

# MediaWiki to L<sup>A</sup>T<sub>E</sub>X

Dirk Hünninger  
Physicist

# Content

- Goals
- Users Perspective
- Alternative Approaches
- Technical Details



# Goals

- Professional PDF documents using the LaTeX typesetting system
- Byproducts EPUB, ODT, LaTeX files

# Easy Web Interface

Enter URL  
+  
Click Start!

No Cookies!  
No Java Script!

<https://mediawiki2latex.wmcloud.org/>

## Create Your PDF

To compile MediaWiki pages via LaTeX to PDF choose any URL from [Wikibooks](#) or any other website running MediaWiki. If you intent to compile a wikibook make sure you use the link to the printable version of the book.

### Send your request

URL to the Wiki to be converted	<input type="text"/>
Output Format	<input type="text" value="Compiled PDF"/>
Template expansion	<input type="text" value="Print"/>
Paper	<input type="text" value="A4"/>
Vector graphics	<input type="text" value="Rasterize"/>
<input type="button" value="Start!"/>	

### Please note:

The LaTeX source code will be compiled several times to make sure all references are resolved. The whole process will usually take about one minute but can take up to an hour depending on the extend of your request.



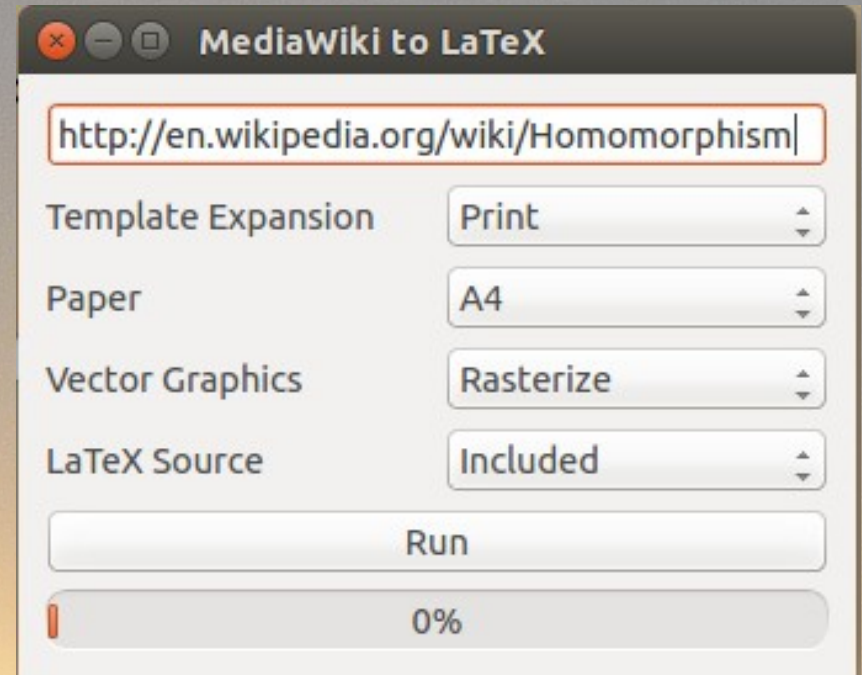
# locally installable GUI

Enter URL

+

Click Run

Technologies:  
Python 3 and Qt 5



# Debian Package

also for Ubuntu, Mint etc.

Docker Container for all Operating systems

GUI and command line interface



# Input modes

- Process Wiki-Text and
  - expand templates in a user defined way
  - expand templates by MediaWiki
- Process HTML generated by MediaWiki
- Collections (multiple Wiki pages as list of links)

# Alternative Approaches

Wikipedia own PDF function

Pedia Press

Pandoc

and many others



# Pedia Press

- Limited to only printed books
- No files for download
- Not an open source project

# Wikipedia PDF function

- No professional typography
- No ODT, EPUB, LaTeX source files
- No page numbering / no internal references
- No list of contributors / no list of figures



# Pandoc

- Very similar technology
- Open Source
- Requires significant modifications to work well with MediaWiki

# Basic Algorithm

- Read the Input to a tree structure.
- Run some calculations on it.
- Write the tree to LaTeX code.



Wikiheading	'=='	See also
-------------	------	----------

# Parse Tree

Itemgroup	'*'	
-----------	-----	--

Item	'*'	
------	-----	--

WikiLink	"	Quasimorphism
----------	---	---------------

Item	'*'	
------	-----	--

WikiLink	"	Diffeomorphism
----------	---	----------------

Wikiheading	'=='	Notes
-------------	------	-------

Wiki-Sourcecode:

== See also ==

\* [[Quasimorphism]]

\* [[Diffeomorphism]]

== Notes ==

# Creation of the Tree

Using Parsec library (just like Pandoc)

concept: „generic Bracket“ (only here)



# Generic Bracket

```
boldp = baseParser {  
    Start = string "<b>"  
    End = string "</b>"  
}
```

# Bracketed the wrong way

Brackets closed the wrong way like:

`<i><b>Hello</i></b>`

are often found in the wiki code on Wikipedia

=> The Wiki language is not context free !

(Proof by Pumping Lemma)



# Corollary: There is no BNF for the Wiki

A Backus Naur Form the wiki does not exist.  
Especially there is not regular expression for it.

=> all projects trying to parse MediaWiki using  
regex or BNF cannot work.

# User Defined Template Processing

*wiki source text:*

```
{{BoxWithFrame|content=boxed Hello}}
```

*user supplied configuration file:*

```
["BoxWithFrame","fbox","content"]
```

*LaTeX Output:*

```
\fbox{boxed Hello}
```



# Thank You for Listening!

Background Image Credits:

Jean-Louis VENET

CC-BY-SA 3.0

[https://commons.wikimedia.org/wiki/  
File:La\\_plage\\_de\\_Saint-Georges-de-  
Didonne.JPG](https://commons.wikimedia.org/wiki/File:La_plage_de_Saint-Georges-de-Didonne.JPG)

(derived work by me)

# Discussion

Questions?

<https://mediawiki2latex.wmcloud.org/>