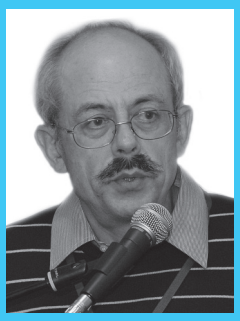

ШВАРЦМАН Михаил Ефремович *- начальник отдела исследования компьютерных систем Российской государственной библиотеки*
Адрес: 119019, г. Москва, ул. Воздвиженка, 3/5
e-mail: shvar@rsl.ru

*Формат METS/ALTO как альтернатива PDF для сложно структурированных научных текстов XVIII–XIX веков**

Изучая издания, отсканированные в соответствии с работой по гранту РФФИ **11-07-00750-а** «Разработка технологии создания и поддержки электронной коллекции уникальных произведений великих русских ученых XVII-XIX веков», мы стали задумываться о том, в каком виде лучше всего предоставлять эти издания. Для этого сначала нужно было ответить на вопрос: а в чем, собственно, могут быть их отличия от других электронных изданий и коллекций.

Во-первых, следует отметить разносторонние интересы многих авторов. Так, например, в каждом из томов полного собрания сочинений М.В. Ломоносова одновременно могут быть стихи, научные статьи, письма и др. Для полноценного их изучения недостаточно просто отсканированных страниц. Как показывает опыт работы по изучению творчества Ломоносова в рамках проекта Фундаментальная электронная библиотека «Русская литература и фольклор» (*http://feb-web.ru/feb/lomonos/default.asp*), для детального анализа творческого наследия нужно иметь множество вспомогательных указателей. Для подготовки таких указателей первым делом следует максимально полно разметить отсканированный документ, разделить его на составляющие части. Каждую из этих частей следует детально каталогизировать для будущих исследователей. Это относится не только к ученым-энциклопедистам. Если мы посмотрим на материалы съезда русских естествоиспытателей и врачей (1889-1890; Петербург) *http://dlib.rsl.ru/01003926249*, то увидим среди выступающих Менделеева, Склифосовского, Столетова (**рис. 1**) и других ведущих российских ученых. Несомненно, такие материалы нельзя давать целиком. Они должны быть описаны постатейно.

Во-вторых, многие издания, напечатанные в XVII-XVIII веках, имеют специфические шрифтовые особенности, которые затрудняют автоматическое распознавание текста. Так, например, в приведенной выдержке (**рис. 2**) из издания «Дневныя записки путешествия доктора и Академии наук адъюнкта Ивана Лепехина по разным провинциям Российскаго государства» *http://dlib.rsl.ru/viewer/01004093992#?page=6* мы видим, что часть букв смазана, некоторые буквы пропечатаны не полностью. Поэтому для полноценного анализа текста необходимо проводить редактирование текста, полученного после распознавания.

Третьей особенностью является наличие большого количества иллюстративного материала (географических карт, рисунков и т.п.), которые для полноценной передачи имеющейся на них информации должны быть отсканированы в большом разрешении, с максимальной цветностью и сжаты без потерь. Таковы, например, гравюры в издании «П.С. Палласа, доктора медицины, профессора натуральной истории и члена Российской императорской Академии наук, и Санктпетербургскаго Вольнаго экономическаго общества, также Римской императорской академии изпытателей естества и Королевскаго Аглинскаго ученаго собрания, Путешествие по разным провинциям Российской империи. - Санктпетербург: При Имп. Акад. наук, 1773-1788. - 4°» или в книге «Записки о Монголии, сочиненные монахом Иакинфом с приложением карты Монголии и разных костюмов. - Санкт-Петербург: Тип. Карла Крайя, 1828. - 22 см.» Для работы с такими рисунками недостаточно включать их в общий с остальными страницами файл в формате PDF, как обычно делается в электронных библиотеках. Большой размер этих страниц будет увеличивать размер файла и время, необходимое для скачивания.

* *Работа выполнена в рамках гранта РФФИ 11-07-00750-а.*

Общій отдѣлъ. 1-е общее собраніе (1); гр. *Дѣляновъ*—привѣтствіе (1), телеграммы (2); *Бекетовъ* — привѣтствіе и отчетъ крымскаго комитета (3); городской голова *Лихачевъ* — привѣтствіе и о санитарномъ состояніи Петербурга (8); *Менделѣевъ*—пріемы естествознанія въ изученіи цѣнъ (11); *Склифасовскій* — объ одной изъ нуждъ нашего врачебнаго образованія (17); выборы (19).

2-е общее собраніе (20); *Столѣтовъ* — эфиръ и электричество (20); *Фаминцынъ* — о психической жизни простѣйшихъ живыхъ существъ (32, таблица рисунковъ); *Вагнеръ* — взглядъ физіологіи и психологіи на явленія гипнотизма (39); *Густавсонъ* — о микробіологическихъ основаніяхъ агрономіи (46).

Рис. 1. *Материалы съезда русских естествоиспытателей*

В настоящее время общепринятой является практика использования формата PDF для представления полных текстов в электронных библиотеках. У этого формата, безусловно, есть много преимуществ. Он утвержден как стандарт *http://www.adobe.com/enterprise/standards/*, имеет возможность совмещения в одном файле текста и изображения. Уже разработано большое количество программного обеспечения для работы с файлами в этом формате для практически всех операционных систем. Однако существует ряд задач, для решения которых возможностей этого формата недостаточно. Так, например, при создании электронной библиотеки газет, технических отчетов, научных журналов и других сложно структурированных документов, как показано выше, необходимо иметь образ как целой страницы, так и отдельных ее составляющих (статей, глав и т.п.). При этом часто необходимо показать не только изображение этого отдельного объекта, но и распознанный текст, соответствующий этому объекту. Часто задача еще более усложняется. Поскольку такие материалы очень часто напечатаны на плохой бумаге, автоматическое распознавание дает не очень хорошие результаты. Для их улучшения нужен сервис редактирования текста, позволяющий одновременно видеть и изображение, и соответствующий изображаемому полигону текст. Для этого нам нужен формат, в котором бы каждому распознанному слову был поставлен в соответствие полигон в изображении страницы. Если мы посмотрим на опыт российских коллег, то увидим, что такой подход был применен при создании Научной педагогической электронной библиотеки Российской академии образования [1]. При этом авторами был разработан собственный формат представления данных. Многочисленные зарубежные проекты, связанные с оцифровкой газет, часто используют в таких случаях формат METS/ALTO (*http://www.loc.gov/standards/alto/*). Этот формат был разработан Библиотекой конгресса США для своей программы создания электронной библиотеки газет (*http://www.loc.gov/ndnp/*) и оказался настолько удачным, что нашел применение в большом количестве похожих проектов [2].

Пригородокъ Табынскъ, служившій намъ зимовьемъ, построенъ на Нагайской дорогѣ въ Башкирской Кситабынской волости, на нагорномъ или на правомъ берегу рѣки Бѣлой, въ самомъ томъ мѣстѣ, гдѣ рѣчка Усолка соединяется съ рѣкою Бѣлою. Отъ города Уфы почитается до него съ девяносто верстъ, а отъ Оренбурга съ двѣсти пятьдесятъ. Первымъ началомъ къ заведенію сего пригородка служили соленые ключи, въ 12 верстахъ отъ нынѣшняго Табынскаго поселенія

Пригородокъ Табынскъ.

Рис. 2. *Выдержка из издания «Дневныя записки...»*

Несколько слов о самом формате. Эта XML-схема для анализа разметки и текстовых объектов (The Analyzed Layout and Text Object (ALTO)) была первоначально разработана группой проекта METS для использования в стандарте METS. В то время как METS предназначен для описания структуры объектов, схема ALTO была связана с содержимым и разметкой каждой части объекта. В настоящее время эта XML-схема детализирует технические метаданные, связанные с разметкой и содержимым физических текстовых ресурсов, таких как книги или газеты. Обычно она используется как расширение для METS в секции административных данных. Однако она может существовать и в виде описания отдельных документов независимо от METS. В 2010 году была утверждена версия 2.0 этой схемы. К сожалению, в настоящее время существует очень небольшое количество программного обеспечения, которое способно понимать этот формат и хорошо его интерпретировать. Лучше всего это делает программное обеспечение Veridian *http://www.dlconsulting.com/veridian/*, разработанное новозеландской компанией DL Consulting, специализировавшейся ранее на поддержке свободного программного обеспечения Greenstone.

Сравнив особенности электронной коллекции уникальных произведений великих русских ученых XVII-XIX веков и возможности, предоставляемые нам форматом METS/ALTO и хранением отсканированных страниц в графических форматах, а не в PDF, мы решили проверить на практике их реализацию. Первой задачей стало представление российского печатного издания XVII-XIX веков в формате METS/ALTO. Нужно сказать, что это нетривиальная задача. Выше я уже писал про небольшое количество программ, способных понимать METS/ALTO. Как оказалось, программ, способных их создавать, тоже мало. Нам удалось найти только одну. Немецкая компания CCS Content Conversion Specialists GmbH *http://www.content-conversion.com* разработала программное обеспечение docWorks, которое обеспечивает весь цикл сканирования, распознавания и представления данных в формате METS/ALTO. Этой разработкой весьма активно пользуются крупнейшие библиотеки всего мира, в основном, для оцифровки газет. Для более детального изучения возможностей этого программного обеспечения Российская государственная библиотека 22 августа 2012 года провела научно-практический семинар «Оцифровка документов и долгосрочное хранение: опыт работы председателя секции ИФЛА Ф. Зарндта (США)». Видеозапись семинара можно посмотреть на сайте Ассоциации электронных библиотек *http://www.aselibrary.ru/conference/conference43/conference433498/*. На семинаре присутствовал один из руководителей компании CCS Content Conversion Specialists GmbH, с которым мы договорились об эксперименте по обработке наших документов в docWorks. Одной из интересных функций docWorks является автоматическое разбиение страницы на блоки. Это в какой-то степени похоже на работу FineReader, только в отличие от него разбиение проводится на статьи, и в статьях автоматически выделяются автор и заглавие. Для проверки этой функции на русских текстах мы использовали газету «Санкт-Петербургские ведомости» за 1847 год. В этой газете довольно сложная структура колонок, и некоторые из статей переходят на следующую страницу. В научных изданиях структура обычно бывает проще.

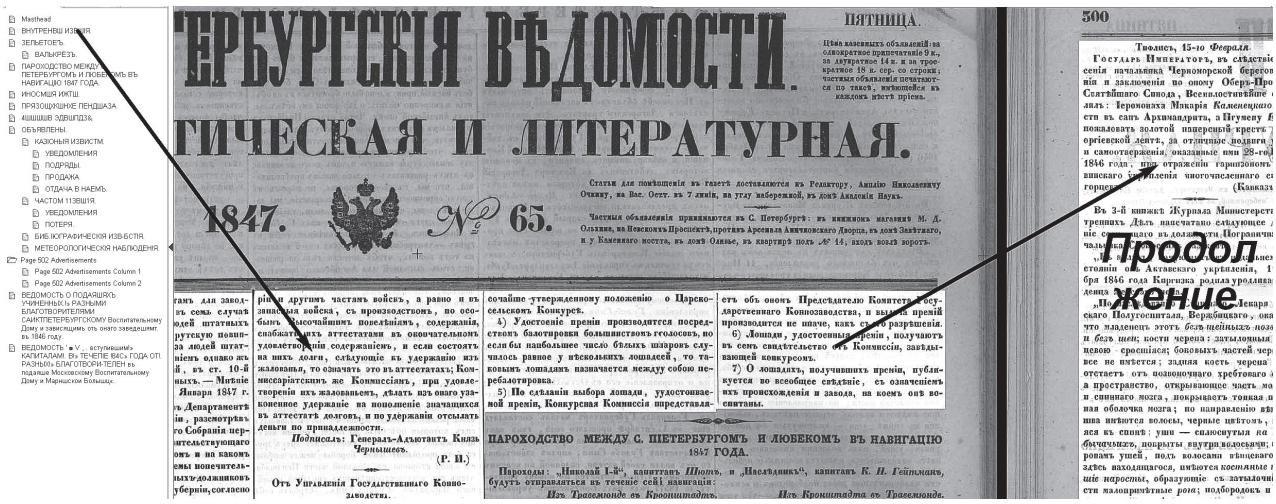

Рис. 3. *Результат работы программы docWorks*

Результаты обработки были представлены в программе Veridian. Как видно из **рис. 3** (помечены стрелками), программа успешно определила структуру газеты, выделила заголовки статей и правильно соотнесла колонку на второй странице с первой статьей про внутренние известия.

В результате обработки формируется файл в формате METS, в котором описывается структура всей газеты (*для краткости приведено с пропусками*).

<mets xsi:schemaLocation="**http://www.loc.gov/METS/ http://schema.ccs-gmbh.com/metae/mets-metae.xsd**" xmlns:xsi="**http://www.w3.org/2001/XMLSchema**

.....

<structMap>

.....

<div ID="**DIVL22**" TYPE="**TEXT**" ORDER="**1**" LABEL="**СЛнктпвтврБургг, 20 Марта. Высочайшим» Приказом» отъ 1С Марта,**

Для каждой страницы создается файл в формате ALTO, в котором каждый текстовый блок разбивается на строчки, а строчки в свою очередь на слова. Для каждого элемента обозначаются его координаты и размеры. Содержание элементов приводится в подполе CONTENT (*в примере выделено мной*).

<TextBlock ID="**P1_TB00010**" HPOS="**84**" VPOS="**843**" WIDTH="**552**" HEIGHT="**1904**" STYLEREFS="**TXT_0 PAR_BLOCK**">

- <TextLine ID="**P1_TL00029**" HPOS="**196**" VPOS="**843**" WIDTH="**326**" HEIGHT="**26**">

<String ID="**P1_ST00179**" HPOS="**196**" VPOS="**844**" WIDTH="**185**" HEIGHT="**23**" CONTENT="**СЛнктпвтврБургг,**" WC="**0.47**" CC="**4764627176672253**" />

<SP ID="**P1_SP00151**" HPOS="**381**" VPOS="**867**" WIDTH="**18**" />

<String ID="**P1_ST00180**" HPOS="**399**" VPOS="**845**" WIDTH="**25**" HEIGHT="**18**" CONTENT="**20**" WC="**0.17**" CC="**96**" />

<SP ID="**P1_SP00152**" HPOS="**424**" VPOS="**867**" WIDTH="**14**" />

<String ID="**P1_ST00181**" HPOS="**438**" VPOS="**843**" WIDTH="**84**" HEIGHT="**26**" CONTENT="**Марта.**" WC="**0.97**" CC="**862640**" />

</TextLine>

В итоге мы получаем детальное описание отсканированного документа, который может быть впоследствии обработан различными программами в зависимости от задач исследователя. Первой же задачей встает редактирование распознанного текста.

Мы видим, что качество распознавания текста (использовался стандартный FineReader), конечно, оставляет желать лучшего. Но мы и не рассчитывали на особенно хороший результат. Для нас главное - это получить основу для последующего редактирования. Для реализации таких желаний в программе Veridian предусмотрена возможность редактирования текста распознанного текста (**рис. 4**).

Каждая строка текста соотносится со своим полигоном, и редактору выводится одновременно два окна: распознанное изображение и оригинал. При помощи нехитрых операций текст редактируется и сохраняется в исправленном файле формата ALTO. Нужно отметить, что Veridian - уже не единственное средство редактирования плохо распознанных изображений. Задача эта востребо-

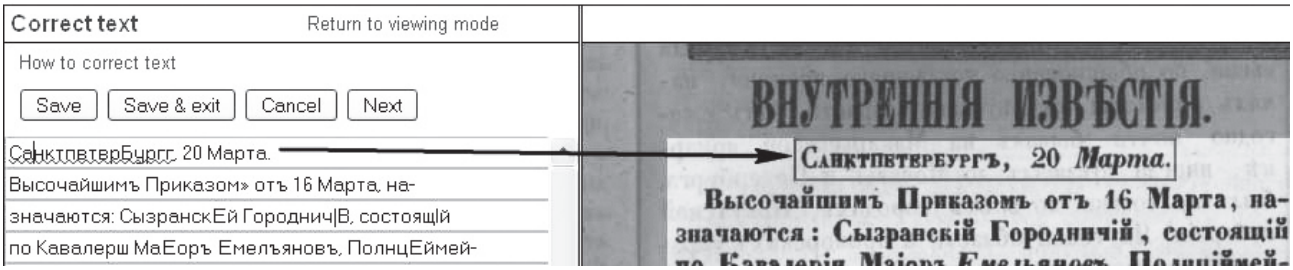

Рис. 4. *Редактирование распознанного текста в Veridian*

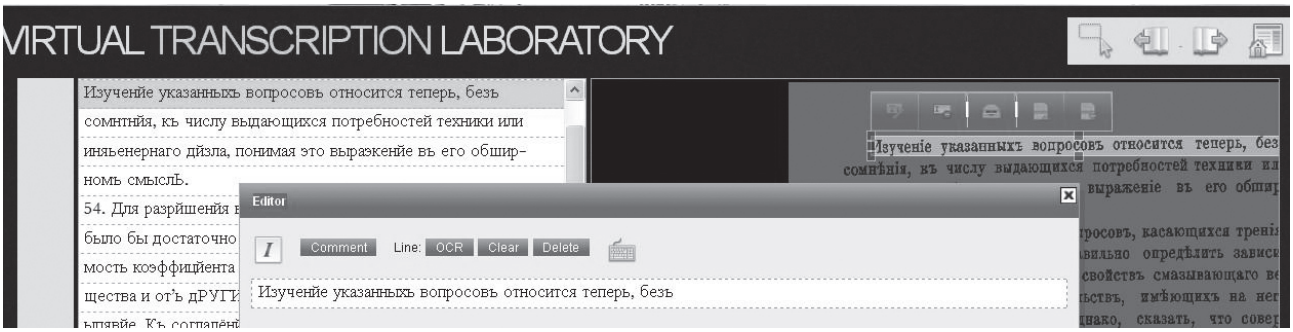

Рис. 5. *Интерфейс Virtual Transcription Laboratory*

вана во многих проектах, и уже имеется успешный опыт создания веб-сервисов для решения этой задачи методом краудсорсинга.

Одним из наиболее продвинутых в этом направлении работ можно считать проект Virtual Transcription Laboratory *http://wlt.synat.pcss.pl/wlt-web/index.xhtml*, созданный сотрудниками суперкомпьютерного центра в Познани (Польша). Цель этого проекта - предоставить площадку для загрузки туда оцифрованных изображений, автоматического распознавания их имеющимися на площадке средствами (*Tesseract*) и предоставить веб-сервис для редактирования результатов распознавания. Более подробно об этой работе можно прочитать в докладе Адама Дудчака (Adam Dudczak) [3]. Мы провели эксперимент и загрузили к ним на площадку те же «Санкт-Петербургские ведомости» (желающие могут сравнить результат работы *Tesseract* и FineReader) и одну из книг нашей коллекции *Петров, Николай Павлович (1836-1920).* ***Трение*** *в* ***машинах*** *и влияние на него смазывающей жидкости / Н. Петров, проф. Николаевской инженерной акад. и С[анкт]-Петербургского практического технологического ин-та. - Санкт-Петербург: Тип. А.С. Суворина, 1883. - [4], IV, 210, [2] с. : ил., табл.; 24 см.*

На **рис. 5** приведена одна из страниц книги в режиме редактирования распознанного текста. Сервис вполне работоспособный, доступ к рисункам открытый, и все желающие могут поэкспериментировать с этим сервисом. Однако проект находится еще в стадии разработки, и поэтому еще далеко не все планы реализованы. Например, в ближайших планах - умение экспортировать и импортировать файлы в формате METS/ALTO. Сейчас разработчики используют свой внутренний формат, довольно похожий на METS/ALTO.

Подводя итоги нашим экспериментам, можно сказать, что формат METS/ALTO вполне может быть использован для электронной коллекции уникальных произведений великих русских ученых XVII-XIX веков. Единственным препятствием является малое количество программного обеспечения, поддерживающего работу с этим форматом. Будем надеяться, что скоро положение изменится, и тогда мы сможем воспользоваться преимуществами, которые он предоставляет.

Литература:

1. Трифонов С.И. Комбинированное электронное представление печатных изданий // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды тринадцатой Всероссийской конференции, Воронеж, 2011. - С. 89-94.

2. Belaid, A., Falk, I.; Rangoni, Y. XML Data Representation in Document Image Analysis // Document Analysis and Recognition. ICDAR 2007. Ninth International Conference on 23-26 Sept. 2007. - V.1. - P. 78-82.

3. Dudczak, A., Kmieciak, M., Werla, M. Creation of Textual Versions of Historical Documents from Polish Digital Libraries // Theory and Practice of Digital Libraries. Lecture Notes in Computer Science. - Volume 7489. - 2012. - P. 89-94.

НАША ИНФОРМАЦИЯ

19 марта 2013 года в Центре «Digital October» журнал «ИКС» (ИнформКурьер-Связь) проводит конференцию для руководителей и сотрудников корпоративных департаментов ИТ и сервис-провайдеров.

Развитие облачных вычислений и мобильности остаются основными движущими силами, оказывающими сегодня влияние на трансформацию корпоративных ИТ-архитектур. Использование частных, публичных и гибридных облаков, кросс-платформенных мобильных приложений, управление гетерогенными ИТ и информационная безопасность являются одними из наиболее обсуждаемых сегодня тем в среде ИТ-профессионалов.

На второй ежегодной конференции Cloud & Mobility 2013 будет продолжено изучение опыта первопроходцев, поиск лучших практик для минимизации рисков наблюдаемых процессов трансформации.

Цели конференции:

- рассмотреть ключевые аспекты перехода к облачной инфраструктуре;
- профессионально обсудить вопросы применения облаков и доступа к ним;
- рассмотреть мобильные решения и устройства как инструмент бизнеса;
- изучить лучшие бизнес-кейсы и примеры отдачи от внедрения;
- обсудить перспективы развития облачных услуг в мире и в России;
- узнать от поставщиков и пользователей о последних отечественных и зарубежных примерах использования облаков и мобильных облачных решений.

Аудиторию конференции составляют ИТ-директора, ведущие отраслевые эксперты и аналитики, владельцы и руководители центров обработки данных, представители крупнейших сервис-провайдеров и поставщиков решений.

Основные направления конференции:

- Стратегии трансформации.
- Облака (задачи, которые ставят облака, и примеры их решения).
- Мобильность (влияние мобильных устройств и приложений на изменения в ИТ-ландшафте).

Участие в прошлогодней конференции приняли более 230 профессионалов, представлявших различные отрасли: энергетику и транспорт, финансовые организации, металлургию, ритейл. Около четверти участников представляли государственные структуры. Ожидаемое количество участников в этом году - 300 человек.

Подробная информация:

http://www.cloudmobility.ru; http://www.linkedin.com/groups/Cloud-Mobility-4673878.