wiki2book Aus Wikipedia eigene Bücher bauen

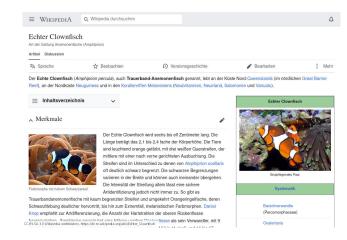
Hauke Stieler

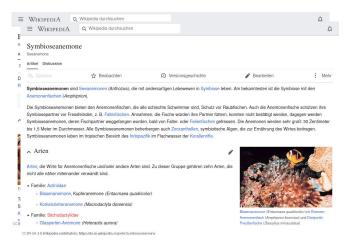
nauke96

11. Dezember 2022

Kennt ihr das?

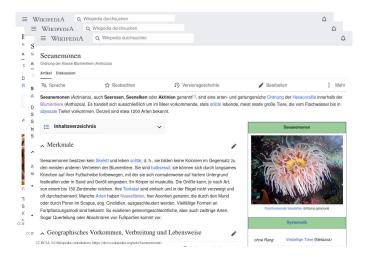




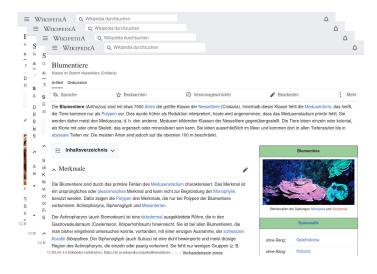




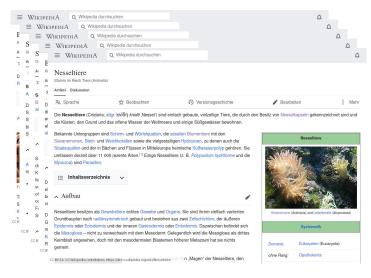
Motivation



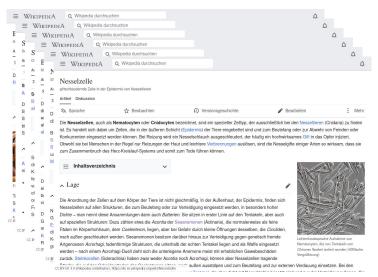












"Going on to Wikipedia to look something up, then unexpectedly being sucked into a seemingly **endless series of link clicking** to end up in a completely different part of wikipedia than you ever meant to go to."

— Urban Dictionary











Existierende Tools

- pandoc
- mediawiki2latex / wb2pdf
- epub-press
- w2eb
- percollate





Existierende Tools

Warum gehen die nicht?

Inhaltliche & visuelle Gründe:

- Formatierung, Schriftgrößen, etc. stimmt nicht
- Templates werden nicht/uneingeschränkt evaluiert
- LaTEX/Math wird nicht in Bild gerendert
- Tabellen funktionieren nicht



Existierende Tools

Warum gehen die nicht?

Technische Gründe:

- Kann nicht mehrere Artikel gleichzeitig
- Bilder werden nicht heruntergeladen
- Wird nicht mehr maintained
- Ist in JavaScript
- Ist in einer Programmiersprache, die ich nicht kann / mag
- Ergebnis ist kein EPUB
- Ergebnis lief nicht auf meinem Tolino

Was will ich haben?

Generierte und gekaufte eBooks sollen sich qualitativ nicht unterscheiden.



Was will ich haben?

Generierte und gekaufte eBooks sollen sich qualitativ nicht unterscheiden.

Allgemeine Anforderungen:

- Formatierung stimmig
- Korrekte Übersetzung/Einbindung von Tabellen, Bilder, Listen, Quellenangaben, etc.
- Wikipedia-spezifische Templates & Kategorien ignorieren





Was will ich haben?

Generierte und gekaufte eBooks sollen sich qualitativ nicht unterscheiden.

Allgemeine Anforderungen:

- Formatierung stimmig
- Korrekte Übersetzung/Einbindung von Tabellen, Bilder, Listen, Quellenangaben, etc.
- Wikipedia-spezifische Templates & Kategorien ignorieren

Persönliche Anforderungen:

- Soll auf meinem Tolino eBook-Reader laufen
- Go als Programmiersprache
- Caching aller heruntergeladenen Daten (zum Coden im Zug)



Formatierung

```
Wikitext ''kann'' auch '''Formattierung'''.
```

Und '''sogar ''alles''' durcheinander'' geht.

Wikitext kann auch Formattierung.

Und sogar alles durcheinander geht.

Links

Interne [[Hyperlink|Links]] gehen.

Auch ins Internetz [https://externe-links], sogar mit [https://foo.bar Namen].

Interne Links gehen.

Auch ins Internetz [1] ☑, sogar mit Namen ☑.

Referenzen & Templates

```
Hi<ref name="foo">{{Internetquelle|url=http://bar.de
|abruf=2022-10-12|titel=Ref mit Template}}</ref>!
```

Die selbe Ref. nochmal!<ref name="foo" />

Hi^[1]!

Die selbe Ref. nochmal![1]

1. ↑ ^{a b} Ref mit Template. ☑ Abgerufen am 12. Oktober 2022.



Überschriften

= Level 1 =

Wird nicht aktiv benutzt, da Titel der Seite h1 ist.

==== Level 4 ====

Die hier wird benutzt.

Level 1

Wird nicht aktiv benutzt, da Titel der Seite h1 ist.

Level 4

Die hier wird benutzt.

Listen

- * Listen
- ** gibt
- es
- # auch

noch

- Listen
 - o gibt

es

- 1. auch
 - 1. noch

Tabellen

```
{| class="wikitable"
|-
! Spalte 1 !! Spalte 2
|-
| Hier
| könnte
|-
| ihre || Werbung stehen
|}
```

Spalte 1	Spalte 2
Hier	könnte
ihre	Werbung stehen



Bilder

Hier ein Bild:

[[Datei:Full moon partially obscured by atmosphere.jpg |mini|Mit Unterschrift.]]

Hier ein Bild:





Und vieles mehr

- Description list
- Zitate
- Einrückungen
- Code
- LATEX-Mathe-Zeug
- Musiknoten
- Gallerien
- Inline Bilder
- Diverse Parameter an allen möglichen Dingen



Instanzen - Artikel

- Instanz pro Land/Sprache \rightarrow z.B. [en|de|nds].wikipedia.org
- Verlinkungen ggf. zu anderen Instanzen möglich



Instanzen – Bilder

- Wikimedia commons (commons.wikimedia.org)
- Normal: upload.wikimedia.org/wikipedia/commons/0/06/Foo.jpg
- Aber auch: upload.wikimedia.org/wikipedia/de/2/26/Son-3.jpg



Instanzen - Bilder

- Wikimedia commons (commons.wikimedia.org)
- Normal: upload.wikimedia.org/wikipedia/commons/0/06/Foo.jpg
- Aber auch: upload.wikimedia.org/wikipedia/de/2/26/Son-3.jpg
- Redirects möglich
 - ▶ Beispiel: File:MET00506.jpg
 - lacktriangle Ggf. ist Dateiname im Artikel eq Dateiname bei Wikimedia commons
 - Nach Bild-Artikel suchen
 - redirects=true Parameter nicht vergessen



Instanzen - Bilder

- Wikimedia commons (commons.wikimedia.org)
- Normal: upload.wikimedia.org/wikipedia/commons/0/06/Foo.jpg

Wikipedia

- Aber auch: upload.wikimedia.org/wikipedia/de/2/26/Son-3.jpg
- Redirects möglich
 - Beispiel: File:MET00506.jpg
 - ▶ Ggf. ist Dateiname im Artikel ≠ Dateiname bei Wikimedia commons
 - Nach Bild-Artikel suchen
 - redirects=true Parameter nicht vergessen
- In Deutschen Artikeln wird natürlich Datei:Sol-3.jpg benutzt



Wikipedia

Instanzen & APIs

APIs – Artikel abfragen

```
Anfrage:
```

```
Puren Wikitext in JSON Antwort verpackt:
GET
     de.wikipedia.org/w/api.php
       ?action=parse
       &format=json
```

&prop=wikitext &page={article name}

Antwort:

```
"parse": {
    "title": "Erde",
    "wikitext": {
        "*": "..."
```



APIs – Templates evaluieren

Anfrage:

Wie bei Artikeln nur andere Parameter.

Antwort:

```
{
    "expandtemplates": {
        "wikitext": "..."
    }
}
```



APIs - Bilder

Aufbau:

```
 upload.wikimedia.org/wikipedia/\{instance\}/ \\ \{MD5[0]\}/\{MD5[0]MD5[1]\}/\{filename\}
```

MD5:

MD5[i] = Das i-te Zeichen des MD5-Hashes von filename



APIs – LATEX-Mathe in Bild umwandeln

- 1. Math-check API für Resource location anfragen
- 2. Eigentliches Bild abfragen



LATEXZU Bild: 1. Resource location bekommen

Anfrage:

URL: POST https://wikimedia.org/api/rest_v1/media/math/check/tex Body: URL encoded form Element q mit dem LaTeX-Code:

q:\sqrt{x}

Antwort:

Header x-resource-location auslesen:

x-resource-location: 73b85c4ec364802ad746381712d10a43f073d50a

LATEXzu Bild: 2. Bild abfragen

Anfrage:

Einfaches GET mit Hash an

wikimedia.org/api/rest_v1/media/math/render/{svg|png}/73b85c4...



Idee & Annahmen

Annahmen:

- lacktriangle Heruntergeladener wikitext ist korrekt ightarrow Keine Syntaxprüfung nötig.
- Formatierung vom HTML (Einrückung, Leerzeilen, etc.) ist egal



Idee & Annahmen

Annahmen:

- lacktriangle Heruntergeladener wikitext ist korrekt ightarrow Keine Syntaxprüfung nötig.
- Formatierung vom HTML (Einrückung, Leerzeilen, etc.) ist egal

Idee:

- Elemente im wikitext rekursiv durch Token ersetzen
- Token in Map speichern: [token] → [token-content]
- Token-Content kann weitere Token enthalten
- Token sollen einfach zu HTML ersetzt/relaxiert werden können



Meine "Compiler" "Pipeline"

wikitext

```
Die '''Erde''' ist ein [[Planet]] ...
```

Meine "Compiler" "Pipeline"

$\mathtt{wikitext} \longrightarrow \mathsf{Token}/\mathsf{AST}$

```
Die '''Erde''' ist ein [[Planet]] ... Die $$TOK_BOLD.1$$ ist ein $$TOK_LINK_2$$ ...
```

Token Map: \$\$TOK_BOLD_1\$\$ → Erde \$\$TOK_LINK_2\$\$ → Planet



Meine "Compiler" "Pipeline"

$\mathtt{wikitext} \longrightarrow \mathsf{Token}/\mathsf{AST} \longrightarrow \mathtt{HTML}$

Meine "Compiler" "Pipeline"

 $\mathtt{wikitext} \longrightarrow \mathsf{Token}/\mathsf{AST} \longrightarrow \mathtt{HTML} \longrightarrow \mathtt{EPUB}$

> Token Map: \$\$TOK_BOLD_1\$\$ → Erde \$\$TOK_LINK_2\$\$ → Planet

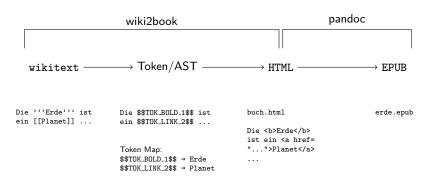
buch.html

erde.epub

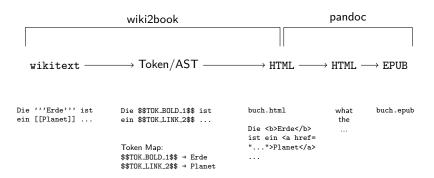
Die Erde
ist ein <a href=
"...">Planet



Meine "Compiler" "Pipeline"



Meine "Compiler" "Pipeline"



Meine "Compiler" "Pipeline"

