Wikipedia to LaTeX, PDF, EPUB and ODT

Dirk Hünniger

Abstract

The MediaWiki2Latex program converts wiki content to LaTeX and other formats. It has been more than 10 years since we last mentioned it in *TUGboat*,[1] so there is quite a bit of news to report.

1 Introduction

A wiki is a great way of working on a document with a distributed group of authors. MediaWiki is the Wiki system used by Wikipedia, owned by the Wikimedia Foundation.

Wikipedia used to provide ways to download the contents in various formats. Their idea was to finance the service by selling printed copies of articles as books; this did not prove cost effective, due to a lack of demand. In turn the download functions were not maintained, and eventually removed. Many efforts were undertaken to re-enable exporting, but all such development stopped years ago.

Our open source approach, as a hobby project with no financing, recently celebrated its tenth anniversary, and is still under active development. It is deployed on a server kindly provided to us by the Wikimedia Foundation.

2 User experience and web service

We dropped the development of a specific binary package for Windows. Instead we offer a docker file that can be run on any operating system. In addition, we still update and support the package for the Debian Linux distribution, which is included in many other Linux distributions too. We furthermore provide an online conversion service at `mediawiki2latex.wmflabs.org`.

This service takes the url to a wiki article and outputs a resulting file for download. We also added additional output file formats. You can choose between a PDF file compiled with LaTeX, its respective LaTeX source code as a zip file, as well as the word processing format ODT and the ebook format EPUB.

Some Wiki pages extensively use HTML tricks to create browser-viewable graphics, such as election diagrams or maps with marked positions. For those cases we optionally offer rendering the tables via the Chromium engine. It is also possible to process collections of wiki articles to a single output file.

3 Command line interface

We offer a command line interface that can be installed locally as a docker container. This provides the same features as the web service described above.

There is a feature in MediaWiki called "templates" which is similar to `\newcommand` in LaTeX. In the command line interface you can specify parsing of the wiki source code, instead of the HTML generated by MediaWiki. Here you can provide a mapping of templates to LaTeX commands which allows you to customize the output in various ways. Also you can let MediaWiki expand the templates to Wiki syntax and use it as input for MediaWiki2LaTeX. When wiki syntax is processed, MediaWiki2LaTeX will resolve references inside the document to sections and page numbers.

4 Technical details

MediaWiki2LaTeX is written entirely in the purely functional programming language Haskell. The image processing is done by ImageMagick in C++. Recently we added http2 multiplexing and compression using curl for the download of images and their respective contributor information, which resulted in a speedup of a factor of two to five, with the highest speedups on articles containing many small images. We also attempted to implement http multiplexing with multi-threading, but this did not work due to servers denying multiple connections.

Furthermore, many smaller bugs have been fixed; they are tracked in a detailed change log in the source package, so we will not discuss them here. The runtime and memory usage of various parts of the program were improved, also tracked in the change log. The documentation has also been improved.

Due to the remaining problem of no free font covering the whole Unicode range, we implemented an algorithm to switch between various fonts during the XƎLaTeX run as needed. This allowed MediaWiki2LaTeX to become an official part of the Debian Linux distribution, which was not possible with the computationally combined font which we used before. A man page and Makefile were added to support easy packaging of the software. Support for http has been dropped in favor of https. Finally, a progress bar was added in the web interface as well as a graphical user interface for Debian.

◇ Dirk Hünniger
Emil-Schweitzer Straße S 10
D-47506 Neukirchen-Vluyn, Germany
`dirk.hunniger (at) googlemail dot com`
`https://de.wikibooks.org/wiki/Benutzer:Dirk_Huenniger`

[1] *TUGboat* 34:2, "Converting Wikipedia articles to LaTeX", `tug.org/TUGboat/tb34-2/tb107huenniger.pdf`

doi.org/10.47397/tb/45-1/tb139huenniger-mediawiki2latex