

Электронные библиотеки - новый класс информационных систем

(От составителей выпуска)

М.Р. Когаловский
Институт проблем рынка РАН
e-mail: kogalov@cemi.rssi.ru

Б.А. Новиков
Санкт-Петербургский госуниверситет
e-mail: borisnov@acm.org

В статье кратко обсуждаются истоки нового направления в области информационных систем, связанного с так называемыми электронными библиотеками. Интенсивные исследования и разработки в этом направлении начались в первой половине 90-х годов. Рассматриваются особенности систем электронных библиотек и проблематика связанных с ними исследований, представлены помещенные в этом специальном выпуске журнала статьи отечественных авторов, подготовленные по материалам их докладов на состоявшейся в октябре 1999 г. первой российской национальной научной конференции, посвященной электронным библиотекам.

Информационные системы, которые стали называть *электронными библиотеками*, появились несколько лет назад благодаря широкому распространению Интернет, стремительному росту емкости запоминающих устройств, а также значительным достижениям в области Web-технологий, в технологиях 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, на наш взгляд, он неточно отражает смысл англоязычного источника ("цифровые библиотеки"), авторы которого стремились подчеркнуть в нем определяющую роль цифровых технологий в информационных системах такого рода.

В соответствии с Концепцией межведомственной программы "Электронные библиотеки России" [1], "под электронной библиотекой понимается распределенная информационная система, позволяющая надежно накапливать, сохранять и эффективно использовать разнообразные коллекции электронных документов, доступные в удобном для пользователя виде через глобальные сети передачи данных".

Это определение, хотя и в некоторой степени неявно, характеризует в самом общем виде основные функциональные возможности систем электронных библиотек - доступность под-

держиваемых информационных ресурсов в интерактивном режиме, предоставление различных услуг (сервисов), связанных с их использованием, распределенная архитектура с сетевым доступом пользователей. Вместе с тем, как нам кажется, оно нуждается в нескольких важных уточнениях.

Прежде всего, не следует понимать в нем *электронный документ* традиционным для документальных информационных систем образом. Здесь электронный документ - это, вообще говоря, оцифрованный информационный ресурс любого рода как по его содержанию (текст на естественном языке, музыкальное произведение, карта местности, исходный текст программы, чертеж изделия или даже коллекция таких ресурсов и т.п.), так и по характеру среды представления - текст, графика, аудио, видео. Документы могут различаться и по степени структурированности (от "свободного" текста на естественном языке без выделения его элементов до формализованного представления информации из баз данных).

Говоря о функциях накопления и хранения *коллекций* электронных документов, не следует забывать о так называемых *виртуальных коллекциях* - коллекциях, составленных из информационных ресурсов, принадлежащих другим коллекциям или другим электронным библиотекам, без создания новых их копий для данной коллекции. Речь о хранении и накоплении в таких случаях, естественно, может идти лишь условно, поскольку при этом вместо документов хранятся лишь ссылки на них.

Далее, в приведенном определении подчерки-

вается *распределенный характер* системы и возможность доступа к ее информационным ресурсам через глобальные коммуникационные сети. Эти свойства, безусловно, необходимы, если речь идет об электронных библиотеках национального или международного масштаба, на создание которых нацелена межведомственная программа. Однако технологии электронных библиотек находят применение и в более ограниченном масштабе. Так, могут, например, создаваться электронные библиотеки отдельных организаций или даже личные библиотеки для поддержки научных исследований. Масштабы конкретной электронной библиотеки, характер доступа к ее информационным ресурсам и круг пользователей зависят, главным образом, от социальной значимости поддерживаемых в ней коллекций информационных ресурсов.

Наконец, требует пояснения понятие *коллекции информационных ресурсов*, представляющей основную форму организации информационных ресурсов в электронных библиотеках. Однако не будем касаться здесь этой темы, поскольку ей посвящена специальная статья [2].

Интенсивные исследования и разработки в области электронных библиотек во многих странах были стимулированы объявленной в США в конце 1993 г. по инициативе Национального научного фонда (NSF), Агентства перспективных исследований Министерства обороны (ARPA) и Национального агентства по космическим исследованиям (NASA) программой "Digital Libraries Initiative" (DLI) [3], 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 Стенфордскому университету (интеграция механизмов неоднородных сервисов для обеспечения унифицированного доступа к различным сетевым коллекциям информационных ресурсов) [4]. К 1998 г., когда началась вторая фаза работ [5] по программе и было предоставлено уже 24 гранта, был достигнут значительный

прогресс в этой области.

Для поддержки исследований и разработок в области электронных библиотек были организованы также научные форумы, которые обеспечили контакты и обмен идеями в сообществе ученых и специалистов, работающих в рассматриваемой области. С 1995 г. стала ежегодно проводиться конференция IEEE 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 (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. Более подробные сведения об этих конференциях можно найти в библиографии по информатике М. Лея [6].

Кроме того, издается ряд специализированных периодических изданий по проблематике электронных библиотек. Наиболее авторитетные из них - выпускаемый ежеквартально с 1997 г. 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.

В настоящее время электронные библиотеки представляют собой область интенсивных исследований, проводимых не только в США, но и во многих других странах. К числу основных проблем, которыми озабочено научное сообщество в этой области, относятся интеграция неоднородных информационных ресурсов, в том числе, поддерживаемых средствами различных технологий, модели данных и архитектуры для поддержки такой интеграции, управление метаданными и их использование в интеграции и обработке информационных ресурсов, пользовательские интерфейсы и их персонализация, методология формирования и поддержки коллекций информационных ресурсов, обеспечение многоязыковой среды, интероперабельность информационных ресурсов, методы обнаружения необходимой информации, классификация, кластеризация и фильтрация информационных ресурсов, их визуализация, эффективные методы по-

иска, создание исследовательских испытательных стендов, технологии электронных публикаций, защита прав интеллектуальной собственности на информационные ресурсы, социально-экономические последствия использования электронных библиотек, создание новых классов их приложений. Этот список, конечно же, - далеко не исчерпывающий.

В последние годы электронные библиотеки стали вместе с тем и сферой активной практической деятельности. Многие научные издания дублируются в электронной форме и представляются в виде Web-сайтов, ведется работа по созданию электронных версий научных журналов за прошлые годы. Появились крупные электронные библиотеки, такие как ACM Digital Library [7], Антология ACM SIGMOD [8] и электронная библиотека IEEE Computer Society [9], которые содержат уже многие сотни томов трудов крупных конференций и периодических изданий, охватывают по некоторым изданиям период более двух десятилетий.

Наряду с представленными в электронной форме публикациями традиционного типа, электронные библиотеки используются также для хранения и распространения других классов документов, прежде всего, мультимедийных. Для примера можно привести уже упоминавшийся электронный журнал по проблематике баз данных DiSC [10], выпускаемый под эгидой ACM SIGMOD в двух редакциях - для Web и на компакт-дисках. В журнале публикуются, в частности, выступления известных ученых на конференциях, представленные средствами видео и аудио.

Начали создаваться крупные электронные библиотеки для поддержки научных исследований, в том числе, и в рамках международного научного сотрудничества. Одним из отечественных примеров таких систем может служить электронная библиотека, созданная и развиваемая в Институте цитологии и генетики СО РАН [11].

Еще одно важное направление в разработке электронных библиотек связано с представлением динамически получаемых документов (например, ответов на запросы к базам данных). В этом контексте электронная библиотека рассматривается как метод структуризации и организации любой информации, опубликованной в Интернет.

В нашей стране исследования и разработки под флагом электронных библиотек ведутся с 1998 г. Начало им было положено состоявшимся в апреле 1998 г. Российско-американским семинаром по электронным библиотекам и объявленным совместно РФФИ и Министерством науки РФ конкурсом исследовательских проектов, посвященных этой проблематике. В развитие этой деятельности в 1999 г. разработана также упо-

минавшаяся ранее межведомственная программа "Электронные библиотеки России" [1].

Ежегодным научным форумом по электронным библиотекам национального масштаба в нашей стране станет цикл конференций, первая из которых состоялась в Санкт-Петербурге в октябре 1999 г. с участием ученых из США и Германии [12]. В рамках одной конференции невозможно охватить все аспекты создания и использования электронных библиотек. На первой национальной российской конференции обсуждались следующие темы:

- Теоретические аспекты создания электронных библиотек (модели данных, языки представления, методы и архитектуры интеграции информационных ресурсов, алгоритмы и методы проверки и поддержки целостности).
- Методы семантического анализа документов на естественном языке и их приложения.
- Методы и средства поиска информационных ресурсов в Интернет и генерации тематических коллекций.
- Систематика коллекций информационных ресурсов в электронных библиотеках.
- Создание конкретных коллекций, при этом обсуждались, главным образом, различные научные коллекции.
- Электронизация традиционных библиотечных технологий.

Ряд представленных на конференции работ отечественных авторов, которые, по нашему мнению, будут представлять интерес для читателей, был отобран для этого специального выпуска и доработан авторами для журнальной публикации.

В статье Л.А. Калиниченко, Н.А. Скворцова, Д.О. Брюхова, Д.В. Кравченко и И.А. Чабан предлагается подход к проектированию виртуальных электронных библиотек, интегрирующих слабоструктурированные данные существующих Web-сайтов и обеспечивающих персонализацию представления информационных ресурсов библиотеки для пользователей.

Работа Е.А. Горшковой, И.С. Некрестьянова, Б.А. Новикова и Е.Ю. Павловой посвящена проблеме согласованности данных, получаемых из неоднородных независимых источников, таких, как узлы WWW, публикующие информацию из баз данных. Используя слабоструктурированную модель данных как средство абстрактного представления таких источников данных, авторы обсуждают возможные пути решения

проблемы согласованности и более детально развывают подход, основанный на модели многоуровневых транзакций.

Методологические вопросы разработки коллекций информационных ресурсов в электронных библиотеках рассматриваются в статье М.Р. Когаловского. Обсуждаются, в частности, основные общие свойства коллекций, применяемые в них методы систематизации информационных ресурсов, вопросы генезиса коллекций, роль в них метаданных, специфические особенности научных коллекций. Приводится также обзор перспективных технологий и стандартов, которые используются для создания, поддержки и использования коллекций.

В работе А.Г. Марчука и А.Е. Осипова рассматривается проблема создания системы идентификации электронных документов и коллекций в глобальной среде Интернет на основе объектно-ориентированного подхода, обсуждается экспериментальная реализация такой системы для XML-документов.

Некоторый подход к решению проблем создания интеллектуального сетевого робота, обеспечивающего сбор информационных ресурсов для тематических коллекций, предлагается в статье Е.В. Романовой, М.В. Романова и И.С. Некрестьянова.

В статье М.Р. Когаловского, Е.Н. Ефимовой, Т.А. Рыбиной и В.Б. Брахина предложены формальные методы и алгоритмы верификации макроструктуры существующих Web-сайтов, основанные на реляционной алгебре и теории графов.

И. Кураленок и И. Некрестьянов представляют в своей работе метод классификации естественно-языковых документов относительно заданного множества тем, основанный на латентно-семантическом анализе этих документов, который позволяет извлечь сведения о семантических зависимостях между содержащимися в них терминами.

М.А. Осипов, О.Л. Мачульский и Л.А. Калинин рассматривают в своей статье отображение структуры документов, представленных средствами расширяемого языка разметки XML, в объектную модель проекта СИНТЕЗ, разрабатываемого в Институте проблем информатики РАН. Эта модель служит в проекте канонической моделью данных электронной библиотеки, интегрирующей неоднородные распределенные коллекции научных информационных ресурсов.

Наконец, в статье Н.А. Колчанова, Ф.А. Колпакова и др. обсуждаются функциональные возможности созданной в Институте цитологии и генетики СО РАН крупной научной электронной библиотеки по структурам и функциям ДНК, РНК и белков, ее информационные ресурсы, а также методы их интеграции.

В одном выпуске журнала в связи с ограниченным его объемом невозможно опубликовать все перечисленные статьи. Некоторые из них будут опубликованы в ближайших последующих выпусках.

Работы в области электронных библиотек в нашей стране только начинают разворачиваться. Редколлегия считает, что эта проблематика заслуживает самого серьезного внимания и приглашает читателей к ее обсуждению на страницах журнала.

Список литературы

- [1] Концепция межведомственной программы "Электронные библиотеки России". Проект, версия от 19.07.99.
- [2] Когаловский М.Р. Систематика коллекций информационных ресурсов в электронных библиотеках. Статья в настоящем выпуске.
- [3] Research on Digital Libraries. NSF 93-141. Program Guideline. September 16, 1993. <http://www.nsf.gov/pubs/stis1993/nsf93141/nsf93141.txt>
- [4] Digital Libraries. Comm. of ACM. Special Issue. Vol. 38, No. 4, April 1995.
- [5] Digital Libraries Initiative - Phase 2. NSF 98-63. July 15, 1998. <http://www.nsf.gov/pubs/1998/nsf9863/nsf9863.htm>
- [6] Computer Sciences Bibliography. Universität Trier. <http://dblp.uni-trier.de>
- [7] ACM Digital Library. <http://www.acm.org/dl/>
- [8] The ACM SIGMOD Anthology. <http://www.acm.org/sigmod/dblp/db/anthology.html>
- [9] Welcome to the Computer Society Digital Library. <http://computer.org/epub/>
- [10] ACM SIGMOD DiSC. <http://www.cs.wpi.edu/~ifc/disc/>
- [11] Колчанов Н.А., Колпаков Ф.А. и др. ГЕНЭКСПРЕСС: электронная библиотека по структурам и функциям ДНК, РНК и белков. Статья в данном выпуске.
- [12] Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Первой Всероссийской научной конференции. С.-Петербург, 19-22 октября 1999 г. Изд. С.-Петербургского университета, 1999.