РАЗРАБОТКА ТЕХНОЛОГИИ СОЗДАНИЯ ЭЛЕКТРОННЫХ КОЛЛЕКЦИЙ ИЗДАНИЙ XIX - НАЧАЛА XX ВЕКА.

Рузанова Н.С., Леонтьев А.А., Иванова С.В.
Петрозаводский государственный университет
г. Петрозаводск, пр. Ленина, д. 33
ruzanova@mainpgu.karelia.ru

В Национальной библиотеке Республики Карелия (НБ РК) и Научной библиотеке Петрозаводского госуниверситета (НБ ПетрГУ) собираются и хранятся издания, отображающие этапы развития республики во всем его многообразии от первых рукописных документов до современных изданий. Многие документы, зачастую существующие в единственном экземпляре в библиотеках Карелии и России, активно используются и могут исчезнуть из-за несвоевременных мер по обеспечению их сохранности. Закон “О библиотечном деле” Республики Карелия гарантирует права человека на “свободный доступ к информации, знаниям, приобщение к ценностям национальной и мировой культуры” Вместе с тем, на сегодняшний день остро стоит проблема обеспечения доступа к уникальным изданиям для широкого круга читателей.

Один из наиболее предпочтительных выходов из сложившейся ситуации – оцифровка и распознавание редких материалов и организация доступа к ним через сеть Интернет. Такого рода технологии успешно внедряются во многих библиотеках России, и Карелии в частности, для организации доступа к современным текстам. Однако на сегодняшний день не существует оптимального подхода к оцифровке и организации доступа к текстам в орфографии XIX-начала XX века. Перед библиотеками встают две основных проблемы:

- сложность распознавания такого рода текстов с помощью стандартных средств;
- невозможность организации контекстного поиска по текстам.

Несмотря на то, что русский язык не претерпел кардинальных изменений на протяжение XIX-XX веков, существенная перестройка произошла в системе русского письма. В 1918 году из нее были удалены некоторые буквы, унифицированы флексии (окончания), введены новые написания слов. В связи с этим поиск многих слов в полнотекстовой базе данных, содержащей слова в их дореволюционной форме, в значительной степени затруднен и требует от пользователя знания орфографических норм XIX века.

Существующие решения, такие как хранение текстов в виде графических файлов или подготовка текстовых файлов и автоматическая замена в них “нестандартных” символов (фита, ижица, и десятеричное и ять) на современные, имеют ряд недостатков. В первом случае невозможно организовать поиск по тексту и предоставить пользователю возможность работать с источником как с полноценным компьютерным текстом. Во втором случае также очень сложно организовать поиск, т.к. эти тексты не отвечают правилам орфографии ни XIX, ни XX века.

В рамках проекта, профинансированного проектом "Прожект Хармони, Инк", специалистами Центра Интернет и Национальной библиотеки Республики Карелия, была разработана технология оцифровки, хранения и предоставления доступа к текстам в орфографии XIX века, отвечающая следующим требованиям:

1. Простота создания электронных материалов – на всей цепочке “книга – электронный документ” от создателя электронного ресурса практически не требуется знания языков HTML, JavaScript и т.п. Функции “перевода” текста в современную орфографию, распознавания, размещения в сети Интернет берет на себя разработанное программное обеспечение.

2. Открытость и соответствие стандартам – доступ к текстам предоставляется по стандартным протоколам Интернет (http и ftp). Получаемые пользователями материалы соответствуют стандартам, принятым в Интернет (HTML, XML, JPEG).
3. Максимальное удовлетворение запросов пользователей – пользователю предоставляется широкий спектр функциональных возможностей (контекстный поиск и поиск по описанию документа, импорт текстов в различных форматах и т.п.); учет потребностей работы с текстом максимально широкого круга пользователей.

Результат работы программного обеспечения – три типа файлов:

1. Описание документа в формате RUSMARC и USMARC.
2. Коллекция графических файлов в формате JPEG, являющихся цифровыми копиями страниц книги.
3. Набор файлов в формате HTML – переведенный в современную орфографию текст книги. Именно по этому файлу осуществляется контекстный поиск.

Представление документа в виде набора файлов разного формата позволяет максимально удовлетворить запросы пользователя и делает технологию применимой для текстовых источников разного типа, не ограничивая сферу ее использования только книгами XIX века.

Работы в рамках проекта:

1. Разработан алгоритм перевода текстов XIX века в современную орфографию.
2. Разработан программный комплекс – автоматизированное рабочее место “Сектор редкой книги – создание полнотекстовых коллекций”, позволяющий в автоматическом режиме выполнять операции:
    - оцифровки книг, распознавания и проверки орфографии;
    - перевода текстов в современную орфографию;
    - конвертирования текстов в формат HTML (для текста в современной орфографии);
    - публикации текстов в сети Интернет.
3. Разработано серверное программное обеспечение для автоматизации процессов:
    - импорта и экспорта документов в электронную коллекцию;
    - поиска по описаниям и контекстного поиска по электронным документам коллекции.

Алгоритм перевода текстов XIX века в современную орфографию.

Для облегчения подготовки текстов разработан алгоритм, позволяющий “переводить” слова из орфографии прошлого века в современную. Алгоритм включает в себя следующие шаги:

- “Перевод” букв “ять”, “и десятеричное”, “фита в соответствующие современные графемы (Е, И, Ф).
- “Перевод” ряда окончаний в современный вид (напр., “ыя” - “ые”, “ея” - “её”).
- Устранение буквы “ер” (Ъ) на конце слов (напр., “предъ” в “пред”).
- Приведение к современным орфографическим нормам написания слов типа “безъименный”, “безподобно” и др.

Использование подобного алгоритма позволяет автоматизировать примерно 80% работы по приведению графики в соответствие с современными орфографическими нормами.

Автоматизированное рабочее место “Сектор редкой книги – создание полнотекстовых коллекций”

АРМ “Сектор редкой книги – создание полнотекстовых коллекций” предоставляет библиотекарю возможность в автоматизированном режиме перевести печатный материал в цифровую форму. Работа по подготовке материалов состоит из следующих этапов:

- Перевод издания в цифровой вид с помощью сканера или цифрового фотоаппарата (цифровой фотоаппарат используется для особо ценных изданий).
- Распознавание и проверка орфографии текста (на основе OCR FineReader Professional Developer Edition и специально разрабатываемых модулей и словарей);
- Перевод текстов в современную орфографию.
- Конвертирование текста в формат HTML с предварительной ручной разметкой содержания документа.
- Публикация документа на сервере электронных коллекций и добавления записи о нем в базу данных.

С использованием технологии был подготовлена полнотекстовая версия книги " Описание Олонецкой губернии в историческом, статистическом и этнографическом отношениях, составленное В. Дашковым. - СПб.,1842.-222 с.". Книга включает в себя 219 страниц текста, титульный лист книги, карту Олонецкой губернии и гравюру "Вид города Петрозаводска". Книга опубликована на сайте “Библиотеки Карелии” (http://libraries.karelia.ru) в разделе “Электронная библиотека”.

Дальнейшее развитие проекта видится в расширении коллекции электронных документов. Новое направление – использование геоинформационных технологий при анализе карт XVIII-XIX века в процессе проведения различных научных исследований.

ELABORATION OF THE TECHNOLOGY OF CREATING OF ELECTRONIC COLLECTIONS OF XIX-THE BEGINNING OF XX-CENTURY ISSUES.

Ruzanova N.S., Leontyev A.A., Ivanova S.V.
Petrozavodsk State University
Petrozavodsk, Lenin Av., 33
ruzanova@mainpgu.larelia.ru

The article deals with the technology of creating of electronic collections, which has been elaborated by the Internet Center of Petrozavodsk State University. It includes books digitizing, recognition, translating texts into the modern spelling , converting of texts into the HTML format (for texts in modern spelling) and texts' publication in the Internet .