Баранов В.А., Вотинцев А.А., Гнутиков Р.М., Зуга О.В., Миронов А.Н., Никифорова С.А., Ощепков С.В., Романенко В.А, Рябова Е.В. (Ижевск)

ИНФОРМАЦИОННО–ПОИСКОВАЯ СИСТЕМА "МАНУСКРИПТ": НОВЫЕ ВОЗМОЖНОСТИ ЭЛЕКТРОННОГО ИЗДАНИЯ ДРЕВНЕРУССКИХ РУКОПИСЕЙ [1]

1. Информационно-поисковая система "Манускрипт" представляет собой программный комплекс, состоящий из: 1) базы данных структурированной текстовой информации, 2) специализированного редактора для ввода и редактирования текстов, хранящихся в базе данных, 3) нескольких Web-сайтов, публикующих тексты в различных представлениях, и некоторых других специализированных модулей. ИПС предназначена для ввода, редактирования, хранения и обработки древнерусских текстов любой графической и структурной сложности и для получения материалов, необходимых для научных исследований.

2. ИПС "Манускрипт" дает возможность пользователю после фрагментирования древних славянских рукописей/текстов с учетом параметров и значений явно представленных и формально отсутствующих составляющих (авторство, создатели, время, оформление, функциональная, тематическая, композиционная и другие структуры) проводить лингвистические, текстологические, археографические, исторические и другие исследования. Список исследуемых объектов в системе может быть дополнен новыми типами единиц. В процессе работы с ИПС пользователь может расширить этот список, выделив интересующую единицу или группу единиц, обладающих определенными свойствами, в новый тип. Последующие упорядочивание, выборки, организация запросов могут производиться как на основе формальных свойств единиц, так и на основе их характеристик, существующих в базе как свойства и значения. Примеры доступа к базе данных через Интернет: http://io.udsu.ru/ptm/ (Путятина минея, XI в.) и http://io.udsu.ru/pev/ (Пантелеймоново Евангелие, XII–XIII вв.).

3. В системе реализуется возможность ведения различных словарей. Элементом словаря может являться любая единица. Применение словарей призвано значительно снизить трудоемкость описания единиц и работы с ними, уменьшить дублирование информации. Например, единица "словоформа" может наследовать часть характеристик единицы-элемента словаря начальных форм, в частности, те характеристики, которые не указаны для нее явно. Словари позволяют рассматривать тексты как реализацию неких инвариантов единиц различных типов: инвариантов лингвистических единиц, инвариантов текстологических единиц (погодные записи летописей, главы и стихи евангельских текстов и т. п.). Инвариант может существовать как в виде описания идентичных свойств и значений некоторых текстовых вариантов множества единиц, так и в виде реконструированного инварианта (архетипа).

[1] Работа осуществляется при поддержке РФФИ (проект № 02-07-90424в), Минобразования (научная программа "Университеты России", проект № ур.10.01.042) и гранта Президента РФ (№ МК-1742.2004.6).

Последний вариант словарей позволяет выявлять разночтения как в пределах одной рукописи (при наличии повторяющихся единиц), так и между рукописями, содержащими аналогичные фрагменты.

4. Между выявленными составляющими текста (единицами) может существовать множество разнообразных связей. Возможность описания и исследования этих связей является одной из основных задач, решаемых ИПС. Множество типов связей в существующей модели определяется пользователем в соответствии с поставленными исследовательскими задачами. Связь определяется количеством единиц (концов связи) и характером вхождения единицы в связь (типом конца связи); связь, в свою очередь, может обладать и собственными характеристиками. Примеры таких связей: связь вхождения, связь следования, связь с элементом словаря.

5. Необходимость хранения в многотекстовой базе специфичных для древнеславянских языков символов и включений фрагментов на других языках привела к необходимости разработки собственного набора символов в стандарте UNICODE.

Кодировочно-шрифтовая система "Манускрипт" включает в себя набор символов UTF8LAPREXT1 и семейство шрифтов Menaion. Набор символов обеспечивает хранение, сортировку и необходимые преобразования всех применяемых в многотекстовой базе символов. Воспроизведение вариантов начертаний одного и того же знака (вплоть до индивидуальных особенностей почерка писцов) осуществляется с помощью необходимых шрифтов. Предложенная система классификации вариантов начертания символов позволяет последовательно отнести символ к одной из базовых групп основных символов в КШС, определить его кодовое значение и, наконец, определить конкретный шрифт, которым должен быть отображен данный символ.

6. Для обеспечения режимов ввода и корректировки данных ИПС создан специализированный текстовый редактор "Манускрипт", который позволяет пользователю эффективно работать с визуализированными данными текстов/рукописей – единицами, связями, их свойствами и значениями.

Редактор позволяет отображать существующие в рукописи/тексте иерархии и текст в "плоском" виде, который представляет собой преобразованную геометрическую иерархию. В "плоском" режиме редактор позволяет добавлять знаки в текст, удалять их, разбивать на фрагменты. В режиме просмотра иерархий редактор отображает информацию о связях между единицами текста, структурными и словарными единицами; позволяет устанавливать или удалять связи между единицами; просматривать и корректировать их свойства, создавать новые единицы (в том числе и тексты).

В связи с необходимостью создания средств доступа к источнику данных в виде отдельного компонента первоначальная традиционная архитектура редактора "клиент-сервер" заменяется на трехуровневую: клиент – сервер приложений – сервер базы данных.

Доступ к базе данных организован с помощью технологии ADO (ActiveX Data Objects), что позволяет в достаточной степени абстрагироваться от источника данных и, например, получать доступ к данным, хранящимся в различных СУБД и в различных форматах (XML). ADO также поддерживает модель работы briefcase, не требующую постоянного соединения с базой

данных, что важно при обеспечении доступа через Интернет. Возможен доступ удаленных клиентов из специализированного редактора к текстовой базе через Интернет, в том числе и с использованием коммутируемых каналов связи (по модему).

7. Основой для многопользовательской работы является открытая архитектура ИПС и применяемая для хранения данных СУБД Oracle. Предусмотрена возможность организации многопользовательского доступа к текстам как для просмотра (с помощью Web-сайтов), так и для корректировки средствами редактора "Манускрипт".