

Электронные библиотеки

77-48211/500497

11, ноябрь 2012

Белоус В. В.

УДК 027:004:37

Россия, МГТУ им. Н.Э. Баумана

belous@wwcdl.bmstu.ru

Первые упоминания об электронных библиотеках (ЭБ) стали появляться в начале 70-х годов. Принято считать, что начало работ по созданию ЭБ положила общественная инициатива, известная под названием «Проект Гуттенберга» (Project Gutenberg, PG). Это пионерское начинание стартовало в 1971 году и имело целью оцифровку и сохранение в электронном формате известных литературных текстов на английском языке, находящихся в свободном обороте. К 2010 данная коллекция насчитывала более 40 000 единиц хранения, главным образом на английском языке. В русской части сети Интернет первой и долгое время единственной было собрание оцифрованных изданий, известное под названием библиотека Максима Мошкова.

В начале инициаторы понимали оцифровку печатных текстов только как способ хранения и резервного копирования. Термин «электронная библиотека» не использовался, а круг потребителей цифрового контента ограничивался специалистами и случайными клиентами. Сейчас электронные библиотеки превратились в один из самых мощных информационных и образовательных ресурсов, а их потребителей можно встретить в любых strатах современного общества: от домохозяек до творческой и научной элиты.

Понятие «электронная библиотека» не имеет пока сложившейся традиции понимания и устоявшегося определения. Приведем несколько наиболее содержательных толкований этого феномена (приводится по [6]):

- «Автоматизированная электронная система, включающая совокупность первичных источников информации и разного уровня организации вторичную информацию, полученную на основе обработки этих (и только этих) первичных источников.
- Тематически ориентированная (или структурированная иным образом) система доступа к удаленным или локальным электронным ресурсам, способная обслуживать электронными ресурсами локальных или удаленных пользователей.
- Локальные или распределенные электронные ресурсы, объединенные единой идеологией структуризации и доступа.
- Система, реализующая унифицированный подход к производству, хранению и организации разнообразной информации с целью поиска, анализа и доступа к ней с использованием глобальных компьютерных сетей».

С нашей точки зрения современное состояние электронных библиотек наиболее точно описывается следующим определением. Электронная библиотека – это автоматизированная информационная система, предназначенная для надежного хранения и удобного использования коллекций электронных документов (текстовых, графических, звуковых, мультимедийных и др), обладающая средствами актуализации и поддержки целостности репозитория и предоставляющая доступ к своему хранилищу при помощи единого Web-интерфейса.

В качестве синонимов термина «электронная библиотека» иногда используются словосочетания «цифровая библиотека» (digital library) и «виртуальная библиотека» (virtual library).

Как следует из приведенного определения, основными целями электронной библиотеки являются интеграция информационных ресурсов и

эффективная навигация в информационном пространстве ЭБ. Под интеграцией понимается такое объединение источников, которое превращает совокупность данных в некоторую информационную целостность. При этом объединение ресурсов необязательно должно осуществляться физически, оно может быть и виртуальным. Эффективная навигация – это возможность удобного поиска релевантных данных во всем доступном информационном пространстве ЭБ. Некоторые исследователи в области электронной библиографии считают, что любая ЭБ должна обладать свойством профильности и включать в себя подсистемы каталогизации и инвентаризации электронных документов [1].

Единицы хранения ЭБ – электронные документы могут представлять собой текстовые произведения, изображения, фонограммы, мультимедийные ролики, базы данных, гипертексты, чертежи, трехмерный контент и другие типы электронных данных. Организация информационного пространства как совокупности объектов и однозначная идентификация последних необходимы для обеспечения эффективной навигации и выполнения любых видов информационных поисков. Для описания и организации совокупности независимых объектов в единое целое информационное поле библиотеки необходимо пополнить метаданными, например библиографическими данными, если речь идет о книге.

Опыт разработки и эксплуатации электронных библиотек позволил выявить общие типологические функции, характерные для этого класса информационных систем:

- поддержка системы гипертекстовых и гипермедийных связей различных типов;
- навигация во всем доступном информационном пространстве;
- наглядное представление логической структуры информационного пространства;
- информационный поиск с использованием лексики русского языка и языков, основанных на латинском алфавите. Использование

логических операторов подстановочных символов для составления сложных запросов;

- символьный поиск – разновидность информационного поиска, где поисковым образцом служит последовательность допустимых символов, например, дата;
- атрибутивный поиск – информационный поиск объектов по значениям их характеристик, например, книги в формате PDF;
- просмотр содержания информационного объекта и его структуры, например, ссылок, сносок оглавления и т.д.;
- многооконный режим работы с возможностью одновременного вывода на экран нескольких различных документов;
- протоколирование сеанса работы пользователя с возможностью перехода в любое из ранее существовавших состояний системы;
- кастомизация системы;
- установка закладок в тексте информационных объектов и средства обработки закладок (сортировка, переименование, удаление и пр.);
- экспорт информации из системы с указанием на источник;
- сортировка списков и перечислений по значениям любого доступного поля;
- подсистема, обеспечивающая продажу электронных издания и проведение платежей для коммерческого фрагмента электронной библиотеки.

С позиции системного администратора ЭБ должна быть открытой системой, доступной для изменения состава информационных объектов, реорганизации структуры информационного пространства, настройки системы внешних связей.

Электронные библиотеки являются способом структурирования и использования цифровых информационных ресурсов. По определению энциклопедического словаря, ресурсы – это наличные запасы продукта или резервные средства, которые могут быть использованы при необходимости

или в чрезвычайной ситуации. Информационные ресурсы (ИР) – объемы информации, записанной на некотором носителе, которая пригодна для сохранения и практического использования. Информационные ресурсы являются важнейшим видом обеспечения жизнедеятельности, наравне с финансовыми, сырьевыми, материально-техническими ресурсами.

Все многообразие существующих информационных ресурсов можно, разделить на группы по их целевому назначению [7].

- *Универсальные.* Аккумулируют электронные версии текстовых документов, относящихся к любым общественно значимым темам.
- *Справочные.* Информационные ресурсы, содержащие краткие сведения научного и прикладного характера. Обычно такие ИР не предназначены для последовательного чтения, поэтому они должны обладать мощными средствами поиска релевантных контекстов. В Рунет наиболее крупные хранилища этого типа – «Рубрикон», «Кирилл и Мефодий».
- *Правовые.* Наиболее крупными изданиями в этой области являются «КонсультантПлюс», «ГАРАНТ», «Кодекс». На рынке правовых электронных систем активно работают такие фирмы, как INTRALEX, «Информационные системы и технологии», компания «1С» и др.
- *Учебные:* В настоящее время в нашей стране наблюдается ситуация фрагментарности электронных учебных ресурсов. Многие учебные заведения владеют ресурсами такого типа или издают диски учебного содержания, но не существует единого центра, аккумулирующего электронные публикации этого направления. Среди ресурсов, предоставляющих свои услуги на коммерческой основе, лидирующее положение занимает Интуит.
- *Научные:* Научные ИР представлены в Рунет достаточно широко. Главным образом, они представляют собой библиографические, реферативные и полнотекстовые базы данных, содержащие различные виды документов: реферативные журналы, диссертации, авторефераты,

книги, нормативно-техническую документацию, патентную информацию, периодику и др. В настоящее время в мире издается более 100 тыс. научных журналов, причем ежегодно появляются новые издания. Как правило, содержание научных журналов представляется на сайте издательства или издающей организации в Интернет, а многие журналы имеют свои полнотекстовые электронные аналоги. Некоторые зарубежные издательства предоставляют доступ к полным текстам своих журналов через серверы других организаций.

В России одним из известных информационных агентств, предоставляющим библиотекам, университетам, институтам, доступ к научной и образовательной информации в электронной форме, является «Национальный электронно-информационный консорциум» (НЭИКОН). Основная цель НЭИКОН – сделать доступной актуальную научную информацию для российских организаций, а также упорядочить и оптимизировать условия подписки на различные ресурсы по гуманитарным и естественным наукам, предлагаемые как зарубежными издательствами и агентствами, так и российскими поставщиками информации.

Другое крупное информационное агентство, предоставляющее российским библиотекам доступ к информационным ресурсам, – научная электронная библиотека E-LIBRARY. Фонды библиотеки содержат выпуски научных журналов ведущих зарубежных издательств: Elsevier Science Kluwer Academic Publishers, Springer, Blackwell, Royal Society Of Chemistry, Institute Of Physics (London), Academic Press и др. Библиотека включает в себя базы данных научного цитирования Института научной информации США, базы данных по математике и медицине, журналы некоторых российских издательств, число которых постоянно пополняется.

Форматы электронных изданий

По ГОСТ 7.83-2001 «Электронные издания. Основные виды и выходные сведения», электронный документ определяется как

зафиксированная на материальном носителе информация в машиночитаемой форме с метаданными, позволяющими ее идентифицировать. Не все форматы электронных документов, находящиеся в настоящее время в обращении отвечают этому требованию.

Перечислим самые распространенные форматы, подходящие для публикации в ЭБ:

- простой текст (plain text) – TXT (.txt);
- текстовые форматы с оформлением и графическими включениями – HTML (.htm, .html), XML (.xml, .xhtml), открытый формат электронных книг Electronic Publication (.epub), OpenDocument, Microsoft Document (.doc, .docx), RTF (.rtf), FictionBook (.fb2), TeX (.tex, .dvi), PDF (.pdf), Microsoft HTMLHelp (.chm), eReader (.lit), PostScript (.ps, .eps), ExeBook, Mobipocket (.prc) и др.;
- графические растровые – TIFF (.tif), JPEG (.jpg), GIF (.gif) и пр.
- форматы сканированных документов (книг, журналов, рукописей и пр.) с включением формул, схем, рисунков и рукописных символов – DjVu, PDF;
- мультимедиа и аудиокниги – SWF (.swf), EXE (.exe);
- специализированные форматы для книг в виде java-мидлетов, предназначенных для просмотра на мобильных устройствах (смартфоны, телефоны с поддержкой java, КПК) – например, J2ME;
- форматы книг в виде приложений под определённую операционную систему, включающие в себя контент и специальную программу просмотра.

Форматы хранения размещаемых в электронных библиотеках документов можно разделить на две категории – форматы, ориентированные на чтение текста в режиме онлайн и форматы, предназначенные для скачивания на компьютер читателя. Это деление весьма условно, поскольку современные программные и технические средства снимают различие между

ними, предоставляя способ использования электронного документа на усмотрение пользователя.

Форматы, предназначенные преимущественно для скачивания, – заархивированный TXT, RTF, DOC, Mobipocket, FictionBook. Последний формат создан специально для хранения литературных произведений. Он получил широкое распространения на устройствах, предназначенных специально для чтения с экрана, – электронных ридерах. Формат ориентированный главным образом на просмотр – это HTML и его многочисленные версии и диалекты (XML, XHTML и др.).

Многие форматы разрешают любую политику использования, удобную для потребителя в данной ситуации. Таковым является, например, PDF. Все современные браузеры (Google Chrome, Firefox, Explorer и пр.) располагают встроенной программой просмотра PDF-файлов, поэтому документ этого типа можно скачать на пользовательский компьютер или вывести для чтения непосредственно в окне браузера.

Выбор формата диктуется, кроме того, традицией или аргументами, которые в наше время потеряли силу. Например, первая электронная библиотека – Проект «Гутенберг» – требует в файлах для скачивания использовать только текстовый формат в архивах ZIP. Основанием для такого выбора является защита от потери специализированных программ просмотра. Библиотека Мошкова выкладывает тексты в формате TXT. Это дань традиции. Во время появления библиотеки в 1994 году скорости работы в сети Интернет были весьма низки, а основным носителем являлась дискета, полностью потерявшая актуальность в наше время.

Технические документы и книги, насыщенные математическими формулами и сложными схемами, принято хранить в форматах PDF, DjVu и в форме растровой графики.

Кроссплатформенный формат PDF обладает очень большими возможностями. Он позволяет сохранять в одном файле текст, растровую, векторную и трехмерную графику, мультимедийные ролики, шрифты,

внутренние и внешние ссылки, формы, мультимедийный и трехмерный контент, JavaScript-включения и другие типы данных. Допускает аутентичное воспроизведение опубликованного произведения. Если в ЭБ должны быть точно воспроизведены все без исключения части книги: структура, орфография и пунктуация, распределение материала по страницам, иллюстративный ряд, шрифтовое оформление и другие особенности издания, то PDF не имеет альтернативы.

Долгое время формат не пользовался популярностью. Это объясняется несколькими объективными причинами:

- большими размерами файлов;
- платностью программного обеспечения фирмы-разработчика Adobe;
- наличием конкурентноспособных альтернатив, например, Envoу и Common Ground Digital Paper.

В наше время все эти проблемы преодолены, и формат становится все более популярным для публикации в сети сложноструктурированных и гетерогенных текстов.

Для публикации в ЭБ сканированных документов (книг, журналов, рукописей и пр.) применяется формат DjVu. Он обладает несколькими особенностями, делающими его уникальным среди других форматов, используемых для публикации электронных документов.

Документ в формате DjVu разбивается на три вспомогательных слоя, каждый из которых обрабатывается по-своему. Это техническая операция, необходимая для достижения компромисса между высоким сжатием и сохранением аутентичности оригинала. Слойная структура документа пользователю не видна. Для него фон, текст, а также графические и формульные включения располагаются на одном доступном для обозрения слое. При записи в DjVu-файлы используется несколько алгоритмов сжатия: с потерями и без потерь. Специальный алгоритм программы-конвертера выполняет расслоение сканированного оригинала. Программа пытается

выделить фон изображения и содержательные текстовые и графические фрагменты. Для текста и векторной графики применяются методы сжатия без значительных графических деформаций. При этом текст делается контрастнее, сохраняются четкие линии и углы, границы не расплываются. Для фона и растровых графических фрагментов применяются алгоритмы сжатия с потерями, напоминающие упаковку в формате JPEG. Точки фона, накрытые текстом, обрабатываются отдельно, в результате символы не расплываются и не интерферируют с окружением.

Файлы DjVu могут быть снабжены отдельным текстовым слоем, что делает возможным полнотекстовый поиск по документу, который выглядит как оцифрованная картинка. Формат оптимизирован для передачи по сети таким образом, что страницу можно просматривать ещё до завершения загрузки файла. Кроме того, DjVu-файл может содержать встроенное интерактивное оглавление и активные области — ссылки, что позволяет реализовать удобную навигацию в DjVu-книгах.

Формат предназначен для хранения электронных документов с обилием сложных графических включений, не допускающих точного распознавания: формул, схем, рисунков и рукописных символов и др. Он стал основой для нескольких библиотек научных книг. Формат довольно популярен, и в нём публикуется большое количество документов различной тематики.

Если сравнить форматы PDF и DjVu, то документы первого формата, как правило, лучше масштабируются, точнее передают полиграфические особенности оригинала, подвержены меньшим искажениям, занимают больше места. PDF более богатый формат. Тексты PDF можно переносить при помощи буфера обмена в другие редакторы и документы, часто с сохранением всех особенностей форматирования. Документы DjVu такой возможностью не обладают. Кроме того, редактирование файлов DjVu намного более сложное дело, чем файлов PDF.

Метаданные

Вся информация, хранящаяся в электронных библиотеках, может быть разделена на два класса: данные и метаданные. Данные – это знания, представленные в формализованном виде, пригодном для передачи, хранения и использования в некотором информационном процессе. Метаданные можно определить как данные о данных. В современных ЭБ метаданные используются для каталогизации и систематизации библиотечных ресурсов. В традиционных «бумажных библиотеках» эти функции, были возложены на экспертов (библиографов и каталогизаторов) и требовали большого труда и высокой квалификации.

Система метаданных является центральным логическим компонентом любой электронной библиотеки. По своему положению в составе ЭБ она напоминает традиционный библиотечный каталог – т.е. превращает множество единиц хранения в целостную систему библиотечных фондов. В наше время все основные технологические процессы обработки электронных документов в составе ЭБ основываются на использовании метаданных:

- навигация в информационном пространстве ЭБ;
- поиск единиц хранения, а также их фрагментов или совокупностей;
- обновления и актуализации репозитория ЭБ;
- управление правами доступа к документам ЭБ.

В наше время для поддержки работоспособности ЭБ различного типа применяются различные типы и виды метаданных. В нескольких публикациях, посвященных теоретическим аспектам функционирования электронных библиотек, предпринималась попытка построить полную и непротиворечивую классификацию типов метаданных [6]. Приведем несколько вариантов наиболее содержательных таксономий.

Классификация по типу определяемой информации:

- метаданные, которые описывают атрибуты электронного документа;
- метаданные о документе, которые предоставляют потенциальному потребителю информацию об основных характеристиках данного информационного ресурса;
- метаданные, предназначенные для автоматизированной компьютерной обработки.

Классификация по свойствам описываемой информации:

- описательные метаданные. Это, прежде всего, библиографическая информация об источнике;
- структурные метаданные. Сведения об используемых форматах и внутренней структуре электронного документа;
- административные метаданные. Права и условия доступа, ограничения на использование и цитирование информации, содержащейся в документе.

Классификация, в которой основным классификационным признаком является способ использования информации, кодируемых метаданными:

- метаданные, создаваемые web-службами индексирования и поиска. Это, главным образом, сведения об электронных сетевых ресурсах, которые собираются программными роботами и поисковыми системами;
- метаданные, используемые для описания информационного ресурса (например, форматы Dublin Core и IAFA/WHOIS);
- метаданные, используемые для задания месторасположения, анализа, оценки, документирования информационного ресурса.

Различие между данными и метаданными ситуативно и зависит от актуального контекста. Например, каталожное описание и реферат обычно рассматривают как метаданные, поскольку они описывают другие данные. Если эти сведения размещены в онлайн-каталоге или реферативной базе,

то они представляют собой данные. Относительный характер метаданных особенно ярко проявляется в текстовых электронных ресурсах. Например, заглавие книги, аннотация статьи, имя автора и другие сопроводительные сведения входят в метаописание большей части электронных ресурсов. В составе самого документа эти рубрики представляют собой полноценные данные. Для других типов электронных ресурсов, например изображений, подобный релятивизм не является характерным.

Состав и содержание метаданных регламентируется многочисленными стандартами. Упомянем самые важные из них.

- Теги <meta> размеченных документов в формате HTML, XML и их диалектов.
- Общие форматы метаданных. Dublin Core – набор элементов для описания различных сетевых ресурсов, доступных в Интернете и локальных сетях.
- Метаданные описания контента. Охватывают описание всех аспектов информационного объекта, как отдельной сущности:
 - Архивная информация: AIF, CWM, DDI, EAD, ISAD, OAIS, RKMS.
 - Библиографическая информация: MARC, RUSMARC.
 - Географическая информация: CIP, CSDGM, DIGEST, GDF, GISED, Global Map, ISO 8211, NGDF, NTF, OpenGIS.
 - Дистанционное обучение: ADL SCORM, IEEE LOM, IMS CP, IMS Digital Repositories, IMS LIP, IMS QTI.
 - Музейная информация: CDWA MIDAS Object ID VRA.
 - Научная информация: CDF, CERIF, CIF, GRIB, HDF, JCAMP-DX, NetCDF, SEDRIS, VICAR.
 - Новостная информация: PRISM, NewsXML, RSS.

Многие исследователи в области электронных коммуникаций и сетевых ресурсов считают [6, 9, 10], что наиболее перспективной и общеупотребительной моделью описания метаданных является система RDF (Resource Description Framework), созданная международной организацией W3C (World Wide Web Consortium). RDF дает описание ресурсов в виде, пригодном для машинной обработки. Ресурсом в RDF может быть любая сущность – как информационная, например, сайт, изображение или документ, так и неинформационная (персона, административная единица, абстрактное понятие и пр.). Декларация о ресурсе в RDF имеет стандартный вид «субъект – предикат – объект» и называется триплетом. Например, утверждение «песок желтого цвета» можно представить в виде триплета, где «песок» – субъект, «желтый цвет» – объект, а предикатом служит утверждение «имеет цвет». связывающим

Для обозначения субъектов, предикатов и объектов в RDF используются URI (от англ. Uniform Resource Identifier) унифицированный идентификатор ресурса. URI – это символьная строка, позволяющая идентифицировать электронный ресурс: документ, изображение, файл, ролик, службу, ящик электронной почты и др. В отличие от URL, URI способен идентифицировать ресурс в определенном пространстве имен, не указывая его физического расположения. Это даёт возможность описывать с помощью RDF сущности, доступ к которым не может быть предоставлен глобальной или локальной сетями, например, личность, город, страна и пр.

Список литературы

1. Акеройд Дж. Управление развитием электронных библиотек // Науч. и технич. б-ки.-2001.– №2 – с.119-127.
2. Вдовицын В. Т. Вопросы формирования и использования электронных научных информационных ресурсов / В. Т. Вдовицын, А. Д. Сорокин // Информационные ресурсы России. – 2004. – №4. – с. 7-12

3. Внешние информационные ресурсы // Справочник информационного работника / Под ред. Р. С. Гиляревского, В. А. Минкиной. – СПб.: Профессия, 2005. – с. 135-235
4. Кемпель, Г. Н. Электронная библиотека вуза в помощь самостоятельной работе студентов [Электронный ресурс] / Г. Н. Кемпель // Международная научно-практическая конференция «Библиотеки и образование-2006»: Тезисы докладов / МУБиНТ. – Вологда, 2006. – Режим доступа:
<http://elib.mubint.ru/bio/2006/materials/?id=ts> . – Загл. с экрана.
5. Лавренова О.А. На пути к электронной библиотеке // Науч. и технич. б-ки.-2001. – №2 – с.85-90.
6. Лапо П.М., Соколов А.В. Введение в электронные библиотеки [Электронный ресурс] Москва – 2005. – 92.
7. Хорошилов, А. В. Мировые информационные ресурсы / А.В. Хорошилов, С. Н. Селетков. – СПб.: Питер, 2004. – 176.
8. Шарон Т. Электронные библиотеки в Интернете. / Т.Шарон, А.Франк // Науч. и технич. б-ки. – 2001. – №2 – с.127-133.
9. Шрайберг Я.Л. Библиотеки и информационные технологии: десять лет спустя: Ежегодный доклад конференций «Крым», 2003.
10. Электронные учебники и электронные библиотеки: Тезисы докладов 3-й Всероссийской конференции. – М.: Издательство МЭСИ, 2002, 255.