2002 | Том 5 | Выпуск 3

Проект создания электронной библиотеки диссертаций в РГБ

Лаврёнова О.А.

Российская государственная библиотека

Описывается проект электронного депозитария диссертаций, получивший поддержку РФФИ. Предполагается, что электронная библиотека диссертаций будет включать: действующий в РГБ электронный каталог авторефератов и защищенных в стране диссертаций, расширяемый путем ретроспективной конверсии карточных каталогов и позволяющий вести поиск по любым элементам библиографических записей и их сочетаниям; полнотекстовую базу данных авторефератов, свободно предоставляемую в теледоступе на Web-сервере РГБ; полнотекстовую базу данных текстов содержаний (оглавлений) из диссертаций; перечни приведенной в диссертациях литературы (некий указатель цитируемой литературы по темам или отраслям знаний); полнотекстовую базу данных самих диссертаций с удаленным доступом или локальным доступом для читателей в РГБ в соответствии с договорами с авторами диссертации; данные об авторах.В электронной библиотеке программными средствами обеспечивается единый интерфейс доступа из одной точки ко всем включаемым в систему информационным ресурсам. Более того, данная электронная библиотека станет одной из электронных коллекций электронной библиотеки РГБ.

Российская государственная библиотека, являющаяся единственным хранилищем подлинников диссертаций, защищенных в стране с 1944 года по всем специальностям, кроме медицины и фармации, имеет фонд свыше 800 тыс. томов таких документов. Необходимость обеспечения широкой доступности и сохранности этого фонда на основе современных информационных технологий и средств передачи данных привели РГБ к выводу о целесообразности создания федеральной электронной (цифровой) библиотеки диссертаций (ЭБД).

Под "электронной библиотекой" (ЭБ), или "цифровой библиотекой", мы понимаем разновидность автоматизированных информационных систем (АИС), в которых документы хранятся и могут использоваться в машиночитаемой ("электронной") форме, причем программными средствами обеспечивается единый интерфейс доступа из одной точки к электронным документам, содержащим тексты и изображения /1, 2/.

База данных ЭБ может состоять из различного вида электронных коллекций документов, но при этом должно соблюдаться приведенное требование

относительно единого интерфейса доступа для пользователя. Таким образом, ЭБ представляет собой, с одной стороны, АИС, а с другой стороны - библиотеку. Как и для традиционной библиотеки, для ЭБ важны три основные группы проблем: формирование фондов (их состав и источники комплектования), обеспечение сохранности и доступности фондов (в том числе - информации о них, т.е. справочно-поискового аппарата).

Проект федеральной электронной библиотеки диссертаций, которая при включении ее в электронную библиотеку РГБ будет рассматриваться как ее часть, т.е. электронная коллекция диссертаций, разрабатывается с 2001 г. Проект поддерживается Российским фондом фундаментальных исследований, а также финансируется из бюджета Российской государственной библиотеки.

В зарубежной практике все шире используются словосочетания "electronic theses and dissertations" (ETDs - электронные авторефераты и диссертации) и "electronic theses and dissertations digital libraries" (электронные библиотеки авторефератов и диссертаций). Надо сказать, что сейчас это один из наиболее "модных" типов проектов в информационно-библиотечной деятельности. Назовем наиболее известные проекты: коммерческий проект предоставления в доступ электронных диссертаций компании UMI в Bell and Howell Information and Learning (http://www.umi.com/hp/Support/Dservices); некоммерческий международный проект Networked Digital Library of Theses and Dissertations (NDLTD), начинавшийся как корпоративный проект
американских университетов (http://www.ndltd.org, http://www.theses.org), и проект формирования связанных нормативных записей для авторов диссертаций для данного проекта (http://alcme.oclc.org:4342/ndltd/AuthLink.html); два университских проекта во Франции, один из которых становится международным (http://www.sudoc.abes.fr, http://www.univ-mlv.fr); проект в Австралии (http://www.library.unsw.edu.au/thesis/thesis.html); проект "Dissertationen Online" (DissOnline) в Германии, который координирует Die Deutshe Bibliothek (http://www.dissonline.de); проект представления диссертаций в XML и стандарта DTD в Берлинском университете им. Гумбольдта. Деятельность по созданию ЭБ диссертаций поддерживается ЮНЕСКО (http://www.eduserver.de/unesco).

Что касается нашего проекта, то на момент подготовки доклада разработаны следующие основные проектные решения:
1. Структура электронной библиотеки диссертаций:
в рамках ЭБД выделяются:

- электронный депозитарий диссертаций и авторефератов, не предоставляемый свободно в сетевом доступе;
- ЭБ диссертаций и авторефератов в интранет, предоставляемая читателям и сотрудникам РГБ в стенах Библиотеки;
- ЭБ диссертаций и авторефератов в Интернет, где для документов определяются различные уровни доступа.

Электронная библиотека диссертаций будет включать:

- электронный каталог (ЭК) защищенных в стране диссертаций и их

авторефератов /3/;
- электронные копии авторефератов;
- тексты содержаний (оглавлений) из диссертаций;
- перечни приведенной в диссертациях литературы;
- электронные копии диссертаций в различных форматах;
- данные об авторах (файлы нормативных записей для имен ученых в ЭК).

Особое внимание будет уделено развитию структуры метаданных, обеспечивающих ориентацию в информационных массивах, в частности, средств тематического поиска в полнотекстовой базе данных диссертаций на основе специальной модели представления знаний /4/, позволяющей повысить семантическую силу обычно применяемых информационно-поисковых языков (классификаций, дескрипторных ИПЯ) и совместить их в рамках целостного лингвистического обеспечения системы.

2. Требования к электронному каталогу электронной библиотеки диссертаций:

- ЭК должен обеспечивать поиск по любым элементам библиографических записей и их сочетаниям, в том числе - по кодам и наименованиям специальностей, индексам и наименованиям делений библиотечно-библиографической классификации (ББК и, возможно, другим классификациям) с учетом смысловых отношений между понятиями, а также по свободным ключевым словам, как это происходит в ЭК РГБ в настоящее время;
- для электронного документа (ЭД), представляющего собой копию диссертации или автореферата, составляется отдельная библиографическая запись (БЗ) в соответствии с форматом MARC21 для электронных ресурсов;
- БЗ для ЭД включаются в общую базу данных ЭК;
- данные об авторах (фамилия, имя, отчество, место защиты и выполнения диссертации) будут формироваться в ЭК как файлы нормативных/авторитетных записей, принятых в современных ЭК.

3. Требования к представлению электронных документов в базе данных ЭБД:

- форматы представления документов в ЭБ выбираются на стадии проектирования. Обязательным форматом считается PDF. Предусматривается также использование языка разметки текстов XML;
- электронные копии авторефератов, тексты содержаний (оглавлений) из диссертаций будут формироваться с распознаванием образов знаков;
- перечни приведенной в диссертациях литературы будет структурирован как некий указатель цитируемой литературы по темам или отраслям знаний;
- электронные копии основного текста диссертации и его частей могут бытьпредставлены как в распознанном виде, так и форме образов в зависимости от целого ряда обстоятельств; иллюстрации представляются в виде образов.

4. Предполагаемые условия доступа к содержанию базы данных:

- электронные копии диссертаций в различных форматах могут быть в локальном или удаленном доступе или храниться исключительно в закрытом электронном депозитарии в зависимости от условий договора с автором;
- электронный каталог и, видимо, данные об авторах предоставляются в открытом удаленном доступе на сайте РГБ.

5. Предполагается использовать следующие источники комплектования ЭБД:

- передача вместе с "бумажным" подлинником диссертации и автореферата или пересылка по почте их электронных копий на дискетах, если таковые создавались при подготовке диссертационной работы и идентичны печатному варианту;
- передача в РГБ электронных копий диссертаций и авторефератов по электронной почте;
- оперативный перевод в электронную форму потока вновь поступающих авторефератов и диссертаций, если автором не представлены их электронные копии;
- ввод автором в режиме он-лайн текстов диссертации и автореферата, включая иллюстративный материал, по специальной форме, заданной на Web-сайте РГБ (в будущем);
- ретроспективное сканирование диссертаций и авторефератов для создания тематических массивов, прежде всего за последние годы, по наиболее часто спрашиваемым направлениям и отраслям (проведено исследование спрашиваемости диссертаций в РГБ по отраслям знаний);
- возможное привлечение в качестве соисполнителя ВНТИЦ как держателя микрокопий диссертаций.

В последнее время диссертации и авторефераты в основном печатаются на основе машиночитаемых текстов, но вероятность появления автора, не имеющего возможности подготовить текст на компьютере, не позволяет обязать всех представлять машинные версии работ и ставить хотя бы одного в неравные условия. Кроме того, многие авторы не имеют машинного варианта диссертации или автореферата, полностью идентичного печатному варианту, представленному к защите.

Комплектование ЭБ диссертаций строится на четкой правовой основе. Диссертант, при желании, будет заключать с РГБ авторский договор, в котором оговариваются права Библиотеки на хранение и использование электронных версий диссертации и автореферата. В электронной библиотеке на Web-сайте РГБ будет закрыта возможность несанкционированного копирования файла диссертации пользователем Интернет. В интранет РГБ тексты будут открыты читателям также только для поиска и просмотра, копироваться исключительно в качестве платной услуги сотрудником Библиотеки при наличии в договоре согласия автора. Доступ к электронному депозитарному хранилищу диссертаций будет разрешен только для специального персонала. В то же время, автор гарантирует в договоре идентичность передаваемой в Библиотеку электронной версии диссертации ее печатному варианту, поступившему на депозитарное хранение в фонд РГБ. Вопрос об идентичности файла печатному тексту также не является тривиальным и

нуждается в тщательной проработке.

В настоящее время проводится работа с различными вузами. Готовятся первые договоры с ними. Содержание договоров может быть различным: ВУЗ может взять на себя всю работу с авторами диссертаций, заключая с ними договоры и передавая электронные версии работ в ЭБ РГБ, или только информирование диссертантов о наличии указанных выше возможностей. С вузом, имеющим авторефераты или диссертации в своей ЭБ, можно договориться о передачи копий в архив РГБ. Надо отметить также, что с ВАК у нас имеется предварительная договоренность о создании депозитарной ЭБ диссертаций именно в нашей библиотеке.

Далее последует работа с советами по защите диссертаций в надежде на их содействие.

Для выяснения отношения авторов к вопросу о передаче в ЭБ диссертаций электронных версий своих работ проводится их анкетирование во время приема печатных работ на хранение, а также в читальном зале отдела диссертаций РГБ. Уже по первым результатам можно рассчитывать на то, что ЭБ будет активно пополняться. Многие авторы готовы немедленно предоставить свои работы на сайт РГБ.

В анкете спрашивается, согласен ли автор выставить автореферат на нашем сайте в сети Интернет, а также предлагается выбрать один из вариантов ответа относительно условий передачи текста диссертации:

- "согласен, для открытого доступа через Интернет (только для чтения),
- согласен, только для доступа читателям РГБ,
- согласен, только для хранения, без обеспечения открытого доступа,
- согласен, при условии включения в электронную библиотеку через полгода после передачи в РГБ,
- согласен, при условии включения в электронную библиотеку через год после передачи в РГБ,
- не согласен".

Сегодня это просто способ выяснения позиции диссертантов и оценка потока поступлений в ЭБ, а завтра такие условия будут вноситься в текст договора РГБ с автором. Кроме того, необходимо оговаривать в нем условия доступа пользователей к различным частям диссертации (содержанию, списку литературы, приложениям, основному тексту) и получить разрешение на электронное копирование, в частности, для службы электронной доставки документов.На обороте анкеты положительные моменты передачи диссертации в машиночитаемом виде формулируются следующим образом: "электронная копия диссертации будет храниться в РГБ вечно, что надежно; Ваша работа появится на Web-сайте РГБ или в ее интранет, что престижно; в ФЭБД Вашу работу пользователи смогут найти не только по библиографической записи в электронном или карточном каталоге, но и по ключевым словам в тексте автореферата, содержания (оглавления) или полном тексте диссертации, что расширит круг Ваших читателей; вместо десятков - сотен человек, которые смогут

добраться до нашего читального зала Вашу работу, увидит любой заинтересованный специалист в любой географической точке, где возможен вывод на экран русского алфавита; Вам как специалисту гарантирована известность; круг Ваших профессиональных контактов существенно расширится; не исключены приглашения к сотрудничеству, к участию в конференциях и совместных проектах, предложения о внедрении результатов; заимствование материалов из Вашей диссертации без ссылки на нее может быть легко обнаружено в любой момент хотя бы одним из сотен и тысяч читателей электронной библиотеки, в особенности - рецензентами других диссертаций; при передаче копий диссертаций и/или авторефератов, их фрагментов по заказам абонентов Библиотека гарантирует автору отчисления от суммы оплаты данной услуги (в соответствии с тем прейскурантом, который будет принят)".

Мы приводим здесь этот текст, так как он лучшим образом выражает наши взгляды на полезность электронной библиотеки диссертаций не только для пользователей, что очевидно, но и для авторов качественных диссертационных работ. Согласитесь, что некоторые отказы от публикации диссертаций в сети Интернет могут быть основаны просто на неуверенности авторов в достоинствах своих работ, хотя достаточно и других причин. Как правило, диссертант стремится к известности и формированию все более широких научных связей с коллегами. Очевидно, что широкая доступность диссертаций и хорошо организованный поиск в базе данных ЭБ будет способствовать сокращению дублирования исследований на стадии выбора научной темы и постановки задачи, формированию научных контактов между отдельными учеными и организациями, а также виртуальных научных коллективов. База данных станет основой для автоматизированного анализа тематики работ, научных тенденций, выявления междисциплинарных связей и анализа использования определенных методов в различных областях знания.

В зарубежной практике можно выделить два варианта организации ЭБ диссертаций: в распределенной ЭБ формируются связи между базами данных, скажем, в различных университетах, где поддерживаются собственные ЭБ диссертаций, и создается некий центр управления такой виртуальной библиотекой с общим электронным каталогом; второй вариант заключается в создании единой базы данных ЭБ в одной организации, причем при наличии развитых ЭБ в университетах такая база данных представляет собой общий архив диссертаций. В наших условиях, реален только второй вариант, так как слишком мало университетов готово к таким нагрузкам, не говоря о наших сетях, а, главное, РГБ и "в традиционном режиме" выполняет функцию депозитарного хранилища диссертаций, так что роль электронного депозитария для нее вполне органична.

Уделим некоторое внимание средствам реализации ЭБ диссертаций, которые мы сейчас разрабатываем с учетом существующего опыта других разработчиков.

Необходимость тщательной проработки метаинформации для электронной библиотеки диссертаций обусловлена, главным образом, разнородностью коллекций, разнообразием решаемых задач и большим объемом слабоструктурированных данных. Метаданные для ЭБ формируются поэтапно. На

первом этапе используются:

- библиографические записи в ЭК РГБ в формате MARC21 (подготовлена структура описания электронных диссертаций и авторефератов);
- описания электронных документов в формате Dublin Core;
- индексы классификации (ББК) в БЗ;
- свободные ключевые слова, дополняющие индексы в БЗ, как это принято в ЭК РГБ;
- наименования и коды специальностей ВАК.

На втором этапе добавляются следующие метаданные:

- иерархические и другие связи между делениями классификации в форме машиночитаемых таблиц ББК;
- нормативные/авторитетные записи для имен лиц, наименований коллективов, заглавий, географических названий.

На третьем этапе (к которому нужно готовиться во время работы на 1-ом и 2-ом этапах), средства тематического поиска будут связаны с семантической моделью. Возможно, будет рассматриваться вопрос об установлении соответствий с другими классификациями и языками предметных рубрик.

Для описания текстов в ЭБ планируется использовать язык разметки XML в качестве формата обмена данными, для импорта - экспорта информации. Для всех типов документов, включаемых в ЭБ, планируется определять стандартные или специфические DTD в качестве грамматик, описывающих комплекс меток XML и их взаимосвязей. В этом плане постараемся ориентироваться на международную деятельность в данной сфере, в частности, на разработку конвертора "MARC21 -> XML", DTD для диссертаций и авторефератов и т.д.

Что касается программного обеспечения, то сейчас мы используем имеющиеся в РГБ программные продукты. Однако в недалеком будущем нам придется при выборе программного обеспечения предъявить к нему следующие требования:

- учет принятых в международной практике стандартов представления, хранения и передачи информации;
- обеспечение адекватного использования выбранных или разрабатываемых в рамках проекта метаданных;
- возможность автоматизированного расчета с пользователями, владельцами и авторами электронных ресурсов;
- минимальные требования к программному обеспечению рабочего места клиента (стандартное программное обеспечение клиента);
- независимость функционирования ЭБ от платформы на стороне клиента и независимость от используемой им СУБД;
- возможность обработки многоязычной текстовой информации с использованием оригинальной графики документов и метаданных;
- спользование UNICODE;
- поддержка распределенных систем хранения информации.

В этом вопросе сложным оказывается выбор между средствами, наилучшими для данной ЭБ, и широко используемыми стандартами, которые обладают меньшей функциональностью.

Основу архитектуры ЭБ РГБ составит принцип Internet/intranet технологий. Основные проектные решения и нормативные документы (в том числе - договоры с авторами), связанные с созданием ЭБД, публикуются на Web-сайте РГБ в разделе "Исследования публикации" (http://www.rsl.ru/h_pub.htm), а полные тексты диссертаций и авторефератов будут помещаться в Открытую русскую электронную библиотеку OREL (http://orel.rsl.ru).

Литература

1. Концепция электронной библиотеки Российской государственной библиотеки // Библиотековедение. - 2001. - N6. - с.33-43
2. Груздев И.А., Лавренова О.А., Перли Б.С. Электронная библиотека РГБ - составная часть РГБ // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Сб. докладов. Третья Всероссийская конференции RCDL'2001, Петрозаводск, 11-13 сентября 2001 г. - Карельский центр
РАН, 2001. - С.232-236
3. Лавренова О.А., Аветисова Т.В. Электронные каталоги Российской государственной библиотеки - реальность // Библиотековедение. - 2000.- N1. - С.52-60
4. Лавренова О.А. Семантическое представление текста на основе модели системы знаний // Научно-техническая информация, сер.2. - 1984. - N4. - С.18-24

Об авторе

Лавренова Ольга Александровна - кандидат филологических наук по специальности "структурная, прикладная и математическая лингвистика", заведующая отделом развития компьютерных технологий и лингвистического обеспечения Российской
государственной библиотеки.
e-mail: lavr@rsl.ru

101000, Москва, Воздвиженка, 3/5.
