Format follows function: Agile editing of scientific manuscripts with markdown

Robert Winkler

Publishing is an essential part of academic life. The material for books, seminar lectures and journal manuscripts has to be prepared in specific formats. Scientific manuscripts consist of contents, including text, figures, formulas, tables and citations, which are presented in a certain format. This format depends on the intended use, e.g. for for submission to a particular journal, publication as a printed or electronic book, or for a webpage. Incompatible file formats, markdown with different target formats. This article demonstrates the feasability to edit the contents for various academic pubication formats in a common format. Markdown files containing the content with some simple formatting rules in plain text, which facilitates the writing. The final document can be exported to high-quality publications in different formats using Pandoc.

**Correspondence**: robert.winkler@cinvestav.mx, CINVESTAV Unidad Irapuato, Department of Biochemistry and Biotechnology, Laboratory of Biochemical and Instrumental Analysis (labABI, <http://www.ira.cinvestav.mx/lababi.aspx>), Km. 9.6 Libramiento Norte Carr. Irapuato-León 36821 Irapuato, Gto. Mexico, Tel.: +52 (462) 623 96 35, Fax +52 (462) 624 58 46

**Keywords**: markdown, latex, publishing, typesetting

# Introduction

Even if a submitted manuscript might be accepted by a journal ‘as is’, it still needs to be adjusted to the particular publication style in the production stage. Generally speaking, a scientific manuscript is composed from contents and formatting. Whilst the content, i.e. text, figures, tables, citations etc., may remain the same between different publishing forms and journal styles, the formatting can be rather different. Current publishing formats PDF HTML EPUB. Typesetting software, word processors such as Microsoft Word, LibreOffice, WPS Office, What You See Is What You Get (WYSIWYG), LaTeX What You See Is What You Want (WYSIWYW), hybrids such as LyX What You See Is What You Mean (WYSIWYM). Because of the sometimes complicated syntax specifications, simple conversions between file formats can be difficult or impossible. In academic publishing, the following types of works require the creation of different output formats from the same source text:

* For the publishing of a book, with a print version in PDF and an electronic version in EPUB.
* For distributing of a seminar script, with an online version in HTML and a print version in PDF.
* For submitting a journal manuscript for peer-review in DOCX, as well as a pre-print version with another journal style in PDF.

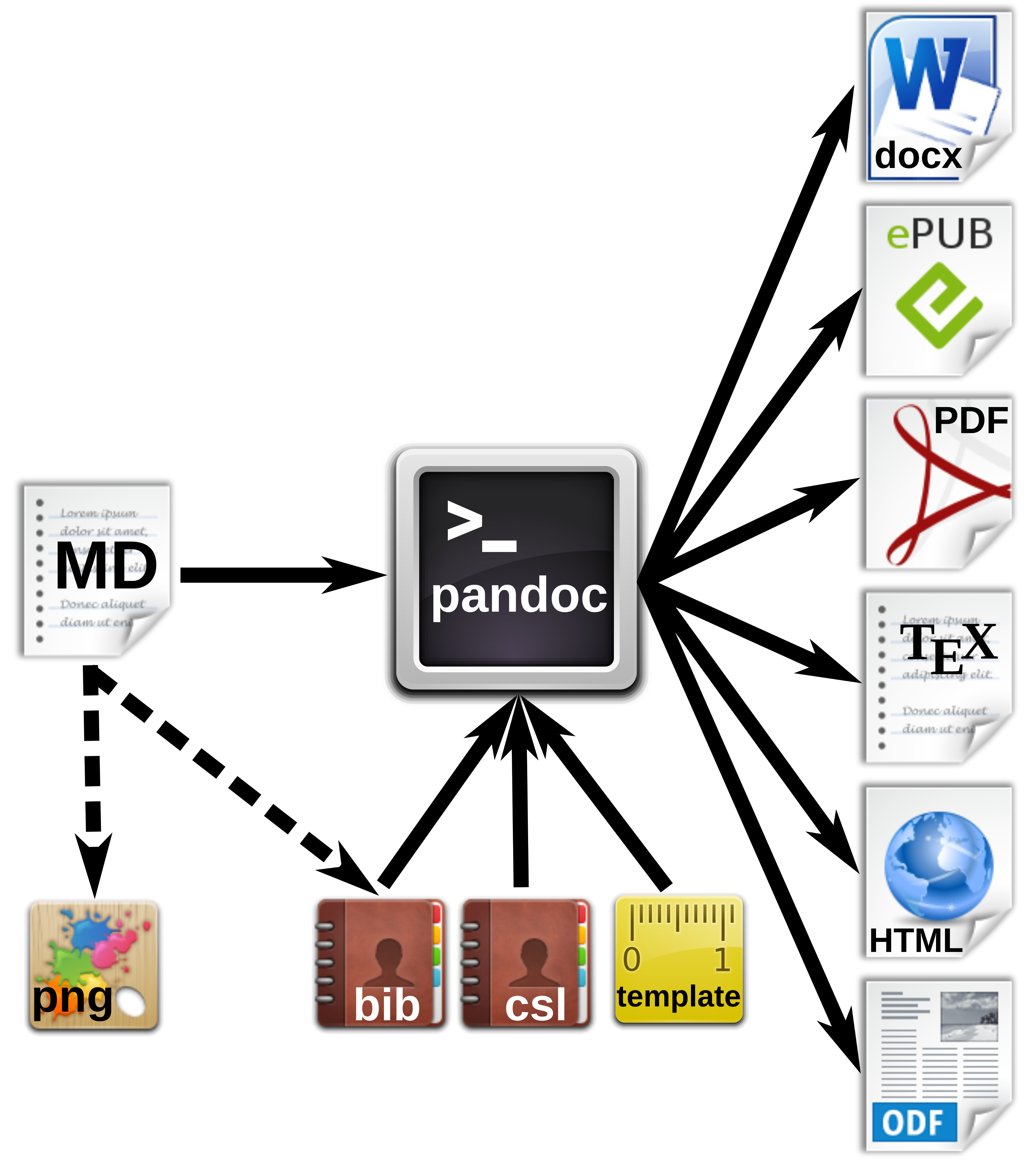
## Comparison of different markup languages

**Table 1.** Formatting elements and their implementation in different document types.

|  |  |  |  |
| --- | --- | --- | --- |
| Element | Markdown | LaTeX | HTML |
| **structure** |  |  |  |
| section | # Intro | \section{Intro} | <h1><Intro></h1> |
| subsection | ## History | \subsection{History} | <h2><History></h2> |
| **text formatting** |  |  |  |
| bold | \*\*text\*\* | \textbf{text} | text |
| **cross references** |  |  |  |
| to section |  |  | text |
| http link |  |  | text |

Documents with the commonly used Office Open XML (DOCX Microsoft Word files) and OpenDocument (ODT LibreOffice) file formats can be opened in a standard text editor after unzipping. However, content and formatting information is distributed into various folders and files.  
Overall, markdown displays the simplest structure, which facilitates the editing of documents.

# Concepts of markdown and Pandoc

 **Figure xx.** Workfow for the generation of multiple document formats with Pandoc.

# Markdown editors and online editing

## Editing programs

Because of the simple syntax, basically any text editor is suitable for editing markdown files. For several popular text editors, such as vim (<http://www.vim.org/>), GNU Emacs (<https://www.gnu.org/software/emacs/>), atom (<https://atom.io/>) or geany (<http://www.geany.org/>), plugins provide additional functionality for markdown editing, such as syntax highlighting, live preview or structure browsing.  
On the other side, in the last years plenty of special mardown editors have been published. Many of those are cross-platform compatible, e.g. Abricotine (<http://abricotine.brrd.fr/>), Ghostview (<https://github.com/wereturtle/ghostwriter>) and CuteMarkEd (<https://cloose.github.io/CuteMarkEd/>).  
xx Writing on the go mobile devices, Even for tablets, Android and iOS devices, numerous free and low-cost applications exist. Parts of this text were written in xx JotterPad dictation and swipe softwarexx Various of those applications support the cloud storage of documents.

**Figure xx.** Coding, preview and table of contents view using the CuteMarkEd editor.

## Online editing and collaborative writing

xx Google Docs test editing.

In recent years, several platforms were developed for collaborative writing. Google Docs. OwnCloud with Markdown Editor plugin (see section xx).

**Figure 2.** Direct online editing of this manuscript with live preview using the ownCloud Markdown Editor plugin by Robin Appelman.

## Document versioning and change control

Integrated in editing software or cloud server, low overhead of the files diff, git.

# Pandoc markdown for scientific texts

## Tables

Pipe tables are less strict in their syntax

| Left | Center | Right | Default |  
|:-----|:------:|------:|---------|  
| LLL | CCC | RRR | DDD |

gives

|  |  |  |  |
| --- | --- | --- | --- |
| Left | Center | Right | Default |
| LLL | CCC | RRR | DDD |

## Figures

## Formula

Formula can be inserted in LaTeX mode using delimiters ($ for Pandoc, $$ adds compatibility for online preview rendering with CuteMarkEd/MathJax). E.g. the formula for calculating the standard deviation of a random sampling would be written as:

$s=\sqrt{\frac{1}{N-1}\sum\_{i=1}^N(x\_i-\overline{x})^{2}}$

and gives:

with the individual observations, the sample mean and the total number of samples.

## Code listings

# Citations and biography

## Reference databases

bibtex databases, which are supported by almost any reference management program.

## Database of cited references

A bibtex data base for the cited references only using BibTool:

1. Extract refrence keys. This can be done by a simple Perl (<https://www.perl.org/>) command: ~ perl -ne ‘print “$1,” if /(?<=@)(.+?)(?=[],])/’ agile-editing-pandoc.md ~ The command prints out the keys of the file agile-editing-pandoc.md, separated by comas. The domains of email adresses also will be returned, but this does not affect the creation of the final database.
2. Write an bibtex .aux file (e.g. bibextract.aux) containing the name of the database (here: zotero.bib) and the extracted keys, separated by comas (from the previous step):
3. Process the complete database to extract the cited references into a now database (called bibshort.bib): ~ bibtool -x bibextract.aux -o bibshort.bib ~

## Styles

Whereas natbib bibtex is supported by Pandoc, it is incompatible with DOCX xx. The Citation Style Language (CSL) <http://citationstyles.org/> is used for the citations and bibliographies. This file format is supported e.g. by the reference management programs Mendeley <https://www.mendeley.com/>, Papers <http://papersapp.com/> and Zotero <https://www.zotero.org/>.  
CSL styles for particular journals can be found from the Zotero style repository <https://www.zotero.org/styles>.  
The bibliography style, which Pandoc should use for the target document can be chosen or in the YAML block of the markdown document or can be passed as an command line option. The later is more recommendable, because different bibliography style may be used for different documents.

## Creation of natbib citations in LaTeX

For citations in scientific manuscripts written in LaTex, the natbib package is widely used. To create TEX output file with natbib citations, Pandoc simply has to be run with the --natbib option.

# Document code and software availability

The software used was cited according to (Smith, Katz, & Niemeyer, 2016; Winkler, 2010, p23). Since unique identifiers are missing for most software projects, we only refer to the project homepages or software repositories:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Software | Use | Authors | Version | Release date | Homepage/ repository |
| Pandoc | universal markup converter | John MacFarlane | 1.16.0.2 | 2016/01/13 | <http://www.pandoc.org> |
| pandoc-citeproc | library for CSL citations with Pandoc | John MacFarlane, Andrea Rossato | 0.9.1 | 2016/03/19 | <https://github.com/jgm/pandoc-citeproc> |
| ownCloud | personal cloud software | ownCloud GmbH, Community | 9.1.1 | 2016/09/20 | <https://owncloud.org/> |
| Markdown Editor | plugin for ownCloud | Robin Appelman | 0.1 | 2016/03/08 | <https://github.com/icewind1991/files_markdown> |
| BibTool | Bibtex database tool | Gerd Neugebauer | 2.63 | 2016/01/16 | <https://github.com/ge-ne/bibtool> |

xx CuteMarkEd xx xx JotterPad Prof xx Pandoc is available for Windows, Mac OS X, Linux, BSD and as source code.

The source code of this manuscript, as well as templates and the Pandoc script have been deposited to xx.

# Definition of output formatting

command line parameters and templates xx

# Example: Manuscript with output of DOCX format and TEX/PDF for PeerJ pre-print submission

## Development of DOCX template

A first DOCX document with bibliography in APA format is created with Pandoc DOCX output:

pandoc -S -s --csl=apa.csl --filter pandoc-citeproc  
-o pandoc-manuscript.docx agile-editing-pandoc.md

The document settings and styles of the resulting file pandoc-manuscript.docx can be modified, and following it can be used as document template (--reference-docx=pandoc-manuscript.docx).

pandoc -S -s --reference-docx=pandoc-manuscript.docx  
--csl=apa.csl --filter pandoc-citeproc -o outfile.docx agile-editing-pandoc.md

It is possible to directly re-use a previous output file as template (i.e. template and output file have the same file name):

pandoc -S -s --columns=10 --reference-docx=outfile.docx --csl=apa.csl --filter pandoc-citeproc -o outfile.docx agile-editing-pandoc.md

In this way, the template can be incrementally adjusted to the desired document formatting. The final document may be employed later as Pandoc template for other manuscripts with the same specifications. In this case, running Pandoc the first time with the template, the contents of the new manuscript would be filled into the provided DOCX template. A page with DOCX manuscript formatting of this article is shown in figure xx.

**Figure xx.** DOCX output with a modified document template.

## Development of a TEX/PDF template

pandoc -D latex > template-peerj.latex

# Conclusions

Writing scientific manuscripts in markdown format helps to focus on the content rather than on the format. Lightweight format facilitates file editing and handling. With the same source file, multiple output files for different uses or publisher’s specifications c can be generated with Pandoc. Therefore, a workflow based on markdown format is certainly an attractive option for scientific writers. Therefore, scientific publishers should consider to support the submission of documents in markdown format.

# Acknowledgments

We cordially thank Dr. Gerd Neugebauer for his help in creating a subset of a bibtex data base using BibTool. The work was funded by the Consejo Nacional de Ciencia y Tecnología (CONACyT) Mexico, with the grant FRONTERAS 2015-2/814 and by institutional funding of the Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV).

# Bibliography

Smith, A. M., Katz, D. S., & Niemeyer, K. E. (2016). Software citation principles. *PeerJ Computer Science*, *2*, e86. <http://doi.org/10.7717/peerj-cs.86>

Winkler, R. (2010). ESIprot: A universal tool for charge state determination and molecular weight calculation of proteins from electrospray ionization mass spectrometry data. *Rapid Communications in Mass Spectrometry*, *24*(3), 285–294.