Предлагаем вниманию читателей новую рубрику под эгидой Российской ассоциации электронных библиотек, материалы которой будут посвящены созданию и функционированию электронных библиотек, оригинальным проектам, ресурсам и методам работы в этой сфере. Открывает рубрику статья В.А. Осиповой о новом проекте межмузейной распределённой библиотеки.

Полнотекстовый поиск в новой распределённой межмузейной электронной библиотеке

При поддержке Российской ассоциации электронных библиотек ряд организаций: ООО «Константа», Архангельский областной центр повышения квалификации специалистов культуры (АОЦПК), Музеи Московского Кремля, Государственный исторический музей и Библиотека истории русской философии и культуры «Дом А.Ф. Лосева», начинают новый уникальный для России проект – создание распределённой межмузейной электронной библиотеки.

Автор

Варвара Андреевна ОСИПОВА
редактор сайта Российской ассоциации электронных библиотек

Идея организации распределённого доступа к ресурсам различных фондов в удалённом режиме не нова, однако воплощение, реализованное в данном случае, практически не имеет аналогов (подобная система разработана только в университете Святого Лаврентия в Швеции). Особенность его заключается в том, что система осуществляет полнотекстовый поиск не по каталогу (заглавие/автор), а по полному содержанию работы, выдавая по запросу пользователя не гиперссылку и не полный текст документа с выделением искомых терминов, а абзацы и контекст к ним.

Презентация проекта была проведена в Музеях Московского Кремля на конференции «Музейные библиотеки в современном обществе: некнижные материалы в фондах музейных библиотек» 5 апреля 2011 г. Для более широкой аудитории аналогичное представление созданного ресурса прошло в формате интернет-трансляции на сайте Российской ассоциации электронных библиотек 8 апреля 2011 г.

У инициаторов проекта, заместителя директора по информационным и компьютерным технологиям Государственного исторического музея А.В. Дремайлова, директора АОЦПК С.Х. Ляпина и заместителя генерального директора ООО «Константа» А.В. Куковякина мысли о создании ресурса возникли в условиях того, что формирующиеся электронные библиотеки музеев, содержащие коллекции оцифрованных книг, иллюстраций и других предметов искусства, не имеют связи между собой. В то же время традиционные библиотеки создают объединения, позволяющие, во-первых, координировать процесс сканирования (исключая возможность перевода в электронный вид одной и той же книги разными организациями), а, во-вторых, создавать для пользователя удобную площадку с распределённым поиском по всем имеющимся ресурсам. Инициаторы проекта задавались вопросом: «Возможна ли Российская национальная музейная электронная библиотека?» И постепенно приходили к выводу о том, что да.

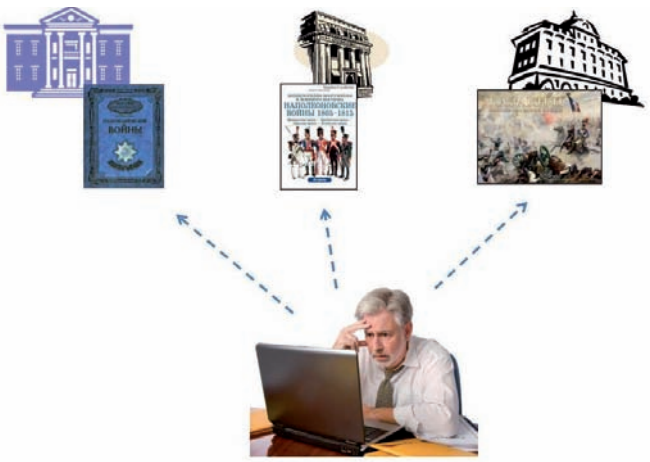

Рис. 1. Представление идеи распределённой электронной музейной библиотеки

Распределённая межмузейная электронная библиотека являет собой связь — «пользователь — поисковый запрос — различные удалённые коллекции ресурсов — ответ на поисковый запрос» (рис. 1). Электронные коллекции в музеях существуют независимо и отдельно, каждая на своём сервере. Когда пользователь начинает работу, формируется третья, временная система, которая генерирует ответы из различных источников, независимо от их «физического» местонахождения. Без пользователя не существует объединённой системы. Преимущества этого в том, что нет единоличного правообладателя ресурсов, нет сложной системы передачи файлов в центральную базу данных и каталог. Для осуществления распределённого поиска по библиотеке даже не обязательно создавать единую точку входа. Можно с сайтов электронных библиотек музеев или других организаций — участников проекта выбрать не локальный, а распределённый поиск. Однако планируется, что в разделе «Проекты» сайта Российской ассоциации электронных библиотек www.aselibrary.ru будет опубликована информация о межмузейной библиотеке, содержащая в том числе и альтернативную ссылку для входа в поисковую систему.

Возможности системы T-Libra

За системную основу было принято программное обеспечение T-Libra, разработанное архангельским ООО «Константа»; в настоящий момент используется версия 6. Информационная система предназначена для создания многомодальных и многофункциональных электронных полнотекстовых библиотек. Она имеет следующие возможности.

1. Импорт / экспорт метаданных (библиографической информации) в форматах ISO 2709 UniMarc/RusMarc из существующих АБИС («электронных каталогов») в систему и обратно, в том числе вместе с присоединёнными файлами ресурсов.

2. Импорт данных:

- импорт и индексация электронных полнотекстовых ресурсов в формате rtf или pdf (символьный) для полнотекстового представления ресурса, сохраняющего базовое форматирование текста (разбивка на абзацы, сноски, вложенные изображения и таблицы, жирность, курсив, подчёркивание, разрядка);
- импорт постраничного графического представления ресурса (в виде совокупности jpg-файлов);
- импорт файлов других форматов для депозитарных представлений ресурса.

При импорте происходит автоматическое пополнение многоязычного электронного словаря словоформ; возможно также и «ручное» редактирование словаря.

3. Ведение собственного каталога, создание и корректировка библиографических описаний (библиографических записей) в диалоговом режиме.

4. Поиск по каталогу с настраиваемыми полями и встроенной в них булевой алгеброй (поддерживаются три формы поиска: полная, краткая, однострочная), а также встроенным в функционал одного из полей предметным мультирубрикатором, содержащим библиотечные классификаторы и рубрикаторы (УДК, ББК, ВАК, ГРНТИ и т.д.) с интерфейсами для их редактирования.

5. Использование для целей хранения и презентации пополняемого файлового хранилища (депозитария) с файлами произвольного вида и формата и собственным настраиваемым рубрикатором ресурсов.

6. Гибкий тематизируемый многоязычный полнотекстовый поиск двух типов, восьми видов с сортировкой, группировкой и различными формами презентации его результатов.

7. Мультимодальное расширение (графика, аудио, видео), используемое как для расширения функциональных возможностей программы, так и для взаимодействия и интеграции с другими информационными системами, модулями и оболочками (электронными коллекциями, электронными экспозициями, мультимедиа-энциклопедиями и т.д.).

Возможности поиска

Распределённый полнотекстовый поиск осуществляется не по каталогу, в отличие от подавляющего числа библиотечных систем, поэтому при необходимости для сужения зоны поиска вводится такое понятие, как «корзина ресурсов». Перед тем, как делать поисковые запросы, нужно отобрать ресурсы путём ввода ключевого слова, «слова-ограничителя» (например, это может быть «региональная культура», «библиотечная деятельность», «живопись», «Ломоносов», и т.д.). Полнотекстовая система анализирует всю внутреннюю семантику текстов и за секунду вычисляет именно те документы, в которых «слова-ограничители» часто встречаются, и будет работать, таким образом, только по ним. Можно искать и по всем имеющимся в базе ресурсам, не выделяя предметную область.

Распределённый поиск в системе управляется пользователем. На рис. 2 представлены поля вкладки «Параметры и результаты», которые иллюстрируют, что пользователь может: ▶

1) применить многоярусный поиск, вводя разные искомые термины в «слои». Минимально необходимое количество слоёв — 2, максимально возможное в тестовой версии — 8;

2) самостоятельно задать максимальное расстояние между искомыми терминами, которое измеряется количеством слов;

3) устанавливать тайм-аут запроса, в секундах, на время ожидания ответа серверов;

4) выбрать отображение результатов поиска «по документам» или «по абзацам».

При выдаче результатов система указывает количество найденных книг и абзацев.

При просмотре результатов «по документам» данные выводятся в 4 столбца: 1-й — порядковый номер, 2-й — название организации, которой принадлежит найденный ресурс, 3-й — данные о документе (заглавие, автор, издательство, дата издания и т.д.), 4-й — количество абзацев книги, в которых встречается искомое слово. На рис. 3 представлено отображение итога поиска с группировкой «по документам».

Внутри описания документа по мере доступности открыты следующие вкладки:

- ресурсы из депозитария — это доступные для скачивания полные тексты книг или статей в форматах MS Word, pdf. Если есть, то и графические файлы в формате jpg;
- термины индексирования — ключевые слова к тексту;
- аннотация на книгу;
- содержание найденной работы в соответствии с главами оригинала.

Заглавие каждого выданного ресурса активно, по нему осуществляется переход к списку всех абзацев с искомым термином, найденных в данной книге.

При просмотре результатов «по абзацам» найденная информация представляется в виде таблицы, включающей следующие столбцы: 1-й — порядковый номер; 2-й — название библиотеки, в которой физически находится документ; 3-й — первые строки найденного абзаца и извлечения из него со встречающимся искомым термином; 4-й — сколько слов из заданных для поиска встречено в данном абзаце; 5-й — сколько раз в абзаце употреблены исковые термины и слова. На рис. 4 представлен результат запроса по трём словам «Москва», «Кремль», «книга».

В данной схеме активен 3-й столбец, по нему можно перейти к полному тексту абзаца, который можно по встроенной шкале оценить от одного до пяти баллов, прокомментировать.

Распределённая библиотека, работающая с абзацами, позволяет положительно оценённые пользователем абзацы собрать в один файл. Эта функция в интерфейсе называется «Создать тему». Скомплектованный из абзацев текст с

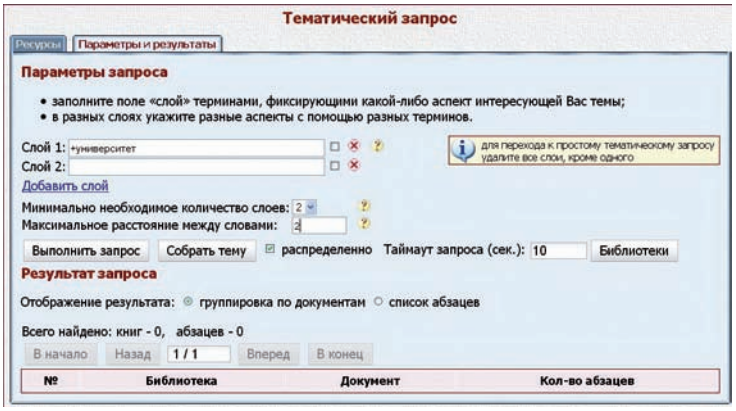

Рис. 2. Вкладка «Параметры и результаты»

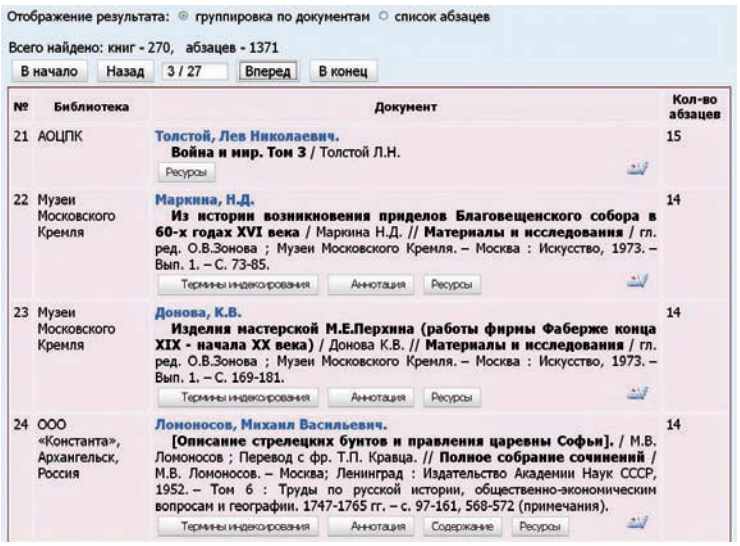

Рис. 3. Структурирование результатов поиска «по документам»

№	Библиотека	Абзац	Кол-во искомых слов различных	Кол-во искомых слов всего
1	ООО «Константа», Архангельск, Россия	Баженов (Василий ... и в Москве, в Кремле, здание ...	3	6
2	АОЦПК	Баженов (Василий ... и в Москве, в Кремле, здание ...	3	6
3	ООО «Константа», Архангельск, Россия	Филарет - патриарх ... в Москве для перевода книг и для ...	2	22
4	АОЦПК	Филарет - патриарх ... в Москве для перевода книг и для ...	2	22
5	ООО «Константа», Архангельск, Россия	Образцом учености московских ... привез в Москву книгу свою Оглашение ...	2	18
6	АОЦПК	Образцом учености московских ... привез в Москву книгу свою Оглашение ...	2	18
7	ООО «Константа», Архангельск, Россия	Мужественно вынося испытание, ... чтоб исправлять книгу Потребник на Москве в печатное ... а привезешь книгу исчерня в Москву, то ...	2	15
8	АОЦПК	Мужественно вынося испытание, ... чтоб исправлять книгу Потребник на Москве в печатное ... а привезешь книгу исчерня в Москву, то ...	2	15

Рис. 4. Структурирование результатов поиска «по абзацам»

указанием всех библиографических данных источников представляется в формате HTML, его можно распечатать, отправить по электронной почте, скопировать на флешку или диск. Примечательно, что работа с абзацами, фрагментами текста, в отличие от работы с полной книгой, освобождена от вопросов авторского права.

Пополнение распределённой библиотеки

В настоящий момент в состав распределённой межмузейной электронной библиотеки входят ресурсы Государственного исторического музея, Музеев Московского Кремля, Библиотеки истории русской философии и культуры «Дом А.Ф. Лосева», ООО «Константа» и Архангельского областного центра повышения квалификации специалистов культуры. Описаний ресурсов представлено 1331; всего ресурсов — 1297; ресурсов, имеющих полнотекстовое представление — 1267; ресурсов, имеющих депозитарное представление — 1295.

Для эффективного пополнения коллекций предполагается создание консорциума музейных библиотек — некоммерческой организации, координирующей вопросы оцифровки фондов музейных библиотек и предоставления их пользователям.

Конечно, консорциум не предполагает выполнения работ по оцифровке и их финансирования. Каждый из участников комплектует свой электронный фонд самостоятельно. Взаимодействие отдельной библиотеки с консорциумом осуществляется в большей степени на подготовительной ступени, когда происходит согласование описи документов, предполагаемых в первую очередь для оцифровки. Далее участники на своём оборудовании и с помощью своих программ производят сканирование, обработку, распознавание, вычитку, разметку, сборку текстов и публикацию их через библиотечную систему с выводом или запретом на вывод в сеть Интернет.

Администратор при загрузке ресурсов вправе самостоятельно определять перечень возможных действий с документами (как по группе ресурсов, так и по каждой отдельной книге) для различных типов пользователей (администраторы; самостоятельно зарегистрировавшиеся пользователи, работающие по сети Интернет; самостоятельно зарегистрировавшиеся пользователи, работающие в локальной среде; анонимные пользователи).

Администратор системы может предоставлять возможность следующих основных действий с ресурсом:

- видеть ресурс / запретить видеть ресурс. Запрет часто ставится на внутреннюю отчётную документацию, которая используется в служебных целях коллектива библиотеки;
- искать полнотекстовое предоставление ресурса, то есть проводить полнотекстовый поиск по документу;
- получить абзацы. Можно ставить полнотекстовый поиск с указанием количества абзацев с искомым термином, но без права просмотра найденных абзацев. Интерес данной опции в том, чтобы читатели для работы с ресурсом лично приходили в помещение библиотеки, поддерживая статистику посещений;
- получить мини-контекст. Показываются три абзаца до и три после найденного. При данном действии пользователь увидит тексты семи абзацев, где абзац с искомыми терминами будет посередине;
- получить макси-контекст. Это максимально возможное, в соответствии с российским законодательством об авторском праве, количество текста, окружающее найденный абзац. В настоящее время чётко определённого показателя в процентах от объёма работы нет, каждая библиотека определяет его по-своему, например, в Российской государственной библиотеке разрешено использовать не более 15% от общего содержания книги или статьи. Данные показатели по всем библиотекам всё же сводятся к тому, что фрагмент, разрешённый для копирования, не должен быть более 30% работы. В системе управления распределённым поиском каждый из участников проекта также имеет возможность «заложить» свой процент макси-контекста;
- давать право на депозитарное предоставление. Можно предоставлять для хранения полные тексты, а можно только метаданные. Можно положить символьный pdf, а можно графический;
- представить изображения с большим разрешением. Чем больше объём графического файла, тем лучше его качество;
- предоставить изображения с маленьким разрешением. Право могут использовать те библиотеки, которые либо экономят объём памяти для хранения документов, либо опасаются копирования изображений третьими лицами для несанкционированного использования в коммерческих целях.

В настоящее время распределённая межмузейная электронная библиотека работает по адресам: **http://demo.tlibra.ru**, предоставлен ООО «Константа», и **http://tlibra.kreml.ru:8383/bin/tauc.exe?DSN=tlibra&ObjectId=1002&MethodId=98** — страница электронных ресурсов библиотеки Музеев Московского Кремля. Разработчики проекта приветствуют любые пожелания, предложения и замечания по работе с системой. ■