MediaWiki to LaTeX

Dipl. Phys. Dirk Hünniger

Inhalt

- Ziele
- Benutzersicht
- Alternativen
- Technische Details

Ziele

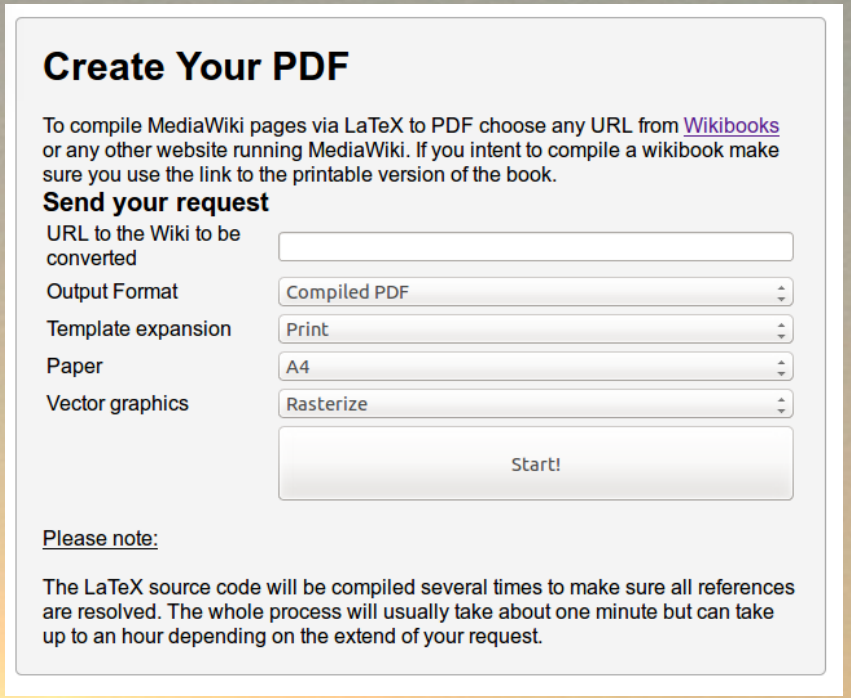
- Hochwertige PDF Dokumente
  mit dem LaTeX Schriftsatzsystem

- Nebenprodukt EPUB, ODT, LaTeX Dateien

Web Interface

URL Eingeben
+
Start! Klicken

Keine Cookies!
Kein Java Script!

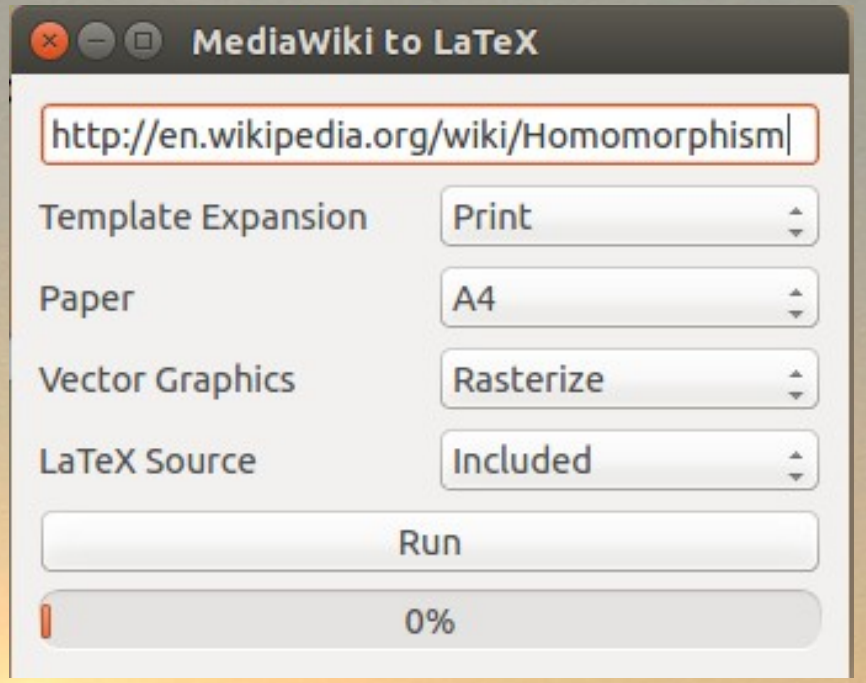


https://mediawiki2latex.wmcloud.org/

Lokal installiertes GUI

URL Eingeben
+
Run Klicken!

Technologien:
Python 3 und Qt 5



Debian Paket

auch für Ubuntu, Mint etc.

Docker Container für alle Betriebssysteme

GUI und Kommandozeileninterface

Eingabemodi

- Wiki-Text
  - eigene Vorlagenexpansion (auch benutzerdefiniert)
  - Vorlagenexpansion durch MediaWiki
- HTML durch MediaWiki erzeugt
- Sammlungen (mehrere Wiki Seiten als Linkliste)

Verarbeitungsmodi

- Tabellen
  - durch LaTeX (für die meisten Fälle schöner)
  - durch Chromium (kann alle HTML Tricks)

- Papierformate (A4, A5, B5, letter, legal, executive)

- SVG als Raster- oder Vektorgrafik ins PDF

Alternativen

Wikipedia PDF Funktion

PediaPress

Pandoc

und viele mehr

Pedia Press

- Ausschließlich gedruckte Bücher
- Keine Dateien zum Download
- Nicht Quelloffen

Wikipedia PDF Funktion

- keine professionelle Typographie
- kein ODT, EPUB, LaTeX Quelltext
- keine Seitenzahlen / buchinternen Verweise
- keine Autorenliste / kein Bildnachweis

Pandoc

- Sehr ähnliche Technologie
- Quelloffen
- Benötigt Modifikation um zu Laufen

Technisches

- Programmiersprachen
- Verwendete Programme
- Bibliotheken
- Algorithmen

Programmiersprachen

- Haskell (funktional, genau wie bei Pandoc)
- Python 3 (für die Benutzeroberfläche)
- LaTeX (Satzsystem, genau wie bei Pandoc)

Verwendete Programme

- curl (paralleler und komprimierte Bilddownload)
- lualatex (moderne LaTeX Implementierung)
- ImageMagick / RSVG (Bildverarbeitung)

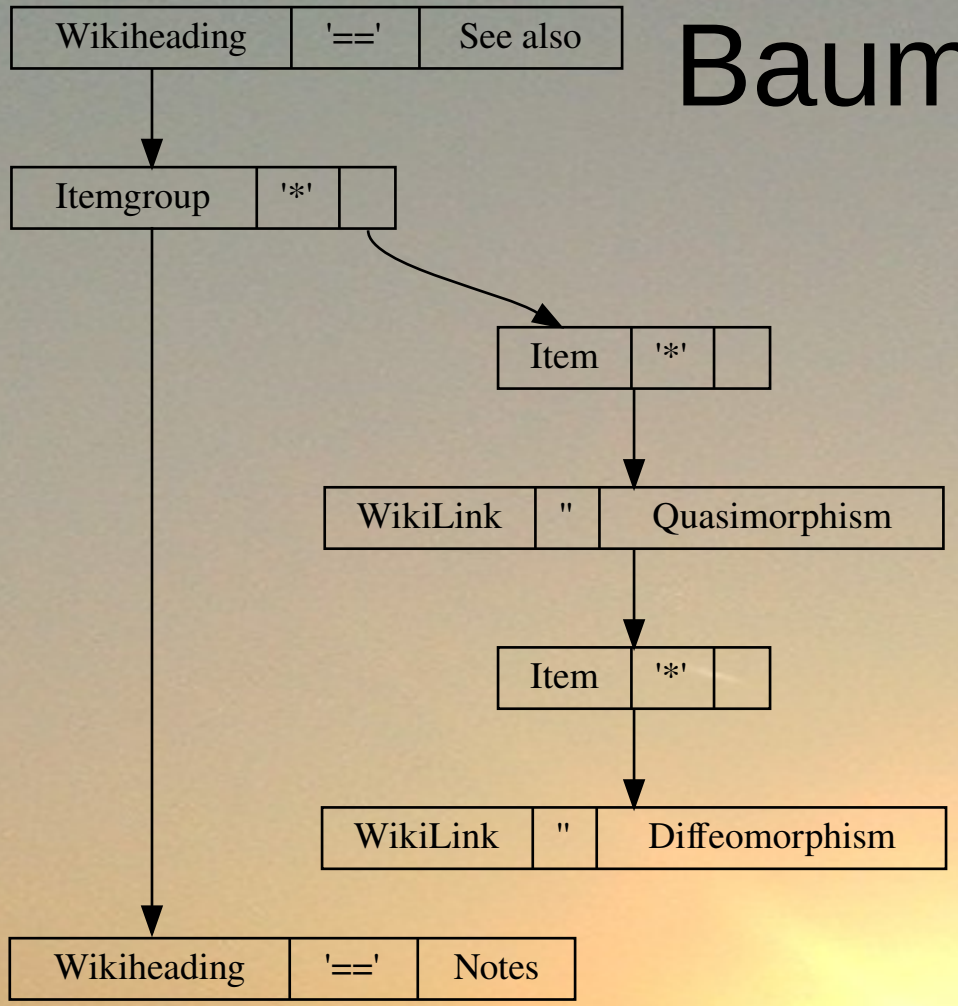
Technologien für den Webservice

- Happstack (Haskell Webserver)
- Blaze HTML (Haskell HTML Bibliothek)
- HTML 5 (Standard Auszeichnungssprache)

Grundlegender Algorithmus

- Lese die Eingabe in eine Baumstruktur.
- Führe einige Berechnungen auf ihr aus.
- Schreibe aus dem Baum nach LaTeX.

Baumstruktur



Wiki-Quelltext:

```
== See also ==
* [[Quasimorphism]]
* [[Diffeomorphism]]
== Notes ==
```

Erstellen des Baums

Bibliothek Parsec (genau wie bei Pandoc)

Konzept: „generische Klammer“ (nur hier)

Generische Klammer

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

Falsch herum geklammert

Falsch herum geschlossene Klammer wie:

Hallo

Findet man häufig im Wikitext der Wikipedia

=> Die Wikisprache ist nicht kontextfrei !

(Beweis über Pumping Lemma)

Beweis mit dem Pumping Lemma

Wenn eine Sprache L kontext-frei ist, dann gibt es eine natürliche Zahl p ≥ 1 so dass man jede Zeichenfolge s in L mit |s| ≥ p schreiben kann als s = uvxyz
Wobei für die Zeichenfolgen u, v, x, y und z, gilt

1. |vxy| ≤ p,
2. |vy| ≥ 1,
3. uv^nxy^nz ist auch in L für alle natürlichen Zahlen n ≥ 0.

$$h \in L \text{ iff } \exists n \in \mathbb{N} : h = \left(^n\left[^n\right)^n\right]^n$$

$$s := \left(^p\left[^p\right)^p\right]^p = uvxyz \Rightarrow \exists n : uv^nxy^nz \notin L$$

Widerspruch!

Korollar: Es gibt keine BNF fürs Wiki

Es gibt keine Backus Naur Form
und schon gar keinen regulären Ausdruck
für die Wiki Syntax.

=> Alle Projekte die versuchen das Wiki so einzulesen funktionieren nicht.

Klammerkorrektur

Aus:
<u>Hallo</u>neueWelt

Wird:
<u>Hallo</u>
neueWelt

Automatische Spaltenbreite

- Drucke jede Spalte auf „unendliches Papier“
  => max. Breite

- Berechne Spaltenbreiten aus max. Breiten

Manuelle Vorlagenverarbeitung

Wiki Quelltext:
{{KastenMitRahmen|inhalt=ein Inhalt}}

Benutzereditierbare Datei:
["KastenMitRahmen","fbox",“inhalt“]

LaTeX Ausgabe:
\fbox{ein Inhalt}

Unicode (Fremd-schriftliches)

- lualatex
- Aufwendiger Font Wechsel im LaTeX Dokument
- Nachteil: Unübersichtlicher LaTeX Quellcode
- Vorteil: Programm vollständig Teil von Debian

Index und Interne Referenzen

Funktioniert!
Aber LaTeX Dokument wird 4 mal kompiliert was den Großteil der Laufzeit des Programms in Anspruch nimmt.

Autorenbestimmung

Freie Lizenzen erfordern meist Nennung des Autors. Text und Bildautoren werden von den Wikiservern geladen. Das von diesen generierte HTML wird hierzu ausgewertet.

Diskussion

Fragen?

https://mediawiki2latex.wmcloud.org/