УДК 004.4

ФОРМИРОВАНИЕ И ПОДДЕРЖКА ФИЗИКО-МАТЕМАТИЧЕСКИХ ЭЛЕКТРОННЫХ НАУЧНЫХ ИЗДАНИЙ: ПЕРЕХОД НА ТЕХНОЛОГИИ СЕМАНТИЧЕСКОГО ВЕБА

В. Г. Веселаго, А. М. Елизаров, Е. К. Липачёв, М. А. Малахальцев

Аннотация

Представлены результаты исследований в области научных электронных библиотек и электронного книгоиздания, проведенных за последнее пятилетие с участием авторов при выполнении ряда проектов Российского фонда фундаментальных исследований (РФФИ) и Научной электронной библиотеки (НЭБ) E-library.ru. Приведены примеры практического применения технологий семантического веба в электронных научных коллекциях. Изложены подходы к формированию и поддержке физико-математических электронных научных изданий на основе XML, RDF и других технологий семантического веба, в частности, предложены решения ряда проблем, связанных с генерацией метаданных, организацией хранения и поиска данных в электронных научных журналах.

1. Введение

Одной из черт происходящего сегодня перехода от индустриального общества к информационному является то, что информация и знание становятся одним из основных факторов развития. Сохранение, развитие и рациональное использование этого стратегического ресурса будущего имеют огромное значение для любого общества и государства. Отличительной чертой современного общества является представление информации и знаний не только в традиционной печатной, но и в электронной, цифровой форме, что позволяет принципиально иначе создавать, хранить, организовывать доступ и использовать информацию. Кроме того, современные информационно-коммуникационные технологии (ИКТ) привели к тому, что сегодня большинство современных информационных ресурсов сразу создается в электронном виде, т. е. формируются электронные библиотеки (ЭБ) (см., например, [1]).

В условиях формирования информационного общества чрезвычайно важным инструментом устойчивого экономического и социального развития является обеспечение публичного (в том числе удаленного) доступа к социально значимой информации, в первую очередь научного, образовательного и культурного характера. Общепризнано, что степень доступности информационных ресурсов России, имеющих огромную ценность и столь же огромные объемы, до сих пор остается слишком низкой. Поэтому стала очевидной необходимость эффективной кооперации российских научных учреждений, информационных центров, библиотек, архивов, музеев и других учреждений, занимающихся созданием и распространением информации и знания. В результате сформировалось новое, в определенной степени синтетическое направление деятельности — электронные библиотеки, объединяющее специалистов в области информационных технологий, библиотекарей, работников музеев и архивов, издателей, теле– и радиовещателей и многих других. В этой деятельности востребован весь спектр ключевых технологий управления информацией,

которые используются в современных информационных системах. Поэтому тенденции развития ИКТ, сформировавшиеся в последние годы, существенно влияют на функциональные возможности электронных библиотек.

В наши дни, когда большая часть информации поступает, хранится и перерабатывается в электронном виде, заметные изменения произошли и в сфере научного обмена. В результате современного исследователя невозможно представить без компьютера, который из средства набора текстов давно стал инструментом получения знаний. Поскольку электронная форма представления информации стала основной, для математиков проблема представления и обработки математических текстов в электронной форме приобрела особую актуальность. В значительной степени эта проблема касается представления математических формул. Перевод математических документов в pdf–формат или html–файлы, когда каждая формула является ссылкой на графический ресурс, делает структурную обработку такой информации крайне затруднительной. Поэтому наиболее распространенное на данный момент решение — представление формул в виде графических файлов — неудовлетворительно с точки зрения структурной обработки математических текстов.

Международная организация World–Wide Web Consortium (сокращенно W3C, см. www.w3.org), разрабатывающая технологии глобальной сети, предложила новую концепцию развития интернета — Semantic Web (семантический веб), направленную на изменение основных принципов функционирования всемирной сети. Главная цель этой концепции — обеспечение машинного управления информационным пространством. В своей основополагающей работе [2] (Дорожная карта семантического веба) Тим Бернерс-Ли утверждает: «Веб разрабатывался как информационное пространство, полезное не только для коммуникации человека с человеком, но и как пространство, в котором эффективное содействие могут оказывать также и машины; ... подход семантического веба базируется на разработке языков для выражения информации в форме, пригодной для машинной обработки».

Таким образом, сегодня актуальным является использование технологий семантического веба как при формировании ЭБ, так и в системе электронного книгоиздания. Обсуждению этих вопросов посвящена настоящая работа.

2. Технологии управления информационными ресурсами и электронные библиотеки

Для современного информационного общества характерны стремительное развитие и активное использование таких ИКТ, которые обеспечивают не только сетевой информационный обмен, но и возможность интеграции локальных информационных ресурсов в единое общедоступное информационное пространство. Эти ресурсы существенно влияют на интенсивность процессов обучения и научных исследований, поэтому обеспечение доступа к ним стало одной из первоочередных задач информационного обслуживания образования, науки и культуры. Сегодня общепризнанно, что наиболее эффективный путь решения этой задачи связан с созданием электронных библиотек — «распределенных информационных систем, позволяющих надежно сохранять и эффективно использовать разнообразные коллекции электронных документов (текст, графика, аудио, видео и др.), доступные в удобном для конечного пользователя виде через глобальные сети передачи данных» [3]. Подразумевается, что электронная библиотека должна обеспечивать хранение информации в цифровой форме практически неограниченное время, а также расширение доступа к уникальным изданиям и предоставление всем заин-

тересованным потребителям качественно новых возможностей работы с большими объемами информации. К таковым, например, можно отнести последовательный, выборочный или параллельный просмотр множества документов; многоаспектный поиск во всем объеме информации, хранимой в данной ЭБ, с использованием как формальных признаков, так и лексики естественного языка; копирование необходимых документов или их фрагментов как на бумагу, так и на современные носители; создание собственных документов и, наконец, производство нового знания. Создание и использование ЭБ реализуются путем накопления, хранения, учета и структурирования электронной информации; организации навигации во всем информационном пространстве, доступном через электронную библиотеку; обеспечения эффективного доступа к ней любого числа пользователей по телекоммуникационным сетям, а также обучения пользователей. В этом смысле университетская среда является наиболее оптимальной для использования существующих и создания новых информационных ресурсов, для развития новых ИКТ, так как именно в университетах одновременно и в различных формах, в обучении и научных исследованиях создаются и используются такие информационные ресурсы и технологии. Создаваемые в университетах информационные ресурсы имеют различную природу — это научные издания и учебно-методические пособия, диссертации и их авторефераты, библиографические указатели и обзоры, справочная литература, материалы теле– и видеоконференций, электронные журналы и электронные версии «бумажных» научных изданий, электронные учебники, научные базы данных и еще многое другое. Создание единой среды и механизма формирования и использования этих ресурсов, а также единый их учет и классификация — актуальные проблемы, еще не нашедшие своего решения.

Информационные системы, которые стали называть электронными библиотеками, появились в 1990–х годах благодаря широкому распространению интернета, стремительному росту емкости запоминающих устройств, а также значительным достижениям в области веб–технологий, технологий баз данных и документальных систем. Эти предпосылки обеспечили возможности создания больших коллекций документов, в первую очередь, публикаций в электронной форме и массового к ним доступа. Нужно заметить, что термин «электронная библиотека» имеет отечественное происхождение. Он стал широко использоваться в нашей стране как эквивалент англоязычного термина «Digital Library». Употребление этого термина в русскоязычной литературе быстро устоялось, хотя оно неточно отражает смысл англоязычного источника («цифровые библиотеки»), авторы которого стремились подчеркнуть определяющую роль цифровых технологий в информационных системах такого рода.

В настоящее время общепринятого или стандартизованного определения электронной библиотеки не существует. Из целого ряда известных определений наиболее адекватным нам представляется приведенное выше определение из [3]. Употребление слова «библиотека» в составе термина «электронная библиотека» весьма условно и не означает необходимости соотносить его с содержанием термина «библиотека», закрепленным в ГОСТе. Создание распределенной ЭБ означает, что электронные документы или их коллекции, имеющиеся в распоряжении библиотек, архивов, информационных центров, издательств, музеев и иных фондодержателей, сохраняя автономию, интегрируются в единую распределенную среду, дополняя и взаимообогащая друг друга.

Известны и другие определения понятия «электронная библиотека», например: электронная библиотека состоит исключительно из цифровых материалов и услуг и исключает аналоговые материалы типа видеозаписей; все материалы обработаны и перемещены через цифровые устройства и сети [4]; ЭБ — «это сосредоточенная

совокупность цифровых объектов, включающих текст, видео, аудио наряду с методами доступа и поиска информации, а также выбора, организации и обслуживания их совокупности» [5].

К достоинствам электронных библиотек можно отнести то, что они способны обеспечить миграцию электронной информации в условиях постоянного развития вычислительной техники и программных средств, а также усовершенствованный сервис, который уже сейчас доступен пользователями, к примеру: индивидуальное обслуживание пользователей, основанное на «пользовательских профилях», в которых отражены информационные потребности пользователей или имеется специальная система поиска информации, управляемая пользователем и облегчающая ему процесс принятия решения о релевантности запросу найденного документа; наличие кооперативной инфраструктуры, позволяющей группам пользователей производить индексирование и оценку документов, относящихся к определенным темам; осуществление информационного поиска по нескольким языкам или запросам в многоязычных базах данных, многоязычный интерфейс.

В той или иной форме идея электронной библиотеки уже давно работает во многих университетах и крупных библиотеках ведущих стран мира (см. [6 – 8]). Например, электронная «библиотека XXI века» создается в Японии соединением усилий Агентства по внедрению новых технологий, Национальной парламентской библиотеки, целого ряда министерств, более 20 библиотек и культурных центров. Библиотека Конгресса США реализует национальную программу создания электронных библиотек.

Интенсивные исследования и разработки в области ЭБ во многих странах были стимулированы объявленной в США в конце 1993 г. по инициативе Национального научного фонда (NSF), Агентства перспективных исследований Министерства обороны (DARPA) и Национального агентства по космическим исследованиям (NASA) программой Digital Libraries Initiative (DLI) (см. [9]), направленной на поддержку фундаментальных исследований в этой области. На первой фазе работ по этой программе, начавшейся в 1994 г., на конкурсных началах были предоставлены гранты шести крупным университетам США — Калифорнийскому университету в Беркли (создание прототипа электронной библиотеки для планирования окружающей среды на примере Калифорнии); Калифорнийскому университету в Санта-Барбара (электронная библиотека Александрия, оперирующая с пространственно-индексированной и графической информацией); Университету Карнеги-Меллон (электронная библиотека Информедиа, использующая цифровые видеоресурсы); Иллинойскому университету в Урбана-Чемпейн (федеративные репозитории научной литературы); Мичиганскому университету (архитектура электронных библиотек с интеллектуальными агентами) и Стэнфордскому университету (интеграция механизмов неоднородных сервисов для обеспечения унифицированного доступа к различным сетевым коллекциям информационных ресурсов) [10]. К 1998 г., когда началась вторая фаза работ по программе [11, 12] и было предоставлено уже 24 гранта, был достигнут значительный прогресс в этой области. На второй стадии развития в начале 1998 г. эти программы были объединены в единую межведомственную программу (DLI-Phase 2), в которой, кроме названных, участвуют Национальная медицинская библиотека, Агентство по статистике США, Национальный гуманитарный фонд, Национальный архив США и другие федеральные агентства.

С 1995 по 2000 гг. осуществлялась национальная программа Великобритании eLib. В других странах (Канада, Германия и др.) многочисленные разрозненные проекты в последние годы также стали превращаться в национальные и международные программы создания электронных библиотек. Ряд проектов по созданию и использованию ЭБ выполнялся в рамках 4-й Рамочной Программы Комиссии

Европейских сообществ (КЕС) и программы «Технологии информационного общества» 5-й Рамочной Программы КЕС.

В России идеи создания электронных библиотек приняты многими университетами, академическими научными центрами и крупными библиотеками (см., например, [13 – 16]). Вместе с тем актуальность названных проблем для университетского сообщества столь значительна, что работы в этих направлениях ведутся практически во всех университетских центрах страны.

В 1998 г. РФФИ и Российский фонд технологического развития начали пилотную программу по созданию и использованию электронных библиотек. В 1999 г. в соответствии с решениями Правительства России была поставлена задача развертывания полномасштабной межведомственной программы по созданию и использованию ЭБ с участием большого числа министерств и ведомств.

Для поддержки исследований и разработок в области ЭБ были организованы также научные форумы, которые обеспечили контакты и обмен идеями в сообществе ученых и специалистов, работающих в рассматриваемой области. С 1995 г. IEEE (Institute of Electrical and Electronics Engineers) стала ежегодно проводиться конференция по перспективам развития электронных библиотек (IEEE Advances in Digital Libraries Conference, ADL). Начиная с 1996 г., аналогичная международная конференция по электронным библиотекам проводится под эгидой ACM (Association for Computing Machinery, ACM International Conference on Digital Libraries, ACM DL). Ежегодно, начиная с 1997 года, проводится также представительная конференция по этой тематике в Европе (European Conference on Digital Libraries, ECDL). Тематика всех этих конференций весьма широка и включает как теоретические проблемы реализации систем электронных библиотек, так и организационные, правовые и другие вопросы, связанные с практическими их разработками. Отметим, что с 1999 г. ежегодно проводится Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL), материалы последней (девятой) конференции см. в [17].

Кроме того, издается ряд специализированных периодических изданий по проблематике электронных библиотек. Наиболее авторитетные из них выпускаемый ежеквартально с 1997 г. издательством Springer-Verlag (Германия) Международный журнал по электронным библиотекам (International Journal on Digital Libraries, JODL); интерактивный журнал DISC (Digital Symposium Collection), выпускаемый с 1999 г. ACM SIGMOD. а также ежемесячный интерактивный журнал D-Lib Magazine, выпускаемый в США с 1995 г. компанией Corporation for National Research Initiatives.

3. Научные электронные библиотеки и цели их создания

В условиях быстрого развития и внедрения современных ИКТ во всех областях фундаментальных научных исследований создание научных ЭБ, позволяющих обеспечить быстрый и эффективный доступ к знанию, рассредоточенному в разных странах и различных хранилищах информации, является чрезвычайно важным. В ведущих зарубежных странах уже развернута активная деятельность по реализации программ и проектов создания и использования научных ЭБ на национальном уровне. С появлением политематических и предметно-ориентированных ЭБ значительно возросла актуальность проблемы семантического моделирования и обеспечения поиска вербальной и невербальной информации. Появились электронные коллекции и библиотеки научных документов, включающих, кроме вербальной информации, математические и химические формулы, таблицы, карты, схемы, рисунки, диаграммы, данные для научных исследований.

Целями создания научных ЭБ могут быть: обеспечение научным работникам возможности быстрого доступа к необходимым информационным ресурсам; предоставление результатов фундаментальных научных исследований широкому исследовательскому сообществу; создание новых технологий научных исследований и эффективного инструментария для их проведения; предотвращение утраты ценных научных коллекций для будущих поколений ученых; обеспечение возможностей для научного сотрудничества не только в региональном, ведомственном, национальном, но и в международном масштабе.

Состояние разработок научных ЭБ на современном этапе характеризуется следующими особенностями [18 – 20]:

по их масштабу — от поддержки отдельного периодического издания или группы изданий, ЭБ отдельного научного (образовательного) учреждения, до электронной библиотеки региона, ведомства, национальной/международной ЭБ;

по тематике информационных ресурсов — от универсальных до ЭБ для конкретной области знаний или научного направления, ресурса личных архивов крупных ученых;

по технологиям — от отдельных веб-сайтов, баз данных или систем текстового поиска до систем, основанных на интеграции технологий;

по архитектуре — от сосредоточенных до распределенных, от материализованных до виртуальных.

Естественно возникает вопрос, имеются ли у научных ЭБ особенности, требующие особого подхода к их разработке. Для ответа на него нужны критерии оценки, каковыми могут служить: пользователи и их информационные потребности, функциональные возможности ЭБ, характер и содержание информационных ресурсов; используемые методы и технологии. Выделим особенности научных ЭБ в соответствии с приведенными критериями.

Пользователи и их информационные потребности: «продвинутые» пользователи; нерегламентируемые информационные потребности; динамичный характер информационных потребностей; стремление пользователей к сотрудничеству; необходимость обмена информацией; многоязыковая среда научного сообщества; разнообразие предметов исследования и технологий научного исследования; демократичность научного сообщества.

Характер информационных ресурсов: неоднородность используемых информационных ресурсов в различных аспектах; использование данных на разных уровнях абстракции.

Функции научных электронных библиотек: хранение и актуализация коллекций информационных ресурсов; поддержка взаимосвязей информационных ресурсов; обеспечение доступа пользователей к коллекциям; поддержка метаданных, в том числе каталогов коллекций; формирование коллекций (оцифровка, регистрация результатов наблюдений и экспериментов непосредственно в процессе их проведения); интеграция издательских технологий и технологий формирования коллекций; поддержка различных сведений о пользователях; предоставление различных встроенных или надстроенных сервисов-приложений, превращающих ЭБ в исследовательский полигон (виртуальная обсерватория, виртуальная химическая или биологическая лаборатория и т. п.).

Содержание информационных ресурсов научных ЭБ: научные публикации в различных формах (статьи, доклады, монографии, диссертации, авторефераты и др.); библиографическая информация; персоналия; событийная информация (календарь конференций и т. п.); результаты различного рода экспериментов, наблюдений, измерений, моделирования исследуемой реальности; модели исследуемых

процессов, явлений, феноменов, представленные в разнообразных формах, и метаданные, описывающие такие ресурсы; разнообразные научные коллекции и их элементы; каталоги коллекций и описания их элементов, классификаторы и другие средства систематизации.

Свойства информационных ресурсов научных ЭБ: неоднородность информационных ресурсов (в различных аспектах); разнообразие сред представления — текст, числовые данные, статические изображения, видео, аудио, мультимедиа; сверхбольшие объемы информационных ресурсов; свойства информационных ресурсов описываются метаданными для системы и для пользователей; виртуальные коллекции информационных ресурсов; часто используются не реальные, а гипотетические данные; существенное значение имеет фактор старения информационных ресурсов; представление информационных ресурсов предметной области исследования в форме, позволяющей непосредственно проводить исследование (не информационная поддержка исследований, а исследовательский полигон).

4. Электронные научные журналы — проблемы развития и интеграции

Одним из элементов научных ЭБ, обеспечивающих формирование новых видов информационных ресурсов и обмен научной информацией на базе современных ИКТ, являются сегодня электронные научные журналы, в том числе и чисто электронные.

В конце 20-го века развитие ИКТ и средств представления информации привело к тому, что появились электронные версии научных журналов, чтение которых сначала было возможно лишь на отдельных компьютерах в локальных сетях, а вскоре и через интернет. Западные издатели стали предлагать своим подписчикам электронные версии журналов как дополнительную услугу. Например, только на электронные издания издательства Elsevier (www.elsevier.ru) зарегистрировано около 6500 подписчиков, из которых более двухсот из России и СНГ (см. http://lsl.ksu.ru/images/konf2006/ yakshonok2.pdf). Впоследствии электронные версии стали реализовываться независимо от их печатных аналогов.

В 1996 году в мире существовало всего около 250 электронных журналов, однако уже спустя два года их количество увеличилось до 2 тысяч. Сегодня издается, по разным оценкам, от 20 до 30 тысяч электронных научных журналов. Все крупнейшие западные издательства публикуют свои журналы в электронном виде. Появилось значительное количество электронных журналов открытого доступа (см., например, http://www.doaj.org/). Подавляющее большинство электронных журналов доступно через интернет и объединено в крупные базы данных. По данным издательства Elsevier электронный архив Science Direct включает около 2500 полнотекстовых электронных журнала, содержащих около 8 млн. статей.

Создание и широкое распространение электронных журналов стало возможным благодаря развитию технологий электронного книгоиздания, специализированных форматов (в основном на основе SGML и XML), средств телекоммуникаций и интернет, программных средств обработки данных. Появление электронных журналов обуславливалось еще рядом факторов, например, высокой стоимостью подписки на печатные версии журналов и ограничением распространения традиционных изданий. Из-за высокой стоимости печатных изданий многие журналы и книги не приобретаются российскими библиотечными учреждениями или доступны в регионах в одной - двух библиотеках; 75% объема рынка российских печатных СМИ, по оценкам Гильдии издателей периодической печати, приходится на Москву. Преобладающий объем столичного рынка обусловлен также тем, что издания

не могут пробиться в регионы из-за отсутствия системы распространения печатной периодики.

В области электронного книгоиздания Россия не является лидером. В 2005 году нами была проведена экспертиза почти тысячи журналов, входящих в список Высшей аттестационной комиссии (ВАК) РФ. Из них в интернете так или иначе присутствовало около 300 журналов. Детальное изучение сайтов показало, что большинство журналов представлено только в виде оглавлений с аннотациями. Часть журналов представлена 1 – 2 годами изданий, публикация которых в электронном виде была осуществлена несколько лет назад. Количество «действующих» электронных журналов из списка ВАК (т. е. тех, которые были опубликованы в 2004 – 2005 гг.) не превышало 100 наименований. Практически на всех сайтах отсутствует возможность поиска по авторам, названиям статей, ключевым словам, аннотациям и особенно — по полным текстам статей. Тексты представлены в разных форматах — html, pdf, DjVu, doc и т. д. Недостаточное развитие российских электронных научных журналов в целом снижает рейтинг российских изданий и приводит к тому, что ученые стали отдавать предпочтение публикации своих работ в западных журналах. Отметим несколько причин недостаточного развития электронного книгоиздания.

Прежде всего, не разработаны форматы представления электронных изданий, обеспечивающие структурирование текстов статей и учет этой структуры для загрузки журналов в базы данных. Форматы записей, давно и успешно применяемые в библиотечном деле, такие, как US Marc, Unimarc или Rusmarc, предназначены лишь для описания печатных источников (книг, журналов в целом, статей) на библиографическом уровне. Для описания полных текстов журналов или отдельных публикаций эти форматы не могут быть использованы из-за отсутствия необходимых спецификаций. Поскольку библиотечная каталогизация направлена только на описание печатных источников, то указанные форматы (при всей их детализации) зачастую не включают описания таких элементов, как, например, подробные сведения об индивидуальных авторах (ученое звание, место работы, почтовый и электронный адреса). Ни один из библиотечных форматов не поддерживает описаний полного текста статьи из журнала, главы из книги или текста книги целиком. Полностью отсутствует описание (тем более детализированное) библиографических списков, что весьма важно при построении индексов научного цитирования. Точно так же не могут применяться описания, построенные на языках метаописания, подобных Dublin Core (DC) (см. http://purl. oclc.org/dc/), используемых в основном для библиографического описания интернет-страниц. Поэтому требуется разработка форматов, предназначенных специально для описания полнотекстовых электронных изданий.

Кроме того, отсутствует специализированное программное обеспечение, позволяющее производить углубленную обработку электронных версий журналов с целью их последующей загрузки в базы данных. Применяется достаточно много издательских пакетов программ (PageMaker, Word и др.), однако почти все они предназначены лишь для набора текста, оформления и создания макетов изданий, с которых производится типографское тиражирование. Использование таких макетов для загрузки электронных изданий в базы данных не предусматривается и не может быть выполнено. Необходима разработка программно-технологических комплексов, предназначенных для структурирования (разделения на поля) электронных версий печатных изданий. При разработке следует учитывать особенности представления специализированных текстов, например, по математике.

Издательства и редакции не всегда идут на публикацию электронных изданий. Возможно, это связано с опасением утратить подписку на печатный вариант из-

дания. Кроме того, юридическая база электронных публикаций недостаточно проработана и, как следствие, традиционные (печатные) варианты публикаций пока считаются более привлекательными. Необходимым условием функционирования электронного журнала является признание научным сообществом равноценности электронной и традиционной (в «бумажном» научном журнале) публикаций. Это накладывает на редакции электронных журналов ту же ответственность, что и в любом традиционном научном журнале, в частности, по организации независимого научного рецензирования.

Необходимо предложить владельцам журналов разумные схемы взаимодействия с электронными библиотеками, включая подготовку пакетов юридических документов по прямому лицензированию. Кроме того, необходимо подготовить юридическую документацию, регулирующую проблемы авторского права применительно к электронным публикациям — на уровне «автор – редакция» или «автор – издательство».

Таким образом, актуальной является задача включения российского научного сообщества в уже сложившуюся научную общемировую систему электронных научных публикаций. Вместе с тем, как отмечено в [21], издание электронных научных журналов в России является скорее инициативой отдельных научных или образовательных организаций, не связанной с предпринимательской деятельностью, чем развитием системы электронного книгоиздания в стране. Несмотря на это, многие академические научные и образовательные организации (научно-исследовательские институты и университеты), традиционно издающие бумажные научные журналы, нашедшие своего читателя и пользующиеся заслуженным авторитетом, создают электронные версии своих изданий, выставляя их в открытый доступ сразу после выпуска бумажного экземпляра или с некоторой задержкой, а иногда распространяя электронную версию также по подписке. Примером могут служить многие авторитетные журналы, издаваемые Российской академией наук, в частности, математические журналы (см. раздел «Математические ресурсы России» на сайте http://libserv.mi.ras.ru/journ_RAN.html).

Особую группу электронных научных журналов составляют чисто электронные издания, не имеющие бумажных версий. Сегодня существует ярко выраженная потребность в организации таких журналов, прежде всего, потому, что такие издания позволяют быстро публиковать поступающие статьи (т. е. обеспечивают оперативность публикаций); дают возможность оперативного ознакомления с публикуемыми научными материалами (сразу после принятия этих материалов в печать) самой широкой аудитории при самой широкой географии охвата; публикуемые материалы, как правило, не ограничиваются по объему, их доступность широкой аудитории определяется лишь доступностью интернет для читателя.

Кроме того, в чисто электронных научных журналах резко ускорен и упрощен весь цикл подготовки, пересылки и рецензирования статей, а их издание существенно дешевле издания бумажных журналов, так как оно исключает все типографские проблемы. Однако издание чисто электронных журналов неизбежно встречает целый ряд серьезных проблем, часть из которых подробно обсуждена в [22]. Там же сделан краткий обзор существующих в России чисто электронных научных журналов, проведен их анализ и указаны особенности. Этот анализ показал, что несмотря на отсутствие каких-либо стандартов в их организации практически все электронные журналы «устроены» одинаково: деятельность журналов лицензируется; для управления журналами создаются редколлегии, осуществляется рецензирование представляемых работ; все журналы представлены в интернете, причем часто не на единственном сайте, а на различных «зеркалах» и в научных электронных библиотеках (например, в НЭБ e-library.ru). Основные варианты представления жур-

нала: общая информация о журнале, библиография (оглавления по томам и номерам), сведения для авторов (правила представления работ) и полные тексты статей (чаще всего в формате *.pdf, который требует использования свободно распространяемого программного обеспечения Adobe Acrobat Reader). Общая информация о журнале, а также библиография (и иногда abstracts) обычно предоставляются свободно, а полные тексты статей — ограниченно (в пределах подписки на бумажные версии изданий или за отдельную плату), для чисто электронных журналов вся информация доступна полностью. Печатные издания чисто электронных научных журналов предназначены только для обязательной рассылки.

Как же найти интересующий нас электронный журнал? Информация о журналах чаще всего размещается на сайтах издательств, в рамках которых эти журналы выходят в свет, или научных обществ и организаций соответствующего направления. Поэтому простейший способ найти в интернете интересующий журнал — обратиться к серверу издающей его организации. Ссылки на ряд российских электронных журналов расположены также на странице «Электронные научные журналы» сайта БЕН РАН (http://www.benran.ru/).

Другой возможный источник информации — «Научная электронная библиотека» НЭБ e-library.ru, созданная на базе консорциума российских библиотек при поддержке РФФИ, которая включает несколько тысяч полнотекстовых научных журналов. Среди многих проектов в области электронных библиотек, которые были реализованы в России за последние годы, проект НЭБ занимает особое место. По объему электронного фонда и количеству читателей уже сегодня НЭБ может сравниться с крупнейшими мировыми электронными библиотеками, содержащими научную литературу. Пользователями библиотеки являются практически все ведущие научные организации и университеты страны.

В сотрудничестве с Казанским государственным университетом (КГУ) Научная электронная библиотека (НЭБ) уделяет большое внимание развитию российских электронных публикаций, решая, прежде всего, технологические, программные и юридические проблемы. В рамках грантов РФФИ и Министерства образования и науки было разработано несколько программно-технологических комплексов подготовки электронных журналов, с помощью которых только в 2005 – 2006 гг. в НЭБ были размещены более 200 российских научных журналов (см. [23]).

Далее, из анализа деятельности чисто электронных научных журналов выяснилось, что все они были правильно позиционированы среди многих других видов электронной научной информации. Эти журналы являются средством быстрой публикации новых научных результатов, причем в них осуществляется предварительное рецензирование статей. Некоторые журналы не делятся на выпуски, и статьи добавляются на сервер по мере их поступления. Это роднит такие журналы с так называемыми электронными архивами, но отличает от них именно рецензированием статей. Чисто электронные научные журналы хорошо дополняют существующие бумажные журналы, в том числе имеющие электронные версии. При организации полностью электронных научных журналов их учредители, как правило, не прибегали к сколько-нибудь массированной рекламе нового издания.

Заметный вклад в развитие электронного книгоиздания в России внесли ученые КГУ. Неотъемлемой стороной научно-исследовательской деятельности Казанской физико-математической школы является издание научных журналов и формирование электронных математических коллекций. В настоящее время в КГУ издаются:

«Lobachevskii Journal of Mathematics» (LJM) (http://ljm.ksu.ru) — первый отечественный электронный журнал по математике — учрежден в 1996 г. совместно с Отделением математики РАН; публикует с 1998 года работы по геометрии и топологии, алгебре, комплексному анализу, функциональному анализу, теории вероят-

ности и математической статистике, оптимальному управлению, теории алгоритмов. Статьи российских авторов занимают примерно половину объема журнала, также в журнале публикуются авторы из Армении, Испании, Канады, Марокко, Сербии, США, Финляндии, Японии. Журнал представлен на сервере Европейского Математического общества, реферируется в Mathematical Review, включен в базы данных Science Direct издательства Elsevier и НЭБ; информация о журнале регулярно появляется в Notices of AMS; с 2008 года LJM издается издательством Springer;

«Magnetic Resonance in Solids, Electronic Journal» (MRSej) (http://mrsej.ksu.ru); учрежден в 1996 году, публикует статьи, посвященные фундаментальным исследованиям в области магнитного резонанса в твердых телах и связанных с ним явлений; все статьи проходят рецензирование внутри журнала; MRSej является бесплатным как для авторов, также и для читателей; в основном издается на английском языке, хотя допустимы отдельные публикации и на русском языке (как правило, мемориального характера);

«Web Journal of Formal, Computational and Cognitive Linguistics» (FCCL) (http://fccl.ksu.ru); учрежден в 1997 году, является единственным электронным журналом по указанной проблематике и одним из пяти электронных журналов по лингвистике в мире; публикует оригинальные исследования по актуальным проблемам теоретической и прикладной лингвистики.

Остановимся теперь на результатах последнего пятилетия по созданию и реализации технологии автоматизированной обработки и включения в соответствующие базы данных электронных научных журналов, развиваемые сегодня в рамках проекта РФФИ 06-07-89132.

Одной из целей проведенных исследований была разработка полнофункциональной программной среды электронного научного журнала по математике, позволяющей автоматизировать ряд процессов, стандартных для научного издания (см. [24 – 27]). Современные физико-математические электронные журналы в своей работе придерживаются тех же высоких стандартов качества публикаций, что и классические математические журналы. В частности, статьи проходят научное рецензирование, что отличает электронный научный журнал от электронных коллекций, например, всемирно известного архива физических и математических препринтов xxx.lanl.gov. Это обстоятельство влияет на одно из главных преимуществ электронных журналов — скорость выхода публикации. Поэтому возникает задача такой организации информационных потоков в процессе функционирования электронного журнала, которая минимизирует временные задержки. Сокращение времени выхода публикации в электронном журнале по сравнению с обычным достигается не только за счет оперативного обмена информацией на шаге рецензирования, но, главным образом, в переходе на новые стандарты электронных публикаций, связанные с семантическим вебом.

5. Семантический веб

Как указано в программных документах консорциума W3C (см. W3C Semantic Web Activity Statement, http://www.w3.org/2001/sw/Activity), семантический веб — это «...расширение традиционного веба в направлении существенно лучшего определения смысла информации, позволяющего компьютерам и людям эффективнее выполнять совместную работу. Мы хотим, чтобы данные на вебе были определены и связаны ссылками так, чтобы их можно было легче находить, интегрировать, автоматизировать и повторно использовать в различных приложениях, ... чтобы данные были разделяемыми и могли обрабатываться как автоматизиро-

ванными средствами, так и людьми». Конечная амбициозная цель состоит в создании такой среды, где программные агенты могут динамически обнаруживать и опрашивать ресурсы, а затем взаимодействовать с ними. Такие агенты должны уметь справляться с возникающими виртуальными проблемами интеллектуализированной среды, обнаруживать новые факты и выполнять изощренные задания, получаемые от людей [28]. Для математического сообщества наибольший интерес представляет MathML (Mathematical Markup Language) — технология, предназначенная для представления математических формул. Разработка этой технологии ведется консорциумом W3C с 1999 года. Поскольку технологии семантического веба предоставляют новый способ организации веб-информации, который дает возможность на более высоком уровне решать задачи программной обработки документов (в частности, задачу поиска), MathML изменяет принципы организации и управления электронными публикациями по математике. В настоящее время язык MathML фактически стал стандартом представления математической информации в электронной форме в силу следующих причин:

технология обработки данных на основе языка MathML реализует одну из основных тенденций современной информатики — разделение разметки и данных, поэтому она представляет широкие возможности многоуровневого структурирования данных и расширенного поиска;

появилась возможность создания программного обеспечения, использующего технологию MathML;

созданы и продолжают совершенствоваться программные средства, позволяющие конвертировать в MathML документы, подготовленные с помощью имеющихся стандартных технологий (таких, например, как LaTeX, Mathematica, Maple, Word). Из наиболее распространенных инструментов отметим редактор WebEq (www.dessci.com/en/ products/webeq), конверторы TeX→ MathML: TtM (http:// hutchinson. belmont.ma.us/tth/mml/), tex4ht (входит в стандартный пакет MikTeX), а также конвертор XML → TeX(MikTeX). MathML поддерживается основными браузерами (просмотрщиками): Internet Explorer (при установке соответствующего плагина), Mozilla, Firefox;

технология MathML поддерживается системами Maple® и MathCAD 2001, а компания Wolfram Research предложила собственную концепцию использования технологии MathML, которая реализована в пакете Mathematica®, в частности, в этом пакете предусмотрено сохранение документов в формате MathML.

Переизбыток и, одновременно, недостаток информации — типичная ситуация, с которой сталкивается современный пользователь интернета. Повсеместное распространение компьютеров во все сферах деятельности и развитие интернета делают доступным все больший объем информации, поэтому отлаженные приемы обработки информации становятся менее эффективными. Поиск нужной информации стал серьезной проблемой: поисковые машины, индексирующие html-страницы, находят много ответов на запрос и охватывают большую часть всего веба, однако количество неудовлетворительных возвращаемых ответов велико, поскольку не существует понятия «правильности» ответов на запросы. Дело в том, что поиск основан на сравнении строки запроса со строками документов, но при этом никак не учитывается смысл информации, ради которой и был организован поиск. Такая ситуация возникла сравнительно недавно, когда развитие интернета привело к тому, что объемы информации, получаемой при поиске, стали намного превышать возможности человеческого восприятия. Отметим, что совокупность технологий первоначальныго интернета, получившая впоследствии название Web 1.0, была ориентирована только на формальное содержание документов (контент). В настоящее время используется набор технологий, который принято обозначать

Web 2.0 и который также оперирует в основном с контентом [2, 29].

Ведущие производители программного обеспечения, в числе которых Oracle, IBM, Adobe, Sun, Microsoft, Mozilla Inc., в качестве основного направления развития интернета на ближайшие годы разрабатывают новую систему, обозначенную как Web 3.0 и основанную на семантической обработке информации. Разработка этого комплекса технологий координируется консорциумом W3C. Особенность этой системы состоит в том, что программные модули (а не пользователи!), опираясь на метаданные и метабазы, осуществляют поиск информации по содержимому, включая поиск по видео- и цифровым изображениям. Основная задача Web 3.0 заключается в решении самой сложной проблемы развития интернета — поиска значимой информации, отделения её от информационного мусора. Семантический веб позволит рассматривать интернет в целом как глобальную базу данных (БД). Точно так же, как разработчик может запрашивать сведения из обычной БД и создавать приложения, оперирующие этой информацией, любой человек получит возможность собирать данные во всей интернет-сети и в соответствии со своими нуждами строить приложения, обрабатывающие взаимосвязанные, но разрозненные сведения из различных источников [30]. Это соответствует программному заявлению Т. Бернерса-Ли: «...основной ролью технологий семантического веба является интеграция данных, содержащихся в различных приложениях».

Название Semantic Web появилось в 2001 году в статье [31], аналогичное название получил проект консорциума W3C. В русскоязычной компьютерной литературе в последнее время используют термины «семантический веб», «семантическая сеть» и «семантическая паутина». Слово «семантика» дало название проекту и определило общее направление развития. Общее определение понятия «семантика» — это изучение значений. Слово «семантика» происходит от греческого semantikos, т. е. «важное значение». Компьютер должен понимать семантику документа в том смысле, что он не просто интерпретирует набор символов, содержащихся в документе, а выделяет смысл документа.

Имеется несколько определений понятия Semantic Web, наиболее подходящим из которых, на наш взгляд, является определение, приведенное в электронной энциклопедии Википедия (см. [32]): «Семантическая паутина — часть глобальной концепции развития сети Интернет, целью которой является реализация возможности машинной обработки информации, доступной во Всемирной паутине. Основной акцент концепции делается на работе с метаданными, однозначно характеризующими свойства и содержание ресурсов Всемирной паутины, вместо используемого в настоящее время текстового анализа документов. В семантической паутине предполагается повсеместное использование, во-первых, универсальных идентификаторов ресурсов (URI), а во-вторых — онтологий и языков описания метаданных». Предполагается, что семантика способна однозначно охарактеризовать найденный контент по ряду характерных признаков. Архитектура семантической сети предполагает наличие у любой информации, находящейся в сети, связанного с этой информацией точного смысла, который нельзя перепутать даже в случае совпадения фраз или слов, встреченных в разных контекстах. Фактически это означает, что любая информация связана с некоторым неотделимым от нее контекстом [33]. Семантический веб использует несколько основных технологий для выявления смысла данных.

Для трактовки данных он использует универсальный идентификатор ресурсов (URI). При этом URI рассматривается в более широком смысле — не только как ссылки на электронные адреса и веб-страницы, как в традиционной схеме, но и для обозначения любых объектов и ресурсов (люди, города, предметы и др.). Сейчас большая часть информации в сети совершенно не приспособлена для ком-

пьютерной обработки, поэтому не удалось создать программы, которые были бы способны разобраться в смысловой составляющей текста и, например, сгруппировать несколько текстов в одну общую категорию. В семантической паутине предлагается использовать форматы описания, доступные для машинной обработки и позволяющие решить эту задачу.

Для определения собственной структуры документов в семантической сети используют язык XML (eXtensible Markup Language), а для формализации метаданных, а также сведений о контексте — RDF (Resource Definition Framework). Для записи конструкций RDF можно использовать RDF/XML, являющийся подмножеством языка XML, а для описания структуры документов — RDF Schema. Для построения семантически связанной сети недостаточно только технологий XML и RDF, поэтому консорциумом W3C и был создан язык онтологии OWL (Ontology Web Language). Возможность OWL создавать онтологии играет ключевую роль в категоризации и классификации групп взаимосвязанных (related) данных [30].

Семантический веб создается как надстройка над уже существующими системами сетей, при этом поиск и обработка информации программируются как машиноориентированные. Чтобы это стало возможным, производится дублирование содержания контента в метабазы. Информация, предназначенная для людей, готовится в виде текста, образов и звуков, а для машин — в виде специальных кодов. Семантический веб объединяет эти виды информации в единую структуру, в которой каждому элементу «человеческой» информации будет соответствовать машинный код — специальный смысловой тэг. Метаданные должны в обязательном порядке включать сведения о том, как, где и кем была собрана данная информация и как она структурирована.

При описании многослойной архитектуры семантической сети обычно используют диаграмму, впервые предложенную Т. Бернерсом–Ли в презентации [34] и получившую название layer–cake, а в русскоязычной литературе — «пирог Тима Бернерса–Ли» (см. рис. 1). Дадим краткое описание этих слоев.

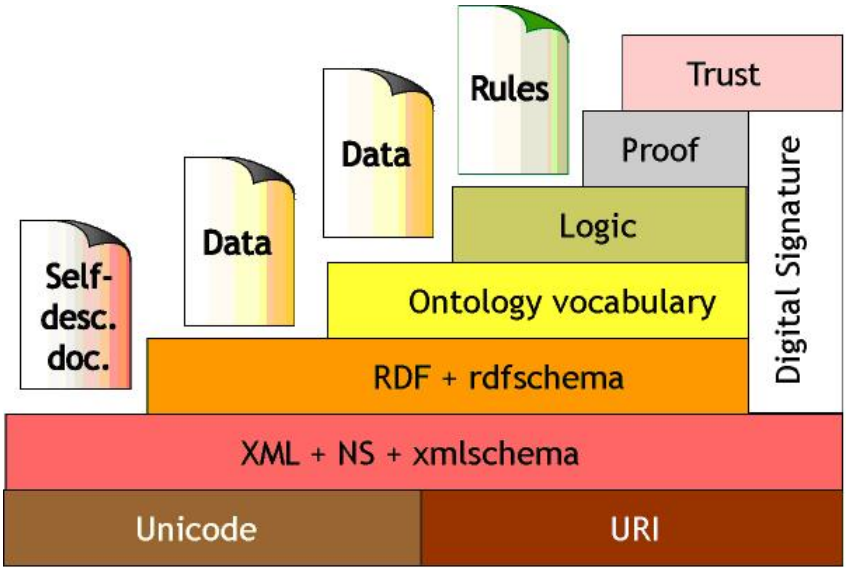

Рис. 1. «Пирог Тима Бернерса-Ли» (из презентации [34])

Первый слой *Unicode и URI.* Unicode — это стандартная кодировка для пред-

ставления символов; URI (Uniform Resource Identifier) — универсальный идентификатор ресурсов — это просто идентификатор ресурсов, с которыми мы постоянно сталкиваемся при работе в интернете (например, http://www.ksu.ru).

Слой XML и связанных с ним стандартов. XML предоставляет синтаксис для определения структуры документа, подлежащего машинной обработке. Синтаксис XML не несёт семантической нагрузки. XML Schema определяет ограничения на структуру XML-документа. Стандартный синтаксический анализатор языка XML в состоянии проверить произвольный XML-документ на соответствие его структуры так называемой схеме документа, описанной в XML Schema. «Схема» — это просто документ или фрагмент кода, управляющий множеством терминов в другом документе или фрагменте кода, как главный контрольный список или грамматика определений.

RDF представляет собой простой способ описания экземплярных данных в формате субъект – отношение – объект, в котором в качестве любого элемента этой тройки используются только идентификаторы ресурсов. Существует стандартизованное отображение этих троек на XML-документы предопределённой структуры (т. е. консорциумом W3C определена схема XML-документов, содержащих RDF-описания), а также на другие форматы представления (например, в нотацию NS). Схема RDF была разработана как простая модель типизации данных для RDF. RDF Schema описывает набор атрибутов (здесь их точнее назвать отношениями) для определения новых типов RDF-данных.

Слой онтологий (Ontologe vocabulary). Онтологии предназначены для описания более сложных конструкций, включающих в себя типы ресурсов и их свойства. Язык онтологий веба OWL расширяет возможности описания новых типов (в частности, добавлением перечислений), а также позволяет описывать новые типы данных RDF Schema в терминах уже существующих (например, определять тип, являющийся пересечением или объединением двух существующих).

Слой логики и доказательства (Logic and Proof). Логический слой необходим как средство для формулировки в документах логических выражений. Это позволяет, например, записывать правила вывода документов одного типа из документов другого типа; проверять соответствие содержимого документа некоторому множеству правил непротиворечивости; преобразовать запросы для замены неизвестных терминов известными. Примером приложения этого уровня является преобразование запросов к одной базе данных в запросы к другой базе данных в случае, когда базы данных на вебе построены независимо и объединены с помощью семантических ссылок.

Слой управления доверием (Trust) — это завершающий слой. Этот компонент находится еще на этапе разработки, одним из его элементов является, например, технология цифровой подписи.

Графический способ представления основных технологий семантического веба, предложенный Т. Бернерсом-Ли, получил развитие — можно найти схемы с другой организацией слоев. Т. Бернерс-Ли предложил также следующий способ определения, действительно ли в том или ином продукте реализована технология семантического веба: следует лишь обратить внимание на поддержку стандартов. Если продукт не поддерживает такие основополагающие стандарты, как RDF, OWL или SPARQL, то он к семантическому вебу не имеет никакого отношения [30]. В настоящее время уже работают сайты, созданные по технологии Web 3.0 (например, http://www.sun.com/servers/wp.html/, http://www.forum.nokia.com/, http://pressroom.oracle.com/, http://www.harpers.org/).

6. Технологии семантического веба в электронных изданиях

В настоящее время технологии семантического веба используется в ряде зарубежных библиотек для долговременного хранения данных. В то же время, отсутствуют российские форматы представления электронных изданий, предполагающие детальное разбиение элементов изданий, которые могут быть использованы для загрузки полнотекстовых журналов и книг в электронные хранилища. Издатели электронных журналов сейчас применяют, как правило, html-разметку выпусков своих изданий, не следуя при этом никаким общепринятым правилам, поскольку их просто не существует. Кроме того, такая разметка не может быть использована для загрузки в электронные хранилища.

Для структурирования изданий в рамках проекта «Научная электронная библиотека eLibrary.ru» были разработаны описание структуры XML-документа на языке DTD (Sarcticle.dtd) и программа Sarticle, использующая это описание. Программное обеспечение разметки электронных журналов и загрузки в базу данных электронной библиотеки, использующие названные форматы, подробно описаны в [23]. Оно прошло успешное тестирование в ряде редакций научных журналов. Следующим шагом было создание новой версии программы, позволяющей работать с большим количеством исходных файловых форматов и максимально автоматизирующей процесс структуризации текста. Модифицированная версия программы позволила расширить перечень обрабатываемых файловых форматов. Кроме того, были добавлены два новых специальных модуля для программного распознавания группы авторов с их описаниями (место работы, почтовый и электронный адреса и т. д.), ключевых слов, а также библиографических списков. Отметим, что программа обрабатывает диакритические, математические и другие специализированные символы и позволяет сохранять XML-файлы как в формате ASCII, так и в формате Unicode (UTF-8 и UTF-16). В результате разработки и использования программы время обработки усредненного выпуска журнала (115 страниц, 3 автора каждой статьи, 17 – 18 статей и 15 – 20 пристатейных ссылок) составляет 2 – 3 часа рабочего времени.

Загрузка данных в базу данных реализована с использованием технологии Windows Script Host. Создан универсальный загрузчик на VBScript, осуществляющий парсинг данных и записывающий результат в базу данных. При этом исходные тексты статей в форматах pdf и/или html записываются на файловый сервер. Апробация программ загрузки данных в НЭБ выполнялась на 60 выпусках журналов издательства Института научной информации по общественным наукам. Пользователи электронной библиотеки могут проводить поиск статей из журналов по следующим параметрам: авторы, названия статей, аннотации (рефераты), ключевые слова, слова из полных текстов статей, библиографические описания источников из пристатейных списков литературы.

По разработанной технологии в НЭБ были размещены более 200 научных журналов. Среди них такие авторитетные издания, как «Успехи физических наук», «Успехи химии», «Биология моря», «Ученые записки Казанского университета», «Известия вузов. Авиационная техника», «Известия вузов. Математика», «Известия вузов. Радиофизика», «Казанский медицинский журнал», «Вестник Казанского государственного технического университета», «Вестник Дальневосточного отделения РАН», «Дальневосточный математический журнал», «Тихоокеанская геология» и ряд других. Большинство из этих журналов до 2005 года не было представлено в интернете.

Некоторые разработки в рамках совместных проектов НЭБ и КГУ были положены в основу создаваемого Научной электронной библиотекой Российского индекса научного цитирования (см. http://www.elibrary.ru/projects/ citation/

proposal.doc).

7. Технологии семантического веба в электронном математическом журнале

Выбор технологий семантического веба для организации работы математического электронного журнала объясняется наличием в них инструментов, позволяющих учитывать структурную и семантическую составляющие информации. Программные решения, применяемые в настоящее время при подготовке научных изданий, основаны на технологиях HTML и, в значительной степени, предполагают участие человека в обработке информации, поскольку теговая разметка языка HTML позволяет структурировать текст только в части отображения документа. Подготовка математических текстов производится в tex-формате (обязательное требование большинства математических журналов). Это текстовый формат с теговой разметкой, обеспечивающей форматирование документа и включение математических формул (http://ctan.org). Tex-формат также является слабоструктурированным. Автоматизация процедур обработки слабоструктурированной информации затруднительна и не всегда эффективна. В частности, поиск информации нельзя полностью автоматизировать из-за сложности процедуры извлечения данных.

Отдельной задачей, стоящей перед научными изданиями, является обеспечение возможности извлечения метаинформации («информации об информации») поисковыми системами, большая часть которых для индексирования применяет программы-роботы, использующие метаописание ресурсов сети. Метаданные содержат обобщенную информацию о структуре и содержании информационного источника (автор, дата, источник, ключевые слова, предметная область и т. д.). В формате HTML описание метаданных возможно только через метатеги. Для наиболее унифицированного описания и каталогизации ресурсов в сети создаются специальные метаязыки, наиболее распространенным из них является Dublin Core (DC). Автоматизация генерации метаданных затруднена в силу слабой структурированности представления информации. Более того, для составления блока метаданных каждого документа требуется участие квалифицированного специалиста. Основная проблема при создании, хранении и отображении электронных публикаций по математике касается представления математических формул. Для решения этой проблемы нами была использована технология MathML, описанная выше.

Работа по автоматизации электронного журнала велась нами на основе CASE-моделирования. Была создана модель электронного математического хранилища, состоящая из комплекса UML-диаграмм.

Разработаны методы автоматической обработки электронной математической информации на основе архитектуры семантического веба. Семантическая разметка веб-страниц выполняется на основе стандарта XML, при этом для разметки математической части текста использован MathML. Разработана иерархическая модель метаданных для ресурсов электронного математического журнала в соответствии со стандартом RDF. Разработаны методы использования технологии XSLT для преобразования математических текстов и служебной информации, представленной в XML-формате, в частности, MathML. Спроектирована архитектура таблиц MySQL базы данных электронного математического журнала. Согласно разработанной схеме, в XML-файле накапливается информация, поступающая в результате пользовательского запроса, далее применяется XSLT-преобразование и результат перенаправляется в MySQL-базу. Это технологическое решение позволило существенно снизить нагрузку (количество запросов) на MySQL-сервер, поскольку

каждый запрос работает с отдельным XML-файлом, без постоянного подключения к системам баз данных.

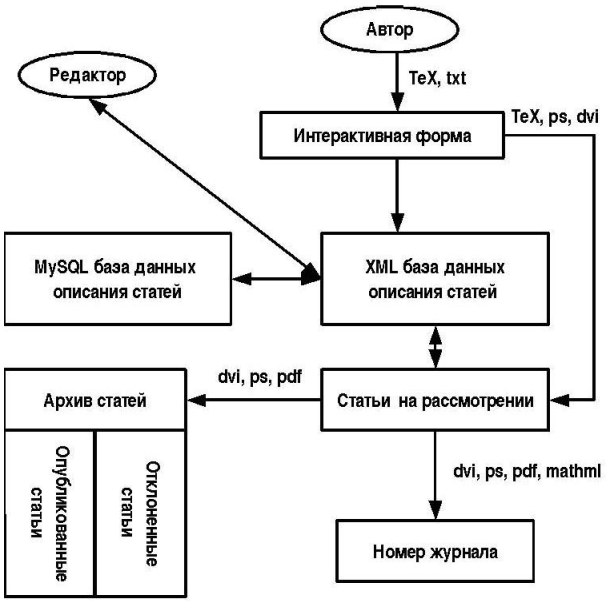

Рис. 2. Схема процесса редакционной обработки статьи

На рис. 2 представлена принципиальная схема процесса редакционной обработки статьи. Каждому прямоугольнику отвечает программный модуль, выполняющий соответствующую процедуру. На стрелках указаны форматы, участвующие в операциях обмена данными.

Указанные подходы были реализованы в виде программного комплекса для обработки и управления потоками данных в электронном журнале на основе XML/XSLT и MathML-технологий. Одной из составляющих этого комплекса является система PHP-скриптов, которая включает в себя, в частности, сервисы автоматизированного представления рукописей статей в журнал и автоматизированного прохождения рукописи. Ряд задач был решен с помощью скриптов на языках AWK и Perl. Для генерации и обработки запросов используется язык RDQL [35]. Создано программное обеспечение, осуществляющее автоматический перевод в RDF-формат метаданных, поступающих из интерактивной формы, предоставленной пользователю. Преобразование XML/RDF-информации в html-формат производится с помощью скриптов (Java, PHP). Осуществлена программная реализация управления потоками данных с веб-сайта в MySQL-базу с использованием XML-файлов в роли буфера.

Сервис автоматизированного представления рукописей включает в себя регистрацию авторов и внесение их данных в MySQL-базу данных; сбор метаданных из формы, заполняемой авторами публикаций; компиляцию tex-файлов в форматы dvi, ps, pdf для редакторской работы; внесение основных данных о рукописи (название, AMS-классификация (AMS — American Mathematical Society), ключевые слова, дата подачи рукописи) в MySQL-базу; извещение секретаря журнала о поступившей рукописи.

Сервис автоматизированного прохождения рукописей позволяет получить информацию о статусе статьи (наличие отзывов с датами представления), дает возможность технического редактирования рукописи, генерирует файлы в форматах dvi, ps, pdf, mathml, а также генерирует html-файлы с информацией о статьях и о томе журнала. Также организовано автоматическое генерирование метаданных, представляемых средствами Dublin Core и RDF.

Разработанное программное обеспечение позволило осуществить перевод в формат MathML полных текстов ранее опубликованных статей. В настоящее время статьи хранятся как в формате MathML, так и в стандартных форматах (dvi, ps, pdf). Кроме того, на сайте журнала LJM размещены аннотации статей в формате MathML.

Работа с сайтом журнала предполагает, что клиент выполнит необходимые настройки своего браузера. Прежде всего, требуется обеспечить поддержку MathML. На сайте журнала (http://ljm.ksu.ru) имеются инструкция по настройке наиболее распространенных браузеров и список узлов с необходимым программным обеспечением.

В связи с тенденцией перехода на открытые Unix-подобные операционные системы разработанное программное обеспечение адаптировано к работе под управлением ОС Linux и использует средства Unix-ориентированных систем.

В заключение отметим, что разработанные программные средства внедрены в работу электронного журнала Lobachevskii Journal of Mathematics. В настоящее время все публикуемые в журнале статьи снабжаются комплексом метаданных, включающим смешанный набор метатегов. Этот набор состоит из традиционного набора метаданных и метаданных в формате DC. Это позволяет индексировать страницы журнала как поисковым системам, роботы которых распознают только традиционный формат метаданных, так и поисковым системам, рассчитанным на метаданные в формате DC.

Summary

V. G. Veselago, A. M. Elizarov, E. K. Lipachev, M. A. Malakhaltsev. Formation and support of physical and mathematical electronic scientific publications: the transition to Semantic Web technologies.

We describe the results of research for scientific digital libraries and electronic publishing. These results were obtained by the authors for the last five years in carrying out projects of the Russian Foundation for Basic Research and the Scientific Digital Library E-library.ru. We have demonstrated the use of Semantic Web technologies in electronic scientific collections. We presented the approaches to the formation and support of physical and mathematical electronic scientific publications based on XML, RDF and other Semantic Web technologies.

The work has been supported by RFBR under grants 06-07-89132.

Литература

1. *Армс В.* Электронные библиотеки / пер. с англ. – М.: ПИК ВИНИТИ, 2001. – 276 с.

2. *Berners-Lee T.* Semantic Web Road map. — http://www.w3.org/DesignIssues/Semantic.html; русский перевод: http://gridclub.ru/library/publication.2007-04-23.2195467714/view.

3. *Ершова Т.В., Хохлов Ю.Е.* Межведомственная программа «Российские электронные библиотеки»; подходы и перспективы// Электронные библиотеки. – 1999. – Т. 2. – Вып. 2. – http:// www.iis.ru/el-bib.

4. *Библиотека памяти американцев Конгресса США.* – http://memory. loc.gov.

5. *Виттен И. Баинбридж Д.* Как строить цифровую библиотеку. – ftp:// infoserv. inist.fr/wwsympa.fcgi/info/diglib.

6. International Institute for Electronic Library Research: The current projects. – http://www.iielr.dmu.ac. uk/Projects/projsum.html.

7. eLib: Electronic Libraries Programme. – http://www.ukoln.ac.uk/services/ elib.

8. Publications from DLI1 projects. – http://www.iielr.dmu.ac.uk/ Projects/ projsum.html.

9. *Griffin S.M.* The Digital Libraries Initiative. An USA Federal Program of Research and Applications. – http://dl.ulis.ac.jp/Dljournal/No_18/1-sgriffin/1-sgriffin. html.

10. *Griffin S.M.* NSF/DARPA/NASA Digital Libraries Initiative. July – August 1998.

11. Вторая фаза проекта DLI2. – http://www.dli2.nsf.gov/.

12. Digital Libraries Initiative – Phase 2. NSF 98-63. July 15, 1998. – http:// www.nsf.gov/pubs/1998/nsf9863/nsf9863.htm.

13. Российско-Британский семинар «Электронные библиотеки». – Институт развития информационного общества, Москва, 16 – 17 июня 1999 г. – Выступления участников семинара. – http://www.iis.ru/rbdlw99 (см. также журнал «Электронные библиотеки». – 1999. – Т. 2. – Вып. 2 – 4. – http:// www.iis.ru/el-bib).

14. Седьмая Международная конференция «Крым 2000», Судак, Автономная Республика Крым, Украина, 3 – 11 июня 2000 г. Тр. конф. – ГПНТБ России, 2000. – Т. 1, 2. – http://www.gpntb.ru/win/inter=eventys/crimea2000.

15. Международная конференция «Интернет, общество, личность (ИОЛ-2000). Новые информационно-педагогические технологии». С.-Петербург, 28 февраля – 3 марта 2000 г. Тез. докл. Институт Открытое Общество – С.-Петербург, 2000. – http://iol.spb.osi.ru.

16. Международная конференция «Управление электронным будущим библиотек», Москва, 17 – 19 апреля 2000 г. Материалы конф. – http://www.rsl.ru.

17. Труды 9-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль-Залесский, Россия, 2007.

18. *Когаловский М.Р., Новиков Б.А.* Электронные библиотеки — новый класс информационных систем //Программирование. – 2000. – № 3. – С. 3–8.

19. *Когаловский М.Р.* Электронные библиотеки — развитие продолжается // Программирование. – 2002. – № 4.

20. *Когаловский М.Р.* Научные коллекции информационных ресурсов в электронных библиотеках. – Тр. 1-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – С.-Петербург, 19 – 22 окт. 1999 г. – http://www.iki.rssi.ru/.

21. *Глухов В.А.* Проект «Научная электронная библиотека eLibrary.ru» и перспективы развития электронного книгоиздания в России// Educational Technology and Society. – 2005. – V. 8. – No 1. – P. 191–197. – http://ifets.ieee.org/russian/periodical/V_81_2005EE.html.

22. *Веселаго В.Г., Елизаров А.М., Сюнтюренко О.В.* Российские электронные научные журналы: новый этап развития, проблемы интеграции// Журнал Электронные библиотеки. – 2005. – Т. 8. – Вып. 1. – http:// www.iis.ru/el-bib (см. также Сб. Научный сервис в сети Интернет: технологии распределенных вычислений. Тр. Всерос. науч. конф., г. Новороссийск, 19 – 24 сент. 2005 г. – М.: Изд-во МГУ, 2005. – С. 187–194).

23. *Глухов В.А., Елизаров А.М.* Проект «Научная электронная библиотека

eLibrary.ru» и российские электронные журналы: новый этап развития// Тр. 8-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, 17 – 19 окт. 2006 г. – Ярославль: Ярославский государственный ун-т им. П. Г. Демидова, 2006. – С. 203–207.

24. *Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Электронные журналы по математике и рекомендации консорциума W3// Тр. Всерос. науч. конф. «Научный сервис в сети Интернет», г. Новороссийск, 22 – 27 сент. 2003 г. – М.: Изд-во МГУ, 2003. – С. 75–76.

25.*Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Технология MathML как стандарт построения математического Web-пространства// Юбилейная X конференция представителей региональных научно-образовательных сетей Relarn-2003. Сб. тез. докл., 16 – 20 июня 2003 г., С.-Петербург. – С-Пб, 2003. – С. 207–208.

26. *Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Основы MathML Представление математических текстов в Internet. Практическое руководство. – Казань: Изд-во Казан. матем. общества, 2004. – 60 с.

27. *Елизаров А.М., Липачёв Е.К., Малахальцев М.А.* Технологии Semantic Web в практике работы электронного журнала по математике// Труды 8-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, 17 – 19 окт. 2006 г. – Ярославль: Ярославский государственный ун-т им. П. Г. Демидова, 2006. – С. 215–218.

28. *Hendler J.* Agents and the Semantic Web// IEEE Intelligent Systems J. – 2001. – V. 16, No 2. – P. 30–37.

29. *Стренталл Д.* Третий Веб. – http://www.xakep.ru/post/40176/.

30. *Herman I.* Questions (and Answers) on the Semantic Web // XML – Days, Berlin, Germany, 2006-09-20.

31. *Berners-Lee T., Hendler J., Lassila Ora.* The Semantic Web // Scientific American, May 17, 2001; русский перевод: Семантическая Сеть. – http://ezolin.pisem.net/logic/semantic_web_rus.html.

32. Семантическая паутина. – Электронная энциклопедия Википедия, http://ru.wikipedia.org/wiki/.

33. Язык онтологий в Web. – http://wmast.com.ua/article.php.

34. *Berners-Lee T.* Semantic Web on XML. – http://www.w3.org/2000/Talks/ 1206-xml2k-tbl/slide1-0.html.

35. RDQL – A Query Language for RDF. – http://www.w3.org/Submission/2004/ SUBM-RDQL-20040109/.

36. *Глухов В. А., Елизаров Е. К., Липачев Е. К., Малахальцев М. А.* Электронные научные издания: переход на технологии семантического Веба// Электронные библиотеки. – 2007. – Т. 10. – Вып. 1 (http://www.elbib.ru)

Сведения об авторах

Веселаго Виктор Георгиевич — д. ф.-м. н., профессор, лауреат Государственной премии, главный научный сотрудник НИИММ;

E-mail: *v.veselago@relcom.ru*

Елизаров Александр Михайлович — д. ф.-м. н., профессор, заслуженный деятель науки РТ, лауреат премии им. Х. М. Муштари АН РТ, директор НИИММ;

Липачёв Евгений Константинович — к. ф.-м. н., доцент, ведущий научный сотрудник НИИММ;

E-mail: *lipachev@ksu.ru*

Малахальцев Михаил Арменович — к. ф.-м. н., доцент, ведущий научный сотрудник НИИММ; E-mail: *mikhail.malakhaltsev@ksu.ru*