БИБЛИОСФЕРА, 2008, № 4, с. 57–62

Обзоры

УДК 02 : 004 + 025.2
ББК 78.34 (2) + 78.36

ОБЗОР УНИВЕРСИТЕТСКИХ И БИБЛИОТЕЧНЫХ ПРОЕКТОВ ПО ОЦИФРОВКЕ КНИЖНЫХ ФОНДОВ БОЛЬШИХ И СРЕДНИХ РАЗМЕРОВ

© М.С. Угаров, А.В. Шабанов, 2008

Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук
630200, г. Новосибирск, ул. Восход, 15

Проведен сравнительный анализ основных целей современных проектов (2005–2008 гг.) по созданию цифровых библиотек больших и средних размеров, их отличительных особенностей и технологий.

Ключевые слова: цифровая библиотека, технологии оцифровки книжных фондов.

Тематика, связанная с созданием и функционированием цифровых библиотек, активно развивается уже более пятнадцати лет. Из новых обзоров на эту тему можно выделить [1] с прекрасным списком источников информации.

Среди организаций, занимающихся созданием цифровых библиотек большого объема, можно выделить следующие группы: крупнейшие государственные библиотеки и университеты, коммерческие организации, такие как Google и Microsoft, и некоммерческие организации, включающие в себя Open Content Alliance и Million Book Project (MBP). Несмотря на то что основной мотивацией этих участников является желание расширить доступ к научной литературе, их цели различаются в зависимости от их организационной миссии [1]. Количество цифровых библиотек среднего объема на 2008 г. весьма велико, функции и состав таких электронных хранилищ охватывают самые разные информационные области.

Лидером в создании цифровых библиотек является США. Соединенные Штаты Америки – одна из первых стран, где стали активно применяться цифровые технологии в библиотечной сфере деятельности [9]. Чуть позднее эти технологии вошли в практику работы крупнейших европейских библиотек, университетов и различных архивов, в последние годы наблюдаются аналогичные процессы в ряде российских организаций. Число же «любительских» инициатив в этой области почти необозримо, причем качество результата зачастую весьма велико, и позднее эти работы могут переходить в сферу интересов профессиональных информационных работников.

В отличие от других обзоров задача данной работы – сравнительный анализ основных целей современных проектов (2005–2008 гг.) по созданию цифровых библиотек больших и средних размеров, их отличительных особенностей и технологий, с помощью которых они добиваются этих целей.

Проекты научных, университетских и национальных библиотек

Ассоциация Research Libraries [2] – это некоммерческая организация, включающая в себя 123 учреждения, которыми являются университеты и научно-технические библиотеки, находящиеся на территории США и Канады, имеющие схожие исследовательские миссии, стремления и достижения. Ассоциация основана в 1932 г. и изначально включала в себя 42 учреждения, со временем этот список существенно расширился. Research Libraries стала первой в разработке совместных решений для проблем обнаружения, доступа, доставки и хранения информации. С точки зрения оцифровки своих фондов и создания на их базе цифровой библиотеки среди учреждений, входящих в ассоциацию, стоит выделить несколько проектов.

Совместный проект Университета штата Мичиган [7] и Google направлен на оцифровку всего фонда университетской библиотеки. Полученная в результате цифровая библиотека будет именоваться MBooks и будет доступна как через университетский библиотечный каталог Myrlin, так и через Google Book Search. Также будут доступны полные тексты научных работ, на которые не распространяется авторское право. Сообщается, что уже оцифрован 1 млн книг из 7,5 млн, хранящихся в университетской библиотеке.

Библиотека Гарвардского университета [3] и Google совместно работают над проектом по оцифровке большого количества книг Гарвардской библиотеки для того, чтобы сделать их доступны-

ми через Интернет. Проект будет приносить пользу студентам и ученым, где бы они не находились, значительно увеличивая доступ к фондам библиотеки, которая, являясь одной из крупнейших библиотек в мире, содержит более 15,8 млн томов. В ходе проекта будет отсканировано более 1 млн научных работ, на которые не распространяется авторское право, причем эти работы можно будет получить по сети в PDF формате. Так как фонд библиотеки размещен в 80 отдельных хранилищах, проект позволит получать оперативный доступ к книгам без их физического изъятия. Сам процесс оцифровки будет производиться сотрудниками компании Google совместно с гарвардскими библиотекарями, проект планируется завершить за несколько лет.

Оксфордский университет подписал соглашение с Google [4], согласно которому в течение трех следующих лет будет оцифрован 1 млн книг из фонда Библиотеки имени Бодлея. Все оцифрованные книги будут доступны через Интернет при помощи популярных поисковых служб Google и на сайте университета. Проект находится на самой ранней стадии, сотрудники библиотечной службы Оксфордского университета и персонал Google будут работать вместе, рассматривая более подробно различные детали проекта. Соглашение будет охватывать только те книги, которые были изданы до 1885 г. По ходу развития проекта круг источников может расшириться и включать в себя журналы, диссертации, государственные издания, материалы на иностранных языках, но пока еще нет определенного решения сделать это. Старопечатные книги, манускрипты, архивы и карты, которые могут быть повреждены в процессе сканирования, не будут оцифровываться. Сканирование будет производиться сотрудниками Google в помещениях университета.

Публичная библиотека Нью-Йорка [5] – это система, состоящая из 89 библиотек, уникальная в плане совмещения основных исследовательских лабораторий и разветвленной системы библиотек с одной общей структурой. Четыре основные научные библиотеки (Research Libraries) имеют фонд более чем на 3 тыс. языках и диалектах. Остальные 85 библиотек (Branch Libraries) расположены в трех районах Нью-Йорка и содержат находящиеся в обращении материалы. В 2004 г. Публичная библиотека Нью-Йорка совместно с Google запустила пилотный проект, в результате которого часть книг из фонда библиотеки станет доступна в сети в полнотекстовом виде. Это сотрудничество будет значительно расширять доступность фондов библиотеки и поддерживать миссию, суть которой заключается в предоставлении свободного и неограниченного доступа к информации и знаниям для любого человека.

Стэнфордский университет [6] также имеет аналогичный проект по оцифровке книг своей библиотеки, фонд которой содержит более 8 млн книг. Сотрудники университета и Google совместно работают над многими практическими аспектами проекта, особое внимание уделяется защите авторских прав. Этот проект напрямую вносит вклад в миссию библиотеки Стэнфордского университета, которая заключается в поддержке исследований и обучения в Стэнфорде и за его пределами путем предоставления небывалого доступа к своим фондам. Более того, сотрудники библиотеки получат опыт в обращении с действительно большим количеством цифрового материала и смогут разрабатывать инструментальные средства для работы с уже готовыми файлами, которые Google сделает доступными для публики как результат этого проекта.

Среди университетских и библиотечных инициатив по оцифровке стоит также отметить проекты средних размеров, например проект Британской библиотеки и электронную библиотеку «ОРЕЛ» Российской государственной библиотеки.

Британская библиотека приступила к переводу в цифровой формат более 100 тыс. старинных книг, ранее не доступных широкой публике [13]. В основном речь идет о книгах XIX в., которые потом больше не переиздавались, а их авторы зачастую оказались забыты. «Отсканировав всю коллекцию, мы предоставим свободный доступ к произведениям, минуя фильтр последующих оценок, которые основывались как на вкусовых, так и на экономических предпочтениях издателей» – заявил представитель Британской библиотеки Кристиан Йенсен. Кроме того, по его мнению, перевод книг в цифровой формат значительно облегчит работу преподавателей, которые не имеют практической возможности использовать их на своих курсах. Проект начат в 2007 г., ежедневно планируется сканировать около 50 тыс. страниц. Сообщается, что сканирование первых 25 млн страниц займет около двух лет. Для хранения всей коллекции в цифровом формате понадобится 30 терабайт.

Российская государственная библиотека [14] создала свою электронную библиотеку. Эта библиотека носит название «ОРЕЛ» (Открытая русская электронная библиотека). Фонд Открытой русской электронной библиотеки состоит из электронных документов, раскрывающих культурные богатства РГБ, и электронных версий наиболее значительных произведений мировой и русской литературы, заимствованных из Интернета. Фонд включает электронные копии книг, журналов, карт, нот, изобразительных материалов, диссертаций и авторефератов диссертаций (при разрешении авторов) по различным отраслям знаний. Документы находятся в свободном доступе через Ин-

тернет. В депозитарии хранятся художественная литература, периодика, научная литература по всем отраслям знаний, книжные памятники. Общее количество источников на данный момент – около 10 тыс. книг и 2,4 тыс. диссертаций.

Из новейших проектов можно выделить Президентскую электронную библиотеку имени Бориса Ельцина, которую планируется открыть в конце 2008 г. [15]. Она будет представлять собой уникальный общественно-информационный центр, связанный в режиме онлайн с библиотеками всей страны и крупнейшими книгохранилищами мира. Библиотека станет важным центром, стержнем, цементирующим единое гуманитарное пространство России, равно важным и для больших городов, и для маленьких сельских поселений. Ее первоочередной задачей станет рассказ о тысячелетней истории Российского государства и, разумеется, в значительной степени о новейшей российской государственности. Вероятно, ее основу составят фонды Российского государственного исторического архива.

Проекты Google

Проект Google Book Search [8] запущен 6 октября 2004 г. Цель проекта Google – оцифровка как можно большего числа книг со всего мира, причем как находящихся в свободном доступе, так и защищенных авторским правом. Совместно с несколькими крупными библиотеками Google работает над тем, чтобы включить собрания этих библиотек в Google Book Search и предоставлять пользователям информацию о книге и в некоторых случаях – несколько фрагментов из нее (предложений, содержащих текст запроса) как в библиотечном каталоге.

Также целью проекта является упрощение поиска книг – особенно тех, которые нельзя найти другим способом, например давно не переиздававшихся книг – при тщательном соблюдении прав авторов и издателей. Задачей проекта является работа с издательствами и библиотеками, направленная на создание всеобъемлющего виртуального каталога для поиска книг на всех языках, который поможет пользователям находить новые книги, а издателям – новых читателей.

Представители компании Google объявили о том, что количество книг, доступных в рамках проекта Google Books, превысило 1 млн экз.

Изначально участниками проекта были следующие университеты и библиотеки: Гарвардский университет, Университет штата Мичиган, Публичная библиотека Нью-Йорка, Оксфордский университет, Стэнфордский университет. Позже к этому списку присоединились: Баварская государственная библиотека, Библиотека Корнеллского университета, Национальная библиотека Каталонии, Принстонский университет, Университет Калифорнии, Университетская библиотека Лозанны и др. По состоянию на июнь 2008 г., к проекту присоединились в общей сложности 18 библиотек с целью отсканировать все или часть своих фондов и сделать их доступными по сети.

Большинство библиотек до сотрудничества с Google уже имели опыт оцифровки своих фондов. В качестве примера можно привести Университет штата Мичиган, который до сотрудничества с Google оцифровывал 5 тыс. томов в год, а также Библиотеки Корнеллского университета и Университета штата Висконсин (Мэдисон), которые за прошедшие пятнадцать лет оцифровали от двух до трех миллионов страниц, что равно примерно 7–10 тыс. книг. При такой производительности потребуется порядка несколько сотен лет для того, чтобы оцифровать все фонды полностью. Так как подобные мероприятия являются весьма дорогостоящими и трудозатратными, многие библиотеки поняли, что ускорить процесс оцифровки своих фондов можно с помощью сотрудничества с коммерческими организациями, такими как Microsoft и Google, тем самым поднимая планку производительности от миллионов страниц до миллионов книг. Таким образом, в ходе совместного проекта Университета штата Мичиган и Google в данный момент оцифровывается 30 тыс. томов в неделю, при таких темпах весь фонд, за исключением тех книг, которые не соответствуют требованиям по качеству (сильно ветхие книги и рукописи), будет преобразован в электронный формат за пять лет.

Также одна из миссий Google – организовать мировую информацию и сделать ее полезной и универсально доступной. Для этого Google Book Search предоставляет несколько различных способов сбора информации о книгах в Google и ссылок на них. Существует как стандартный формат ссылок, так и динамические ссылки. Остановимся подробнее на каждой разновидности ссылок.

Стандартный формат ссылок, или статические ссылки, позволяют разработчикам ссылаться на книги, используя номера ISBN, LCCN и OCLC. Ссылаться можно не только на саму книгу, но и на описание, титульный лист, страницу об авторском праве, оглавление, алфавитный указатель.

Динамические ссылки реализованы при помощи интерфейса программирования Book Viewability API. Данный интерфейс позволяет разработчикам:

- ссылаться на книги в Google Book Search, используя номера ISBN, LCCN и OCLC;
- знать о том, имеется ли заголовок книги в Google Book Search и какой возможностью просмотра этот заголовок обладает;
- создавать ссылки на страницу с информацией о книге.

Microsoft

В 2005 г. корпорация Microsoft запустила свой проект Live Search Books (LSB) [10] при поддержке Open Content Alliance. Целью LSB является создание базы данных полных текстов книг. В 2006 г. корпорация расширила свою деятельность в этой области путем привлечения дополнительных библиотек-партнеров и заключила контракт с Kirtas Technologies для участия в оцифровке материалов. На данный момент Microsoft включила в свою базу данных только те книги, на которые не распространяется авторское право (public domain), а также позволила научным учреждениям предоставлять электронные копии для совместного использования некоммерческим организациям до тех пор, пока эти организации согласны не делать эти файлы доступными для коммерческих поисковых интернет-служб. Также Microsoft предлагает программу Live Search Books Publisher Program для добавления контента через прямое сотрудничество с издателями.

Live Search, проводя различия между собой и проектом Google Book Search, акцентирует внимание на получении результатов поиска с помощью уникального интерфейса и предоставлении пользователю продвинутых инструментальных средств для поиска и получения данных. Все найденные книги имеют логотип Live Search.

В настоящее время проект находится в стадии бета-тестирования и работает в тестовом режиме.

Проекты Open Content Alliance

Основанный на сотрудничестве культурных, технологических, некоммерческих и правительственных организаций, Open Content Alliance (OCA) [12] был задуман в 2001 г. организацией Internet Archive и компанией Yahoo! Его целью является создание цифровых коллекций и предоставление доступа к ним через Internet Archive и The Open Library. OCA характеризует себя как проект, продвигаемый библиотеками. В отличие от инициатив Google и Microsoft, OCA фокусируется на создании «постоянного хранилища» (permanent archive) оцифрованных текстов на нескольких языках и мультимедийных файлов. Также OCA включит в свой архив коллекции, предоставленные следующими организациями: European Archive, Internet Archive, National Archives (UK), O'Relly Media, Prelinger Archives, Университетом Калифорнии и Университетом Торонто. Весь контент OCA является доступным для поиска через все основные машины поиска, за исключением контента, оцифрованного и предоставленного Microsoft Live Books. Физическим размещением файлов занимается Internet Archive, Microsoft и Библиотека Александрии (Library of Alexandria). Другие копии этих файлов хранятся во множестве разных систем хранения и, может быть, будут общедоступными в будущем. OCA сообщает, что хранение и обслуживание данных с помощью множества систем хранения позволит хранить файлы, тестировать работу механизма хранения и восстанавливать утраченные файлы. Проект частично финансируется Microsoft и Adobe.

Проект Университета Карнеги–Меллона Million Book Project

Проект Million Book Project (MBP) [11] возглавляет Университет Карнеги–Меллона. Отличительной чертой MPB является обширный план мероприятий по исследованию цифровых библиотек, включающий в себя исследования в таких областях, как: хранение и управление большим количеством информации; поисковые машины для многоязыковых данных; обработка изображений; оптическое распознавание символов для языков, не относящихся к романской группе, и т. д. Изначально при создании проекта Национальный научный фонд (National Science Foundation) выделил грант в размере 3 млн долл. для покупки оборудования и путешествий, затем MBP привлек международных партнеров и бюджет проекта превысил 100 млн долл. Официально финансирование проекта закончилось в июле 2007 г., но, несмотря на это, партнеры продолжают работать вместе. Начиная с 2001 г. было отсканировано 1,4 млн книг в Китае, Индии и Египте. Проект включает в себя 26 учреждений-партнеров, из которых некоторые содействуют в создании контента, другие – в проведении исследовательской работы в области цифровых библиотек. Internet Archive является одним из партнеров проекта и помогает приобретать книги для сканирования. Основные страны, которые предоставляют материалы для оцифровки (Индия, Китай и Египет), предпочитают размещать электронные копии отсканированных книг на своей территории. Они могут со временем сделать свой контент доступным для совместного использования с Internet Archive и Интерактивным компьютерным библиотечным центром (Online Computer Library Center), но в настоящее время четких планов в отношении этого нет.

В табл. 1 показана общая сводка по целям участников и их отличительные особенности.

В качестве примера в табл. 2 приведены сравнительные характеристики изображений, оцифрованных разными организациями, участвующими в создании цифровых библиотек.

Анализ программного обеспечения, используемого для создания и функционирования цифровых библиотек, – отдельная узкоспециальная тема,

Таблица 1

Основные цели и отличительные особенности проектов

Проекты и организации	Основные цели	Отличительные особенности
Научные библиотеки (Research Libraries)	• Поддержка в распространении знаний, • преобразование способов поиска и доступа к содержимому библиотек, • гарантия того, что материалы библиотек останутся доступными для будущих поколений, • использование цифровых копий в качестве резервных, • разработка продвинутых инструментальных средств для поиска и получения данных, эксперименты с интеллектуальным анализом текста (text mining)	• Сохранение своей, проверенной временем, руководящей роли в сборе, организации, управлении, хранении и предоставлении доступа к информации для поддержки изучения, обучения и исследований
Google Book Search	• Предоставление наиболее обширного списка книг на нескольких языках, • сделать более простым поиск и обнаружение релевантных книг с помощью машины поиска Google, • предлагая поисковую машину, которая осуществляет широкозахватный поиск (far-reaching search engine), привлечь как можно больше пользователей	• Оцифровка широко распространенных книг для того, чтобы их было проще искать
Microsoft Live Books	• Создание базы данных полных текстов, • сделать более простым нахождение релевантных книг, • получение результатов поиска с помощью уникального интерфейса и предоставление пользователю продвинутых инструментальных средств для поиска и получения данных	• Преобразование поиска по сети в информационный поиск путем создания надежного индекса аутентичного контента
Open Content Alliance	• Создание электронных коллекций с открытым доступом и предоставление доступа к ним через Internet Archive и Open Library, • поддержка разработки постоянного хранилища (permanent archive) оцифрованных текстов на многих языках и мультимедийного контента, • хранение и обслуживание файлов при помощи множества репозитариев	• Создание «постоянного хранилища» (permanent archive) научной информации, поиск по которому может быть осуществлен всеми основными машинами поиска
Million Book Project	• Предоставление пользователям быстрого и удобного доступа к ресурсам высокого качества путем оцифровки и обеспечения доступа к ним через Web, • обслуживание тестовой системы для стимулирования и поддержки исследований в таких областях, как: хранение и управление информацией; поисковые машины; обработка изображений и автоматический перевод	• Исследование круга вопросов, относящихся к поиску и управлению большими и многоязычными собраниями текстов в электронном виде

не входящая в сферу интересов настоящей публикации. Кратко можно отметить, что основная, наиболее сложная задача – обработка первичных данных в целях получения конечного (пользовательского) изображения, наиболее адекватного функциям реализуемой системы. Для создания клиент-серверных приложений чаще используются скриптовые языки программирования.

Таким образом, современное разнообразие различных технологий оцифровки позволяет создавать цифровые библиотеки разной тематики, назначения и объемов коллекций.

Список литературы

1. Preservation in the Age of Large-Scale Digitization: a white paper [Электронный ресурс]. – Режим доступа : http://www.clir.org/pubs/reports/pub141/pub141.pdf
2. Association of Research Libraries : About ARL [Электронный ресурс]. – Режим доступа : http://www.arl.org/arl/index.shtml
3. Harvard-Google Project [Электронный ресурс]. – Режим доступа : http://hul.harvard.edu/hgproject/index.html
4. Oxford University-Google Digitization Programme: FAQs [Электронный ресурс]. – Режим доступа : http://www.bodley.ox.ac.uk/google/

Таблица 2

Сравнительные характеристики данных и способа оцифровки

Проекты и организации	Разрешение	Формат данных	Аппаратура
Библиотека Университета штата Мичиган (цифровой контент предоставлен Google)	Большинство страниц имеют разрешение 600 dpi	TIFF Страницы с иллюстрациями хранятся в формате JPEG2000 с разрешением 300 dpi	Не разглашается
Библиотека Корнеллского университета (цифровой контент предоставлен Microsoft)	300–400 dpi	JPEG, рассматривается переход на JPEG2000	Сканнер Kirtas APT 2400 с автоматическим перелистыванием страниц или система Scribe, используемая OCA, требующая оператора для перелистывания страниц, и монитор для изображения
Open Content Alliance	400–600 dpi	JPEG2000	Scribe
Million Book Project	Преимущественно 600 dpi	TIFF	Различные, включая сканирование с помощью сканера Minolta PS7000

5. NYPL Partners with Google [Электронный ресурс]. – Режим доступа : http://www.nypl.org/press/2004/google.cfm
6. Stanford Google Library Project: FAQ [Электронный ресурс]. – Режим доступа: http://www-sul.stanford.edu/about_sulair/special_project/google_project_faq.html
7. Michigan Digitization Project [Электронный ресурс]. – Режим доступа : http://en.wikipedia.org/wiki/ Michigan_digitization_project
8. Google Book Search [Электронный ресурс]. – Режим доступа : http://books.google.com/
9. *Редькина Н.С.* Цифровые библиотеки: опыт США // Библиосфера. – 2008. – № 1. – С. 57–64.
10. Microsoft Live Search Books Publisher Program [Электронный ресурс]. – Режим доступа : http://publisher.live.com/
11. Million Book Project [Электронный ресурс]. – Режим доступа : http://www.library.cmu.edu/Libraries/MBP_FAQ.html
12. Open Content Alliance [Электронный ресурс]. – Режим доступа : http://www.opencontentalliance.org/faq.html
13. Проект Британской библиотеки [Электронный ресурс]. – Режим доступа: http://about-all.ucoz.ru/publ/70-1-0-552
14. Открытая русская электронная библиотека [Электронный ресурс]. – Режим доступа : http://orel.rsl.ru/
15. Библиотека имени Ельцина [Электронный ресурс]. – Режим доступа : http://www.pravoslavie.ru/news/080514145418

Материал поступил в редакцию 19.09.2008 г.

Сведения об авторах: *Угаров Михаил Станиславович – аспирант ГПНТБ СО РАН,*
тел.: (383) 266-75-79, e-mail: ugarovms@ spsl.nsc.ru
Шабанов Андрей Васильевич – кандидат технических наук, старший научный сотрудник отдела редких книг и рукописей,
тел.: (383) 266-10-91, e-mail: shabanov@spsl.nsc.ru