Развитие технологий формирования электронной коллекции периодической печати на примере казанских газет 19-го века

© Абросимов А.Г.

Научная библиотека им. Н.И. Лобачевского Казанского государственного университета
E-mail: aga@ksu.ru

Аннотация. Статья посвящена проблеме формирования коллекции периодической печати 19-го – начала 20-го веков, создаваемой в Научной библиотеке Казанского государственного университета (НБ КГУ) при поддержке Российского гуманитарного научного фонда (проект № 04-01-12032в).

Введение. За двухсотлетнюю историю в Научной библиотеке Казанского государственного университета (НБ КГУ) сформировалось богатейшее собрание документов 9-го – 20-го веков. Большую ценность представляет коллекция цензорских экземпляров казанских газет конца 19-го – начала 20-го веков.

Коллекция редких книг и рукописей активно используется в учебном процессе и научных изысканиях и поэтому приоритетным направлением в создании электронной библиотеки КГУ является формирование коллекции периодической печати.

Метод формирования коллекции периодической печати – сканирование изданий с использованием планетарного бесконтактногосканера.

Наиболее удобным форматом данных для читателя является текстовый. Но для изданий 19-го века основной проблемой становится качество печати самих изданий. Трудозатраты на исправление распознанного текста выбранных периодических изданий сравнимы с трудозатратами на набор этого текста с оригинала (результат – 5-10 % распознанных слов) и в связи с этим коллекция создается из электронных образов страниц.

Профиль метаданных [2]**.** Особенностью предлагаемой коллекции периодической печати является определение электронного документа. Естественной единицей информации является статья, которая может быть и меньше страницы, и располагаться на нескольких страницах. Другой особенностью газет того времени является то, что значительная часть публикаций не озаглавлена и не подписана авторами, что делает не возможным полное библиографирование по заголовкам и фамилиям. Газета состоит из более или менее кратких сообщений, объявлений и других видов информации. Поэтому в качестве описываемой единицы содержания выбрана часть текста, ограниченная определенной темой – сообщение о международных отношениях, сообщение о научных открытиях, всевозможные объявления, статьи и т. д. В качестве электронного документа – номер издания.

Таким образом, структура метаданных представляет собой:

- ***описание коллекции***, содержащее описание структуры коллекции, список разделов коллекции и т. д.;
- ***описание разделов коллекции*** – содержащее общее описание конкретного издания;
- ***описание электронных документов***;
- ***описание единиц содержания*** – статей, объявлений, сообщений и т. д.

При создании настоящей коллекции было принято решение разрабатывать свой профиль метаданных, используя Dublin Core и добавляя свои дополнительные уточняющие классификаторы. Моделью описания метаданных выбрана система RDF (Resource Description Framework). Для обеспечения интероперабельности и описания синтаксиса XML-документа, содержащего метаданные, разработаны XML Schema для каждого типа XML-документов, определено XML-пространство имен.

Выбор лингвистических средств [3]**.** Одной из ключевых задач развития электронной коллекции является повышение эффективности навигации и поиска в массиве электронных документов. Решение этой проблемы предусматривает создание информационной системы, ориентированной на поиск и представление информации пользователю. Центральное место в такой системе занимает лингвистическое обеспечение.

В качестве основного инструмента для систематизации и поиска газетных публикаций был выбран подход, основанный на многоаспектной классификации текстов – фасетная система классификации, позволяющая выбирать признаки классификации независимо друг от друга и многоаспектно охарактеризовать специфический газетный материал.

В процессе аналитико-синтетической переработки текста изданий описывалась информация каждой единицы содержания. Систематизация содержания издания осуществлялась по следующим аспектам: ***виду информации, сфере общественной жизни, персонам*** (именам, встречающиеся в газете), ***учреждениям, географическим названиям мест, датам событий***, приведенным в тексте.

Так же предусмотрена группировка текстов и по такому формальному признаку, как газетная рубрика. Дополнительно в процессе аналитико-синтетической обработки газетного текста формируются авторитетные файлы ***авторов, жанров периодики, список периодических изданий***, из которых перепечатан новостной материал.

Применение опыта создания коллекции. Газеты и журналы, входящие в формируемую коллекцию, самых различных типов и направлений. Так, например, содержание газет представляет собой многоплановый синтетический материал, включающий в себя самую разнообразную по жанру, происхождению, содержанию информацию. Наряду с официальными сообщениями и документами, законодательными актами, объявлениями, некрологами, письмами, городской хроникой в них публикуется обширный научно-образовательный и литературно-художественный материал. Журнал «Заволжский муравей» – литературный. Таким образом, коллекция представляет широкий спектр изданий, за исключением только научных.

Таким образом, использованный подход к созданию коллекции периодической печати 19-го – начала 20-го веков может быть предложен и для создания других коллекций периодики, исключая коллекции научных изданий. Для научных изданий гораздо естественнее использовать уже существующие классификационные языки и имеющиеся проработанные профили метаданных.

В процессе работы над коллекцией периодической печати 19-го – начала 20-го веков был определен алгоритм создания коллекции:

- анализ издания, определение типа издания;
- определение электронного документа, определение описываемой единицы содержания;
- анализ структуры издания, выделение типовых рубрик, формирование классификационных справочников, определение схемы лингвистического обеспечения, формирование списков значений признаков, формирование профиля метаданных;
- формирование электронных документов – сканирование, обработка изображений, распознавание текста;
- формирование метаданных;
- разработка навигационного и поискового программного обеспечения.

Применительно к современным изданиям алгоритм формирования коллекции не требует модификации. Более того, состав и структура фасетов предлагаемой системы классификации не зависят от типа и времени выхода издания. Принцип, использованный при разработке профиля метаданных – схема Dublin Core с уточняющими квалификаторами – можно считать универсальным. В подтверждение можно привести слова А. Б. Антопольского [1]: «Таким образом, Дублинское ядро представляется как вершина иерархии систем метаданных, которая развивается более детально в конкретных коллекциях или сервисах системы электронной библиотеки при помощи частных систем метаданных».

Заключение. Подводя итоги, можно сделать следующий вывод – опыт создания коллекции периодической печати 19-го – начала 20-го веков может быть использован при формировании других коллекций периодической печати. Более того, развитие электронных коллекций основанных на перспективных технологиях Semantic Web, будет способствовать созданию эффективной инфраструктуры для поддержки научных исследований, образования и других сфер деятельности.

Литература:

[1] Антопольский А.Б. Лингвистическое обеспечение электронных библиотек.// М.: ФГУП Научно-технический центр «Информрегистр», 2003. – 302 с.

[2] Абросимов А. Г. Метаданные описания коллекции периодической печати [Электронный ресурс] // Электронные библиотеки: рос. науч. электронный журн. – 2005. – Т. 8, вып. 2. – http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2005/part2/Abrosimov

[3] Абросимов А. Г., Кузьмина В. Ю., Салосина И. И. Применение фасетной классификации для систематизации газетных публикаций в электронной коллекции казанской периодической печати 19-го – начала 20-го веков [Электронный ресурс] // Электронные библиотеки: рос. науч. электронный журн. – 2006. – Т. 9, вып. 1. – http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2006/part9/Abrosimov

Technologies development of the periodical press electronic collection forming on the example of Kazan newspapers of the 19th century

© Abrosimov A.G.
Scientific library named after N.I.Lobachevsky
of Kazan State University
E-mail: aga@ksu.ru

Annotation. The article is devoted to the problem of forming of the periodical press collection of 19th – beginning of 20th century, creating in Scientific library of Kazan State University (SL KSU) with the help of Russian Humanitarian Scientific Fund (project # 04-01-12032в).