Использование технологии XML для введения в научный оборот комплекса документов «Moscowitica-Ruthenica»♣

© Иванов А.С.[1], Варфоломеев А.Г.[2]

[1]Даугавпилсский университет (Латвия)
[2]Петрозаводский государственный университет (Россия)
history@dau.lv, avarf@psu.karelia.ru

Аннотация

Среди наиболее ценных исторических источников Латвийского государственного исторического архива особое место занимает комплекс средневековых документов "Moscowitica-Ruthenica". В статье рассматриваются возможности введения этого комплекса в научный оборот путем электронной публикации документов в виде, наиболее удобном для решения исследовательских задач. Для достижения этой цели предлагается использование технологии XML, предполагающей разработку языка разметки на базе существующего стандарта TEI. Один раз размеченный документ может затем многократно использоваться – как для просмотра на экране монитора или подготовки печатного издания, так и для поиска информации или её интерпретации.

1 Введение

Статья посвящена проекту создания полнотекстовой базы данных исторических источников, входящих в комплекс документов «Moscowitica-Ruthenica» (далее – MR) Латвийского государственного исторического архива (далее – ЛГИА).

Цель проекта – введение в научный оборот уникальной коллекции документов как в форме печатной публикации, так и в виде электронной библиотеки, предоставляющей историкам и языковедам не только сами источники, но и инструменты их исследования – поиск документов по различным критериям, подсчёт частот встречаемости в текстах тех или иных объектов и признаков, сравнение формуляров документов между собой. Для современной науки представляется актуальной как сама публикация большого комплекса средневековых документов в форме, удобной для исследования, так и апробация на этом примере определённой технологии, позволяющей публиковать таким образом и другие текстовые источники.

Для достижения этой цели предлагается разработать язык разметки текстов средневековых источников, входящих в MR. Для совместимости с уже существующими стандартами такой язык должен быть основан на XML-схеме разметки TEI, широко применяемой историками, лингвистами и литературоведами. Размеченный текст далее может быть многократно использован – для вывода на экран монитора в виде HTML-документа, для преобразования в удобный для печати формат PDF или RTF, а также как объект для выполнения произвольных запросов.

2 О комплексе MR

2.1 Общая характеристика комплекса

Коллекция документальных памятников MR представляет собой естественный, исторически сложившийся комплекс исторических источников об отношениях Риги, Ливонии, Ганзы и отдельных немецких городов с древнерусскими землями, княжествами и городами, Московским государством, а также с Великим княжеством Литовским и Речью Посполитой с конца XII в. и до начала XVIII в. К сожалению, комплекс MR, как единое целое, лишь в редких случаях становился объектом специального изучения. Лишь отдельные документы комплекса, преимущественно – древнейшие, издавались неоднократно, благодаря чему активно используются в научных исследованиях. При этом большая часть уникальных документов по-прежнему находится вне научного оборота.

Труды 6ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2004, Пущино, Россия, 2004.

2.2 История формирования комплекса

Состав документов MR обусловлен функциями высшего органа управления Риги – магистрата: с момента присоединения Риги к торговому и политическому союзу северо-немецких городов (Ганза) в 1282 г. важной областью деятельности рижского магистрата стало поддержание экономических и политических связей с русскими землями, посредничество в торговле Руси с Ганзой. Как результат документирования этой функции магистрата и возник комплекс MR.

Первоначально документальные свидетельства об отношениях Риги и Ливонии с восточными соседями концентрировались в двух различных структурных подразделениях архива рижского магистрата. Во-первых, это было хранилище особо важных документов – так называемый «внутренний архив». Отдельно хранилась текущая документация магистрата, включающая переписку с русскими и литовскими городами, княжествами и землями, отдельные договоры, грамоты русских и литовских князей. Во второй половине XVIII в. эта часть архива рижского магистрата стала называться «внешним архивом». Именно во внешнем архиве и сложился самостоятельный отдел MR, название которого впервые было зафиксировано в описи, составленной в XVII в.[1]

С 1962г. документы бывшего внешнего и внутреннего архива рижского магистрата составили два отдельных фонда ЛГИА. Несмотря на это разделение, документы, с исследовательской точки зрения, по-прежнему образуют единый комплекс.

2.3 Состав комплекса

В комплексе MR можно выделить следующие разновидности источников: международные политические и торговые договоры русских княжеств, городов и земель, а также великих княз ей литовских с Ригой, немецкими городами, Ливонией; подтвердительные грамоты русских городов, русских и литовских князей; рядные грамоты и другие частные акты; жалованные грамоты и привилегии русских и литовских князей; судные грамоты; указные грамоты; памяти, инструкции, распоряжения, постановления и приказы рижского магистрата; опасные, проезжие и верительные грамоты; паспорта; доклады, донесения и отчеты рижских послов из русских княжеств, Новгорода, Пскова, а также из Литвы; грамоты наместников, епископов, воевод русских княжеств и городов, адресованные рижскому магистрату, ганзейским городам, магистру Ливонского ордена; положения и торговые уставы немецких торговых дворов в Новгороде и Полоцке; инструкции немецким купцам; протоколы учреждений города Риги и Ливонии; различные описи и реестры; жалобы, челобитные и прошения купцов, как ливонских, так и русских. Таким образом, в комплексе MR представлена большая часть видов и разновидностей актового и делопроизводственного материала, характерного как для Руси, так и для регионов, входивших в сферу немецкого политического и культурного влияния.

В комплексе MR преобладают оригиналы. Наряду с подлинниками, сохранились и современные списки важнейших актов, сделанные в канцелярии рижского магистрата. Оригинальные документы составлены на разных языках: на древнерусском, немецком (нижненемецком, средненемецком, верхненемецком диалекте), латинском языке, а отдельные документы – также на польском и шведском языках. В употреблении языков можно выявить определенные закономерности: на средненемецком были написаны грамоты литовских великих князей; на нижненемецком рижский магистрат вел переписку с другими учреждениями Ливонии, с немецкими купцами в Новгороде, Пскове, Полоцке, Витебске, а также со своими должностными лицами и с ганзейскими городами. Документы на верхненемецком диалекте в комплексе MR появляются лишь со второй половины – конца XVI в. Немногочисленные документы на шведском языке относятся к периоду шведского господства в Риге (1621-1710гг.). Большая часть источников на польском языке датируется 1581 – 1621гг. – временем вхождения Риги в состав Польско-литовского государства. Документы на русском языке в массиве MR представлены довольно равномерно, хотя и можно отметить, что их удельный вес был наибольшим в XIII – XIV вв. и наименьшим – в XVII в.

Объем комплекса (документального массива) MR значителен: 787 документов (свыше 1500 листов) во внешнем архиве[2], около 80 документов во внутреннем архиве[3]. и отдельные документы в фонде коллегии ландратов Лифляндии[4].

2.4 История введения документов MR в научный оборот

Описание документов комплекса MR, начатое еще в XVI – XVII вв., особенно активно велось со второй половины XIX в. Прибалтийско-немецкие историки и архивисты удовлетворительно датировали и описали большую часть данного документального массива. Однако нельзя утверждать, что эта работа завершена: атрибуция и датировка некоторых источников требует пересмотра или уточнения, ряд документов по-прежнему не описан и не датирован. Нельзя считать законченной и работу по выявлению документов, относящихся к комплексу MR, и в других фондах ЛГИА.

Публикация источников из собрания MR началась в XIX в. За два века археографической деятельности удалось опубликовать свыше половины документов комплекса. При этом наивысшая степень полноты представления источников MR в сборниках документов достигнута за XIII – начало XVI в., хотя и в этих временных границах можно констатировать пробелы.

Источники XVI – XVII вв. в сборниках документов представлены выборочно. Качество большинства публикаций не соответствует современному уровню археографии, поэтому их можно считать устаревшими. В целом публикации не позволяют создать адекватное представление о комплексе MR.

История введения документов комплекса MR в научный оборот поучительна и противоречива: в археографической деятельности ученые пытались следовать принципу всесторонности и полноты при представлении источников в сборниках, на практике же лишь относительно небольшая часть источников публиковалась и, соответственно, интенсивно использовалась в исторических исследованиях. Таким образом, комплексный подход, обоснованный в теоретическом источниковедении [6], был фактически заявлен в различных планах издания документов рижского магистрата, однако в практике введения комплекса MR в научный оборот не был реализован. Как отмечают археографы и источниковеды, основная причина разрыва теории и практики – подчинение археографической деятельности господствующим в исторической науке представлениям и интерпретациям, в результате чего археографы вынуждены заниматься иллюстрированием основных положений своей школы в историографии [2]. Это приводило и приводит к тому, что исторические источники искусственно «вырываются» из контекста своей эпохи, игнорируется место документов в системе делопроизводства, равно как и взаимосвязь их с другими документами комплекса.

Иллюстративному подходу в археографии можно противопоставить метод «сплошной» публикации архивных фондов, являющихся исторически сложившимися комплексами источников. Это позволяет уменьшить элемент субъективности при отборе источников к публикации и, соответственно, при формировании источниковой базы исследования. Реализация данного подхода будет также способствовать применению адекватных методов к изучению больших групп исторических источников, начиная с традиционных методов и приемов источниковедческой критики, формулярного анализа, контент-анализа и заканчивая новейшими компьютерными технологиями представления, связывания и агрегирования источниковой информации [3].

Представляется, что реализация проекта по созданию полнотекстовой базы данных на основе всех документов комплекса MR и будет реальным воплощением комплексного подхода в источниковедении и археографии.

3. Электронная библиотека документов комплекса MR

3.1 Технология XML и стандарт TEI

Как уже отмечалось, полнотекстовая база данных должна обеспечить представление комплекса MR научной общественности как в форме «сплошной» печатной публикации документов, так и в виде электронной библиотеки, снабжённой всеми необходимыми инструментами исследования текстов.

На наш взгляд, наиболее оптимальное решение задачи одновременной подготовки печатной и электронной публикации комплекса MR возможно на базе технологии XML. Эта технология предусматривает разметку текстов исторических источников с помощью определенной системы тегов. Разметка выделяет некоторые смысловые единицы текста, которые могут вкладываться друг в друга и описывать структуру источника с любым уровнем детальности, вплоть до отдельных слов или символов. Смысловые единицы снабжаются атрибутами, относящими их к тем или иным классам объектов, или задающими для них стандартные значения (это касается, например, дат или географических названий). Разметка может быть использована также для связи фрагментов текста с их комментариями и переводами (что особенно актуально для публикации документов комплекса MR). Технология XML дает в руки исследователей сколь угодно гибкое средство структурирования и анализа текста, оставляющее источник неизменным, не «разбитым» на ячейки реляционных таблиц. Можно сказать, что это технология источнико-ориентированных баз данных [1], доведенная до совершенства.

С целью совместимости нашего проекта с общепринятыми на сегодняшний день стандартами мы предлагаем использовать схему разметки TEI (Text Encoding Initiative) [10]. TEI – один из самых известных XML-стандартов разметки, широко используемый в проектах по созданию электронных коллекций для гуманитарных наук. Изначально этот стандарт был ориентирован прежде всего на описание конкретного печатного издания некоторого литературного произведения – с «разбивкой» текста на главы и абзацы, стихов – на строфы, с указанием концов страниц, выделением прямой речи героев и т.д. Но возможности, заложенные в стандарт, намного шире, что позволяет активно использовать его не только для создания адекватных электронных копий печатных изданий, но и для выделения системы логических фрагментов в текстах произвольной природы, а также для связи этих фрагментов между собой. В результате TEI можно считать прежде всего инструментом анализа и интерпретации текста, что делает его незаменимым для решения источниковедческих задач.

На сайте Консорциума TEI приводится длинный список проектов, в той или иной мере использующих этот стандарт. Есть среди них и проекты, посвященные созданию электронных коллекций средневековых исторических документов [7,8,9]. Как правило, в них предлагаются специфические схемы разметки, основанные на TEI, но ориентированные на конкретные особенности текстов, входящих в коллекции, и на решение определённых задач, стоящих перед публикаторами и исследователями.

3.2 Принципы разметки, предлагаемой в нашем проекте

В источниковедческой практике давно сложились определённые правила описания, репрезентации информации и критики (анализа и синтеза) средневековых источников. Эти правила рассматривают исторический источник как многоуровневый объект. На самом верхнем уровне находится метаинформация об источнике – вид, язык, место хранения. Второй (палеографический) уровень составляют внешние признаки документа – материал, дефекты, печати, филиграни, особенности почерка. Сюда же относятся маркеры, задающие физическое разделение текста источника на листы и строки. Третий уровень можно назвать дипломатическим, он выделяет в тексте определённые логические фрагменты, характерные для документов (актового материала) данного вида. Так, в средневековых грамотах выделяют invocatio (посвящение Богу), intitulatio (обозначение лица, от которого исходит документ), inscriptio (обозначение адресата), satulatio (приветствие), и т.д.[4]. Далее располагаются уровни, характеризующие стилистические, синтаксические, лексические особенности документов. Они предусматривают деление текста на предложения, устойчивые словесные обороты и отдельные слова. Наконец, на самом нижнем уровне находятся символы, составляющие текст.

Разметка исторического источника, ориентированная как на многообразное представление, так и на различные виды анализа текста, должна позволять выделять объекты на каждом из рассмотренных уровней источника. Соединение всех уровней в одной схеме разметки вполне возможно, поскольку объекты более низких уровней оказываются «вложенными» в объекты более высоких уровней, и именно такая иерархическая структура документа является оптимальной для применения технологии XML.

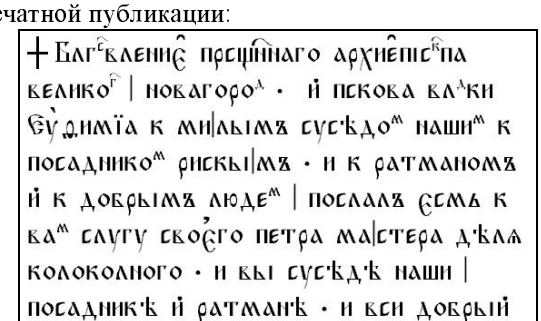
Документ с разметкой всех уровней будет первичным документом нашей базы данных. На его основе с помощью преобразований, описанных на языках XSLT и XSL-FO [5], будут получаться:

1) файлы в форматах HTML, PDF, RTF, представляющие документ на экране монитора или формирующие оригинал-макет для издательства,

2) XML-файлы, содержащие разметку только одного уровня. Такие файлы будут служить основой для анализа и интерпретации текста, который, на наш взгляд, должен заключаться в ручном или автоматическом распределении объектов текста по различным типам, определяемым исследователем, в подсчёте частоты встречаемости объектов (признаков) тех или иных типов, в установлении связи между фрагментами разных документов, в комментировании или переводе этих фрагментов. Результаты такого анализа, проведённые для одного или нескольких текстов, также можно оформить с помощью разметки TEI, что позволит исследователям обмениваться друг с другом как своими выводами, так и научными методиками.

Следует сказать несколько слов о проблеме представления символов и диакритических знаков, характерных для средневековых текстов. Мы предлагаем использовать методику проекта Menota [8], согласно которой все символы, встречающиеся в документах комплекса MR, должны быть представлены в первичных документах нашей базы данных как XML-сущности (строки вида «&name;»), связанные с определёнными кодами стандарта Юникод. На наш взгляд, диакритические знаки следует рассматривать как отдельные символы, чтобы в максимальной степени постараться соответствовать существующим в настоящее время стандартным диапазонам символов Юникода . Для того, чтобы тексты на разных языках могли одновременно отображаться на экране монитора, требуется создать специальный шрифт, содержащий в себе все символы, используемые в проекте.

В качестве примера рассмотрим один из документов комплекса MR – грамоту архиепископа новгородского и псковского Евфимия о присылке колокольных дел мастера, отправленную в Ригу в 1456г.[5] На рисунке представлено начало документа в том виде, который должен быть достигнут при печатной публикации:



Приведем фрагмент XML-документа, содержащего некоторую разметку нижних уровней и соответствующего первым четырем словам текста источника:

```
<doc><docBody>
<cross/>
<w n="1">&BU;&lu;&gla;<top>&slo;&pal;</top>
```

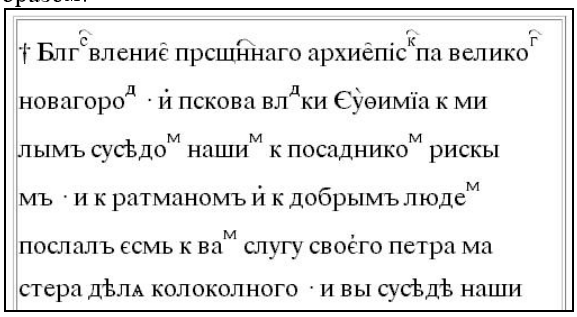
```
&ve;&lu;&est;&na;&izhe;&est2;&ibr;</w>
<w n="2">&po;&rts;&slo;&sht;&na;&pal;&na;
&az;&gla;&on;</w>
<w n="3">&az;&rts;&her;&izhe;&est;&ibr;&po;
&id;&slo;<top>&ka;&pal;</top>&po;&az;</w>
<w n="4">&ve;&est;&lu;&izhe;&ka;&on;
<top>&gla;&pal;</top></w><lb n="1"/>
...
</docBody></doc>
```

В этом фрагменте жирным шрифтом выделены теги стандарта TEI, подчеркиванием – дополнительные теги (в частности, <top> </top> для выносных букв и <cross/> для креста). XML-сущностям типа &BU; или &lu; в DTD документа ставятся в соответствие коды стандарта Юникод:

```
<!ENTITY BU "&#1041;">
...
<!ENTITY lu "&#1083;">
```

Несколько XML-сущностей соответствуют диакритическим знакам: &pal; (“палатализация”), &tit; (“титло”), &ibr; (“inverted breve”), и др.

Применяя к XML-документу соответствующее XSLT-преобразование, мы можем получить HTML-документ, отображающийся в окне браузера с помощью шрифта TITUS Cyberbit Basic следующим образом:



XML- и XSL-файлы рассмотренного примера, а также более подробную информацию о принципах разметки и преобразованиях документов можно получить на сайте нашего проекта по адресу: http://hist-docs.cs.karelia.ru/MR.

Литература

[1] Гарскова И.М. Базы и банки данных в исторических исследованиях. М.-Гёттинген, 1994.

[2] Добрушкин Е.М. Археографическая теория и источниковедческая практика (к вопросу о взаимосвязи) // Россия в IX – XX веках: Проблемы истории, историографии и источниковедения. М., 1999. С.131-132.

[3] Иванов А.С. Проблемы введения в научный оборот обширных комплексов источников: коллекция «Moscowitica – Ruthenica” в Латвийском государственном историческом архиве // XXI век: Актуальные проблемы исторической науки. Минск, 2004. С.100-101.

[4] Каштанов С.М. Русская дипломатика. М., 1988.

[5] Холзнер С. XSLT. СПб., 2002.

[6] Шмидт С.О. Путь историка: Избранные труды по источниковедению и историографии. М., 1997. С.50-53.

[7] CELT Project. http://www.ucc.ie/celt

[8] Menota (Medieval Nordic Text Archive). http://www.menota.org

[9] Repertorium of Old Bulgarian Literature and Letters. http://clover.slavic.pitt.edu/~repertorium

[10] TEI P4. Guidelines for Electronic Text Encoding and Interchange. TEI Consortium, 2001. http://www.tei-c.org/P4X/

XML-technologies in Introduction into Scientific Circulation of the Document Collection “Moscowitica-Ruthenica”

A.Ivanov, A.Varfolomeyev

Among the most valuable historical sources of the Latvian State Historical Archives there is a collection of documents, which provides historians with firsthand information about relations of Muscovy Rus, Russian, Belorussian lands and towns (Smolensk, Novgorod, Pskov, Polotsk, etc.), as well as Lithuania with Riga, Livonia, Hanseatic League and some German towns in the late 12th – early 18th centuries. The historical name of this document collection is “Moscowitica-Ruthenica”. Although this collection as a department of the Latvian State Historical Archives doesn’t exist any more, its documents constitute the natural complex of historical sources, which should be studied as a whole.

In order to stimulate the circulation of these documents within the histor ical science, a new complete edition (both – a paper one and an electronic one) of the collection “Moscowitica-Ruthenica” seems to be urgent. In achieving this purpose XML-technologies should be used. The application of the technologies mentioned above to the texts from the collection “Moscowitica-Ruthenica” requires the development of the specialized mark-up language. As a result, any XML-document can be used repeatedly in order to achieve different scientific aims, representation of a document as a HTML-page and preparation of the printed document edition included. The further information about principles of marking-up and transformation of the document texts is accessible in: http://hist-docs.cs.karelia.ru/MR

♣ Статья написана в рамках проекта по подготовке публикации источников «Moscowitica-Ruthenica», поддержанного администрацией Латвийского государственного исторического архива и Фондом культурного капитала Латвии.

[1] ЛГИА, ф.673, оп.1, д.1482 (Ruthenica), 1483 (Moscowitica).

[2] ЛГИА, ф.673, оп.4, Kasten 18-20.

[3] ЛГИА. ф.8, оп.3, capsula A, №№14-18, 41, 72; capsula B, №42; capsula C, №№1-11, 23, 27, 34, 43; оп.4, №№6-58.

[4] ЛГИА, ф.214, оп.6, д.114, 115, 116.

[5] ЛГИА, ф.673, оп.4, Kasten 18, №159.