Анна Юрьевна Филиппович
к.т.н., доцент кафедры Медиасистем и технологий
МГУП им. Ивана Федорова

Информационные технологии сохранения культурно-исторического наследия России

Введение

Сегодня, живя в век высоких технологий, человек теряет связь со своей историей. Для формирования высоконравственного общества требуется создание информационного пространства, которое будет базироваться в том числе на исторических текстах. Это будет иметь культурно-нравственное воспитательное значение для молодежи, оторванной от славянских корней и будет способствовать расширению сферы влияния славянской культуры в мире.

Актуальной проблемой нашего общества является сохранение культурно-исторического письменного наследия, представленного печатными и рукописными источниками в архивных фондах и библиотеках России. В настоящее время эта проблема находит решение в создании цифровых факсимильных копий исторических книг и рукописей. Однако факсимильное копирование не позволяет в полной мере решить проблему доступа к конкретному историческому документу, т.к. будучи неиндексированными исторические тексты выброшены из информационного Интернет-пространства.

Целью представляемого проекта является создание современных информационных технологий и инновационных разработок для сохранения исторических и культурных ценностей России. Основные задачи – расширить возможности читателя-исследователя, используя современные системы поиска и анализа текста, автоматизировать создание электронных копий исторических книг и рукописей, сформировать современные электронные ресурсы для научного исследования старинных книг и рукописей.

В зоне нашего внимания оказалось несколько культурно-исторических пластов: музыкальные произведения XI-XVI вв., скорописные тексты XVII века, печатные источники XVIII – нач. XIX в.

В настоящий момент проект реализуется в трех самостоятельных направлениях:

1. Информационная технология переиздания печатных источников XVIII – нач. XIX вв.
2. Компьютерная Семиография.
3. Компьютерный фонд древнерусской скорописи.

Информационная технология переиздания печатных источников XVIII – нач. XIX вв.

Представляемая технология основана на выделении группы исторических источников с определенными свойствами (в числе которых способ печати, используемые средства оформления, шрифтовые гарнитуры и т.п.) и обработке их системами распознавания и индексирования текстов. Основным компонентом таких систем является лингвистическая база данных, ядро которой – лексическая система языка рассматриваемого исторического периода. Основная идея проекта: необходимо сформировать в электронной форме лексическое ядро языка коллекции документов.

Материалами для исследований выступают печатные источники XVIII – нач. XIX вв. В качестве основы ядра используется Словарь Академии Российской 1789-1794 гг., Российской 1789-1794 гг. (САР), содержащий более 200000 лексических единиц, который в 2001-2005 гг. был переиздан с использованием современных информационных технологий. В течение 2006-2008 гг. в рамках проекта РГНФ было создано электронное издание САР, содержащее БД объемом более 44 тысяч структурных единиц.

Основные задачи проекта:

1. Анализ печатных источников XVIII - нач. XIX вв. и выявление их палеографических и лексических характеристик.

2. Исследование современных технологий оцифровки книг, методик ввода и обработки текстов и изображений, систем оптического распознавания, систем информационного поиска и автоматического индексирования документов.

3. Создание лексического ядра коллекции документов XVIII - нач. XIX вв. на основе БД Словаря Академии Российской 1789-1794 гг.,

4. Разработка информационной технологии переиздания источников XVIII - нач. XIX вв. и исследование ее эффективности.

5. Подготовка учебно-научных материалов для исследования эффективности представленной технологии и подготовки ее отдельных компонентов.

В настоящее время работы по данному проекту поддерживаются советом по грантам Президента Российской Федерации: проект 2010-2011 «Разработка словарных компонентов интегрированной информационной технологии переиздания печатных источников XVIII – нач. XIX вв.». (грант № МК 3732.2010.9.)

Результаты проекта могут быть использованы для переиздания значительного массива конкретных источников XVIII – нач. XIX вв., разработки систем распознавания исторических текстов, решения практических задач электронного и полиграфического издания древних памятников.

Руководитель проекта: Анна Ю. Филиппович, к.т.н., доцент.

Сайт проекта: http://philippovich.ru/Projects/DicXVIII/DicXVIIImain.htm

Компьютерная Семиография

Музыкальные произведения XI-XVI вв. писались с помощью специальной музыкальной системы (нотации), которую принято называть знаменной или семиографической. Она представляла собой сложную систему знамен (семиографических знаков, крюков). Количество этих знамен более 300 и каждому из них соответствует определенная последовательность звуков различной длительности и высотности. В ходе последующих реформ музыкальных нотаций был утрачен «ключ» к их расшифровке, вместе с тем записи более поздних музыкальных произведений, позволяет анализировать семиографические песнопения более ранних периодов.

Московский государственный университет Печати совместно с Московский государственный технический университет им. Н.Э. Баумана и Московская государственная консерватория им. П.И. Чайковского более 10 лет назад организовали проект «Компьютерная семиография», который направлен на многостороннее исследование музыкальной системы знаменных песнопений.

В рамках проекта создается информационная система, которая позволит представить специалистам и заинтересованным пользователям Интернет необходимые ресурсы и средства автоматизации по электронному ведению рукописей, интерактивному воспроизведению песнопений, их синтаксического и семантического анализа, генерации и анализу вариантов расшифровки (перевода в современную нотацию).

В проекте принимают активное участие студенты и аспиранты, которые на сегодняшний день разработали специализированные компьютерные шрифты и с их помощью осуществили ввод "Круга церковного древнего знаменного пения в шести частях" (под редакцией Д.В. Разумовского)" - фундаментального собрания певческих книг русской церкви, полностью нотированных знаменами и содержащего в шести томах свыше 1500 песнопений; Сборник попевок Соловецкого собрания, двузнаменники, некоторые известные азбуки и другие значимые рукописи.

На сегодняшний день созданы и развиваются компьютерные программы, которые позволяют создавать частотные словники, знаменники и конкордансы различной размерности; библиотеки азбук для расшифровки; попевочный фонд и другие лингвосемиотические конструкции. Разработаны уникальные технологии для статистического анализа, выявления синтаксической и семантической структуры семиографических песнопений; интерактивный музыкальный проигрыватель в веб-среде и др.

По результатам проекта подготовлено около десяти бакалаврских и магистерских работ, свыше 20 научных публикаций. В 2010 году проект «Компьютерная семиография»

стал «лучшим IT-проектом по сохранению культурных ценностей» в конкурсе IT ПРОРЫВ-2010, организованного партией «Единая Россия» совместно с государственной корпорацией «Ростехнологии» и компанией Softline. В 2011 году проект получил грант Российского гуманитарного научного фонда.

Работы по данному проекту выполняются коллективом авторов, в числе которых Андрей Ю. Филиппович, к.т.н., доцент – руководитель проекта; Б.Г. Смоляков, музыковед, профессор – главный научный консультант; М.В. Даньшина, И.В. Даньшина, С.А. Писарев, Е.А. Выломова.

Сайт проекта: http://philippovich.ru/Projects/Semio/SemioMain.htm

Компьютерный фонд древнерусской скорописи.

Данный проект посвящен созданию электронного фонда древнерусской скорописи и разработке информационной технологии распознавания скорописных древнерусских текстов и документов.

Скоропись в русских документах появляется с XIV в. и к XVII в. становится основным почерком деловых документов. Скоропись – это почерк, рассчитанный на существенное ускорение процесса письма. Ее отличают более свободные взмахи, росчерки, большое разнообразие графических вариантов отдельных букв.

Сложность создания электронного фонда древнерусской скорописи основана на ограничении круга людей, способных к чтению скорописных текстов, трудоёмкостью ручного перевода рукописей в электронное представление и отсутствии современных средств распознавания, работающих с древнерусской скорописью.

Материалы проекта – скорописные книги XVII в.

1. Книга отводная Карельского села Онежского Крестного монастыря приказчика старца Тихона старцу Иринарху (РГАДА, фонд 1195. Оп.1, ед.хр. 34, л.1–12. 3 октября 169 г. = 1660 г.).
2. Книги отводные Онежского Крестного монастыря казначея старца Иринарха Каменева новому казначею старцу Игнатию (РГАДА, фонд 1195. Оп.1, ед.хр. 86, л.1–26. 1 апреля 173 г. = 1665 г.)
3. Книга записная властелинским указам Онежского Крестного монастыря (РГАДА, фонд 1195. Оп.1, ед.хр. 387, л.1–12об. 1 января 195 г. = 1687 г.).
4. Отводная книга Белого двора Онежского Крестного монастыря (РГАДА, фонд 1195. Оп.1, ед.хр. 412, л.1–16об. 6 апреля 196 г. = 1688 г.).

Основные задачи проекта:

1. Разработка концепции и архитектуры компьютерного фонда древнерусской скорописи.

2. Разработка информационной технологии автоматизированного перевода (распознавания) скорописных древнерусских текстов и документов из растровых изображений в вид электронную форму.

3. Выявление особенностей древнерусской скорописи.

4. Исследование и разработка методов, алгоритмов, процесса распознавания; проектирование и реализация информационной технологии.

5. Создание автоматизированной системы распознавания древнерусской скорописи (АСРДС).

В рамках данного проекта введены и переведены в электронный вид более 300 скорописных листов. Создана графическая база данных, включающая изображения слов, отдельных букв и страниц текста.

Разработана теоретическая модель и алгоритмы распознавания скорописных текстов, ключевыми аспектами которых являются: структурный подход; двухуровневая схема распознавания (слово-буква); распознавание, управляемое гипотезами; нечёткость в описании; использование базы знаний и участие эксперта.

Создан макет системы распознавания, основные компоненты которой: *трассировщик*; *распознаватель*, состоящий из *распознавателей слов и букв; база знаний*, содержащая информацию о буквах и словах; *модуль обучения*.

Руководитель проекта: Ю.Н.Филиппович, к.т.н., профессор; разработчик системы распознавания: И.А. Зеленцов.

Сайт проекта: http://philippovich.ru/Projects/Skoropis/Skoropismain.htm

Заключение

Работа над проектом «Информационные технологии сохранения культурно-исторического наследия России» выполняется созданными коллективами из числа студентов и аспирантов в МГУП им. Ивана Федорова на кафедре Медиасистем и технологий творческой мастерской «Компьютерная лингвистика и семиотика» и в МГТУ им. Н.Э.Баумана на кафедре Систем обработки информации и управления. Материалы проекта используется в учебном процессе, ведутся научные работы, проводятся занятия, готовятся к выходу учебные пособия, основанные на материалах проекта.