Компьютерные инструменты в образовании, 2018
№ 6: 39–46
УДК: 004.91
http://ipo.spb.ru/journal
doi:10.32603/2071-2340-2018-6-39-46

АВТОМАТИЧЕСКАЯ ВЁРСТКА И ОФОРМЛЕНИЕ НАУЧНОЙ И ПРОГРАММНОЙ ДОКУМЕНТАЦИИ

Павлов Д. А.

Институт прикладной астрономии РАН, Санкт-Петербург, Россия

Аннотация

В статье описано новое средство автоматической вёрстки и оформления научной и программной документации в соответствии с требованиями ГОСТ. Средство называется GOSTdown (ГОСТ + Markdown) и рассчитано на документы крупного размера, создаваемые совместно несколькими участниками. Как правило, такие документы разрабатываются в Microsoft Word. GOSTdown — набор шаблонов и скриптов, позволяющий пользователям работать над содержательной частью документа в формате Markdown, получая на выходе автоматически собранный документ в формате DOCX, не требующий ручной доработки. GOSTdown основан на универсальном конвертере документов Pandoc и скриптовом языке Powershell.

Ключевые слова: *документация, генерация документов, Pandoc, Markdown.*

Цитирование: Павлов Д. А. Автоматическая вёрстка и оформление научной и программной документации // Компьютерные инструменты в образовании, 2018. № 6. С. 39–46. doi:10.32603/2071-2340-2018-6-39-46

1. ВВЕДЕНИЕ

Идея использования языков разметки для получения форматированного текста возникла в конце 1960-х годов; первой широко используемой реализацией стала программа roff, включённая в первые версии UNIX [1], и её дальнейшие воплощения nroff, troff, groff. В последующие годы появились языки разметки TeX [2] (1984) и HTML (1993), известные по сей день. Менее известны, но используются в узких нишах языки разметки Texinfo (1986), Wiki (1994) и другие.

Бурное развитие WYSIWYG-программ для создания документов в 1980-х годах частично снизило популярность языков разметки, но не сделало их ненужными — по причинам, о которых будет рассказано в следующем разделе. В 2004 г. блогер Джон Грубер и интернет-активист Аарон Шварц создали Markdown — язык разметки, максимально лёгкий для чтения. Новый язык быстро завоевал сердца публики, в настоящее время имеется множество реализаций и расширений языка.

Одним из таких расширений является *Pandoc Markdown* [3], используемый программой *Pandoc*. Эту программу создал в 2006 г. профессор философии Джон МакФарлан. Pandoc — универсальный конвертор документов между десятками форматов, включая многие языки разметки, а также нетекстовые форматы, в частности DOCX.

В силу исторических и организационных причин российские государственные заказчики научной и программной документации предпочитают DOCX в качестве формата электронных версий документов. Pandoc — одно из немногих (а с учётом возможностей Pandoc Markdown — и вовсе уникальное) средство генерации DOCX из языка разметки. Иными словами, Pandoc позволяет исполнителям использовать язык разметки, при этом не создавая проблем с требованиями заказчиков к итоговому продукту.

2. МОТИВАЦИЯ

Перечислим рутинные задачи, стоящие перед авторами документации:

— нумерация глав и разделов, генерация оглавления,
— вставка и нумерация формул, таблиц и рисунков,
— нумерация и форматирование многоуровневых списков,
— создание и поддержка ссылок на главы, разделы, формулы, таблицы и рисунки,
— ведение перечня использованных источников, создание списка литературы и ссылок на него в тексте,
— подсчёт и вставка в текст количества страниц, таблиц, рисунков, приложений,
— вставка текста из других документов и оформление его под стиль,
— слияние изменений, сделанных разными людьми в своих копиях документа,
— соблюдение требований к оформлению по ГОСТ 7.32 (Отчёт о научно-исследовательской работе) или ГОСТ 19 (Единая система программной документации, ЕСПД).

Разумеется, при длительной работе над крупным документом требуется максимальная автоматизация этих операций. Не должно быть ручной работы по нумерации ссылок, настроек стиля форматирования для каждого абзаца в отдельности и пр.

Формально говоря, в Word есть средства для решения всех этих задач. А при интеграции Word с сервером для совместной работы SharePoint достигается автоматизация параллельной работы нескольких пользователей над документом.

В реальной жизни людей, которые успешно освоили все средства Word (не говоря уже о SharePoint), крайне мало. Одна из причин неудобства использования Word для автоматизации работы с текстом — пресловутый принцип WYSIWYG. Содержание документа в Word не отделено от оформления. Из-за этого, решая основную задачу редактирования содержания, пользователь в любой момент может совершить неосторожный шаг и нарушить оформление. Попытка исправления, как правило, делает всё только хуже, потому что нарушает централизованную структуру стилей в документе. Наиболее известные примеры: постоянные поломки форматирования списков (например, при вставке текста); невозможность отслеживать изменения в документе, потому что от «Отформатировано: русский» у пользователя рябит в глазах; борьба с ручными настройками форматирования абзацев «поверх» стилей.

Заметим, что пользователь при создании отчётной документации в принципе не должен работать над оформлением. Оформление стандартизировано ГОСТом и реализовано единожды в корпоративном образце. Таким образом, Word предоставляет пользователю возможности, которые ему в данном случае не нужны, при этом затрудняя осуществление действительно необходимых операций.

При использовании языка разметки вместо Word проблема оформления исчезает, а точнее, переносится на плечи программы, компилирующей DOCX из разметки. Пользователь работает только с содержимым, но не со стилями (не считая минимального набора средств выделения текста жирным и курсивным начертаниями, верхними и нижними индексами).

Представление документа в виде текстового файла даёт дополнительные преимущества:

— возможность использовать привычный текстовый редактор,
— возможность редактировать документ на любой операционной системе или в онлайне,
— возможность использовать средство совместной работы с текстовыми файлами и слияния изменений (Git или аналогичное),
— невозможность «поломки» и гарантированная воспроизводимость документа.

3. PANDOC

Диалект Markdown, поддерживаемый Pandoc, включает:

— заголовки разных уровней,
— внутритекстовые и выключные формулы в формате LaTeX, при конверсии в DOCX преобразующиеся в формат Office Math Markup Language (OMML [4]), с которым работает встроенный редактор формул Word 2007 и более поздних версий,
— таблицы,
— растровые рисунки в форматах JPEG и PNG, векторные в форматах EPS и EMF,
— нумерованные и маркированные списки, в том числе вложенные,
— фрагменты программного кода (форматируются моноширинным шрифтом),
— жирный/курсив/зачёркнутый текст, нижние/верхние индексы.

Пример, иллюстрирующий часть возможностей Pandoc, приведён на рис. 1 и 2.

Схема работы Pandoc такова, что любые преобразования идут через внутреннее промежуточное представление документа в самом Pandoc. Модель внутреннего представления не содержит всех возможностей всех исходных форматов (и, наоборот, содержит возможности, которые присутствуют не во всех исходных форматах). Только Pandoc Markdown в полной мере соответствует внутреннему представлению документа.

Pandoc обладает расширениями, так называемыми фильтрами, работающими непосредственно с документом во внутреннем представлении. Вместе в Pandoc поставляется фильтр pandoc-citeproc (для обработки библиографических ссылок). Фильтр pandoc-crossref [5] для нумерации и ссылок на таблицы, рисунки, разделы и формулы) разрабатывается независимо Николаем Якимовым (МАИ).

```
## Таблица с ячейками из многих строк

-------------------------------------------------------------------------------
  Номер  Описание параметра
 -------  ------------------------------------------------------------------
  188-190  Углы $\varepsilon_x$, $\varepsilon_y$ и $\varepsilon_z$ ориентации эфемерид в ICRS

  201-700  Гравитационные параметры ($Gm$) планет и астероидов

  701-706  Элементы орбиты Меркурия: $\ln(a)$,
           $\sin i\cdot\cos(\Omega)$, $\sin i\cdot\sin(\Omega)$,
           $e\cdot\cos(\varpi)$, $e\cdot\sin(\varpi)$, $l$,
           где $a$ — большая полуось, $i$ — наклон орбиты,
           $\Omega$ — долгота восходящего узла, $e$ — эксцентриситет,
           $\varpi$ — долгота перицентра, $l$ — средняя долгота

      707  $\dot{a}/a$ Меркурия
-------------------------------------------------------------------------------

Table: Таблица с многострочными (но одноабзацевыми) ячейками. {#tbl:mytable-multiline}
```

Рис. 1. Пример исходного текста в Markdown

7.3 Таблица с ячейками из многих строк

Таблица 7.5 – Таблица с многострочными (но одноабзацевыми) ячейками.

Номер	Описание параметра
188-190	Углы ε_x, ε_y и ε_z ориентации эфемерид в ICRS
201-700	Гравитационные параметры (Gm) планет и астероидов
701-706	Элементы орбиты Меркурия: $\ln(a)$, $\sin i \cdot \cos(\Omega)$, $\sin i \cdot \sin(\Omega)$, $e \cdot \cos(\varpi)$, $e \cdot \sin(\varpi)$, l, где a — большая полуось, i — наклон орбиты, Ω — долгота восходящего узла, e — эксцентриситет, ϖ — долгота перицентра, l — средняя долгота
707	$\dot{a}\,/a$ Меркурия

Рис. 2. Скомпилированный результат

4. БИБЛИОГРАФИЯ

Существует два ГОСТа для библиографии: 7.1 (Библиографическая запись. Библиографическое описание) и 7.0.5 (Библиографическая ссылка). При создании списка литературы в программной и научно-технической документации формально требуется применять ГОСТ 7.1. Но реально он применяется редко из-за неудобных требований:

— вставлять пометку «[Текст]» в ссылки на любые тексты (то есть практически во все ссылки);
— повторять имя первого автора дважды (включая случай, если автор всего один);
— писать полный список авторов после названия (до трёх включительно), а если их четыре и более, то упоминать только первого «и др».

ГОСТ 7.0.5 обходится без этих требований, поэтому его более охотно используют вместо 7.1 — ту его часть, которая называется «затекстовые ссылки», про которые, впрочем, сказано, что «Совокупность затекстовых библиографических ссылок не является библиографическим списком».

Стиль библиографии в Pandoc настраивается файлом формата CSL 1.0.1 [5]. Сама библиография может быть задана в различных форматах: Bibtex, EndNote, RIS и др. В GOSTdown используется файл библиографии в формате Bibtex, что, как и запись формул в формате LaTeX, привычно многим сотрудникам научных организаций, которые занимаются написанием статей.

В дистрибутив pandoc-citeproc входит стилевой файл [6], основанный на ГОСТ 7.0.5 с отдельными элементами ГОСТ 7.1. В целом, правила ГОСТов (как 7.1, так и 7.0.5) настолько сложны, что автоматизировать библиографию по всем правилам невозможно. Формат CSL 1.0.1, задуманный как универсальный окончательный формат для формирования всей библиографии по любым правилам, не подходит в полной мере для этих ГОСТов.

В частности, в ГОСТе есть разумное требование: английские ссылки должны быть с английскими вспомогательными словами (Vol., No., pp., ed. и прочие), а русские — с русскими. В CSL 1.0.1 такой опции не предусмотрено. Опция появилась в расширении CSL-M, которое не поддерживается в pandoc-citeproc. Это препятствие удалось преодолеть в GOSTdown ценой установки вспомогательного поля «note» для русскоязычных статьей в библиографии и дополнительных изменений в CSL-файле. Для перевода

«et al.» в «и др.» сделанное решение не работает по причине недостаточной гибкости процедур обращения с данными, поддерживаемых CSL. Поэтому в литературных источниках GOSTdown перечисляются все авторы.

При необходимости строжайше следовать ГОСТу в случаях, когда CSL не способен помочь, у пользователей есть возможность использовать запись вида @misc, которая позволяет задать произвольный текст библиографической ссылки.

5. ШАБЛОН

Pandoc работает с основным текстом документа. Такие вещи, как титульный лист и колонтитулы, легко создаются пользователями в Word в режиме WYSIWYG, тогда как их создание в Pandoc затруднено или невозможно. Для разрешения этого противоречия в GOSTdown существует файл-шаблон в формате docx, который содержит следующее:

— Формат колонтитулов (Pandoc «подхватывает» этот формат при генерации документа),
— Перечень стилей заголовков и абзацев, также для Pandoc,
— Небольшая часть содержимого, которая не генерируется из Markdown: лист утверждения (при наличии), титульная страница, заключительный «лист регистрации изменений».

6. ПОСТОБРАБОТКА

Комбинирование шаблона и основного текста — задача, которую Pandoc выполнить не может. Этим занимается скрипт постобработки, вставляя основной текст вместо специального маркера %MAINTEXT% в шаблоне. И не только: файл, сгенерированный Pandoc, хотя и содержит необходимые стили, всё же не является готовым продуктом. Для получения готового продукта необходимо следующее:

1. Установка стилей нумерации и отступы для списков и вложенных списков (Pandoc в настоящий момент не обладает механизмом шаблонизации списков).
2. Исправление выравнивания номеров формул (нумерованные формулы pandoc-crossref генерирует как таблицы из двух колонок).
3. Исправление горизонтального выравнивания ячеек таблиц, в которых есть только формула и больше ничего (Word в таком случае берёт настройки выравнивания из формулы, а не из ячейки, а Pandoc не заботится о том, чтобы выравнивание в формуле совпадало с выравниванием в ячейке.)
4. Установка стилей таблиц (Pandoc не умеет этого делать).
5. Установка стилей ненумерованных заголовков (они отличаются от стилей нумерованных, а Pandoc присваивает всем заголовкам одного уровня один стиль).
6. Установка ненумерованных стилей заголовков.
7. Вставка произвольных DOCX-файлов, заданных по желанию пользователя маркерами %INCLUDE(file.docx)%.
8. Вставка оглавления средствами Word вместо специального маркера %TOC%.
9. Подсчёт количества страниц (%NPAGES%), рисунков (%NFIGURES%), таблиц (%NTABLES%), приложений (%NAPPENDICES%), литературных источников (%NREFERENCES%), и вставка этих показателей вместо соответствующих маркеров.
10. Сохранение полученного документа не только в DOCX, но и в PDF.

Всё это делается с помощью механизма COM (Component Object Model). Фактически, любая инсталляция Word на Windows открывает возможности для программного управления содержимым документа так же, как если бы пользователь делал это вручную в программе. Широко известно применение этой технологии в макросах (Visual Basic for Applications, VBA), но макросы не очень удобны тем, что сами хранятся в документе. Менее известен следующий факт: с COM-объектами без труда можно работать в скрипте Powershell. Это и реализовано в GOSTdown.

7. ДОПОЛНИТЕЛЬНЫЕ ВОЗМОЖНОСТИ

Минимальные средства форматирования, позволенные в Markdown, иногда недостаточны. Наиболее известное ограничение связано с таблицами. В Pandoc (в отличие от Word и LaTeX) невозможно создать таблицу с объединёнными ячейками. Невозможно задать различное выравнивание в строках таблицы (только в каждом столбце индивидуально целиком). Также нет средств для создания подчёркнутого или цветного текста. Наконец, Pandoc не позволяет вставить два и более изображений с единым номером и подписью.

Всё это преодолимо различными обходными путями.

1. Сложные таблицы или рисунки можно редактировать непосредственно в Word в отдельных файлах, после чего вставлять эти файлы в основной текст с помощью %INCLUDE(file.docx)%.
2. Составные рисунки, содержащие изображения из двух и более файлов, реализованы в уже упоминавшемся (и используемом в GOSTdown) фильтре pandoc-crossref под названием «подграфики» (Subfigures, см. [7]).
3. При необходимости набрать текст каким-то стилем, для которого не хватает средств Markdown, можно в шаблоне создать пользовательский стиль (например, MyCustomStyle) и использовать его в документе. Pandoc позволяет применять пользовательские стили к словам и абзацам с помощью декларации вида {custom-style="MyCustomStyle"}.

8. РЕДАКТИРОВАНИЕ ИСХОДНОГО КОДА

Большая часть текстовых редакторов обеспечивают минимальную поддержку формата Markdown. Наиболее современные, такие как Visual Studio Code с плагином «Markdown Preview Enhanced (MPE)», позволяют редактировать текст одновременно с просмотром результата (рис. 3).

9. РЕАЛИЗАЦИЯ

Реализация GOSTdown вместе с руководством пользователя и примерами, охватывающими большую часть сценариев применения, свободно доступна в репозитории программного кода Института прикладной астрономии РАН [8]. Для работы GOSTdown необходимы Powershell и Word с поддержкой COM, что исключает возможность компиляции документов на ОС, отличных от Windows. Однако редактирование исходных кодов возможно на любой ОС, включая мобильные. При наличии сервера контроля версий Git и системы постоянной интеграции (continuous integration, CI) значительно облегчается совместная работа и появляется возможность удалённой компиляции документа на единой для всех пользователей инсталляции GOSTdown.

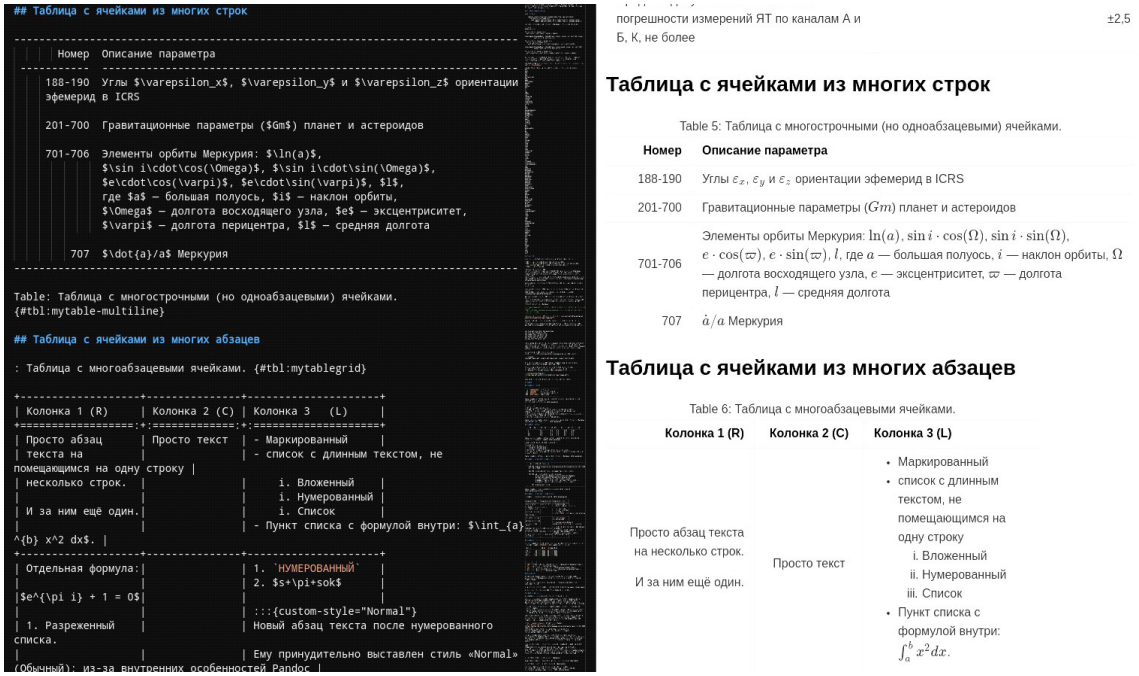

Рис. 3. Редактирование документа в Visual Studio Code с просмотром результата в правой панели

10. ЗАКЛЮЧЕНИЕ

Формат Markdown, заслуженно пользующийся популярностью в документации современного ПО, в веб-программировании и в электронных изданиях, теперь может использоваться для создания полноценных документов, соответствующих требованиям российских государственных организаций. Текстовая основа формата позволит использовать богатые возможности систем контроля версий. Полностью автоматизированная процедура оформления избавит пользователей от многих забот.

Pandoc и его фильтры продолжают активно развиваться, что даёт надежду на расширение возможностей GOSTdown по оформлению таблиц и библиографии.

Благодарность. Создание GOSTdown не было бы возможно без участия А. Г. Водолагиной (ИПА РАН), проделавшей большую работу по настройке стилей и постобработке таблиц и списков для соответствия ГОСТ.

Список литературы

1. *Thompson K., Ritchie D. M.* Unix Programmer's Manual. Bell Labs, 1971.
2. *Knuth D. E.* The TEXbook. Reading. Mass.: Addison Wesley, 1984. 483 pp.
3. *MacFarlane J.* Pandoc User's Guide (2018). URL: http://pandoc.org/MANUAL.pdf (Retrieved 16.12.2018).
4. ECMA. ECMA-376: Office Open XML File Formats. ECMA (European Association for Standardizing Information and Communication Systems), Geneva, Switzerland (2006). URL: http://www.ecma-international.org/publications/standards/Ecma-376.htm (Retrieved 16.12.2018).
5. *Rintze M.* Zelle, CSL 1.0.1 Specification. URL: http://docs.citationstyles.org/en/1.0.1/specification.html (Retrieved 16.12.2018).
6. URL: https://github.com/citation-style-language/styles/blob/master/gost-r-7-0-5-2008-numeric.csl (Retrieved 16.12.2018).

7. pandoc-crossref. URL: http://lierdakil.github.io/pandoc-crossref/ (Retrieved 16.12.2018).
8. URL: https://gitlab.iaaras.ru/iaaras/gostdown (Retrieved 16.12.2018).

Поступила в редакцию 30.11.2018, окончательный вариант — 16.12.2018.

Computer tools in education, 2018
№ 6: 39–46
http://ipo.spb.ru/journal
doi:10.32603/2071-2340-2018-6-39-46

AUTOMATIC LAYOUT AND PUBLISHING OF SCIENTIFIC AND SOFTWARE DOCUMENTATION

Pavlov D. A.

Institute of Applied Astronomy RAS, Saint Petersburg, Russia

Abstract

The article presents a new tool for automatic layout and publishing of scientific and software documentation in accordance with the requirements of GOST, the Russian national standards. The tool is called GOSTdown (GOST + Markdown) and is designed for large documents created jointly by several participants. Usually such documents are developed in Microsoft Word. GOSTdown is a set of templates and scripts that allow users to work on the content of a document in Markdown format, resulting in an automatically assembled document in DOCX format that does not require manual modification. GOSTdown is based on a universal document converter Pandoc and the Powershell scripting language.

Keywords: *documentation, document processing, Pandoc, Markdown.*

Citation: D. A. Pavlov, “Automatic Layout and Publishing of Scientific and Software Documentation”, *Computer tools in education*, no. 6, pp. 39–46, 2018 (in Russian). doi:10.32603/2071-2340-2018-6-39-46

Received 30.11.2018, the final version — 16.12.2018.

Dmitry A. Pavlov, Senior researcher at the Laboratory of Ehemeris Astronomy, Institute of Applied Astronomy RAS; 191187 Russia, Saint Petersburg, Kutuzova Embankment 10, IAA RAS, dpavlov@iaaras.ru

**Павлов Дмитрий Алексеевич,
кандидат физико-математических наук,
старший научный сотрудник Лаборатории эфемеридной астрономии Института прикладной астрономии РАН;
191187 Санкт-Петербург, наб. Кутузова 10, ИПА РАН,
dpavlov@iaaras.ru**