ОЦИФРОВКА КНИГ КАК СПОСОБ СОХРАНЕНИЯ ЛИТЕРАТУРНОГО НАСЛЕДИЯ

П. В. Гляков,
заведующий кафедрой информационных технологий в культуре БГУКИ, кандидат физико-математических наук, доцент

Проблема сохранения мирового культурного наследия особенно актуальной стала после предложения ЮНЕСКО в 1992 г. программы «Память мира», провозгласившей необходимость принятия неотложных действий по предотвращению разрушения документального наследия мира. Затем в 1995 г., Генеральной конференцией ЮНЕСКО было принято «Общее руководство по сохранению наследия», которое инициировало крупномасштабные проекты по сохранению литературного наследия [2].

Основу работ по сохранению литературного наследия составляет оцифровка книг, представляющая собой процесс перевода бумажных книг в электронный вид. Электронные копии книг могут образовывать электронные библиотеки и распространяться в Сети. Электронные книги можно легко распространять, воспроизводить и читать на экране [4].

Электронные книги оформляются в виде, свойственном печатным книгам. Так, в электронных книгах обычно текст делят на нумерованные страницы одинакового размера; их типографика соответствует уровню печатных изданий. Среди электронных книг различают верстанные и сканированные.

Верстанные книги – материал, подготовленный авторами в издательской системе либо распознанная и вручную вычитанная и отформатированная бумажная книга. Исходный файл представлен в формате наглядного текстового процессора или на языке разметки LaTeX. Результатом является электронная книга в формате PDF, реже ПостСкрипт или DVI.

Такие файлы обычно содержат векторные шрифты и иллюстрации высокого качества, поэтому они пригодны для печати в любом разрешении, для просмотра на экране и для поиска по тексту книги, включая возможность выделять и копировать фрагменты текста и иллюстрации. Файлы этого вида кратко называют векторными. Типичные векторные PDF-файлы имеют размеры от 3 до 10–15 килобайт на страницу, в зависимости от числа формул и иллюстраций [5].

Сканированные книги – это файлы, хранящие целые электронные изображения каждой страницы книги. Такие файлы

создаются путем сканирования бумажной книги постранично и дальнейшей обработки с целью улучшения качества и уменьшения размеров файла. Поскольку каждая страница хранится в виде ряда точек (растра), то такие книги можно кратко называть растровыми, чтобы отличать их от векторных.

Основные форматы, употребляющиеся для растровых файлов, – PDF и DJVU. В этих форматах можно добавлять также распознанный текст, закладки и гиперссылки, чтобы были возможны быстрые переходы по книге и автоматический поиск текста. Поэтому качественно сделанные растровые книги не менее удобны в использовании, чем векторные, и несущественно проигрывают им в качестве распечатанного текста. Типичный размер растровой книги – от 5 до 10–15 килобайт на страницу, в зависимости от разрешения и качества текста или иллюстраций.

Сканирование изображений может происходить вручную или автоматически. В обычных сканерах книга располагается на стекле, на книгу падает свет, и оптический механизм сканирует книгу, двигаясь под стеклом. Другие книжные сканеры используют V-образную раму и фотографируют страницы сверху.

Страницы могут переворачиваться вручную или с помощью автоматических устройств подачи бумаги. После сканирования программа корректирует изображение документа, выравнивая его, обрезая, редактируя и преобразовывая его в текст и окончательную форму электронной книги. Обычно отсканированное изображение требует визуальной проверки и устранения ошибок. Сканирование 300 точек на дюйм является нормой для преобразования в цифровой вид текста, однако для редких и сложных книг необходимо использование более высокого разрешения.

Для оцифровки могут использоваться три подхода: обязательный, опциональный и смешанный. При обязательном подходе получают копии страниц в виде графических (обычно растровых) изображений. Он осуществляется путем сканирования или фотографирования с последующей обработкой и сохранением в одном из форматов графических файлов. В этом случае полностью сохраняется оригинальная верстка книги, и исключаются какие-либо ошибки, однако невозможен поиск или извлечение фрагментов текста, например для целей цитирования.

При опциональном подходе распознают текст (иначе его называют оптическим распознаванием символов – OCR), а затем сохраняют распознанный текст в одном из форматов электронных книг. В этом случае становится возможен полнотекстовый поиск по книге и индексация больших массивов электронных книг, однако

при этом затрудняется воспроизведение оригинальной верстки, изображений, схем и формул, практически неизбежными становятся ошибки распознавания.

В последнее время все чаще применяется смешанный подход: текст книги распознается в автоматическом режиме и подкладывается под оригинальные растровые изображения страниц, что позволяет совместить преимущества обоих подходов.

Для оцифровки книг используют следующие типы сканеров: планшетные, планетарные и роботизированные. Планшетные сканеры ориентированы на домашнего пользователя, но сконструированы так, чтобы облегчить процесс сканирования книг. Планетарные – это профессиональные высокопроизводительные сканеры. Называются они так из-за расположения камеры, как спутника над планетой, которой является сканируемый оригинал. Роботизированные сканеры являются промышленными сверхвысокопроизводительными сканерами, оборудованными устройствами различных конструкций для автоматического переворачивания страниц.

Планетарные и роботизированные сканеры позволяют достичь производительности 500–2000 страниц в час, у лучших моделей производительность достигает 2500–3000 страниц в час. Эти сканеры используют цифровую камеру и источники света по обе стороны от камеры, что обеспечивает легкий доступ к книге.

Широкое распространение получила технология сканирования без вмешательства человека. Она основана на использовании цифровой фотокамеры и обеспечивает оцифровку как сшитых, так и расшитых изданий. Подходит для оцифровки как относительно новых, так и ветхих изданий за счет специальной V-образной колыбели, позволяющей не раскрывать книгу на 180 градусов, что сводит к минимуму вредное воздействие на издание. Книга остается в одной и той же позиции. Скорость сканирования в цветном режиме около 500–700 страниц в час. Перелистывание страниц происходит вручную.

В настоящее время есть модели сканеров с автоматическим перелистыванием, однако ценные, ветхие книги, составляющие основу библиотечного фонда, не рекомендуется оцифровывать на таком оборудовании во избежание повреждений. Разрешение получаемых изображений 130–470 dpi.

В ходе масштабных проектов по оцифровке книг, как правило, обрабатываются книги, перешедшие в общественное достояние. Среди наиболее крупных проектов по оцифровке на сегодня можно выделить два: Google Books и проект Open Library.

На конец 2010 г. в базе Google Books находилось более 15 миллионов книг, из них около миллиона – в общественном достоянии. Согласно подсчету Google, в мире издано около 130 миллионов уникальных книг, не считая переизданий. Google заявил, что отсканирует их все к концу десятилетия [1].

Проект Open Library ведет некоммерческая организация Архив Интернета. Эта организация осуществляет бесплатный доступ к своим базам данных для широкой публики. Декларируемой целью Архива Интернета является сохранение культурно-исторических ценностей цивилизации в эпоху интернет-технологий, создание и поддержка электронной библиотеки.

В настоящее время библиотека Архива Интернета содержит в открытом доступе более 2 миллионов книг, а в каталог библиотеки занесено больше 25 миллионов изданий. Коллекция Архива Интернета постоянно растет, так как библиотека сканирует около 1000 книг в день [3].

В заключение отметим, что оцифровка книг высвобождают текст из оков материальности, а Интернет и другие формы электронной коммуникации снимают пространственные ограничения на распространение информации. Цифровые медиа стали принципиально новой формой существования знания [1].

Книга как материальный объект все больше замещается носителями книжного содержания в абстрактной форме: она становится эфемерной, материальны лишь устройства для чтения.

Традиционные каналы, обеспечивающие доступ к книгам, – книжные магазины и библиотеки – вступают в неоднозначные отношения от конкуренции до конвергенции со своими электронными аналогами.

Электронное чтение приобретает все большую популярность. Увидеть сегодня человека, читающего книгу или журнал с экрана планшета, ридера или коммуникатора стало делом не менее привычным, чем человека с печатным изданием в руках.

---

1. Горный, Е. Проблемы сохранения культурного наследия в эпоху цифрового текста / Е. Горный [Электронный ресурс]. – Режим доступа: www. netslova.ru/gorny/digtext.html?2012.

2. К вопросу сохранения книжного наследия [Электронный ресурс]. – Режим доступа: http://rumchten.rsl.ru/assets/files/2008/doc/1200572469.doc.

3. Компьютерные Вести On-line. Оцифровка книг: общественный проект [Электронный ресурс]. – Режим доступа: http://old.kv.by/ index2005440602.htm.

4. Оцифровка книг – Википедия [Электронный ресурс]. – Режим доступа: http://ru.wikipedia.org/wiki/Оцифровка_книг.

5. Оцифровка печатных текстов – Викиучебник [Электронный ресурс]. – Режим доступа: http://ru.wikibooks.org/wiki/.