

## Отчет по работе в проекте «Открытая Электронная библиотека г.Томска для высшего образования России»

Группа создания полнотекстового контента

Руководитель группы – Карауш А.С. (заместитель директора МИБС Томска)

Для осуществления работ по созданию и представлению полнотекстовых документов программными средствами участников консорциума была образована группа специалистов от библиотек. В нее вошли представители библиотек и организаций:

- Томского политехнического университета;
- Томской областной универсальной научной библиотеки;
- Томской областной детско-юношеской библиотеки;
- Научно-медицинской библиотеки СГМУ;
- МИБС Томска;
- МУБИС г.Нижевартовска;
- Библиотека Томского университета систем управления и радиоэлектроники;
- Центр телекоммуникаций Томского государственного университета.

На начальном этапе работы группы были обозначены цели и задачи, а также намечены основные этапы выполнения проекта.

Так, планировались 4 основных задачи- временные этапы работы группы:

- 1) Изучение опыта работы групп представления полнотекстового контента в г. Томске и России (май-август 2002 г.);
- 2) Разработка технологической схемы работы представления полнотекстового контента (июнь-сентябрь 2002 г.);
- 3) Создание полнотекстового контента библиотеками-участниками проекта (август 2002 – февраль 2003 г.);
- 4) Отчет о работе группы за 2002-2003 гг. (январь-март 2003 г.).

Для решения **первой задачи** необходимо было изучить опыт работы коллег из других библиотек, а также из организаций, занимающихся по роду своей деятельности созданием (оцифровкой) полнотекстовых документов. Для решения этих задач были проведены консультации во время мероприятий:

- 1) Международная конференция «Крым –2002»;
- 2) Международная конференция «Либком-2002»;
- 3) Научно-практический семинар «Проблемы применения современных технологий для представления старопечатной и рукописной (старообрядческой) книги как информационного ресурса»;
- 4) Тренинги Томского тренинг-центра, посвященные вопросам создания и представления полнотекстов;
- 5) Тренинги других тренинг-центров России;
- 6) Интернет-форум в рамках сервера проекта «Открытая электронная библиотека г.Томска».

Во время работы **над первой задачей** был обработан большой массив информации, касающийся форматов хранения полнотекстовых ресурсов, форматов архивирования данных документов, а также особенностей представления ресурсов для конечного пользователя, исходя из опыта работы различных организаций.

В результате было принято решение о приоритетном использовании для представления полнотекстовой информации формата файла, разработанного фирмой Adobe и известного, как Adobe Acrobat. Причем было принято решение об использовании последней на то время редакции этого формата в версии программы Adobe Acrobat 5.0, по причине отсутствия проблем с представлением кириллических символов, которые наблюдались в более ранних версиях этой программы.

Выбранный нами формат позволяет использовать все виды источников информации для представления как текстового, так и графического наполнения с использованием защиты документов от несанкционированного извлечения текстовой и иной информации из файлов. Решение об использовании одного формата файлов для представления полнотекстовой информации позволило существенным образом сократить и упростить разработку технологического процесса подготовки полнотекстовых документов.

Для решения **второй задачи** была разработана технологическая схема создания и представления полнотекстовых документов с учетом особенностей выбранного формата файлов и правил каталогизации в стандарте и коммуникативном формате Rusmarc.

Технологическая схема создания полнотекстовых документов состоит:

1. Отбор литературы (информационных источников) из фондов библиотеки, отвечающих требованиям теме цифровой коллекции и всем юридическим моментам использования для создания цифровой копии (законы об авторских и смежных правах).
2. Подготовка технологического процесса сканирования – получение готового файла, содержащего распознанный или графический образ документа.
3. Подготовка технологического процесса описания полученных и приобретенных полнотекстовых ресурсов в собственном электронном каталоге.
4. Подготовка составляющих для представления и доступа пользователей к электронному каталогу с полнотекстовыми ресурсами.
5. Изучение использования и дополнительного спроса на полнотекстовые ресурсы библиотеки.

На первом этапе необходимо определиться с типами источников, типами описаний в соответствии с правилами каталогизации, коммуникационными форматами, типами файлов для оптимального представления информации. Данный этап занимает от 30% до 50% всего времени создания полнотекстовых ресурсов.

На втором этапе производится сканирование, оцифровка и сохранение полученных образов в виде файла. Приняты следующие технологические параметры:

2.1. Использование разрешения сканирования 300 dpi.

2.2. Использование единого цикла распознавания и сохранения программы ABBYY FineReader 6 в пакетном режиме в формате Adobe Acrobat.

2.3. Отсутствие технологической операции редактирования и коррекции при использовании режима замены неуверенно распознанных букв их графическими образами.

2.4. Сохранение образа сканированного документа в графическом формате tiff на компакт-дисках для дальнейшего использования других форматов хранения полнотекстовых документов.

Второй этап занимает не более 10-15% от всего времени создания полнотекстовых ресурсов, но требует максимального количества аппаратных средств относительно других этапов.

Третий этап целиком зависит от технологических возможностей программного обеспечения автоматизации библиотечных процессов. Однако следует придерживаться правил создания библиографической информации для последующего преобразования в коммуникационные форматы. И не «изобретать велосипед» в плане придумывания точек доступа к ресурсам и форматов описаний [1-3]. Библиографическое описание на электронный ресурс, как правило, обладает многоуровневыми ссылками. Третий этап занимает от 10% до 20% общего времени создания и представления ресурса.

На четвертом этапе предстоит решить множество проблем, связанных с системой доступа к полученным ресурсам и определить стоимость и открытость полнотекстовых

документов. Система доступа пользователей к полнотекстовым ресурсам должна содержать несколько уровней, например, библиотека: вуз: город: все остальные. Расширения (типы) отдельных файлов полнотекстовых ресурсов должны использовать только латинские буквы без пробелов, а в наименованиях файлов желательно ограничиться использованием только цифр, как знаков, нечувствительных к регистру. Четвертый этап занимает от 10% до 20% общего времени создания полнотекстовых документов.

Пятый этап не связан напрямую с затратами времени на создание ресурсов, но от качественной работы этого этапа и обеспечения технологической цепочки обратной связи зависят оптимальные показатели объема полнотекстовых документов в фондах библиотеки. При наличии технических возможностей необходимо проводить мониторинг выполненных запросов и отказов, опросы мнения читателей по использованию ресурсов, а также анализировать планы организаций и различных структур населенного пункта на предмет опережающего накопления полнотекстовых ресурсов с целью удовлетворения запросов пользователей.

Нормально работающая технологическая схема данного процесса позволяет создавать 3000 страниц полнотекстовых документов путем сканирования силами одного работника отдела создания баз данных (каталогизации) при использовании вычислительной техники «средней» мощности.

Особенностями предложенного технологического процесса создания полнотекстовых документов следует считать то, что не должно содержаться этапов редактирования распознанного документа. Современное программное обеспечение позволяет «прятать» неуверенно распознанные символы за графическое изображение слова. Таким образом, пользователь видит текст книги «правильно» и без ошибок распознавания.

В результате использования данной схемы библиотека со «средними техническими возможностями», подготовив необходимые документы на основные технологические этапы сканирования и, применив современное программное обеспечение, получает технологический процесс создания полнотекстовых ресурсов с минимальными материальными и производственными затратами.

**Третья задача** решалась партнерами самостоятельно по предложенным алгоритмам и технологическим схемам. Результатом явилось представление полнотекстовой информации в свободный доступ для задач составления метаописаний и использования для работы полнотекстового хранилища документов, созданного в рамках консорциума «Открытая электронная библиотека г.Томска».

Во время решения задачи создания полнотекстовых документов велась активная работа в рамках группы по консультированию партнеров правильному использованию выбранной технологической схемы. Сложность применения единой технологической схемы во всех библиотеках объясняется:

1. Различной степенью подготовки технических специалистов и специалистов отдела каталогизации;
2. Особенности организационной структуры библиотеки, где в одном случае для создания полнотекстовых ресурсов были отвлечены технические специалисты, а в другом вся работа была сделана силами отделов каталогизации (обработки документов).

**Четвертая задача** состояла в представлении результатов работы группы. В рамках работы международной конференции «Крым-2003» Караушем А.С. был сделан доклад на тему «Опыт создания полнотекстовых ресурсов в Муниципальной информационной библиотечной системе города Томска», где были кратко обозначены основные этапы технологических решений, принятых в организации- участниках проекта.

Считаю, что слаженная работы представителей группы, в том числе удаленная, позволила создать и представить требуемые полнотекстовые ресурсы с минимальными затратами сил специалистов, средств и времени.

### **Литература**

1. Российский коммуникативный формат. Электронные ресурсы. Материалы к обсуждению [Электронный ресурс]/ Национальная Служба развития системы форматов RUSMARC, СПб: 2002, – . – Режим доступа: <http://www.rba.ru:8101/rusmarc/discuss/er/er.htm>, свободный. – Загл. с экрана.
2. IFLA Universal Bibliographic Control and International MARC Core Programme (UBCIM). UNIMARC Guidelines no. 6. Electronic Resources [Электронный ресурс]/ International Federation of Library Associations and Institutions (www.ifla.org), 2000,-.- Режим доступа: <http://www.ifla.org/VI/3/p1996-1/guid6.htm>, свободный. – Загл. с экрана. (перевод на русский язык: <ftp://library.tomsk.ru/pub/IRBIS/mils-docs/er-unimarc-rus.zip>)
3. Guidelines for Distinguishing Cartographic Electronic Resources from other Electronic Resources. [Электронный ресурс]/ Library of Congress: 2002, – . – Режим доступа: <http://www.loc.gov/marc/cfmap.html>, свободный. – Загл. с экрана.