов М.А.	Агеев М.С.
Адаскина Ю.В.	Аджиев А.С.	Азизов Т.Р.	Азриель С. Мораг	Акимов С.И.	Алексеев А.Н.
Аленин В.В.	Алимов А.Ф.	Алимова И.С.	Алитова Р.Ф.	Аляутдинов А.Р.	Алëshин Л.И.
Амелин М.Н.	Антоз Юбер	Антопольский А.Б.	Антропова С.Е.	Антофеев С.В.	Апанович З.В.
Арнаутов С.А.	Атаева О.М.	Атаманова О.Ю.	Ахлестин А.Ю.	Ахметов Д.Ю.	

Б

Бабина О.И.	Бабяк П.В.	Багажков А.К.	Байгарова Н.С.	Баранов В.А.	Баранов В.С.
Барахнин В.Б.	Баринова Т.В.	Бартунов О.С.	Барышев Р.А.	Барышева О.В.	Батова Л. Л.
Бахмин А.В.	Бездушный А.А.	Бездушный А.Н.	Бекенев В.А.	Беляев В.О.	Беляков В.А.
Бенедетти Бенедетто	Беннетт Рик	Бирман Дэвид	Бирюков А.Л.	Блинов П.Д.	Богданов А.В.

Рис. 3. Раздел «Авторский указатель»

ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ

Главная О журнале Редакционная политика Авторам Последний выпуск Читателям

Издается с 1998 года ISSN 1562-5419 English

КЛЮЧЕВЫЕ СЛОВА

3D (1)	ASP.NET (1)
ArcGIS for INSPIRE (1)	CERIF (1)
CRF (1)	CSW (1)
ESA (2)	HMA (2)
INSPIRE (1)	JEL (1)
Linked Open Data (1)	OGC (2)
Open Journal System (1)	Open Journal Systems (1)
Open Linked Data (1)	SOA (1)
SSE (1)	SentiRuEval (2)
WPS (1)	Web Processing Service (1)
Z39 (1)	смарт-библиотека (2)
е-Инфраструктура (1)	ebRIM (1)
level design (1)	unity3D (1)
АРБИКОН (1)	Веб-ГИС технологии (1)
Великий Болгар (1)	ГРНТИ (1)
ДЗЗ (1)	Дистанционное зондирование Земли из космоса (1)
ЕСИМО (1)	Европейский Портал Исследований Земли из космоса (1)
Инфраструктура научных информационных ресурсов (1)	Комитет по спутникам для исследования Земли из космоса (1)
Научная библиотека Пермского национального исследовательского политехнического университета (1)	Онтология по естественным наукам и технологиям ОЕНТ (1)
	Программа глобального мониторинга для экологии и безопасности (1)

Рис. 4. Раздел «Ключевые слова»

Архивные коллекции журнала (<http://elbib.ru/index.phtml?page=elbib/rus/journal>) являются набором html-страниц, представляющих содержание номеров журнала и ссылки на статьи. Каждая статья размещена на отдельной странице сайта.

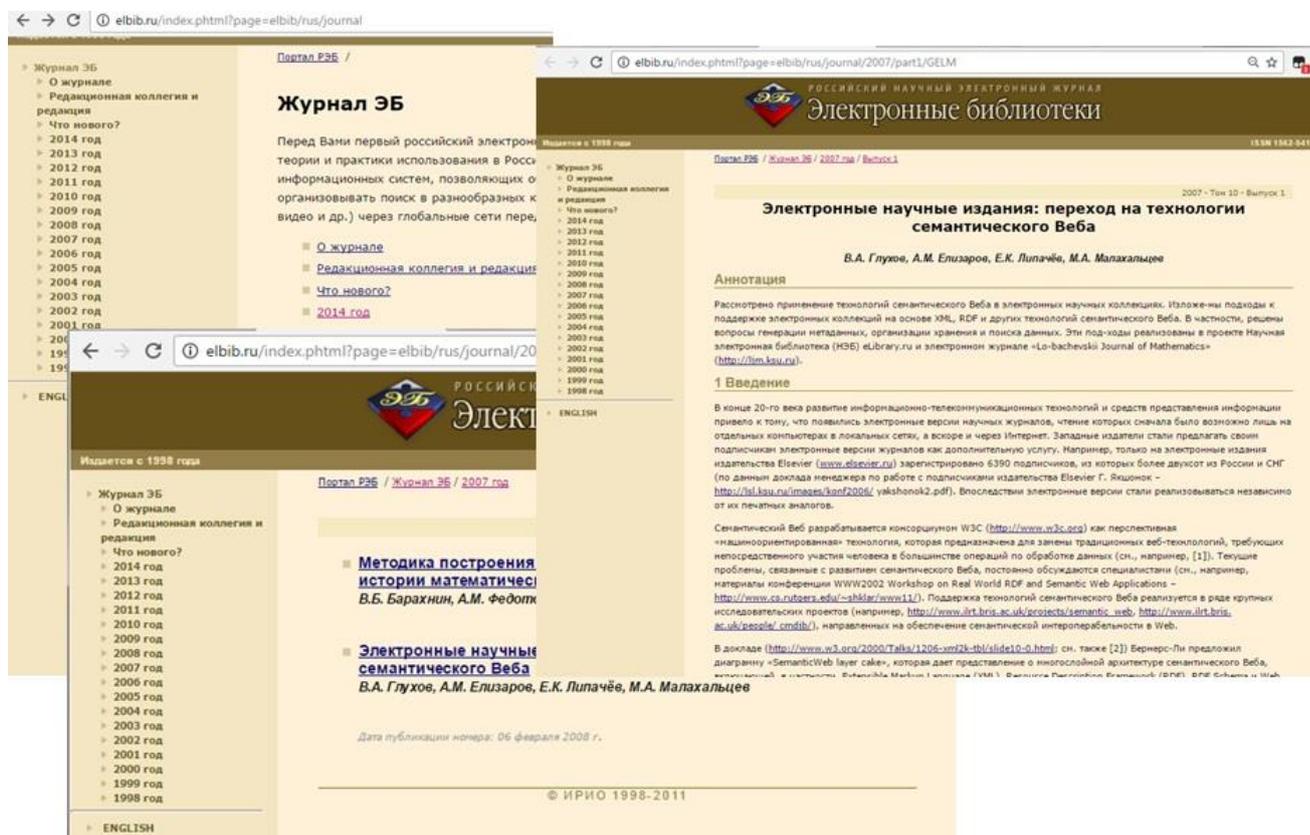


Рис. 5. Страница со статьей из архива elbib.ru

Для управления электронным контентом разработан программный комплекс, состоящий из модулей: выделения и обработки метаданных; формирования списка авторов со ссылками на их статьи, опубликованные в журнале; создания базы ключевых слов со ссылками на соответствующие публикации; преобразования статей архива в единый формат (см. <http://elbib.kpfu.ru>). Модули реализованы на языке PHP с использованием технологий CURL (<https://curl.haxx.se/>), html dom, htmsql. Статьи из архива журнала преобразованы в автоматическом режиме в формат .pdf с помощью технологий tcpdf, mpdf, а также онлайн-сервисов по конвертации данных. Выбор способа хранения метаданных этих статей в xml-файлах позволяет структурированно хранить информацию в виде текстовых файлов без применения какой-либо базы данных.

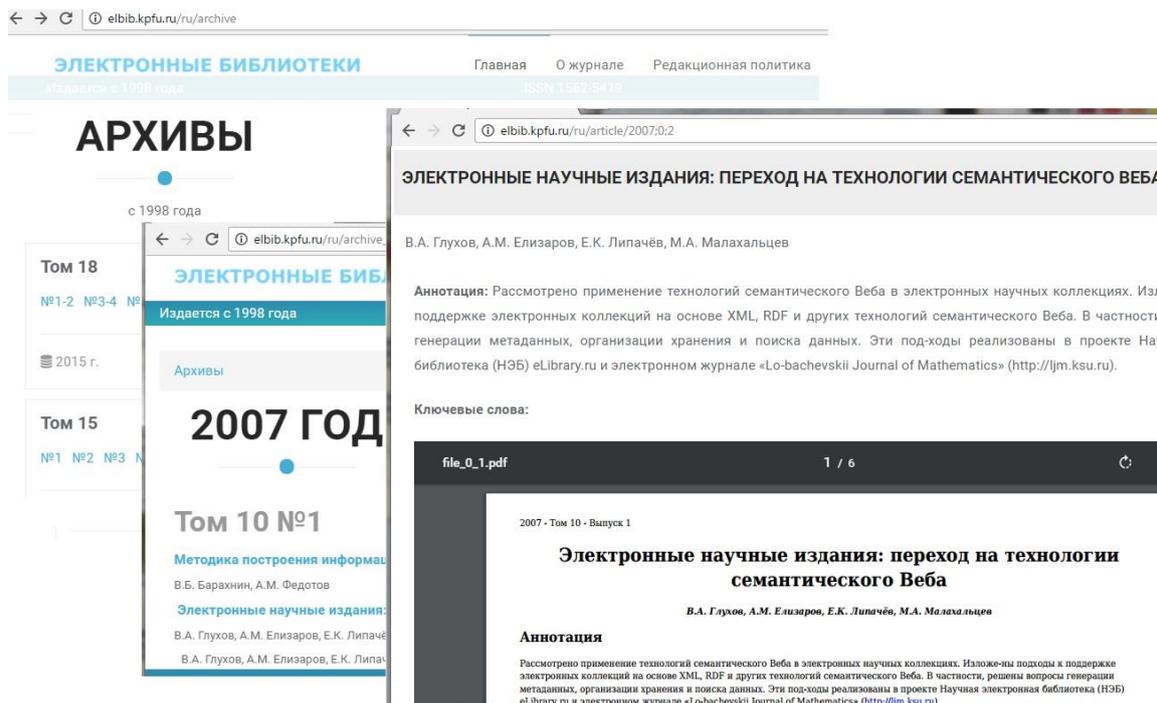


Рис. 6. Страница со статьей из архива elbib.kpfu.ru

Каждый номер журнала переведен в xml-формат (см. рис. 7). Для каждого автора создан xml-файл (рис. 8), содержащий информацию обо всех его работах, опубликованных в журнале. Для каждого ключевого слова формируется xml-файл (рис. 9), в который заносится информация об использовании этого слова в статьях архива.

```
<?xml version="1.0" encoding="utf-8"?>
<issue>
  <volume>18</volume>
  <title>№1-2</title>
  <articles>
    <article>
    <article>
    <article>
      <id>3</id>
      <title>АВТОМАТИЗАЦИЯ РЕДАКЦИОННЫХ ПРОЦЕССОВ В ИНФОРМАЦИОННОЙ СИСТЕМЕ УПРАВЛЕНИЯ
      ЭЛЕКТРОННЫМИ НАУЧНЫМИ ЖУРНАЛАМИ</title>
      <authors>Д.Ю. Ахметов, А.М. Елизаров, Е.К. Липачев</authors>
      <annotation>Исследованы особенности использования информационных систем в процессе
      издания электронных научных журналов и проведено их сравнение с точки зрения
      автоматизации редакционных процессов. Описаны программные модули, созданные для
      расширения функционала платформы Open Journal Systems в целях автоматизации ряда
      редакционных процессов электронного научного журнала. Приведены алгоритмы автоматической
      стилиевой валидации текстов на этапе регистрации автором статьи в информационной системе
      электронного научного журнала, автоматического подбора рецензентов, рассылки уведомлений
      и контроля сроков рецензирования.</annotation>
      <keywords>издательские системы; электронный научный журнал; интеграция электронных
      ресурсов; данных научного цитирования; экстракция метаданных; Open Journal Systems
      </keywords>
      <pages>32-45</pages>
```

Рис. 7. Фрагмент xml-файла номера журнала

```
<?xml version="1.0" encoding="utf-8"?>
<articles>
  <article>
    <id>2015;1;3</id>
    <title>АВТОМАТИЗАЦИЯ РЕДАКЦИОННЫХ ПРОЦЕССОВ В ИНФОРМАЦИОННОЙ СИСТЕМЕ УПРАВЛЕНИЯ
    ЭЛЕКТРОННЫМИ НАУЧНЫМИ ЖУРНАЛАМИ</title>
    <authors>Д.Ю. Ахметов, А.М. Елизаров, Е.К. Липачев</authors>
    <volume>18</volume>
    <number>№1-2</number>
  </article>
</articles>
```

Рис. 8. Фрагмент xml-файла, содержащего список статей автора

```
<?xml version="1.0" encoding="utf-8"?>
<keyword>
  <name>издательские системы</name>
  <articles>
    <article>
      <id>2015;1;2</id>
      <year>2015</year>
      <title>Формирование метаданных для международных баз цитирования в системе управления
      электронными научными журналами</title>
      <authors>А.Н. Герасимов, А.М. Елизаров, Е.К. Липачев</authors>
      <volume>18</volume>
      <number>№1-2</number>
    </article>
    <article>
      <id>2015;1;3</id>
      <year>2015</year>
      <title>АВТОМАТИЗАЦИЯ РЕДАКЦИОННЫХ ПРОЦЕССОВ В ИНФОРМАЦИОННОЙ СИСТЕМЕ УПРАВЛЕНИЯ
      ЭЛЕКТРОННЫМИ НАУЧНЫМИ ЖУРНАЛАМИ</title>
      <authors>Д.Ю. Ахметов, А.М. Елизаров, Е.К. Липачев</authors>
      <volume>18</volume>
      <number>№1-2</number>
    </article>
    <article>
      <id>2015;1;5</id>
```

Рис. 9. Фрагмент xml-файла, сформированного для ключевого слова
«издательские системы»

ЗАКЛЮЧЕНИЕ

Описана автоматизированная система научного журнала «Электронные библиотеки», реализующая семантические сервисы поиска и обработки документов, ориентированные на авторов, редакторов, рецензентов и читателей журнала. Представлены методы автоматического извлечения метаданных статей журнала, а также формирования и обработки наукометрической информации. Полученные результаты позволяют реализовать взаимодействие с международными и отечественными базами цитирования. В настоящее время в автоматическом режиме организована загрузка метаданных в базы данных Российского индекса научного цитирования.

Благодарности

Работа выполнена при финансовой поддержке РФФИ (проекты №№ 15-07-08522, 15-47-02472).

СПИСОК ЛИТЕРАТУРЫ

1. *Ершова Т.В.* Информационное общество – это мы! М.: Институт развития информационного общества, 2008. 512 с.
2. *Щур Л.Н.* Роль инфокоммуникационных технологий в развитии процесса глобализации научных исследований // Информационное общество. 2012. № 5. С. 16–24.
3. *Елизаров А.М., Зувев Д.С., Липачёв Е.К.* Управление жизненным циклом электронных публикаций в информационной системе научного журнала // Вестник Воронежского государственного университета. Серия: системный анализ и информационные технологии. 2014. № 4. С. 81–88.
4. *Горбунов-Посадов М.М.* Живая публикация // Открытые системы. 2011, № 4. С. 48–49. Обновляемая («живая») публикация: URL: <http://keldysh.ru/gorbunov/live.htm>.
5. *Паринов С.И., Козаловский М.Р.* «Живые» документы в электронных библиотеках // Прикладная информатика. 2009. № 6 (24). Авторская версия: <http://socionet.ru/publication.xml?h=repec:rus:isyigw:article-215>.
6. *Горбунов-Посадов М.М.* Интернет-активность как обязанность ученого // Информационные технологии и вычислительные системы. 2007. № 3. С. 88–93. URL: <http://keldysh.ru/gorbunov/duty.htm>.
7. *Olver P.J.* Journals in flux // Notices Amer. Math. Soc. 2011. V. 58, No 8. P. 1124–1126.
8. *Olver Peter J.* Journals in flux. Presentation: <http://www.math.umn.edu/~olver>.
9. *Елизаров А.М., Зувев Д.С., Липачёв Е.К.* Информационные системы управления электронными научными журналами // Научно-техническая информация. Серия 1. 2014. № 3. С. 31–38 (Elizarov A.M., Zuev D.S., and Lipachev E.K. Electronic scientific journal-management systems // Scientific and Technical Information Processing. 2014. V. 41, No 1. P. 66–72).

10. Best current practices for journals // Notices Amer. Math. Soc. 2011. V. 58. P. 62–65. URL: <http://www.ams.org/notices/201101/rtx110100062p.pdf>, <http://www.mathunion.org/fileadmin/CEIC/bestpractice/bpfinal.pdf>.

11. Моисеев Е.И., Муромский А.А., Тучкова Н.П., Меденников А.М. Организация персонального информационного пространства на основе модели предметной области автора // Научный сервис в сети Интернет, многообразии суперкомпьютерных миров: труды Международной суперкомпьютерной конференции (г. Новороссийск, 22–27 сентября 2014 г.). М.: Изд-во Московского университета, 2014. С. 429–431. URL: <http://agora.guru.ru/abrau2014/pdf/429.pdf>.

12. Елизаров А.М., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К. Семантическое аннотирование в системе управления физико-математическим контентом // Научный сервис в сети Интернет: труды XVII Всероссийской научной конференции (г. Новороссийск, 21–26 сентября 2015 г.). М.: ИПМ им. М.В. Келдыша, 2015. С. 98–103.

13. *Berners-Lee T.* Linked data – design issues. 2006. URL: <http://www.w3.org/DesignIssues/LinkedData.html>.

14. *Elizarov A., Kirillovich A., Lipachev E., Nevzorova O., Solovyev V., Zhiltsov N.* Mathematical knowledge representation: semantic models and formalisms // *Lobachevskii Journal of Mathematics*. 2014. V. 35, No 4. P. 347–353.

15. *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMathPro ontology: a linked data hub for mathematics // *Communications in Computer and Information Science*. 2014. V. 468. P. 105–119.

16. *Ахметов Д.Ю., Елизаров А.М., Липачёв Е.К.* Информационные системы и сервисы комплексной поддержки периодических научных изданий // Научный сервис в сети Интернет: труды XVII Всероссийской научной конференции (г. Новороссийск, 21–26 сентября 2015 г.). М.: ИПМ им. М.В. Келдыша, 2015. С. 16–25.

17. *Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Липачёв Е.К., Невзорова О.А., Соловьёв В.Д.* Методы анализа семантических данных математических электронных коллекций // Научно-техническая информация. Сер. 2. Информационные процессы и системы. 2014. № 4. С. 12–17.

18. *Елизаров А.М., Жильцов Н.Г., Иванов В.В., Кириллович А.В., Липачёв Е.К., Невзорова О.А.* Семантический рекомендательный сервис в профессиональной

деятельности математика // Ученые записки Института социальных и гуманитарных знаний. 2015. Вып. 1 (13). С. 190–197.

19. *Astafiev A., Prokofyev R., Guéret C., Boyarsky A., Ruchayskiy O.* ScienceWISE: a Web-based interactive semantic platform for paper annotation and ontology editing // In: Extended Semantic Web Conference 2012, Greece, 2012. URL: http://2012.eswc-conferences.org/sites/default/files/eswc2012_submission_316.pdf.

20. *Когаловский М.Р., Хохлов Ю.Е.* Стандарты XML для электронного правительства. М.: Институт развития информационного общества, 2008. 416 с.

21. *Когаловский М.Р., Хохлов Ю.Е.* Стандарты Всемирной паутины в разработках электронного правительства // Информационное общество. 2009. № 2. С. 21–32.

22. *Елизаров А.М., Липачёв Е.К., Хохлов Ю.Е.* Научные электронные коллекции как составляющая вузовской информационно-образовательной среды // Учёные записки Института социальных и гуманитарных знаний. 2013. № 2 (11). С. 102–111.

23. *Елизаров А.М., Зуев Д.С., Липачёв Е.К.* Свободно распространяемые системы управления электронными научными журналами и технологии электронных библиотек // Труды XV Всероссийской научной конференции – RCDL-2013, г. Ярославль, 14–17 октября 2013. Ярославль: Изд-во: Ярославский государственный университет им. П.Г. Демидова, 2013. С. 227–236 (CEUR Workshop Proc. Selected Papers of the 15th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections", Yaroslavl, Russia, October 14–17, 2013. V. 1108. P. 102–111. URL: <http://ceur-ws.org/Vol-1108/paper13.pdf>).

24. *Ахметов Д.Ю., Елизаров А.М., Липачёв Е.К.* Автоматизация редакционных процессов в информационной системе управления электронными научными журналами // Электронные библиотеки. 2015. Т. 18. № 1–2. С. 32–45.

25. *Barbera M., Donato F., Morbidoni C., Tummarello G.* HyperJournal software, PHP scripting and Semantic Web technologies for the Open Access// In ESWC: European Semantic Web Conference, Heraklion (Greece), 2005. URL: <http://eprints.rclis.org/8295/>.

26. *Candela L., Castelli D., Dobрева 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. URL: http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf.

27. *Елизаров А.М., Зуев Д.С., Липачёв Е.К.* Сервисы поддержки жизненного цикла электронных научных публикаций // Научный сервис в сети Интернет: многообразие суперкомпьютерных миров. Сб. трудов Международной суперкомпьютерной конференции. г. Новороссийск, 22–27 сентября 2014 г. М.: Изд-во Московского ун-та, 2014. С. 436–438.

28. *Chýla Ch.* What open source webpublishing software has the scientific community for e-journals? // CASLIN 2007, Stupava (Slovak Republic). URL: <http://eprints.rclis.org/10055/>.

29. *Czyzyk M., Choudhury S.* A survey and evaluation of open-source electronic publishing systems. 2008. URL: <http://jhir.library.jhu.edu/handle/1774.2/32737>.

30. Tools and Platforms. URL: http://www.openoasis.org/index.php?option=com_content&view=article&id=353&Itemid=379/.

31. Library publishing directory 2014 / Ed. S.K. Lippincott. Library Publishing Coalition, 2013. 245 p.

32. *Козаловский М.Р.* Метаданные, их свойства, функции, классификация и средства представления // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. XIV Всерос. науч. конф. RCDL-2012 (г. Переславль-Залесский, 15–18 октября 2012 г.). Переславль-Залесский, 2012. С. 3–14. URL: <http://rcdl.ru/doc/2012/paper3.pdf>.

33. *Глухов В.А., Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Электронные научные издания: переход на технологии семантического Веба // Электронные библиотеки. 2007. Т. 10, Вып. 1. URL: <http://www.elbib.ru>, <http://www.elbib.kpfu.ru/ru/article/2007;0;2>.

34. *Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Веб-технологии для математика: Основы MathML. М.: Физматлит, 2010. 192 с.

35. *Enge E., Spencer S., Stricchiola J., Fishkin R.* The Art of SEO. O'Reilly Media, 2012. 714 p.

36. *Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Организация взаимодействия языков разметки в системе автоматизации электронных научных хранилищ: семантический подход // Научный сервис в сети Интернет: масштабируемость, параллельность, эффективность: Тр. Всерос. суперкомпьютерной конференции (г. Новороссийск, 21–26 сентября 2009 г.). М.: Изд-во Московского ун-та, 2009. С. 456–457.

37. *Елизаров А.М., Липачёв Е.К., Малахальцев М.А.* Веб-технологии в работе электронного математического журнала Lobachevskii Journal of Mathematics // Научный сервис в сети Интернет: многоядерный компьютерный мир. 15 лет РФФИ. Тр. Всерос. науч. конф. М.: Изд-во Московского ун-та, 2007. С. 355–356.

38. *Васильев А., Козлов Д., Самусев С., Шамина О.* Извлечение метаинформации и библиографических ссылок из текстов русскоязычных научных статей // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. VIII Всерос. науч. конф. RCDL-2007 (г. Переславль-Залесский, 15–18 октября 2007 г.). Переславль-Залесский, 2007. С. 175–184. URL: http://rcdl.ru/doc/2007/paper_24_v1.pdf.

39. *Елизаров А.М., Зувев Д.С., Липачёв Е.К., Малахальцев М.А.* Сервисы структурирования математического контента и интеграция электронных математических коллекций в научное информационное пространство // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XIV Всероссийской научной конференции RCDL-2012 (г. Переславль-Залесский, 15–18 октября 2012 г.). Переславль-Залесский, 2012. С. 309–312. URL: <http://rcdl.ru/doc/2012/paper47.pdf> (Elizarov A., Zuev D., Lipachev E., Malakhaltsev M. Services structuring mathematical content and integration of digital mathematical collections at scientific information space // Proceedings of the 14th All-Russian Scientific Conference "Digital Libraries 2012: Advanced Methods and Technologies, Digital Collections". CEUR Workshop Proceedings, 2012. P. 309–312).

40. *Елизаров А.М., Липачёв Е.К., Хохлов Ю.Е.* Семантические методы структурирования математического контента, обеспечивающие расширенную поисковую функциональность // Информационное общество. 2013. № 1–2. С. 83–92.

41. Жижимов О.Л., Пестунов И.А., Федотов А.М. Структура сервисов управления метаданными для разнородных информационных систем // Электронные библиотеки. 2012. Вып. 6. URL: <http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2012/part6/ZPF>.

42. Григорьева Е.И., Курсанов А.С., Ситдииков И.М. Сайт научного журнала. URL: <http://www.isras.ru/publ.html?id=3058>.

43. Григорьева Е.И., Курсанов А.С., Ситдииков И.М. Каким должен быть сайт научного журнала // Полис. Политические исследования. 2014. № 5. С. 177–187. doi: 10.17976/jpps/2014.05.14.

44. Елизаров А.М., Зуев Д.С., Липачёв Е.К. Инфраструктура электронного научного журнала и облачные сервисы поддержки жизненного цикла электронных публикаций // Тр. 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL-2014, г. Дубна, Россия, 13–16 октября 2014 г. Дубна: ОИЯИ, 2014. С. 283–286. URL: http://rcdl.ru/doc/2014/paper/RCDL2014_283-286.pdf (CEUR Workshop Proceedings. Vol-1297. Selected Papers of XVI All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections" (RCDL 2014). P. 156–159. URL: http://ceur-ws.org/Vol-1297/156-159_paper-23.pdf).

45. Willinsky J., Stranack K., Smecher A., MacGregor J. Open Journal Systems: a complete guide to online publishing. Simon Fraser University Library, 2010. 273 p.

SERVICE-ORIENTED INFORMATION SYSTEM OF “RUSSIAN DIGITAL LIBRARIES JOURNAL”

D.Yu. Akhmetov¹, A.M. Elizarov², E.K. Lipachev³

N.I. Lobachevskii Institute of Mathematics and Mechanics.

Kazan (Volga Region) Federal University

¹akhmetov.dy@gmail.com, ²amelizarov@gmail.com, ³elipachev@gmail.com

Abstract

We investigated the current models of the publication and dissemination of scientific knowledge. We also presented different approaches to the organization of the

personal information space of scientist. We describe the modern information management system of scientific publications and services that determine their functionality. We have created software modules for automating of electronic scientific journal workflow. These modules are an extension of Open Journal Systems. We describe the architecture of a universal information management system of electronic scientific journals and management platform of “Russian Digital Libraries Journal”.

Keywords: publishing systems, advanced models of publication and dissemination of scientific knowledge, the information society, information space of scientist, electronic scientific journal, service-oriented information system, modern information management system of scientific publication, integration of electronic resources, extraction of metadata, infrastructure of “Russian Digital Libraries Journal, the platform Open Journal Systems.

REFERENCES

1. *Ershova T.V.* Informacionnoe obshhestvo – jeto my! M.: Institut razvitija informacionnogo obshhestva, 2008. 512 s.
2. *Shhur L.N.* Rol' infokommunikacionnyh tehnologij v razvitii processa globalizacii nauchnyh issledovanij // Informacionnoe obshhestvo. 2012. № 5. S. 16–24.
3. *Elizarov A.M., Zuev D.S., Lipachev E.K.* Upravlenie zhiznennym ciklom jelektronnyh publikacij v informacionnoj sisteme nauchnogo zhurnala // Vestnik Voronezhskogo gosudarstvennogo universiteta. Serija: sistemnyj analiz i informacionnye tehnologii. 2014. № 4. S. 81–88.
4. *Gorbunov-Posadov M.M.* Zhivaja publikacija // Otkrytye sistemy. 2011. № 4. S. 48–49. Obnovljaemaja («zhivaja») publikacija: URL: <http://keldysh.ru/gorbunov/live.htm>.
5. *Parinov S.I., Kogalovskij M.R.* «Zhivye» dokumenty v jelektronnyh bibliotekah // Prikladnaja informatika. 2009. № 6 (24). Avtorskaja versija: <http://socionet.ru/publication.xml?h=repec:rus:isyigw:article-215>.
6. *Gorbunov-Posadov M.M.* Internet-aktivnost' kak objazannost' uchenogo // Informacionnye tehnologii i vychislitel'nye sistemy. 2007. № 3. S. 88–93. URL: <http://keldysh.ru/gorbunov/duty.htm>.

7. *Olver P.J.* Journals in flux // *Notices Amer. Math. Soc.* 2011. V. 58, No 8. P. 1124–1126.

8. *Olver Peter J.* Journals in flux. Presentation: <http://www.math.umn.edu/~olver>.

9. *Elizarov A.M., Zuev D.S., Lipachev E.K.* Electronic scientific journal-management systems // *Scientific and Technical Information Processing.* 2014. V. 41, No 1. P. 66–72.

10. Best current practices for journals // *Notices Amer. Math. Soc.* 2011. V. 58. P. 62–65. URL: <http://www.ams.org/notices/201101/rtx110100062p.pdf>, <http://www.mathunion.org/fileadmin/CEIC/bestpractice/bpfinal.pdf>.

11. *Moiseev E.I., Muromskij A.A., Tuchkova N.P., Medennikov A.M.* Organizacija personal'nogo informacionnogo prostranstva na osnove modeli predmetnoj oblasti avtora // *Nauchnyj servis v seti Internet, mnogoobrazie superkomp'juternyh mirov: trudy Mezhdunarodnoj superkomp'juternoj konferencii* (g. Novorossijsk, 22–27 sentjabrja 2014 g.). M.: Izd-vo Moskovskogo universiteta, 2014. S. 429–431. URL: <http://agora.guru.ru/abrau2014/pdf/429.pdf>.

12. *Elizarov A.M., Zhiltsov N.G., Kirillovich A.V., Lipachev E.K.* Semanticheskoe anotirovanie v sisteme upravlenija fiziko-matematicheskim kontentom // *Nauchnyj servis v seti Internet: trudy XVII Vserossijskoj nauchnoj konferencii* (g. Novorossijsk, 21–26 sentjabrja 2015 g.). M.: IPM im. M.V. Keldysha, 2015. S. 98–103.

13. *Berners-Lee T.* Linked data – design issues. 2006. URL: <http://www.w3.org/DesignIssues/LinkedData.html>.

14. *Elizarov A., Zhiltsov N., Kirillovich A., Lipachev E., Nevzorova O., Solovyev V.* Mathematical knowledge representation: semantic models and formalisms // *Lobachevskii Journal of Mathematics.* 2014. V. 35, No 4. P. 347–353.

15. *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMathPro ontology: a linked data hub for mathematics // *Communications in Computer and Information Science.* 2014. V. 468. P. 105–119.

16. *Akhmetov D.Yu., Elizarov A.M., Lipachev E.K.* Informacionnye sistemy i servisy kompleksnoj podderzhki periodicheskikh nauchnyh izdanij // *Nauchnyj servis v seti Internet: trudy XVII Vserossijskoj nauchnoj konferencii* (g. Novorossijsk, 21–26 sentjabrja 2015 g.). M.: IPM im. M.V. Keldysha, 2015. S. 16–25.

17. *Birjal'cev E.V., Elizarov A.M., Zhiltsov N.G., Lipachev E.K., Nevzorova O.A., Solovyev V.D.* Metody analiza semanticheskikh dannykh matematicheskikh jelektronnykh kollekcij // Nauchno-tehnicheskaja informacija. Ser. 2. Inform. processy i sistemy. 2014. № 4. S. 12–17.

18. *Elizarov A.M., Zhiltsov N.G., Ivanov V.V., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* Semanticheskij rekomendatel'nyj servis v professional'noj dejatel'nosti matematika // Uchenye zapiski Instituta social'nyh i gumanitarnyh znaniy. 2015. Vyp. 1 (13). S. 190–197.

19. *Astafiev A., Prokofyev R., Guéret C., Boyarsky A., Ruchayskiy O.* ScienceWISE: a Web-based interactive semantic platform for paper annotation and ontology editing // In: Extended Semantic Web Conference 2012, Greece, 2012. URL: http://2012.eswc-conferences.org/sites/default/files/eswc2012_submission_316.pdf.

20. *Kogalovskij M.R., Hohlov Yu.E.* Standarty XML dlja jelek-tronnogo pravitel'stva. M.: Institut razvitija informacionnogo obshhestva, 2008. 416 s.

21. *Kogalovskij M.R., Hohlov Yu.E.* Standarty Vsemirnoj pautiny v razrabotkah jelektronnogo pravitel'stva // Informacionnoe obshhestvo. 2009. № 2. S. 21–32.

22. *Elizarov A.M., Lipachev E.K., Hohlov Yu.E.* Nauchnye jelektronnye kollekcii kak sostavljajushhaja vuzovskoj informacionno-obrazovatel'noj sredy // Uchjonye zapiski instituta social'nyh i gumanitarnyh znaniy. 2013. №2 (11). S. 102–111.

23. *Elizarov A.M., Zuev D.S., Lipachev E.K.* Open scientific e-journals management systems and digital libraries technology // CEUR Workshop Proc. Selected Papers of the 15th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections", Yaroslavl, Russia, October 14–17, 2013. V. 1108. P. 102–111. URL: <http://ceur-ws.org/Vol-1108/paper13.pdf>.

24. *Akhmetov D.Yu., Elizarov A.M., Lipachev E.K.* Avtomatizacija redakcionnykh processov v informacionnoj sisteme upravlenija jelektronnymi nauchnymi zhurnalami // Jelektronnye biblioteki. 2015. T. 18. № 1–2. S. 32–45.

25. *Barbera M., Donato F., Morbidoni C., Tummarello G.* HyperJournal software, PHP scripting and Semantic Web technologies for the Open Access // In ESWC: European Semantic Web Conference, Heraklion (Greece), 2005. URL: <http://eprints.rclis.org/8295/>.

26. *Candela L., Castelli D., Dobрева 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. URL: http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf.

27. *Elizarov A.M., Zuev D.S., Lipachev E.K.* Servisy podderzhki zhiznennogo cikla jelektronnyh nauchnyh publikacij // Nauchnyj servis v seti Internet: mnogoobrazie superkomp'yuternyh mirov. Sb. trudov Mezhdunarodnoj superkomp'yuternoj konferencii. g. Novorossiysk, 22–27 sentjabrja 2014 g. M.: Izd-vo Moskovskogo un-ta, 2014. S. 436–438.

28. *Chýla Ch.* What open source webpublishing software has the scientific community for e-journals? // In CASLIN 2007, Stupava (Slovak Republic). URL: <http://eprints.rclis.org/10055/>.

29. *Czyzyk M., Choudhury S.* A survey and evaluation of open-source electronic publishing systems. 2008. URL: <http://jhir.library.jhu.edu/handle/1774.2/32737>.

30. Tools and Platforms. URL: http://www.openoasis.org/index.php?option=com_content&view=article&id=353&Itemid=379/.

31. Library publishing directory 2014 / Ed. S.K. Lippincott. Library Publishing Coalition, 2013. 245 p.

32. *Kogalovskij M.R.* Metadannye, ih svojstva, funkcii, klassifikacija i sredstva predstavlenija // Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii: Tr. XIV Vseros. nauch. konf. RCDL-2012 (Pereslavl'-Zalesskij, 15–18 oktjabrja 2012 g.). Pereslavl'-Zalesskij, 2012. S. 3–14. URL: <http://rcdl.ru/doc/2012/paper3.pdf>.

33. *Glukhov V.A., Elizarov A.M., Lipachev E.K., Malakhaltsev M.A.* Jelektronnye nauchnye izdaniya: perehod na tehnologii semanticheskogo Veba // Jelektronnye biblioteki. 2007. T. 10, Vyp. 1. URL: <http://www.elbib.ru>; <http://www.elbib.ru>, <http://www.elbib.kpfu.ru/ru/article/2007;0;2>.

34. *Elizarov A.M., Lipachev E.K., Malakhaltsev M.A.* Veb-tehnologii dlja matematika: Osnovy MathML. M.: Fizmatlit, 2010. 192 s.

35. *Enge E., Spencer S., Stricchiola J., Fishkin R.* The Art of SEO. O'Reilly Media, 2012. 714 p.

36. *Elizarov A.M., Lipachev E.K., Malakhaltsev M.A.* Organizacija vzaimodejstvija jazykov razmetki v sisteme avtomatizacii jelektronnyh nauchnyh hranilishh: semanticheskij podhod // Nauchnyj servis v seti Internet: masshtabiruemost', parallel'nost', jeffektivnost': Tr. Vseros. superkomp'juternoj konferencii (g. Novorossijsk, 21–26 sentjabrja 2009 g.). M.: Izd-vo Moskovskogo un-ta, 2009. S. 456–457.

37. *Elizarov A.M., Lipachev E.K., Malakhaltsev M.A.* Veb-tehnologii v rabote jelektronnogo matematicheskogo zhurnala Lobachevskii Journal of Mathematics // Nauchnyj servis v seti Internet: mnogojadernyj komp'juternyj mir. 15 let RFFI. Trudy Vseros. nauch. konf. M.: Izd-vo Moskovskogo un-ta, 2007. S. 355–356.

38. *Vasil'ev A., Kozlov D., Samusev S., Shamina O.* Izvlechenie metainformacii i bibliograficheskikh ssylok iz tekstov russkojazychnyh nauchnyh statej // Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii: Tr. VIII Vseros. nauch. konf. RCDL-2007 (Pereslavl'-Zalesskij, 15–18 oktjabrja 2007 g.). Pereslavl'-Zalesskij, 2007. S. 175–184. URL: http://rcdl.ru/doc/2007/paper_24_v1.pdf.

39. *Elizarov A., Zuev D., Lipachev E., Malakhaltsev M.* Services structuring mathematical content and integration of digital mathematical collections at scientific information space // Proceedings of the 14th All-Russian Scientific Conference "Digital Libraries 2012: Advanced Methods and Technologies, Digital Collections". CEUR Workshop Proceedings, 2012. P. 309–312.

40. *Elizarov A.M., Lipachev E.K., Hohlov Yu.E.* Semanticheskie metody strukturirovanija matematicheskogo kontenta, obespechivajushhie ras-shirennuju poiskovuju funkcional'nost' // Informacionnoe obshhestvo. 2013. № 1–2. S. 83–92.

41. *Zhizhimov O.L., Pestunov I.A., Fedotov A.M.* Struktura servisov upravlenija metadannymi dlja raznorodnyh informacionnyh sistem // Jelektronnye biblioteki. 2012. Vyp. 6. URL: <http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2012/part6/ZPF>.

42. *Grigor'eva E.I., Kirsanov A.S., Sitdikov I.M.* Sajt nauch-nogo zhurnala. URL: <http://www.isras.ru/publ.html?id=3058>.

43. *Grigor'eva E.I., Kirsanov A.S., Sitdikov I.M.* Kakim dolzhen byt' sajt nauchnogo zhurnala // Polis. Politicheskie issledovanija. 2014. № 5. С. 177–187. doi: 10.17976/jpps/2014.05.14.

44. *Elizarov A.M., Zuev D.S., Lipachev E.K.* Infrastruktura jelektronnogo nauchnogo zhurnala i oblachnye servisy podderzhki zhiznennogo cikla jelektronnyh publikacij // CEUR Workshop Proceedings. Vol-1297. Selected Papers of XVI All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections" (RCDL 2014). P. 156–159. URL: http://ceur-ws.org/Vol-1297/156-159_paper-23.pdf.

45. *Willinsky J., Stranack K., Smecher A., MacGregor J.* Open Journal Systems: a complete guide to online publishing. Simon Fraser University Library, 2010. 273 p.

СВЕДЕНИЯ ОБ АВТОРАХ



АХМЕТОВ Дмитрий Юрьевич – аспирант Института математики и механики им. Н.И. Лобачевского Казанского (Приволжского) федерального университета.

Dmitriy Yurevich AKHMETOV, received MS degree in Mathematics from Kazan Federal University (2012). Currently is a graduate student at the N.I. Lobachevskii Institute of Mathematics and Mechanics of Kazan Federal University. Current scientific interests: knowledge extraction technologies, integration of scientific resources.

email: akhmetov.dy@gmail.com



ЕЛИЗАРОВ Александр Михайлович – доктор физико-математических наук, профессор, заслуженный деятель науки Республики Татарстан, зам. директора по научной деятельности Института математики и механики им. Н.И. Лобачевского Казанского (Приволжского) федерального университета.

Alexander Mikhailovich ELIZAROV – Doctor of Physics and Mathematics, Professor, Honoured Worker of Science of the Republic of Tatarstan, Deputy Director of the Lobachevskii Institute of Mathematics and Mechanics of Kazan Federal University. Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: amelizarov@gmail.com



ЛИПАЧЁВ Евгений Константинович – кандидат физико-математических наук, доцент кафедры теории функций и приближений Института математики и механики им. Н.И. Лобачевского Казанского (Приволжского) федерального университета.

Evgeny Konstantinovich LIPACHEV – Candidate of Physics and Mathematics, Associate Professor, Lobachevskii Institute of Mathematics and Mechanics of Kazan Federal University. Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: elipachev@gmail.com

Материал поступил в редакцию 14 декабря 2015 года

УДК 004.6

ИНТЕЛЛЕКТУАЛЬНЫЙ ПОИСК СЛОЖНЫХ ОБЪЕКТОВ В МАССИВАХ БОЛЬШИХ ДАННЫХ

А.М. Гусенков¹

*Институт вычислительной математики и информационных технологий
Казанского (Приволжского) федерального университета*

¹ gusenkov.a.m@gmail.com

Аннотация

Предложен подход к интеллектуальному поиску сложных объектов в различных типах структурно размеченных текстов, который может быть применен для обработки Больших данных (Big Data). Исследуются два вида представления информационных объектов: реляционные базы данных (РБД), которые структурно размечены своими схемами, и полнотекстовые естественнонаучные документы, содержащие математические выражения (формулы). Для таких полнотекстовых документов предлагается дополнительная автоматизированная разметка для организации поиска формул. В обоих случаях источником информации для построения онтологии и, в дальнейшем, организации поиска являются тексты на естественном языке, которые относятся к слабоструктурированным данным. Для РБД это комментарии к наименованиям таблиц и их атрибутов, а для естественнонаучных документов (статей, монографий и т. д.) – текстовое содержимое размеченных документов.

Ключевые слова: *большие данные, семантический поиск, слабоструктурированные данные, онтологии, реляционные базы данных, естественнонаучные тексты, разметка математических выражений.*

ВВЕДЕНИЕ

Экспоненциальный рост накопленных массивов данных, объём которых в настоящее время измеряется в зеттабайтах, привел к качественным изменениям в IT-технологиях сбора, хранения, управления, обработки и анализа информации. Одновременно в научный оборот вошло понятие Big Data (Большие данные).

Например, международная исследовательская компания Forrester [1] определяет это понятие как технологию в области аппаратного и программного обеспечения, которая объединяет, организует, управляет и анализирует данные, характеризующиеся «четырьмя V»: объемом (Volume), разнообразием (Variety), изменчивостью (Variability) и скоростью (Velocity):

- Volume – очень большой объем информации, накопленный в базах данных, трудоемко обрабатывать и хранить традиционными средствами СУБД;
- Variety – разнообразие форматов данных: способность приложения обрабатывать большие массивы данных, поступающие из разных источников в различных форматах, является главным критерием отнесения его к технологии Big Data; обычно такие приложения объединяют данные из разных источников (как внутренних, так и внешних по отношению к организации) и разной степени структурированности (структурированные, слабоструктурированные и неструктурированные); многие бизнес-задачи и научные эксперименты требуют совместной обработки данных различных форматов – это могут быть табличные данные в СУБД, иерархические данные, текстовые документы, видео, изображения, аудиофайлы и т. д.;
- Variability – изменчивость информации: например, информация, непрерывно поступающая с датчиков некоторых устройств или из интернета, имеющая важное значение для анализа, прогнозирования и принятия решений;
- Velocity – скорость накопления и обработки данных; в ряде задач востребованы технологии обработки данных в реальном времени.

В настоящее время разработаны технологии работы с Big Data, наиболее известными из них являются следующие:

- NoSQL [2] – ряд подходов к реализации хранилищ баз данных, имеющих отличия от моделей, используемых в РБД; их удобно использовать при работе с данными, структура которых не может быть жестко определена;
- MapReduce [3]– модель распределения вычислений, используется для параллельных вычислений при обработке очень больших наборов данных;
- Hadoop [4] – фреймворк с набором утилит и библиотек для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов; используется для реализации поисковых и контекстных механизмов

высоконагруженных сайтов; система защищена от выхода из строя любого из узлов кластера путем дублирования данных на других узлах;

- SAP HANA [5] – высокопроизводительная NewSQL-платформа для хранения и обработки данных; сочетание технологий OLAP и OLTP в базе данных SAP HANA создает унифицированный ракурс данных, полученных из систем обработки транзакций, систем анализа, принятия решений и планирования.

Проблема интеграции гетерогенных электронных ресурсов чрезвычайно многоаспектна и многообразна. М.Р. Когаловский предложил следующую классификацию систем интеграции данных [6]: интеграция информации на физическом, логическом и семантическом уровнях. Интеграция на физическом уровне сводится к конверсии данных из различных источников в требуемый единый формат их физического представления. Интеграция на логическом уровне предусматривает возможность доступа к данным, содержащимся в различных источниках, в терминах единой глобальной схемы, которая описывает их совместное представление с учетом структурных и, возможно, поведенческих (при использовании объектных моделей) свойств данных. Интеграция на семантическом уровне обеспечивает поддержку единого представления данных с учетом их семантических свойств в контексте единой онтологии предметной области. Достоинство семантического подхода заключается в том, что основой пользовательского интерфейса является при этом высокоуровневая модель данных, а возможность рассуждений в терминах онтологии выступает в качестве концептуальной модели. В качестве средства описания онтологий обычно используется язык OWL, разработанный рабочей группой Semantic Web Activity и рекомендованный консорциумом W3C [7].

Одной из основных целей интеграции ресурсов является возможность организации эффективного поиска информации в интегрированных электронных ресурсах.

В настоящее время активно развиваются исследования в области компьютерной лингвистики и обработки информации, представленной на естественном языке. Это работы и прикладные системы, связанные с созданием электронных словарей, тезаурусов, онтологий, а также алгоритмы автоматического извлечения фактов из текста. В рамках этого направления разработано большое количество специализированных систем поиска в различных предметных областях.

В статье рассматривается единый подход к интеллектуальному поиску сложных объектов в различных типах структурно размеченных текстов, который может быть применен для обработки Big Data.

Исследуются два вида представления информационных объектов: реляционные базы данных (РБД), которые структурно размечены своими схемами, и полнотекстовые естественнонаучные документы, содержащие математические выражения (формулы). Для таких полнотекстовых документов предлагается дополнительная автоматизированная разметка для организации поиска формул. В обоих случаях источником информации для построения онтологии [8] и, в дальнейшем, организации поиска являются тексты на естественном языке, которые относятся к слабоструктурированным данным [9]. Для РБД это комментарии к наименованиям таблиц и их атрибутов, а для естественнонаучных документов (статей, монографий и т. д.) – текстовое содержимое размеченных документов.

ИНТЕГРАЦИЯ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ

Предлагается следующий подход для интеграции реляционных баз данных с целью организации эффективного поиска в РБД больших информационных систем, насчитывающих несколько десятков локальных баз данных с различными логической структурой и физической организацией, но относящихся к одной предметной области. Для интеграции используются информация, извлекаемая из самих баз данных, а также более общая информация, относящаяся к предметной области в целом. Таким образом, для успешного решения задачи интеграции РБД использованы вспомогательные информационные ресурсы, содержащие физические модели баз данных, логические модели предметной области и тезаурусы пользовательской терминологии, представленные в формализме онтологий.

ОНТОЛОГИЯ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ

Для эффективной интеграции реляционных баз данных при организации поиска необходима информация о структуре самой базы данных. Анализ структурных проблем интеграции РБД приведен в работах [10–12]. Связность и близость значений атрибутов РБД определяются связностью кортежей через общие ключи. Таким образом, для выдачи пользователю осмысленной информации по-

исковая машина должна иметь информацию о возможных соединениях кортежей по ключам. Такую информацию достаточно легко получить непосредственно из схем баз данных, так как современные СУБД хранят информацию о ключах в своих системных данных. Восстановление всех возможных осмысленных комбинаций соединения таблиц РБД по ключам является достаточно простой задачей. Кроме того, большинство современных СУБД содержит информацию о возможных соединениях таблиц по ключам РБД для поддержки целостности данных.

Для описания структуры реляционных баз данных существует достаточное количество формализмов: DDL SQL, ER-диаграммы [13], UML [14] и другие диаграммные техники. Их мощности вполне достаточно для решения задач проектирования баз данных и поддержки доступа к базам данных, рассматриваемых изолированно, каждая в своей собственной системе атрибутов. При решении задач интеграции информации из двух или более реляционных баз данных возникают проблемы [15], связанные с особенностями наименований артефактов баз данных, в первую очередь, таблиц и столбцов. Эти особенности требуют рассмотрения наименований артефактов не просто как атомарных имен, а как самостоятельных объектов, возможно обладающих собственной структурой и имеющих между собой семантически нагруженные связи. Такие связи невыразимы в традиционных формализмах представления структуры реляционных баз данных. Таким образом, мы не можем не только решить задачу интеграции реляционных баз данных, но и сформулировать ее.

Известны подходы к описанию структуры конкретных реляционных баз данных с помощью онтологий. В целом варианты представления структуры реляционных баз данных сводятся к двум основным подходам.

Первый подход [16] предполагает преобразование множества таблиц конкретной базы данных во множество одноименных концептов со слотами, соответствующими столбцам определенной таблицы, и проецирование связей между таблицами (в виде мигрирующих ключей) в отношения между концептами. В набор концептов онтологии также включается множество доменов (встроенных или пользовательских типов данных). Экземплярами концептов онтологии в данном случае выступают записи таблиц, значения ключей и состав доменов перечисляемого типа.

Второй подход, предложенный в работах [11, 12] и описанный в данной статье, предполагает более высокую степень абстракции концептов. При представлении структуры РБД выделяются универсальные концепты (не зависящие от конкретной базы данных): ТАБЛИЦА, СТОЛБЕЦ, КЛЮЧ, ДОМЕН, соответствующие основным объектам баз данных, и универсальные отношения между ними:

ТАБЛИЦА содержит СТОЛБЕЦ

ТАБЛИЦА имеет первичный КЛЮЧ

ТАБЛИЦА имеет внешний КЛЮЧ

КЛЮЧ содержит СТОЛБЕЦ

СТОЛБЕЦ имеет тип ДОМЕН

Объекты (таблицы, столбцы, ключи и домены) конкретной базы данных представляются как экземпляры универсальных концептов соответствующего типа.

Вводятся две функции интерпретации (ФИ) [8]:

ФИ1: Если ТАБЛИЦА1 имеет первичный КЛЮЧ1 и ТАБЛИЦА2 имеет внешний КЛЮЧ1, то существует ТАБЛИЦА3, содержащая столбцы, принадлежащие ТАБЛИЦА1 и ТАБЛИЦА2.

ФИ2: Если ТАБЛИЦА1 содержит СТОЛБЕЦ1, то существует ТАБЛИЦА2, содержащая все остальные столбцы ТАБЛИЦА1, кроме СТОЛБЕЦ1.

Первая функция интерпретации соответствует операции соединения по ключу, вторая – операции проекции реляционного отношения, необходимой для сокращения множества столбцов, получаемого при соединении таблиц, до искомого.

Для РБД задача интеграции сводится к нахождению способов извлечения заданных атрибутов (столбцов) из таблиц интегрируемых БД. Для задачи построения совместного представления это будет собственно целевая таблица, для задачи переноса информации – это таблицы целевой базы. Таким образом, задача сводится к нахождению такой последовательности применения ФИ1 и ФИ2, которая даст в результате искомое множество столбцов $\{C\}$ из онтологии O , т. е. принадлежит известному классу задач о блуждании по ориентированному графу, где вершинами являются экземпляры концепта ТАБЛИЦА, а дугами – наличие общего ключа, ориентированного посредством отношений «имеет первичный» и «имеет

вторичный». Подходы к решению данного класса задач хорошо известны [17] и представляют лишь вычислительную сложность при большой размерности задачи.

Оба подхода имеют как преимущества, так и недостатки. В частности, первый подход позволяет сохранить в онтологии полный экстенционал базы данных, во втором случае это нереализуемо. Вместе с тем, первый подход подразумевает создание для каждой конкретной базы данных уникальной онтологии, что не дает возможности построить универсальные аксиомы.

Отмеченные особенности получаемых онтологий определяют необходимость выбора второго подхода при построении онтологий больших баз данных для задач их интеграции, так как мы имеем дело с базами данных произвольной и заранее неизвестной структуры.

ОНТОЛОГИЯ ПРЕДМЕТНОЙ ОБЛАСТИ

Создание онтологии предметной области является трудоёмким процессом, который, как правило, выполняется командой высококвалифицированных специалистов, как в конкретной предметной области, так и в области компьютерной лингвистики. В этом направлении проводится большое количество исследований, в том числе и в Казанском университете, среди которых можно выделить следующие [35–38]. В настоящей статье обсуждается подход, предложенный в работах [11, 12, 18, 19, 34], а именно, методика построения онтологии предметной области на основе логической модели «сущность–связь», представленной в виде ER-диаграмм, для нефтедобывающей корпорации. В качестве прототипа онтологии предметной области выбрана логическая модель данных Epicentre нефтетехнической корпорации Petrotechnical Open Software Corporation (POSC) [20], имеющая статус отраслевого стандарта. Модель представлена в виде ER-диаграмм, а также набора текстовых файлов на языке EXPRESS. В модели данных Epicentre определено более 1000 реально существующих технических и бизнес-объектов, связанных с разведкой и добычей нефти. В терминологии POSC эти объекты названы сущностями (entities). В модели определены характеристики, которые могут содержать сущности, названные атрибутами сущностей (attributes). Наиболее важными являются атрибуты, определяющие взаимосвязи между сущностями. Обь-

ектно-ориентируемая концепция наследования классов является частью архитектуры Epicentre. Так как модель данных достаточно велика, концепция наследования классов обеспечивает эффективный способ организации всех сущностей в логически связанную структуру.

В подходе, рассматриваемом в статье, в качестве единого средства представления онтологий выбран язык OWL [21], разработанный рабочей группой Semantic Web Activity и рекомендованный консорциумом W3C, а именно, диалект языка OWL-DL (Description Logic). Выбор языка OWL-DL обусловлен чисто практическими аспектами, в частности, его поддержкой в существующих сегодня системах описания знаний и системах логического программирования. Разработана схема конвертации модели Epicentre в язык описания онтологий OWL. Применялись следующие основные подходы:

- любой сущности Epicentre соответствует простой именованный класс OWL-онтологии с сохранением в именах классов приставок, позволяющих идентифицировать сущности-свойства и сущности-справочники; все эти классы располагаются в корне таксономического дерева онтологии;
- степени связности сущностей (один-к-одному, один-ко-многим, многие-ко-многим) в OWL соответствует определение простых свойств-атрибутов, если связанная сущность не является типом данных, и свойств-значений в противном случае; указание степени связи между классами реализовано с помощью понятия кардинальности в OWL.

В OWL отсутствуют структурные элементы, которые в полном объеме описывают определение уникальности Epicentre. Поэтому в определение каждого класса на языке OWL добавлено новое предопределенное свойство, в котором перечислены все атрибуты, образующие уникальный ключ. Аналогичным образом решена проблема сохранения условий ограничений.

Для каждой категории данных Epicentre в OWL-онтологии построены отдельные классы, в свойствах которых использовались встроенные типы данных языка OWL. Построена формальная LR(1)-грамматика модели Epicentre, на основе которой реализовано семантическое преобразование модели Epicentre, описан-

ной на языке EXPRESS, в онтологию на языке OWL. Выполнена русификация описания сущностей и атрибутов модели Epicentre, а также соответствующих им классов и свойств на OWL.

ЛИНГВИСТИЧЕСКИЙ ТЕЗАУРУС

Для создания лингвистической онтологии природно-технических объектов выбран подход, основанный на построении тезаурусов WordNet [22]. Словарь предметной области построен путем объединения словоформ из описаний сущностей и атрибутов модели Epicentre со словоформами из описаний атрибутов таблиц и доменов таблиц-справочников реляционных баз данных нефтедобывающей корпорации ОАО «Татнефть». Лексико-семантические характеристики баз данных, использованных при построении словаря, описаны в [11, 23]. Словарь содержит около 6000 словоформ. При построении тезауруса для каждого слова определен входной синонимический ряд (синсет). На лексико-семантических вариантах слов и синсетах определены следующие отношения: гипонимия, часть – целое, несовместимость, антонимия, конверсивность, омонимия.

Описаниям синсетов построенного таким образом тезауруса присущи особенности, вытекающие из специфики описания атрибутов реальных баз данных: наличие коротких фраз, сокращений, технических аббревиатур, орфографических ошибок. Однако в отличие от анализа сплошных текстов для баз данных имеется возможность уточнения распознавания входных слов путем просмотра содержимого доменов атрибутов таблиц и сопоставления их с онтологией предметной области.

ПОИСК НА ЕСТЕСТВЕННОМ ЯЗЫКЕ. ГЕНЕРАЦИЯ SQL-ЗАПРОСОВ

В данном разделе излагается методика автоматической генерации SQL-запросов, требующая от пользователя знаний только в своей предметной области. Данный подход использует онтологию предметной области, экземпляр онтологии РБД и тезаурус пользовательской терминологии, описанные в предыдущих разделах. Следует отметить, что автоматизация построения онтологии предметной области и языка предметной области (лингвистического тезауруса), а также способ создания экземпляра онтологии произвольной РБД обуславливают применимость данного подхода к его использованию технологиями сбора, хранения

и обработки информации больших данных [1, 2]. Автоматизация построения перечисленных объектов интеграции позволяет строить системы, более устойчивые к модификациям в процессе их эксплуатации. При изменениях логической модели предметной области онтологию предметной области можно перестроить, а лингвистический тезаурус пополнить. Все это относится не только к нефтедобывающей предметной области, но и к любой другой, имеющей представление в виде ER-диаграмм логической модели «сущность–связь».

Из существующих реализаций можно выделить систему InBase, разработанную школой А.С. Нариньяни [24], как наиболее близкую к предложенному подходу. Отличительной особенностью InBase является применение опережающего семантического анализа при разборе и понимании запросов. Разбор производится на основе объектной модели предметной области, которая привязывается проектировщиком к модели РБД. Данный подход обладает рядом существенных преимуществ по сравнению с применением синтаксических шаблонов. В частности, он избавляет от необходимости задавать в шаблонах запросов все возможные словоформы и порядки следования слов в запросе. Вместе с тем, следует признать, что InBase не перешла в стадию широкого применения.

Как нам представляется, подход, развиваемый в системе InBase, имеет ряд проблем. Наиболее существенной из них является трактовка столбцов РБД как некоторых атрибутов классов объектной модели предметной области. В реальных базах данных довольно часто используются конструкции, которые не могут быть адекватно отображены на объектные модели, так как имеют смысл скорее подзапроса, чем классического атрибута. Этот вопрос подробно рассмотрен в [15, 23], поэтому ограничимся одним примером. Так, в одной из исследованных промышленно эксплуатируемых РБД встречались столбцы с комментариями «Дебит скважины до ремонта» и «Дебит скважины после ремонта». Очевидно, что такие конструкции в объектных моделях соответствуют не атрибуту класса, а достаточно сложному выражению.

Второй проблемой является необходимость ручной привязки столбцов к атрибутам объектной модели. Помимо большой сложности этого процесса при первоначальном внедрении данной системы, это создает сложности при ее сопро-

вождении. При изменении РБД, что в реально эксплуатируемых системах – достаточно частый случай, требуются квалифицированная корректировка объектной модели и ее проекция на столбцы таблиц РБД.

В корпоративных информационных системах РБД могут содержать сотни таблиц и тысячи столбцов. Изучение вопроса, в какой конкретно РБД и в каких конкретно столбцах и таблицах содержится требуемая информация, может занять несколько дней, тогда как формулировка собственно запроса на естественном языке потребует нескольких минут.

В предлагаемом подходе обработка информации строится на следующих принципах:

- ведение диалога между системой и пользователем в табличном виде;
- использование семантических подходов для поиска столбцов;
- автоматический поиск возможных соединений по ключам БД;
- использование визуальных процедур для задания операций селекции.

Пользователь формулирует запрос в виде таблицы на естественном языке, используя термины тезауруса (языка предметной области). В табл. 1. приведён пример модельного запроса. Очевидно, что подобный запрос, сформулированный в виде предложения на естественном языке, был бы весьма громоздким не только для машинного анализа, но даже для понимания человеком.

Название столбца	Условие
Номер скважины	10*
Дата ввода в эксплуатацию	> 1.06.2001
Дата КРС	
Дебит нефти ожидаемый	>0
Дебит нефти фактический	

Таблица 1. Модельный запрос

Семантическая информация представлена в системе в виде семантической сети (онтологии) предметной области, которая содержит набор возможных связей между концептами. Операция семантического разбора комментариев к названиям столбцов РБД почти полностью автоматическая, что позволяет доста-

точно быстро привязать к предметной области большую корпоративную базу данных. Семантический разбор запроса пользователя заключается в выявлении подграфов на семантической сети, которые связывают концепты, именуемые словами и словосочетаниями, употребленными в названии столбцов. На искомые подграфы наложено условие минимальности либо по количеству связей, либо по их суммарной длине.

Поиск столбцов БД, наиболее релевантных названиям, задаваемым пользователем в своем запросе, производится путем сопоставления подграфов семантического разбора столбцов запроса и столбцов БД. При сопоставлении могут использоваться различные метрики, например, разность подграфов или расстояние между подграфами.

Для выявленных релевантных кандидатов, которые могут содержаться в различных таблицах, автоматически выявляются подмножества, для которых возможно построение соединений по ключам. Это алгоритмически несложная, но ресурсоемкая операция, в результате которой среди кандидатов выявляются столбцы, имеющие между собой связь по миграции ключей. Если комбинаций связанных релевантных столбцов возникает несколько, то дополнительно учитывается количество записей в построенном соединении. Построенный запрос предъявляется пользователю для задания дополнительных условий в режиме визуального конструктора.

Предложенная в статье методика и реализованная по этой методике система прошли тестирование на реальных базах данных ОАО «Татнефть». По большинству несложных запросов система показала достаточно хорошую релевантность результатов.

Реактивность системы сильно зависит от количества столбцов, задаваемых в запросе. Основное время тратится на поиск связей столбцов в БД, которые выполнялись в данной реализации простым перебором. Таким образом, предлагаемая технология в текущей реализации эффективно работает с запросами, содержащими небольшое количество столбцов, что вполне достаточно для запросов справочного характера. Для ограничения вариантов перебора соединений таблиц

используются связи, сохраненные в контексте пользователя, что позволяет повысить реактивность системы. Более подробно система описана в работах [11, 12, 25].

ПОИСК В ЕСТЕСТВЕННОНАУЧНЫХ ТЕКСТАХ, СОДЕРЖАЩИХ МАТЕМАТИЧЕСКИЕ ВЫРАЖЕНИЯ (ФОРМУЛЫ)

Подход, представленный в настоящей работе, направлен на интеграцию функциональных возможностей полнотекстового поиска и поиска по математическим формулам, при котором конечному пользователю предлагается формулировать поисковый запрос на поиск математической формулы в форме ключевых слов или словосочетаний (именных групп). Наиболее близкий подход предложен в математической поисковой системе EgoMath (доступна по URL: <http://egomath.projekty.ms.mff.cuni.cz>), которая в данный момент предоставляет возможности традиционного формульного поиска в синтаксисе LATEX и механизм переформулирования запроса (от ключевых слов к символьным обозначениям), однако алгоритм этого связывания не раскрыт в оригинальной статье авторов EgoMath [26]. Достаточно близкие подходы были предложены в работах [38–40]

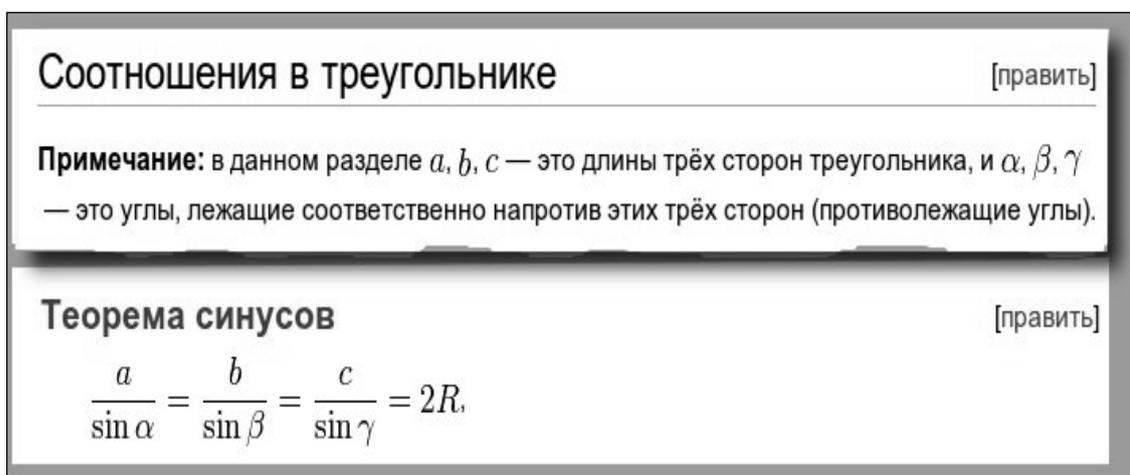
Новизна подхода, предлагаемого в данной статье, состоит в том, что в качестве поискового объекта рассматривается сложный нелинейный нетекстовый объект, включающий собственно математическое выражение формулы в нотации LATEX и набор определений символьных обозначений, участвующих в математическом выражении, которые извлекаются из всего анализируемого текста [12, 28, 29].

Данная постановка позволяет рассматривать предлагаемый подход как новый тип запросов к коллекциям математических документов. Действительно, в отличие от известной задачи поиска математической формулы по ее фрагменту в формулировке запроса должны использоваться не математические конструкции, а словесные наименования переменных, входящих в искомую формулу. В предлагаемом подходе используется дополнительная разметка естественнонаучных текстов, смысл которой заключается в следующем.

В естественнонаучных текстах выделяются виды сущностей: естественнонаучные термины, символьные условные обозначения терминов (переменные), ма-

тематические фрагменты (формулы). Для них определяются отношения: «термины – переменные» и «переменные – формулы». Первое отношение есть текстовое определение значения символа в некотором контексте с помощью терминов, второе отношение указывает на вхождение символа в формулу. Предполагается, что появление текстового определения переменной в окрестностях её символического представления указывает на семантическую связь между ними.

Все перечисленные сущности и отношения между ними составляют контекст формулы.



Соотношения в треугольнике [править]

Примечание: в данном разделе a, b, c — это длины трёх сторон треугольника, и α, β, γ — это углы, лежащие соответственно напротив этих трёх сторон (противоположные углы).

Теорема синусов [править]

$$\frac{a}{\sin \alpha} = \frac{b}{\sin \beta} = \frac{c}{\sin \gamma} = 2R.$$

Рис. 1. Пример формульного контекста (фрагмент статьи Википедии)

Пример расширенного формульного контекста приведен на рис. 1, где определения переменных даны в сплошном тексте, а формула является нетекстовым объектом.

Введем метод разметки математических выражений.

Шаг 1. Классификация математических выражений (МВ). МВ считается любой текст между специализированными тегами разметки `$` и `$` (рис. 2). В качестве инструмента анализа использован язык регулярных выражений. В МВ выделяются: символы арифметических и логических операций, переменные, переменные с индексами, ключевые слова, числа. Если МВ содержит только одну переменную или переменную с индексом, то оно классифицируется как переменная. Иначе МВ классифицируется как формула.

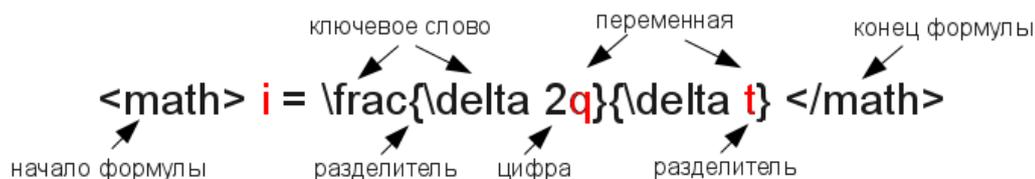


Рис. 2. Структура математической формулы в Википедии

Шаг 2. Связывание формул с переменными. Во всем анализируемом тексте для каждой переменной производится поиск вхождения этой переменной в каждую формулу. Пусть $\{F\}$ – множество формул, а $\{P\}$ – множество переменных. Для $\forall p_i \in P$, если $p_i \subset f_k \in F$, устанавливается отношение $\langle p_i, f_k \rangle$. Для каждого отношения в качестве атрибута запоминаются позиции формул и переменных в тексте.

В результате разметки получаем отношение много-ко-многим между формулами и переменными, входящими в состав этих формул.

ПОИСК В ВИКИПЕДИИ

Для проверки базовых концепций предложенного выше метода разметки математических выражений была реализована система семантического поиска математических выражений, специализированная для поиска математических формул в статьях русскоязычной Википедии. Выбор Википедии связан, с одной стороны, с тем, что последняя является одной из крупнейших коллекций научных текстов из различных областей знаний, с другой стороны, математические выражения в Википедии имеют унифицированную форму представления.

Для поиска формул в статьях Википедии была выполнена программная реализация предложенного подхода. Пользовательский интерфейс системы представляет собой веб-приложение, доступ к которому осуществляется через любой современный браузер с поддержкой JavaScript.

На странице ввода запроса пользователь может указать в текстовых полях одно или более названий параметров, которые должны присутствовать в искомой формуле, и после этого инициировать поиск.

Результаты поиска

Поиск формул в Википедии

Результаты поиска по фразам: "сила тока", "напряжение", "сопротивление":

- [Электрический ток](#)

$$I = \frac{U}{R}$$
 - ...в Амперах По закону Ома сила тока I пропорциональна приложенному...
 - ...приложенному напряжению U и обратно пропорциональна...
 - ...и обратно пропорциональна сопротивлению проводника R ::...
- [Закон Ома](#)

$$U = R \cdot I$$
 - ...или разность потенциалов, I – сила тока, R – сопротивление. Закон...
 - ...где: U – напряжение или разность потенциалов, I – ...
 - ...сила тока, R – сопротивление. Закон Ома также применяется ко всей цепи, но в...
- [Электромагнитная энергия](#)

$$U = I \cdot R$$
 - ... R можно выразить как через ток: $W = I(t)^2 \cdot R$...
 - ... так и через напряжение: $W = \frac{U(t)^2}{R}$...
 - ... выделяемую на сопротивлении R можно выразить как через [[сила...
- [Схемы на переключаемых конденсаторах](#)

$$I = \frac{U}{R}$$
 - ...(1) где: I – сила тока, U – напряжение или разность ...
 - ...(1) где: I – сила тока, U – напряжение или разность потенциалов, R – ...
 - ... – напряжение или разность потенциалов, R – сопротивление. Сопротивление цепи рассчитывается по...
- [Электродный котёл](#)

$$J = \frac{U}{R}$$
 - ... - мощность котла, Вт; J - сила тока, А; U - напряжение, ...
 - ... - сила тока, А; U - напряжение, В. Согласно закону Ома $U = JR$, ...
 - ... R - сопротивление жидкости, Ом, которое определяется согласно...
- [Электрическая мощность](#)

$$p(t) = u(t) \cdot i(t)$$

Готово

Рис. 3. Закон Ома. Поиск по параметрам: сила тока, напряжение, сопротивление

Результаты поиска отображаются на странице результатов, где выводятся запрос, а также ранжированный по релевантности список ссылок на страницы Википедии. Результаты поиска представляются наборами фрагментов, содержащих нетекстовый объект (формулу), и фрагментов, связывающих переменные формулы и их текстовые определения вне зависимости от их местонахождения в тексте (см. рис. 3).

Система поиска в Википедии включает следующие взаимосвязанные подсистемы: загрузки и анализа данных Википедии; полнотекстового индексирования; индексирования математических формул и переменных; поиска и ранжирования.

Производится импорт данных русскоязычной Википедии из предварительно полученного архива. Единицей анализируемой и загружаемой информации является html-страница. Страницы, содержащие формулы, сохраняются в базе данных для дальнейшего использования, остальные отбрасываются.

В системе производятся индексирование входных документов как текстовых данных (библиотека Apache Lucene [27]), а также дополнительное позиционное индексирование, построенное в результате применения метода разметки математических выражений.

Поиск и ранжирование выполняются в два этапа. На первом этапе производится полнотекстовый поиск всех вхождений ключевых словосочетаний (терминов) в тексты. Для каждого вхождения определяется, существует ли в некоторой окрестности ключевой фразы переменная. Для определения окрестности введено понятие максимально допустимого расстояния (МДР) – расстояния в символах влево и вправо от термина, в пределах которого может находиться переменная. По найденным переменным определяется соответствующая формула. Для каждой формулы строится группа текстовых фрагментов, включающих термины и переменные.

На втором этапе производится поиск наилучшей группы текстовых фрагментов для всей совокупности введенных ключевых фраз. Для этого составляются и проверяются по критерию близости все возможные сочетания полученных текстовых фрагментов в документе. В качестве критерия близости использован минимум среднеквадратичного отклонения найденных фрагментов с определениями переменных от позиции соответствующей формулы. В результате для каждого документа получим оптимальную группу текстовых фрагментов (потенциальных определений) и относящуюся к ним формулу. Результаты для всех документов сортируются по критерию близости и дополнительным критериям релевантности, связанным с полнотой вхождения выделенных переменных в формулу.

Реализованный в системе алгоритм поиска математических формул по введенным ключевым фразам (названиям параметров формулы) показал достаточную релевантность в сочетании с высокой скоростью поиска. Проведенное тестирование выявило, что выдаваемые результаты практически всегда имеют непосредственное отношение к задаваемому запросу, и на первой странице поиска находится формула, отвечающая запросам пользователя. Для дальнейшего улучшения релевантности необходимо модифицировать механизм поиска, а не механизм ранжирования.

Наиболее очевидным путем представляется синтаксический анализ текста с целью выделения терминологических словосочетаний – именных групп (ИГ) – и дальнейшего анализа отношения выделенных ИГ и заданных пользователем наименований параметров, также являющихся ИГ. С этой целью был разработан подход к разметке для поиска по онтологии с использованием связывания формул с ИГ.

РАЗМЕТКА ДЛЯ ПОИСКА ПО ОНТОЛОГИИ

Для организации поиска в связной коллекции математических текстов автоматизировано строится онтология, включающая в себя, кроме самих компонентов онтологии, связанные с ними переменные и формульные выражения [30, 31]. Поиск запрос пользователя на естественном языке переводится в термины онтологии, по которым затем формируется запрос на языке запросов к RDF-документам SPARQL [32]. Автоматизация построения онтологии упрощает пополнение коллекции математических текстов, что играет важную роль при работе с Big data [1, 2].

В данной статье рассматривается только дополнение связной коллекции математических текстов, размеченных для построения онтологии, разметкой формул и входящих в них переменных, а также их связей с ИГ [12, 29, 33].

Модуль формульной разметки реализован на языке Java в формате плагина к текстовому процессору GATE [41], для разметки используются средства работы с аннотациями библиотеки Gate и оригинальные алгоритмы.

Формульная разметка применяется к XML-документу, предварительно размеченному стандартными аннотациями (Token, Sentence, Math и др.) и NLP-аннотациями (TERM, ENDS). Ключевыми аннотациями для работы алгоритма являются аннотации Math, размечающие формульные фрагменты, и аннотации TERM, соответствующие именованным группам.

На основе аннотаций Math строится внутренняя модель документа, содержащая набор разобранных, классифицированных, связанных между собой формульных фрагментов.

Обработка XML-документа включает в себя следующие действия:

- выделение и анализ формульных фрагментов;
- определение связей между переменными и формульными фрагментами;
- определение связей между формульными фрагментами и ИГ;
- дополнение аннотаций Math атрибутами формульной разметки.

В первых двух действиях использован метод разметки математических выражений.

На третьем шаге вводится понятие максимально допустимого расстояния (МДР) между аннотациями Math и TERM, которое определяется как наибольшее расстояние в символах между концом левой аннотации и началом правой, при котором может быть выполнено связывание. МДР является параметром, который оказывает непосредственное влияние на точность связывания и может различаться для разных коллекций документов.

В документах встречается различное взаимное расположение формул и ИГ:

- ИГ содержит формулу, тогда она – единственный кандидат для связывания. В простейшем случае ИГ состоит из единственного главного слова. В более сложном случае она содержит более одного слова, и рассматривается расстояние между формулой и главным словом. Если это расстояние составляет более трех символов, формула считается дополнением и не связывается;

- Формула и ИГ следуют друг за другом (в пределах одного предложения). В этом случае основой анализа является концепция МДР. Алгоритм позволяет связывать формулу только с одной именной группой, но с одной и той же ИГ может быть связано более одной формулы.

На заключительном этапе связи, построенные на внутренней модели, переносятся в обрабатываемый документ и используются при построении онтологии связанной коллекции документов.

Рассмотрим вышесказанное на примере аннотирования математического текста, изображенного на рис. 4.

Пусть $\overline{\alpha}$ — вторая фундаментальная форма n -поверхности \overline{M} , $\overline{\nabla}$ — связность Леви-Чивита метрики \overline{g} . Имеет место [1] равенство

$$\partial_X dfY - df\overline{\nabla}_X Y = \overline{\alpha}(X, Y), \quad (2)$$

Рис. 4. Фрагмент математического текста

В результате обработки документа, фрагмент из которого приведён на рис. 4, аннотации, соответствующие фрагменту, будут дополнены атрибутами формульной разметки (выделены жирным шрифтом):

Math Id=960 mode=inline, **termid=965**, tex=\overline{\alpha}, text=overline@(\alpha), **varid=3**, xml:id=p10.m1

TERM Id=965 Form=вторая фундаментальная форма \$\$\$-поверхности \$\$\$, HeadBegin=0, HeadEnd=27

Math Id=966 mode=inline, tex=n, text=n, **varid=10**, xml:id=p10.m2

Math Id=969 mode=inline, **termid=965**, tex=\overline{M}, text=overline@(M), **varid=0**, xml:id=p10.m3

Math Id=974 mode=inline, **termid=979**, tex=\overline{\nabla}, text=overline@(\nabla), **varid=8**, xml:id=p10.m4

TERM Id=979 Form=связность Леви-Чивита метрики \$\$\$, HeadBegin=0, HeadEnd=20

Math Id=980 mode=inline, tex=\overline{g}, text=overline@(g), **varid=7**, xml:id=p10.m5

TERM Id=987 Form=равенство \$\$\$, HeadBegin=0, HeadEnd=8

Math Id=989 mode=display, **termid=987**,
tex=\partial_{\{X\}}dfY-df\overline{\nabla}_{\{X\}}Y=\overline{\alpha}(X,Y),,
text=(partial-differential _ X)@(d * f * Y) - d * f * (overline@(\nabla)) _ X * Y =
overline@(\alpha) * open-interval@(X, Y), **vars=3;8**, xml:id=S0.E2.m1

Выделенная в тексте формула очищается от служебных символов языка разметки и лишних пробельных символов. Затем формульный фрагмент разбивается на элементы: разделителями считаются различные символы скобок, символы арифметических и логических операций, знаки пунктуации, пробельные символы и т. п. Полученные элементы анализируются на принадлежность к специальным группам – ключевые слова (начинаются с символа «\»), нижние индексы (начинаются с символа «_»), числа и т. п. Если элемент на этом этапе не классифицирован, то с большой долей вероятности его можно считать переменной. К таким элементам дополнительно применяется проверка на соответствие правилам именования переменных – не начинается с цифры, может быть буквой греческого алфавита (например, “\alpha”), может содержать индекс. В результате все формульные фрагменты документа будут разделены на три типа: переменные, формулы, служебные (содержащие только символы разметки).

Одиноким переменным, выявленным в тексте, сопоставляются уникальные идентификаторы, которые будут использованы при дополнении аннотаций. Следует отметить, что алгоритм различает строчные и прописные буквы (X и x), индексированные и неиндексированные переменные (Y и Y_i), при этом варианты обозначения индекса не сказываются на идентификации переменной (Y_0 и Y_{i+1} определяются как одна и та же переменная). Для переменных, входящих в состав формульных фрагментов, строится индекс вхождения переменных в формулы. Для формул также определяются связи между фрагментами вида “содержит” и “содержится в”. Служебные фрагменты не рассматриваются как не несущие смысловой нагрузки.

Так, фрагмент текста на рис. 4 содержит переменные $\overline{\alpha}$, n , \overline{M} , $\overline{\nabla}$, \overline{g} и формулу, в которую входят идентифицированные переменные $\overline{\alpha}$, $\overline{\nabla}$. Переменные X , Y не присутствуют в тексте вне формул, поэтому идентификаторы им не сопоставлены.

Формулы и именные группы могут идти последовательно друг за другом, не пересекаясь. Другой вариант – когда формула содержится внутри именной группы. Частичное пересечение формул и именных групп не встречается, так как это противоречило бы структуре XML. Алгоритм связывания построен с учётом особенностей взаимного расположения формул и именных групп.

Для каждого формульного фрагмента с помощью средств работы с аннотациями библиотеки Gate определяется тип расположения и набор именных групп – кандидатов на связывание, из которых по заданным критериям отсеиваются неподходящие и отбираются наиболее близкие.

Если именная группа содержит формулу внутри себя, она становится единственным кандидатом на связывание. В простом случае, когда именная группа не содержит других слов, кроме главного слова, связывание будет проведено с высокой степенью достоверности. Например, в примере, приведённом на рис. 4, во втором предложении выделена ИГ “равенство \$\$\$”, где “\$\$\$” обозначает вхождение соответствующей формулы в именную группу. В данном случае формула, обозначенная (2), будет связана с ИГ “равенство \$\$\$”.

В более сложных случаях анализируется расстояние между формулой и главным словом именной группы (атрибуты HeadBegin, HeadEnd аннотации TERM). Если оно оказывается больше допустимого интервала между словами (3 и более символа), считается, что формула является дополнением к основному понятию именной группы. В этом случае связывание не производится. Например, переменная \overline{g} из текста на рис. 4 не будет связана с ИГ “связность Леви-Чивита метрики \$\$\$”, так как главное слово здесь – “связность Леви-Чивита”. Также не связываются формулы, входящие в конструкции с дефисом, как не имеющие самостоятельного значения. Например, для фразы “вторая фундаментальная форма n-поверхности \overline{M} ” формула “n” останется не связанной, а формула “ \overline{M} ” будет связана с ИГ “вторая фундаментальная форма \$\$\$-поверхности \$\$\$”.

Если формула не содержится внутри именной группы, задача определения набора ИГ, с которыми возможно связывание, усложняется. Сначала определяются границы области, в которой должны располагаться аннотации-кандидаты. За левую границу принимается позиция в документе, соответствующая ближайшей левой аннотации ENDS и отстоящая от начала формулы влево не более чем на МДР. Аналогично за правую границу принимается позиция в документе, соответствующая ближайшей правой аннотации ENDS и отстоящая от конца формулы вправо не более чем на МДР. Кроме того, если формула входит в группу равенств

(аннотация `equationgroup`), отсчёт ведётся от начала и конца всей группы. Это необходимо, чтобы все уравнения группы были привязаны к одной и той же ИГ.

По границам области связывания определяются левый и правый наборы аннотаций `TERM`, из которых затем выбирается одна аннотация, находящаяся на минимальном расстоянии от формулы. Для повышения достоверности связывания при обнаружении распространённой в математических текстах конструкции “<формула> – <ИГ>” приоритет отдаётся правому набору (например, определение $\overline{\alpha}$ и $\overline{\nabla}$ на рис. 4).

Следует отметить, что алгоритм позволяет связывать формулу только с одной именной группой, но с одной и той же ИГ может быть связано более одной формулы. Это позволяет учитывать перечисления формул, относящихся по семантике к одной ИГ, но в то же время может давать некоторое количество недостоверных связываний.

На заключительном этапе связи, построенные на внутренней модели, переносятся в обрабатываемый документ. В аннотации `Math` в зависимости от типа формульного фрагмента и наличия связанных ИГ добавляются новые атрибуты. Обозначим через `<variable_id>` уникальный идентификатор переменной, присвоенный ей во внутренней модели документа. Отметим, что всем вхождениям какой-либо переменной в формульные фрагменты любого типа будет соответствовать один и тот же идентификатор. Переменные, встречающиеся только в составе сложных формул, идентификатора не имеют. Для формульной разметки используются следующие атрибуты:

`varid=<variable_id>` – идентификатор переменной, добавляется в аннотации к одиночным переменным;

`vars=<variable_id1>;<variable_id2>;...` – список идентификаторов переменных, входящих в формулу, добавляется в аннотации к формулам;

`termid=<annotation_id>` – идентификатор аннотации `TERM` в документе, соответствующей именной группе, с которой связана формула, добавляется к переменным и формулам при наличии связи с ИГ.

Реализованный модуль формульной разметки был протестирован на коллекциях XML-документов, предварительно размеченных стандартными аннота-

циями GATE. Была проведена оценка релевантности и полноты связывания математических выражений и именных групп, выявленных в тексте, которая основана на экспертной оценке качества связывания на корпусах математических текстов. В качестве документов для разметки использовались статьи журнала «Известия вузов. Математика» за 1997–2009 гг. Результаты анализа связывания при изменении МДР от 15 до 40 симметрично в обе стороны представлены в табл. 2.

МДР	Math	Terms	VirOK%	NotVirOK%	TotalOk%	VirBad%	Others%	TotalBad%
15	1247	1357	36,33	30,47	66,80	23,90	9,30	33,20
20	1247	1357	42,34	25,50	67,84	25,66	6,50	32,16
25	1247	1357	40,98	20,69	61,67	23,02	15,32	38,33
30	1247	1357	41,38	21,49	62,87	27,83	9,30	37,13
35	1247	1357	41,86	21,01	62,87	29,03	8,10	37,13
40	1247	1357	42,02	19,65	61,67	29,67	8,66	38,33

Таблица 2. Статистика связывания в зависимости от МДР

Для каждого заданного значения МДР на всем корпусе текстов определялись следующие параметры: Math – количество выделенных формул; Terms – количество выделенных ИГ; VirOK – процент правильных связываний формул с ИГ (полнота связывания); NotVirOK – процент правильных несвязываний формул с ИГ (т. е. констатация того факта, что математическое выражение находится в контексте, не содержащем его определения); TotalOk – общий процент правильно обработанных формул (сумма VirOK и NotVirOK); VirBad – процент неправильных связываний формул с ИГ (из возможных кандидатов на связывание была выбрана не подходящая по семантике ИГ или произошло связывание математического выражения в контексте, не предполагающем связывания); Others – другие ошибки связывания (отсутствие связывания там, где оно должно было быть; неправильное выделение ИГ; нераспознанные ИГ; влияющие на связывание особенности оформления текста автором); TotalBad – общий процент неправильно обработанных формул (сумма VirBad и Others).

Из табл. 2 видно, что процент правильно обработанных формул TotalOk и процент ошибок всех типов TotalBad изменяются незначительно, что свидетельствует об устойчивости применяемого алгоритма. Тем не менее, процент правильно связанных математических выражений VirOK имеет тенденцию к возрастанию с увеличением МДР, что вполне ожидаемо. Общий процент ошибок связывания TotalBad также растет с увеличением МДР. Вместе с тем, изменения этих параметров имеют нелинейную зависимость, что позволяет сделать выбор оптимального МДР, при котором отношение VirOK/TotalBad максимально. Для данного корпуса текстов оптимальное МДР составляет 20 символов.

В целом описанные реализации показали принципиальную работоспособность и хорошую устойчивость результатов при применении предложенных подходов, а также выявили ряд проблем с релевантностью поиска. Для повышения релевантности и полноты связывания можно определить следующие направления дальнейшего развития предлагаемого подхода: дополнительный анализ контекста документа; разработка шаблонов, типичных для математических текстов; использование Байесовских методов; обучение распознаванию ключевых слов и конструкций на основе экспертного связывания.

ЗАКЛЮЧЕНИЕ

Рассмотренные подходы реализации поиска сложных объектов в структурно размеченных текстах, опыт их реализации и проведенные эксперименты показали, что такие тексты представляют собой многосвязные графы, неявно отражающие семантику предметной области.

Эксперт-пользователь, формулирующий запрос, знает, что между введенными им ключевыми словами должна быть некоторая связь, благодаря чему в достаточно большом наборе данных такие связи находятся и имеют содержательный смысл. Этим, вероятно, объясняется высокая эффективность простого переборного механизма поиска, реализованного в описываемом подходе. Таким образом, как структура хранения текстов, так и сформулированный запрос несут в себе неявную семантику, комбинация которых позволяет поисковому механизму найти полное и достаточно релевантное множество ответов, не имея априорной информации о предметной области. Вместе с тем, рассматриваемый поисковый механизм имеет несколько очевидных направлений дальнейшего развития.

Интересным частным результатом является фактическое выявление, при анализе особенностей структуры транзакционных баз данных, нового механизма нормализации данных. Известные нормальные формы, включая наиболее общую доменную нормальную форму, предложенную R. Fagin в 1981 году, не меняют состава атрибутов базы данных, а лишь перераспределяют их по кортежам. Вместе с тем, в реальных базах данных достаточно часто ключ с малым числом значений разворачивается в группу имен атрибутов или таблиц, т. е. множество значение атрибута и множество атрибутов могут трансформироваться друг в друга. Де-факто такая практика распространена в реальных структурах транзакционных БД, что нашло отражение в некоторых практических конструкциях механизмов хранения промышленных БД, например, сегментация объемных таблиц по ключу, поддерживаемая в Oracle, PostgreSQL и ряде других систем. До некоторой степени эта практика легитимизирована в хранилищах, при интеграции данных из различных источников, при версионности данных, их непрогнозируемом изменении во времени и ряде других случаев. Для поддержания хотя бы первой нормальной формы все атрибуты кортежей, обладающих данными особенностями, хранятся в отдельных таблицах с дополнительными атрибутами времени, источника, а также первичным ключом родительского кортежа. В результате форма хранения ряда атрибутов приобретает совершенно унифицированный вид и их можно хранить в одной таблице, введя дополнительный атрибут «Имя атрибута». Довольно часто такой прием используется при хранении группы однотипных по структуре справочников. Таким образом, практика трансформации имени атрибута в значение его супертипа существует, но в теоретическом плане пока получила слабое освещение. На наш взгляд, детальное изучение этого феномена может дать новый вклад в теорию реляционных баз данных.

Применительно к рассматриваемой теме дальнейшая формализация данного механизма может дать ключ к автоматизации построения запросов и репликации данных между транзакционными системами с разной структурой баз данных, а также с хранилищами, многомерными OLAP системами и другими форматами хранения, в том числе на базе XML. Весьма актуально может быть развитие данного направления для поиска в полнотекстовых базах данных, содержащих табличные данные. Учитывая возрастающий интерес к развитию направления Big

Data, для которых в принципе характерна разнородность и разноуровневость информации, формализация поисковых механизмов, учитывающая возможные преобразования имен атрибутов в их значение, может иметь хорошие перспективы дальнейшего развития.

Другим интересным результатом, вытекающим из рассмотренного подхода, является применение комбинаторного механизма к поиску математических зависимостей между параметрами, для которых в анализируемом корпусе текстов нет формулы, содержащей все заданные в запросе параметры. Пока реализован поисковый механизм для формул, ищущий явно прописанную формулу, объединяющую заданные в запросе параметры. В конструкторе запросов к базам данных в сконструированном запросе может участвовать несколько отношений, связь между которыми извлекается из схемы БД. Перенос данного подхода на поиск формул может дать фактически генератор новых знаний в виде цепочки формул, выстраивающих такие связи между переменными запроса, которые семантически подразумеваются, но в анализируемом корпусе текстов не присутствуют явно. Не исключено, что комбинаторный подход позволит получать такие цепочки вывода без предварительного обучения. Возможным препятствием к применению данного подхода является его большая комбинаторная сложность, однако непрерывно возрастающие вычислительные мощности и, возможно, некоторые новые закономерности в естественнонаучных текстах, полученные в новых исследованиях, позволят преодолеть этот барьер.

Неиспользованным резервом является подключение к комбинаторным механизмам поиска путей на графах статистических подходов. Очевидно, что в большом корпусе текстов будет содержаться несколько вариантов искомой информации. Статистический анализ, примененный к множеству результатов, позволяет решать несколько задач. В первую очередь можно повысить релевантность выдаваемой информации путем статистического анализа результатов, отсеивая редко встречающиеся варианты. Это позволит работать на недостаточно выверенных текстах, так как редко встречающиеся варианты скорее всего ошибочны. Опыт работы с естественнонаучными текстами показывает, что для обозначения конкретных физических и технических параметров применяется ограниченный набор имен переменных. Так, например, буквой P может быть обозначено давление или

объем промышленного производства, но никогда не обозначаются расстояние или денежный агрегат. Статистический анализ имен переменных, ассоциированных с запрашиваемыми параметрами, и их сочетаемость позволят, вероятно, еще более повысить релевантность. Статистический анализ сочетаемости также, возможно, позволит автоматизировать построение онтологий предметной области и получить ряд других полезных результатов.

Таким образом, комбинаторный поиск на графах, неявно задаваемых структурой хранения данных, обладает определенной универсальностью в силу того, что он позволяет решать различные поисковые задачи, и является развиваемым, так как имеет очевидные перспективы объединения с лексико-синтаксическими системами разбора текстов и статистическим анализом, а также в связи с бурным развитием таких областей, как Big Data и искусственный интеллект. Автоматизация построения объектов интеграции особенно актуальна при работе с Большими данными и позволяет строить системы, более устойчивые к модификациям в процессе их эксплуатации.

Подход, изложенный в статье, применялся к задачам, совершенно разным как по семантике, так и по формам представления данных, что позволяет говорить о достаточной его универсальности.

СПИСОК ЛИТЕРАТУРЫ

1. *Hopkins B., Evelson B.* Expand your digital horizon with Big Data. URL: http://www.asterdata.com/newsletter-images/30-04-2012/resources/forrester_expand_your_digital_horiz.pdf, 2011.
2. URL: <http://nosql-database.org/>.
3. URL: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html.
4. URL: <http://hadoop.apache.org/>.
5. URL: <https://hana.sap.com/abouthana.html>.
6. *Когаловский М.Р.* Методы интеграции данных в информационных системах // Институт проблем рынка РАН, Москва, 2010. URL: <http://www.ipr-ras.ru//articles/kogalov10-05.pdf>.
7. URL: <https://www.w3.org/TR/2004/REC-owl-features-20040210/>.

8. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. СПб.: Питер, 2001. 384 с.

9. *Buneman P.* Semistructured data // Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Tucson, Arizona, United States, May 11–15, 1997. P. 117–121.

10. *Биряльцев Е.В., Гусенков А.М., Косинов Я.Г.* Представление структуры реляционных баз данных в формализме онтологий // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2006. Казань: Изд-во «Отечество», 2007. С. 32–37.

11. *Биряльцев Е.В., Гусенков А.М.* Интеграция реляционных баз данных на основе онтологий. // Ученые записки Казанского государственного университета. Серия Физико-математические науки. 2007. Т. 149, Кн. 2. С. 13–25.

12. *Гусенков А., Биряльцев Е., Жибрик О.* Интеллектуальный поиск в структурированных массивах информации // LAP LAMBERT Academic Publishing, 2015. 129 с.

13. *Гарсиа-Молина Г., Ульман Дж., Уидом Дж.* Системы баз данных. Полный курс / Database Systems: The Complete Book // Вильямс, 2003. 1088 с.

14. *Буч Г., Рамбо Д., Джекобсон А.* Язык UML. Руководство пользователя. Пер. с англ. М.: ДМК, 2000. 432 с.

15. *Биряльцев Е.В., Гусенков А.М., Галимов М.Р.* Особенности лексико-семантической структуры наименований артефактов реляционных баз данных // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2005. Казань: Изд-во Казанского ун-та, 2006. С. 4–12.

16. *Жучков А.В., Арнауттов С.А., Твердохлебов Н.В.* Новые технологии для понятийных сетей, создаваемых в рамках МНТП «Вакцины нового поколения и диагностические системы будущего» // Электронные библиотеки, 2003. Т. 6, Вып. 6. URL: <http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part6/ZATGS>.

17. *Андерсон Дж.* Дискретная математика и комбинаторика. Пер. с англ. М.: Изд. дом «Вильямс», 2003. 960 с.

18. *Биряльцев Е.В., Гусенков А.М., Хайруллина А.И.* Представление модели данных Episcenter POSC на языке онтологий OWL // Труды Казанской школы по

компьютерной и когнитивной лингвистике TEL-2006. Казань: Изд-во «Отечество», 2007. С. 38–49.

19. *Биряльцев Е.В., Гусенков А.М.* Построение онтологии предметной области на основе логической модели баз данных // Труды Всерос. конф. с международным участием «Знания-Онтологии-Теории» (ЗОНТ-07). Новосибирск: Ин-т математики им. С.Л. Соболева СО РАН, 2007. Т. 1. С. 176–183.

20. URL: <http://www.energistics.org/energistics-standards-directory/epicentre-archive>.

21. URL: <https://www.w3.org/OWL>.

22. *Fellbaum C. (ed.)* WordNet: An electronic lexical database. Cambridge: MIT Press, 1998. 423 p.

23. *Биряльцев Е.В., Гусенков А.М.* Онтологии реляционных баз данных. Лингвистический аспект. // Тр. межд. конф. Диалог'2007, «Бекасово», 30 мая – 3 июня 2007 г. М.: Изд. центр РГГУ, 2007. С. 50–53.

24. URL: http://www.dialog-21.ru/Archive/2001/volume2/2_21.htm.

25. *Биряльцев Е.В., Гусенков А.М., Миронов С.В.* Один подход к реализации нерегламентированного доступа к реляционным базам данных // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2008. Казань, 10–13 декабря 2008 г. Казань: Изд-во Казанского ун-та, 2009. С. 10–23.

26. *Misutka J., Galambos L.* Extending full text search engine for mathematical content // Proceedings of DML. 2008. P. 55–67.

27. URL: <http://lucene.apache.org>.

28. *Биряльцев Е.В., Галимов М.Р., Гусенков А.М., Жибрик О.Н.* Некоторые подходы к повышению релевантности поиска математических выражений в естественнонаучных текстах // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2012. Казань, 25–28 января 2012 г. Казань: Изд-во Фэн Академии наук РТ, 2012. С. 78–93.

29. *Биряльцев Е.В., Гусенков А.М., Жибрик О.Н.* Некоторые подходы к разметке естественнонаучных текстов, содержащих математические выражения // Ученые записки Казанского университета. 2014. Т. 156, Кн. 4. С. 133–148.

30. Биряльцев Е.В., Галимов М.Р., Жильцов Н.Г., Невзорова О.А. Подход к семантическому поиску математических выражений в научных текстах // Материалы межд. науч.-техн. конф. OSTIS-2012. Минск: БГУИР, 2012. С. 245–256.

31. Nevzorova O., Zhiltsov N., Zaikin D., Zhibrik O., Kirillovich A., Nevzorov V., Biryaltsev E. Bringing Math to LOD: a semantic publishing platform prototype for scientific collections in Mathematics // The SemanticWeb – ISWC 2013. 12th Int. SemanticWeb Conference. Sydney, NSW, Australia, October 21–25, 2013. Springer, Lecture Notes in Computer Science, 2013. V. 8218. P. 379–394.

32. URL: <http://www.w3.org/TR/rdf-sparql-query/>.

33. Биряльцев Е.В., Гусенков А.М., Жибрик О.Н. Поиск математических выражений в естественно-научных текстах. Экспериментальная оценка релевантности // Интеллект. Язык. Компьютер. Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. Казань: Изд-во Фэн Академии наук РТ, 2014. С. 34–37.

34. Биряльцев Е.В., Гусенков А.М., Елизаров А.М. О доступе к электронным коллекциям в виде реляционных баз данных на основе онтологий // Труды 9-й Всерос. научн. конф. Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2007, Переславль-Залесский, Россия, 15–18 октября 2007 г. Переславль-Залесский, Ярославль: Изд-во Университет города Переславля, 2007. С. 211–216

35. Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., Solovyev V.D., Zhiltsov N.G. Mathematical knowledge representation: semantic models and formalisms // Lobachevskii Journal of Mathematics. 2014. V. 35, No 4. P. 347–353.

36. Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Липачев Е.К., Невзорова О.А., Соловьев В.Д. Методы анализа семантических данных математических электронных коллекций // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2014. № 4. С. 12–16 (Biryaltsev E.V., Elizarov A.M., Zhiltsov N.G., Lipachev E.K., Nevzorova O.A., Solov'ev V.D. Methods for analyzing semantic data of electronic collections in mathematics // Automatic Documentation and Mathematical Linguistics. 2014. V. 48, No 2. P. 81–85).

37. *Елизаров А.М., Липачёв Е.К., Хохлов Ю.Е.* Семантические методы структурирования математического контента, обеспечивающие расширенную поисковую функциональность // Информационное общество. 2013. № 1–2. С. 83–92.

38. *Елизаров А.М., Липачев Е.К., Невзорова О.А., Соловьев В.Д.* Методы и средства семантического структурирования электронных математических документов // Докл. РАН. 2014. Т. 457, № 6. С. 642–645.

39. *Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Веб-технологии для математика: основы MathML. Практическое руководство. М.: Физматлит, 2010. 192 с.

40. *Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Сервисы электронных естественнонаучных коллекций, построенные на основе технологии MathML // Труды Всероссийской суперкомпьютерной конф. «Научный сервис в сети Интернет: суперкомпьютерные центры и задачи», г. Новороссийск, 20–25 сентября 2010 г. М.: Изд-во Московского ун-та, 2010. С. 533–534.

41. URL: <http://gate.ac.uk/>.

INTELLIGENT SEARCH OF COMPLEX OBJECTS IN BIG DATA

A.M. Gusenkov¹

Institute of Computational Mathematics and Information Technologies.

Kazan (Volga Region) Federal University

¹ gusenkov.a.m@gmail.com

Abstract

This article considers approach to intelligent search of complex objects in different types of texts with structural markup which can be used for Big Data processing. We research two types of data entry: relational databases, which use their schemes as structural markup, and full-text scientific documents containing mathematical expressions (formulae). For such full-text documents we suggest additory automated markup to allow formula search. In both cases we use natural language texts, which are semi-structured data, as data source for building ontology and conducting search at a later stage. For relational databases those are comments to table and table attribute names;

for scientific documents (articles, monographs, etc.) it is a text content of marked up documents.

Keywords: *Big Data, semantic search, semi-structured data, ontology, relational databases, science texts, mathematical expressions markup.*

REFERENCES

1. *Hopkins B., Evelson B.* Expand your digital horizon with Big Data. URL: http://www.asterdata.com/newsletter-images/30-04-2012/resources/forrester_expand_your_digital_horiz.pdf, 2011.
2. URL: <http://nosql-database.org/>.
3. URL: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html.
4. URL: <http://hadoop.apache.org/>.
5. URL: <https://hana.sap.com/abouthana.html>.
6. *Kogalovskij M.R.* Metody integracii dannyh v informacionnyh sistemah // Institut problem rynka RAN, Moskva, 2010. URL: <http://www.ipr-ras.ru/articles/kogalov10-05.pdf>.
7. URL: <https://www.w3.org/TR/2004/REC-owl-features-20040210/>.
8. *Gavrilova T.A., Horoshevskij V.F.* Bazy znaniy intellektual'nyh sistem. SPb.: Piter, 2001. 384 s.
9. *Buneman P.* Semistructured data // Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Tucson, Arizona, United States, May 11–15, 1997. P. 117–121.
10. *Birialtsev E.V., Gusenkov A.M., Kosinov Y.G.* Predstavlenie struktury relyacionnyh baz dannyh v formalizme ontologij // Trudy Kazanskoj shkoly po komp'yuternoj i kognitivnoj lingvistike TEL-2006. Kazan: Izdatel'stvo "Otechestvo", 2007. S. 32–37.
11. *Birialtsev E.V., Gusenkov A.M.* Integraciya relyacionnyh baz dannyh na osnove ontologij. // Uchenye zapiski Kazanskogo gosudarstvennogo universiteta. Seriya Fiziko-matematicheskie nauki. 2007. T. 149. Kn. 2. Kazan: Kazanskij universitet, 2007. S. 13–25.
12. *Gusenkov A., Birialtsev E., Zhibrik O.* Intellektual'nyj poisk v strukturirovannyh massivah informacii. // LAP LAMBERT Academic Publishing, 2015. 129 s. ISBN 978-3-659-76919-1.

13. *Garcia-Molina H., Ullman J., Widom J.* Database Systems: The Complete Book. Williams, 2003. 1088 s.

14. *Booch G., Rumbaugh J., Jacobson I.* The Complete UML Training Course // Per. s angl. Moskva: Izdatel'stvo DMK, 2000. 432 s.

15. *Birialtsev E.V., Gusenkov A.M., Galimov M.R.* Osobennosti leksiko-semanticheskoy struktury naimenovanij artefaktov relyacionnyh baz dannyh // Trudy Kazanskoj shkoly po komp'yuternoj i kognitivnoj lingvistike TEL-2005. Kazan: Izdatel'stvo Kazanskogo universiteta, 2006. S. 4–12.

16. *Zhuchkov A.V., Arnautov S.A., Tverdohlebov N.V.* Novye tekhnologii dlya ponyatijnyh setej, sozdavaemyh v ramkah MNTP «Vakciny novogo pokoleniya i diagnosticheskie sistemy budushchego» // Ehlektronnye biblioteki. 2003. T. 6, Vyp. 6. URL: <http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part6/ZATGS>.

17. *Anderson J.* Discrete Mathematics with Combinatorics. Per. s angl. Moskva: Izd. dom Williams, 2003. 960 s.

18. *Birialtsev E.V., Gusenkov A.M., Khairullina A.I.* Predstavlenie modeli dannykh Epicenter POSC na iazyke ontologii OWL// Trudy Kazanskoj shkoly po kompiuternoj i kognitivnoj lingvistike TEL-2006. Kazan: Izdatel'stvo "Otechestvo", 2007. S. 38–49.

19. *Birialtsev E.V., Gusenkov A.M.* Postroenie ontologii predmetnoj oblasti na osnove logicheskoy modeli baz dannyh // Trudy Vseros. konf. s mezhdunarodnym uchastiem "Znaniya-Ontologii-Teorii" (ZONT-07). Novosibirsk: In-t matematiki im. S.L. Soboleva SO RAN, 2007. T. 1. S. 176–183.

20. URL: <http://www.energistics.org/energistics-standards-directory/epicentre-archive>.

21. URL: <https://www.w3.org/OWL/>.

22. *Fellbaum C. (ed.)* WordNet: An Electronic Lexical Database. Cambridge: MIT Press, 1998. 423 p.

23. *Birialtsev E.V., Gusenkov A.M.* Ontologii relyacionnyh baz dannyh. Lingvisticheskij aspekt. // Tr. mezhdunar. konf. Dialog'2007, Bekasovo, 30 maya – 3 iyunya 2007 g. Moskva: Izd. centr RGGU, 2007. S. 50–53.

24. URL: http://www.dialog-21.ru/Archive/2001/volume2/2_21.htm.

25. *Birialtsev E.V., Gusenkov A.M., Mironov S.V.* Odin podhod k realizacii nereglamentirovannogo dostupa k relyacionnym bazam dannyh // Trudy Kazanskoj shkoly

po komp'yuternoj i kognitivnoj lingvistike TEL-2008. Kazan, 10–13 dekabrya 2008 g. Kazan: Izdatel'stvo Kazanskogo universiteta, 2009. S. 10–23.

26. *Misutka J., Galambos L.* Extending full text search engine for mathematical content // Proceedings of DML. 2008. P. 55–67.

27. URL: <http://lucene.apache.org>.

28. *Birialtsev E.V., Galimov M.P., Gusenkov A.M., Zhibrik O.N.* Nekotorye podhody k povysheniyu relevantnosti poiska matematicheskikh vyrazhenij v estestvennonauchnyh tekstah // Trudy Kazanskoj shkoly po komp'yuternoj i kognitivnoj lingvistike TEL-2012. Kazan, 25–28 yanvaryaya 2012 g. Kazan: Izdatel'stvo "Fehn" Akademii nauk RT, 2012. S. 78–93.

29. *Birialtsev E.V., Gusenkov A.M., Zhibrik O.N.* Nekotorye podhody k razmetke estestvennonauchnyh tekstov, sodержashchih matematicheskie vyrazheniya. Uchenye zapiski Kazanskogo universiteta, Kazan: Kazanskij universitet, 2014. T. 156, kn. 4. S. 133–148.

30. *Birialtsev E.V., Galimov M.P., Zhiltsov N.G., Nevzorova O.A.* Podhod k semanticheskomu poisku matematicheskikh vyrazhenij v nauchnyh tekstah // Materialy mezhdunarodnoj nauchno-tekhnicheskoj konferencii OSTIS-2012. Minsk: BGUIR, 2012. S. 245–256.

31. *Nevzorova Olga, Zhiltsov Nikita, Zaikin Danila, Zhibrik Olga, Kirillovich Alexander, Nevzorov Vladimir, Birialtsev Evgeniy.* Bringing Math to LOD: a semantic publishing platform prototype for scientific collections in Mathematics // The SemanticWeb – ISWC 2013. 12th International SemanticWeb Conference Sydney, NSW, Australia, October 21–25, 2013. Lecture Notes in Computer Science 8218. Springer. P. 379-394.

32. URL: <http://www.w3.org/TR/rdf-sparql-query/>.

33. *Birialtsev E.V., Gusenkov A.M., Zhibrik O.N.* Poisk matematicheskikh vyrazhenij v estestvenno-nauchnyh tekstah. Eksperimental'naya ocenka relevantnosti // Intel'lekt. Yazyk. Komp'yuter. Trudy Kazanskoj shkoly po komp'yuternoj i kognitivnoj lingvistike TEL-2014. Kazan: Izdatel'stvo "Fehn" Akademii nauk RT. 2014. S. 34-37.

34. *Birialtsev E.V., Gusenkov A.M., Elizarov A.M.* O dostupe k ehlektronnym kollekcijam v vide relyacionnyh baz dannyh na osnove ontologij // Trudy 9-j Vseros. nauchn. konf. Elektronnye biblioteki: perspektivnye metody i tekhnologii, ehlektronnye

kollekcii – RCDL-2007, Pereslavl'-Zalesskij, Rossiya, 15–18 oktyabrya 2007 g. Pereslavl'-Zalesskij, Yaroslavl': Izd-vo Universitet goroda Pereslavlya, 2007. S. 211-216.

35. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., Solovyev V.D., Zhiltsov N.G.* Mathematical knowledge representation: semantic models and formalisms // *Lobachevskii Journal of Mathematics*. 2014. V. 35, No 4. P. 347–353.

36. *Birialtsev E.V., Elizarov A.M., Zhiltsov N.G., Lipachev E.K., Nevzorova O.A., Solovyev V.D.* Methods for analyzing semantic data of electronic collections in mathematics// *Automatic Documentation and Mathematical Linguistics*. 2014. V. 48, No 2. P. 81–85.

37. *Elizarov A.M., Lipachev E.K., Hohlov Yu.E.* Semanticheskie metody strukturirovaniya matematicheskogo kontenta, obespechivajushhie ras-shirennuju poiskovuju funkcional'nost'// *Informacionnoe obshhestvo*. 2013. № 1–2. S. 83–92.

38. *Elizarov A.M., Lipachev E.K., Nevzorova O.A., Solovyev V.D.* Metody i sredstva semanticheskogo strukturirovaniya ehlektronnyh matematicheskikh dokumentov // *Dokl. RAN*. 2014. T. 457, № 6. S. 642–645.

39. *Elizarov A.M., Lipachev E.K., Malakhal'tsev M.A.* Veb-tehnologii dlya matematika: osnovy MathML. Prakticheskoe rukovodstvo. M.: FIZMATLIT, 2010. 192 s.

40. *Elizarov A.M., Lipachev E.K., Malakhal'tsev M.A.* Servisy ehlektronnyh estestvennonauchnyh kollekcij, postroennye na osnove tehnologii MathML // *Trudy Vserossijskoj superkomp'yuternoj konf. Nauchnyj servis v seti Internet: superkomp'yuternye centry i zadachi*, g. Novorossiysk, 20–25 sentyabrya 2010 g. M.: Izd-vo Mosk. un-ta, 2010. S. 533–534.

41. URL: <http://gate.ac.uk/>.

СВЕДЕНИЯ ОБ АВТОРЕ



ГУСЕНКОВ Александр Михайлович – старший преподаватель Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета.

Alexander Mikhailovich GUSENKOV – senior lecturer, Institute of Computational Mathematics and Information Technologies of Kazan Federal University. Current scientific interests: knowledge extraction technologies, big data, data mining, parallel computing.
email: gusenkov.a.m@gmail.com

Материал поступил в редакцию 24 декабря 2015 года