Информационная система для работы с полнотекстовыми базами данных исторических документов на основе технологии XML♣

В.О.Филатов И.В.Кравцов А.Г.Варфоломеев

Петрозаводский государственный университет
г. Петрозаводск, Россия

Аннотация

Наша статья посвящена вопросам, связанным с разработкой инструментария, доступного через Интернет, для введения в научный оборот уникальных комплексов документов, как в форме печатной публикации, так и в виде полнотекстовой базы данных, предоставляющей историкам 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 структуры документов между собой. Для исторической науки представляется актуальной как сама публикация больших комплексов документов в форме, удобной для исследования, так и апробация на этих примерах определённой технологии, позволяющей придать традиционным публикациям источников новую форму и новое качество.

1 Существующие проекты

Современная историческая наука уже вступила на этап создания полнотекстовых баз данных исторических документов, доступных в сети Интернет. Коллективом под руководством проф. В.А.Баранова (Ижевск) разработана информационно-поисковая система «Манускрипт» [5], позволяющая вводить, редактировать, хранить и обрабатывать древнерусские тексты, а также просматривать их в окне Web-браузера. Для хранения текстов в этой системе используются реляционные базы данных.

Существует ряд проектов, основанных на использовании технологии XML. Как правило, эти проекты разрабатываются научными коллективами, входящими в консорциум TEI (Text Encoding Initiative)[15]. Среди проектов, использующих стандарт TEI, можно выделить несколько проектов, по своим целям и предметной области наиболее близких к нашему:

Menota (Medieval Nordic Text Archive) [10] – проект, посвященный электронной публикации средневековых рукописей Норвегии, Швеции, Исландии и Дании;

MEP (Model Edition Partnership) [11] – разработка схемы разметки для электронных публикаций писем исторических деятелей;

Newton Manuscript Project [14] – электронная публикация рукописей И.Ньютона;

Repertorium of Old Bulgarian Literature and Letters [12];

CELT Project [8], посвященный созданию электронной коллекции исторических и литературных ирландских текстов.

Характерной чертой этих проектов является то, что они предлагают 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.

Однако в настоящее время нет ни одного Интернет-проекта, посвященного представлению и исследованию исторических документов, который предоставлял бы вместе со средствами создания и просмотра коллекций текстов те инструменты, которые необходимы для анализа информации и структуры текста, а также средства обмена между учеными полученными результатами и применяемыми методиками, описанными в единой форме. Поэтому можно констатировать отсутствие отечественных и зарубежных аналогов для предлагаемого нами проекта.

Кроме того, в наши дни технология XML используется повсеместно во многих прикладных областях как универсальный формат хранения и передачи данных любого характера. Можно уверенно сказать, что каждый второй (если не все) сетевой проект так или иначе пользуется преимуществами XML, и мы в том числе.

2 Источнико-ориентированный подход

Историки всегда ощущали ограниченность технологии реляционных баз данных, ее слабую приспособленность к адекватной и всесторонней передаче информации, содержащейся в исторических источниках. Действительно, эта технология ориентирована на представление

информации в виде таблиц «объект-признак», в которых каждый объект исследования характеризуется определенным набором значений фиксированных признаков. Но для историка также важны сведения о самих источниках как таковых (палеографические особенности, происхождение и т.д.), о способах и контексте репрезентации информации в них. Поэтому историками многократно предпринимались попытки соединить в базах данных значения признаков объектов с информацией об оригинальном (исходном) представлении тех же признаков в источниках. Но в рамках господствовавшей реляционной модели последовательно реализовать источнико-ориентированный подход было делом затруднительным.

Возможности для реализации источнико-ориентированного подхода представляет технология XML, основанная на идее разметки текстов источников служебными словами-тэгами и последующей работы с этими размеченными текстами (XML-документами) как с базами данных. В рамках технологии XML выделяется два вида XML-структур – data-driven и document-driven. Первые напоминают обычные реляционные таблицы, только записанные в текстовом виде. Структуры document-driven – это оригинальные тексты источников с выделенными в процессе разметки логическими фрагментами. Именно эти размеченные тексты, дополненные метаинформацией об источниках, и являются реальным воплощением источнико-ориентированного подхода при создании XML-баз данных.

3 Направление исследования

Использование технологии XML требует от историков и специалистов в области компьютерного источниковедения и исторической информатики рассмотрения следующих вопросов:

- Разработка новых или адаптация существующих схем разметки для различных видов письменных исторических источников;
- Разработка инструментов, облегчающих процесс разметки электронных текстов исторических источников;
- Разработка алгоритмов для реализации запросов к XML-документам, а также средств, облегчающих создание таких запросов;
- Создание информационных систем, предназначенных для введения в научный оборот комплексов исторических источников, представленных как XML-базы данных.

Поставленные вопросы частично разрешались при подготовке к электронной публикации комплекса средневековых исторических источников «Moskowitica-Ruthenica» [2], а также при решении задачи формулярного анализа документов приказного делопроизводства на примере комплекса документов из истории города Динабурга [3]. Для указанных коллекций были разработаны специальные схемы разметки. Мы представляем систему, объединяющую обе эти коллекции документов и разработанные для них принципы организации полнотекстовых источнико-ориентированных данных, и предназначенную для дальнейших научных исследований коллекций. Система изначально ориентирована на сообщество исследователей и совмещает в себе функции исследовательского инструмента и хранилища готовых к публикации источников [4]..

4 Описание системы

4.1 Логические слои

Для удобства описания разобьем систему на три логических слоя: слой пользовательского взаимодействия (user layer), слой организации хранилища коллекции документов (storage layer) и слой организации исследования (research layer).

Содержимое слоев представлено на схеме (рис. 1), на которой отражены пользователи системы, информационные объекты и их взаимодействие, а также компоненты системы, из которых она составлена. Это такие модули как:

- Регистрация в системе;
- Система назначения прав и создания групп;
- Менеджер доступных коллекций документов;
- Компонента просмотра документов;
- Визуальный онлайн редактор для перевода изображений в первоначально размеченные тексты (символьная разметка);
- Редактор для логической разметки текстов;
- Менеджер разметок;
- Менеджер логических модулей;
- Компонента просмотра отчетов;
- MediaWiki платформа сопутствующей исследованиям документации.

4.2 Движение информации в системе

Процесс исследования начинается с момента, когда источник появляется в системе. Скорее всего - это отсканированное изображение в одном из графических форматов. Затем текст распознается в ручном, полуавтоматическом или автоматическом режиме. При распознавании для кодирования текстов используются естественные алфавиты (например - современный русский язык), либо специально разработанные и хранимые в виде XML-сущностей (например – старославянский язык). Полученный текст дополняется метаинформацией о том, что это за текст, откуда он, кто автор и т.д. В таком виде текст документа считается «исходным» в системе дальнейшего анализа, а также готовым к публикации как самостоятельный электронный документ.

Кроме того, в системе заводится для каждого источника специальный файл истории (тоже XML-

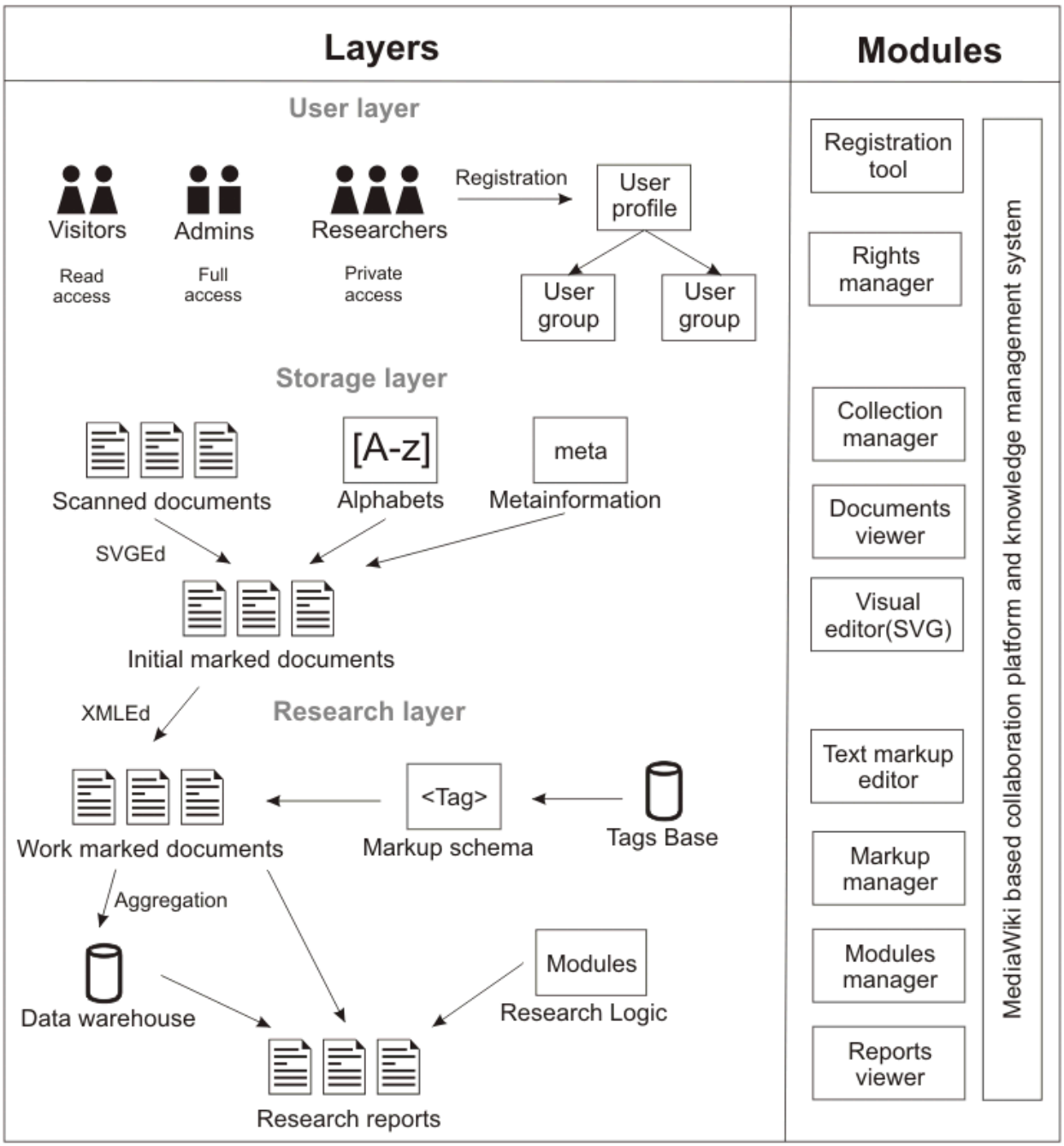

рис. 1

документ), который содержит в себе ссылки на все формы документа, которые появляются в системе (сканированное изображение, физически размеченный текст, логически размеченный текст и др.), а также хронологию его изменения и сопутствующие комментарии исследователя. Также хранится информация о том, кто осуществлял операции сканирования, распознавания, разметки и т.д. Если текст распознавался автоматически, то документ появляется сначала в виде черновика, который потом редактируется человеком.

Далее начинается второй этап – разметка «исходного» текста документа или наложение логической разметки поверх физической. Исходя из целей конкретного исследования, в тексте маркируются объекты, используемые в анализе. Причем, под каждую задачу используется специально созданная разметка (DTD-описание), либо такая разметка создается самим исследователем с нуля. Размеченный документ сохраняется в отдельном файле.

Далее мы работаем уже с размеченными текстами. К ним мы применяем необходимые нам модули анализа, как, например, сбор частотной статистики, создание словаря, глоссария используемых в текстах исторических объектов, кластерный анализ по выбранным параметрам или, например, решается задача идентификации коллекции, описанная далее.

Далее рассмотрим систему более подробно.

4.3 Пользователи системы

В системе существует три группы пользователей, такие как – посетители, администраторы и исследователи. Посетители получают наименьшие права – это в основном, просмотр коллекции документов, а также ознакомление с принятыми к публикации результатами исследований. Для посетителей не требуется регистрация в системе. Администраторы системы в свою очередь обладают полными правами и доступом ко всем материалам и исследованиям, а также выполняют функции модераторов содержания появляющихся текстов. Исследователю необходимо зарегистрироваться в системе, для того чтобы получить возможность начать собственное исследование или присоединиться к существующей исследовательской группе.

4.4 База знаний

Вся сопровождающая исследование документация – постановки и этапы решения задач, комментарии и дискуссии исследователей, описания применяемых методик – оформлены на базе платформы MediaWiki [9]. Wikipedia [16] – всемирная открытая энциклопедия, англоязычная версия которой содержит уже больше миллиона статей, наглядно показывает, что применяемые в ней принципы работают хорошо. Произвольный текст (статья) такого ресурса может быть отредактирован несколькими лицами, зарегистрированными в системе, то есть, можно сказать, что авторами являются несколько человек. Кроме того, храниться вся история изменения текста произвольной статьи и можно в любой момент вернуться к предыдущему или любому другому его состоянию. 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.

5 Подготовка к публикации. Физическая разметка

Вернемся к коллекции. Процесс первоначальной разметки заключается в распознавании изображений документов, расположенных в системе. Принимая во внимание специфику работы исследователей-источниковедов, мы разрабатываем редактор [6], позволяющий совмещать процесс распознавания слов и символов с первоначальной разметкой. Каким образом это происходит? Исследователь видит перед собой в окне браузера изображение документа, выделяет на этом изображении распознанные участки – слова, отдельно стоящие значимые символы, любой другой участок изображения. После этого задает значение для выделенного участка и выбирает его тип (слово, символ, конец предложения и т.д.). На основании введенной информации и выделенного участка изображения создается XML элемент. Упорядоченное в соответствии с представлением документа множество XML элементов будет в последствии представлять первоначально размеченный текст. Для облегчения труда исследователя изображение документа можно приблизить или удалить.

Предполагается, что источники могут быть на произвольных языках (в нашей практике встречались документы на старославянском языке), что создает определенные трудности в процессе разметки. В связи с этим в редакторе существует возможность использовать сущности – конструкции вида &#код; (например, &#1122), которые соответствуют определенным символам Unicode. Для того чтобы избавить исследователя от необходимости точно знать сущности для каждого распознанного им символа, в редакторе присутствует возможность выбрать определенный алфавит из списка предлагаемых. Выбранный алфавит затем отображается в виде виртуальной клавиатуры, которая помогает вводить значения символов.

Редактор создается с применением технологий (стандартов) SVG [13] и AJAX [7]. Первая технология используется для работы с векторными изображениями, она 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. Вторая – привносит удобство в пользование редактором и позволяет работать с ним как с локально установленной программой. Этого удается достичь с помощью подхода к построению пользовательских интерфейсов Web-приложений AJAX (Asynchronous JavaScript and XML), 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.

6 Задачи исследования. Логическая разметка

Если касаться исследовательского слоя, то можно выделить, по крайней мере, три типа характерных задач: построение разметки, классификация документов коллекции и их отдельных частей, а также связывание документов

коллекции. Вместе они образуют логическую разметку документов.

6.1 Построение разметки

Этап создания разметки можно ранжировать как самый важный среди всей цепочки исследования, так как он определяет базис исследовательского аппарата. От того, как исследователь определит структуру документа и интересные ему объекты, зависит и логика всего исследования, и его результаты, и степень точности и адекватности полученной картины реальным событиям. Для максимальной правдоподобности разумно скопировать ту последовательность действий, которой исследователь пользовался бы в ручном, неавтоматизированном варианте исследования. А действовать он мог бы так: отделял бы друг от друга блоки текста, которые казались ему законченными по смыслу, а также несли сходную смысловую нагрузку в разных документах. Кроме того, исследователь помечал бы слова, которые указывали на дату, время, место, фигурантов событий и прочие исторические свидетельства. Акцентируя внимание на наиболее характерных для одного типа документов блоках, исследователь составлял бы абстрактный формуляр этого типа документов. Ориентируясь на этот максимально похожий на остальные документы формуляр, можно составить в итоге необходимую схему разметки.

В системе, не имея готовой разметки, исследователь действует аналогично. Он помечает определенные куски текста тегами, которые сам в тот же момент создает. Изначально тегам присваивается атрибут черновых (draft) и исследователь может произвольно менять их параметры. И лишь добившись необходимой ему системности, а также, проверяя на соответствие стандарту, он фиксирует разметку и сохраняет в виде отдельного DTD-файла, что и является результатом этого этапа.

6.2 Классификация документов и их отдельных частей

Следующий этап один из самых рутинных, он заключается в применении выбранной или построенной на обучающей выборке документов разметки (DTD) ко всей коллекции документов. Далее сюда можно отнести задачи сбора статистики с размеченного текста: сколько раз и где встречаются конкретные упоминания о тех или иных местах, персоналиях, событиях. Или более тонко: как ведут себя отдельные блоки текста, как они выстроены в документах, какие появляются наиболее часто, подтверждается ли абстрактный формуляр, выбранный для данного типа документов.

Или, например, можно поставить задачу определения степени участия отдельных персоналий в историческом процессе периода, охваченного документами. Для этого необходимо будет найти упоминания о персоналии в текстах документов и построить временную линию, на которой будет указано, где и как проявил себя исторический субъект, а также насколько активно было его участие.

Результатом этого этапа будет комплекс документов, размеченных определенным образом, а также хранилище агрегированной метаинформации по этой разметке.

6.3 Связывание документов коллекции

Связывание – это установление какой-либо логической взаимосвязи между документами. Связь может выражаться по-разному, иметь различные типы, веса, может быть направленной или даже циклической.

Если рассматривать конкретную коллекцию документов, то можно поставить более узкую задачу. На данный момент мы работаем с XML-базой данных исторических документов периода 1675-1667 гг., созданной в содружестве с латвийским историком профессором А.С. Ивановым на основе комплекса источников из Российского государственного архива древних актов. Большинство документов комплекса представляют собой «отписки», посланные русскими воеводами из Динабурга (ныне Даугавпилс) в Москву в ответ на царские грамоты. Сами грамоты сохранились частично в виде черновых записей, либо упоминаются или цитируются в текстах отписок.

Возникает задача для указанного комплекса исторических документов произвести вероятностную реконструкцию документооборота, воссоздать содержания документов, используя имеющиеся ссылки в текстах документов, а также некоторые сохранившиеся черновики. Необходимо определить и выстроить все пары «Царская грамота» - «Отписка воеводы» на временной линии. Так как «грамоты» в большинстве своем не представлены в коллекции даже своими черновиками, то они подменяются «фиктивными» документами, в которые заносится информация, найденная в других текстах и связанная с ними, а также предположения и догадки исследователя о содержании документа. Для каждого добавленного сведения в «фиктивном» документе указывается источник информации и степень ее достоверности.

В представленном примере (рис. 2) изображены два реальных документа и три фиктивных, специально введенных в систему. В отписке от 4 августа упоминается грамота с запросом в Друю на муку, а также говорится о том, что оттуда привезли в Динабург рожь, но без повеления царя принять не смеют, да и нечем эту рожь молоть. В грамоте от 12 августа царь велит рожь принять, а также упоминаются распоряжения, отосланные в Дисну и в Друю, с запросами на мельника и жернова. Далее в отписке от 4 августа присутствует помета о том, что получена ответная грамота от царя (от 12 августа), и есть распоряжения в Дисну и Друю.

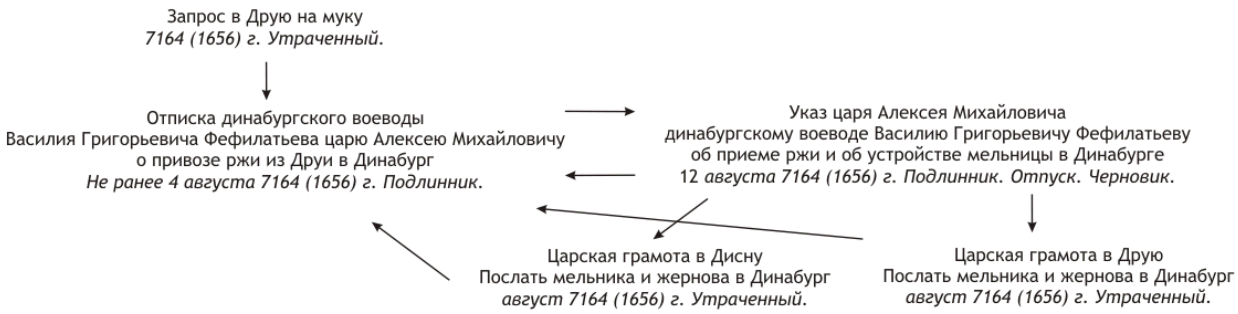

рис. 2

Так как в документе от 4 августа упоминаются все описанные грамоты - имея его, можно ввести в коллекцию четыре остальных. Документы заводятся с определенным вероятностным числом существования. Грамота от 12 августа увеличивает для распоряжений в Друю и в Дисну это число, так как они уже упоминаются дважды.

При связывании документов можно выделить как минимум два типа связей: явные - отписка на конкретную грамоту (указывается в протоколе отписки), и неявные – упоминание о документе в тексте грамоты.

6.4 Описание результата исследования

Как описать результат такого или любого другого исследования [1]? На наш взгляд он должен удовлетворять, по крайней мере, таким требованиям:

- Должен включать в себя постановку задачи, формально описанную, а также отнесенную к какому-либо классу задач;
- Должен содержать описание метода исследования, ссылки на печатные и электронные публикации, а также на описания других исследований в той же информационной системе;
- Должен содержать словесную формулировку результата исследования;
- В этом документе должны быть представлены ссылки на все документы (источники), используемые в анализе;
- Анализируемые документы должны быть представлены таким образом, чтобы в их порядке и описании отражался результат исследования;
- Должны быть указаны все исследователи, работавшие с текстами.

Исходя из этого, результирующий документ содержит 4 основных блока в своем корневом элементе: исследователи, постановка задачи, метод решения, описание результата.

```
<research>
  <researchers group></researchers group>
  <problem></problem>
  <method></method>
  <result></result>
</research>
```

Остановимся подробнее на последнем блоке. В блоке <result>, например, для задачи связывания приводится список документов коллекции, каждый из которых представлен в виде ветви документов, в которой корневым элементом является сам документ, и в него вложены ссылки на все связанные с ним документы.

```
<result>
 <doc id="doc1">
  <linked_docs>
    <link_id="doc3" type="link1"></link>
    …
    </linked_docs>
 </doc>
 <doc id="doc2">…</doc>
</result>
```

Результирующий документ будет содержать избыточную информацию, так как связи между документами указываются для каждого документа, а не для одного из пары связанных документов. Однако к такому документу будет проще выполнять запросы, для детального анализа результата исследования.

Что касается задач второго типа – то там с документа, размеченного по определенной фиксированной схеме, можно собрать обобщенную информацию, специальные агрегированные данные. Данные, подсчитанные какой-либо агрегатной формулой, начиная от суммы элементов одного типа и заканчивая произвольной, специально введенной метрикой.

```
<agregation>
  <shema id="shema1"></shema>
  <doc id="doc1">
    <param id="param1">value</param>
  </doc>
</agregation>
```

Данные сохраняются в отдельном XML-файле, а потом уже из него берутся лишь те из них, которые необходимы для решения конкретных задач и подзадач второго типа - классификации документов коллекции и их отдельных частей. То есть, мы можем посчитать, сколько раз упоминается каждая персоналия в тексте, но в результирующем отчете отражать только тех, кто нам интересен.

```
<result>
 <doc id="doc1">
   <param id="param1">value</param>
 </doc>
 <doc id="doc2">…</doc>
</result>
```

Документы (<doc></doc>) в файле отчета могут содержать блоки, отражающие их свойства, если рассматривается задача классификации, или могут содержать ссылки на другие документы, если решается задача связывания. А могут содержать и те и другие блоки, если постановка задачи более сложная и содержит компоненты разных классов задач.

Заключение

Со времени первой публикации [4], посвященной созданию информационной системы, её описание стало более комплексным и проработанным, а функциональные возможности расширились. Реализованы механизмы ввода документов в систему в виде прототипов визуального SVG-редактора (физическая разметка) и XML-редактора для нанесения логической разметки. Выделены характерные типы исследовательских задач и разработана модель сохранения результатов исследований. Принято решение работать не только с историками и их проблематикой, но и расширить логическую разметку для работы, например, с лингвистами, а в частности с коллективом профессора Баранова [5], который уже начал разработки по обмену данными между различными системами с помощью формата XML-TEI.

Литература

[1] Варфоломеев А.Г., Кравцов И.В. Использование технологии XML для публикации методики и результатов исследования текстов исторических источников // Информационный бюллетень АИК. 2006. №34. С. 64-66.

[2] Иванов А.С., Варфоломеев А.Г. Использование технологии XML для введения в научный оборот комплекса документов «Moscowitica-Ruthenica» // Труды Шестой Всероссийской научной конференции RCDL'2004 (Пущино, 29 сентября - 1 октября 2004 г.). Пущино, 2004. С.285-289

[3] Иванов А. С., Варфоломеев А.Г. Технология XML как инструмент компьютерного источниковедения (на примере формулярного анализа документов приказного делопроизводства) // Круг идей: Алгоритмы и технологии исторической информатики. Труды IX конференции Ассоциации «История и компьютер» / ред. Л.И. Бородкин, В.Н. Владимиров. Москва; Барнаул: Изд-во Алтайского университета, 2005. С.241-281.

[4] Кравцов И.В., Филатов В.О. Проект web-приложения для организации совместной работы с историческими источниками // Информационно-вычислительные технологии и их приложения. Сборник материалов Международной научно-технической конференции. Пенза, 2005. С. 117-119

[5] Манускрипт | Древние славянские памятники, 2005. http://manuscripts.ru/

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

[7] AJAX/ - Wikipedia, the free encyclopedia, 2006 http://en.wikipedia.org/wiki/AJAX/

[8] CELT: The Online Resource for Irish History, Literature and Politics, 2006 http://www.ucc.ie/celt/

[9] MediaWiki, 2006 http://www.mediawiki.org/wiki/MediaWiki/

[10] Medieval Nordic Text Archive (Menota), 2006 http://www.menota.org/

[11] Model Editions Partnership, 2000 http://mep.cla.sc.edu/

[12] Repertorium of Old Bulgarian Literature and Letters, 2005 http://clover.slavic.pitt.edu/~repertorium/

[13] Scalable Vector Graphics (SVG), 2006 http://www.w3.org/Graphics/SVG/

[14] The NEWTON Project | 'Bringing the works of Isaac Newton to life', 2005 http://www.newtonproject.ic.ac.uk/

[15] The Text Encoding Initiative http://www.tei-c.org/

[16] Wikipedia, 2006 http://www.wikipedia.org/

Based on XML information system for working with full-text databases of historical documents

V.O. Filatov, I.V. Kravtsov, A.G. Varfolomeev

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. The system is oriented to the group of researches. They work separately, join into the groups, can correct and add the results of each other. The system is not close and is ready for the information input and output. It is about not only the texts of the origins, but also the intermediate research works results. We also describe a special research toolkit based on two original editors: SVG visual editor and XML mark-up editor.

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