Инструментальное программное обеспечение для создания полнотекстовых электронных библиотек

© Л.Г. Еремеев © А.В. Борисенко © С. В. Исмакаева
ОмГУ им.Ф.М.Достоевского
г. Омск
eremeev@omsu.ru alex@omsu.ru ismakaeva@omsu.ru

Аннотация

Этот документ описывает систему создания полнотекстовых электронных библиотек (ССПЭБ), с помощью которой можно создавать различные по масштабу и тематике электронные библиотеки. Приводится описание набора инструментальных программных средств, предназначенного для конструирования, каталогизации и поиска полнотекстовых электронных материалов конечными пользователями без привлечения IT и библиотечных специалистов. Авторы статьи не ставили перед собой цели сделать обзор и анализ существующих библиотечных систем – такой обзор тема для отдельной статьи. В данной статье описана конкретная система, с помощью которой можно создавать полнотекстовые электронные библиотеки.

Полнотекстовые электронные библиотеки являются в настоящее время необходимым элементом как учебного процесса в любых учебных учреждениях, так и научно-исследовательской и опытно-конструкторской деятельности во всех отраслях науки и техники. Не только преподаватели, научные работники и студенты, но и школьники предпочитают сначала найти необходимую для работы и учебы информацию через Интернет и только потом пойти в традиционную («бумажную») библиотеку. Количество посетителей электронных библиотек с каждым годом становиться все больше и больше по сравнению с посетителями традиционных библиотек.

Однако большинство информационных ресурсов, которые позиционируют себя как электронные библиотеки, на самом деле являются электронными коллекциями полнотекстовых версий бумажных книг, статей, учебных пособий и т. п. или просто электронными изданиями таковых, которые не имеют бумажных аналогов. По нашим данным в Рунете сейчас находится более 500 сайтов, позиционирующих себя как электронные библиотеки [8].

В чем же отличие электронной библиотеки от электронной коллекции? На наш взгляд, тем основным, что отличает электронную библиотеку от электронной коллекции, является то, что электронная библиотека, в отличие от электронной коллекции имеет каталог, аналогичный каталогу традиционной библиотеки. Примерами действительно электронных полнотекстовых библиотек (поиск осуществляется через электронные каталоги) являются: Библиотека Конгресса США, Российская Государственная Библиотека, Президентская библиотека им. Б. Ельцина, ГПНТБ. Ряд ВУЗовских библиотек России также уже начали предоставлять электронные полнотекстовые материалы.

Поиск необходимого документа (книги, статьи, учебного пособия) в электронной библиотеки происходит через электронный каталог. В электронных коллекциях нет электронных каталогов и вследствие этого, поиск в них организован другими способами: либо используется минимум метаданных для визуального поиска по произвольному списку и/или алфавиту, либо различными средствами поиска используемыми на сайтах. Если в электронной коллекции содержится небольшое количество документов – несколько сотен или тысяч, то визуальный поиск не является слишком трудоемким, но при наличии в ней достаточно большого количества документов такой способ поиска уже затруднителен. Для облегчения поиска в электронных коллекциях их создатели вводят дополнительные сервисы, например поиск по буквам алфавита – первых букв фамилий авторов. Такой способ поиска аналогичен поиску в словарях или энциклопедиях. Однако, когда на какую-то букву, например на букву «К» приходится достаточно большое количество документов, далее читателю приходиться вновь переходить к визуальному поиску в достаточно длинном списке. Ряд электронных коллекций имеют собственные поисковые механизмы, которые дают возможность читателю вводить в поисковое предписание и фамилию автора, и название документа. Однако практически все электронные коллекции имеют собственные поисковые механизмы, со своими особенностями. Нет единого стандарта. Поэтому читателю для поиска необходимой информации

приходиться каждый раз разбираться в особенностях системы поиска, реализованной в каждой и таких электронных коллекций.

Примерами таких электронных коллекций являются известные крупнейшие электронные коллекции: Максима Мошкова, Альдебаран, Либрусек.

При поиске в электронной библиотеке читатель использует для составления поискового предписания общепринятые библиотечные стандарты, которые обеспечивают нетрудоемкий способ нахождения необходимого материала через электронный каталог. Важно отметить, что эти стандарты полностью аналогичны процессу поиска через бумажные каталоги традиционных библиотек. Для создания электронных каталогов разработаны международные форматы, например, MARC.

Стоит отметить, что MARC-форматы и RUSMARC в частности полностью соответствует по содержанию библиографической записи и используемой терминологии требованиям, описанным в ГОСТ 7.1-2003, ГОСТ 7.76-96. Библиографические карточки в традиционных библиотеках создаются в соответствии с теми же стандартами.

Почему же создатели электронных коллекций не используют эти стандарты и не создают свои коллекции в виде электронных библиотек, имеющих электронные каталоги?

Основные причины:

- электронные коллекции создаются в большинстве своем энтузиастами создания электронного контента, которые не являются специалистами в библиотечном деле и не знакомы с правилами каталогизации. Поэтому они обычно используют встроенные механизмы поиска тех сайтов, на которых они размещают свои электронные коллекции;

- на пути размещения электронных полнотекстовых материалов в свободный доступ существуют административные барьеры. При этом наиболее часто можно услышать ссылки на законодательство об авторском праве.

На наш взгляд эти ссылки не являются корректными. Вот полная цитата из этого закона: «В соответствии с пунктом 1 статьи 1229 Гражданского кодекса Российской Федерации правообладатель (гражданин или юридическое лицо, обладающее исключительным правом на результат интеллектуальной деятельности) вправе использовать результат деятельности по своему усмотрению, а также распоряжаться принадлежащим ему исключительным правом, в том числе разрешать или запрещать другим лицам использование результата интеллектуальной деятельности».

Из этой статьи Гражданского кодекса следует, что авторы сами принимают решение о свободном доступе к своим работам при размещении их в электронной библиотеке. Следовательно, автор, как правообладатель (если это его право не ограничено какими-либо соглашениями) интеллектуальной собственности, самостоятельно решает вопрос о предоставлении свободного доступа к материалам.

Библиотеки изначально создавались государством как информационные центры, которые дают бесплатный доступ к информации. Для образования и научной работы такой доступ является необходимым условием.

Ведущие библиотеки страны не в состоянии в обозримом будущем (а может быть и вообще никогда) обеспечить представление в цифровом формате полнотекстовых материалов, которые создаются в регионах и в конкретных предприятиях и организациях.

Следовательно, вопросы создания инструментальных средств, которые можно использовать для создания региональных электронных библиотек и таких же библиотек организаций являются актуальными.

На наш взгляд таковые инструментальные средства должны отвечать следующим требованиям:

- возможность построения собственных автоматизированных библиографических баз данных в рамках отдельного учреждения или ведомства;

- возможность создания собственных полнотекстовых баз данных, в том числе и территориально-распределенных;

- возможность пополнения библиографических и полнотекстовых баз данных самими авторами, без привлечения других специалистов (программистов и библиографов);

- возможность оперативного введения новых рубрик, ключевых поисковых слов и терминов;

- возможность получения доступа каждого учащегося или исследователя не только к конкретным библиографическим данным, а и к полнотекстовой информации, в том числе и по ключевым словам;

- возможность автоматизированного составления библиографических списков библиотечных материалов, относящихся к интересующей тематике;

- интуитивно понятный интерфейс пользователя, являющийся универсальным для поиска в достаточно большом количестве электронных библиотек;

- использование международно-признанных технологий;

- технологическая совместимость с базами данных зарубежных электронных библиотек, на основе информационных международно-признанных технологий и обеспечения доступа к библиографической и полнотекстовой информации через Интернет;

- универсальность применения используемых технологий к большинству отраслей науки и техники;

- легкость в освоении – к пользователям системы должны предъявляться минимальные требования – достаточно владеть основами компьютерной грамотности;

- масштабируемость – возможность применения как для отдельной организации, так и для куста организаций и региональных структур;

- приемлемая стоимость: стоимость компьютеров (в зависимости от масштаба применения – мощность серверов), стоимость программного обеспечения (максимальное использование свободно-распространяемого программного обеспечения).

В институте математики и информационных технологий Омского государственного университета разработана система создания полнотекстовых электронных библиотек (ССПЭБ) [5], которая соответствует большинству перечисленных выше требований.

Целью разработки собственной системы было создание недорогой легко осваиваемой инструментальной системы, не требующей привлечения IT-специалистов и каталогизаторов для эксплуатации, с помощью которой можно без больших затрат создавать различные по масштабу электронные библиотеки.

Полнотекстовая электронная библиотека, созданная с помощью ССПЭБ представляет собой базу данных, в которой для индексации используются метаданные, соответствующие набору поисковых атрибутов BIB-1 протокола Z39.50, разработанному в Библиотеке Конгресса США и являющимся общепризнанным международным стандартом.

В базу данных включено два множества D и M. Множество D содержит собственно полнотекстовые материалы, а M – метаданные этих материалов (библиографические записи), по которым осуществляется поиск.

Множество D разделено на два подмножества D1 и D2. Множество D2 можно рассматривать как «мастерскую» авторов, в которой каждый автор имеет личный электронный кабинет. Документ, над которым работает автор, может находиться в D2 неопределенное время (пока автор занимается его созданием - конструированием электронной версии документа).

После того как автор закончил работу над своим произведением, документ из множества D2 переводиться во множество D1, а во множестве M создается элемент, содержащий его библиографическую запись (карточка электронного каталога) и устанавливается взаимно-однозначное соответствие между этой карточкой и собственно документом. С этого момента документ становиться доступным читателю.

Именно наличие взаимно-однозначного соответствия между элементами множеств M и D1 обеспечивает избавление читателя при поиске необходимых материалов от поискового мусора, присущего процедуре поиска на множестве сайтов с применением поисковых систем, таких как Яндекс, Google и обеспечивает быстрое и гарантированное нахождение нужных материалов.

Программная инструментальная среда ССПЭБ предназначена для обслуживания трех категорий пользователей: авторов, администраторов и читателей.

ССПЭБ базируется на специально разработанной системе управления контентом (CMS), работающей в среде PHP и веб-сервера Apache. Кроме того в ней используются специальные средства для работы с протоколом Z39.50.

В качестве исполняющей среды, так и СУБД насколько возможно использовались свободно распространяемые, но достаточно признанные и эффективные продукты: PHP, Apache, PostgreSQL, Zebra (от IndexData). Это позволило существенно удешевить себестоимость создания системы [1,7].

Собственно программное обеспечение электронной библиотеки использует все эти компоненты для реализации функционала АРМ «Автор» и «Администратор» [6] и поискового сервиса.

Для авторов создан АРМ «Автор», с помощью которого авторы могут создавать сетевые электронные версии своих произведений (либо электронные версии, не имеющие бумажных аналогов), не привлекая для этого профессиональных программистов. Автор может использовать различные текстовые редакторы (например, Word, Adobe Acrobat и т.п.), может вставлять в текст изображения, формулы, гиперссылки и т. п., а также аудио- и видеоматериалы. При создании документа автор одновременно вводит метаданные: название работы, ФИО автора (авторов), аннотацию, ключевые слова и год издания. Эти метаданные в дальнейшем используются для создания элемента электронного каталога и автоматической каталогизации.

Этот набор метаданных соответствует перечню обязательных полей формата RUSMARC:

- поле идентификации записи;

- поле кодирования издания (год издания, дата ввода в базу, тип документа, и т.д.);

- язык документа;

- заглавие;

- источник документа (кодовое название организации, где находится документ).

Вместе с автоматически создаваемыми дополнительными полями выбранный нами набор метаданных полностью покрывает этот обязательный набор полей и предоставляет возможность использования других полей - наиболее часто используемых при поиске, таких как

«автор», «год издания», «ключевые слова», «аннотация».

Создаваемый автором документ представляет из себя граф типа дерева. Корневая вершина графа является обязательной и содержит в себе метаданные и часть текста документа (в вырожденном случае – весь текст документа). Функциональные возможности АРМ «Автор» позволяют автору конструировать документ как граф, присоединяя к корневой вершине другие вершины. Ребра графа реализуются посредством гиперссылок. Присоединяемые вершины являются файлами, которые могут быть созданы с помощью широкого спектра программных средств. Например, это могут быть: текстовые файлы форматов doc, pdf, djvu и др., видео-файлы форматов avi, flash, mpg и др., аудиофайлы, графические и т.д. Конечно, автор должен понимать, что читатель только тогда сможет прочесть (просмотреть и прослушать) его произведение, когда на его компьютере будет установлено соответствующее программное обеспечение. Соответственно, более правильно будет выбирать для включения в документ файлы наиболее распространенных форматов.

После того как автор закончил создавать электронный документ, он отправляет в администрацию электронной библиотеки запрос на публикацию. Для этого в АРМ «Автор» предусмотрена специальная функция.

Для администраторов электронной библиотеки создан АРМ «Администратор». С помощью этого АРМ администратор осуществляет модерацию представляемых авторами документов и их автоматическую каталогизацию (без привлечения для этого профессиональных каталогизаторов). Кроме этого он может управлять учетными записями авторов и электронными каталогами.

Администраторы, кроме управленческих, фактически выполняют и функции, аналогичные функциям редакционной коллегии электронного журнала. Инструментальные средства АРМ «Администратор» позволяют им просматривать содержимое электронных материалов, созданных авторами и либо разрешать их публикацию, либо отклонять публикацию, с указанием причин отклонения, либо удалять.

Если принято решение опубликовать заявленный автором материал, администратор осуществляет это решение буквально нажатием одной кнопки «Опубликовать». В этот момент на основе метаданных, которые автор ввел при конструировании документа, автоматически создается электронная карточка каталога (формат RUSMARC). В 856 поле карточки содержится URL, по которому может быть осуществлен доступ к документу по протоколу HTTP. Эта технология позволяет существенно сократить организационные и временные затраты, поскольку можно обходиться без привлечения библиотечных специалистов – каталогизаторов.

Созданная карточка помещается в один из электронных каталогов, обычно в тот, который выбрал автор при конструировании документа. Электронные каталоги размещаются на Z39.50 сервере. Размещение карточки в электронном каталоге является реализацией переноса документа из множества D2 во множество D1. И документ становится доступным для поиска и чтения.

Для того чтобы автор мог иметь личный кабинет, в котором он может конструировать свои электронные материалы, он должен обратиться к администратору библиотеки. Администратор может создать учетную запись для автора и выдать ему логин и пароль для входа в личный кабинет.

Управляя электронными каталогами, администратор может создавать их, редактировать и удалять. Сколько и каких электронных каталогов содержится в библиотеке, решает администрация.

Поисковая система библиотеки создана на основе использования шлюза HTTP – Z39.50 [2]. Для того чтобы использовать поисковую систему достаточно иметь на компьютере (или на планшете) читателя один из распространенных браузеров, например: Mozilla Firefox, Internet Explorer, Chrome.

Читатель может создавать поисковое предписание, содержащее до трех термов, соединенных операторами «И», «ИЛИ», «НЕ». Каждый из термов может быть выбран из списка: «по всем полям», «автор», «заглавие», «ключевое слово», «дата издания», «дата каталогизации».

После того как читатель создал поисковое предписание и выбрал каталог (или каталоги) в которых он решил произвести поиск, он нажимает кнопку «начать поиск». Запрос передается на сервер библиотеки по протоколу HTTP и на сервере шлюз HTTP – Z39.50 преобразует запрос в протокол Z39.50. Поиск осуществляется на Z-сервере в выбранных читателем каталогах. Содержимое найденных записей электронного каталога, которые соответствуют заданному поисковому предписанию, передается на компьютер читателя в виде списка.

Читатель может просмотреть содержимое карточек каталога в четырех вариантах:

- минимальная форма представления;
- библиографическая карточка;
- RUSMARC;
- текстовая форма.

Содержимое электронной карточки, кроме метаданных включает в себе URL документа. Для перехода на просмотр полного текста читатель должен нажать кнопку «переход на полнотекстовый документ».

Так как в данной системе работа с электронными каталогами реализована на основе использования протокола Z39.50, поиск в каталогах можно производить не только с помощью собственно поисковой системы самой библиотеки, но с помощью специального программного обеспечения, установленного на компьютере читателя. Это

программное обеспечение представляет собой Z-клиент, который может связываться с Z-сервером непосредственно по протоколу Z39.50 без преобразования на шлюзе HTTP – Z39.50.

Такое программное обеспечение было разработано в Институте математики и информационных технологий ОмГУ несколько лет назад и получило название «библиотечный браузер LibNavigator» [3,4].

Использование для поиска Z-клиента дает следующие преимущества – можно производить поиск во всех библиотеках, которые содержат свои электронные каталоги на Z-серверах. Для этого необходимо знать следующие параметры соответствующего Z-сервера: URL (или IP-адрес), номер порта, который используется для протокола Z39.50 и имя базы данных. Эти параметры фактически дают адрес Z-сервера библиотеки, на котором в ней размещены электронные каталоги. В настоящее время в мире несколько тысяч библиотек размещают свои электронные каталоги на Z-серверах.

Для нашего Z-клиента LibNavigator был создан список таких Z-серверов, который включен в состав дистрибутивной версии и, соответственно, формируется при установке LibNavigator'а на компьютере пользователя в виде дерева ресурсов. Далее пользователь по своему усмотрению может корректировать это дерево.

LibNavigator можно использовать и для другой интересной возможности. А именно, для организации электронной библиотеки на базе существующей электронной коллекции. Для этого надо разработать утилиту (или набор утилит), которая обработает метаданные полных текстов электронной коллекции, на их основе создаст карточки электронного каталога и разместит их на Z-сервере. Таким образом, для этой электронной коллекции будет создан электронный каталог, и поиск в ней можно будет производить по библиотечным правилам, т.е. с использованием каталога.

Мы разработали такие утилиты для известной электронной коллекции «Техническая библиотека» (http://techlibrary.ru). Теперь на нашем Z-сервере размещен электронный каталог, через который можно производить поиск полных текстов в этой электронной коллекции.

Заметим, что этот пример иллюстрирует тот факт, что использование данной технологии позволяет создавать распределенные полнотекстовые библиотеки. Для этого необходимо только иметь доступ к Z-серверам, на которых размещены электронные каталоги, содержащие к своих карточках URL полных текстов (поле 856 в формате RUSMARC) и совсем необязательно знать, где реально размещается сам полный текст.

ССПЭБ была практически применена при создании полнотекстовой электронной библиотеки ОмГУ им. Ф.М. Достоевского (http://elib.omsu.ru). Сейчас в ней более 2500 полнотекстовых материалов созданных в ОмГУ, большая часть из которых является электронными версиями статей из сборника «Вестник Омского университета». Для читателя доступен поиск по нескольким каталогам, например: «Вестник ОмГУ», «Полнотекстовые материалы по этнографии», «Профилактика экстремизма».

Функциональные возможности АРМ «Автор» достаточно универсальны и это дало возможность авторам создавать разнообразные и по содержанию, и по структуре полнотекстовые материалы: статьи, монографии, учебные пособия, мультимедиа материалы.

Примером использования данной технологии для развития региональных информационных ресурсов является создание с применением ССПЭБ полнотекстовой электронной библиотеки в Северо-восточном Федеральном университете (Республика Саха, Якутия).

Полнотекстовые материалы, представленные в данной электронной библиотеке, содержат информацию о малочисленных народах Севера - долганах, коряках, чукчах, эвенках, эскимосах, юкагирах, в том числе, и на языках этих народов. Создание такой электронной библиотеки помогает решать важные региональные вопросы использования современных информационных и образовательных технологий для сохранения и развития языков, культуры и духовности народов Севера.

Особенностью данной библиотеки, в частности, является то, что языки малочисленных народов Севера используются не только для содержащихся в ней полнотекстовых материалов (а также мультимедиа материалов). Эти языки, наряду с русским и английским также могут быть использованы и для метаданных, которые содержатся в карточках электронного каталога и, соответственно, используются при поиске.

Доступ читателей к материалам библиотеки организован через «Арктический многоязычный портал» (http://arctic-megapedia.ru), посредством «Поисковой системы языкового и культурного наследия коренных малочисленных народов Севера (КМНС)», созданной на базе шлюза HTTP- Z39.50 ССПЭБ.

Еще одним примером применения данной технологии является создание учебного пособия в рамках работ по созданию учебного контента для подготовки космонавтов к выполнению длительных космических полетов. Данная работа выполнялась в сотрудничестве с российским Центром подготовки космонавтов [9].

Коллектив разработчиков продолжает разработку ССПЭБ. Сейчас создается новая версия ССПЭБ, в которой будет расширен список метаданных и введен ряд дополнительных сервисов.

Литература

[1] А.В. Борисенко, Л.Г. Еремеев, А.В. Кузнецов. Создание полнотекстовых электронных библиотек. Информационные технологии в обеспечении нового качества высшего образования. Всероссийская научно-практическая конференция с международным участием. М. 2010 г.

[2] Л.Г. Еремеев, А.В. Борисенко. Полнотекстовые ресурсы омского корпоративного библиотечного консорциума. Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества. Библиотеки и доступность информации в современном мире: электронные ресурсы науке, культуре и образованию, 2003 г., 10 юбилейная Международная конференция Крым 2003.

[3] Л.Г. Еремеев, Д.С. Пашкевич. Библиографический браузер LibNavigator, версия 2, редакция Читатель. Научные и технические библиотеки, 2006 г., № 2.

[4] Л.Г. Еремеев [и др.]. Развитие средств доступа к образовательным библиотечным ресурсам. Одиннадцатая Международная конференция и выставка. LIBCOM 2007.

[5] Л.Г. Еремеев, А.В. Кузнецов, И.П. Стрельчук, Д.С. Пашкевич. Технологии создания учебно-методических материалов с использованием систем управления контентом и предоставления доступа к ним по протоколу Z39. 50. Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса, 2008 г.

[6] Л.Г. Еремеев, А.В. Кузнецов. Программное обеспечение, позволяющее авторам создавать полнотекстовые электронные документы и производить их автоматическую каталогизацию. Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса, 2009 г.

[7] Л.Г. Еремеев, А.В. Кузнецов. Система создания полнотекстовых электронных библиотек. Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: материалы Семнадцатой Международной конференции Крым 2010.

[8] Л.Г. Еремеев, А.В. Кузнецов. Обзор электронных библиотек России. Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: материалы Семнадцатой Международной конференции Крым 2010.

[9] Б.И. Крючков, А.С. Ренжин, В.М. Усов, Л.Г. Еремеев. Особенности формирования учебного контента для подготовки космонавтов к выполнению длительных космических полетов. Человек Земля Космос. Международная конференция, Калуга. 2011 г.

Software to create a full-text digital libraries

L. G. Eremeev, A. V. Borisenko, S. V. Ismakaeva

This document describes the system for creation of full text electronic libraries (SCFEL) which can be used to create electronic collections different in scope and subject. This document describes the software for creating, cataloging and retrieval of full-text electronic materials by users without involving IT and library professionals. The authors did not set a goal to make the review and analysis of existing library systems - a review of a topic for another article. This article describes a particular system, with which you can create a full-text electronic library.