Информационные модели и технологии в организации работы научного сообщества по публикации и анализу коллекций исторических документов♣

© Кравцов И.В.

Петрозаводский государственный университет
ignat@drevlanka.ru

Аннотация

Данная статья представляет собой краткое изложение основных идей и результатов диссертационного исследования, направленного на создание универсальной модели формализации информации, содержащейся в коллекциях текстов исторических документов, и построения информационной системы для упорядочивания и анализа накопленных знаний в рамках работы сетевого сообщества.

Также в работе делается попытка предложить методику построения целого класса информационных веб-систем, предназначенных для цифровой публикации документов культурного наследия нового типа – аналитических и динамических веб-публикаций.

С одной стороны – предметное поле работы достаточно широко и некоторые элементы информационных моделей и инструментов можно встретить во многих существующих Интернет-проектах. В то же время системообразующая модель организации данных и их взаимосвязей позволяет предлагать интерпретации многих существующих инструментов анализа данных, а также предлагать новые инструменты, которые сложно или невозможно построить на классических моделях организации данных в веб-системах.

1 Введение

1.1 Предметная область

Неуклонно растет объем цифровых веб-данных разных форматов, создаваемых и используемых научным сообществом. Прежде всего, это касается естественно-научных областей знаний. Необходимость соединять вычислительные мощности для обработки огромных объемов информации привела к появлению Grid-технологии и развитию e-Science – совокупности программных, технических и методологических средств для обеспечения территориально распределенных научных исследований.

В то же время гуманитарные науки также все чаще оказываются связанными с использованием больших объемов оцифрованных данных. В роли этих данных выступают коллекции текстов в корпусной лингвистике, изображения и тексты печатных источников или рукописей в истории и источниковедении, рисунки и фотографии предметов, привязанные к планам раскопок в археологии, аудиозаписи в устной истории и фольклористике – то, что уже давно получило название «массовые источники».

Достижения в области e-Science приводят к мысли об использовании тех же принципов организации распределенной работы в гуманитарных науках. Но главной задачей в этом случае выступает уже не совместное использование объединенных вычислительных мощностей, а территориальное распределение сбора и хранения самих данных, разработка стандартов для свободного обмена данными, а также сервисов, позволяющих с ними работать. В центре внимания оказывается социальная составляющая – организация работы сетевого научного сообщества.

До сих пор многие исследователи гуманитарии не хотят принимать новые информационные технологии, либо используют их в очень примитивном виде. На текущий момент одной из основных проблем использования информационных технологий является проблема выработки «единого» формата описания метаинформации и содержания текстового источника[11]. В то же время все чаще исследования по истории и лингвистике опираются на большие коллекции текстовых документов. Такие коллекции, представленные в Интернете, составляют основу для формирования сетевых сообществ исследователей, разделяющих между собой

Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

форматы представления данных, а иногда и сами тексты для совместного изучения и редактирования.

1.2 Проекты сетевых публикаций

Всемирная паутина изначально создавалась как среда для научных публикаций, поэтому неудивительно, что веб-сайты, предоставляющие исходные данные для научных исследований, существуют в Сети. Самыми подходящими типами данных для веб-публикаций являются, безусловно, тексты и растровая графика, поэтому с середины 90-х годов широкое распространение получили веб-проекты, посвященные публикациям исторических документов [18-20], а также электронных копий печатных изданий [22]. Такие проекты были реализованы и в России [6,15].

Веб-сайты, посвященные научной публикации исторических документов, можно по способу представления исходной информации разделить на два типа. В одном случае публикация осуществляется в виде базы данных сканированных изображений с метаинформацией об источнике [18,19]. В другом случае сохраняются прежде всего электронные тексты источников в виде полнотекстовых реляционных баз данных или XML-документов [6,15,20]. Конечно, публикации второго типа имеют большие возможности для анализа информации, и, как правило, содержат некоторые инструменты для работы с текстами – хотя бы для отображения текстов на экране в виде, похожем на бумажную публикацию, или для организации полнотекстового поиска. Однако цели таких проектов, как правило, в обоих случаях археографические – сохранение культурного наследия, введение документов в научный оборот, апробация новых технологий публикации источников. Похожие по содержанию и целям веб-проекты объединяются в «консорциумы» [20,22], служащие основой для появления Интернет-сообществ исследователей. Но эти сообщества включают в себя не только тех, кто профессионально занимается публикациями текстовых источников, но и тех, для кого в центре внимания оказываются методы изучения текстов – исследования их структуры, выделения информации, формализации содержания, сравнения и классификации документов. Поэтому естественным направлением развития является переход от археографических к аналитическим веб-публикациям исторических документов.

В таком случае кроме самих коллекций необходимо предоставить сетевой инструментарий для работы с ними. Примером среды для совместной работы с текстами является разрабатываемая в Германии система TextGrid, ориентированная на историко-филологические исследования[23]. Для масштабного проекта «Монастериум»[21], начатого немецкими историками совместно с коллегами из Австрии, Венгрии и других стран, и посвященного созданию электронного архива документов из архивов монастырей Центральной Европы, в университете Кельна (Германия) разрабатывается специализированный редактор EditMom для совместной распределенной работы по оцифровке и ручному распознаванию текстов средневековых грамот.

Обобщение и развитие идей и успехов описанных выше проектов, а также участие в проекте создания системы «Источник» [7], предназначенной для организации работы сетевых сообществ исследователей текстовых исторических источников и реализуемой в Петрозаводском государственном университете, позволило сформировать автору работы концептуальный подход к решению информационных задач в данной предметной области.

Основная идея разработки системы «Источник» и других подобных систем - формировать открытые многофункциональные коллекции текстов, которые могут эволюционировать за счет деятельности организованного вокруг коллекций сообщества с одной стороны, и инструментария для поддержки разносторонней совместной деятельности этого сообщества.

Публикация коллекций документов вместе с методиками и результатами исследований, проведенных на основе этих документов[10], способна изменить традицию и приблизить методологию исторического и историко-филологического исследования к стандартам точных наук. Однако, несмотря на то, что исторические документы публикуются в Интернете уже давно, среди таких публикаций практически отсутствуют проекты, направленные на повышение объективности исследования путем представления его источниковой базы научному сообществу.

1.3 Технологии

Переход к эпохе «Web 2.0» с ее новыми формами взаимодействия пользователей и способами создания контента, а также параллельно развивающаяся XML-революция с повсеместным использованием самоописывающейся, семантической разметки текстов, не могли не затронуть научные веб-публикации текстовых документов. Во-первых, XML-технология оказалась очень удобной как основа для создания полнотекстовых баз данных для источников, не зависимая ни от аппаратного, ни от программного обеспечения пользователей. Создавать свои собственные коллекции текстовых источников, структурируя информацию с помощью XML-разметки, оказалось теперь по силам многочисленным специалистам, традиционно работающим с архивными документами. Если при этом используется какая-нибудь стандартная схема разметки, то пользователи получают возможность обмениваться созданными XML-документами и соединять их в большие коллекции. Подобная технология использовалась и раньше (SGML в проектах TEI и MEP)[20,22], просто сейчас она

стала общедоступной. Во-вторых, появившиеся в последние годы технологии организации работы сетевых сообществ позволяют предоставить возможность формирования веб-публикаций текстовых документов конечным пользователям – нетехническим специалистам в области истории, филологии, социологии и др.

1.4 Цели и задачи работы

Целью данной работы является попытка предложить модель организации многомерного пространства данных и знаний, необходимого для создания современной, аналитической и динамической сетевой публикации, а также архитектуру информационной системы (класса систем) с использованием этой модели. Эти абстракции охватывают собой комплексные методы и технологии автоматизации деятельности научных сообществ гуманитарных дисциплин (лингвистов, историков, источниковедов), способы сохранения исходных исследовательских источников (текстов) и результатов работы исследователей в онлайн пространстве, а также методы связывания исходной и извлеченной информации.

Решаемые задачи:

1) Разработка абстрактной модели описания структуры и семантики источников, а также окружающего их информационного поля;
2) Описание методов и технологий формализации и анализа текстов и коллекций исторических документов, отражение требований этих методов в модели системы;
3) Выработка концепции современной сетевой публикации коллекции исторических источников с учетом возможностей универсальной модели организации данных;
4) Разработка методологии и инструментария взаимодействия в сетевом сообществе;
5) Включение информационного поля сообщества в семантический веб, обеспечение связности с другими системами сети;
6) Проектирование открытой архитектуры информационной системы сообщества, состоящей из набора сервисов и информационных библиотек;
7) Проектирование хранилища данных для консолидации извлеченных из текстов знаний сообщества.

2 Тезисы, выносимые на защиту

2.1 Концепция аналитической публикации

С точки зрения предметной области традиционный подход оформления и издания научной публикации какой-либо коллекции исторических документов всегда связан с весьма продолжительным периодом времени, затраченным, как правило, одним исследователем на анализ большого количества текстовых источников. Большинство работы проделывается исследователем вручную, а в публикации фиксируется только окончательный вариант рассуждений без промежуточных выкладок. Доверие же к результатам исследований обычно базируется на авторитете автора. Кроме того, имеет место проблема практической невоспроизводимости и непроверяемости полученных результатов, так как для проверки необходимо затратить такое же или даже большее количество времени и обладать суммой знаний исследователя.

Для расширения доступности источников историки уже начали использовать пространство Интернет как площадку для размещения электронных копий документов. Публикация осуществляется в виде базы данных сканированных изображений с метаинформацией об источнике, либо сохраняются электронные тексты источников («транскрипции») в виде полнотекстовых реляционных баз данных или XML-документов. Цели таких проектов в обоих случаях археографические – сохранение культурного наследия, введение документов в научный оборот, апробация новых технологий публикации источников. По своему составу и функциональности электронные публикации на данный момент копируют бумажные аналоги или даже уступают им.

На наш взгляд, необходимо существенно шире использовать возможности информационных технологий в процессе подготовки публикаций, а также активнее использовать сетевой инструментарий для анализа оцифрованных текстов. Необходимо повсеместно использовать возможность распределенной удаленной, но в то же время, совместной работы в рамках сетевого сообщества. Сообщества, организованного вокруг специализированного инструментария, позволяющего не только проводить работу онлайн, но и дающего возможность проследить ход каждого исследования, вернуться на произвольный этап исследования, повторить цепочку анализа на примере чужого исследования. Тогда в центре внимания оказываются не сами источники, а методы изучения текстов – исследования их структуры, выделения информации, формализации содержания, сравнения и классификации документов. Меняется сама цель публикации – источники выкладываются в Интернет для обеспечения проведения исследований на их базе[3]. Тогда как оформленные результаты исследований могут стать отдельной цифровой и печатной публикацией.

2.2 Формализация текстов как основа сетевой публикации

Основой практически любого метода исследования текста является некоторая его формализация, то есть замена текста обобщенными

количественными показателями, качественными категориями, либо специальными моделями(графы, деревья)[13], отражающими структуру и тематику текста. Традиционным методом количественного анализа текста является контент-анализ. В этом случае текст формализуется с помощью вектора частот встречаемости составляющих его слов. Анализируя такие вектора, можно, например, решать задачи атрибуции текстов, то есть принадлежности их определенному автору, времени, литературному стилю и т. д. К распространенным методам качественного анализа относятся, например, задачи выделения в тексте ключевых слов, определения тематики текста, составления краткой аннотации документов.

На наш взгляд, использование глубокой разметки текстов с помощью технологий XML позволяет формировать произвольные графовые (сетевые) модели текстов, а также строить более емкие и комплексные инструменты анализа и проводить с помощью компьютера исследования над большими разнородными коллекциями. Кроме того, оформление в виде XML-документов различных интерпретаций или формализаций текстов позволяет считать тексты машиночитаемыми и строить на их основе инструменты автоматического семантического и структурного анализа. Это, например, построение онтологий предметных областей, семантический поиск, всевозможные классификаторы, интеллектуальные агенты, средства поддержки принятия решений на базе многомерных хранилищ данных.

Так как очевидно, что для разных методов анализа и областей применения формализованных текстов потребуются модели разного уровня сложности и детализации, необходимо предложить обобщенный способ описания необходимой единой модели формализации, используемой в системе.

2.3 Модель структурно-семантического пространства данных и знаний

Во многих гуманитарных дисциплинах довольно схожи методики анализа текстовых источников. Все они, так или иначе, пытаются формализовать текст, выделить необходимые категории, построить на уровне абстракции свои модели, провести с построенными моделями характерные исследования и попытаться интерпретировать результат. Для таких задач можно предложить универсальную модель описания формализованного текста и извлеченных из него знаний. Универсальная модель должна удовлетворять следующим требованиям:

- выделять произвольные единицы текста как обособленные объекты;
- формировать связь произвольного числа объектов;
- позволять строить произвольные иерархии объектов и связей;
- соотносить как объекты, так и связи с произвольными смысловыми категориями;
- привязывать к объектам и связям различные показатели (числовые, номинальные, вероятностные и пр.);
- позволять переходить от моделей текстов к моделям более высокого уровня (например, моделям коллекций текстов).

В работе предлагается в качестве такой универсальной модели концепция «структурно-семантического пространства»[9]. Данное пространство состоит из набора измерений, количество которых потенциально бесконечно, и точек в этом пространстве. Каждое измерение является фиксированным набором значений, определяет некоторую шкалу. Каждая точка отражает факт взаимосвязи значений на фиксированном наборе измерений.

При работе с текстом вводится понятие базового измерения – это отложенные на шкале слова, формирующие текст в порядке их появления. Практически любая точка в структурно-семантическом пространстве определяется хотя бы одним базовым измерением и некоторым набором других измерений. Например, для соотнесения элементов текста с определенными смысловыми категориями создается набор точек в двумерном подпространстве, где одно измерение базовое, а другое – шкала категорий объектов. Для создания связи между двумя объектами текста берется два базовых измерения и, если необходимо, измерение категорий связей, и в этом подпространстве ставится точка. Соответственно, для создания N-арной связи берется N базовых измерений.
Таким образом, задача преобразования информации о текстах из любой внутренней структуры хранения в обобщенную модель сводится к разложению информации на оси и точки в подобном пространстве.

2.4 Множественная разметка текстов

Система для научного сообщества исследователей будет обладать мощным научным потенциалом, если будет реализована возможность работы группы пользователей над одним историческим документом, в любых интересующих пользователей дисциплинах. Любое изменение и все исследования исходных исторических документов должны фиксироваться и сохраняться в системе. Таким требованиям вполне удовлетворяет глубокая и множественная XML разметка исходных документов[16, 17].

Все свои идеи и наблюдения исследователь выражает в виде разметки исходного документа, выбрав подходящую для конкретного изучения или же внедрив свою собственную разметку в систему. Размеченные документы сохраняются в базе данных системе в виде отдельных файлов, вариантов исходного текста.

Разметка считается множественной, так как наносится в несколько этапов. Такая разметка состоит из совокупности одноуровневых разметок, которые могут частично пересекаться между собой.

Простейшим этапом разметки является физическая разметка средневековой рукописи. Физическая разметка определяет границы и взаимное расположение словоформ относительно друг друга, а также специальные свойства источника: деление на строки, страницы, описание материала, места хранения, повреждений, встречающихся подписей и печатей и прочее.

На первичную физическую разметку накладывается вторичная и последующие, включающие в себя логические и семантические фрагменты текста. Например, одним из уровней, может быть разметка, выделяющая в тексте упоминания о персоналиях, о географических названиях (города, реки). Другим примером разметки могут выступать лексическая и синтаксическая разметки.

Фрагмент текста, разбитого на словоформы:

```
<doc id="pg002">
 <wf id="wf1">Доброродным</wf>
 <wf id="wf2">и</wf>
 <wf id="wf3">почестливым</wf>
 <wf id="wf4">паном</wf>
 <wf id="wf5">бурмистром</wf>
 ...
 <wf id="wf114">вашей</wf>
 <wf id="wf115">милости</wf>
 <wf id="wf116">листу</wf>
</doc>
```

Далее в этом же тексте выделены крупные блоки двух уровней. На первом уровне блоки «протокол» и «основной текст», на втором уровне сегменты различной семантической окраски(«тип 2», «тип 8»):

```
<doc id="pg002">
<blok type="protocol">
  <wf id="wf1">Доброродным</wf>
  ...
</blok>
<blok type="main_text">
 <seg type="2">
  <wf id="wf32">а</wf>
  ...
  <wf id="wf64">великого</wf>
  <wf id="wf65">короля</wf>
 </seg>
 ...
 <seg type="8">
  <wf id="wf103">и</wf>
  <wf id="wf104">мы</wf>
  ...
 </seg>
</blok>
</doc>
```

Далее пример выделения в тексте произвольной категории, например, персоналии:

```
<cat type="personality"> <wf id="wf102">
Ольбраха</wf></cat>
```

Еще вариант выделения персоналии:

```
<person><wf id="wf102">Ольбраха</wf>
</person>
```

А с добавлением идентификатора такое выделение становится индикатором упоминания в тексте определенного конкретного объекта, а не просто персоналии:

```
<person id="pers01"><wf id="wf102">Ольбраха
</wf></person>
```

Пример выделения индикатора события определенного типа, указание на то, что возможно в тексте речь идет о торговле:

```
<process type="trade"><wf id="wf46">товары
</wf></process>
```

Пример выделения в этом же тексте обращений к адресантам:

```
<doc id="pg002">
 <salutation>
  <wf id="wf1">Доброродным</wf>
  <wf id="wf2">и</wf>
  <wf id="wf3">почестливым</wf>
  <wf id="wf4">паном</wf>
  <wf id="wf5">бурмистром</wf>
 </salutation>
  <wf id="wf6">и</wf>
 ...
  <wf id="wf112">а</wf>
  <wf id="wf113">подлуг</wf>
 <salutation>
  <wf id="wf114">вашей</wf>
  <wf id="wf115">милости</wf>
 </salutation>
  <wf id="wf116">листу</wf>
</doc>
```

Объединении фрагментов, указывающих на один и тот же объект:

```
<doc id="pg002">
<text>
  ...
  <wf id="wf34">от</wf>
 <persona id="pers1">
  <wf id="wf35">господарь</wf>
  <wf id="wf36">наш</wf>
 </persona>
 <persona id="pers2">
  <wf id="wf37">освященный</wf>
  <wf id="wf38">великий</wf>
  <wf id="wf39">король</wf>
 </persona>
 <persona id="pers3">
  <wf id="wf40">его</wf>
  <wf id="wf41">милость</wf>
 </persona>
  <wf id="wf42">сказал</wf>
  ...
</text>
<links>
 <link id="l1" type="join" person_main="pers1"
person_relative="pers2">
 <link id="l2" type="join" person_main="pers1"
person_relative="pers3">
 ...
</links>
</doc>
```

2.5 Многомерное хранилище данных и многомерный анализ

При реализации системы возможно применение некоторых элементов технологии Хранилищ данных (Data warehouse), и тогда модель структурно-семантического пространства представляется в виде многомерной базы данных[9, 13].

В основе хранилищ данных лежит понятие гиперкуба, или многомерного куба данных, в ячейках которого хранятся анализируемые данные.

Факт в терминах хранилищ данных - это числовая величина, которая располагается в ячейках гиперкуба. Измерение - это множество объектов одного или нескольких типов, организованных в виде иерархической структуры и обеспечивающих информационный контекст числового показателя. Измерение принято визуализировать в виде ребра многомерного куба. Объекты, совокупность которых и образует измерение, называются членами измерений. Члены измерений визуализируют как точки или участки, откладываемые на осях гиперкуба.

В реляционном варианте реализации многомерной базы данные распределяются в таблицах двух видов.

Таблица фактов. Является основной таблицей хранилища данных. Как правило, она содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться.

Таблицы измерений. Содержат неизменяемые либо редко изменяемые данные. Каждая таблица измерений должна находиться в отношении «один ко многим» с таблицей фактов.

Схема таблиц, подходящая для разрабатываемой системы, называется «звездой» или «снежинкой». В этих схемах структура данных становится денормализованной, так как в нескольких таблицах дублируются идентификаторы таблиц измерений. Преимуществом же схем «снежинка» является сокращение время получения запросов к часто используемой информации, например, срез гиперкуба по конкретному измерению, по конкретной шкале категорий объектов. Для получения необходимой информации требуется анализ только таблицы фактов.

Измерения структурно-семантического пространства определяют размерность гиперкуба, а точки представляют ячейки гиперкуба. Основной объем данных хранится в таблицах фактов. Если рассматривать разметку текстов в терминах многомерной базы данных, в которой они хранятся, то схемы разметки будут представлены «таблицами измерений», а сама примененная разметка – «таблицами фактов». Ведущим измерением будет поле самого текста, разбитого на словоформы.

Удобство использования подобных структур в том, что при вводе новых измерений структурно-семантического пространства, необходимо лишь затронуть таблицу измерений, не изменяя всей остальной структуры данных. Также, если необходимо добавить какие-то оценочные, весовые и любые другие данные к разметкам текстов, необходимо добавить соответствующие поля в таблицы фактов (приписывание к конкретной реализации разметки текста дополнительных параметров).

В рамках системы многомерная база данных позволит консолидировать данные и знания, полученные через разметку и более удобно реализовывать сервисы анализа данных, сокращая время, требуемое на чтение и разбор XML-файлов.

Кроме того, построение хранилища данных позволит в перспективе применить к нему средства многомерного и интеллектуального анализа данных.

2.6 Информационная система – инструмент, образующий и поддерживающий сообщество

Очевидна необходимость при разработке информационных систем следующего шага – объединения простоты создания новой информации с помощью сетевых технологий и возможностью сделать информацию полезной и качественной с одной стороны, и с другой – способствовать формированию сообщества экспертов, модераторов накапливаемой информации на базе совместной работы в информационном пространстве. Можно сказать[4,12], развитие сообщества такой информационной системы – это постоянный процесс обучения, в котором его члены получают новые знания как в явном (explicit) формализованном виде представленной в системе информации и информационных сообщений друг другу, так и в неявном(tacit) виде - за счет совместного освоения инструментария и методик работы, передачи опыта между участниками сообщества. Работа в сообществе мыслится как работа в групповой сетевой операционной системе. Для каждого исследователя работает правило персонального управления знаниями: «то что я знаю, кого я знаю, и что знают те, кого я знаю». Сама же система является примером организации сообщества практикующих(community of practice).

2.7 Системы аналитической публикации в сети – это системы класса Metaweb

Среди современных тенденций развития Web как такового, то можно выделить две успешные ветви: развитие социальных сервисов, а также ветвь машиночитаемой информации Semantic Web. На наш взгляд, следующей ступенью развития будет являться соединение этих двух подходов в виде области так называемого Metaweb. Если изначальный Web соединял информацию, то Social software соединяет людей, а Semantic Web объединяет машинное знание. Metaweb будет представлять собой человеко-машинные решения для интеллектуальных задач (connects intelligence). Рассмотренный в работе подход как раз определяет описание информации и создание сервисов как

комбинацию человеко-машинного описания извлеченных из текстов знаний, а также использование ручных, автоматических и полуавтоматических инструментов работы.

2.8 Система современной сетевой публикации – это открытая и глобальная система

На текущий момент, даже если публикация готовится в специализированной информационной системе, такой как ИПС «Манускрипт» [15], то дополнительные возможности по работе со структурой и семантикой текстов возможны лишь в рамках этой системы, в большинстве случаев в локальном доступе. Извлечение текстов и информации о них из таких систем осложнено особенностями внутреннего формата хранения. Можно сказать, тексты являются полноправной частью системы, электронной библиотеки и не могут быть отчуждаемы. Поэтому система для сетевого сообщества обязательно должна быть открытой, а тексты и прочая информация - легко извлекаемы.

При соблюдении принципа открытости системы, она становится доступной для использования произвольными сторонними сетевыми сервисами. Особенно обширные возможности несут в себе стремительно развивающиеся инструменты Semantic Web, такие как семантический поиск.

Машинная обработка возможна в семантической паутине благодаря двум её важнейшим характеристикам:

- Повсеместном использовании универсальных идентификаторов ресурсов (URI). Традиционная схема использования таких идентификаторов в современном Интернете сводится к установке ссылок, ведущих на объект, им адресуемый (веб-страница, файл произвольного содержания). Концепция семантической паутины расширяет это понятие, включая в него ресурсы, недоступные для скачивания. Адресуемыми с помощью URI ресурсами могут быть, например, отдельные люди, города и другие географические сущности и т. д. К идентификатору предъявляются несколько простых требований: он должен быть строкой определённого формата, уникальной, а также адресующей реально существующий объект.

- Повсеместном использование онтологий и языков описания метаданных. Таких как семейство форматов, «Semantic Web family»: RDF, RDF Schema или RDF-S, и OWL.

Для создания в системе научного сообщества стандартной машиночитаемой разметки семантической публикации документа используется трансляция данных из XML-формата или хранилища данных в формат RDF.

Также в нашем случае, текстовый исторический объект или его фрагмент, имеющий уникальный идентификатор в рамках глобальной сети (URI, URN и пр.), через этот идентификатор обладает свойством распределенности. Один и тот же объект (источник) может входить в структуру нескольких сетевых информационных систем, изучаться и развиваться динамически усилиями разных людей в разных точках мира.

Понятно, что такое многообразие форм и свободное использование материалов требует регламентов работы и научного упорядочивания, но в то же время такой подход может стать инструментом источниковедения нового поколения.

2.9 Текст-ориентированная разработка

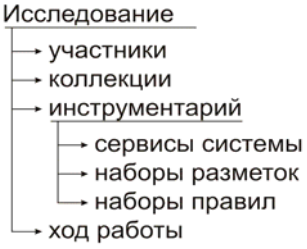
Если попытаться обосновать наш подход к разработке, то он находится в рамках парадигмы Model-driving engineering. Фактически текущим результатом работы является довольно детально описанная модель класса информационных систем. Сам процесс разработки опирается на моделирование структуры и семантики текстов и взаимосвязей текстов и прочих элементов или объектов системы. Моделью считается практически любая формализация текстов и информации сокрытой в текстах, которую мы называем знаниями. Основной технологией записи знаний и прочей информации является XML, потому моделирование определяется и частично ограничено возможностями XML. На данный момент существует большое количество различных проектов на базе XML моделей, но можно уверенно сказать, что в процессе их разработки не использовался подход, предлагаемый нами как основа разработки. Такой подход можно назвать концепцией «текст-ориентированной» разработки (text-driven), когда модули и сервисы системы проектируются так, чтобы передавать друг другу информацию в виде универсальных текстовых документов, файлов в XML-формате.

3. Примеры организации работы сетевых сообществ

3.1 Работа в системе «Источник»

Кроме описанного выше, в рамках системы «Источник» создается библиотека методик и результатов исследований коллекций текстов [8]. Цель такой библиотеки - организация хранилища аналитических публикаций, которые в своем составе содержат не только исходные источниковые материалы и описание полученного результата в виде научной статьи, но и сам инструментарий исследования с промежуточными выкладками.

Рассмотрим структуру информации в библиотеке об одном «типовом» исследовании.



Каждое исследование является фактом связывания набора составляющих его объектов и имеет собственный уникальный идентификатор и также является отдельным объектом системы.

В рамках одного исследования определяется состав участников, работающих над ним. Как правило, это руководитель исследования, который определяет состав остальных объектов исследования, и исполнители, которые получают возможность работать с определенными в исследовании объектами.

Исследование проводится на основе фиксированной коллекции документальных источников. Как правило, это полная обособленная коллекция документов.

Исследование заключается в применении к источниковому материалу специализированного инструментария: программных сервисов и алгоритмических модулей, фиксированных наборов структурных и семантических разметок, определенных на данных разметках наборов правил получения результатов.

Инструментарий применяется последовательно, согласно заранее описанному ходу работы. Например, сначала производится физическая разметка текстов (выделение словоформ), затем – структурно-семантическая разметка, после этого производится обработка полученной структуры с помощью заранее определенных правил (функций, преобразований) с целью получения новой разметки или новых правил, содержащих выявленные закономерности информации.

Результат всего исследования либо отдельной его стадии записывается в виде XML-документа, состоящего из трех частей: результата, правил, посылок. В качестве посылок (аргументов правил) выступают первично размеченные тексты, результатом является новая разметка или новые правила (закономерности).

В случае отсутствия заранее разработанного инструментария для анализа проводится «экспериментальное» исследование, результатом которого являются новые, выработанные исследователем, схемы разметки текстов или правила получения результатов.

Накопленные результаты могут быть рассмотрены как исходный материал для выдвижения и проверки новых гипотез и проведения новых исследований.

Кроме рассмотренного разреза «исследования», навигацию по библиотеке можно будет осуществлять на базе остальных составляющих ее объектов. Например, при просмотре с точки зрения участников можно будет видеть, в каких исследованиях они участвовали. Для коллекций текстов можно будет просмотреть примененные к ним разметки и правила, и, наоборот, для той или иной разметки получить примеры её использования.

3.2 Сообщество «Письменное наследие»

Кроме реализации в системе «Источник», планируется апробация полученных результатов при формировании сообщества исследователей древнерусских текстов. Проект предусматривает создание единого информационного портала "Письменное наследие"[1,11], в рамках которого должны быть подготовлены удобные инструменты для совместного решения учеными России и других стран актуальных сегодня технологических задач в области электронного хранения, представления в Интернете, исследования и популяризации древних и средневековых письменных памятников, для координации работ в области подготовки электронных полнотекстовых ресурсов, электронных описаний и каталогов и электронных словарей, в области выработки стандартов обмена данными. Кроме того, проект предусматривает создание организационной и технологической платформы для развития и поддержки единого исследовательского, учебного и информационного пространства, объединяющего текстовые ресурсы, справочные, аналитические и информационные материалы, рабочие группы, исследовательский инструментарий.

4. Заключение

Упомянутые выше, а также другие идеи и результаты, выносимые на защиту, докладывались на конференциях: RCDL (2003-2008), Современные информационные технологии и письменное наследие (2006, 2008), конференциях Ассоциации «История и Компьютер» (2006, 2008), Интернет и современное общество (2006, 2007), Научный сервис в сети Интернет (2007), Научных чтениях Даугавпилского университета (2008, 2009), и были опубликованы в работах [1-5,8-14,16,17].

Литература

[1] Баранов В.А., Кравцов И.В. Интернет портал «Письменное наследие». Формирование сообщества исследователей древних текстов // Интернет и современное общество : Труды X Всероссийской объединенной конференции. – СПб. : Факультет филологии и искусств СПбГУ, 2007. – С. 57 – 60.

[2] Варфоломеев А.Г., Кравцов И.В, Москин Н.Д. Проект специализированного Интернет-ресурса для представления и анализа фольклорных песен // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Пятой Всероссийской научной конференции RCDL'2003. СПб, 2003. С.339-343.

[3] Варфоломеев А.Г., Кравцов И.В. Аналитические Web-публикации исторических документов // Научный сервис в сети Интернет: многоядерный компьютерный мир. 15 лет

РФФИ: Труды Всероссийской научной конференции. М.:Изд-во МГУ, 2007. С.389-390.
[4] Варфоломеев А.Г., Кравцов И.В. Приобретение и представление знаний в сетевом сообществе исследователей текстов // Вторая Международная конференция "Системный анализ и информационные технологии" САИТ-2007: Труды конференции. В 2 т. Т.1. М., 2007. С.104-106.
[5] Варфоломеев А.Г., Кравцов И.В., Филатов В.О. SVG-визуализация в цифровых библиотеках рукописных документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Девятой Всероссийской научной конференции RCDL'2007. Переславль-Залесский: Изд-во "Университет города Переславля", 2007. С.230-235.
[6] Древнерусские берестяные грамоты. Сайт проекта, 2009. http://gramoty.ru
[7] Источник. Сайт проекта, 2009. http://istochnik.karelia.ru
[8] Каргинова Н.В., Кравцов И.В., Москин Н.Д., Варфоломеев А.Г. Проект электронной библиотеки методик и результатов исследований текстовых коллекций для системы "Источник" // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Десятой Всероссийской научной конференции "RCDL'2008". Дубна: ОИЯИ, 2008. С.239-245.
[9] Кравцов И.В. Моделирование структуры и семантики текста в информационных системах для исследования исторических документов // Системы управления и информационные технологии. - № 1.1(31) – Москва-Воронеж : Научная книга, 2008. – С. 163 – 167.
[10] Кравцов И.В. О возможностях информационных технологий в подготовке публикаций и организации исследований комплексов исторических документов // Vēsture: avoti un cilvēki. XVIII Zinātniskie lasījumi. Vēsture XII. Daugavpils, 2009. P.110-114.
[11] Кравцов И.В., Багимова К.А. Модель обмена знаниями в системах гуманитарных исследований // Материалы международной научной конференции «Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам». Казань. 2008. С. 163-167.
[12] Кравцов И.В., Варфоломеев А.Г. Принципы организации информационного пространства сетевого сообщества исследователей рукописных текстов // Информационное общество. Интеллектуальная обработка информации. Информационные технологии. Материалы 7-ой международной конференции НТИ-2007. Москва : Изд-во ВИНИТИ, 2007. С.383-386.
[13] Кравцов И.В., Филатов В.О. Информационная система для работы с коллекциями рукописных исторических документов. // Информационные технологии моделирования и управления, 2007, №2(36). - С. 188-195.
[14] Кравцов И.В., Филатов В.О. Подходы к организации совместной работы научного сообщества в области публикации и исследования средневековых текстов // Интернет и современное общество. Труды IX Всероссийской объединенной конференции. – СПб: СПбГУ, 2006. – С.77-79.
[15] Манускрипт. Древние славянские памятники. Сайт проекта, 2009. http://manuscripts.ru
[16] Филатов В.О., Кравцов И.В. Технологии создания информационной системы для работы с полнотекстовыми базами данных исторических документов // Материалы международной научной конференции «Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам». Ижевск. 2006. С. 168-173.
[17] Филатов В.О., Кравцов И.В., Варфоломеев А.Г. Информационная система для работы с полнотекстовыми базами данных исторических документов на основе технологии XML // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Восьмой Всероссийской научной конференции (RCDL'2006). – Ярославль: ЯрГУ им. П.Г.Демидова, 2006. - С.337-344.
[18] Codices Electronici Ecclesiae Coloniensis (CEEC) Web site, 2009. http://www.ceec.uni-koeln.de
[19] Codices Electronici Sangallenses (CESG). Web site, 2009. http://www.cesg.unifr.ch
[20] Model Editions Partnership Web site, 2009. http://adh.sc.edu
[21] Monasterium Project Web site, 2009. http://monasterium.net
[22] Text Encoding Initiative Web site, 2009. http://www.tei-c.org
[23] TextGrid Project Web site, 2009. http://www.textgrid.de

Information models and technologies for the web community of researchers in the field of historical documents publication and analysis

I. Kravtsov

This article represents a summary of the basic ideas and results of dissertational research. Researcher describe universal formal model of the information extracted from collections of texts of historical documents. Principles of construction of an intelligence system for marshaling and the analysis of the stored knowledge within the limits of operation of network community are besides described.

♣ Статья подготовлена в рамках проекта, поддержанного грантом РГНФ (проект № 08-01-12136в).