На правах рукописи

АБРОСИМОВ АНДРЕЙ ГЕОРГИЕВИЧ

ЭЛЕКТРОННАЯ КОЛЛЕКЦИЯ ПЕРИОДИЧЕСКОЙ ПЕЧАТИ

05.25.03 – Библиотековедение,
библиографоведение и книговедение

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата педагогических наук

Казань 2006

Работа выполнена на кафедре информатики Казанского государственного университета культуры и искусств

Научный руководитель:	доктор физико-математических наук А. М. Елизаров
Официальные оппоненты:	доктор педагогических наук Г. И. Кирилова
	кандидат педагогических наук О. А Калегина
Ведущая организация:	Государственная публичная научно-техническая библиотека России

Защита состоится 28 апреля 2006 года в 14 часов на заседании Диссертационного совета К 210.005.01 по защите диссертаций на соискание ученой степени кандидата педагогических наук при Казанском государственном университете культуры и искусств по адресу: 420059 Республика Татарстан, Казань, Оренбургский тракт, 3

С диссертацией можно ознакомиться в Научной библиотеке Казанского государственного университета культуры и искусств.

Автореферат разослан «____» марта 2006 г.

Ученый секретарь
Диссертационного совета
кандидат педагогических наук, доцент — Л. Е. Савич

Общая характеристика работы

Актуальность темы исследования. Библиотеки, прежде всего библиотеки ВУЗов, традиционно играли роль как библиотечных, так и информационных центров в информационном обеспечении научных исследований и учебного процесса, используя традиционные технологии. Однако развитие вычислительной техники и сетевых технологий, появление новых носителей информации создали предпосылки для коренного изменения традиционных подходов к процессам информационного обеспечения науки и развития новых видов сервиса.

Требования времени, интересы науки, образования обуславливают процесс повышения профессиональной компетентности библиотечных работников, который сегодня включает и развитие навыков работы с разнообразными электронными ресурсами и информационными технологиями. Поэтому под технологическими инновациями в системе библиотечно-библиографической деятельности на данном этапе подразумевается особая педагогическая деятельность, в результате которой библиотечно-библиографическое пространство становится полем моделирования и апробации эффективных нововведений, инициатив, вызывающих качественные изменения в системе библиотечных услуг, приводящих к более рациональному использованию знаний.

В настоящее время для информационного обеспечения научных исследований основной интерес представляют не отдельные публикации, сообщения, наборы данных (отдельные «информационные продукты»), пусть даже сыгравшие выдающуюся роль в истории науки, а организованные и структурированные собрания информационных продуктов, приспособленных для неоднократного использования относительно широким кругом пользователей, то есть информационные ресурсы.

Круг проблем, связанных с созданием и использованием информационных ресурсов фундаментальной науки, весьма обширен. Достаточно перечислить, например, вопросы сбора информации и формирования информационных ресурсов; описания структурных и содержательных особенностей информационных ресурсов и способов представления в них информации в различных научных областях; технологические, организационные, экономические, правовые вопросы их создания и функционирования; проблемы доступности создаваемых информационных ресурсов.

Одним из таких информационных ресурсов, несомненно, являются собрания периодических изданий, как научных, так и массовых, представляющие большую научную и образовательную ценность. В разные моменты своего существования они не только несут актуальную информацию, но и представляют интерес как исторические документы.

В настоящий момент еще мало уделяется внимания созданию полноценных коллекций периодической печати, особенно издававшейся до 20-го века. Исключение составляют научные издания, собрания которых широко

представлены в Интернете и которые создаются как самими издателями (Elsevier, Kluwer Academic Publishers, Springer и т. д.), так и агрегаторами (EBSCO, JSTOR, НЭБ и т. д.). Электронные версии современных периодических изданий, не являющихся научными, представляют собой архивы публикаций, слабо структурированные и имеющие лишь минимальные возможности поиска.

Электронные коллекции периодической печати, изданной до середины 20-го века (проект «Старые газеты», коллекция «Ведомости» Российской национальной библиотеки, коллекция уральских газет Челябинской областной универсальной научной библиотеки и т. д.), не отвечают современным тенденциям развития информационных технологий и электронных библиотек. Так, в некоторых коллекциях отсутствуют метаданные электронных документов, в некоторых в наличии или только средства навигации, или только поиска по коллекции.

В настоящий момент, когда мировое сообщество перешло к массовой практической реализации проектов, связанных с электронными библиотеками, а в России практически каждая крупная библиотека анонсировала свой проект создания электронной библиотеки, **существует необходимость** разработки концепции и технологии формирования электронной коллекции периодической печати, основанных на современных информационных технологиях. Востребованы разработка общей структуры коллекции, типового профиля метаданных и структуры лингвистического обеспечении, а также принципов организации программного обеспечения.

Разработанность темы. В последние годы исследования в области электронных библиотек активно развиваются. Сказанное в полной мере подтверждается тем, что сегодня возникла острая необходимость использования в научных и социально-культурных целях возможностей, предоставляемых современными информационными технологиями и средствами телекоммуникаций. Решение указанных вопросов позволит удовлетворить возросшие требования к предоставляемой информации, а также информационные потребности науки и образования.

Над проблемами создания электронных библиотек работают многие ученые. Об этом свидетельствуют документы и материалы проектов «Электронные библиотеки России»[1], Федеральной целевой программы «Электронная Россия»[2] и проекта «Электронный Татарстан»[3]. Тематика электронных библиотек постоянно обсуждается на международных и всероссийских конференциях – ежегодных конференциях Международной федерации библиотечных ассоциаций и учреждений (ИФЛА), международных конференциях «Крым», «LIBCOM», серии региональных конференций. В профессиональной периодической печати опубликовано большое количество статей. Особенно

[1] http://www.elbib.ru/index.phtml?env_page=backg/concept/programme1.ru.html

[2] http://www.programs-gov.ru/cgi-bin/index.cgi?prg=134&year=2005

[3] http://www.mcrt.ru/index.php?nodeid=457

следует отметить электронный журнал «Электронные библиотеки»[4] – единственный журнал, целенаправленно развивающий названную тематику.

Основные направления исследований в области электронных библиотек развиваются по следующим направлениям и в работах следующих авторов:

- изучение сущности электронной библиотеки, ее специфики, места и роли в системе информационных коммуникаций – А. Б. Антопольский, В. И. Армс, Г. А. Василенко, А. А. Воронов, Ф. Гай, Р. С. Гиляревский, В. О. Громов, Т. В. Ершова, Н. Е. Каленов, О. А. Лаврёнова, П. М. Лапо, Е. А. Негуляев, А. В. Соколов, А. М. Стахевич, Ю. Н. Столяров, О. В. Сюнтюренко, Л. Б. Хайцева, Ю. Е. Хохлов, А. И. Чугунова, Я. Л. Шрайберг и другие;
- создание коллекций информационных ресурсов – М. Р. Когаловский, О. С. Колобов и другие;
- изучение структуры, метаданных и компонентов – А. Б. Антопольский, С. А. Арнаутов, А. Н. Бездушный, Т. Бейкер, Т. Бернерс-Ли, К. В. Вигурский, К. Лагозе, Р. Мерей, Ю. Е. Хохлов, М. Е. Шварцман и другие;
- формирование лингвистического обеспечения – А. Б. Антопольский, Н. И. Гендина, Р. С. Гиляревский, М. В. Губин, Е. М. Зайцева, И. М. Зацман, К. У. Клевердон, М. Г. Крейнес, Ф. У. Ланкастер, И. А. Лурье, Ш. Р. Ранганатан, Э. Р. Сукиасян и другие;
- обсуждение функциональных, информационных и других характеристик отдельных структурных элементов – А. Ю. Абызгильдин, Б. Вегнер, Г. А. Евстигнеева, А. М. Елизаров, Т. И. Ключенко, Р. Р. Назырова, Н. А. Никифоров и другие;
- международное сотрудничество в области электронных библиотек – Д. Оуденарен и другие;
- развитие электронного книгоиздания и системы электронных научных журналов – О. Б. Арушанян, В. Г. Веселаго, В. В. Воеводин, Б. Р. Гельчинский, А. М. Елизаров, Д. С. Латухин, О. В. Сюнтюренко, Ю. Е. Хохлов и другие.

Целью исследования является разработка принципов формирования электронной коллекции периодических изданий как раздела электронной библиотеки, технологий описания и идентификации электронных документов, навигации и поиска в массиве электронных документов, а также реализация и экспериментальная апробация разработанных теоретических положений на примере коллекции казанской периодической печати 19-го – начала 20-го веков.

---

[4] www.elib.ru. Издание Автономной некоммерческой организации «Институт развития информационного общества».

Объект исследования – собрание периодических изданий 19-го – начала 20-го веков из фондов Научной библиотеки им. Н. И. Лобачевского Казанского государственного университета.

Предмет исследования – развитие технологий формирования электронной коллекции периодических изданий, описания и идентификации, навигации и поиска в массиве электронных документов, создание информационной системы, ориентированной на поиск и представление информации пользователю.

Гипотеза исследования. Электронная коллекция периодических изданий, как составная часть электронной библиотеки Казанского государственного университета, реализованная с учетом содержательного наполнения, структуры, технологии, позволит осуществлять результативный поиск и удовлетворять возросшие требования к предоставляемой информации, обеспечивать эффективное взаимодействие с другими коллекциями электронной библиотеки.

В соответствии с обозначенной целью были определены следующие **задачи исследования**:

- изучение принципов организации, технологии формирования электронных библиотек и электронных коллекций, форматов метаданных как инструмента описания и идентификации электронных документов и принципов создания лингвистического обеспечения электронных библиотек и коллекций;
- разработка технологии создания коллекции периодической печати;
- разработка профиля метаданных и лингвистического обеспечения коллекции периодической печати;
- разработка принципов организации программного обеспечения электронной коллекции периодической печати, ориентированного на поиск и представление информации пользователю;
- создание электронной коллекции периодической печати 19-го – начала 20-го веков;
- разработка принципов включения электронной коллекции периодической печати 19-го – начала 20-го веков в электронную библиотеку Казанского государственного университета.

Методология и методика исследования. Методологической основой исследования выступают теория информационного поиска; теория информации; системный подход.

Для решения задач исследования использовался следующий комплекс методов: анализ источников из различных областей знания, позволивший сделать выводы о теоретической и практической разработанности темы; синтез эмпирического (экспериментальная оценка, статистическая обработка данных) и теоретического материалов (анализ литературы по проблеме исследования, операционализация понятий, сравнительный анализ, информационное моделирование).

Экспериментальной базой диссертационного исследования является собрание периодической печати 19-го – начала 20-го веков из фондов Научной библиотеки им. Н. И. Лобачевского Казанского государственного университета.

Научная новизна и теоретическая значимость диссертационного исследования заключается в следующем:

- разработана концепция создания электронной коллекции периодических изданий как составной части электронной библиотеки Казанского государственного университета;
- определены пути взаимодействия коллекций электронной библиотеки Казанского государственного университета и отношения между их ресурсами;
- разработаны и обоснованы профиль метаданных коллекции периодической печати и структура лингвистического обеспечения;
- разработаны принципы организации программного обеспечения электронной коллекции периодической печати.

Практическая значимость исследования:

- создана электронная коллекция периодической печати 19-го – начала 20-го веков;
- сформировано программное обеспечение электронной коллекции периодической печати;
- предложены практические рекомендации по их использованию.

Достоверность и обоснованность результатов исследования обеспечивались использованием фундаментальных теоретических положений в области теории информации и информационного поиска, комплекса теоретических и эмпирических методов.

Основные положения и выводы диссертационной работы обсуждались на объединенном заседании специальных кафедр информационно-библиотечного факультета Казанского государственного университета культуры и искусств, апробированы на научно-теоретических и научно-практических конференциях международного, всероссийского и регионального уровней.

Международные научные конференции: Восьмая Международная конференция и выставка «LIBCOM 2004»: Информационные технологии, компьютерные системы и издательская продукция для библиотек (Звенигород, Московская область, 15 – 19 ноября 2004 г.); Девятая Международная конференция и выставка «LIBCOM 2005»: Информационные технологии, компьютерные системы и издательская продукция для библиотек (Звенигород, Московская область, 14 – 18 ноября 2005 г.); Международный семинар «Российские электронные ресурсы по науке и технике. Проект РусЭМБ как часть международной электронной библиотеки по математике» (Москва, 1 – 2 февраля 2006 г.).

Всероссийские, региональные и республиканские научно-практические конференции: Девятая ежегодная конференция АДИТ–2005: Культурное многообразие в едином информационном пространстве (Россия, Республика Татарстан, г. Казань, 30 мая – 3 июня 2005 г.); Всероссийская научно-практическая конференция: Университетские библиотеки в глобальном информационном и культурном пространстве (Казань, Научная библиотека КГУ, 7 – 8 декабря 2004 г.).

Положения, выносимые на защиту:

- концепция, технология создания и структура электронной коллекции периодической печати;
- профиль метаданных, позволяющий создать полноценное описание информационных ресурсов коллекции и обеспечить интероперабельность создаваемой информационной системы;
- структура лингвистического обеспечения, позволяющая осуществлять навигацию и результативный поиск по коллекции;
- принципы организации программного обеспечения коллекции периодической печати;
- способы включения электронной коллекции периодической печати в электронную библиотеку.

Структура и объем диссертации подчинены логике диссертационного исследования. Основное содержание диссертации состоит из введения, трех глав, содержащих 8 параграфов, заключения, библиографии, приложений.

В списке литературы отражены цитируемые и упоминаемые источники на русском и английском языках.

В приложениях представлены:

- описание профиля метаданных с примерами XML-документов;
- примеры описаний спецификаций XML;
- значения фасетов лингвистического обеспечения коллекции.

Работы по созданию электронной коллекции периодической коллекции 19-го – начала 20-го веков проводились при финансовой поддержке Российского гуманитарного научного фонда (проект № 04-01-12032в) и Российского фонда фундаментальных исследований (проект № 02-07-90230).

Основное содержание диссертации

Во введении обосновывается актуальность темы, анализируется степень разработанности проблемы, определяются цели и задачи исследования, формулируются его методологические основы и научная новизна.

Глава первая «*Электронные библиотеки и электронные коллекции. Определения, основные свойства*» включает два параграфа и посвящена рассмотрению определений основных понятий, задач и функциональных воз-

можностей электронных библиотек (ЭБ) и электронных коллекций, анализу опыта создания электронных коллекций периодической печати.

Появление в 1980-е годы средств вычислительной техники и информационных технологий, обеспечивающих надежное сохранение, оперативную обработку и эффективное использование больших массивов разнородной информации, привело к активному развитию работ по электронным библиотекам. Конкретизировалось понятие электронной библиотеки, определились и уточнились ее цели, задачи и функции.

Изучение понятия электронной библиотеки необходимо начать с определения связанного с ней термина «электронный документ». Обобщая все известные определения электронного документа, необходимо отметить следующее:

- электронный документ может представлять собой один или несколько файлов любого формата;
- «электронный документ – ограниченный и завершенный на конкретный момент времени» – это наиболее удачная формулировка предложена в проекте концепции Национальной электронной библиотеки[5].

Преимущества электронных документов перед традиционными бумажными носителями очевидны и состоят в следующем:

- электронные документы не локализованы; благодаря телекоммуникационным связям электронный документ может быть использован из любой точки мира;
- электронный документ может использоваться одновременно несколькими пользователями в одно и то же время;
- электронный документ легко копируется;
- электронные документы очень гибкие, их легко можно переформатировать, сочетать с другими документами, изменять и т. д.;
- коллекции электронных документов занимают намного меньшее по объему место, чем их традиционные виды, имеется тенденция к еще большему сжатию и компактному хранению электронных коллекций.

Из целого ряда существующих определений, в той или иной степени отражающих сущность электронных библиотек, наиболее адекватным представляется следующее определение Т. В. Ершовой и Ю. Е. Хохлова[6]: «электронная библиотека – это распределенная информационная система, позволяющая надежно сохранять и эффективно использовать разнородные коллекции электронных документов (текст, графика, аудио, видео и др.), доступные в удобном для пользователя виде через глобальные сети передачи данных».

---

[5] http://www.rsl.ru/pub.asp?13.htm Национальная электронная библиотека: Концепция. Проект.

[6] Ершова Т. В., Хохлов Ю. Е. Межведомственная программа «Российские электронные библиотеки» // Элек-тронные библиотеки. – 1999. – Т. 2. – Вып. 2.

Подразумевается, что ЭБ должна обеспечивать хранение информации в цифровой форме практически неограниченное время и предоставление всем заинтересованным потребителям качественно новых возможностей работы с большими объемами информации. К таковым, например, можно отнести последовательный, выборочный или параллельный просмотр множества документов; многоаспектный поиск во всем объеме информации, хранимой в данной электронной библиотеке; копирование необходимых документов или их фрагментов как на бумагу, так и на современные носители; создание собственных документов и, наконец, производство нового знания.

Создание и использование ЭБ реализуется через:

- накопление, хранение, учет и структурирование электронной информации;
- организацию навигации во всем информационном пространстве, доступном через данную электронную библиотеку;
- обеспечение эффективного доступа к ней любого числа пользователей через телекоммуникационные сети.

Завершая обсуждение термина «электронная библиотека», необходимо привести цитату из материалов симпозиума Report on NSF Workshop «Distributed Knowledge Work Environments»[7]: «Не следует отождествлять электронную библиотеку с совокупностью оцифрованных коллекций и инструментария управления ими. ЭБ нужно понимать более широко, как среду, объединяющую коллекции, сервисы и людей для поддержки полного жизненного цикла создания, распространения, использования и сохранения данных, информации и знаний».

Во многих работах, посвященных ЭБ, описываются возможности и преимущества, которые предоставляют электронные библиотеки:

- решаются три главные библиотечные проблемы: малая экземплярность изданий, нехватка площадей для хранения фонда, сохранность книжного фонда;
- полнее удовлетворяются информационные запросы пользователей библиотеки, пользователь получает информацию независимо от времени и места нахождения;
- открываются новые формы библиотечного и информационного обслуживания пользователей, в том числе обслуживания инвалидов по зрению;
- существенно повышается оперативность предоставления пользователям необходимых документов и данных; для ряда пользователей электронная форма предоставляет единственную возможность получить требуемый документ;

[7] March 9 – 11, 1997, Santa Fe, New Mexico

– предоставляется возможность производить работу с электронными документами, которая выходит за рамки простого чтения текста или просмотра изображения (в том числе редактировать, соединять, добавлять, вводить подразделы, перестраивать электронные документы, создавать на их основе новые);
– повышается уровень информационной культуры и компьютерной грамотности как читателей, так и сотрудников библиотеки, а это имеет не только узкое прикладное, но и более широкое значение, поскольку человек с высоким уровнем информационной культуры значительно легче ориентируется в постоянно меняющемся мире, не боится новаций и перемен;
– использование мультимедийных компьютеров, CD-ROM, предоставляющих текстовую, аудио- и видеоинформацию позволяет лучше усваивать материал, так как информация воспринимается комплексно, несколькими органами чувств одновременно.

С электронной библиотекой связано понятие «коллекция электронных документов» или «коллекция информационных ресурсов». Наиболее подробно эту тему разрабатывает в своих работах М. Р. Когаловский[8], который определяет коллекцию следующим образом: «Коллекция информационных ресурсов представляет собой систематизированную совокупность информационных ресурсов, объединенных по какому-либо критерию принадлежности, например, по общности содержания, источников, назначения, по кругу пользователей, способу доступа и т. д.». В общем случае, как это было видно из определений, приведенных выше, электронная библиотека – это гетерогенная система, то есть объединяющая самые разнообразные данные. С другой стороны, обязательными свойствами ЭБ являются структурированность, систематизированность содержания, что приводит к необходимости разделить ресурсы электронной библиотеки на группы, объединив в них электронные документы по какому-либо признаку. Такие группы по самой своей сути соответствуют определению коллекций информационных ресурсов.

Таким образом, можно утверждать, что электронная библиотека в общем случае представляет собой иерархическую систему и состоит из коллекций электронных документов, которые определяют логическую структуру библиотеки.

На сегодняшний день количество электронных библиотек постоянно возрастает, при этом формируются библиотеки самых различных направлений – научные, литературные, учебные и т. д. Учитывая направленность данной работы, далее предлагается краткий анализ электронных коллекций периодических изданий.

---

[8] Когаловский М.Р. Научные коллекции информационных ресурсов в электронных библиотеках // Первая Всероссийская научная конференция “Электронные библиотеки: перспективные методы и технологии, электронные коллекции”. Санкт-Петербург, 19 - 21 октября 1999 г.

Название коллекции	Профиль метаданных	Характеристики поиска и навигации по коллекции	Формат электронных документов
Проект «Старые газеты»	метаданные отсутствуют	навигация по коллекции позволяет от списка газет перейти к конкретному номеру газеты, поиск по коллекции отсутствует	формат HTML и PDF
Коллекция Российской национальной библиотеки. Первая русская газета «Ведомости»	метаданные в формате RUSMARC	предлагается упрощенный интерфейс к электронному каталогу, то есть отыскиваются библиографические описания статей, содержащие все слова, которые введены в поисковую форму	графический образ страницы
Коллекция Уральских газет из Челябинской областной универсальной научной библиотеки	на сайте метаданные отсутствуют	навигация осуществляется с помощью календаря, что позволяет выйти на номера газеты, изданные в выбранный день, поиск по коллекции отсутствует	графический образ статьи
Коллекция научных журналов JSTOR	свой (оригинальный) профиль метаданных	навигация предлагает возможности выбора конкретного номера журнала, нужной статьи, имеются атрибутный поиск, полнотекстовый поиск	графический образ страницы журнала, существует распознанный текст
Университетская информационная система РОССИЯ Раздел «Средства массовой информации»	свой (оригинальный) профиль метаданных	поисковая система дает возможности поиска по метаданным и по тексту статей, имеются тезаурус и рубрикаторы	текстовые документы
Научная электронная библиотека (НЭБ)	свой (оригинальный) профиль метаданных	поиск производится по метаданным и полным текстам статей	текстовые документы

Анализ приведенных сведений позволяет сделать следующие выводы:

– не существует единого подхода к созданию коллекций периодической печати;
– проект JSTOR – хороший пример того, как должна быть организована коллекция периодических изданий, так как присутствует оригинальный профиль метаданных и имеется возможность атрибутного поиска по ним, электронные документы состоят из графических образов страниц и распознанного текста, что позволяет организовать поиск и по содержимому статей изданий;
– существует потребность в разработке концепции и технологий создания коллекций периодической печати, позволяющих организовать удобные

для пользователей средства работы, в том числе навигации и поиска, атрибутного и по тексту документов.

Во второй главе «*Электронная коллекция периодической печати 19-го - начала 20-го веков*» описаны технологии создания, структура, профиль метаданных, лингвистическое обеспечение и программное обеспечение электронной коллекции периодической печати 19-го – начала 20-го веков, создаваемой в Научной библиотеке им. Н.И. Лобачевского Казанского государственного университета.

За двухсотлетнюю историю в Научной библиотеке Казанского государственного университета (КГУ) сформировалось богатейшее собрание документов 9-го – 20-го веков на русском, западноевропейских и восточных языках, включающее уникальную коллекцию редких книг и рукописей (более 60-ти тысяч единиц хранения). Большую ценность представляет коллекция цензорских экземпляров казанских газет конца 19-го – начала 20-го веков, в которых сохранились наряду с ценнейшим краеведческим материалом первые редактуры произведений А. М. Горького, В. Г. Короленко, Н. Г. Гарина-Михайловского и др., активно печатавшихся на страницах казанской прессы. В Научной библиотеке КГУ сосредоточена основная и наиболее полная коллекция арабографических татарских газет и журналов, издававшихся в различных городах России в начале 20-го века.

Коллекция редких книг и рукописей активно используется в учебном процессе и научных изысканиях. При таком интенсивном использовании часть коллекции (в особенности комплекты газет и журналов, изданных в начале прошлого века на бумаге плохого качества) пришла в негодность и не выдается читателям. Кроме того, сказываются естественное старение и разрушение бумаги. Таким образом, существует реальная угроза потери части коллекции, которая, как исторический источник, имеет не только национальное, но и международное значение.

В связи с вышеизложенным, приоритетным направлением в создании электронной библиотеки КГУ является формирование коллекций электронных документов на основе фондов отдела рукописей и редких книг библиотеки и, в первую очередь, коллекции периодической печати конца 19-го – начала 20-го веков.

При создании коллекций электронных документов формирование электронных документов возможно двумя основными способами – сканирование печатных документов и получение готовых электронных документов от авторов, издательств и т. д. Однако при создании коллекций старых изданий возможен только первый способ, и наилучшим вариантом является сканирование с использованием специализированного планетарного сканера формата А2.

Сканирование документов позволяет решить одновременно две задачи: получение страховых копий; формирование изображений для электронной коллекции.

В первом случае требуется получить изображение высокого качества, которое в дальнейшем может быть использовано для печати, распознавания текста и т. д.

При решении второй задачи – формирования изображений для электронной коллекции – определяющим является удобство работы читателя, то есть качество и скорость вывода изображения на экран. Следовательно, размер изображения должен привязываться к самому популярному типу мониторов, а размер файла – минимизироваться. Если же планировать возможность работы с этими файлами читателей с ослабленным зрением, то должна предусматриваться возможность хотя бы двукратного увеличения изображения.

Таким образом, необходимо наличие обоих вариантов изображений – страховых копий для хранения в архиве и модифицированных изображений для электронной коллекции.

Наиболее удобным форматом данных для читателя является текстовый. Возможности поиска, копирования цитат и т. п. обеспечивают максимальную эффективность обработки электронных документов. Для электронных документов, созданных на базе старых изданий, оптимален следующий вариант – каждая часть документа хранится в двух видах: копия оригинала и текстовый вариант в формате, пригодном для организации поиска по тексту документа. Это позволит представить естественный вид документа и получить максимальные возможности его обработки.

Но для изданий 19-го века основной проблемой становится качество печати самих изданий. Особенно это относится к периодике, которая не только печаталась на низкосортной бумаге, но и качество печати было низким. На сегодняшний день не существует эффективных средств распознавания для изданий 19-го века и более старых. Трудозатраты на исправление распознанного текста сравнимы с трудозатратами на набор этого текста с оригинала.

Технологический процесс создания электронных документов на основе отсканированных изображений периодических изданий можно представить в виде следующей последовательности действий:

- просмотр и подготовка бумажного издания; определение параметров сканирования;
- сканирование, результат файл с изображением формата TIFF;
- постраничная разрезка отсканированных разворотов, склейка изображений страниц, отсканированных частями;
- контроль сканирования и исправление ошибок (пересканирование «бракованных» или пропущенных страниц);
- формирование архивного документа;

– постраничная обработка (удаление дефектов изображения и восстановление истинных размеров страницы);
– конвертирование в формат пригодный для представления читателю;
– распознавание и форматирование текста, создание электронного документа, состоящего из двух частей – текста и изображения.

Коллекции – наиболее распространенная форма организации информационных ресурсов в электронной библиотеке и представляет собой систематизированную совокупность электронных документов, объединенных по какому-либо критерию принадлежности, например, по общности содержания, источников, назначения, по кругу пользователей, способу доступа и т.д. Тем не менее, внутри коллекции можно выделить отдельные группы электронных документов, более тесно связанных между собой и образующих разделы коллекции. В общем случае разделы коллекции не обязательно должны быть явно выражены. В формируемой коллекции периодической печати 19-го – начала 20-го веков выделение явных разделов производится естественным образом – разделом коллекции является группа электронных документов отдельного издания (газеты, журнала).

С функциональной точки зрения информационные ресурсы коллекции подразделяются на данные (электронные документы) и метаданные. Метаданные – это специально подготовленные, машиночитаемые, структурированные сведения о ресурсе, представляющие свойства, которые имеет ресурс, услуги, которые предоставляет ресурс. Соответственно на основе системы метаданных строятся основные технологические процессы, а именно:
– поиск и навигация в информационном пространстве коллекции;
– ввод и изъятие электронных документов, организация их хранения;
– управление правами доступа к электронным документам, включая защиту авторских прав, и т. д.

Особенностью коллекции периодической печати является определение электронного документа. Основной единицей хранения архивных копий служит страница издания, она же могла являться электронным документом. Но естественной единицей информации является статья, которая может быть и меньше страницы, и располагаться на нескольких страницах. Другой особенностью газет того времени, в особенности первой половины 19-го века, является то, что значительная часть публикаций не озаглавлена и не подписана авторами, что делает не возможным полное библиографирование по заголовкам и фамилиям. Газета состоит из более или менее кратких сообщений, объявлений и других видов информации, многообразие которых затрудняет организацию поиска нужных сведений. Поэтому в качестве описываемой единицы содержания выбрана часть текста, ограниченная определенной темой – сообщение о международных отношениях, сообщение о научных открытиях, всевозможные объявления, статьи и т. д.

В этом случае электронный документ не совпадает с единицей описания. Фактически это означает, что помимо описания электронного документа

в структуре метаданных коллекции появляется еще один уровень – описание единиц содержания номера издания, и весь поисковый и навигационный аппарат будет ориентирован на работу именно с этими метаданными.

Таким образом, структура метаданных представляет собой:

- ***описание коллекции***, содержащее описание структуры коллекции, список разделов коллекции с указанием типа, тематики издания, общую информацию о коллекции, подборка статей о коллекции и т. д.;
- ***описание разделов коллекции*** – конкретных изданий, содержащее общее описание издания, описание структуры издания, блок выпускных данных, количественные характеристики, историческую справку с информацией об основании издания, сведений об основателе, и т. д.;
- ***описание электронных документов*** – номеров, страниц, рубрик изданий;
- ***описание единиц содержания*** – статей, объявлений, сообщений и т. д.

Еще одно предназначение метаданных – обеспечение интероперабельности информационной системы. Самый простой способ обеспечения интероперабельности – воспользоваться уже известными схемами метаданных. Второй вариант – создать собственное формальное описание схемы – достаточно сложный и, самое главное, очень сильно зависящий от выбранной модели описания метаданных.

При создании настоящей коллекции было принято решение разрабатывать свой профиль метаданных, используя при формировании метаданных Dublin Core и добавляя свои дополнительные классификаторы.

Использование при создании метаданных XML-технологий обусловлено тем, что язык разметки XML рассматривается сегодня как одна из ключевых технологий для построения современных электронных библиотек. Можно привести следующие доводы в пользу их применения:

- широкое распространение языка XML при разработке Интернет-приложений в последние годы, опора на XML и RDF (Resource Description Framework) в создании Семантической Сети (Semantic Web);
- развитые возможности содержательного структурирования текстов;
- удобство архивации XML-представлений документов для долговременного хранения, обусловленное их текстовым форматом;
- возможность организации полнотекстового поиска по всей коллекции, а не только по отдельному электронному документу;
- электронные документы доступны для индексирования поисковыми системами Интернета;
- XML-документы могут быть легко заимствованы другой информационной системой, то есть создаются условия для обеспечения интероперабельности коллекции;
- представление информационных ресурсов средствами стандартов XML обеспечивает навигационный доступ к электронным коллекциям с

помощью Веб-браузеров, поддерживающих эти стандарты, то есть с помощью средств, привычных для пользователей WEB.

Моделью описания метаданных является система RDF (Resource Description Framework). Спецификация содержания документов средствами стандарта RDF дает возможность семантического поиска информационных ресурсов в среде, поддерживающей такие метаданные.

Метаданные можно рассматривать как набор утверждений о свойствах характеризуемого ресурса, представляющих собой тройку: ресурс, именованное свойство и его значение. Под свойством следует понимать некий аспект, характеристику, атрибут или отношение, используемое для описания ресурса. Каждое свойство имеет свой специфический смысл, допустимые значения, тип ресурсов, к которым оно может быть применено, а также отношения с другими свойствами. Это же понятие тройки – ресурс, свойство, значение – является основой RDF.

Для обеспечения интероперабельности и описания синтаксиса XML-документа, содержащего метаданные, разработаны XML Schema для каждого типа XML-документов, определено XML-пространство имен. Описание XML-документов средствами XML Schema позволяет осуществлять более тонкую верификацию целостности представленных XML-документов.

Одной из ключевых задач развития электронной коллекции является повышение эффективности навигации и поиска в массиве электронных документов. Решение этой проблемы предусматривает создание информационной системы, ориентированной на поиск и представление информации пользователю. Центральное место в такой системе занимает лингвистическое обеспечение – комплекс информационно-поисковых языков и лингвистических процессоров, предназначенных для обработки, представления и поиска электронных документов на семантическом уровне.

Обычно вербальные языки, среди которых широко распространен язык ключевых слов, являются центральным лингвистическим средством для выражения смыслового содержания текста. Достоинства координатного индексирования общеизвестны, но свойства предметной области коллекции и способ представления электронных документов, в случае создаваемой коллекции, делают этот процесс проблематичным и неоправданно трудозатратным. Среди основных причин отказа от применения этого подхода можно выделить следующие:

во-первых, невозможность применения методов автоматической обработки текста для полнотекстового индексирования и основанного на них полнотекстового поиска;

во-вторых, неконтролируемое и неуправляемое применение ключевых слов неизбежно приводит к значительным потерям в характеристиках полноты и точности поиска, для устранения этих недостатков как минимум необходимо решение проблемы дескрипрторизации ключевых слов, а для организа-

ции действительно содержательного поиска предстоит сложный и трудоемкий процесс создания тезауруса.

В качестве основного инструмента для систематизации и поиска газетных публикаций был выбран подход, основанный на многоаспектной классификации текстов. Классификационные языки обладают рядом преимуществ перед другими типами поисковых языков, прежде всего, наглядностью, простотой для пользователя и независимостью от естественного языка. Фасетная система классификации позволяет выбирать признаки классификации (фасеты) независимо друг от друга. Каждый фасет содержит совокупность однородных значений данного классификационного признака. Фасетная система классификации позволяет многоаспектно (всесторонне) охарактеризовать специфический газетный материал.

В процессе аналитико-синтетической переработки текста изданий, вошедших в создаваемую коллекцию, описывалась информация каждой единицы содержания в соответствии с разработанной системой классификации. Систематизация содержания издания осуществлялась по следующим аспектам: *виду информации, сфере общественной жизни, персонам* (именам, встречающиеся в газете), *учреждениям, географическим названиям мест, датам событий,* приведенным в тексте. Внутри фасетов значения признаков либо просто перечисляются, либо образуют иерархическую структуру, если существует соподчиненность выделенных признаков.

Поскольку заранее невозможно определить все необходимые исследователю аспекты рассмотрения материалов периодического издания и необходимо учитывать неизбежный субъективизм систематизации, предусмотрена группировка текстов и по такому формальному признаку, как газетная рубрика. Несмотря на непостоянство названий рубрик в разных номерах газеты и неадекватное отражение ими тематики текстов, размещенных в разделе, такая систематизация наиболее точно и полно отражает структуру номеров газеты и может предоставить ряд дополнительных возможностей, в том числе и при поиске.

Дополнительно в процессе аналитико-синтетической обработки газетного текста формируются авторитетные файлы *авторов, жанров периодики, список периодических изданий,* из которых перепечатан новостной материал.

Выбор XML-технологий в качестве основы формирования метаданных коллекции обусловил дальнейшие шаги по выбору программного обеспечения коллекции, которое должно обеспечить следующие функциональные возможности:

- навигация во всем доступном информационном пространстве – наглядное предоставление пользователю логической структуры информационного пространства;
- атрибутный поиск – информационный поиск объектов по значениям их характеристик;

– просмотр содержания информационного объекта и его структуры: последовательный (например, страница за страницей) и выборочный (переход на любую заданную страницу или на любой элемент, отраженный в структуре).

Навигация во всем доступном информационном пространстве рассматриваемой коллекции предполагает построение иерархической системы, позволяющей выбрать в коллекции необходимое издание, найти требуемый номер издания и просмотреть его целиком, листая страницы, или получить его оглавление, то есть список статей и выбрать требуемую статью.

Визуализация классификационной схемы на странице поиска информационной системы может существенно облегчить тематический поиск информации пользователю, не знакомому с формализмами записи фасетных формул. Система визуализации должна позволить осуществлять выбор и эффективно формировать поисковый запрос при любой последовательности выбора фасетов.

На первом уровне программное обеспечение коллекции можно функционально разделить на две основные части:

– подсистема формирования метаданных, которая обеспечивает ввод и коррекцию метаданных, формирование XML-документов с метаданными и т. д.;
– подсистемы навигации и поиска в информационном пространстве коллекции, которая обеспечивает последовательный и выборочный просмотр газетных номеров, навигацию во всем доступном информационном пространстве представленном в виде иерархической структуры, формирование тематического поискового запроса на основе визуализированной схемы классификации, использование записей авторитетных файлов при формировании запроса и т. д.

Второй уровень программного обеспечения коллекции должен, помимо расширения вышеперечисленных функций, включать:

– средства ведения базы архивных копий, электронных документов и XML-документов с метаданными;
– средства генерации подсистемы формирования метаданных, используя в качестве данных XML Schema, описывающий формат XML-документа для разработанного профиля метаданных;
– средства генерации подсистемы навигации и поиска, используя в качестве данных XML Schema.

Третий уровень программного обеспечения коллекции включает формирование RDF Schema и онтологии коллекции, то есть обеспечение полной интероперабельности информационной системы в том смысле, который предполагает Semantic Web.

Известные программные системы создания электронных библиотек и электронных архивов не обладают достаточными возможностями для реализации разработанного профиля метаданных и выбранной схемы классифика-

ции, следовательно, необходима разработка собственного программного обеспечения сопровождения электронной коллекции.

На сегодняшний день полностью реализован первый уровень программного обеспечения коллекции.

В третьей главе «*Коллекция периодической печати как составная часть электронной библиотеки*» описаны возможности применения результатов, полученных при создании электронной коллекции периодической печати 19-го – начала 20-го веков, при формировании других коллекций периодической печати и пути включения электронных коллекций в электронную библиотеку на примере создаваемой электронной библиотеки Казанского государственного университета.

Газеты и журналы, входящие в формируемую коллекцию, самых различных типов и направлений. Так, например, содержание газеты «Казанские известия» представляет собой многоплановый синтетический материал, включающий в себя самую разнообразную по жанру, происхождению, содержанию информацию. Наряду с официальными сообщениями и документами, законодательными актами, объявлениями, некрологами, письмами, городской хроникой в ней публикуется обширный научно-образовательный и литературно-художественный материал. Журнал «Заволжский муравей» – литературный. Таким образом, коллекция представляет широкий спектр изданий, за исключением только научных.

В связи с этим опыт, полученный при создании коллекции периодических изданий 19-го – начала 20-го веков, может быть использован при создании других коллекций периодической печати, включающих и современные издания.

С другой стороны, коллекции периодической печати 19-го – начала 20-го веков имеет свою специфику:

- в качестве электронных документов выступают отсканированные изображения, так как распознавание текста изданий слишком сложно;
- слабо выражена структура издания, значительная часть публикаций не озаглавлена и не подписана авторами;
- издания состоят из более или менее кратких сообщений, объявлений и других видов информации, многообразие которых затрудняет организацию поиска нужных сведений;

и, как следствие из выше перечисленного:

- электронным документом является отсканированное изображение номера газеты, статьи журнала;
- единица описания и электронный документ не всегда совпадают между собой;
- при отсутствии возможности полнотекстового поиска, возрастает значение лингвистического обеспечения;

– при формировании лингвистического обеспечения упор делается на фасетную классификацию;
– разработанный профиль метаданных удобнее реализовать с помощью XML-технологий;
– использование XML-технологий требует разработки собственного программного обеспечения.

В процессе работы над коллекцией периодической печати 19-го – начала 20-го веков был определен алгоритм создания коллекции:

– анализ издания, определение типа издания;
– определение электронного документа, определение описываемой единицы содержания;
– анализ структуры издания, выделение типовых рубрик, формирование классификационных справочников на основе структуры издания, определение схемы лингвистического обеспечения, формирование списков значений признаков, формирование профиля метаданных;
– формирование электронных документов – сканирование, обработка изображений, распознавание текста;
– формирование метаданных;
– разработка навигационного и поискового программного обеспечения.

Применительно к современным изданиям алгоритм формирования коллекции не требует модификации. Более того, состав и структура фасетов предлагаемой системы классификации не зависят от типа и времени выхода издания. Хотя в настоящий момент и существует более четкая специализация изданий, тем не менее, характер информации, включаемой в отдельный номер издания, самый разнообразный, что и являлось одной из основных причин выбора системы классификации, и определения состава фасетов. Фасеты «Вид информации», «Сферы общественной жизни», «Персона», «Учреждения», «Географические названия мест» одинаково применимы для современных изданий и периодики 19-го века практически всех типов. Естественно, что будут уточняться списки значений, возможно введение новых фасетов, отражающих специфику конкретных изданий и/или коллекций или более общих, таких, как, например, «Жанр публикации», но и предлагаемый список фасетов уже достаточен для полноценного описания материалов, публикуемых в периодических изданиях.

Исключением являются научные журналы, сборники статей. Для научных изданий гораздо естественнее использовать уже существующие классификационные языки – ББК, УДК и т. д. Более того, уже имеются хорошо проработанные профили метаданных для научных статей – например, Sarticle – разработка для Научной электронной библиотеки e-library.ru.

Разработанный в диссертации профиль метаданных предназначен для описания конкретной коллекции и не является универсальным. Но принцип, использованный при его разработке, – использование схемы Dublin Core с

уточняющими квалификаторами – можно считать универсальным. А. Б. Антопольский[9] считает: «Таким образом, Дублинское ядро представляется как вершина иерархии систем метаданных, которая развивается более детально в конкретных коллекциях или сервисах системы электронной библиотеки при помощи частных систем метаданных». Фактически это означает, что для обеспечения первого уровня интероперабельности достаточно корректного использования Dublin Core с необходимыми для конкретной коллекции уточняющими квалификаторами.

Таким образом, использованный подход к созданию коллекции периодической печати 19-го – начала 20-го веков может быть предложен и для создания других коллекций периодики, исключая коллекции научных изданий.

Состав электронной библиотеки в общем случае – набор коллекций электронных документов. А основная цель создания ЭБ – обеспечение универсального доступа к информации и информационным службам. При этом совершенно не обязательно наличие только универсального доступа, каждая коллекция может иметь свой собственный дополнительный аппарат навигации и поиска, ориентированный на специфические особенности, присущие только этой коллекции. Достаточно распространенной является схема, когда пользователю помимо унифицированной поисковой системы предлагается список коллекций, по которому можно перейти к работе со средствами навигации и поиска конкретной коллекции.

Это создает дополнительные удобства для пользователя, так как может возникнуть необходимость последовательного просмотра номеров издания, а такую возможность могут предоставить только средства навигации по конкретной коллекции – последовательный список номеров в коллекции периодической печати, например, что не возможно при работе только с общим поисковым аппаратом.

В реальности создание электронной библиотеки начинается до того, как разрабатывается ее концепция. Так, в Научной библиотеке КГУ в настоящий момент независимо друг от друга формируются коллекция авторефератов диссертаций, коллекция «Труды ученых КГУ», коллекция учебно-методической литературы, коллекция периодической печати 19-го – начала 20-го веков и т. д., для их создания используются различные технологии, профили метаданных и лингвистические средства.

Такой подход естественно создаст определенные трудности в дальнейшем, центральная из которых – объединение существующих коллекций в электронную библиотеку, так как одна из основных задач при создании ЭБ – обеспечить навигацию по электронной библиотеке и поиск электронных документов по всем разнообразным коллекциям, входящим в нее.

---

[9] Антопольский А.Б. Лингвистическое обеспечение электронных библиотек.// М.: ФГУП Научно-технический центр «Информрегистр», 2003. – 302 с.

В настоящий момент в Научной библиотеке КГУ существует два типа коллекций, самым существенным различием между которыми является способ организации. Это коллекции организованные с помощью АБИС «Руслан», эксплуатируемой в Научной библиотеке КГУ, и коллекции ориентированные на использование XML-технологий с профилем метаданных, основанном на Dublin Core.

Объединение коллекций первого типа не составляет труда, они уже существуют в единой системе, с единым поисковым аппаратом, единым лингвистическим обеспечением и единым профилем метаданных – RUSMARC.

Объединение коллекций второго типа производится так же достаточно просто, если в основе их профилей метаданных лежит одна и та же схема – Dublin Core. В этом случае при создании общей поисковой системы, то есть при обеспечении универсального доступа к информации, достаточно игнорировать все дополнительные уточняющие квалификаторы, введенные при разработке профиля метаданных коллекции, и обрабатывать содержащиеся в них данные как единое содержимое элемента Dublin Core.

Основной проблемой является объединение коллекций разных типов. К сожалению, в настоящий момент неизвестны примеры выполнения таких работ. Естественными представляются два простых подхода к решению этой проблемы:

– преобразование метаданных коллекций, созданных в АБИС, в Dublin Core и перевод в XML-документы;
– преобразование метаданных коллекций, созданных с использованием Dublin Core и XML-технологий, в RUSMARC и загрузку в АБИС.

Более сложное решение – разработка программного обеспечения, осуществляющего поиск одновременно и в базах данных АБИС, и в массиве XML документов с метаданными.

В настоящий момент наиболее просто реализуемым представляется второй вариант, поскольку он не требует разработки дополнительного программного обеспечения, а все возможности поиска по электронной библиотеке обеспечивает АБИС.

Наиболее перспективным представляется третье решение. Несмотря на высокую трудоемкость его реализации, оно позволит наиболее полно использовать все возможности для организации поиска и навигации, заложенные при разработке конкретных коллекций. Дополнительной возможностью такого решения является использования данных из электронного каталога, не относящихся к электронной библиотеке.

Подводя итоги, можно сделать следующий вывод – опыт создания коллекции периодической печати 19-го – начала 20-го веков может быть использован при формировании других коллекций периодической печати. Более того, развитие электронных коллекций основанных на перспективных технологиях Semantic Web, будет способствовать созданию эффективной инфра-

структуры для поддержки научных исследований, образования и других сфер деятельности.

Список работ, опубликованных по теме диссертации

1. О комплексе программ «Картотека фондов музея» / А. Г. Абросимов // Исследования по информатике. – Казань, 2000. – Вып. 2. – С. 177–180.
2. Коллекции периодической печати 19 – начала 20 веков / А. Г. Абросимов, В. Ю. Кузьмина // Восьмая Международная Конференция и Выставка Libcom–2004 «Информационные технологии, компьютерные системы и издательская продукция для библиотек»: докл. и тез. докл. – М.: ГПНТБ России, 2004. – С. 5–7.
3. Метаданные описания коллекции периодической печати [Электронный ресурс] / А. Г. Абросимов // Электронные библиотеки: рос. науч. электронный журн. – 2005. – Т. 8, вып. 2. – Режим доступа: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2005/part2/Abrosimov, свободный.
4. Электронная коллекция казанских газет конца XIX – начала XX вв: проблемы создания коллекции / А. Г. Абросимов, В. Ю. Кузьмина // Девятая ежегодная конференция АДИТ–2005 «Культурное многообразие в едином информационном пространстве», 30 мая – 3 июня 2005 г.: тез. докл. – Казань, 2005. – С. 55–56. – Сведения доступны также по Интернет: http://www.adit.ru/rus/conference/adit2005/papers/paper.asp?nomer=50.
5. Электронная версия газеты «Казанские известия» (1811 – 1820) [Электронный ресурс] / А. Г. Абросимов, В. Ю. Кузьмина // Информационные технологии, компьютерные системы и издательская продукция для библиотек: материалы конф. «LIBCOM–2005». – Электрон. текстовые дан. – М.: ГПНТБ России, 2005. – 1 электрон. опт. диск (CD-ROM). – Загл. с этикетки диска. – ISBN 5-85638-100-9. – № гос. регистрации 0320501386. – Сведения доступны также по Интернет: http://www.gpntb.ru/libcom5/disk/doc/3.pdf.
6. Применение фасетной классификации для систематизации газетных публикаций в электронной коллекции казанской периодической печати 19-го – начала 20-го веков [Электронный ресурс] / А. Г. Абросимов, В. Ю. Кузьмина, И. И. Салосина // Электронные библиотеки: рос. науч. электронный журн. – 2006. – Т. 9, вып. 1. – (В печати).