GEOINFORMATICS RESEARCH PAPERS
Proceedings of Geophysical Center RAS

ИССЛЕДОВАНИЯ ПО ГЕОИНФОРМАТИКЕ
Труды Геофизического центра РАН

VOL. 4, BS4001, doi:10.2205/2016BS028, 2016

Технологии документирования научного контента. II. Электронная книга

В. А. Нечитайленко[1]

Получено 20 января 2016 г.; принято 25 января 2016 г.; опубликовано 11 февраля 2016 г.

[1] Статья посвящена проблеме построения конвертера для преобразования LaTeX документов в HTML5 и EPUB3. Суть предложенного и реализованного автором решения заключается в замене макроопределений используемого LaTeX класса новыми макроопределениями, позволяющими эмулировать XML-совместимую структуру документа. Набор таких макроопределений представляет конечное множество, что позволяет построить алгоритм конверсии LaTeX в XML/XHTML. Хотя предложенное решение и ограничивает возможности LaTeX, оно вполне укладываются в стандарты представлений, определенных издателем для конкретного журнала. *КЛЮЧЕВЫЕ СЛОВА:* электронная книга; электронные публикации; информационные технологии; документирование науки; семантические включения; метаописание; CrossRef; eLibrary.

Ссылка: Нечитайленко, В. А. (2016), Технологии документирования научного контента. II. Электронная книга, *Geoinf. Res. Papers, 4,* BS4001, doi:10.2205/2016BS028.

Введение

[2] Начиная с конца 80-х годов прошлого столетия язык разметки научных текстов LaTeX получил широкое распространение и развитие и стал стандартом *де-факто* в научной среде, в особенности в области точных и технических наук, наук о Земле и др. Выбор LaTeX в качестве основного первичного языка разметки был предопределен отсутствием реальных конкурентов, особенно в случае сложных математических и химических текстов.

[3] С развитием телекоммуникаций и появлением первых стандартов языков представления текстов в компьютерных сетях сразу же встал вопрос о необходимости конверсии LaTeX-текстов в HTML и позднее в XML. В общем случае корректный перевод произвольных LaTeX-текстов в XML невозможен, поскольку TeX/LaTeX несовместим с концепцией SGML.

[4] Именно поэтому все известные разработки конвертеров текстов в LaTeX к HTML текстам ограничены фиксированным списком обрабатываемых макроопределений LaTeX. Дополнение же указанных конвертеров механизмами расширения этих списков, путем добавления в программы соответствующих модулей, не устраняет проблему, хотя и весьма полезно для адаптации конвертера к решению конкретной задачи (т.е. конверсии документа с заданным DTD).

[5] Очень важным является понимание принципиального различия между концепцией XML как подмножества SGML, которая определяет объекты, составляющие документ, их структуру и свойства, не заботясь о том, как документ представлен на физическом уровне (носителе), и концепцией TeX, которая сводится по сути к описанию поведения пера на бумаге, не более того.

[6] Распространенное представление о том, что макропакеты, такие как LaTeX, AmsTeX и др. преодолевают это различие, – не более чем заблуждение. В работе известных TeXguru *Миттельбаха и Роули* [1997, p. 196] читаем:[2]

> *Поскольку типичное SGML DTD использует концепцию подобную концепции LaTeX, форматирование часто реализуется простой заменой элементов документа конструкциями LaTeX, вместо того, чтобы использовать непосредственно примитивы TeX'а. Это позволяет использовать сложные аналитические методы, встроенные в LaTeX, и избежать необходимости программировать в TeX.*

[1]Геофизический центр РАН, Москва, Россия

[2]Because a typical SGML Document Type Definition (DTD) uses concepts similar to those of LATEX, the formatting is often implemented by simply mapping document elements to LATEX constructs rather than directly to 'raw TEX'. This enables the sophisticated analytical techniques built into the LATEX software to be exploited; and it avoids the need to program in TEX.

[7] Упомянутое в этой цитате подобие концепций является не прямым, а лишь опосредованным, т.е. не существует 1-1 соответствия.

[8] Для полноты картины процитирую ответ еще одного TEXguru [*Sebastian Rahtz,* 2005; http://www.tug.org/pipermail/tex-live/2005-March/007846.html] на вопрос одного из пользователей, можно ли найти хороший инструмент для конверсии LATEX файлов в XML:[3]

> *Существует несколько инструментов, ни один из которых нельзя легко использовать или гарантировать удовлетворительные результаты. Конверсия LATEX к XML от природы чрезвычайно сложна, чтобы быть 100% правильной. По моему мнению это плохая идея* :-)

От плохой идеи к прагматичному решению

[9] Документ (научная статья) включает множество объектов, которые несут большую или меньшую семантическую нагрузку. Для лучшего восприятия эти объекты (заголовки разного уровня, фигуры, таблицы, списки, в том числе реферативные, текстовые выделения, гипертекстовые ссылки и пр.) каким-то образом акцентируются, правила акцентирования определяются стандартами издателя и культурной традицией и могут быть одинаковыми для разных объектов или разными для одного класса объектов.

[10] Следует ли из этого, что для генерации версий документов, например, в PDF/PS и в HTML/XML нужно иметь различные соответствующим образом сформированные исходные файлы?

[11] Да, если мы хотим использовать всю мощь издательской системы TEX/LATEX с ее многочисленными расширениями и возможностью включения авторских макроопределений непосредственно в исходный файл, а также включить в документ динамические и интерактивные объекты.

[12] Нет, если издатель и автор готовы использовать ограниченный перечень макроопределений, которые, с одной стороны, определяют правила интерпретации исходного текста для TEXengine и, с другой стороны, эмулируют XML совместимую структуру, обеспечивая таким образом возможность программной конверсии исходного файла к форматам HTML5/XHTML/EPUB3.

[13] Использование языков типа XML в принципе позволяет не только учитывать специфику различных предметных областей, но и повышать эффективность структурирования при автоматизированной обработке статей и иных документов для последующей индексации. Однако без полного онтологического описания предметной области использование XML для семантического структурирования так же ограниченно, как и использование LATEX, который, как уже отмечалось, являясь языком представления, содержит семантику лишь опосредованно.

[14] Издаваемые сегодня электронные журналы, равно как и бумажные, ориентируются, в первую очередь, на читателя, а не на машину. В то же время есть все основания полагать, что в обозримом будущем научные публикации будут ориентироваться, в первую очередь, на ввод в информационно-поисковые (интеллектуальные) системы с развитым межмашинным интерфейсом.

[15] А пока... лучшее решение – следовать рекомендации, прозвучавшей на Первой конференции экспертов ЮНЕСКО и МСН по электронным публикациям в науке (Париж, 1996): "Сделай, потом фиксируй!"[4]

[16] Именно такой подход был реализован автором при разработке программного пакета ELXpaper (ELectronic eXtended paper) для изданий, публикуемых ГЦ РАН (*Russian Journal of Earth Sciences*, *Вестник Отделения наук о Земле РАН* и *Geoinformatics Research Papers*).

Программный пакет ELXpaper

[17] Пакет ELXpaper включает три основных компонента: стилевой файл `elxpaper.sty`, Perl скрипт `elxpaper.pl` с конфигурационным файлом `elxpaper.cfg`, Perl скрипт `elxtomml.pl` с конфигурационным файлом `elxtomml.cfg`, а также набор основных предварительно созданных элементов (CSS файлы, логотипы и макеты обложек и титульных страниц публикуемых изданий и т.п.).

[18] **Стилевой файл elxpaper.sty** является расширением стандартного класса LATEX `article.cls` и поддерживает двухколонный журнальный формат, генерацию внутренних и внешних активных гиперссылок, генерацию сообщений об ошибках в исходном файле или предупреждений, генерацию результата трансляции в форматах DVI или PDF, а также генерацию XML-метаописаний публикуемых документов для регистрации в системе CrossRef (в соответствии с XML-схемой CrossRef 4.3.3) и загрузки в Научную электронную библиотеку eLIBRARY.RU (в соответствии с XML-схемой eLibrary CE 7.1.4.1284).

[19] Последняя задача решается непосредственно в процессе трансляции исходного текста благодаря тому обстоятельству, что TEX/LATEX – это не просто один из языков разметки текста, а, по существу, программная система, реализующая концепцию программирования машины Тьюринга. Конечно, возможности программирования в TEX/LATEX существенно уступают возможностям языков

[3] ... there are several, none of them easy to use or guaranteed to give satisfactory results. LaTeX to XML is, inherently, extraordinarily hard to 100% right. In my personal view, it's a Bad Idea :-)

[4] Do it, then fix it! *Из коллекции утверждений, собранных в работе* [*Arnoud de Kemp*, 1996].

программирования высокого уровня, но они оказываются достаточными для построения множества макроопределений, расширяющих стандартные классы LaTeX и эмулирующих семантическую структуру документа. Детали работы `elxpaper.sty` рассмотрены в работе [*Нечитайленко,* 2015]

[20] **Perl скрипт elxpaper.pl** транслирует исходный LaTeX файл в набор HTML5 файлов с генерацией основного файла с динамическим оглавлением и набора файлов, вызываемых из оглавления и из текста статьи, включающих фигуры, таблицы, и ряд вспомогательных файлов. Трансляция является многопроходной и включает следующие шаги:

- выделение фрагментов для последующего формирования из них иконок (для фигур и таблиц);
- нумерация строк исходного LaTeX кода и слияние их в параграфы;
- разделение параграфов на семантически значимые фрагменты (например, если имеется дисплейный элемент внутри параграфа);
- обработка преамбулы с фиксацией основных параметров;
- выделение множества защищенных фрагментов (включая `verbatim` блоки и элементы) для последующей подстановки шрифтов, замены диакритических символов символами UTF-8, подстановки внутристрочной математики и разметки шрифтов;
- извлечение из текста элементов оглавления, метаданных и реферативных списков;
- генерация XML метаданных для CrossRef и eLibrary;
- обработка внутренних и внешних гиперссылок;
- извлечение из исходного документа плавающих элементов (фигур и таблиц), подстраничных ссылок и табулированных фрагментов;
- конверсия основного документа к формату HTML5;
- конверсия дополнительных файлов (фигур, таблиц, оглавления, карты статьи и др.);
- генерация пакетного файла, содержащего набор команд для перевода набора HTML файлов в набор XHTML с последующим циклическим вызовом программы `elxtomml.pl` (см. далее);
- генерация структуры XHTML файлов, предусмотренных стандартом EPUB3, включая TOCNNN, `package.opf` и др.;
- генерация титульной страницы и обложки для электронной книги в формате EPUB3 на основе предустановленных шаблонов и специфической информации конкретного документа (статьи).

[21] **Конфигурационный файл elxpaper.cfg** содержит описания параметров всех обрабатываемых в `elxpaper.pl` команд, включая их маски (patterns), регистрационные данные журналов, публикуемых ГЦ РАН, команды вызова вспомогательных free-domain программ и др. Кроме стандартных команд LaTeX2ε, определенных в `article.cls` (некоторые из них переопределены, например, `title`, `abstract` и др.), включен также ряд команд, определенных в пакетах, расширяющих `article.cls`, в первую очередь в пакете `hyperref.sty`. Перечень используемых пакетов LaTeX включен в преамбулу `elxpaper.sty`.

[22] Скрипт `elxpaper.pl` работает в предположении, что (i) исходный файл статьи в LaTeX не порождает сообщений об ошибках при его трансляции драйверами `latex` или `pdflatex`, (ii) в текущей директории присутствует файл `jobname.aux`, (iii) в преамбуле и тексте статьи не содержатся авторские макроопрепделения, формируемые командами типа `\newcommand`, `\renewcommand`, `\def`, `\gdef`, и т.п.

[23] Все команды в тексте статьи, не имеющие семантической нагрузки и используемые только для форматирования текста, такие, например, как `\space`, `\skip`, `\linebreak`, и т.п. игнорируются, равно как команды, имеющие семантическую нагрузку, но для которых в скрипте не определен алгоритм их обработки. Команды управления шрифтами заменяются соответствующими определениями в CSS файле `elx-online.css`.

[24] **Perl скрипт elxtomml.pl** преобразует сгенерированные на предыдущем шаге файлы в HTML5 формате в файлы формата XHTML, что необходимо для включения последних в структуру EPUB3. При этом некоторые элементы конструкции исходного файла заменяются на более строгие конструкции XHTML. При публикации документов на русском языке осуществляется также перевод исходного текста в Юникод.

[25] Скрипт `elxtomml.pl` решает также задачу преобразования математических и химических формул (как дисплейных, так и встроенных непосредственно в параграфы текста) представленных в нотификациях LaTeX к формату MathML. Как известно, издательская система TeX/ LaTeX/AmsTeX обеспечивает своего рода стандарт *де-факто* для представления математики при документировании научного контента. С переходом от POP (printed-on-paper) публикаций к электронным публикациям возникла необходимость найти способ представления TeX математики в форму, которая могла быть отражена в окне веб-браузера. В первых реализациях подобных решений широко использовалась замена математических и химических формул изображениями. Такое решение хотя и могло быть реализовано достаточно простыми средствами, но имело ряд серьезных недостатков, основными из которых стали проблема масштабирования изображений и проблема интерпретации их поисковыми системами.

[26] С разработкой языка разметки математического текста *MathMl*, который стал частью официально рекомендованного Консорциумом W3C стандарта HTML5 и

в 2015 зарегистрирован в качестве стандарта Международной организации стандартов (ISO/IEC DIS 40314), задача свелась к конверсии математики, представленной в TEX/AmsTEX в формат MathMl. В настоящее время среди многочисленных решений этой задачи особое место занимает система *MathJax*, разработанная и поддерживаемая MathJax консорциумом – совместным предприятием Американского математического общества (AMS) и Общества индустриальной и прикладной математики (SIAM).

[27] MathJax не является традиционным конвертером файлов одного формата в файлы другого. Конверсия осуществляется "на лету" при загрузке в браузер HTML файла, имеющего в теле (`<body>`) математические фрагменты в TEX нотациях, а в преамбуле (`<head>`) команды вызова MathJax скриптов либо из сети серверов MathJax (https://www.mathjax.org/), либо из отдельной копии MathJax, устанавливаемой на пользовательском сервере. MathJax использует CSS с веб-шрифтами или SVG, вместо рисунков (bitmap или Flash), благодаря чему математика масштабируется одновременно с текстом при всех уровнях увеличения/уменьшения. При этом результат трансляции направляется непосредственно на экран.

[28] В пакете ELXpaper конверсия математических текстов для HTML5 и EPUB3 реализована двумя различными способами. Для онлайновой версии статьи в формате HTML5 достаточно сохранить математику в том виде, как она представлена в исходном LATEX файле, добавив в преамбулу (`<head>`) вызов MathJax скриптов.

[29] Заметим, что в реализованной нами версии пакета ELXpaper MathJax скрипты вызываются непосредственно с сайтов сети `cdn.mathjax.org`. Для этого рабочий компьютер должен быть подключен в Интернету. Возможна также установка MathJax на локальном сервере (см. детали). Второе решение в ряде случаев может оказаться более предпочтительным, но оно потребует регулярного обновления локальной версии MathJax, тогда как непосредственное соединение с MathJax CDN избавляет от этой необходимости.

[30] Для электронной книги в формате EPUB3 такое решение непригодно, так как электронная книга в соответствии с рекомендациями IDPF (International Digital Publishing Forum) должна обеспечивать корректную интерпретацию текстов при отсутствии доступа к сети (offline). Устройства чтения документов в формате EPUB3, равно как эмуляторы таких устройств на десктопах, в соответствии с рекомендациями IDPF должны обеспечивать корректную интерпретацию HTML5/XHTML, составной частью которых является MathML.

[31] Как известно, формат EPUB представляет собой сжатый архив всех файлов, составляющих публикацию и отвечающих стандарту XHTML, с добавлением файлов, определяющих структуру архива (`package.opf`)и его тип (`mimetype`). Т.е. при переводе HTML5 файлов в более строгий формат XHTML необходимо конвертировать математические TEX фрагменты (дисплейные и строчные) во фрагменты MathML.

[32] Анализ существующих конвертеров файлов одного формата в файлы другого показал, что по многим параметрам они не могут конкурировать с системой MathJax (ограничения на исходные файлы, отсутствие поддержки, коммерческие условия использования и т.п.) В то же время MathJax, хотя и создает MathML фрагмент для каждого TEX фрагмента с возможностью их просмотра во всплывающих окнах, но не позволяет вывести их в отдельный файл, что связано с ограничениями языка JavaScript.

[33] Нами найдено и реализовано решение, которое по выражению автора TEX Дональда Кнута может быть отнесено к категории "dirty tricks". Суть его в том, что из JavaScript мы не можем записать сформированный MathML файл на локальном компьютере, но можем послать его в качестве содержимого HTML формы на сервер, содержащий программу обработки этой формы с записью файла во внешнюю память. Наиболее просто это можно сделать установив такой специальный веб-сервер непосредственно на `localhost`, где установлен пакет ELXpaper. Тогда результат обработки на этом сервере будет очевидным образом доступен соответствующей программе пакета. Вопрос синхронизации работы этих программ решается вполне ординарно через систему "флагов". Если по каким-то причинам установка такого вспомогательного сервера на `localhost` невозможна или нежелательна, например при использовании разных компьютеров для подготовки статей к публикации, соответствующие скрипты и директорию для записи конвертированных файлов можно установить на любом доступном веб-сервере.

[34] **Конфигурационный файл elxtomml.cfg** определяет порядок связи `elxtomml.pl` со вспомогательными программами (Рис. 1) взаимодействия с системой MathJax, включая адреса директорий `doc` и `cgi-bin` вспомогательного сервера, тип используемого в процессе конверсии браузера и команды его открытия и закрытия, адреса служебных файлов формируемого EPUB3 архива.

[35] **Дополнительные компоненты пакета** представляют набор основных предварительно созданных элементов (рисунки логотипов и макеты обложек, CSS файлы, таблицы соответствия диакритических символов LATEX символам UNICODE, и др.) Перечень вспомогательных программ и порядок установки и использования пакета ELXpaper включены в Приложение.

Заключение

[36] Разработанная технология успешно используется в практике редакционной подготовки изданий Геофизического центра РАН [*Астапенкова и др.* 2015], в первую очередь Russian Journal of Earth Sciences. Каждая из опубликованныч в RJES статей, начиная с июня 2012 г., представлена в четырех форматах: (i) основная версия

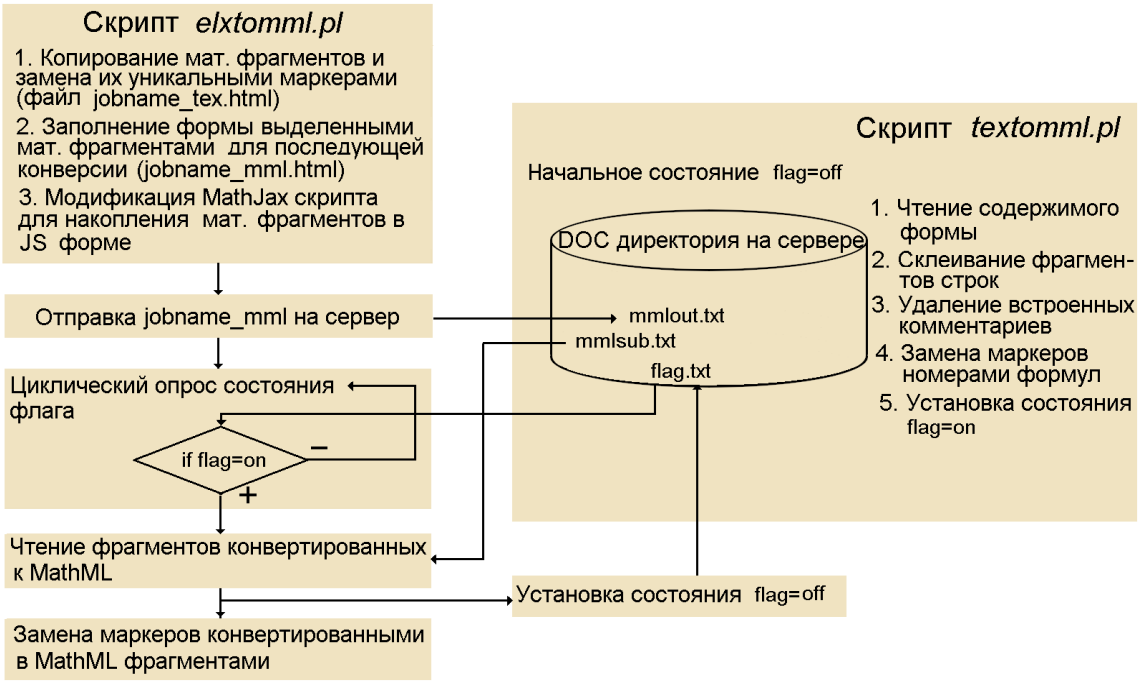

Рис. 1. Схема взаимодействия скриптов `elxtomml.pl` и `textomml.pl` в рабочей директории и в DOC директории вспомогательного сервера.

(version of record) в формате PDF; (ii) версия в формате HTML5; версия в формате EPUB3; а также (iv) версия в формате PDF, адаптированная для малых экранов (е-ридеров и смартфонов).

[37] Следует иметь в виду, что на момент написания этой статьи большая часть устройств и программ чтения документов в формате EPUB3 не обеспечивает корректного воспроизведения, включая, в том числе, и Readium, рекомендованный IDPF в качестве основы для включения в ПО для планшетов, смартфонов и электронных книг. Наши тесты показали, что лучшие результаты обеспечивают приложение iBooks2 на iPad и iPhone, Gyan ePub Reader и некоторые эмуляторы устройств чтения электронных книг, в частности, AZARDI и EPUBReader: Add-on for Firefox.

[38] Все сгенерированные версии в EPUB3 формате подвергаются проверке на соответствие стандарту с использованием программы epubcheck3.0 и рекомендованной IDPF программы EPUB Validator, после чего загружаются на сервер при отсутствии сообщений об ошибках и других типов предупреждений.

[39] **Благодарность.** Автор признателен своим коллегам А. А. Астапенковой и Э. О. Кедрову за ценные критические замечания и помощь в подготовке статьи к печати.

Приложение

[40] Описанный здесь пакет доступен в виде архива elxpaper_1.5.zip на сайте электронных публикаций ГЦ РАН. В архиве файл `ReadMeFirst_rus.pdf` содержит всю необходимую информацию по установке и использованию пакета. Архив включает также исходную версию этой статьи в LaTeX формате, из которой после обработки пакетом ELXpaper были получены HTML и EPUB3 версии данной статьи.

Литература

Астапенкова, А. А., Э. О. Кедров, В. А. Нечитайленко (2015), Документирование научного контента: современные концепции и решения, *Материалы 4-й Международной научно-практической конференции "Научное издание международного уровня – 2015: современные тенденции в мировой практике редактирования, издания и оценки научных публикаций"* p. 18–26, РАНХиГС, СПб. (http://conf.neicon.ru/materials/15-Domestic0515/Materials-0515.pdf)

de Kemp, Arnoud (1996), Options for the Future, *Proceedings of the Joint ICSU Press/UNESCO Expert Conference on Electronic Publishing in Science* p. 149–152, ICSU Press, Paris. (http://eos.wdcb.ru/eps1/dekemp.htm)

Mittelbach, Frank, Chris Rowley (1997), The LaTeX3 Project, *TUGboat, 18,* No. 3–Proceedings of the 1997 Annual Meeting, 195–198. (`https://www.tug.org/TUGboat/tb18-3/l3project.pdf`)

Nechitailenko, V. A. (2015), Record of science technologies. I. Online journal, *Geoinf. Res. Papers, 3,* BS1001, doi:10.2205/2015BS016

В. А. Нечитайленко, Геофизический центр РАН, Москва, Россия (vitaly@wdcb.ru)