Интернет-средства поиска и визуализации данных для лингвистического анализа информационно-аналитической системы "Манускрипт"

А. А. Вотинцев, П. А. Вотинцев, И. С. Соломенников, В. А. Баранов
Удмуртский государственный университет, ООО "Виктори", Удмуртский государственный университет, Ижевский государственный технический университет, Россия

1.1. Начало работы над информационно-аналитической системой «Манускрипт» (ИАС «Манускрипт») было положено в середине 90-х годов, когда на кафедре русского языка Удмуртского государственного университета под руководством профессора Виталия Михайловича Маркова возобновилась работа над словарем языка М. В. Ломоносова. Тогда с помощью созданных программистами лаборатории по автоматизации филологических исследований программ были получены конкордансы писем и поэзии Ломоносова. Одновременно началась работа сначала над печатным изданием Путятиной минеи (РНБ, Соф. 202), а потом и над ее электронной публикацией. То внимание, которое с самого начала было уделено именно тексту, его структуре и лингвистическим составляющим, а не его метаописанию и комментированию, определило принципиальные подходы, которые были положены в основу модели базы данных «Манускрипт», и технологические решения, реализованные в первом интернет-проекте «Путятина минея» (URL сегодня: http://manuscripts.ru/mns/portal.main?p1=19&p_lid=1).

С самого начала интернет-модули системы «Манускрипт» уже были не столько демонстрационными, сколько аналитическими, позволяющими пользователю получать необходимые для исследования текстов данные – перечни словоформ, их количество, конкордансы. Особое внимание было уделено точности передачи рукописей и их единиц и возможностям их трансформации, например визуализации словоформ в более простом виде, чем они хранятся в базе данных (далее – БД).

1.2. В настоящее время на портале «Манускрипт: славянское письменное наследие» (URL: http://manuscripts.ru/) представлено уже несколько десятков текстов XI–XIV вв. и более позднего периода. Кроме

того, на портале представлена коллекция произведений М. В. Ломоносова, которая активно пополняется. Понятно, что значительное увеличение количества текстов поставило новые задачи – задачу организации навигации по корпусу и задачу разработки новых форм и видов демонстрации лингвистических данных.

Принципиальным изменением в подходе к представлению данных является переход от однотекстовых интернет-публикаций к многотекстовым. В настоящее время реализован двушаговый алгоритм, в результате которого пользователь имеет возможность получить так называемые сравнительные справочные материалы (указатели, перечни) по тому количеству текстов, рукописей, фрагментов, которые отобраны для анализа. На первом шаге на основе мета- и аналитических данных осуществляется отбор объектов – текстов / рукописей / фрагментов, на основе которых будут построены сравнительные перечни. На втором – указывается форма представления выборки – текстовая, списочная или табличная, маска лингвистических единиц и некоторые другие параметры запроса.

2. На сегодняшний день ИАС «Манускрипт» представляет собой комплекс взаимосвязанных программных средств, центральным звеном которого является хранилище данных (база данных). Система имеет средства загрузки, ввода, корректировки, анализа и обработки данных, а также средства отображения текстов сложной структуры, реализованные как отдельные интернет-модули. Пользователям Интернета для работы предоставляются модули однотекстовых интернет-публикаций, модуль запросов и выборок и модули морфологического анализатора.

3. Представляемая в данной работе подсистема многотекстового поиска и визуализации документов и их объектов также является интернет-модулем ИАС «Манускрипт» (URL: http://manuscripts.ru/mns/srch.simple).

Шаг 1. Отбор объектов.

Для отбора рукописей / текстов / фрагментов разработаны два поисковых интерфейса: простой и расширенный. Их различия – в

возможности конструировать сложные поисковые вопросы с учетом мета- и аналитических данных.

Интерфейс обычного поиска позволяет задавать текстовую маску, поиск которой осуществляется затем среди множества свойств текстовых объектов (рукописей, текстов, фрагментов) базы данных. Предусмотрена возможность повторного поиска в результатах предыдущего поискового запроса.

Форма расширенного поиска рассчитана на пользователей, желающих получить выборку по более конкретным критериям, чем просто текстовая маска. Отличие ее от формы простого поиска в том, что в ней запрос конструируется в виде логического выражения, операндом которого является критерий отбора. Каждый критерий отбора представляет собой сочетание трех элементов: типа объекта для поиска (рукопись, либо текст, либо фрагмент), свойства данного объекта (например, автор текста, место издания рукописи и др.) и текстовой маски для поиска в выбранном свойстве. Примером такого запроса является: Тип объекта = *Текст*, Свойство = *Автор*, Значение = *Мичка*. Данный запрос вернет все тексты, написанные Мичкой.

Как и в классической логике, критерии между собой сочетаются логическими операторами *И* и *ИЛИ*. Также предусмотрена возможность ввода логического отрицания любого критерия. Используя предыдущий пример: Тип объекта = *Текст*, Свойство = *Автор*, Значение = *НЕ Мичка*. Такой запрос вернет набор текстов, написанных любым автором, но не Мичкой.

Для авторизованных в системе «Манускрипт» пользователей все полученные выборки объектов автоматически сохраняются для последующего их использования. Такой пользователь может в любое время просмотреть список сделанных им ранее выборок, удалить из него те выборки, которые ему больше не нужны.

Шаг 2. Визуализация данных.

В ходе разработки подсистемы было принято решение вынести все сведения о рукописях, текстах и их фрагментах и лингвистических единицах

в обособленную схему данных. Такое решение было обусловлено рядом причин, важнейшей из которых является обеспечение приемлемого для пользователей Интернета быстродействия системы.

Структуры основного хранилища данных ИАС «Манускрипт» ориентированы на хранение максимально детальной информации об объектах документов и не всегда позволяют организовать быстрый сбор информации для удобного представления пользователю. Но именно скорость реакции системы является критичным фактором для web-приложений. Поэтому страницы разработанного модуля формируются на основе вспомогательных структур данных, которые специальным образом спроектированы для целей быстрого построения страниц сайта без обращения к основному хранилищу системы. Данные вспомогательных структур готовятся заблаговременно на основе информации базы данных ИАС «Манускрипт» и на момент выполнения пользовательского запроса уже готовы к использованию. Процедура подготовки данных и их загрузка во вспомогательные структуры (денормализация данных) в системе получила название "публикация текста" и в различных вариантах доступна из редактора OldEd пользователям, имеющим права на редактирование текста.

Помимо прочего, хранение «строительных» данных сайта в отдельной схеме обеспечивает необходимый уровень безопасности: основное хранилище информации скрыто от несанкционированного внешнего доступа.

Минимальными единицами хранения в схеме данных сайта являются фрагменты словоформ и символы, не входящие в состав словоформ, что, с одной стороны, позволяет с достаточной производительностью реагировать на пользовательские запросы в объемах, представленных на сайте, а с другой – накладывает некоторые ограничения на возможности более детального исследования документов (например, исследователь лишен возможности выполнения запросов на основе знаков текста и их свойств). Схема сайта содержит не только информацию о минимальных составляющих текста, из которых может быть собран текст,

представляемый пользователю, но и описание свойств и взаимосвязей более крупных единиц, таких, как, например, фрагменты текста.

Таким образом, часть БД ИАС «Манускрипт», используемая в работе многотекстового модуля, содержит предварительно подготовленные данные в объеме, необходимом для формирования представленной на сайте информации в удобном и понятном для пользователя виде: фрагменты текста, конкордансы, указатели, списки.

4. Основными формами визуализации данных многотекстового модуля являются (а) текст(ы), (б) перечень (список) лингвистических единиц, (в) таблица.

Первый вид дает возможность познакомиться с неразделенными или разделенными на словоформы текстами. Второй и третий – позволяют получить прямой или инверсированный указатели, в которых адреса (аббревиатура текста, номер листа, страницы, строки) даются или в словарной статье, или в ячейках таблицы, каждый столбец которой соответствует тексту или фрагменту текста. В виде таблицы представлен и количественный указатель, который упорядочен по убыванию суммарного встречаемости словоформ и в котором значениями ячеек являются абсолютное количество употреблений словоформы в тексте или фрагменте.

Особым видом представления данных является показ фрагментов, обладающих идентичными значениями, например, являющихся стихами из разных евангельских списков. При наличии в выборке стихов с одним и тем же значением они располагаются рядом друг с другом, при разных значениях – друг за другом. Для каждого фрагмента указывается диапазон адресов (лист, страница, строка) его начала и конца.

Каждый из видов может быть ограничен дополнительными параметрами – диапазоном листов рукописи и/или маской словоформы. В последнем случае в списочный и табличный виды включают только искомые словоформы, а в текстовом представлении они выделяются цветом.

5. Большие сложности для реализации модуля вызвал поиск и выбор варианта отображения результатов многотекстовых запросов. Если упорядоченный список единиц с указанием адресов текстовых прецедентов является стандартным представлением данных, то табличное представление не часто используются для визуализации упорядоченных лингвистических данных. В то же время расположение материала в виде таблицы дает возможность наглядно представить соотношение распределения словоформ и/или их количества в отобранных текстах, дать дополнительные сведения в виде суммарного количества словоформ в прямом и инверсированном указателях. Именно таблица является сегодня базовым представлением данных в многотекстовом модуле.

6. Доработка модуля планируется в нескольких направлениях. По мере увеличения количества лемматизированных текстов будет вводиться показ словоуказателей и предоставляться поиск по грамматическим признакам лингвистических единиц. Увеличение круга фрагментированных текстов и типов фрагментов должно расширить возможности системы для лингвотекстологического анализа данных. Подключение справочников и лингвистических словарей системы должно дать возможность использовать при подготовке запросов структурированные значения единиц базы данных.

Благодарности

Работа по созданию многотекстового модуля выполняется в рамках проектов № 07-04-00369а (электронное издание майских служебных миней), № 07-04-12147в (создание коллекции М. В. Ломоносова в Интернете), поддержанных Российским гуманитарным научным фондом (РГНФ).

Список литературы

Baranov 2004 – *Baranov, Victor*. Old Slavic Manuscript Heritage: Electronic Publications and Full-Text Databases / Victor Baranov, Andrey Votintsev, Roman Gnutikov, Aleksey Mironov, Sergey Oshchepkov, Vitaliy Romanenko // EVA 2004 London (Electronic Imaging, the Visual Arts Conference & Beyond) : Conference Proceedings / University College

London. Institute of Archaeology ; principal Editor James Hemsley. – London, 2004. – P. 11.1-11.8.

Современные информационные технологии и письменное наследие 2006 – *Современные информационные технологии и письменное наследие*: от древних рукописей к электронным текстам : материалы междунар. науч. конф., Ижевск, 13–17 июля 2006 г. / Отв. ред. В. А. Баранов. — Ижевск : Изд-во ИжГТУ, 2006. — 193 с.

Baranov 2006b – *Baranov, Victor*. Information-Analytical System “Manuscript”: technologies and tools of creation of electronic collections of ancient and medieval documents [Электронный ресурс] / Victor Baranov // Dagstuhl Seminar Proceedings 06491: Digital Historical Corpora - Architecture, Annotation, and Retrieval / L. Burnard, M. Dobreva, N. Fuhr, A. Lüdeling; Dagstuhl Seminar 06491, 03.12. – 08.12.2006; Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany. – Режим доступа : http://drops.dagstuhl.de/portals/index.php?semnr=06491, свободный. – Загл. с титул. страницы.

Baranov 2007 – *Baranov, Victor*. The ideology and technology of creating online full-text digital collections of ancient and medieval slavonic manuscripts / Victor A. Baranov // International Conference on Applied Natural Sciences, Trnava, Slovakia, November 7-9, 2007. – Trnava, 2007. – P. 199-207.

Online search and visualization of data for linguistic analysis using the Manuscript informational-analytical system

Andrey A. Votintsev, Pavel A. Votintsev, Igor’ S. Solomennikov, Victor A. Baranov

Udmurtia State University, Victory LLC, Udmurtia State University, Izhevsk State Technical University, Russia

This paper discusses the experience of developing a means to represent data about linguistic units in full-text online databases of medieval Slavic texts. Attention is focused on procedures for querying manuscripts and their parts and

on forms of visualization of data - parts of manuscripts, fragments, and word forms.