на компьютер, на котором не было Siberia, шорские буквы пропадали.

6. В настоящее время разрабатывается и постоянно пополняется шрифт UNICODE, который в итоге должен содержать глифы символов любого языка мира. Шрифт UNICODE основан на четырехбайтовой кодировке, но это не значит, что каждая буква представляется четырьмя числами символов. Самые популярные, часто используемые в мире символы имеют меньше кодов для своей записи. Для латинского алфавита, как и прежде, используется однобайтовое кодирование, если его ASCII-код меньше 128. Затем идет большой раздел двухбайтового соответствия. В этом диапазоне находятся также полностью и русский, и шорский алфавиты.

Чтобы набрать букву этих алфавитов, надо знать соответствующие таблицы кодирования формата UTF8. Их можно узнать как в справочниках, так и в Интернете, например, по адресу http://titus.uni-frankfurt.de/unicode/unitest.htm. Каждый символ должен записываться двумя числами. Это можно делать с помощью цифровых клавиш на малой клавиатуре при нажатой клавише ALT. Например, букве «ғ» соответствует четырехзначный код 1171. Этот медленный способ и требует наличия на компьютере современного стандарта UNICODE UTF8, присутствующего в операционных системах Windows 2000, XP, 2003 и в их офисных программах.

Каждый из способов ввода символов имеет свои достоинства и недостатки, поэтому каждый пользовать решает, как вводить, редактировать и хранить национальный шорский текст[1].

[1] Соседко 2003 — Соседко, О. А. Особенности обработки шорских текстов / О. А. Соседко // Повышение качества профессиональной подготовки будущего учителя информатики : сб. науч. тр. — Новокузнецк : КузГПА, 2003.

Технологии создания информационной системы для работы с полнотекстовыми базами данных исторических документов[1]

В. О. Филатов, И. В. Кравцов
Петрозаводский государственный университет, Россия

Our article covers various terms and development technologies for web information systems for historical documents. The purpose of our system is to prepare electronic and printed editions for various kinds of medieval historical texts collections. We also describe a special research toolkit based on two original editors: SVG visual editor and XML markup editor. The article is built on logical pairs: needs of historians — our technological offers for their problems.

Предлагаемая нами информационная система ориентирована на введение в научный оборот специализированного инструментария и методологий, позволяющих, во-первых, публиковать в сети Интернет комплексы средневековых документов, во-вторых, проводить на базе этих коллекций широкий спектр источниковедческих исследований и, в-третьих, организовать информационное пространство для совместной работы сообщества исследователей.

В данном тексте мы не ставим перед собой задачу аргументировать применение новейших компьютерных технологий в традиционных областях исторического источниковедения, так как это уже описано в других работах [Технология 2005: 241–281], кроме того, у каждого исследователя есть по этому поводу свое мнение. Мы лишь предлагаем платформу, с помощью которой можно это мнение подтвердить реальным экспериментом, перепроверить результаты этого эксперимента и подискутировать о методе.

На данный момент в систему вводятся документы двух комплексов источников: комплекса средневековых исторических источников «Moscowitica-Ruthenica» XII–XVIII вв. [Moscowitica

[1] Работа выполнена при финансовой поддержке РГНФ (проект № 06-01-12124в).

2004, № 3: 47–54, № 4: 94–106], а также комплекса документов приказного делопроизводства из истории города Динабурга XVII в. [Динабург 2002].

Чтобы пояснить выбор используемых информационных технологий, мы будем ссылаться на обобщенные требования историков, предъявляемые к инструментарию подобного рода. Данные требования сформировались в процессе совместной работы с латвийским историком профессором А. С. Ивановым, а также в результате общения с другими историками-источниковедами при обсуждении проекта нашей информационной системы на X конференции Ассоциации «История и компьютер» (Москва, 2006 г.).

Работать лучше в сообществе. Учитывая обширность большинства коллекций и трудоемкость их изучения и подготовки к публикации, исследователь-историк сталкивается с трудно решаемой задачей, по крайней мере, за конечное время. Совершенно оправдано его желание разбить задачу на части, привлечь других заинтересованных специалистов, а также разделить их обязанности и сузить задачи каждого из них. Причем хочется привлечь всех желающих, независимо от их местонахождения, то есть использовать для этого возможности глобальной сети Интернет. Именно поэтому нашу систему изначально было решено реализовывать с помощью клиент-серверных web-технологий на некоммерческой основе при помощи open source средств.

Среди принципов построения системы можно выделить несколько основных, определяющих ее функциональность и архитектуру:

– основной задачей является подготовка коллекций средневековых документов к публикации, причем как электронной, так и печатной;

– система ориентирована на сообщество исследователей, они работают удаленно, объединяются в группы, могут корректировать и дополнять результаты друг друга;

– система не является замкнутой и подготовлена к вводу и выводу информации, это касается не только текстов источников, но и результатов и промежуточных отчетов исследований [Использование 2006: 64–66].

Историки предпочитают работать с полным текстом. Исследователей интересует не только повествовательное содержание

документа, но и его происхождение, палеографические особенности и другие важные сведения об источниках как таковых, о способах и контексте репрезентации информации в них. Причем все это историки видят как одну сущность, неразделимую на отдельные фрагменты (например, на ячейки таблиц реляционной модели данных). Должен быть один документ, содержащий всю эту информацию, а также готовый включить в себя дополнительные сведения, — XML-документ. Действительно, выбранная нами технология XML [http://www.w3.org/XML] позволяет создавать свой произвольный формат данных, ничем не ограничивая исследователя.

Историки хотят работать с оригиналами источников или хотя бы с их изображениями. Работа с изображениями позволяет более точно учесть особенности каждого документа. Именно поэтому разметка документов в нашей системе, в отличие от других систем, может осуществляться тремя способами: можно размечать изображения документов, тексты на языке оригинала или адаптированные тексты. Для разметки изображений создан online-редактор [Филатов 2006: 64–66], позволяющий различными способами помечать участки документа (слова, отдельные символы, физические особенности документа) и затем задавать им значения: написание на языке оригинала, адаптированное написание… При вводе значений для выделенного блока изображения существует возможность автоподстановки на основе уже введенных слов для данной коллекции документов. Если ничего похожего не находится в словаре введенных слов для конкретной коллекции, то тогда просматривается глобальный словарь для всех введенных на данном языке слов. Вышеозначенный редактор реализован с помощью возможностей технологии SVG (Scalable Vector Graphics) [http://www.w3.org/Graphics/SVG], основанном на XML стандарте для описания двумерной графики и графических приложений. Данная технология позволяет «рисовать» прямо на изображении документа, а затем «рисунки-разметку» сохранять в виде SVG-файла для последующей работы с данным изображением или для воспроизведения проделанной работы.

Историкам желательно пользоваться одним инструментом для различных задач. Множественность поставленных задач, различия в структуре и языке размечаемых документов, необходимость модификации разметки для разных коллекций — все это

порождает огромное количество вариантов XML-редактора, приспособленного под конкретный вид документа и способ разметки. Мы предлагаем унифицировать интерфейс и работать с одним настраиваемым редактором, в котором можно опционально подключать те или иные возможности: выбирать схемы разметки из базы готовых разметок, создавать собственную разметку, подключать виртуальную клавиатуру для набора устаревших символов (например, старославянских), а также отображать параллельно несколько вариантов разметки одного текста либо текста и его сканированного изображения.

Историки хотят работать «как в текстовом редакторе». Вполне разумно требование удобства использования и интерактивности информационной системы. В идеале система должна работать как обычное настольное приложение и сразу реагировать на действия пользователя, а не загружать новое окно с очередной web-формой или результатом запроса. Эффекта настольной программы можно достичь с помощью подхода к построению пользовательских интерфейсов Web-приложений AJAX (Asynchronous JavaScript and XML) [http://en.wikipedia.org/wiki/AJAX], в котором с помощью асинхронных запросов с сервера подгружается лишь та информация, которая необходима для текущего изменения web-страницы. Этим достигается необходимая интерактивность, система сразу реагирует на каждое значимое действие пользователя (например, на описанную выше автоподстановку), и не надо нажимать никаких кнопок «отправить», чтобы узнать, что вы вдруг сделали ошибку. Кроме того, XML-документ в редакторе на стороне сервера разбирается в реляционную базу данных (мы используем PHP+MySQL). Это делается для того, чтобы избежать многократного прочтения текстового XML-файла и ускорить запросы к редактируемому документу. При завершении работы с редактором (с конкретным файлом) документ опять записывается в виде XML-файла.

Историки хотят работать с понятным им инструментом. Многие потенциальные пользователи сталкиваются с проблемой правильного, последовательного и корректного использования предложенного инструментария. Кроме того, вокруг коллекций и исследований возникает свое, узкоспециализированное информационное пространство. Нами предлагается оформить это про-

странство в качестве базы знаний, привязанной к системе, реализованной с помощью платформы MediaWiki [http://www.mediawiki.org/]. Wikipedia [http://wikipedia.org] — реализованная на этой платформе всемирная открытая энциклопедия, англоязычная версия которой содержит уже больше миллиона статей, наглядно показывает, что применяемые в ней принципы работают хорошо. В нашем случае, например, удобно, что тематические статьи могут создаваться несколькими авторами совместно, в том числе членами одной исследовательской группы. Кроме того, все употребляемые в обороте термины и методики могут ссылаться на связанные с ними статьи, поясняющие их смысл и назначение. Если какой-то термин не представлен в базе, то пользователь может создать связанную с ним пустую статью, которую оперативно добавляют эксперты-исследователи либо другие пользователи системы. Ход произвольного исследования может комментироваться в двух потоках: в официальном — основной статье, которая позже может входить в публикацию результатов исследования, а также в потоке обсуждения в форме неформального общения.

Список литературы

Варфоломеев и др. 2006 — *Варфоломеев, А. Г.* Использование технологии XML для публикации методики и результатов исследования текстов исторических источников / А. Г. Варфоломеев, И. В. Кравцов // Информационный бюллетень АИК. — № 34. — 2006. — С. 64–66.

Иванов 2004 — *Иванов, А. С.* «Moscowitica-Ruthenica» в Латвийском государственном историческом архиве / А. С. Иванов // Древняя Русь. 2004. — № 3. — С.47–54 ; № 4. — С.94–106.

Иванов и др. 2005 — *Иванов, А. С.* Технология XML как инструмент компьютерного источниковедения (на примере формулярного анализа документов приказного делопроизводства) / А. С. Иванов, А. Г. Варфоломеев // Круг идей: Алгоритмы и технологии исторической информатики : тр. IX конф. Ассоциации «История и компьютер» / ред. Л. И. Бородкин, В. Н. Владимиров. — Москва ; Барнаул : Изд-во Алт. ун-та, 2005. — С.241–281.

Иванов и др. 2002 — *Иванов, А.* Динабург в документах российского государственного архива древних актов (1656–1666) / А. Иванов, А. Кузнецов. — Ч. I–II. — Daugavpils : Saule, 2002.

Филатов 2006 — *Филатов, В. О.* Специализированный XML-редактор для создания полнотекстовых баз данных на основе изображений исторических источников / В. О. Филатов // Информационный бюллетень АИК. — № 34. — 2006. — С. 67–69.

WikipediA 2006 — WikipediA The Free Encyclopedia [Электронный ресурс]. — 2006. — Режим доступа: http://en.wikipedia.org/, свободный. — Загл. с экрана.

WikipediA 2006 — WikipediA The Free Encyclopedia: Ajax (programming) [Электронный ресурс]. — 2006. — Режим доступа: http://en.wikipedia.org/wiki/AJAX, свободный. — Загл. с экрана.

MediaWiki 2006 — MediaWiki [Электронный ресурс]. — 2006. — Режим доступа: http://www.mediawiki.org/, свободный. — Загл. с экрана.

SVG 2006 — Scalable Vector Graphics (SVG) XML Graphics for the Web [Электронный ресурс]. — 2006. — Режим доступа: http://www.w3.org/Graphics/SVG, свободный. — Загл. с экрана.

XML 2006 - Extensible Markup Language (XML) [Электронный ресурс]. — 2006. — Режим доступа: http://www.w3.org/XML, свободный. — Загл. с экрана.

Информационная технология создания электронного издания Словаря Академии Российской 1789–1794 гг.

А. Ю. Филиппович
Московский государственный технический университет им. Н. Э. Баумана, Россия

The report topic: Information technology of creating the electronic edition of the Russian Academy Dictionary 1789–1894.

In the report the features of creating the printed reedition of the Russian Academy Dictionary are considered and the results of study of the efficiency of the proof-reading processes and the analysis of the frequency characteristics of the dictionary text are presented. The structure of the electronic edition of the Russian Academy Dictionary and the technology of designing its principal components (linguistic database, hypertext information system and hypergraphic system of facsimile copies of the dictionary pages) are described.

Сегодня многие старинные книги и рукописи находятся на грани исчезновения. Причина тому — несовершенство средств хранения информации. Эти печатные и рукописные материалы, памятники литературы и письменности, являются предметом и источником научных исследований. Для решения проблемы доступности этой информации для потенциальных исследователей источники вводятся в научный оборот, осуществляется их копирование. Особенности современных (компьютерных) издательских технологий, малые тиражи научной литературы, корпоративные интересы носителей научного знания (книговедов, филологов, историков и др.) сближают процессы копирования источников и их переиздания. Современная доступная копия какого-либо источника — это его печатное и электронное научные переиздания.

Одним из таких источников является Словарь Академии Российской 1789–1794 гг. (САР) [САР 2]. Это первый толковый словарь русского языка. Он был создан в эпоху просвещения. В это время, в 1783 г., по Высочайшему соизволению Российской импе-