К СОЗДАНИЮ РЕГИОНАЛЬНОГО РЕСУРСНОГО ЦЕНТРА[1]

Сулейманов Д. Ш., Невзорова О.А., Аюпов М.М.

E-mail: dvdt@telecet.ru,olga.nevzorova@ksu.ru,madehur@ail.ru

Казанский государственный университет,Татарский государственный гуманитарно-педагогический университет

Аннотация. *В статье обсуждаются актуальные вопросы создания регионального ресурсного центра в целях интеграции и координации деятельности по формированию фонда электронных коллекций, а также разработке многоязычных информационно- поисковых сервисов.*

To the Development of the regional resource center

Suleymanov D., Nevzorova O., Ajupov M.

Abstract. *In this paper the actual problems of the development of the regional resource center are discussed. The main tasks of this center are integration and coordination of activity intended to creation the foundation of electronic collection and also the development of Multilanguage information retrieval service.*

1. *Введение*

Информационный бум, возникший в 20 столетии, предопределил качественно новые подходы к задачам хранения и использования информации. Новые информационные технологии уже в 70-80 годах двадцатого века стали использоваться для реализации концепции машинного фонда (МФ) национальных языков в различных странах. Фактически, Машинный фонд представляет собой сложную, иерархическую, разветвленную автоматизированную систему, способную решать как информационно-поисковые, так и исследовательские лингвистические задачи. Однако, первоначально, основными составляющими Машинного фонда являлись словарные ресурсы, отдельные тематические коллекции текстов, а также достаточно ограниченный программный инструментарий, предназначенный для ведения коллекции.

В настоящее время происходит новая волна информатизации, связанная с внедрением новых технических и программных средств обработки лингвистических данных, а также средств глобальной коммуникации. Существенно возросли объемы электронной информации, созданы новые информационные технологии структурирования и поиска информации. Возникли и активно используются во всем мире электронные библиотеки различных уровней. Актуальным для РТ является создание регионального ресурсного центра, интегрирующего различные полнотекстовые электронные коллекции и соответствующий программный инструментарий для обработки коллекций. Для образовательного пространства РТ необходимо создание сети региональных тематических научно - образовательных коллекций на различных языках. ***2. Региональный ресурсный центр в РТ*** Идея создания региональных ресурсных центров широко обсуждается в литературе [1]. Подобные центры должны выполнять роль координатора всех процессов ведения электронных коллекций – от организационно-технического сопровождения до правового регулирования вопросов охраны интеллектуальной собственности и взаимодействия с внешними информационными системами.

Актуальными для регионального ресурсного центра должны стать задачи интеграции существующих электронных коллекций в РТ, подготовка новых тематических научно - образовательных коллекций, а также развитие соответствующего программного инструментария для взаимодействия с текстами коллекции.

Тематические научные коллекции создаются в РТ различными исследовательскими коллективами, например, можно отметить проект электронного журнала "Lobachevskii Journal of Mathematics"(LJM - http: // ljm.ksu.ru) [2], в котором решены важные вопросы, связанные с автоматической конвертацией статей в наиболее популярные форматы; вопросы генерации метаданных, поиска данных и др., ряд других электронных журналов, а также сайтов различных образовательных учреждений, поддерживающих собственные тематические образовательные коллекции.

При этом можно отметить, что создаваемые коллекции текстов, в основном, являются русскоязычными. Фактически, отсутствуют системные коллекции на татарском языке, а также программные средства их обработки. Именно на решение задачи разработки программных средств, обеспечивающих поиск в электронных двуязычных коллекциях текстов, направлен совместный проект РФФИ и ИВФ РТ "Разработка прикладной грамматической модели татарского языка для задач интеллектуального информационного поиска в многоязычных корпусах текстов", выполняемый коллективом авторов настоящей статьи.

1. *Обеспечение двуязычного информационной поиска в тематической электронной коллекции*

[1] Работа выполнена при поддержке Российского Фонда Фундаментальных Исследований, грант № 05-07-90257 и Инвестиционно-венчурного фонда РТ

Проект выполняется совместно с разработчиками Университетской информационной системы РОССИЯ (НИВЦ МГУ, www.cir.ru). Современные поисковые технологии позволяют адаптировать основные поисковые механизмы для любого нового языка, для которого разработана соответствующая лингвистическая поддержка на уровне развитых морфологических и частичных синтаксических моделей. Такой уровень поддержки татарского языка практически полностью обеспечивается функциональностью прикладной грамматической модели татарского языка, разработанной в рамках проекта. Прикладная грамматическая модель татарского языка, адаптированная к задачам поиска, содержит две основные компоненты: морфологический анализатор и модуль частичного синтаксического анализа, предназначенный для распознавания аналитических конструкций в татарском тексте.

Интегрирование прикладной грамматической модели татарского языка в УИС РОССИЯ позволяет эффективно поддерживать многоязычный поиск в татарско - русской электронной коллекции текстов.

Функциональные возможности УИС РОССИЯ позволяют:
- вести большие полнотекстовые базы данных для интеграции общественно - политической информации о жизни региона и федерации в целом;
- реализовывать стандартные возможности полнотекстового поиска - контекстный поиск по документам базы данных с учетом морфологии, включая подсветку результатов поиска;
- автоматически выделять формальную метаинформацию для обрабатываемых документов – автор, заглавие, дата и т.п. Язык запросов позволяет включать в запрос условия по любой метаинформации одновременно для нескольких классов документов.

Аналитические возможности УИС РОССИЯ позволяют автоматически выделять тематическую метаинформацию (построение терминологического индекса по общественно - политическому тезаурусу, автоматическая рубрикация одновременно по нескольким рубрикаторам, автоматическое аннотирование). Уникальной особенностью является возможность качественной автоматической рубрикации по иерархическим рубрикаторам большого размера (более 500 рубрик). Язык запросов позволяет включать в запрос логические условия любой сложности, где частным условием является любой элемент метаинформации или контекста. УИС РОССИЯ поддерживает обработку многоязыковых текстов [3]. При этом под многоязыковой поддержкой понимается:

1. (a) возможность построения запросов на различных языках к текстам коллекции;
   (b) возможность использования многоязычного тезауруса по общественно - политической тематике для обработки и поиска документов;
   (c) возможность подсветки результатов запроса для обоснования релевантности документа;
   (d) возможность анализа содержания иноязычного документа средствами на родном языке.

Возможность построения запроса на татарском языке к русско - татарской коллекции текстов базируется на морфологии разбора документа и построении морфологического поискового индекса. Многоязычный тезаурус по общественно-политической тематике для обработки и поиска документов, снабженный синонимическими рядами концептов на татарском языке, обеспечивает механизмы смены языка запросов.

1. ***Формирование тематических научно-образовательных электронных коллекций на татарском языке***

В рамках указанного выше проекта авторами была подготовлена экспериментальная электронная коллекция татарских текстов по общественно-политической тематике размером 90 Мб.

Основные источники информации для экспериментальной коллекции:

1. http://www.tatar.ru – официальный сервер Республики Татарстан
2. http://www.tatarlar.ru – всемирный татарский сервер
3. http://tatar-kongress.org – официальный сайт Исполкома Всемирного Конгресса Татар;
4. http://www.tatar-inform.ru - официальный сайт информационного агентства РТ
5. http://www.tatarca.boom.ru и другие интернет-сайты
6. издания Академии Наук Республики Татарстан (журналы «Фэнни Татарстан» и «Научный Татарстан»)
7. издания национального издательства «Магариф»
8. издания Молодежного общественного фонда «Сэлэт»
9. научно-популярный журнал «Ф?н ??м тел» («Наука и язык»).

Анализ основных проблем, связанных с подготовкой татарского сегмента коллекции, показывает, что в настоящее время общественно-политическая тематика представлена в изданиях на татарском языке в небольшом объеме. Существуют определенные проблемы, связанные с установлением авторства и выходных данных ряда публикаций. В первую очередь это относится к публикациям, выставленным в Интернет. Одним

из требований к документам, помещаемым в коллекцию текстов, является их полная идентификация, что означает, либо установление точных выходных данных документов, либо их размещение на официальных сайтах.

Основными источниками информации в этой области могли бы стать официальные органы власти и управления РТ (коллекция официальных документов на русском и татарском языках); издательства, выпускающие периодические издания, прежде всего газеты и журналы на татарском языке. Актуальной проблемой является обеспечение эффективного взаимодействия с этими структурами, решение организационно-правовых вопросов.

Основное содержание коллекции по рубрикам было классифицировано по следующим рубрикам:

«Политика» - официальные документы Госсовета РТ, Правительства РТ, Конституция РТ, протоколы заседаний Исполнительного комитета Всемирного конгресса татар (2000 г.), тексты по истории религии, татарский толковый словарь общественно-политических терминов, Интернет-статьи с официального сайта РТ.

«Экономика. Экономические науки» - программа социально - экономического развития РТ на 2005-2010 годы, газетные публикации (социально-экономические обзоры).

«Образование. Воспитание. Обучение. Организация досуга» - документы Республиканского Молодежного общественного Фонда «Сэлэт».

«Этнография. Нравы. Фольклор. Обычаи» – татарские народные песни, публицистика.

Статьи журналов АН РТ «Фэнни Татарстан» (на татарском языке) и «Научный Татарстан» (на русском языке) за 2000-2005 г.г.

«Пресса» - статьи из периодических газет и журналов РТ, Интернет-статьи с сайта информационного агентства РТ.

Сводная статистическая информация по составу экспериментальной коллекции:

Количество файловКоличество слов в файлах

Русские тексты145 > 284832

Татарские тексты4812289761

При подготовке экспериментальной коллекции была решена актуальная задача конвертации татарских текстов в стандартную кодировку. В настоящее время существует несколько видов кодировок татарских букв, отличных от стандартной кодировки (Постановление Кабинета Министров РТ № 1064 от 12 декабря 1996 г.). Следует отметить, что коллективом проекта разработаны стандартные кодировки татарского алфавита на основе кириллической и латинской графики для различных операционных систем: DOS (OEM-кодировка), WINDOWS 95 (ANSI-кодировка), WINDOWS_98 (ANSI и UNICODE кодировки), WINDOWS_NT (2000, XP, и т.д.) (UNICODE кодировка). Новый стандарт кодировки национальных алфавитов (UNICODE) предусматривает 16-битовую кодировку (вместо 8-битовой) и поддерживает алфавиты большинства национальных языков (в частности, уже внедрен в ОС WINDOWS NT, UNIX). Стандарт UNICODE был выбран в качестве базового стандарта текстовой коллекции, и исходные тексты, представленные в различных кодировках, были конвертированы в стандарт UNICODE. Дополнительно был подготовлен вариант коллекции в кодировке Win1251, который используется программным обеспечением татарского морфологического анализатора.

Заключение

Формирование ресурсного регионального центра является актуальной задачей сегодняшнего дня. Необходимость интеграции и кооперации усилий всех заинтересованных сторон является достаточной очевидной. В РФ с 2005 года идет процесс институционализации общероссийской ассоциации на базе Некоммерческого партнерства "Электронные библиотеки"(НП ЭЛБИ, www.elibra.ru). В число основных направлений деятельности и программы действий НП ЭЛБИ входят проекты по разработке и внедрению реальных механизмов координации и комплектования электронных коллекций, в том числе и на основе региональных ресурсных центров.

Результаты представленного в статье проекта обеспечивают поддержку программного инструментария двуязычного (русско - татарского) информационного поиска в электронных коллекциях на базе УИС РОССИЯ, который может быть включен в качестве информационного сервиса в систему регионального ресурсного центра.

Литература

1. Чугунов А.В. Формирование тематических научно - образовательных коллекций: интеграция данных и координация комплектования // Труды Восьмой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", 2006. – С. 353-358.

2. Елизаров А.М., Липачев Е.К., Малахальцев М.А. Технологии Semantic Web в практике работы электронного журнала по математике // Труды Восьмой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", 2006. – С. 215-218.

3. Добров Б.В., Лукашевич Н.В. Организация двуязычного поиска в Университетской системе РОССИЯ // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Четвертой Всероссийской научной конференции RCDL'2002 (Дубна, 15-17 октября 2002 г.): В 2 т. – Дубна: ОИЯИ, 2002. – Т.2. – С.148-158.