SVG-визуализация в цифровых библиотеках рукописных документов♣

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

Петрозаводский государственный университет
avarf@psu.karelia.ru

Аннотация

Данная статья посвящена применению формата векторной графики SVG в качестве технологической основы для визуализации информации в цифровых библиотеках рукописных исторических документов. Основное внимание уделено четырем вариантам использования SVG: для связи XML-разметки текстов с изображением исходных документов, для определения шрифтов, адекватных оригиналу, для визуализации аналитических запросов к коллекции документов, и для создания визуальных инструментов исследования текстов.

Рассмотренный подход используется авторами на практике в разработке информационной системы «Источник», предназначенной для организации работы сетевых сообществ исследователей текстовых исторических источников.

1 Введение

Задача оцифровки исторических документов и организации доступа к их электронным копиям – одна из важнейших на современном этапе развития информационных технологий. Помимо сугубо научных целей – введения в научный оборот больших массивов источников в машиночитаемой форме, у этой задачи есть и общегуманитарная составляющая – сохранение для потомков культурного наследия, накопленного за много веков в архивах. Для того, чтобы обеспечить доступ к этому наследию, необходимо использовать Интернет-технологии. Существует множество проектов создания цифровых библиотек рукописных документов, например, [1-3]. Полнотекстовый характер исходной информации и естественная иерархическая структура документов приводят к выбору XML в качестве основы для организации хранения текстов. Ценность XML состоит в его универсальности и самоописываемости, благодаря чему различные XML-форматы можно использовать в рамках одного проекта для различных целей, произвольно их комбинируя и при этом всегда оставаясь в рамках стандарта. Особенности разработки электронных архивов рукописных исторических документов на базе XML-технологии подробно описаны в [4-10]. В данном докладе мы хотели бы подробнее остановиться на важном аспекте предлагаемого нами подхода – широком использовании SVG (XML-формата векторной графики) [11] для визуализации информации на разных этапах работы с документами. Данный подход используется нами в разработке информационной системы «Источник» [12], предназначенной для организации работы сетевых сообществ исследователей текстовых исторических источников.

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.

2 Варианты использования SVG в цифровых библиотеках

2.1 Использование SVG для связи XML-разметки с изображением документа

Работа с электронной копией рукописного исторического документа начинается с ввода в компьютер растрового изображения оригинала, полученного с помощью сканера или цифрового фотоаппарата. Такое изображение считается «оригиналом документа» в нашей системе и служит отправной точкой для всех дальнейших преобразований документа. Хранение оригинала, безусловно, необходимо для того, чтобы иметь возможность в любой момент провести преобразования заново для проверки их правильности или изменения методики исследования. Первым преобразованием является распознавание текста с разбиением его на строки, словоформы, предложения, выделением надстрочного текста и диакритических знаков. Это преобразование выполняется вручную, поскольку современные программы OCR не способны корректно распознавать рукописный текст. В результате мы должны получить XML-файл с первичной разметкой. Технология SVG позволят связать этот XML-файл с изображением-

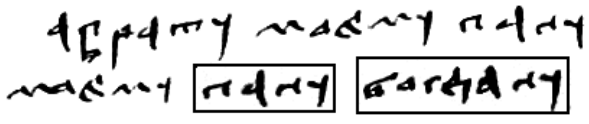
оригиналом. Для этого мы предлагаем создавать дополнительный SVG-файл, содержащий в себе объекты-прямоугольники, отмечающие расположение словоформ, надстрочного текста и диакритических знаков на изображении документа. Связь прямоугольников с соответствующими элементами XML-разметки реализуется с помощью атрибутов-идентификаторов. Прямоугольники соединяются в группы объектов, обозначающие расположение строк и предложений. Впоследствии, при добавлении в XML-файл логической и семантической разметки, она также может быть визуализирована на изображении-оригинале, поскольку любой фрагмент текста состоит из набора словоформ, местоположение которых определено в SVG-файле.

SVG-связывание XML-разметки текста с изображением документа позволяет решить две важные задачи:

заменить ручное распознавание всего текста в целом на распознавание значения отдельных элементов изображения – словоформ, которое поддается некоторой автоматизации. Так, в нашей системе уже распознанные словоформы записываются не только в XML-файл, но также в базу данных словоформ, и затем, при вводе очередной словоформы, самые подходящие из них появляются в виде динамически изменяющегося списка подсказок, реализованного с помощью технологии AJAX [13], что значительно облегчает трудоемкий процесс распознавания текста.

визуализировать результаты выполнения запросов к XML-документам, используя для этого их изображения. Речь идет о запросах на поиск определенных словоформ, персоналий, географических названий, логических частей документов. Вывод результатов таких запросов в виде соответствующих фрагментов изображений-оригиналов повышает аналитические возможности цифровой библиотеки рукописей, имитируя работу с настоящими документами в настоящем архиве.

Рассмотрим пример выделения областей на изображении документа и связывания их с XML-файлами с оригинальным и адаптированным текстами. На следующем рисунке представлен фрагмент одного из документов[i] комплекса «Moscowitica-Ruthenica» [4], находящегося в Латвийском государственном историческом архиве (ЛГИА), с выделенными на нем двумя прямоугольными областями, соответствующими двум словоформам.



Для этого рисунка ниже представлен фрагмент соответствующего SVG-файла.

```
<?xml version="1.0" encoding="UTF-8"
standalone="no"?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG
1.0//EN"
"http://www.w3.org/TR/SVG/DTD/svg10.dtd">
<svg version="1.0"
xmlns="http://www.w3.org/2000/svg"
xmlns:xlink="http://www.w3.org/1999/xlink"
zoomAndPan="enable"
width="1280" height="1280">
<image x="0" y="0" width="900" height="997"
xlink:href="005_MIN.JPG">
<g id="markup" stroke="black" stroke-width="0.5"
fill="black"
fill-opacity="0.1">
....
  <rect id="r104" x="322" y="296" width="95"
height="27">
  <rect id="r105" x="420" y="299" width="111"
height="26">
....
</g>
</svg>
```

В приведенном SVG-коде теги <rect> задают прямоугольные области, тэг <g> - группу прямоугольников, которая может быть строкой или законченной фразой. Тэг <image> используется для вывода растрового изображения исходного документа.

Данному SVG-файлу соответствуют два XML-файла, содержащие оригинальный текст на древнерусском языке и адаптированный современный текст. В обоих файлах вместо прямоугольных областей используются тэги словоформ <w>, имеющие такое же значение атрибута id, что и прямоугольники Так, например, выглядят две выделенные на рисунке 1 словоформы в XML-файле с адаптированным написанием:

```
<w id="r104">пану</w>
<w id="r105">Богдану</w>
```

Так как все способы представления одной словоформы связаны одним и тем же идентификатором, исследователи всегда могут перейти от одного способа к другому. Выделять логические фрагменты, сравнивать структуру документов удобнее с использованием адаптированного текста, но сами тексты часто требуется отображать в форме, максимально близкой к оригиналу.

2.2 Использование SVG-шрифтов для описания и визуализации символов

При распознавании текста средневековых рукописей часто требуется прежде всего определить множество используемых в тексте символов, связать их с определенными кодами стандарта Юникод [14] и подобрать соответствующий шрифт, содержащий в своем составе все необходимые символы. Эта задача может оказаться непростой для больших коллекций документов, содержащих

тексты на разных языках. Во-первых, варианты написания одного и того же символа могут зависеть от времени и места создания документа. Во-вторых, далеко не для всех вышедших из употребления символов можно найти однозначное соответствие в стандарте Юникод. В-третьих, шрифты, использующие кодировку Юникод и имеющие в своем составе подходящие символы, могут не подойти по начертанию глифов, а разработка собственного TrueType-шрифта является сложной задачей.

Альтернативой использования TrueType-шрифтов с жесткой привязкой к стандарту Юникод является применение формата SVG для описания желаемой формы символов в виде набора кривых Безье. Каждый символ представляется в SVG-файле в виде группы объектов, снабженной идентификатором. Конечно, символ может иметь и другие характеристики, в том числе и коды близких символов Юникода. Набор идентификаторов, представляющих символы словоформы, записывается в базу данных наряду с ее значением. В результате мы получаем возможность выводить текст из XML-файла в окно браузера не в виде HTML с использованием TrueType-шрифта, а непосредственно в виде SVG-графики.

В качестве примера рассмотрим одну из букв «аз» из документа, фрагмент которого приведен на предыдущем рисунке:

Для описания этого символа с помощью кривых Безье достаточно следующего короткого фрагмента SVG-кода:

<g id="az2"><path d="M 113,126 C 107,78 108,79 80,85 C 66,88 51,91 47,91 C 33,91 38,76 55,63 C 74,49 84,53 71,68 C 64,76 65,78 72,78 C 94,78 100,67 93,43 C 87,25 87,17 95,8 C 103,-2 104,-2 117,12 C 124,20 130,36 131,46 C 132,57 133,75 134,86 C 136,115 130,157 123,157 C 120,157 116,143 113,126 z " style="fill:#000000"/></g>

Отметим, что использование SVG-шрифтов, как и SVG-связывания, приводит к усилению аналитических возможностей цифровой библиотеки, поскольку предоставляет пользователям не только произвольно формировать множество символов для электронных текстов, но и сравнивать особенности начертания одних и тех же символов в разных документах – контуры, описанные с помощью SVG, становятся доступными для количественного анализа.

2.3 SVG-визуализация результатов аналитических запросов к документам

Ценность цифровой библиотеки, на наш взгляд, состоит в том, что она не только предоставляет научному сообществу оцифрованные тексты, снабженные средствами поиска, но и дает возможность соединять коллекции текстов документов со средствами их анализа. Под анализом текстов мы понимаем не только количественный анализ, основанный на замене текстов или их частей математическими объектами (векторами, графами) и операциях с этими объектами, но и качественный анализ, основанный на выборе формы представления текстов, удобной для обнаружения исследователем скрытых закономерностей. Одним из методов качественного анализа является визуализация информации.

Графические образы для визуализации информации в XML-базах данных должны быть существенно изменены по сравнению со стандартными графиками и диаграммами, используемыми для визуализации табличных данных. Мы предлагаем использовать в нашей системе несколько видов визуализации информации, показанных и перечисленных ниже:

«облака тэгов», получаемые при запросах на поиск определенных элементов в XML-документах. Облака состоят из SVG-прямоугольников и показывают местоположение интересующих нас элементов в документе. С помощью такой визуализации можно сравнивать структуру различных документов между собой.

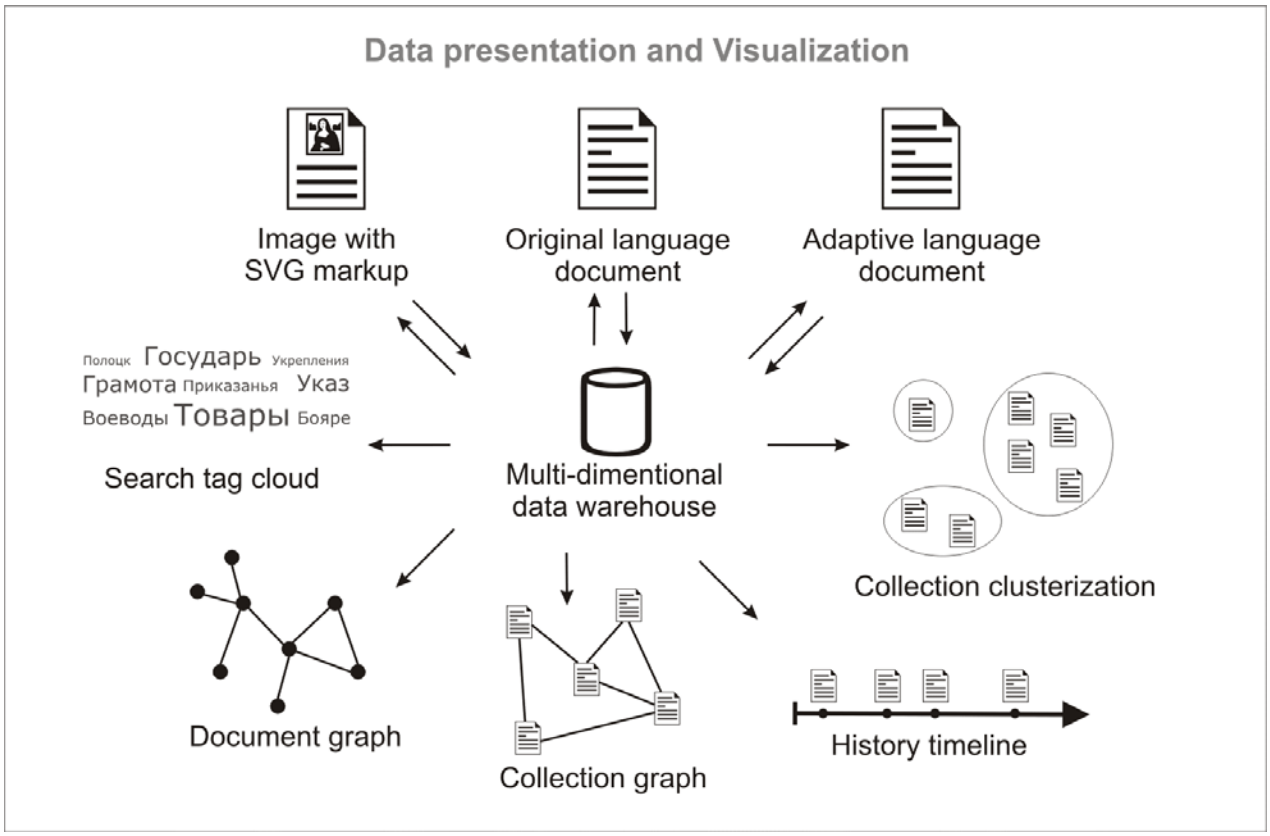
«граф документа» - показывает порядок следования логических и семантических элементов и связи между элементами внутри XML-документа

«граф коллекции документов» - демонстрирует связи между документами. Связи могут быть явными (в виде прямых ссылок) или неявными (например, упоминания одних и тех же событий или персоналий).

«временная линейка» - показывает расположение документов на оси времени, если известна их датировка

«кластеризация коллекции» - эта визуализация основана на предварительном кластерном анализе коллекции документов. Существует много алгоритмов определения расстояния между строками, графами, деревьями и XML-документами. Если на основе какого-либо из таких алгоритмов получить матрицу расстояний между документами коллекции и применить один из методов кластерного анализа, то можно распределить XML-документы на группы похожих между собой документов и визуализировать это распределение в виде SVG-изображения.

Все собранные с текстов коллекций знания кроме XML файлов сохраняются в нашей системе в реляционной многомерной базе данных, которая активно используется для анализа текстов. Если рассматривать разметку текстов в терминах многомерной базы данных, то схемы разметки будут представлены «таблицами измерений», а сама примененная разметка – «таблицами фактов». Ведущим измерением будет поле самого текста, разбитого на словоформы.



Информация базы данных позволяет формировать исследовательские XML-документы произвольного вида, которые выступают с одной стороны как публикация результатов какого-либо исследования, либо как исходные файлы для проведения дальнейшего исследования или построения различных визуализаций информации.

Например, для построения визуализации «Временная линейка» (timeline) необходимо создать XML-файл, содержащий даты создания документов коллекции, для которых строится визуализация. Также он должен содержать элемент, определяющий начало и конец периода, для которого строится визуализация. Такой файл может выглядеть, например, так:

```
<report>
  <period>
     <start>1650</start>
     <end>1670</end>
  </period>
  <doc doc_id=”d146” date=”13.08.1656”>
  <doc doc_id=”d148” date=”17.02.1663”>
   …
</report>
```

Имея такой документ, достаточно несложно с помощью простого XSLT-преобразования [15] построить SVG-документ, воспроизведение которого в окне браузера даст необходимую картинку.

Преимущества использования SVG-визуализации на этапе анализа текстов цифровой библиотеки в том, что алгоритмы визуализации могут быть технически реализованы в виде XSLT-преобразований из XML в SVG, а значит, открыты для модификаций. Более того, сообщество исследователей может разрабатывать свои собственные алгоритмы визуализации и добавлять их в систему.

2.4 Визуальные инструменты исследования

Визуализацию можно использовать не только для различных форм отображения исходной и полученной в процессе исследования информации, но и в качестве инструментального средства формирования новой информации и знаний. Людям, как правило, намного проще оперировать визуальными образами при формировании своих умозаключений, при построении взаимосвязей между объектами или упорядочивания их, чем использовать для этого таблицы или размеченные тексты.

В рамках нашей информационной системы мы предлагаем использовать инструмент визуального анализа текстов. Он построен на основе идеи использования специальных графов, которые, следуя [16], мы будем называть «моделями» и «историями». Они строятся для некоторого набора объектов (узлов), в качестве которых могут выступать элементы текста, выделенные с помощью какой-либо произвольной семантической разметки, например, персоналии, географические объекты и предметы обихода. «Модели» не линейны, определяют произвольные ненаправленные отношения между узлами. «Истории» линейны, определяют однонаправленные связи между узлами, выражающие их следование друг за другом в тексте или во времени.

Довольно часто необходимо отбросить контекст для упрощения информационного поля и работы с выделенными объектами на «чистом листе». Так, предлагаемый нами инструмент, представляет собой специализированный SVG-редактор для объектов в виде прямоугольников, которые можно перемещать и соединять линиями-связями. Если все же контекст для построения связей необходим, то можно отображать область текста и графическую область редактора одновременно и работать с ними параллельно и интерактивно.

На рисунке изображен фрагмент перевода грамоты из коллекции приказных документов по истории города Динабурга [17, с.10], выделенные в тексте объекты, а также SVG-схемы «модель» и «история», построенные для этих объектов.

Построенные SVG-схемы можно разбирать и преобразовывать в XML-файл, отражающий результаты исследования.

3. Заключение

В качестве заключения мы хотели бы показать формы участия визуализаций информации в процессе формирования и развития в нашей системе новых знаний о текстах.

Все чаще, рассматривая вопросы публикации и поиска информации в сети Интернет, говорят о семантической паутине или Semantic Web. Эта концепция развития всемирной паутины нацелена на изменение процесса публикации информации, на добавление к ней описания, понятного компьютеру. Такие описания компьютер сможет анализировать с точки зрения смыслов, семантической нагрузки. Информационное сообщение становится «двойным текстом», рассказывающим не только человеку, но и компьютеру о смысле своего содержимого. Достигается это на современном этапе, как правило, с помощью XML-разметки текста.

Под текстом для компьютера в нашей системе мы понимаем всю информацию, добавленную людьми с помощью различных XML-схем. Анализируя и комбинируя эти новые данные, компьютер выдает нам сведения или результаты их анализа в визуальной форме, на основании которых мы уже можем сформировать следующую порцию знаний о текстах, строить догадки на основе визуальных образов, которые никогда не смог бы сделать сам компьютер. Эти заключения исследователь опять же оформляет в виде новой или дополнительной разметки текстов.

Таким образом, благодаря применению визуализации цепь трансформации знаний о текстах «человек-компьютер-человек» замыкается и превращается в спираль развития.

Литература

[1] Манускрипт. Древние славянские памятники (http://manuscripts.ru)

[2] Новгородские берестяные грамоты (http://gramoty.ru)

[3] Menota (Medieval Nordic Texts Archive) (http://menota.org)

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

[5] 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). Суздаль, 17 - 19 октября 2006 г. – Ярославль: ЯрГУ им. П.Г.Демидова, 2006. - С.337-344.

[6] 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.

[7] 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.
[8] Варфоломеев А.Г., Кравцов И.В. Использование технологии XML для публикации методики и результатов исследования текстов исторических источников // Информационный бюллетень Ассоциации «История и компьютер». № 34. Материалы X конференции АИК. Май 2006 г. – М.;Тамбов: Изд-во Тамбовского университета, 2006. – С.63-64.
[9] Hansen, A. M. Text Encoding of Manuscripts: Danish Prayer Books from the 16th Century // Le Médiéviste et l'ordinateur. no.41. 2002. (http://lemo.irht.cnrs.fr/41/mo41-09.htm).
[10] The Menota Handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources. Version 1.0. Bergen: Medieval Nordic Text Archive, 2003. (http://helmer.aksis.uib.no/menota/guidelines/)
[11] Scalable Vector Graphics (http://www.w3.org/Graphics/SVG/)
[12] Источник. Информационная система для работы сообществ исследователей текстов. (http://istochnik.karelia.ru)
[13] Крейн Д., Паскарелло Э., Джеймс Д. AJAX в действии. М.: Издательский дом «Вильямс», 2006. 640 с.
[14] Unicode (http://www.unicode.org)
[15] Холзнер С. XSLT / Библиотека программиста. СПб.: Питер, 2002. 544 с.
[16] Ryan Coleman. An Introduction to Visual Thinking (http://www.slideshare.net/rycoleman/an-introduction-to-visual-thinking)
[17] Иванов А., Кузнецов А. Динабург в документах Российского государственного архива древних актов (1656-1666). Т.2. Даугавпилс, 2002.

SVG-visualization for digital libraries of hand-written documents

A.Varfolomeyev, I.Kravtsov. V.Filatov

Our article covers various terms and development technologies for Scalable Vector Graphic using in digital libraries of handwritten historical documents.

Full-text nature of the initial information and natural hierarchical structure of the documents define XML-technology as a choice and a basis for texts storing. But, if we use XML for allocation of logic elements in the texts of our library, it looks logically to apply the same technology to other purposes - for example, for building of queries to a collection of the documents or for visualization of the information at different stages of our work with the texts.

In this article, the basic attention is given to four variants of SVG using. We talk about making dependencies between XML-markup and images of initial documents. We describe special vector fonts definition for adequate representing of original texts. We demonstrate different forms of visualization as results of analytical queries to a collection of the documents. We also offer to use SVG-based editor for graphs models creating.

The considered approach is used in practice in development of special toolkit for information system "Istochnik" ("Source") intended for network community of the researchers.

♣ Статья написана в рамках проекта, поддержанного грантом Российского гуманитарного научного фонда (проект № 06-01-12124в).

[i] Грамота полоцкого воеводы Олехны Судимонтовича рижскому магистрату о готовности литовских послов начать переговоры с ливонскими послами в Индрице 15 сентября. 30 августа [1471 г.] ЛГИА, ф.673, оп. 4, ящик 19, № 7, л. 2 – 2об.