

Стандарты XML и электронные библиотеки

М.Р. Когаловский

Институт проблем рынка РАН

В статье обсуждаются особенности электронных библиотек как нового класса информационных систем, рассматриваются наиболее злободневные технологические проблемы их развития, а также направления использования стандартов платформы XML в таких системах. Работа частично поддерживается РФФИ, проект 01-07-90444.

Введение

Благодаря развитой коммуникационной инфраструктуре и удобным возможностям представления информационных ресурсов, допускающим вместе с тем интеграцию данных из внешних источников, Всемирная паутина (или Веб, как ее теперь часто называют в отечественной литературе) стала средой, в которой поддерживается функционирование электронных библиотек и обеспечивается доступ к их ресурсам. По этим причинам процессы технологического переоснащения Веб на основе комплекса стандартов XML и его окружения [1] не могут не оказывать влияния на развитие технологий электронных библиотек. В свою очередь, электронные библиотеки стали одной из важных областей применения стандартов платформы XML. Главная задача этой статьи заключается в обсуждении наиболее важных направлений использования технологий XML в электронных библиотеках.

Электронные библиотеки и технологические проблемы их развития

Разработки и исследования электронных библиотек являются одним из актуальных направлений развития информационных систем в последние годы, привлекающим внимание специалистов различного профиля.

Специалисты в области библиотечного дела видят в электронных библиотеках новые возможности для совершенствования автоматизированных библиотечных систем, превращения их в публичные электронные библиотеки нового поколения с развитыми средствами представления разнообразных цифровых информационных ресурсов и доступа к ним, создаваемые с учетом необходимости интеграции издательских и библиотечных технологий.

Музейные работники получают в новых технологиях возможности сохранения национального культурного наследия и превращения его в общечеловеческое достояние благодаря обеспечению глобального доступа в среде Веб с помощью функционально развитых сервисов к создаваемым ими электронным коллекциям цифровых образов музейных экспонатов.

Сотрудники образовательных учреждений различных ступеней совместно с библиотечными специалистами ведут большую работу по созданию принципиально новой технологии интерактивного обучения (Digital Libraries Education, DLE) [2], основанной на новой

информационной инфраструктуре образовательного процесса, ядром которой должны стать электронные библиотеки. При этом предусматриваются средства адаптации программ обучения и его информационной поддержки к потребностям и возможностям конкретного учащегося без каких-либо ограничений на его возраст, географическое местоположение, распорядок дня и т.д.

Научных сотрудников, занимающихся исследовательской работой в различных областях знаний, технологии электронных библиотек привлекают возможностями обеспечения эффективного распространения результатов исследований в среде научного сообщества, поддержки научного сотрудничества коллективов исследователей, для которого не являются препятствием административные, географические и национальные границы. Информационные ресурсы электронных библиотек и специально разработанные для оперирования ими функциональные сервисы становятся основой исследовательских стендов в различных областях науки, заменяя натурные эксперименты экспериментами с моделями реальных сущностей, процессов или явлений [3].

Наконец, специалисты в области информационных систем, в свою очередь, рассматривают электронные библиотеки как новый класс информационных систем, базирующихся на самых передовых достижениях информационных технологий и технологий телекоммуникаций [4]. Разработки таких систем порождают разнообразные сложные теоретические и технологические проблемы, требующие своего исследования.

Функциональные возможности электронных библиотек варьируются в довольно широком диапазоне. Вообще говоря, предполагается, что они предоставляют пользователю глобальный доступ с помощью различного рода сервисов в среде Веб к их коллекциям цифровых информационных ресурсов, которые могут быть распределенными и в различных аспектах неоднородными. При этом могут обеспечиваться разнообразные возможности их интеграции на техническом и/или на семантическом уровне, развитые средства каталогизации и индексирования. В некоторых электронных библиотеках предусматриваются пользовательские интерфейсы с повышенным уровнем семантики, например, с возможностями семантического поиска требуемых информационных ресурсов, мультязыкового доступа и визуализации данных, средства персонализации пользовательских интерфейсов и т.д.

Разнообразный характер могут иметь и информационные ресурсы электронных библиотек - от традиционных библиотечных электронных каталогов до сложных интегрированных коллекций информационных ресурсов, включающих полнотекстовые документы, числовые данные, графические, аудио и видео ресурсы, геоданные и т.д., снабженных разнообразными метаданными.

Направления исследований и разработок в области электронных библиотек охватывают технологические, лингвистические, экономические, правовые, социальные и другие аспекты систем этого класса, а также методы и инструментарий создания поддерживаемых в них коллекций разнообразных цифровых информационных ресурсов.

Разрабатываемые в настоящее время электронные библиотеки базируются на передовых достижениях Веб-технологий (платформа XML, работы по созданию стандартов семантического Веб, в частности, языка описания онтологий и др.), технологий баз данных (объектные и объектно-реляционные базы данных, XML-ориентированные базы данных), технологий текстового поиска (полнотекстовый поиск, модели семантического поиска, новые подходы, ориентированные на текстовый поиск в Веб), достижениях в области методов

представления и обнаружения знаний, технологий создания и поддержки электронных публикаций, моделирования данных и метаданных. При этом уделяется особое внимание поддержке метаданных и их использованию для повышения эффективности поиска требуемых информационных ресурсов, для разработки приложений высокого уровня, решения проблем интеграции ресурсов и для других целей.

К числу наиболее злободневных технологических проблем развития электронных библиотек можно отнести:

- Развитие методов представления информационных ресурсов электронных библиотек.
- Определение состава метаданных, независимых от применений и специфических для различных сфер приложения, разработка средств их представления.
- Развитие новых подходов к каталогизации информационных ресурсов электронных библиотек.
- Разработка техники индексирования информационных ресурсов различной природы (текст, аудио, видео и т.п.), методов поиска и обнаружения релевантных ресурсов, а также принципов и средств их анализа.
- Интеграция неоднородных коллекций информационных ресурсов на логическом и семантическом уровне.
- Разработка подходов к интеграции метаданных и методов их реализации.
- Создание функционально развитых пользовательских интерфейсов (многоязыковый доступ, визуализация данных, персонализация функций, поддержка семантического уровня общения пользователей с системой).
- Эффективное использование новых Веб-технологий, основанных на стандартах платформы XML.
- Исследование архитектурных аспектов электронных библиотек.
- Обеспечение безопасности информационных ресурсов электронных библиотек.

Созданию прототипов электронных библиотек для исследования перечисленных проблем и различных подходов к их решению посвящены многочисленные исследовательские проекты, выполняемые во многих научных центрах. Активно ведутся также разработки действующих систем. В ряде этих проектов и практических разработок успешно используются стандарты платформы XML. В следующем разделе кратко обсуждаются основные направления использования функциональных возможностей этих стандартов в рассматриваемой области.

Стандарты XML в электронных библиотеках

Поскольку, как уже отмечалось, Веб стал "средой обитания" электронных библиотек, радикальные технологические сдвиги, которые происходят в этой среде, связанные, прежде всего, с активно развиваемыми консорциумом W3C технологиями и стандартами платформы XML [1, 5-7], оказывают существенное влияние на развитие информационных систем этого класса.

Можно выделить следующие наиболее важные направления использования средств платформы XML в разработках электронных библиотек:

- *Представление коллекций электронных информационных ресурсов в электронных библиотеках.* Разметка публикуемых в Веб научных публикаций и экспериментальных

данных с помощью языка XML обеспечивает новое качество создаваемых информационных ресурсов, отражая их содержательную структуру. Это не только улучшает их восприятие пользователем, но и обеспечивает автоматическую верификацию целостности содержательной структуры документов в хранимых коллекциях.

- *Обеспечение навигационного доступа к информационным ресурсам с помощью средств, привычных для пользователей Веб.* Аналогично технологиям HTML, представление информационных ресурсов в Веб средствами стандартов XML обеспечивает навигационный доступ к XML-ресурсам электронных библиотек с помощью привычных Веб-браузеров, поддерживающих эти стандарты.
- *Обеспечение интерфейса языка запросов для доступа к информационным ресурсам, представленным как XML-документы, на основе элементов их содержания.* Создание стандарта языка запросов XQuery для XML-данных, разработка которого ведется консорциумом W3C, позволит создавать в Веб, а также в системах электронных библиотек поисковые сервисы для обработки таких запросов. Результатами обработки запросов при этом могут быть не только полные XML-документы, но и их фрагменты, удовлетворяющие заданному критерию поиска, а также производные документы, представляющих различного рода трансформации документов, найденных в результате поиска.
- *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 не только средства самого языка XML, но и языковые средства стандартов XML Schema и RDF. Описание XML-документов средствами XML Schema позволяет осуществлять более тонкую верификацию целостности представленных XML-документов. Спецификация содержания документов средствами стандарта RDF дает возможность семантического поиска информационных ресурсов в среде, поддерживающей такие метаданные. Структурированное представление метаданных коллекций информационных ресурсов электронных библиотек с помощью указанных средств позволяет автоматизировать обработку метаданных. Например, во многих системах поиск нужных элементов в коллекции информационных ресурсов сводится к поиску в коллекции XML-документов, представляющих собой метаописания отдельных элементов рассматриваемой коллекции.
- *Предоставление разработчикам электронных библиотек инструментальных средств систем баз данных нового зарождающегося класса (XML-ориентированных баз данных), обеспечивающих эффективную поддержку коллекций информационных ресурсов XML и развитие возможности доступа к ним.*
- *Обеспечение интеграции информационных ресурсов из различных независимых источников.* Цель состоит при этом в обеспечении единого пользовательского интерфейса для доступа к ним. Интеграция информационных ресурсов в электронных библиотеках возможна на различных уровнях - физическом, логическом и семантическом. Интеграция на физическом уровне требует обеспечения единого формата представления информационных ресурсов и метаданных, заимствованных из различных

источников. "Логическая" интеграция предусматривает единое представление интегрируемых информационных ресурсов и/или описывающих их метаданных в терминах некоторой интегрирующей модели данных и доступ к ним с помощью единого интерфейса, основанного на этой модели. Во многих разработках в качестве такой модели используется язык XML как язык описания данных в сочетании с каким-либо из языков платформы XML и ее окружения - для манипулирования данными в XML-представлении. Для этой цели чаще всего используются языки XPath, XSLT, XQuery. В качестве интегрирующей модели часто используется также объектная модель данных, определяемая стандартом DOM. Необходимыми компонентами архитектуры систем, обеспечивающих логическую интеграцию информационных ресурсов, являются механизмы отображения модели данных (метаданных) источника ресурсов в интегрирующую модель данных. В некоторых случаях разработчики электронных библиотек ограничиваются лишь интеграцией метаданных, их описывающих. Примером могут служить корпоративные каталоги распределенных информационных ресурсов, представленные средствами языка XML. Интеграция информационных ресурсов на семантическом уровне является весьма сложной задачей. Системы, обеспечивающие такие возможности, используют онтологические спецификации предметной области электронной библиотеки и источников ресурсов, методы интеграции онтологий, технику отображения информационных ресурсов в интегрирующую модель с помощью адаптеров и посредников. Исследовательские проекты, разрабатываемые в этой области, используют в качестве интегрирующей модели данных мощные модели представления знаний. В электронных библиотеках, обеспечивающих семантическую интеграцию информационных ресурсов, могут использоваться стандарты платформы XML - RDF, RDFS, OWL и др.

- Семантическая интеграция информационных ресурсов в электронных библиотеках тесно примыкает к проблеме *создания новой интеллектуализированной среды представления информационных ресурсов электронных библиотек следующего поколения* на основе инструментария семантического Веб (стандартов RDF, RDFS, языка описания онтологий OWL и др.), активно разрабатываемого консорциумом W3C.

Заключение

Использование технологий XML в разработках электронных библиотек является весьма перспективным, особенно в библиотеках научного назначения, многие из которых оперируют неоднородными коллекциями информационных ресурсов и используют сложные технологии их анализа.

На основе XML созданы языки разметки информационных ресурсов в математике, химии, астрономии, геоинформатике и в других областях знаний, в области электронного бизнеса, и они довольно широко применяются на практике. Предложены различные подходы к структурированному представлению библиографической информации с помощью языка XML и к созданию на этой основе электронных библиотек с поисковыми машинами, оперирующими такой информацией. В ряде разработок используются средства платформы XML для представления метаданных, определенные стандартами RDF, RDFS, OWL.

Имеются многочисленные примеры локальных и распределенных систем электронных библиотек, в которых используются технологии XML. Как было показано выше, стандарты XML

обеспечивают новые важные возможности для представления, интеграции и оперирования информационными ресурсами в таких системах.

Литература

1. World Wide Web Consortium. <http://www.w3.org/>
2. Kalinichenko L. Digital Libraries in Education: State of the Art. Обзорный доклад. Сб. трудов Третьей Всероссийской конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". Петрозаводск, 11-13 сентября 2001 г. - Петрозаводск: Карельский научный центр РАН, 2001. - С. 172-173.
3. Когаловский М.Р. Особенности научных электронных библиотек. Тезисы докладов научной конференции, посвященной 10-летию РФФИ "Электронные библиотеки и информационное обеспечение научной деятельности", Москва, 25-26 ноября 2002 г. Российский фонд фундаментальных исследований, Центр фотохимии РАН, 2002.
4. Когаловский М.Р., Новиков Б.А. Электронные библиотеки - новый класс информационных систем. МАИК/Наука "Интерпериодика". Программирование. - 2000. - № 3.
5. Когаловский М.Р. Стандарты платформы XML и базы данных. Обзорный доклад. Сб. трудов Третьей Всероссийской конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", Петрозаводск, 11-13 сентября 2001 г. - Петрозаводск: Карельский научный центр РАН, 2001. - С. 108-117.
6. Когаловский М.Р. Энциклопедия технологий баз данных. - М.: Финансы и статистика, 2002. - 800 с.
7. Когаловский М.Р. Перспективные технологии информационных систем. - М.: ДМК, 2003. - 288 с.