От оцифрованных коллекций средневековых рукописей к электронным многофункциональным интернет-библиотекам[1]

В. А. Баранов
Ижевский государственный технический университет, Россия

The paper is devoted to the possibilities of the Manuscript system (http://manuscripts.ru/) designed for preparation of electronic scientific publications of ancient manuscripts on the Internet.

The primary consideration is given to the specialized modules of the system ensuring 1) input, storage, editing and processing of materials in the database, 2) textologic, linguistic and paleographic analyses of manuscripts/texts and 3) preparation of dummy copies and publication of manuscripts and research apparatus. All modules interact with a common database allowing processing text/manuscript units organized into hierarchies and nets, their relationships and values that adequately reflect modeled objects and their relationships.

1. В настоящее время полнотекстовые библиотеки развиваются в Интернете в нескольких направлениях: 1) имеется большое количество коллекций, содержащих оцифрованные страницы печатных изданий и рукописей, 2) еще более широко представлены транскрипции современных текстов; в то же время 3) единичны примеры многофункциональных web-модулей транскрипций древних и средневековых текстов.

Объективная сложность создания библиотек последнего типа обусловлена собственно материалом — чрезвычайно сложными по структуре и составу документами, а также высокими требованиями, которые предъявляются к таким библиотекам: они должны предоставлять возможность комплексного использования данных — позволять осуществлять поиск и выборку объектов библиотеки (не только текстов и рукописей, но и их структурных и семантических единиц — фрагментов, выделенных на любых значимых для описания, исследования и демонстрации текста/рукописи основаниях), накопление, упорядочивание и сравнение найденного материала.

[1] Работа по созданию ИПС «Манускрипт» ведется при поддержке Российского фонда фундаментальных исследований (грант № 05-07-90217в), работа по созданию автоматизированного морфологического анализатора — при поддержке Российского гуманитарного научного фонда (грант № 05-04-12408в).

Требования к электронным полнотекстовым библиотекам средневековых рукописей с точки зрения пользователя можно сформулировать следующим образом:

– навигация по коллекции с использованием метаданных документов,

– навигация по фрагментам рукописей/текстов,

– навигация по нескольким рукописям одного текста,

– выборка текстовых (например, лингвистических) единиц, в том числе составных, организованных с помощью связей,

– упорядочивание и сохранение выборок,

– сравнение аналогичных выборок между собой,

– знакомство с упрощенными видами средневековых текстов,

– просмотр оцифрованных изображений страниц и некоторые другие.

Требования к инструментальным средствам создания таких библиотек с точки зрения авторов коллекций:

– распределенный и удаленный ввод, редактирование и анализ данных,

– автоматические и автоматизированные средства трансформации текстовых единиц,

– использование авторитетных файлов (словарей) для описания, фрагментирования и анализа данных,

– передача/конвертирование данных, созданных в различных форматах, при распределенной работе над библиотекой,

– возможности для альтернативного описания, комментирования и фрагментирования данных различными исследователями,

– указание, редактирование соответствий единиц одного текста в разных рукописях и навигация по этим связям,

– публикация подготовленных транскрипций текстов и рукописей, словарей и комментариев в Интернет,

– защита данных от несанкционированного доступа.

Одним из наиболее эффективных инструментов для решения перечисленных задач являются специализированные базы данных, содержащие структурированную информацию о глубоко фрагментированных текстах, рукописях и их объектах. Полнотекстовая база данных (a) должна поддерживать многошрифтовость[2], многоязычность[3], многотекстовость[4], (b) должна обеспечивать хранение метаданных[5], аналитических[6], лингвистических[7] данных, комментирующей и справочной информации, (c) должна иметь средства работы со словарями[8], со связями единиц[9] и с графическими объектами[10].

В электронной библиотеке, основанной на базе данных, указанные конструктивные особенности должны использоваться (a) при организации запросов и (b) осуществлении поиска, (c) при

[2] Многошрифтовость — использование символов различных национальных алфавитов.

[3] Многоязычность — обеспечение хранения информации о единицах (свойства, значения, комментарии) и наличие интерфейса на различных языках.

[4] Многотекстовость — возможность создания запроса с указанием текстов/рукописей, в которых будет осуществляться поиск; выборка единиц из всего корпуса рукописей/текстов или из отобранной пользователем части корпуса.

[5] Метаданные — свойства и значения рукописей/текстов. Метаданные должны обеспечивать (a) создание запроса с указанием свойств и значений текстов/рукописей, в которых будет осуществляться поиск, а также (b) выборку единиц из рукописей/текстов, обладающих указанными в запросе значениями.

[6] Аналитические данные — свойства и значения фрагментов рукописей/текстов. Анатические данные должны обеспечивать возможность создания запроса с указанием свойств и значений фрагментов, в которых будет осуществляться поиск, а также выборку единиц из фрагментов, обладающих указанными в запросе значениями.

[7] Лингвистические данные — информация о форме, свойствах и значениях лингвистических единиц текстов.

[8] Словари — компоненты базы данных, содержащие информацию о форме, свойствах и значениях словарных единиц.

[9] Связи — информация об отношениях между объектами текстов/рукописей и их единицами, а также о свойствах и значениях отношений.

[10] Графические файлы используются для хранения скан-копий страниц рукописей и графических элементов текста.

визуализации результата выборки и (d) упорядочивании данных, (e) при создании сравнительных перечней и (f) критического сравнения рукописей одного текста, (g) при создании критического аппарата издания и (h) воспроизведении разных видов текста.

2.1. Полнотекстовая база данных «Манускрипт» и ее специализированные модули, разрабатываемые в Удмуртском государственном университете и Ижевском государственном техническом университете, представляют собой единую информационно-поисковую систему (ИПС) и предназначены для ввода, редактирования, хранения, обработки сложных по структуре и составу документов, а также для научных исследований и создания электронных публикаций древних рукописных памятников в сети Интернет.

ИПС «Манускрипт» состоит из нескольких разнофункциональных модулей, взаимодействующих с единой базой данных, в которой хранятся иерархически структурированные данные:

– специализированный редактор,

– модуль запросов и выборок,

– модуль электронных изданий,

– модуль лингвистических словарей,

– модуль загрузки данных из внешних источников, имеющих иной формат,

– модуль выгрузки текстов и справочных материалов для печатных публикаций,

– модуль хранения и обмена документами.

Сегодня модули системы «Манускрипт» связаны между собой в единую технологическую цепочку, которая обеспечивает подготовку электронной коллекции, удовлетворяющей многим из выше названных требований, актуальных для лингвистов, текстологов, историков, археографов и ученых других специальностей.

2.2. Работа над электронной коллекцией или над отдельной публикацией состоит из нескольких этапов:

– *подготовительный*, включающий палеографическое, орфографическое, археографическое, текстологическое изучение рукописи и текста;

– *этап ввода, сверки и редактирования*, включающий подготовку транскрипции текста, сверку с оригиналом, правку текста;

– *этап подготовки справочного аппарата*: фрагментирование текста/рукописи, анализ единиц текста, установление необходимых связей между единицами, подготовка словарей, индексов, инципитов и др.;

– *этап разработки и создания веб-приложений доступа к данным*: создание страниц публикации, справочных и комментирующих материалов, создание запросных форм, создание форм вывода результатов запроса;

– *этап обеспечения использования коллекции* в ходе научных исследований.

Существующие в настоящее время компоненты ИПС «Манускрипт» обеспечивают все этапы работы над коллекцией и с коллекцией рукописных памятников.

2.3. *Идеология модели базы данных.* Данные хранятся в виде иерархически связанных объектов, принадлежащих одной области: это, например, объекты, описывающие расположение единиц в рукописи — лист, страница, столбец, строка, это текстологические единицы — фрагмент, созданный в иное, чем вся рукопись время, фрагмент, созданный иным писцом, это лингвистические единицы — словоформа, синтагма, фраза, текст и др.

Минимальной единицей иерархии всегда является символ, максимальной — текст или рукопись. Наиболее сложной является лингвистическая иерархия, которая включает объекты, необходимые для адекватного реальному положению вещей представления структуры и семантики текста, для описания отношений между единицами текста и для работы с объектами и отношениями. Структурные отношения единиц описываются связями в виде дерева, семантические — связями, образующими сеть.

2.4. *Специализированный редактор OldEd*. Редактор предназначен: (a) для ввода и редактирования единиц рукописи, (b) для фрагментирования рукописи и текста на единицы, (c) для установления между единицами иерархических и иных связей, (d) для присвоения единицам свойств и значений, которыми они обладают, (e) для установления связей единиц рукописей с единицами словарей и для других операций.

Специально разработанная кодировочно-шрифтовая система позволяет максимально приблизить транскрипцию текста к оригиналу и передать в публикации любые графические варианты символов.

Редактор обеспечивает фрагментирование рукописей и текстов и присвоение выделенным объектам значений. Если фрагмент представляет собой вариант объекта, который существует или может существовать во многих рукописях или текстах, с помощью редактора можно связать фрагмент с его инвариантом, хранящимся в соответствующем словаре.

Одной из важных операций редактора является макетирование будущего электронного издания. Оно позволяет расположить фрагменты текста на экране в соответствии с оригиналом, что обеспечивает соответствующую визуализацию страницы в электронной публикации на сайте.

Обеспечение редактором распределенной и дистантной работы над текстами, хранящимися в одной базе данных, дает возможность коллективу авторов увидеть и оценить результаты непосредственно в ходе редактирования.

2.5. *Web-модуль электронных публикаций.* Основным компонентом электронного издания является страница запроса, которая позволяет выбрать коллекцию рукописей, текст коллекции и указать критерии запроса. Результатом запроса могут быть (a) различные перечни: прямые и инверсированные указатели начальных форм и словоформ, количественные указатели, указатели инципитов; (b) тексты различного вида — оригинальный, нормализованный, дипломатический, транслитерированный. Основным среди лингвистических указателей является полный указатель слов и словоформ, в котором даются грамматические признаки единиц, их количество и адреса.

2.6. *Модуль выборок и запросов.* Если в предыдущем модуле возможности запроса предопределены и пользователь не может их изменить или расширить, то модуль выборок и запросов предназначен для формирования достаточно свободного запроса и позволяет (a) выбрать единицы, их свойства и значения; (b) указать отношения между единицами; (c) указать состав результата запроса, приоритет его единиц и их сортировки, (d) вывести выборку на экран, (e) осуществить операции над выборками, (f) использовать результат запроса для следующей выборки.

В целом модуль выборок и запросов ориентирован на предоставление пользователю гибких возможностей для формирования запросов на основе любых имеющихся в базе данных свойств и значений единиц и их связей.

2.7. *Модуль словарей.* Центральным модулем системы «Манускрипт» является разрабатываемый модуль словарей. Модуль должен обеспечить использование единых словарей при фрагментировании и разборе текстов и рукописей.

Наиболее сложной по исполнению является система лингвистических словарей, которая ориентирована на автоматизированный морфологический анализ древних и средневековых славянских текстов. В основе грамматических словарей лежит принцип оперирования изменяемыми и неизменяемыми частями словоформы — окончаниями и основами. В системе «Манускрипт» используются два известных в настоящее время подхода к делению словоформы на эти компоненты: с сохранением в основе чередований (нормализованные основы и окончания) и без сохранения чередований (псевдоосновы и псевдоокончания). Это сделано для того, чтобы комплексно решить главную задачу автоматического морфологизатора — задачу разбора очень сильно варьирующейся графической и морфологической структуры древней славянской словоформы.

Основными возможностями модуля в настоящее время являются (a) хранение элементов словарей, связей, а также их свойств и значений, (b) редактирование элементов словарей, (c) установление связей между элементами одного словаря и элементами различных словарей, (d) установление связи между основами и их окончаниями, (e) построение парадигм на основе леммы.

3. Таким образом, в настоящее время ИПС «Манускрипт», имея развитые средства для ввода, редактирования, разбора, макетирования и публикации сложных по структуре и составу документов, а также для создания запросов и работы с выборками, удовлетворяет многим требованиям, предъявляемым к многофункциональным полнотекстовым исследовательским системам, а использование системы при лингвистических и лингвотекстологических исследованиях позволило получить материал для анализа древних славянских текстов, их фрагментов и сделать значимые фундаментальные выводы.

Описание проблемной области в интеллектуальных информационных технологиях

И. А. Барков
Ижевский государственный технический университет, Россия

The main principles of creating and using the semantic describing of the problem sphere in the intellectual informational technologies are considered.

Введение

Основной тенденцией создания современных информационных технологий, в том числе и информационных технологий обработки текстов, является попытка придать ей признаки интеллектуальности. Под интеллектуальностью в этом случае понимается обеспечение автоматизированной системы возможностями обработки смысловой, семантической информации. Традиционно в интеллектуальных информационных технологиях смысловая составляющая информация фиксируется в модели предметной области, а для реализации процессов обработки используется логический вывод. В общем случае под предметной областью памятника письменного наследия будем понимать смысловое описание изложенной в памятнике темы. Как правило, созданием описания предметной области занимается эксперт. Современный уровень развития научного и прикладного знания характерен развитой системой наук, поэтому задача описания предметной области является многопрофессиональной: один и тот же объект или явление в глазах различных специалистов получает совершенно различное смысловое описание. Различие смысловой интерпретации окружающего мира привело к дифференциации наук, усиливающейся с течением времени. Используя понятие семантического треугольника можно представить задачу описания проблемной области следующим образом (рис. 1).