Комбинированное электронное представление печатных изданий

© С.И. Трифонов

Научная педагогическая Библиотека им. К.Д. Ушинского РАО
Trf@ya.ru

Аннотация

Комбинированное электронное представление позволяет читателю пользоваться как графическим изображением страниц печатного издания, так и полным массивом текстовой информации. В статье ставятся и обсуждаются различные вопросы поддержки этого представления в рамках электронной библиотеки. Предлагаемые решения применяются как в технологии подготовки информации, так и в архитектуре сервера электронной библиотеки.

1. Введение

В современных электронных библиотеках, как правило, используется либо текстовое, либо графическое представление изданий [1]. В общих словах, комбинированное представление издания строится, как объединение этих двух основных представлений, а также включает информацию для их синхронизации. В результате такой комбинации, пользователь библиотеки имеет возможность использования достоверного графического представления с одной стороны, и все возможности для полнотекстового поиска — с другой.

Формирование электронной библиотеки на основе комбинированного представления не является сегодня распространённым. Единственным широко известным проектом данного класса является электронная библиотека компании Google [2]. Вызвано это определённой сложностью технологии подготовки комбинированного представления, нестандартными задачами автоматизированной поддержки форматов.

В статье представлены решения по данной проблеме, разработанные коллективом Лаборатории Разработки и Внедрения Информационных Технологий Научной Педагогической Библиотеки им. К.Д.Ушинского.

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

2. Функциональное описание комбинированного представления

Комбинированное представление издания строится на основе двух представлений различной природы: графического и текстового[1].

Графическое представление издания (будем также употреблять термин **графический слой**) — это комплект изображений страниц в каком-либо графическом формате, в нашем случае это JPG. Графическое представление даёт достоверную, но не формализованную информацию об издании, непригодную к использованию в задачах информационного поиска.

Для текстового представления (**текстового слоя**) издания мы используем формат HTML, причём, с большим числом различных ограничений. Ограничения отсекают избыточные свойства общего формата до функциональности, достаточной для воспроизведения печатной продукции. Автоматический контроль ограничений заметно улучшает качество разметки текстового слоя. Учитывая то, что полное представление издания в библиотеке содержит графический слой вместе с текстовым, разметка используется с функциональными целями, для выделения логической структуры текста.

Требования к комбинированному представлению издания определяются возможностями, которые web-сервер электронной библиотеки предоставляет конечному пользователю библиотеки в отношении данного издания

Основными режимами работы пользователя с изданиями являются два: это режимы просмотра и поиска.

[1] Здесь стоит отметить, что для электронного представления единичных изданий прекрасно подходит формат PDF, обеспечивающий в определённом смысле все три представления: текстовое, графическое и комбинированное. Тем не менее, использование PDF в электронных библиотеках, предоставляющих сервис полнотекстового поиска по всей коллекции изданий, порождает слишком много технических проблем. Подробнее см. [1].

В режиме просмотра пользователь читает текст издания. Как правило, это удобнее всего делать, используя графический слой. Тем не менее, в отдельных случаях пользователю удобно переключиться в просмотр текстового слоя, например для копирования фрагмента текста. Соответственно, требованием к комбинированному представлению является *строгое соответствие сегментов текстового слоя страницам издания.*

В режиме поиска пользователь имеет возможность найти документы по различным характеристикам: атрибутам и полным текстам. Для реализации этой возможности необходимо текстовое представление издания. Чтобы воспользоваться результатами поиска, пользователю требуется войти в специальный режим.

В режиме просмотра результатов поиска пользователь видит то же самое издание, что и при обычном просмотре. Но с крайне существенным для нас отличием: найденные в результате полнотекстового поиска слова должны быть выделены визуально, мы будем говорить **подсвечены**, и по этим подсветкам должна быть реализована навигация (передвижения вперёд/ назад). Для этого средствами браузера на изображение страницы накладываются цветные полупрозрачные прямоугольники.

Для организации этой функциональности требуется соответствующая информация, а именно соответствие между словами, расположенными на страницах и прямоугольниками, расположенными на графических изображениях страниц. Будем называть это соответствие **графическим индексом страницы.**

Таким образом, мы приходим к самому существенному требованию к комбинированному представлению: оно *должно содержать графические индексы*.

Суммируем требования к комбинированному представлению издания. Такое представление должно содержать:

- текстовый слой; в его разметку, в частности, должно входить разделение на страницы),
- графический слой,
- графические индексы страниц.

Сформулируем также два существенных свойства графического индекса. Часто слова в изданиях набраны с переносами — соответственно, графический индекс должен поддерживать *соответствие одного слова не одному, а вообще говоря нескольким прямоугольникам* [2] . При реализации подсветки размеры прямоугольников слегка увеличиваются для удобства восприятия. Соответственно, в графическом индексе *высокая точность для координат прямоугольников не требуется*. Точность в 2-4 пикселя оказывается достаточной для всех практических задач.

3. Подготовка графических индексов

Для подготовки информации в виде графических индексов, нами разработан специализированный текстовый формат. Его описание приведено ниже.

Логика использования графических индексов в нашей технологии вписывается в следующую последовательность операций:

- страницы печатного издания сканируются и обрабатываются; так получается графический слой (заметим, что устранение перекосов на изображениях играет в последствии существенную роль при визуализации подсветок),
- в программе, где происходит распознавание текста по изображениям и первичная вычитка (в нашем случае ABBY FineReader), выполняется экспорт результата в форматы PDF и HTML,
- результат в формате PDF преобразуется в комплект файлов графического индекса,
- параллельно результат распознавания в формате HTML подвергается разметке и, возможно, дополнительной вычитке; так получается текстовое представление,
- созданные в результате предыдущих операций файлы проходят валидацию и совместное редактирование в специализированной программе H2 — внутренней разработке Лаборатории,
- в случае отсутствия ошибок, по общему комплекту накопленной информации происходит генерация так называемого **серверного пакета** - комплекта файлов, предназначенного для эксплуатации издания на сервере библиотеки.

Валидация в програме H2 [3] представляет собой комплексную финальную фазу подготовки издания. В отношении графических индексов, она решает следующие задачи:

- пословное соответствие текстов, включённых в текстовое представление и в графические индексы,
- отработка автоматических контрольных пометок в графических индексах: исправление проблем либо подтверждение правильности,
- визуальная (в перспективе — автоматизированная) валидация соответствия графических индексов изображениям страниц.

Практика показывает, что программа H2 позволяет решать эти задачи вполне эффективно.

[2] Кроме обычных переносов слов между строками, в изданиях нередко встречаются переносы слов между страницами. Аккуратная отработка этого эффекта приводит к серии трудностей, при чём не только технического характера. В разметке текстового слоя для этого нужно регламентировать разрыв страницы внутри слова. В нашей технологии мы отрабатываем эти переносы корректно.

4. Особенности архитектуры сервера библиотеки

Как правило, поддержка полнотекстовых поисковых возможностей сервера реализуется через автономный специализированный модуль полнотекстового поиска (в нашем случае — Яндекс Сервер), и этот модуль работает в терминах слов и словопозиций текстовых документов. Однако, в нашем контексте нас интересуют подсветки в терминах прямоугольников на изображениях страниц. Это требование нестандартно, и приводит к необходимости выделения в архитектуре сервера специального модуля — **сервера подсветок**.

Модуль сервера подсветок предназначен для решения задачи подсветки на графическом слое. Для этого он взаимодействует с модулем полнотекстового поиска, извлекая словопозиции найденных слов, и вычисляет соответствие этих словопозиций наборам прямоугольников в соответствии с информацией графических индексов страниц издания.

Заметим, что непосредственно текстовые файлы графического индекса на сервере не используются. Вместо них в серверный пакет входит двоичный файл, содержащий информацию из графических индексов в виде, оптимизированном под быстрое решение задач сервера подсветок.

5. Формат файла графического индекса

Пример файла графического индекса:

```
 W  177  386  217  587 |АКАДЕМИЯ
 W  177  600  217  953 |ПЕДАГОГИЧЕСКИХ
_W  177  965  217 1066 |НАУК
_W  177 1079  217 1180 |СССР

 W  477  378  579  440 |Л
 .  477  440  579  464 |.
_W  477  489  579  551 |С
 .  477  551  579  574 |.
 W  477  598  579 1186 |ВЫГОТСКИЙ

 W  643  541  739 1023 |СОБРАНИЕ

 W  754  500  851 1069 |СОЧИНЕНИЙ

 W  869  717  929  840 |ТОМ

 W  937  643  994  909 |ПЕРВЫЙ

 W 1974  654 2034  880 |МОСКВА

 . 2040  563 2102  582 |‘
 W 2040  582 2102  974 |ПЕДАГОГИКА
 . 2040  974 2102  994 |’

 W 2117  711 2171  831 |1982
```

Формат разработан специально для комбинированного представления данных. Файл в этом формате - текстовый, в нём в понятной, но обширной форме приводится информация обо всех значимых элементах текста страницы. В принципе, формат допускает прямое редактирование файла в текстовом редакторе, но это крайне трудоёмко и рекомендовано только в исключительных ситуациях. Редактировать файлы с помощью адекватных средств можно в редакторе H2.

Модель расположения текста на странице, используемая в формате, представляет собой существенное упрощение модели, используемой в формате PDF. Текст на странице представляется в виде последовательности строк, при этом каждая строка представляет собой последовательность атомов двух видов: словных и несловных. Словные атомы состоят из последовательности букв и цифр, несловные - из остальных возможных символов, за исключением символа «пробел».

Каждому атому соответствует 4 целочисленных координаты (в пикселях), определяющие прямоугольник на изображении. Кроме того, у каждого атома могут быть указаны атрибуты из фиксированного списка:

- «_» - перед атомом стоит пробел
- «W» - атом словный
- «.» - атом несловный
- «!» - пометка «проверить текст»
- «?» - пометка «проверить графику»
- «-» - «перенос слова»
- «/» - разрыв абзаца

Кроме непосредственной задачи представления корректного графического индекса, формат поддерживает возможности для поэтапного редактирования файла. Так атрибуты «проверить текст» и «проверить графику» выставляются программными средствами автоматически, но подразумевают исправление или снятие редактором при дальнейшей обработке.

Атрибут «перенос слова» означает, продолжение слова атомом на следующей строке.

Информация об атоме занимает одну строку файла. Последовательности атомов, формирующих строки, разделяются друг от друга пустой строкой. В правильно оформленном индексе прямоугольники атомов каждой строки расположены по горизонтали слева направо, а по вертикали на одном и том же уровне.

Заметим, что в формате графического индекса отсутствует информация о пробелах, вместо этого используется флаг «перед атомом стоит пробел».

Дублирование текста в текстовом слое и в графических индексах на практике не приводит к техническим проблемам. Наоборот, при наличии соответствующих программных средств сличения текста мы получаем эффективный способ автоматического контроля за достоверностью воспроизведения оригинала в текстовом представлении.

В то же время, по файлу графического индекса текстовый слой не воспроизводится: информации о группировке строк в абзацы в нём отсутствует. При сличении текста в программе H2 структура абзацев восстанавливается, и появляется возможность автоматического контроля за расположением строк одного абзаца сверху вниз. Атрибут «разрыв абзаца» используется при этом для отключения

этого контроля. Это полезно в редких случаях, например при разрыве абзаца двухколонной вёрсткой, или при обтекании рисунка текстом.

6. Заключение

Предлагаемые решения внедрены в технологию подготовки информации нашей Лаборатории, и в настоящее время проходят апробацию в рамках работы по созданию Научной педагогической электронной библиотеки Российской Академии Образования [4, 5].

Литература

[1] Вигурский К.В., Трифонов С.И.. Представление печатных изданий в электронных библиотеках. *Межотраслевая информационная служба,* выпуск 2(155), стр. 17-28, 2011.

[2] Google Books http://books.google.com

[3] Трифонов С.И., Поляков А.Е. Технологический процесс подготовки изданий на примере Фундаментальной электронной библиотеки «Русская литература и фольклор». Текущее состояние и принципы модернизации. RCDL'2009. Петрозаводск: КарНЦ РАН, стр. 475-478, 2009..

[4] Прототип Научно-педагогической электронной библиотеки РАО http://bibrao.gnpbu.ru

[5] Бусев В.М., Вигурский К.В., Маркарова Т.С., Трифонов С.И. Проект создания Научной педагогической электронной библиотеки Российской академии образования. В печати.

Combined Electronic Representation of Printed Publications

© S.I. Trifonov

Combined electronic representation of printed publications allows a reader to work with both the set of graphic images of pages and the wide array of textual information. In this article we formulate and discuss various problems related to the support of this representation. The proposed solutions effect the data preparation technology as well as the architecture of the digital library's server.