Тенденции развития технологий управления информационными ресурсами в электронных библиотеках ♣

© М.Р. Когаловский

Институт проблем рынка РАН
kogalov@cemi.rssi.ru

Аннотация

В разработках коллекций информационных ресурсов электронных библиотек, обеспечении их поддержки и доступа к ним оказался востребованным весь спектр ключевых технологий управления информацией, используемых в современных информационных системах, – технологии баз данных, технологии текстового поиска и, конечно же, веб-технологии. Поэтому не случайно, что сложившиеся и формирующиеся в последние годы тенденции развития указанных технологий оказывают весьма существенное влияние на функциональные возможности электронных библиотек. В докладе рассматриваются наиболее значимые из этих тенденций.

1 Введение

Существующие в настоящее время и разрабатываемые новые системы электронных библиотек характеризуются большим разнообразием поддерживаемых в них информационных ресурсов, способов организации их коллекций, функциональными возможностями пользовательских интерфейсов, архитектурных особенностей этих систем и других их технологических характеристик. Не случайно поэтому, что в разработках информационных систем этой категории востребован практически весь спектр ключевых технологий управления информацией, созданных научно-техническим сообществом и индустрией программного обеспечения в области баз данных, текстового поиска, Всемирной паутины и Интернет.

Действительно, веб-технологии являются непременной составной частью технологического оснащения многих электронных библиотек. Веб является средой «обитания» электронных библиотек, обеспечивающей доступ пользователей к их ресурсам. Электронные коллекции информационных ресурсов многих электронных библиотек организованы в виде веб-сайтов. Кроме того, Веб является средой доступа к различным системам баз данных, содержащим метаданные и/или коллекции структурированных данных, непосредственно интересующих пользователей электронной библиотеки. Более того, Веб может рассматриваться как уникальная гигантская общечеловеческая универсальная (по предметной области) электронная библиотека.

Вероятно, наиболее распространенным видом информационных ресурсов электронных библиотек являются тексты на естественных языках. Этим обусловлено широкое применение в таких системах технологий текстового поиска. Они используются при этом не только в системах, построенных по принципу традиционных текстовых систем, но и для поиска в коллекциях, организованных в виде веб-сайтов, а также для поиска в глобальной среде Веб. Технологии текстового поиска, созданные первоначально для использования в рамках централизованно администрируемых коллекций текстовых документов уже более десятилетия назад стали применяться в децентрализованной среде Веб. В последнее время адаптированные к Веб технологии текстового поиска возвращаются в централизованно администрируемую информационную среду. Так, компанией Яндекс разработана свободно распространяемая поисковая система для поиска ресурсов на платформе персональных компьютеров [10]. Аналогичную систему создала компания Google.

Нужно отметить также важную роль технологий баз данных в электронных библиотеках. В рамках электронных библиотек используются СУБД, основанные на различных моделях данных – реляционные, объектные, объектно-реляционные, XML-ориентированные системы. Управляемые ими базы данных поддерживают в электронных библиотеках разнообразные коллекции структурированных данных и обеспечивают эффективный доступ к ним. Это, например, данные, полученные в результате научных экспериментов, наблюдений и измерений, компьютерного моделирования реальных процессов, экономическая статистика и т.п. Системы баз данных обеспечивают в электронных библиотеках поддержку разнообразных структурированных ме-

Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006.

таданных (например, классификаторов, каталогов, тезаурусов, словарей и др.). Создание XML-ориентированных систем баз данных позволило поддерживать в таких системах слабоструктурированные и структурированные XML-данные. Помимо этого, целый ряд коммерческих СУБД позволяет еще с середины 90-х годов хранить текстовые документы, осуществлять их полнотекстовое индексирование и на этой основе осуществлять поиск хранимых текстовых документов по элементам их содержания.

Все перечисленное показывает ключевую роль технологий управления информацией в электронных библиотеках. Развитие этих технологий обогащает функциональные возможности электронных библиотек. В свою очередь, возникающие в многочисленных разработках электронных библиотек различного назначения новые требования к технологиям управления информацией стимулируют их дальнейшее развитие.

Конечно же, сфера применения технологий управления информацией не ограничивается электронными библиотеками. Однако анализ наиболее значимых тенденций развития этого пласта информационных технологий, сформировавшихся и зарождающихся в последние годы, даст возможность оценить перспективы развития разработок в области электронных библиотек. Именно этот анализ является целью данной работы.

2 Интенсивный рост объема информационных ресурсов

Создание Всемирной паутины, развитие информационных технологий, процессы формирования информационного общества – все это стимулирует быстрый рост объема информационных ресурсов, поддерживаемых в информационных системах, в частности, и в электронных библиотеках. Темпы роста информационных ресурсов особенно интенсифицировались в последнее десятилетие. Именно в этот период сформировались указанные выше предпосылки.

Достигнутые масштабы объема информационных ресурсов, генерируемых, хранимых и обращающихся в различных сферах жизнедеятельности, уже не позволяют обойтись перечнем единиц измерения, которые стали привычными и широко используются на практике – биты, байты, килобайты (Kb), мегабайты (Mb), гигабайты (Gb), терабайты (Tb) и, наконец, петабайты (Pb). Введены в лексикон специалистов новые единицы измерения объема информации - экзабайт (Exabyte, Eb = 1K петабайтов), зетабайт (Zettabyte, Zb = 1K экзабайтов) и йотабайт (Yottabyte, Yb= 1K зетабайтов).

Исследования, направленные на получение оценки объемов накопленных человечеством информационных ресурсов и темпов их ежегодного роста, проводятся в Калифорнийском университете (Беркли) в Школе управления информацией и информационных систем (School of Information Management and Systems) при поддержке компаний Microsoft, Intel, Hewlett-Packard и EMC.

В 1999 и 2002 гг. в рамках указанного проекта были получены оценки хранимых на машиночитаемых носителях информационных ресурсов, а также объемов потоков информации – телефон, радио, TV, Интернет, печатных изданий и документооборота организаций. Объемы информации, представленной в аналоговом виде, для сопоставимости пересчитывались в объемы эквивалентной оцифрованной информации. По материалам этого проекта поддерживается «живой» документ в Веб [33].

Приведем лишь несколько оценок, представленных в этом документе. В 2002 г. продуцировано около 5 Eb новой информации. Из них около 92% хранится на магнитных носителях. В период 1999–2002 гг. объем хранимой информации возрастал в среднем на 30% в год и за три года примерно удвоился. Объем оцифрованных информационных ресурсов Библиотеки Конгресса США мог бы составить 10 терабайтов. Примерно в 2 Pb можно оценить объем ресурсов всех университетских библиотек США. Объемы информационных ресурсов Веб в 2002 г. могут быть приблизительно оценены следующим образом: стандартный гипермедийный Веб (“Surface” Web) 167 Tb; «скрытый» Веб (FTP-архивы и базы данных, доступные в среде стандартного Веб) - 92 Pb.

Без сомнения, можно предполагать, что в составе этих гигантских объемов информационных ресурсов значительную долю составляют информационные ресурсы электронных библиотек. Так, например, чрезвычайно крупной коллекцией информационных ресурсов обладает прототип «цифровой Земли» - Alexandria Digital Earth Prototype (ADEPT) [13], разработанный в рамках проекта электронной библиотеки Alexandria совместно университетами в Санта Барбара и Лос-Анджелесе (Калифорния), Техническим научно-исследовательским институтом и университетом штата Джорджия (США). Другими крупными коллекциями информационных ресурсов обладают электронные библиотеки, созданные в области космического зондирования земной поверхности и экологического мониторинга [7].

В ряде источников отмечается характерный для многих областей научных исследований в последние годы экспоненциальный рост данных, полученных в результате научных экспериментов, наблюдений, измерений, компьютерного моделирования. Так, в работе [22] этот факт отмечается в области молекулярной биологии для последнего десятилетия. В астрономических исследованиях также имеют место высокие темпы роста объема данных, накапливаемых в обсерваториях. Объем этих данных примерно удваивается за период от шести двенадцати месяцев [1, 16]. Крупнейшими «генераторами» информационных ресурсов являются исследования в области физики частиц, проводимые в ряде крупных исследовательских центров (ЦЕРН, Стэнфордский университет и др.). Как известно, именно потребности управления гигантскими объемами

данных, которые генерируются на современных ускорителях, привели к рождению грид-технологий и концепции грида данных.

3 Глобализация формирования и использования информационных ресурсов

Возможности использования коммуникационной среды Интернет и веб-технологий в разработках электронных библиотек и других информационных систем стимулировали процессы распределения и глобализации как формирования коллекций информационных ресурсов, так и доступа к ним. Глобальный доступ к информационным ресурсам системы в любой точке, где имеется доступ в Интернет, и в любое время является одним из необходимых условий отнесения такой информационной системы к категории систем, которые принято называть электронными библиотеками.

Наряду с электронными библиотеками, коллекции информационных ресурсов которых администрируются централизованно, создаются такие электронные библиотеки, коллекции которых поддерживаются во множестве автономно формируемых и администрируемых децентрализованным образом источников, доступных в глобальной среде.

Примерами крупных электронных библиотек такого вида являются международная электронная библиотека по общественным наукам RePec и выступающая в ней в качестве одного из источников информационных ресурсов, а также и в виде крупной самостоятельной научной электронной библиотеки Отделения общественных наук РАН, отечественная система Соционет [9, 38].

К электронным библиотекам рассматриваемого вида можно отнести и другие многочисленные системы регионального, национального и международного уровня, например, корпоративные библиотечные системы. К их числу относятся, в частности, крупнейшая международная система OCLC WorldCat [47], а также отечественная автоматизированная система Российского сводного каталога по научно-технической литературе [11].

В электронных библиотеках рассматриваемого вида и в других информационных системах используются различные подходы и методы интеграции информационных ресурсов, которые обсуждаются далее (разд. 6).

4 Интеграция технологий управления информационными ресурсами

Важной тенденцией последнего десятилетия в развитии технологий управления информацией стала интеграция таких технологий в реализациях многочисленных конкретных систем. Наряду с информационными системами вообще и системами электронных библиотек, в частности, основанными на каком-либо одном из пластов технологий управления информацией (технологии баз данных, веб-технологии, технологии текстовых систем) имеются многочисленные примеры совместного использования различных сочетаний этих технологий в рамках одной системы.

Многие организации стали обладать источниками структурированных данных наряду с текстовыми системами. Стремление к упрощению технологических процессов в организации в таких ситуациях и необходимость интеграции информационных ресурсов привели к производству СУБД, способных поддерживать наряду со структурированными данными также и текстовые документы и выполнять их поиск по запросам пользователей. Развитыми средствами текстового поиска обладают в настоящее время многие серверы баз данных, например, DB2 (IBM), Oracle (Oracle Corp.), SQL Server (Microsoft Corp.) и др.

Другое развивающееся направление интеграции технологий управления информационными ресурсами – это интеграция технологий баз данных и веб-технологий. Доступность коммуникационных возможностей Интернет и комфортный доступ пользователей в среду Веб с помощью легко осваиваемых программ просмотра – веб-браузеров - стимулировали обеспечение удаленного доступа к базам данных в этой среде многих пользователей без необходимости разработки специальных средств пользовательского интерфейса. Разработки новой технологической платформы Веб, основанной на языке XML, привели к созданию нового класса систем баз данных, называемых XML-ориентированными системами [3, 5]. Разработка технологий семантического Веб и создание широко признанных стандартных средств описания онтологий создают предпосылки для решения одной из важных перспективных задач развития технологий баз данных – создание пользовательских интерфейсов в системах баз данных, основанных на онтологиях предметной области системы. Актуальность решения этой задачи была отмечена на состоявшейся в июне 2003 году Лоуэллской дискуссии (штат Массачусетс, США) о перспективах развития технологий баз данных, в которой участвовал ряд крупнейших специалистов в области технологий баз данных [12].

Развитые комплексы инструментальных средств систем баз данных, соответствующих стандартам платформы XML, поддерживаются в настоящее время SQL-серверами баз данных компаний Oracle, IBM, Microsoft и других поставщиков программного обеспечения систем баз данных. Углублению интеграции технологий баз данных и веб-технологий способствует также завершенная в 2003 году ISO разработка новой версии стандарта объектно-реляционного языка запросов для систем баз данных SQL-2003. В составе этого стандарта имеется компонент SQL/XML [21], обеспечивающий интеграцию технологий SQL-баз данных и XML-технологий.

Нужно отметить также еще одно активно развиваемое направление интеграции технологий управления информационными ресурсами. Оно связано с веб-технологиями и технологиями текстового поис-

ка. После создания Всемирной паутины и интенсивного наращивания ее информационных ресурсов стало ясно, что навигационный доступ к информационным ресурсам, который обеспечивается технологиями этой системы, не может эффективно удовлетворять информационные потребности пользователей. Для решения этой проблемы в Веб начали использоваться традиционные технологии текстового поиска. Стали создаваться поисковые машины Веб, которые сегодня активно используются многими миллионами пользователей этой гигантской электронной библиотеки.

Сегодняшние версии таких систем радикально отличаются от ранних их версий функциональными возможностями, учитывают особенности поиска ресурсов в Веб [8], существенно отличающиеся от условий поиска в традиционных системах текстового поиска. Действительно, в отличие от традиционных систем текстового поиска, в Веб нет централизованного администрирования информационными ресурсами, не поддерживаются метаданные коллекций, существенную роль играют взаимосвязи между документами с помощью гиперссылок, огромные объемы пространства поиска, высокая динамичность информационных ресурсов - изменчивость состава коллекции и отдельных документов (веб-страниц). При поиске в Веб необходимо учитывать также низкое качество документов, связанное с легкостью публикации ресурсов в этой среде и отсутствием администрирования, многоязычность ресурсов, значительная избыточность коллекций - наличие многих копий документов, содержащихся на разных веб-сайтах и т.д.

В разработках информационных систем с использованием рассмотренных вариантов интеграции технологий управления информационными ресурсами каждый из базовых пластов таких технологий привносит свои специфические возможности в создаваемые системы. Их можно кратко охарактеризовать следующим образом.

- На основе технологий баз данных обеспечиваются полнофункциональное управление структурированными данными, обработка запросов в терминах поддерживаемой модели данных и в транзакционном режиме, хранение традиционных текстовых ресурсов и XML-документов и эффективный доступ к ним в среде хранения с использованием техники индексирования данных и других методов прямого доступа.
- Технологии текстового поиска привносят возможности поддержки естественных языков в качестве языков пользовательского интерфейса, использование различных подходов к структуризации содержания текстовых документов, представленных в системе, и пользовательских запросов, сформулированных на естественных языках.
- Вклад веб-технологий состоит в обеспечении распределения информационных ресурсов между узлами Интернет и возможности децентрализованного управления ими, глобального доступа к информационным ресурсам в среде Веб без предъявления высоких требований к квалификации пользователей благодаря существованию средств навигационного доступа, обеспечении поддержки семантики информационных ресурсов средствами технологий Веб нового поколения, основанного на стандартах платформы XML (семантический Веб), и, соответственно, доступа к ним на семантическом уровне.

Одним из следствий указанных тенденций интеграции технологий стало индустриальное производство ряда серверов баз данных, которые уже неправомерно, строго говоря, относить к продуктам указанной категории. Это, скорее, теперь уже технологические «комбайны». Действительно, такие, например, продукты, как сервер баз данных Oracle Database 10g или флагманский программный продукт для систем баз данных компании IBM – сервер баз данных DB2 Universal Database v.8 – способны не только выполнять функции управления традиционными объектно-реляционными SQL-базами данных. Они могут эффективно оперировать текстовыми, пространственными и мультимедийными данными. Как уже отмечалось, они поддерживают также важнейшие стандарты платформы XML, управляют XML-ориентированными базами данных, обладают веб-интерфейсами, поддерживают технологии потоков работ, интеграции бизнес-процессов и выполняют многие другие функции.

5 Конвергенция технологий управления информационными ресурсами

В развитии технологий управления информационными ресурсами можно проследить также тенденцию конвергенции, идейного сближения разных пластов указанных технологий, их взаимного влияния, миграции проверенных временем идей и концепций из одних областей в смежные технологические области. Эта тенденция наиболее масштабно проявляется в разработках технологий Веб нового поколения. Рассмотрим кратко, каким образом это происходит на примере стандартов платформы XML, где можно обнаружить воплощение многих идей, заимствованных из технологий баз данных.

Прежде всего, о значительном влиянии традиционных «базоданновых» подходов на эту область убедительно свидетельствует активное применение в ее техническом лексиконе таких терминов, как «модель данных», «база данных», «схема», «метаданные», «ограничение целостности», «язык запросов» и др.

Как и в системах баз данных, в Веб нового поколения предусматривается многоуровневая архитектура данных – различаются хранимые данные («хранимые сущности» XML, файлы – физический уровень) и XML-документы (логический уровень). Физическое и логическое представления данных определяются по принципу самоописываемости с помощью встроенных метаданных, выраженными средствами XML-разметки. Для логического представления XML-данных может быть определена

отчужденная от них схема (DTD и/или XML Schema). Более высокий уровень абстракции данных в архитектуре XML-данных – семантический уровень. Для описания семантики XML-документов используются RDF-спецификации [39] в терминах понятий, определяемых описанием онтологии предметной области. Онтологии описываются средствами языков RDFS [40] или OWL [37], и это описание представляет онтологический уровень архитектуры.

Со структурной точки зрения, XML-документ является частным случаем записи базы данных CODASYL, представляющей собой иерархию элементов данных, которые могут быть простыми (атомарными), повторяющимися группами, в том числе, и с переменным числом повторений. В записи базы данных CODASYL, однако, могут содержаться производные (виртуальные) элементы данных. Более развитым является и множество типов данных, представляющих значения атомарных элементов данных записи.

Как и в технологиях баз данных, фундаментальным понятием в рассматриваемых веб-технологиях является понятие модели данных, хотя оно и интерпретируется в отличие от современной «базоданновой» трактовки [2] не как инструмент моделирования данных, а как его результат – как структура XML-документа. В рамках многоуровневой архитектуры данных Веб, основанного на платформе XML, поддерживается комплекс моделей данных (в «базоданновом» смысле). На логическом уровне используются альтернативные модели: (XML + XQuery), DOM, XPath, (XML + XSLT). На семантическом уровне предоставляется модель данных (RDF + SPARQL). Наконец, для уровня онтологий создается вариант полнофункциональной модели (OWL + язык правил). Язык SPARQL – это язык запросов консорциума W3C в терминах RDF-спецификации [42]. Язык правил для семантического Веб находится в настоящее время в стадии разработки. Требования к одному из возможных претендентов на роль стандарта такого языка в настоящее время обсуждаются в W3C, и он получил название Rule Interchange Format (RIF) [41].

Приведенные факты, на наш взгляд, убедительно подтверждают наличие тенденции конвергенции технологий управления информационными ресурсами.

6 Интеграция информационных ресурсов

Хотя проблема интеграции данных в различных ее постановках привлекает внимание специалистов по управлению данными уже около трех десятилетий, до недавнего времени связанные с нею разработки все еще не выходили из стен исследовательских лабораторий. Однако в последние годы создание систем интеграции данных стало весьма актуальным направлением практических разработок информационных систем различного назначения, в том числе и электронных библиотек.

Под интеграцией данных в информационных системах понимается обеспечение единого унифицированного интерфейса для доступа пользователей к совокупности автономных источников данных, которые, как правило, обладают неоднородностью относительно некоторых их свойств.

Проблема интеграции данных характеризуется большим разнообразием постановок задач, подходов и методов, используемых для их решения. Обсуждению различных аспектов технологий интеграции данных посвящены многочисленные публикации в периодике, в трудах многих авторитетных научно-технических конференций. Содержательный обзор проблематики интеграции данных и используемых в этой области подходов можно найти, например, в обзоре [31] и в презентации [19]. Можно упомянуть здесь также нашу работу [6]. Важные аспекты проблемы интеграции данных обсуждаются в работе [1].

В исследования систем интеграции данных чаще всего рассматриваются случаи интеграции структурированных данных либо комбинации структурированных и слабоструктурированных данных. При этом принимаются во внимание логическая (различие моделей данных источников, различие схем и т.п.) и/или семантическая неоднородность источников данных (различие онтологий). Состав источников интегрируемых данных может быть статическим и динамическим. Содержимое источников может быть неизменным или изменяемым.

Рассматриваются разнообразные способы интеграции – материализованная и виртуальная, а также разные уровни интеграции - логическая и семантическая интеграция.

При использовании материализованной интеграции данных создается новый материализованный источник интегрированных данных, который используется автономно от породивших его источников. При необходимости его состояние приходится синхронизироваться с их актуальным состоянием.

В случае виртуальной интеграции, напротив, не предусматривается создание нового материализованного источника интегрированных данных, Система интеграции поддерживает виртуальный источник, который в любой момент времени «содержит» актуальные данные интегрируемых источников, и синхронизации его состояния не требуется. Права владельцев исходных интегрируемых источников сохраняются. Они продолжают автономно поддерживать их в своих интересах, предоставляя вместе с тем права доступа к их ресурсам пользователям системы интеграции данных в соответствии с установленным регламентом. Авторизованные пользователи системы интеграции получают непосредственный доступ только к виртуальному источнику, воплощаемому данной системой.

Используются различные подходы к построению архитектуры данных систем виртуальной интеграции. Наиболее популярной является архитектура

посредника-адаптеров. Посредник – это функциональный компонент системы интеграции данных, который обеспечивает поддержку глобальной схемы для интегрированного виртуального источника и организует обработку пользовательских запросов, выраженных в терминах глобальной схемы, декомпозируя их на подзапросы, адресуемые соответствующих источникам, осуществляя композицию получаемых частичных результатов и выдачу полного результата пользователю. Адаптеры источников обеспечивают их «гомогенизацию», представляют информационные ресурсы источников однородным образом в терминах глобальной модели данных, принимают на обработку подзапросы от посредника, активизируют их обработку источником и возвращают полученные результаты посреднику.

На практике чаще всего используются две разновидности архитектуры данных систем виртуальной интеграции с посредником - Global as View и Local as View [32, 34]. Они различаются способами определения отображений между схемами данных источников и глобальной схемой.

Первая из них (Global as View) предусматривает определение глобальной схемы в терминах схем локальных источников. Такой подход более эффективен в случае, когда множество всех используемых источников предопределено. При использовании второй разновидности рассматриваемой архитектуры (Local as View) предполагается, что схема для каждого из локальных источников данных определяется в терминах глобальной схемы. Хотя в этом случае усложняется отображение пользовательских запросов в среду локальных источников данных, такой подход имеет важное достоинство – он допускает динамичность состава множества интегрируемых источников данных. Новые источники данных могут подключаться к системе как на стадии разработки, так и на стадии функционирования.

Обратимся теперь к уровням интеграции данных. В системах логической интеграции данных преодолевается неоднородность интегрируемых источников информационных ресурсов относительно поддерживаемых ими моделей данных и/или схем данных. Эта неоднородность преодолевается динамически - на стадии исполнения. В то же время, семантическая неоднородность данных, принадлежащих разным источникам, преодолевается на стадии разработки. В системах семантической интеграции данных семантическая неоднородность данных из разных интегрируемых источников преодолевается на стадии исполнения.

В проблематике семантической интеграции данных важное место занимают разработки, связанные с использованием онтологических спецификаций предметной области. При этом в системе интеграции поддерживаются общая онтология системы и частные онтологии отдельных интегрируемых источников, обеспечиваются отображения между частными онтологиями и общей онтологией [35]. Исследования и разработки по семантической интеграции данных в последние годы весьма активно проводятся в области молекулярной биологии [26]. Оригинальный подход к семантической интеграции информации с использованием развитой семантической модели данных в качестве канонической модели данных посредника, а также разработанного авторами метода построения предметных посредников, реализуется в проекте Института проблем информатики РАН [29].

Интеграция неструктурированных данных также стала попадать в последние годы в сферу проблематики систем интеграции данных.

Своеобразный класс систем интеграции представляют системы, основанные на технологии Инициативы открытых архивов (Open Archives Initiative, OAI) [45]. В большинстве известных систем этой категории их информационные ресурсы представляют собой коллекции текстовых документов, чаще всего научных публикаций, которые автономно формируются в узлах глобальной сети, поддерживаются и администрируются их владельцами. Важно заметить, однако, что информационные ресурсы открытого архива не обязательно должны быть текстовыми документами. Это могут быть также структурированные или слабоструктурированные данные, а также смесь структурированных, слабоструктурированных и/или неструктурированных данных.

В соответствии с технологией OAI, предусматривается материализованная интеграция в едином репозитории не самих информационных ресурсов, интересующих пользователей системы интеграции, а представленных некоторым стандартным образом метаданных, описывающих коллекции информационных ресурсов источников данного архива и отдельные элементы этих коллекций. Сбор таких метаданных для репозитория осуществляется в соответствии со специально разработанным протоколом Open Archives Initiative Protocol for Metadata Harvesting, описанным в [45]. Централизованно поддерживаемый репозиторий метаданных доступен сервису открытого архива, который обрабатывает запросы его пользователей.

Примерами электронных библиотек, основанных на принципах и технологии OAI, являются уже упоминавшиеся выше международная электронная библиотека по общественным наукам RePec и отечественная электронная библиотека по общественным наукам Соционет [9, 38].

Одним из важных аспектов систем интеграции данных является архитектура таких систем. В многочисленных известных проектах систем интеграции данных можно обнаружить не только различные подходы к архитектуре данных, но и некоторое разнообразие других аспектов их архитектуры - архитектуры взаимодействия функциональных компонентов системы интеграции, их сетевой архитектуры и др. Так, на практике часто используется не только архитектурный подход «клиент-сервер», но и децентрализованная архитектура P2P [17]. В этом случае обычно не поддерживается глобальная схема интегрированных информационных ресурсов, и используются попарные отображения представлений

данных узлов (peer), обменивающихся данными. Кроме того, часто используются архитектура промежуточного слоя [28, 36], а также веб-сервисная архитектура [18, 24].

В связи с востребованностью и активным развитием грид-технологий, особое внимание уделяется в настоящее время технологиям интеграции данных на основе гридов данных. Практическая реализация возможных в этой области подходов существенным образом связана с созданием комплекса стандартов, необходимых для разработки основанных на них систем интеграции данных. Важную роль в этом направлении играет деятельность консорциума Global Grid Forum (GGF) – признанного органа стандартизации грид-технологий. Недавно Рабочая группа консорциума Database Access and Integration Services Working Group опубликовала спецификации WS-DAI (Web Service Data Access and Integration), определяющие интерфейсы веб-сервисов, обеспечивающих доступ к источникам данных, независимо от модели данных, в терминах которой представляются их информационные ресурсы. Кроме того, разработаны расширения этих спецификаций для реляционных и XML-ориентированных систем баз данных (WS-DAIR и WS-DAIX). Тем самым созданы основы стандартизации доступа к информационным ресурсам указанного вида в среде, основанной на грид-технологиях, которая, как известно, базируется на веб-сервисной архитектуре. Обзор указанного семейства спецификаций можно найти в работе [14]. Полные их тексты доступны на веб-сайте консорциума GGF (http://www.ggf.org).

В последнее время проблеме интеграции информационных ресурсов уделяется большое внимание поставщиками индустриальных технологий. Наиболее развитые средства для решения этой проблемы основаны на архитектуре промежуточного слоя. К этой категории относится, например, IBM WebSphere Information Integrator [28] - технология компании IBM для интеграции неоднородных структурированных, слабоструктурированных и неструктурированных данных. Продукты семейства Data Hub [36] компании Oracle обеспечивают интеграцию структурированных данных из множества неоднородных источников с использованием большого набора конверторов данных, ориентированных на преобразование многочисленных форматов представления данных.

7 Обеспечение доступа к информационным ресурсам на уровне семантики

Важнейшей тенденцией развития технологий управления информацией в последние годы стала конструктивно осуществляемая на уровне индустриальных технологий попытка обеспечения доступа пользователей к информационным ресурсам на уровне семантики. Исследовательские работы в этой области проводятся с разной степенью интенсивности уже более трех десятилетий. В технологиях баз данных в 70-80-х гг. создавались семантические модели данных, велись работы на стыке технологий баз данных и баз знаний. Были созданы различные прототипы. Однако результаты этих исследований не привели к созданию индустриальных технологий. Эта проблема вновь была поставлена на повестку дня участниками упоминавшейся ранее Лоуэллской дискуссии о перспективах развития технологий баз данных [12]. Цели дискуссии состояли в том, чтобы оценить вызовы времени и сформулировать перспективные, с точки зрения экспертов-участников, направления развития технологий баз данных. В отчете о дискуссии в качестве одного из таких направлений признается использование подходов текстовых систем и семантического Веб, позволяющих формулировать запросы на основе онтологий в терминах предметной области.

В области технологий текстового поиска еще во второй половине 60-х годов под руководством основателя современных технологий текстового поиска Дж. Сэлтона проводились исследования и разработки методов поиска текстовых документов на основе их содержания, была предложена векторная модель поиска [3]. Сегодня эти подходы широко используются во многих создаваемых системах текстового поиска в качестве основы поисковых механизмов. В последние годы в области систем текстового разработаны подходы, использующие в процессе поиска документов формальные или полуформальные онтологии предметной области.

Что касается веб-технологий, то именно с указанной целью создателем Всемирной паутины Т. Бернерсом-Ли во второй половине 90-х годов была провозглашена задача создания семантического Веб [15, 46] – Веб нового поколения, который, в отличие от действующей версии Веб, ориентирован на взаимодействие не только с человеком, но и способен обеспечить совместную работу с его ресурсами человека и компьютерных агентов. Технологии семантического Веб успешно разрабатываются консорциумом W3C.

Для решения указанной проблемы необходимо стандартизовать средства явного описания семантики информационных ресурсов и средства пользовательского интерфейса семантического уровня. В настоящее время консорциум W3C располагает стандартом RDF описания контента информационных ресурсов [39, 40]. Завершается разработка языка запросов в терминах RDF-спецификации ресурсов (язык SPARQL [42]). Кроме того, разработаны полуформальные и формальные языковые средства описания онтологий – стандарты RDFS [40] и OWL [37]. Наконец, создается язык правил [41] для работы на уровне онтологий, который позволит реализовать методы логического вывода в среде, поддерживающей указанные стандарты семантического Веб.

Нужно, наконец, отметить, что в системах текстового поиска, а также в рамках веб-технологий, широкое применение для описания контента информационных ресурсов в настоящее время находит неформальное средство - набор элементов метаданных Дублинского ядра [20], который имеет статус

официальных стандартов ISO (стандарт ISO:15836-2003) и ANSI (стандарт ANSI/NISO Z39.85-2001). Дублинское ядро весьма привлекательно благодаря его простоте. Однако оборотной стороной этих его достоинств является некоторая размытость описания, связанная с неоднозначностью трактовки смысла отдельных входящих в него элементов метаданных, а также отсутствие стандартизации представления значений некоторых из них.

8 Уменьшение гранулярности доступа

Одна из проблем, возникающих в информационных системах при выдаче пользователю информации в ответ на введенный им запрос, заключается в том, чтобы предоставлять пользователю информацию именно в том объеме, в котором он ее запрашивал. Иначе говоря, гранулярность доступа должна соответствовать информационной потребности пользователя.

Выполнение этого требования всегда обеспечивается в системах баз данных. Так, в реляционной системе базы данных результатом обработки запроса всегда является таблица (или представление), включающая только столбцы, которые указаны в целевом списке запроса. Если в запросе предусмотрена операция селекции, то результирующая таблица будет включать только строки, соответствующие заданному критерию селекции.

Указанное требование не всегда выполняется в системах текстового поиска и при доступе к информационным ресурсам в Веб. В традиционных системах текстового поиска в результате обработки пользовательского запроса всегда выдается результирующее множество полных документов, даже если пользователя интересуют лишь какие-либо фрагменты этих документов. Подобным образом, в действующей версии Веб пользователь может, используя навигационный доступ, всегда извлечь только полную веб-страницу или получить с помощью поисковой машины Веб список гиперссылок на полные веб-страницы.

Создаваемые в настоящее время новые технологии управления информационными ресурсами позволяют обеспечить более мелкую гранулярность доступа. С этой целью для систем текстового поиска разрабатываются технологии «вопрос-ответ». Используя их, можно получать в ответ на запрос не полные документы, а их фрагменты, содержащие ответы на сформулированные в запросах пользователей вопросы. В тематике международной конференции Text Retrieval Conference (TREC) [44], которая является движущей силой деятельности по сопоставимой сравнительной оценке эффективности разрабатываемых систем текстового поиска путем проведения сопоставимых испытаний на тестовых коллекциях, предусматривается специальная дорожка, посвященная указанной проблеме.

Что же касается уменьшения гранулярности доступа в Веб, то эта проблема решается средствами XML-технологий. В частности, при использовании в качества языка запросов XQuery или XPath в ответ на запрос можно получать не только специфицированные в нем полные XML-документы, но и их фрагменты. Интерфейсы, поддерживающие указанные языки, в настоящее время уже используются в целом ряде XML-ориентированных СУБД. В дальнейшем они будут использоваться и собственно в среде Веб, а также в различных репозиториях, поддерживающих XML-данные.

9 Использование стандартов платформы XML в электронных библиотеках

Уже отмечалось, что Веб является “средой обитания” электронных библиотек. Поэтому радикальные технологические сдвиги, осуществляемые в этой среде, связанные, прежде всего, с созданием для нее новой технологической платформы, не могут не оказывать влияния на развитие информационных систем этого класса.

Ограничимся здесь кратким перечислением наиболее существенных направлений использования стандартов платформы XML в электронных библиотеках. Более подробно этот вопрос обсуждается в [4].

К числу указанных направлений относятся:

- Представление коллекций электронных информационных ресурсов в электронных библиотеках.
- Обеспечение навигационного доступа к информационным ресурсам по гиперссылкам с помощью средств, привычных для пользователей Веб.
- Обеспечение интерфейсов языков запросов для доступа к информационным ресурсам, представленным в виде XML-документов, на основе элементов их содержания. В качестве языков запросов могут использоваться XQuery, XPath, XSLT, SPARQL.
- Использование представленных с помощью стандартов XML информационных ресурсов в рамках продвинутых веб-приложений, являющихся функциональными компонентами электронных библиотек.
- Использование XML как языка-посредника для обмена данными между различными компонентами распределенных электронных библиотек или различными взаимодействующими электронными библиотеками, в которых Веб служит средой транспорта данных.
- Использование стандартов платформы XML для представления метаданных, описывающих свойства информационных ресурсов электронных библиотек. Для этих целей могут использоваться как средства самого языка XML (описание типов документов DTD), так и языковые средства стандартов XML Schema и RDF.
- Предоставление разработчикам электронных библиотек инструментальных средств систем баз данных нового класса (XML-ориентированных баз данных), обеспечивающих эффективную поддержку

коллекций информационных ресурсов XML и развитые возможности доступа к ним.

- Использование XML-ориентированных моделей данных в качестве интегрирующих моделей для интеграции данных в электронных библиотеках.
- Предоставление средств описания онтологий (стандарты RDFS, OWL) для электронных библиотек, позволяющих оперировать информационными ресурсами на семантическом уровне.

Заключение

Представленный в данной работе анализ тенденций развития технологий управления информационными ресурсами, составляющих обширную область информационных технологий, не является исчерпывающим. К сожалению, в силу ограниченности рамок статьи за его пределами осталось обсуждение активно развиваемых в последние годы технологий потоков данных [43] и сенсорных сетей [25], новых достижений в мобильных технологиях [27], весьма перспективной новой концепции пространств данных [23], которая, без сомнения, найдет применение в электронных библиотеках, оперирующих неоднородными информационными ресурсами и требующих разнообразных сервисов для их использования. Заслуживают также серьезного внимания те вызовы времени в рассматриваемой области, которые обозначены в Лоуэллском отчете [12].

Литература

[1] Калиниченко Л.А. Методология организации решения задач над множественными источниками информации. Труды 1-й международной научно-практической конференции «Современные информационные технологии и ИТ-образование, Москва, МГУ, 2005». – М.: Макс Пресс, 2005.

[2] Когаловский М.Р. Абстракции и модели в системах баз данных //СУБД, 4-5, 1998.

[3] Когаловский М.Р. Перспективные технологии информационных систем. - М.: ДМК, АйТи-Пресс, 2003. – 288 с.

[4] Когаловский М.Р. Стандарты XML и электронные библиотеки //Электронные библиотеки. ИРИО. – 2003. - Том 6 - Выпуск 2.

[5] Когаловский М.Р. Технологии XML и XML-данные. Базы данных и информационные технологии XXI века. Материалы международной научной конференции, Москва, 29-30 сентября 2003 г. – М.: РГГУ, 2004.

[6] Когаловский М.Р. Интеграция данных в информационных системах. Сб. трудов Третьей Всероссийской конференции “Стандарты в проектах современных информационных систем”, Москва, 23-24 апреля 2003 г.

[7] Кудашев Е.Б., Балашов А.Д. Организация информационной распределенной среды и интеграция спутниковых архивов. Труды 7-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Ярославль, 2005.

[8] Некрестьянов И., Пантелеева Н. Системы текстового поиска для Веб //Программирование. МАИК «Наука/Интерпериодика». – 2002. - №4.

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

[10] Персональный поиск Яndex. http://desktop.yandex.ru/

[11] Положение об автоматизированной системе российского сводного каталога (АС РСК) по научно-технической литературе. – М.: ГПНТБ России, 1995. – 14 с.

[12] Abiteboul S., Agrawal R., Bernstein P., Carey M., Ceri S., Croft B., DeWitt D., Franklin M., Garcia Molina H., Gawlick D., Gray J., Haas L., Halevy A., Hellerstein J., Ioannidis Y., Kersten M., Pazzani M., Lesk M., Maier D., Naughton J., Schek H., Sellis T., Silberschatz A., Stonebraker M., Snodgrass R., Ullman J., Weikum G., Widom J., and Zdonik S. The Lowell Database Research Self Assessment. Commun. of the ACM, v.48, no.5, 2005. (Первоначально отчет об Лоуэллской дискуссии, состоявшейся в июне 2003 года, был опубликован на сайте одного из ее участников – Дж. Грея, сотрудника компании Microsoft. http://research.microsoft.com/~gray/lowell/).

[13] Alexandria Digital Library Project. Alexandria Digital Earth Prototype (ADEPT). http://www.alexandria.ucsb.edu/adept/adept.html

[14] Antonioletti M., Krause A., Laws S., Paton N.W., Malaika S., Pearson D., Eisenberg A., Melton J. The WS-DAI Family of Specifications for Web Services Data Access and Integration. ACM SIGMOD Record, Vol. 35, No.1, March 2006.

[15] Berners-Lee T., Hendler J., and Lassila O. The Semantic Web. Scientific American, May 2001.

[16] Briukhov D.O., Kalinichenko L.A., Zakharov V.N., Panchuk V.E., Vitkovsky V.V., Zhelenkova O.P., Dluzhnevskaya O.B., Malkov O.Yu., Kovaleva D.A. Information Infrastructure of the Russian Virtual Observatory (RVO). 2nd Edit. IPI RAN, 2005.

[17] Calvanese D., Damaggio E., De Giacommo G., Lenzerini M., and Rosati R. Semantic data Integration in P2P Systems. K. Aberer et al. (Eds.): VLDB 2003 Workshop DBISP2P, LNCS 2944, 2004.

[18] Carey M., Blevins M., Takacsi-Nagy P. Integration, Web Services Style. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering/ IEEE CS, 2002.

[19] Cruz I.F. Semantic Data Integration. http://www.cs.uic.edu/~ifc/Talks/Minnesota/cruz.pdf

[20] Dublin Core Metadata Element Set Reference Description, Version 1.1, 1999-07-02. http://purl.org/dc/ documents/rec-dces-19990702.htm

[21] Eisenberg A., Melton J. Advancements in SQL/XML. ACM SIGMOD Record, Vol. 31, No. 2, June 2002.
[22] EMBL Nucleotide Sequence Database Statistics. http://www3.ebi.ac.uk/Services/DBStats. 2003.
[23] Franklin M., Halevy A., and Maier D. From Databases to Dataspaces: A New Abstraction for Information Management. ACM SIGMOD Record, Vol. 34, No.4, December 2005.
[24] Hanzen M., Madnick S., and Siegel M. Data Integration Using Web Services. In Proc. of Efficiency and Effictiveness of XML Tools and Techniques and data Integration over the Web. VLDB 2002 Workshop. LNCS 2590, Springer, 2003.
[25] Hellerstein J.M., Honh W., Madden S.R. The Sensor Spectrum: Technology, Trends, and Requirements. ACM SIGMOD Record, Vol. 32, No.4, December 2003.
[26] Hernandez T., Kambhampati S. Integration of Biological Sources: Current systems and Challenges Ahead. ACM SIGMOD Record, Vol. 33, No.3, September 2004.
[27] Imielinksi T. and Badrinath B.R. Wireless Mobile Computing: Challenges in Data Management, CACM 37, No. 10 (October 1994).
[28] Information Integration. IBM Corporation. http://www-306.ibm.com/software/data/integration/
[29] Kalinichenko L.A., Skvortsov N.A. Extensible Ontological Modeling Framework for Subject mediation. Труды Четвертой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Том 1. – Дубна: ОИЯИ, 2002.
[30] Laks V.S. Lakshmanan, Fereidoon Sadri. Information Integration and the Semantic Web. IEEE Computer Society Technical Committee Data Engineering Bulletin. December 2003, Vol. 26, No. 4.
[31] Lenzerini M. Data Integration: A Theoretical Perspective. Proc. of the ACM Symposium on Principles of Database Systems (PODS), 2002.
[32] Levy A.Y. Logic-Based Techniques in Data Integration. Logic-based Techniques in Data Integration. In: Logic Based Artificial Intelligence. Ed. by J. Minker. Kluwer Publishers, 2000.
[33] Lyman P. and Varian H. How much information? 2003. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/
[34] Manolescu I., Florescu D., Kossman D. Answering XML Queries over Heterogeneous Data Sources. Proc. Of the 27th VLDB Conference, Roma, Italy, 2001.
[35] Noy N.F. Semantic Integration: A Survey of Ontology-Based Approaches. ACM SIGMOD Record, Vol. 33, No. 4, Dec. 2004.
[36] Oracle Data Hub. Oracle, 2006. http://www.oracle.com/global/ru/data_hub/index.html
[37] OWL Web Ontology Language Guide. W3C Recommendation 10 February 2004. http://www.w3.org/ TR/2004/REC-owl-guide-20040210/
[38] Parinov S., Krichel T. RePEc and Socionet as partners in a changing digital library environment, 1997 to 2004 and beyond. Труды Шестой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Пущино, 2004.
[39] Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
[40] RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/
[41] RIF Use Cases and Requirements. W3C Working Draft 23 March 2006. http://www.w3.org/TR/2006/WD-rif-ucr-20060323/
[42] SPARQL Query Language for RDF. W3C Candidate Recommendation 6 April 2006. http://www.w3.org/TR/2006/CR-rdf-sparql-query-20060406/
[43] Stonebraker M., Cetintemel U., and Zdonik S. The 8 Requirements of Real-Time Stream Processing. ACM SIGMOD Record, Vol. 34, No.4, Dec. 2005.
[44] Text Retrieval Conferences (TREC). http://trec.nist.gov/
[45] The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html
[46] Updegrove A. The Semantic Web: an Interview with Tim Berners-Lee. The Consortium Info.org, June 2005. http://www.consortiuminfo.org/bulletins/ semanticweb.php
[47] WorldCat. Window to the World's Libraries. http://www.oclc.org/worldcat/

Trends in Evolution of Information Resources Management Technologies for Digital Libraries

Mikhail R. Kogalovsky

The elaborations of information resources collections for digital libraries, activities provided their support and access to these resources need all spectrum of key information management technologies used in modern information systems: database technologies, text search technologies as well as web technologies. So significant impact on digital libraries functionalities influenced by formed and aborning in last years trends of mentioned technologies evolution is not casual. Most significant of these trends are considered in the paper.

♣ Работа поддержана грантами РФФИ 04-07-90184-в, РФФИ 05-07-9025-в и РГНФ 06-02-12205-в.