

Проблема оцифровки текстового документа для хранения в информационно поисковой системе музея.

1. Введение
2. Существующие системы
3. Задачи системы
4. Проблема выбора формата
5. В каком направлении ведётся работа
6. Список литературы

Введение

Хранение текстовых документов в электронном виде способно дать огромные возможности. И самая главная из них - это поиск нужного документа по обрывку содержащейся в нём фразы. Вот простой пример: если мы хотим найти в музейном фонде документ, посвящённый какому-нибудь событию, мы должны либо знать название документа и найти его в каталоге, либо просмотреть все документы. Это займёт в лучшем случае несколько десятков минут. А в худшем случае документ, содержащий уникальную информацию, может десятилетиями не попадаться на глаза исследователям. Если же мы храним документы в цифровом виде, то нам достаточно запросить из базы все документы, в которых упоминается нужное нам событие. Компьютер за считанные секунды просмотрит все имеющиеся документы и предоставит нам информацию о тех документах, в которых упоминается интересующее нас событие. А дальше мы можем либо взять оригиналы документов из архива, либо посмотреть нужную нам информацию в электронном виде, не посещая лишний раз архив.

Существующие системы

Было бы наивно полагать, что мы первые, кого заинтересовала эта проблема. На сегодняшний день в сети можно найти большое количество библиотек, занимающихся оцифровкой своих фондов, а также электронных библиотек, собирающих электронные копии произведений со всего мира.

Одна из самых старых и наиболее известных Российских электронных библиотек – «Библиотека Мошкова». Огромное количество произведений разных жанров. И все они хранятся в виде текста, предварительно отформатированного пробелами и переносами строк. На современных устройствах такой формат неудобен даже для чтения.

Возможно, сказался возраст проекта. Если мы посмотрим на следующие по популярности электронные библиотеки «Журнальный зал» или «Альдебаран», то увидим уже оформленные html страницы или даже документы распространённых форматов (rtf и fb). Но мы не сможем увидеть, как выглядел печатный вариант произведения.

Принципиально другой подход использует большинство зарубежных проектов электронных библиотек. За основу берётся отсканированное изображение книги, которое затем обрабатывается системами OCR. По такому принципу пошла компания Google, создавая свою библиотеку. К сожалению, и в этом случае чуда не произошло. В результате мы можем увидеть, как выглядел оригинал книги, но не можем получить текст произведения. При этом сам текст хранится и используется для полнотекстового поиска, но результаты поиска предоставляются в виде фрагментов страниц документов, часто очень неудачно обрезанные.

Задачи системы

Подобные решения недостаточны для ИПС музея. Ведь мы хотим не только хранить документы, но и получить средство для автоматизации работы с ними.

Мы хотим искать документы по задаваемым пользователем фразам. Причём не только по полному тексту документа, но и по выбранным пользователем элементам документа. Для этого система должна хранить не только текст документа, но и информацию о таких элементах, как: заголовки, сноски, таблицы, подписи под рисунками, эпиграфы и другие элементы, поиск по которым может быть востребованным.

Мы хотим искать в документах упоминания объектов из других разделов ИПС. Такие

как: упоминание персоналии или фотографий, на которых изображён этот человек. Для возможности такого поиска необходимы специальные отметки в документах и это должен учитывать формат хранения документов.

В процессе поиска важно иметь возможность быстро получить общее представление о найденном документе. Для этого система должна уметь выдавать полную библиографическую информацию о документе, а также фрагмент текста, в котором система нашла поисковую фразу, в достаточном для понимания объёме. Вероятно, не меньше полного предложения, а может быть, и целый абзац.

Система должна предусматривать возможность предоставить полный текст документа в удобном для изучения виде. Удобный для чтения шрифт, корректная графическая разметка, постраничная разметка, соответствующая оригиналу, и т.п.

Система должна позволять извлекать из базы данных фрагменты документов или документы целиком для вставки в виде цитат в другие документы. А также уметь преобразовывать извлекаемые документы и их фрагменты в формат выбранной для публикации системы вёрстки или в один из распространённых форматов публикаций в интернете.

И как итоговая цель создания системы - автоматизированная «сборка» группы выбранных документов в формат готового издания с формированием необходимых для изданий подобных типов именных, предметных и прочих указателей.

Проблема выбора формата

Как уже было сказано выше, нам необходим не только полнотекстовый поиск по всему тексту документа, но и поиск по некоторым частям документа. Документы встречаются очень разнообразные. Они могут содержать внутри себя таблицы, сноски, рисунки и даже нотные записи и другую специфическую информацию. Рассмотренные выше требования к ИПС предполагают дополнительно к этому необходимость хранить маркеры для автоматической сборки указателей.

Во многих электронных библиотеках сегодня можно встретить книги в формате FB2. Этот формат зарекомендовал себя для электронных книг, где важен текст и иллюстрации с минимальным оформлением. Формат удобен для чтения, но не позволяет воспроизвести внешний вид документа. В формате не предусмотрено даже постраничного разбиения.

Аналогичная ситуация с форматом DocBook. Формат нацелен на создание структуры документа, оставляя вопрос внешнего вида документа средству визуализации. Это удобно для хранения (занимает мало места) и чтения (подстраивается под устройство, с которого ведётся чтение), но не позволяет реализовать все задачи ИПС.

Формат Open Document так же, как и Office Open XML позволяют полностью воссоздать внешний вид документа, но цена этой возможности весьма высока. Данные форматы сильно избыточны для задач хранения документов и неудобны для ручного редактирования. А существующие редакторы этих форматов часто заносят в документ большое количество лишней информации. Это не позволяет использовать данные форматы для задач ИПС.

В итоге мы оказываемся в неприятной ситуации. Существующие открытые форматы не обеспечивают нам нужной функциональности. И мы вынуждены разрабатывать свой формат. Поскольку ИПС будет работать длительное время, формат должен быть переносимым, расширяемым и обеспечивать воспроизведение документа. Такой формат мы будем разрабатывать также на базе XML, заимствуя из рассмотренных выше открытых форматов разметки удобные для нас элементы и добавляя собственные, необходимые для решения задач ИПС. При этом мы будем стремиться к простоте формата разметки, чтобы получить потенциальную возможность редактировать документ любым текстовым редактором.

Проблемы ввода документов

Но даже если у нас будет формат – сами собой документы в ИПС не появятся. Их необходимо ввести в систему. И сделать это можно двумя способами (см. рисунок):



В первом случае нам требуется минимум программных средств, но огромное количество времени - на ручное оформление текста. К сожалению, для некоторых документов, хранящихся в фондах музеев (особенно рукописей), данный способ будет единственно возможным.

Второй способ требует значительно меньше усилий. Но чтобы его можно было реализовать, необходимо подбирать программное обеспечение, а также разрабатывать конвертеры из различных форматов в формат документов ИПС.

В каком направлении ведётся работа

На данном этапе основной проблемой оцифровки стал формат хранения документов. Было принято решение создать формат хранения документов на основе XML, подходящий для решения нашей задачи. Часть тегов планируется заимствовать из формата FB2. При этом многие теги будут дополнены атрибутами. Например, тег заголовка предполагается дополнить атрибутом с оригинальным текстом буквенно-цифрового маркера заголовка.

Для указания границ бумажных страниц будет использован самозакрывающийся тег. Это позволит избежать проблем с абзацами, пересекающими страницу. Для особо сложных случаев, таких как сноски, имеющие продолжение через страницу, в теге предусмотрены атрибуты с номерами закрываемой и открываемой страницы. Предполагается, что указание номеров закрываемых или открываемых страниц в теге разделения страниц должно дать дополнительные возможности при работе с фрагментами документа.

Параллельно с разработкой формата разметки документов ведётся разработка XSLT схем для получения из хранящихся документов страниц в форматах HTML и TeX (последний выбран в качестве переносимой системы вёрстки для обеспечения автоматизированного процесса публикации документов), а также необходимых файлов "обвязки" отображения: стилевых, макропакетов, скриптов сборки указателей. Данные задачи необходимо решать параллельно, чтобы иметь возможность оценить формат и увидеть недостатки принятых решений.

Список литературы

1. XML схема FictionBook2.1 [Электронный ресурс] Режим доступа: URL: http://www.fictionbook.org/index.php/XML_схема_FictionBook2.1, свободный — Загл. с экрана (дата обращения: 14.10.2011)
2. The DocBook Schema Version 5.0 [Электронный ресурс] Режим доступа: URL:

- <http://docs.oasis-open.org/docbook/specs/docbook-5.0-spec-os.html>, свободный — Загл. с экрана (дата обращения: 22.10.2011)
3. Lib.Ru: Библиотека Максима Мошкова [Электронный ресурс] Режим доступа: URL: <http://lib.ru>, свободный — Загл. с экрана (дата обращения: 15.11.2011)
 4. OpenOffice.org XML File Format [Электронный ресурс]/ Sun Microsystems, Inc Режим доступа: URL: http://www.openoffice.org/xml/xml_specification.pdf, свободный — Загл. с экрана (дата обращения: 18.11.2011)