Технология подготовки электронных изданий для электронной библиотеки «Научное наследие России»

А. Н. Сотников, С. А. Кириллов
(МСЦ РАН)

Электронная библиотека «Научное наследие России» (ЭБ ННР) с организационной точки зрения представляет собой распределенную информационную систему, которая объединяет множество дистанционно удаленных участников, работающих на оборудовании различного типа, иногда по собственным технологиям, но по общим правилам, которые определены централизованно для всех участников проекта.

Базовыми компонентами такой информационной системы являются типовые центры сканирования и обработки, размещенные у держателей информационных фондов в различных городах России. Основная задача этих центров — обеспечивать полный цикл создания электронной книги с помощью локальных ресурсов и централизованных серверов обработки и хранения данных.

Архитектура типового центра сканирования и обработки предполагает наличие следующих технологических участков: участок подбора контента, участок описания контента, участок сканирования, участок обработки сканированных изображений, участок верстки электронных изданий.

Участок подбора контента

Первичный подбор книг для проекта ЭБ ННР осуществляют специалисты библиотек–участников проекта. Первым шагом выбора книг для оцифровки является анализ фонда библиотеки и подбор (определение по каталогам) публикаций — кандидатов для включения в электронную библиотеку. Сотрудники библиотек при этом руководствуются «Правилами по отбору информационных объектов для ввода в ЭБ ННР» и требованиями законодательства об охране авторского права.

Получив книги из библиотечного хранения, специалисты производят оценку соответствия содержания издания тема-

тике библиотеки, а также физического состояния книги, позволяющего выполнять её безопасное сканирование на имеющемся оборудовании.

Книги, соответствующие определенным выше критериям, поступают на участок описания контента.

Участок описания контента

Участок описания контента обеспечивает формирование и ввод набора метаданных, описывающих издание. В состав метаданных входят имена авторов, их биографии, год издания книги, издательство и др. Метаданные заносятся в специализированную базу данных, расположенную на технологическом сервере ЭБ ННР. (http://meta.e-heritage.ru). Представитель организации-держателя информационных фондов, наделенный правом входить в систему диспетчеризации ЭБ ННР, проверяет, не зарегистрирована ли уже в ней данная публикация. Если в системе публикация отсутствует, то необходимо ввести описывающий её набор метаданных. Процедура ввода метаданных в технологическую базу данных содержит два этапа:

- ввод биографической информации об авторах книг (при необходимости, если биография еще не создана участниками проекта);
- ввод метаданных по публикациям.

Остановимся на них подробнее.

Биографическая информация

В базу данных осуществляется ввод следующей информации об авторе: полные фамилия, имя, отчество, возможные псевдонимы, варианты написания имени и фамилии, необходимые пояснения относительно рода деятельности, годы жизни и пр. Составляется и вводится биография автора. Следует отметить, что вся информация, размещенная в ЭБ ННР, формируется вокруг персоны ученого. Другими словами, сначала мы определяем ученого, создаем биографическую справку и затем помещаем в ЭБ ННР его научные труды, архивные данные, сведения о связанных с ним музейных коллекциях. При этом основной особенностью проекта *является требование создавать развернутую биогра-*

фическую справку по всем, без исключения, авторам, произведения которых включены в ЭБ ННР.

Одновременно с вводом в БД биографических сведений ведется работа по поиску внешних источников информации по этим авторам. В частности, в ЭБ ННР вводятся ссылки на сайты музеев, различные электронные страницы в интернете, которые содержат значимую информацию о жизни и деятельности выдающихся учёных. Одно из полей базы данных биографической информации (поле «Архивная информация») обеспечивает взаимодействие проекта с Архивом РАН. В это поле Архив РАН вносит ссылки на личные фонды учёных, представленные на сайте АРАН (http://www.arran.ru)

Библиографическая информация

Формирование в технологической базе данных метаинформации об издании включает в себя: ввод основных элементов библиографического описания (заглавие, год издания, издательство и т.п., авторы приводятся в виде ссылок на метаданные соответствующих персон), индекса издания по ГРНТИ, перевода названия на русский язык (для изданий на иностранных языках), вида издания, дополнительных сведений (указания на возможные отличия от предыдущих изданий как в содержании, так и в названии книги, служебная информация).

Участок сканирования

На участке сканирования и предварительной обработки осуществляется перевод изображений страниц публикаций в цифровую форму. Печатная информация (например, страницы книги) переводится в цифровой вид с использованием планетарных, планшетных или иных типов сканеров, которые различаются разрешающей способностью, возможностью сканирования различного типа изображений и другими свойствами. В связи с этим при реализации проекта ЭБ ННР сформировались 5 основных типов комплексов сканирования:

- комплекс сканирования на основе планетарного сканера MINOLTA PS7000 и планшетного сканера EPSON GT 15000;

- комплекс сканирования на основе цветного планетарного сканера ПланСкан С-600;
- комплекс сканирования на основе цветного сканера Пауэрскан Д14000 А0-20/25;
- комплекс сканирования на основе цветного сканера ПланСкан А2-VC-B;
- комплекс сканирования на основе сканера микрофиш Kodak ABR 3000 DSV.

Приведем краткое описание и основные характеристики данных комплексов:

Комплекс сканирования на основе планетарного сканера MINOLTA PS7000 и планшетного сканера EPSON GT 15000A

Данный комплекс был определен в качестве базового на первом этапе становления проекта ЭБ ННР в 2005 году. В основе этого комплекса лежит принцип совместной работы черно-белого планетарного сканера и цветного планшетного сканера. Как правило, одного планшетного сканера достаточно для обслуживания 3–5 станций «Minolta PS 7000».

Оператор сканера Minolta PS7000 сканирует книги формата от А5 до А3 с разрешением 600 dpi. Большие книги (формата А2) сканируются с разрешением 400 dpi. Отсканировав книгу, оператор переходит к проверке и первичному редактированию файлов на соединенной со сканером рабочей станции. При этом он использует встроенные функции программы SRZ Proscan, такие как: автоматическое или ручное выравнивание страницы, автоматическое удаление мелких (1–3 пикселя) черных и белых точек, обрезку страницы по шаблону типичной страницы. Если в книге имеются страницы, которые не могут быть представлены в виде бинарного черно-белого изображения, например, иллюстрации, то список таких страниц передается оператору планшетного сканера.

Рисунки, карты и страницы, содержащие цветные или полутоновые изображения, в соответствии с технологическими требованиями, оцифровываются на планшетном сканере с разрешением 600 dpi. После того, как текстовые страницы и иллюстрации книг отсканированы, результирующие файлы собираются в папки с названием присвоенных книгам ID и передаются на следующий этап технологиче-

ской цепи — участок обработки отсканированных изображений.

Состав оборудования типового комплекса: сканеры MINOLTA PS7000 — 5 шт., компьютеры (рабочие станции) управления планетарным сканером — 5 шт., сканер EPSON GT 15000A — 1 шт., компьютеры (рабочие станции) управления планшетным сканером — 1 шт.

Базовое программное обеспечение: ПО SRZ Proscan, Adobe Photoshop.

Комплекс сканирования на основе цветного планетарного сканера ПланСкан С2-ЦА-600

Этот комплекс сканирования пришел на смену устаревшему комплексу на основе сканера Minolta PS7000. Оператор сканера ПланСкан С2-ЦА-600 имеет возможность сканировать любые типы публикаций до формата А2 (в цветном, сером и черно-белом режиме). Книги формата А5-А3 сканируются с разрешением 600 dpi. Издания формата А2 сканируются с разрешением 400 dpi. Встроенная автоматическая книжная колыбель позволяет сканировать оригиналы толщиной до 10 см и весом до 6 кг. Имеется прижимное стекло с системой автоматического подъема.

Состав оборудования комплекса: сканер ПланСкан С2-ЦА-600, компьютер (рабочая станция) управления сканером.

Базовое программное обеспечение: встроенное в сканер управляющее ПО с функциями коррекции изображения.

Внедрение данной модели сканера позволило улучшить качество сканирования и снизить трудоемкость выполнения работ по получению сканов с различного вида печатных изданий. В настоящее время этими сканерами оснащены участки сканирования МСЦ РАН, БЕН РАН, АРАН, ИЭА РАН, ИНИОН РАН в Москве, а также центры сканирования в Санкт-Петербурге.

Комплекс планетарного цветного сканирования ПланСкан А2-VC-B

Комплекс планетарного цветного сканирования ПланСкан А2-VC-B по характеристикам аналогичен ПланСкан С2-ЦА, однако он более компактен, оборудован встроенной светодиодной системой подсветки.

Комплекс планетарного цветного сканирования ПланСкан А2-VC-B позволяет сканировать оригиналы до формата А2+ с разрешением до 600 dpi, режимы сканирования: черно-белый — 2 бит, серый — 8 бит, цветной — 24 (30) bit/pix; имеет встроенную V–образную книжную колыбель с возможностью раскрытия до 180 градусов, сенсорную русскоязычную панель управления с графическим меню. Может использоваться без управляющего компьютера с просмотром отсканированного изображения на встроенном мониторе.

Состав оборудования комплекса: сканер ПланСкан С2-ЦА, компьютер (рабочая станция) управления сканером.

Базовое программное обеспечение: встроенное в сканер управляющее ПО с функциями коррекции изображения.

Комплекс сканирования на основе сканера Пауэрскан Д14000 А0-20/25

Основной особенностью данного комплекса является возможность сканировать материалы особо большого формата (вплоть до размера А0) с высокой разрешающей способностью, например, газеты, карты, плакаты, картины.

Комплекс Пауэрскан Д14000 А0-20/25 с режимами сканирования «цветной», «серый», «черно-белый» и максимальным форматом сканирования до А0 позволяет сканировать печатные материалы больших форматов и особо ветхие издания. Для этого комплекс имеет эргономичный стол сканирования, книжную колыбель для книг до 20 см толщиной и весом до 25 кг., а также систему холодной подсветки и сканирующую головку CCD 3x14000 pix.

Состав оборудования комплекса: сканер Пауэрскан Д14000 А0-20/25 , компьютер (рабочая станция) управления сканером.

Базовое программное обеспечение: ПО управления процессом сканирования и управления сканером, включая модуль построения ICC-профилей для калибровки цветового пространства сканера.

Комплекс сканирования на основе сканера микрофиш Kodak ABR 3000 DSV

Данный комплекс служит для оцифровки материалов, хранящихся на микрофишах и микропленке.

Сканер Kodak ABR 3000 DSV имеет объектив «8731408» с увеличением 7,5х; объектив «1677293» с увеличение 13–27х; ручную универсальную каретку «3846201» UC-5 для микрофиш, микропленки 16/35 мм и апертурных карт с ручной загрузкой и позиционированием микрофильмов.

Состав оборудования комплекса: сканер Kodak ABR 3000 DSV, компьютер (рабочая станция) управления сканером.

Базовое программное обеспечение: ПО PowerFilm V5.3 для сканирования и управления сканером.

Участок обработки сканированных изображений

Полученные после сканирования цифровые изображения нуждаются в дополнительной обработке для повышения качества отображения страниц книги, устранения дефектов изображения и преобразования формата данных в наиболее удобный для дальнейшей работы вид.

Обработка сканированных изображений включают в себя три технологических этапа: автоматическая обработка сканов, проверка результатов автоматической обработки и ручная коррекция. Остановимся более подробно на каждом из них:

Этап автоматической обработки. На данном этапе происходит автоматическая обработка сканов на сервере пакетной обработки страниц с установленным ПО BookRestorer. Задача этого этапа — исправить типичные дефекты сканирования книги до приемлемого уровня.

Получив файл с отсканированным изображением, ПО BookRestorer обрабатывает его с помощью скрипта, содержащего в себе набор макрокоманд, которые последовательно выпрямляют строки, абзацы и страницу в целом; удаляют следы пыли и не слишком крупные пятна; поворачивают страницу, обрезают поля. После этого обработанный файл перемещается в отдельную директорию.

Такой способ позволяет максимально эффективно задействовать доступные вычислительные ресурсы, в том числе и в ночное время, и требует лишь периодического присутствия оператора.

На этом этап автоматической обработки отсканированных образов завершается, и начинается их проверка оператором в ручном режиме. Это связано с тем, что на сего-

дняшний день существующие программные алгоритмы обработки отсканированных образов не позволяют полностью автоматизировать этот участок работы.

Этап проверки результатов автоматической обработки. На данном этапе выполняется проверка электронных образов страниц, прошедших автоматическую обработку. Этот этап включает в себя:

- проверку последовательного отображения страниц (нумерация страниц должна быть последовательной) поиск пропущенных при сканировании страниц;
- проверку качества электронной страницы (степень читаемости текста, не менее 99 % информации, представленной на странице, должно быть читаемо);
- проверку качества автоматической обработки отсканированных страниц (корректная обрезка страниц, геометрическая коррекция текста, изгибов текста и иных искажений);
- дополнительная правка страниц (обрезка, удаление посторонних элементов).

Этап ручной коррекции электронных образов страниц. На этом этапе, при необходимости, производится ручная обработка страниц в программе Adobe PhotoShop. Этот этап предусмотрен для наиболее сложных страниц, содержащих многочисленные формулы, таблицы, иллюстрации и т.п. Он включает в себя:

- обрезку страниц;
- геометрическую коррекцию текста, изгибов текста и иных искажений;
- удаление посторонних элементов на страницах (следы пальцев оператора, полосы, тени и др.);
- цветовую коррекцию, коррекцию яркости и контрастности.

Состав оборудования участка: рабочая станция графического редактора, сервер пакетной обработки страниц с установленным ПО BookRestorer.

Базовое программное обеспечение: BookRestorer, Adobe PhotoShop.

Участок верстки электронных изданий

На этом участке из отдельных изображений страниц осуществляется верстка электронной книги в режиме on-line на технологическом сервере верстки электронного издания http://media.e-heritage.ru.

Обработанные и отредактированные по формату страницы экспортируются в программу загрузки изображений и составления электронного оглавления NNMetaDis. При необходимости, оператор производит распознавание страниц оглавления книги, заносит и редактирует информацию о самом документе и его структуре в виде электронного оглавления документа. По завершении верстки выполняется предварительная публикация издания на технологическом сервере.

На участке выполняются следующие работы:

- формирование структуры графических образов страниц;
- распознавание страниц оглавления книги;
- формирование навигационной системы (оглавления);
- публикация электронной книги на технологическом сервере.

Состав оборудования участка: рабочая станция графического редактора — 1 шт.

Базовое программное обеспечение: NNMetaDis, Adobe PhotoShop.

Участок контроля и координации наполнения ЭБ ННР

Поскольку ЭБ ННР представляет собой распределенную систему с множеством удаленных участников, возникает необходимость формирования структуры, которая организует и контролирует весь процесс создания электронных книг: от подачи заявок на сканирование до публикации электронной книги на Интернет-портале и отправки файлов обработанных книг на архивное хранение. Эту роль в настоящее время выполняет Центр контроля и координации ЭБ ННР. Центр аккумулирует все заявки на сканирование в единой базе данных и производит проверку поступивших заявок по различным параметрам. Координация работы

участников проекта, в основном, происходит методом текущего контроля поставляемых электронных публикаций. Каждая электронная книга перед публикацией на сайт проходит проверку выпускающего редактора участка контроля. Проверяется качество сканирования книги, наличие страниц, порядок размещения страниц, правильность и полнота метаданных и т. д. В конце работы выпускающий редактор принимает решение о публикации книги на сайте или об информировании поставщика о необходимости исправления допущенных ошибок (с их указанием).

Участок контроля и координации состоит из трех редакторских групп:

Редакторская группа проверки верстки электронной книги на сервере http://media.e-heritage.ru

Работа этой группы построена как на анализе графических образов страниц, так и на проверке системы навигации.

Проверка графических образов страниц включает:

- проверку последовательного отображения страниц;
- проверку качества сканирования (не менее 99 % информации, представленной на странице, должно быть читаемо);
- проверку качества обработки отсканированных страниц (корректная обрезка страниц, геометрическая коррекция текста, изгибов текста и иных искажений);
- проверку на отсутствие посторонних элементов на страницах (следы пальцев оператора, полосы, тени и т.п.).

Проверка системы навигации (ссылок) в электронных публикациях включает:

- проверку ссылок на предмет их открытия;
- проверку ссылок на соответствие главам и содержанию;
- проверку ссылок на правильность написания (проверка орфографии: прописных и строчных букв; проверка пунктуации, пробелов).

Редакторская группа проверки библиографических данных на технологическом сервере http://meta.e-heritage.ru

Работа этой группы состоит в проверке правильности ввода метаданных и включает:

- проверку соответствия изданию полей записей БД: фамилия автора, источник, заголовок, перевод названия на русский язык, выходные данные издания;
- проверку оформления метаданных — их орфографию, пунктуацию и формальные признаки (единообразие в сокращениях, наличиях пробелов и т.п.)
- проверку соответствия оригиналу сведений, внесенных в поля «вид издания», «язык», «страницы». Поле «страницы» выверяется строго по электронной версии книги и включает в себя общее количество файлов в электронной версии, подготовленной к загрузке на сайт, проверку наличия и правильного оформления индекса ГРНТИ, его соответствие тематике публикации;
- проверку оформления библиографического описания (согласно ГОСТ 7.1–2003)

Редакторская группа проверки качества представления биографических данных (данных по персоналиям) на технологическом сервере http://meta.e-heritage.ru

Работа этой группы заключается в проверке правильности ввода биографических данных и включает:

- проверку наличия биографии автора;
- выверку и редактирование биографий;
- полная вычитка биографии;
- проверка и редактирование (при необходимости) основных данных;
- устранение орфографических ошибок, ошибок пунктуации, повторов;
- приведение к единому стилю формальных элементов текста;
- удаление лишних пробелов или вставка недостающих;
- единообразие в использовании кавычек и тире;
- единообразие в сокращениях;

- проверку орфографии в названиях публикаций и журналов на иностранных языках, упоминаемых в биографиях;
- внесение недостающих слов, а также исправление стилистических неточностей и погрешностей в тексте;
- выверку библиографии в тексте биографии.

Завершив все стадии проверки, выпускающий редактор осуществляет публикацию — загрузку книги на Интернет-портал электронной библиотеки «Научное наследие России».

Заключение

Выполнить полный цикл создания электронной книги — идеальная задача, к которой должны стремиться все участники проекта. Однако часто возникает ситуация, когда участники располагают фондами, необходимыми для электронной библиотеки, но по разным причинам, например, из-за отсутствия оборудования, сотрудников и т.п. не могут выполнить полный цикл работ по созданию электронной публикации. В таком случае предусмотрен сокращенный цикл, по которому выполняется только часть работ, а завершение работ передается другим участникам проекта. Например, организация поставляет только биографии учёных и ссылки на внешние или внутренние источники (Архив РАН); организация осуществляет полный ввод метаданных, сканирование и графическую обработку публикации (ЦНБ УрО РАН); организация осуществляет полный ввод метаданных и сканирование публикации (Дарвиновский музей Москвы) и т. д. Первоначально завершение всех работ централизованно принимал на себя Отдел информационных ресурсов и систем МСЦ РАН, который и сейчас контролирует весь процесс создания электронных книг от подачи заявок на сканирование до публикации электронной книги на Интернет-портале. Но в ходе развития проекта возникают различные формы межведомственного взаимодействия. Например, Государственный геологический музей им. В. И. Вернадского РАН осуществляет полный ввод метаданных и передает книги в центр сканирования и обработки БЕН РАН, который завершает цикл работ по созданию электронной

публикации. Другие формы горизонтальной кооперации реализованы в Санкт-Петербурге.

В заключение необходимо отметить, что организация полного цикла производства электронных книг не представляет сложности для наших научных учреждений.

В любой библиотеке или другой научной организации имеются сотрудники, которые могут, пользуясь внутренним каталогом, подобрать литературу, составить биографию автора и ввести в электронный каталог сведения о публикации.

Организация участков сканирования и обработки — это новое дело для некоторых библиотек, однако эпоха ксероксов закончилась, и перевод книг в цифровой вид (с учетом авторского законодательства) уже вчера стал актуальной необходимостью. Современные системы сканирования максимально дружественны пользователю, а наработанные технологии электронной библиотеки «Научное наследие России» превращают книжный сканер и компьютер обработки в аналог промышленного станка, с которым наилучшим образом справляются даже сотрудники со средним образованием.

Организация участка верстки электронного издания в формате электронной библиотеки, хотя и является новым видом работ для традиционной библиотеки, также не представляет особой сложности и после небольшого курса обучения сотрудников основам верстки для этого достаточно одного рабочего дня. Впрочем, и эту работу можно представить как традиционную сферу библиотечной деятельности — новый способ представления издания читателю.

В заключение отметим, что электронная библиотека «Научное наследие России» является открытой для всех организаций, заинтересованных в популяризации и сохранении отечественного научного наследия.