Редактор OldEd как специализированный инструмент для редактирования документов в базе данных «Манускрипт»[1]

Р. М. Гнутиков, В. А. Баранов
Удмуртский государственный университет, Ижевск, Россия

In the process of work with the ancient texts a comprehensive study of ancient texts gains in importance. Any research undertaken needs to a considerable extent conclusions of scientists from various fields.

To ensure the comprehensive study of unique manuscripts and succession in the work of various experts, the text/manuscript should exist in the form that is maximum close to the original as a set of minimal structural components (symbols) that are grouped into larger units — word forms, fragments, sections etc. that are assigned values proper to them. As mentioned above this can be achieved with the help of full-text databases.

To process data of the full-text database, it makes sense to create specialized editors. The editor of this type is a module of the system "Manuscript" that provides access to the databases.

При обращении ученых к уникальным по своей культурной и исторической значимости древним текстам важнейшее значение приобретает их комплексное изучение. Любое из предпринятых исследований, будь то текстологическое, палеографическое, фонетическое, грамматическое, лексическое, литературоведческое, культурологическое или иное, в значительной мере нуждается в использовании выводов ученых разных специальностей.

Так, при лингвистическом анализе древнейших и средневековых памятников для получения объективных выводов оказывается важным учитывать состав и структуру рукописей и входящих в них текстов, авторство и жанр последних, переводность и ориги-

[1] Работа по созданию ИПС «Манускрипт» ведется при поддержке Российского фонда фундаментальных исследований (грант № 05-07-90217в).

нальность разделов и многие другие факторы. Текстологические исследования являются в данном случае необходимым этапом при изучении рукописных памятников. Практически же применение каждым следующим исследователем уже имеющихся сведений вновь начинается с чтения, фрагментирования и членения текста и выборки анализируемых единиц в соответствии с выдвигаемой гипотезой, а каждый новый поворот исследования требует перегруппировки данных в зависимости от необходимости учета или игнорирования тех или иных параметров текста. В любом случае практическое применение имеющихся текстологических данных представляет собой достаточно трудоемкое занятие.

Для обеспечения комплексного исследования уникальных рукописных памятников и преемственности в работе различных специалистов текст (рукопись) должен существовать в максимально приближенной к оригиналу форме в виде набора минимальных структурных составляющих — знаков, которые группируются в более крупные единицы — словоформы, фрагменты, разделы и т. п. (иногда нелинейного состава), которым присваиваются свойственные им значения и по аналогии с которыми могут быть созданы другие единицы, отсутствующие в самом тексте, но связанные с фактически существующими (например, некие протофрагменты, по отношению к которым текстовые фрагменты могут быть исследованы с точки зрения разночтений, нормализованный грамматический или современный эквивалент текста и т. п.).

Понятно, что на всем протяжении работы к электронному образу рукописи должен быть обеспечен доступ, в какое бы время она ни осуществлялась и где бы текст ни хранился. Кроме того, работа пользователя с электронным документом может быть полноценной и эффективной в том случае, если набор, редактирование, фрагментирование, присвоение значений, выборка, группировка, упорядочивание текстовой информации осуществляется с помощью дружественного интерфейса, который вне зависимости от сложности структуры текста и связей между единицами представляет пользователю достаточно привычный образ рукописи.

Все сказанное может быть реализовано с помощью полнотекстовых баз данных.

Полнотекстовая база данных ИПС «Манускрипт»[2], в которой хранятся описания текстов и рукописей, организована в соответствии со специально разработанной моделью. Модель позволяет хранить и описывать практически любые объекты, которые интересуют или могут заинтересовать исследователя. Так, объектами-единицами являются *тексты*, *текстовые словоформы*, *морфемы*, *пунктуационные знаки*, *синтагмы*, *предложения*, *фрагменты*, выделенные на определенном основании, и др. Любая единица может обладать уникальным набором характеристик, определяемых ее типом. В ИПС отсутствуют какие-либо ограничения на список исследуемых объектов. Возможно добавление новых интересующих объектов, например, выделение единицы, обладающей определенными свойствами, в новый тип.

Единицы, формально представленные в тексте в виде знака или в виде набора расположенных в определенном порядке знаков, обладают координатами и размерами (высота и ширина). Кроме того, существуют единицы, не имеющие явного графического представления в тексте рукописи, это *начальные формы*, *морфологические*, *синтаксические*, *семантические* и аналогичные признаки; *эквиваленты текстовых единиц* и т. п.

Между единицами существуют различного рода связи. Совокупность единиц и связей между ними представляет собой *сеть единиц*. В сети единиц на основании свойств определенной предметной области выделяются *подсети*. Подсети имеют структуру иерархической подчиненности, иначе, являются *иерархиями*. Например, в иерархии геометрической подчиненности единицей самого нижнего уровня является «знак», а единицей самого высокого уровня — «рукопись».

Приведем примеры иерархий.

– *Геометрическая*: рукопись состоит из листов, листы из страниц, страницы из слоев, слои из строк, строки из знаков.

[2] Баранов и др. 2002 — *Баранов, В. А.* Структура и функции информационно-поисковой системы «Манускрипт» / В. А. Баранов, А. А. Вотинцев, Р. М. Гнутиков, О. В. Зуга, А. Н. Миронов [и др.] // Информационный бюллетень Ассоциации «История и компьютер». № 30 : Специальный выпуск : материалы XIII конф. Ассоциации АИК (Санкт-Петербург, 26–29 июня 2002 г.) — М., 2002. — С. 87–89.

– *Лингвистическая*: текст состоит из синтагм, словоформ и знаков, синтагмы состоят из словоформ и знаков, словоформы состоят из знаков.

– *Функционально-структурная*: текст состоит из разделов, которые содержат подразделы и знаки. Пример: канон состоит из заголовка и нескольких песен; песня разделяется на заголовок, тропарь (один, или несколько), канон, богородичен, ирмос.

Выявление иерархий нового типа не вызывает принципиальных проблем.

Для работы с данными, хранящимися в полнотекстовых базах данных, целесообразно создание специализированных редакторов. Разработка и создание подобного редактора в настоящее время осуществляется творческим коллективом Лаборатории по автоматизации филологических исследований Удмуртского госуниверситета. Редактор является модулем системы «Манускрипт» и обеспечивает доступ к базам данных.

Редактор реализует следующие ключевые функции: (a) набор и редактирование текста и дополнительной информации о нем при непосредственном взаимодействии с базой данных; (b) представление текста рукописи в близком к оригиналу виде; (c) выделение и создание единиц текста и работа с их свойствами и значениями; (d) работа с взаимосвязями единиц (создание, переподчинение, удаление, изменение свойств и типов связей, работа с иерархическими структурами); (e) поддержка многопользовательского режима работы.

Доступ к базе данных организован с помощью технологии ADO (ActiveX Data Objects). Это позволяет в достаточной степени абстрагироваться от источника данных, например, получать доступ к различным версиям СУБД Oracle (Oracel 8, Oracle 9), файлам XML. ADO также поддерживает модель работы briefcase, что подразумевает работу с базой данных без поддержки постоянного соединения.

Возможности провайдера для Oracle позволяют посредством редактора получить доступ к текстовой базе удаленным клиентам через Интернет, в том числе с использованием модемной связи. К сожалению, это требует достаточно широкого канала и постоянного соединения. В настоящее время ведется работа по совершенствованию редактора.