Formatting Open Science

Albert Krewinkel, Robert Winkler

Publishing is an essential part of academic life. Currently, a silent revolution is going on, because the concepts or pre-print and ‘Open Access’ publishing with rapid electronic distribution are challenging traditional print media. To facilitate the handling of manuscripts during the editorial process, the publishers usually specify the format of submissions to some degree. Thus, academic writers have to spend a notable amount of time to shape their material into the required formats and specifications. These document specifications depend on the intended publishing format of the manuscript (EPUB, HTML, PDF, ..), the publisher and the journal. Technically, scientific manuscripts consist of contents (text, figures, formulas, tables, code, citations, etc.) and formatting rules. This article demonstrates the feasibility to edit the contents for various academic publication formats in a common format. Markdown files contain the content with some basic formatting rules in plain text, which facilitates the writing. The final document can be exported to high-quality publications in different formats such as LATEX, EPUB, PDF, DOCX or HTML using Pandoc. Therefore, we recommend a workflow based on mardown files to focus on content rather than format. Since the clean syntax facilitates the processing of of the documents into different formats, also publishers could profit from accepting markdown manuscripts.

**Correspondence**: Prof. Dr. Robert Winkler, [robert.winkler@cinvestav.mx](mailto: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

Although 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.

The cost related to the editing and production of articles is the justification scientifc publishers to either charge high subscription fees or the author for making their articles freely available (‘Open Access Option’). (Van Noorden, 2013). Ironically, even after paying for making the work free to the plublig, the author may only have limited rights for distribution and re-use of his own work.

Open Journal Systems of the Public Knowledge Project (<https://pkp.sfu.ca/ojs/>)

## Cost of document production and impact on distribution

Benkler predicts a new form of economy, which is based on the distributed creation of goods with not exclusively commercial motives (Benkler, 2006). In science literature production, an interesting trend to open access articles and non-commercial publishing models can be observed xx. PLoS journals, but associated with high cost for the authors. PeerJ has a different business model, which stimulates the early reporting of results with a preprint server, as well as an accessible pricing scheme for the authors.  
Examples as the *Journal of Statistical Software* (https://www.jstatsoft.org/) demonstrate the possiblity of completely community supported open access publications.

## Current standard publishing formats

Ignoring the **Guide for Authors**, e.g. by submitting a manuscript with a different reference style, gives a negative impression with a journal’s editorial staff. Too carelessly prepared manuscripts can even provoke a straight ‘desk-reject’ (Volmer & Stokes, 2016). Scientific journal accept either DOC(X), LATEX and/ or PDF submissions.

a 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. 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.

**Table xx.** Current standard formats for publishing

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type | Description | | Use | Syntax | Reference |
| DOCX | Office Open XML | | WYSWYG editing | XML, ZIP |  |
| ODT | |  | | | |
| PDF | |  | | | |
| EPUB | |  | | | |
| LATEX | |  | | | |
| HTML | |  | | | |
| MD | |  | | | |

**Table xx.** Examples for formatting elements and their implementations in different markup languages types.

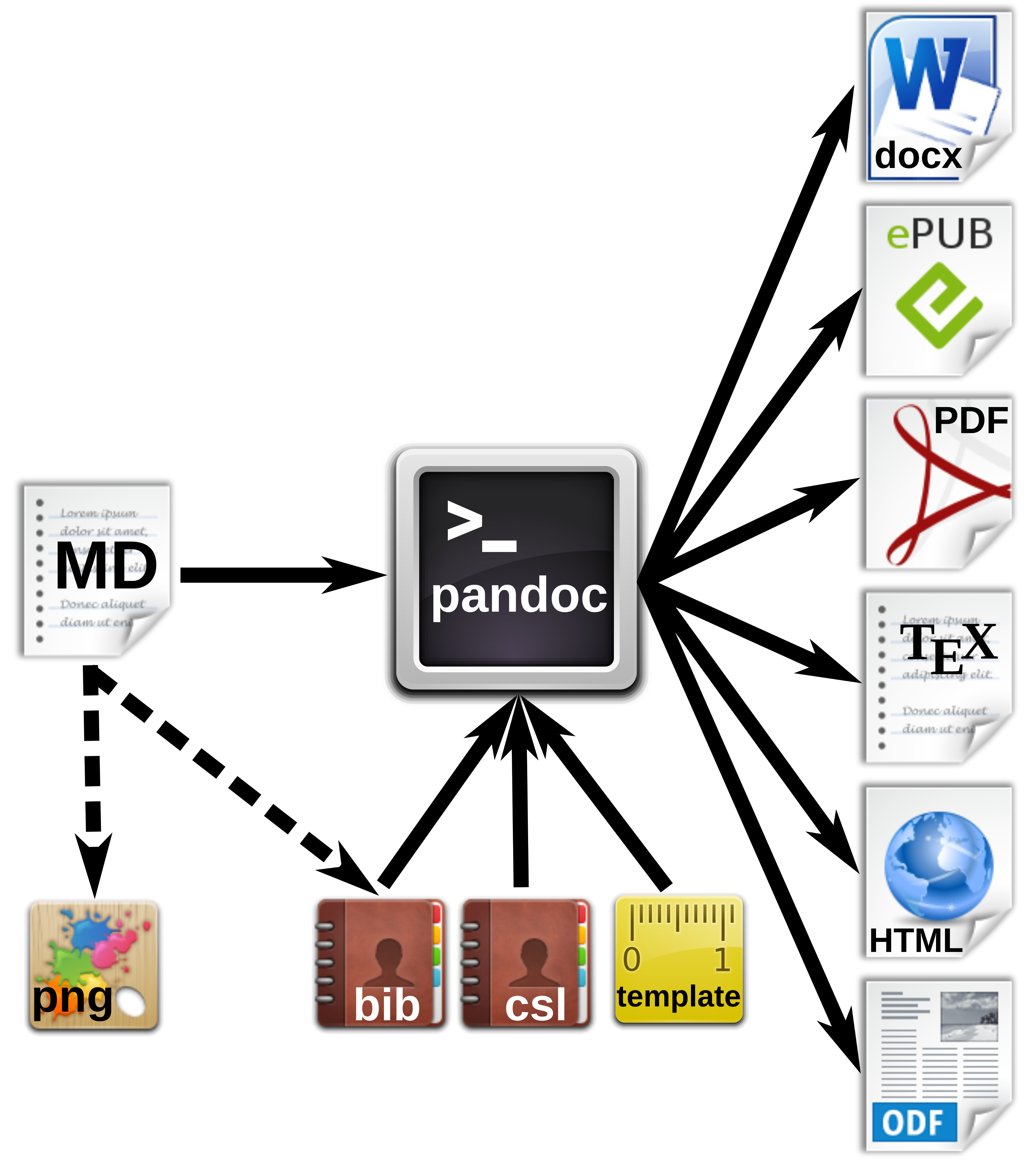
|  |  |  |  |
| --- | --- | --- | --- |
| Element | Markdown | LaTeX | HTML |
| **structure** |  |  |  |
| section | # Intro | \section{Intro} | <h1><Intro></h1> |
| subsection | ## History | \subsection{History} | <h2><History></h2> |
| **text formatting** |  |  |  |
| bold | \*\*text\*\* | \textbf{text} | text |
| italics | \*text\* | textit{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.

Several programs for the conversion between documents formats exist, such as the e-book library program calibre <https://code.google.com/archive/p/faenza-icon-theme/>. But the results of such conversions are often not satisfactory and require substancial manual corrections. Therefore, we were looking for a solution, which enables the creation of scientific manuscripts in a simple format, and the subsequent generation of multiple output formats.

# Concepts of markdown and Pandoc

Markdown was originally developed by John Gruber to simplify the writing of HTML documents <http://daringfireball.net/projects/markdown/>.

 **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 xx.** 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

Following the potential of typesetting scientific manuscripts with Pandoc is demonstrated with examples for typical document elements, such as formulas, figures, tables, code listings and references. The complete Pandoc User’s Manual can be found at <http://pandoc.org/MANUAL.html>.

## 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

## Formulas

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

Verbatim code blocks are indicated by three tilde symbols: ~ verbatim code ~ Typeseting inline code is possible by enclosing text between back ticks ``.

# Citations and biography

## Reference databases

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

## 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.

## Database of cited references

To share the bibliography for a certain manuscript with co-authors or the publisher’s production team, it is often desirable to generate a subset of a larger database, which only contains cited references.  
If the --natbib option was used for creating a PDF file (xx LATEX-PDF?), LATEX creates an AUX file, which can be extracted using BibTool: ~ bibtool -x md-article.aux -o bibshort.bib ~ In this case the new database would be called bibshort.bib. If no AUX file is present, it has to be created manually: 1. The reference keys have to be extracted from the manuscript. This can be done by a simple Perl (<https://www.perl.org/>) command: ~ perl -ne ‘print “$1,” if /(?<=@)(.+?)(?=[],])/’ article.md ~ The command prints out the keys of the file article.md, separated by comas. Domains of email adresses also will be returned, but this does not affect the creation of the final database. 2. A bibtex .aux file (e.g. bibextract.aux) has to be created, containing the name of the database (here: zotero.bib) and the extracted keys, separated by comas (from the previous step):  3. Now the new database can be generated with BibTool as mentioned above.

We wrote the Perl script mdbibexport with a simple user dialog, which asks asking for the relevant files and performs the generation of an article specific BIB database.

# Definition of output formatting

command line parameters and templates xx

# Example: Manuscript with output of DOCX format and TEX/PDF for submission to different journals.

DOCX the most common, but also LATEX only journals, PeerJ pre-print server

## 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 also 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).

# Software and code availability

The relevant software for creating this manuscript used is cited according to (Smith, Katz, & Niemeyer, 2016). 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](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> |
| mdbibexport | bibliography of cited references | Robert Winkler | 0.10 | 2016/10/31 | https://github.com/robert-winkler/mdbibexport |

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.

Icons for document types and applications have been adopted from Calibre <http://calibre-ebook.com/> and the GNOME Theme Faenza <https://code.google.com/archive/p/faenza-icon-theme/>.

# Bibliography

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press.

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>

Van Noorden, R. (2013). Open access: The true cost of science publishing. *Nature*, *495*(7442), 426–429. <http://doi.org/10.1038/495426a>

Volmer, D. A., & Stokes, C. S. (2016). How to prepare a manuscript fit-for-purpose for submission and avoid getting a “desk-reject”. *Rapid Commun. Mass Spectrom.*, n/a–n/a. <http://doi.org/10.1002/rcm.7746>