DARIAH Working Paper Workflow

Work in Progress

Thorsten Vitt Mirjam Blümm

DARIAH-DE
Working Papers

Thorsten Vitt, Mirjam Blümm: „DARIAH Working Paper Workflow“. *DARIAH-DE Working Papers* Nr. 0. Göttingen: DARIAH-DE, 2016. URN: TODO: urn.

Zusammenfassung

Für die Publikation der DARIAH-Working-Papers gibt es einen Workflow auf der Basis von Markdown, das mit Pandoc und LuaLatex formatiert wird.

Dieser Artikel beschreibt die Installation und einige Spezifika der Working-Paper-Vorlage; Details zur Markdown-Syntax findet man z. B. auf der Pandoc-Homepage.

Inhaltsverzeichnis

1 Artikel schreiben

Die Texte sollen mit Markdown ausgezeichnet werden. Zum Übersetzen wird Pandoc verwendet, es sind entsprechend also Konstrukte aus Pandoc's Markdown[1] möglich.

Diese Datei stellt lediglich die wichtigsten Konstrukte sowie einige Besonderheiten der *DARIAH Working Papers* zusammen, für eine umfassende Dokumentation sei auf die o.g. Pandoc-Dokumentation verwiesen.

1.1 Text

Markdown-Dateien sind einfache Textdateien so wie diese. Zeilenumbrüche werden wie Leerzeichen behandelt, für einen Absatzwechsel schreibt man eine Leerzeile in den Text.

Kursivierungen werden erzeugt, indem man die zu kursivierenden Passagen `_mit Unterstrichen_` (oder alternativ Sternchen) umschließt. Für **Fettdruck** verwendet man `__doppelte Unterstriche__` oder doppelte Sternchen. Überschriften sind Absätze, die (je nach Ebene) mit einem bis drei Rautenzeichen **#** beginnen, gefolgt von einem Leerzeichen:

```
## Überschrift zweiter Ebene
```

1.1.1 Listen

Um eine Liste zu erzeugen, beginnt man eine Zeile mit einem Aufzählungszeichen: *****, **-** oder **+**, gefolgt von einem Leerzeichen. Das Aufzählungszeihen darf bis zu drei Leerzeichen eingerückt sein. Einzelne Leerzeilen zwischen den Listeneinträgen sind erlaubt.

Besteht ein Listeneintrag aus mehreren Absätzen, so sind Leerzeilen obligatorisch und man rücke die Folgeabsätze um vier Leerzeichen ein. Verschachtelte Listen werden ebenfalls um vier Leerzeichen eingerückt.

Nummerierte Listen folgen derselben Syntax:

```
1. Beispieleintrag
2. Noch ein Eintrag.

   Im Gegensatz zum vorherigen besteht dieser Eintrag aus mehreren Absätzen.
   Man beachte die Einrückung.

3. Hier nun eine untergeordnete Liste:

   * Eintrag,
```

[1] http://pandoc.org/MANUAL.html#pandocs-markdown

```
    * noch ein Eintrag,
    * weiterer Eintrag.
```

erzeugt

1. Beispieleintrag
2. Noch ein Eintrag.

   Im Gegensatz zum vorherigen besteht dieser Eintrag aus mehreren Absätzen. Man beachte die Einrückung.
3. Hier nun eine untergeordnete Liste:

   - Eintrag,
   - noch ein Eintrag,
   - weiterer Eintrag.

1.1.2 Code und Blockformate

Um inmitten eines Absatzes ein Stück Code in Festbreitenschrift zu formatieren, umschließt man das enstprechende Stück Code mit Backticks. Ganze Codeblöcke können entweder um vier Leerzeichen eingerückt werden oder – diese Variante empfehlen wir – mit Zeilen aus je drei Backticks umgeben werden. Unmittelbar hinter der einleitenden Backtickreihe kann der Sprachenname angegeben werden, um Syntax-Highlighting zu erreichen:

```
```python
def foo():
    return "bar"
```
```

```python
def foo():
    return "bar"
```

Soll für Gedichte o.ä. der Zeilenfall erhalten bleiben, aber ansonsten normaler Text formatiert werden, beginnt man die Zeilen mit |. Zeilen mit Blockzitaten wird > vorangestellt.

1.1.3 Formeln

Mathematische Formeln können in LaTeX-Syntax eingegeben werden. Inline-Formeln wie in $x_i, i < n$ werden zwischen einfachen Dollarzeichen geschrieben: `$x_i, i < n$`. Für abgesetzte Formeln verwendet man doppelte Dollarzeichen:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

1.1.4 Bilder

Bilder sollten als PDF-, PNG- oder JPEG-Datei mitgeliefert werden. Sie werden über eine Bildreferenz eingebunden, die in einem eigenen Absatz stehen sollte (Leerzeile davor und danach):

```
![Ein Beispielbild](img/Logo_Working-Papers.pdf)
```

Abbildung 1: Ein Beispielbild

In den eckigen Klammern steht die Bildunterschrift, die durchaus Formatierungen enthalten kann.

Bitte beachten Sie, dass Groß- und Kleinschreibung in Bildreferenz und Dateinamen zueinander passen und verzichten Sie möglichst auf Leer- und Sonderzeichen in den Dateinamen.

Ohne weitere Angaben wird eine in den Bildmetadaten hinterlegte Druckgrößenangabe berücksichtigt, die Bildgröße jedoch auf die Größe des Textbereichs begrenzt. Da die entsprechenden Metadaten oft falsch sind, sollten sie bei Bildern in Seitengröße überprüft und ggf. korrigiert werden. Das geht z. B. mit ImageMagick[2], das folgende Kommando setzt z.B. die Auflösung aller JPEG-Bilder auf 300 dpi:

```
mogrify -density 300 -units PixelsPerInch *.jpg
```

Bei Gimp heißt die entsprechend Option *Print Size*. Alternativ sind Größenangaben beim Einbinden des Bilds möglich:

Abbildung 2: 25% der Textbreite

```
![25% der Textbreite](img/Logo_Working-Papers.pdf){width=25%}
```

[2] http://www.imagemagick.org/script/command-line-options.php#density

1.1.5 Links und Fußnoten

Verweise auf Webseiten[3] bestehen i.d.R. aus einem Linktext in eckigen gefolgt von der vollständigen URL (mit http://!) in runden Klammern: `[Webseiten](https://de.dariah.eu/working-papers)`. Im Text wird der Linktext anklickbar, die URL kommt zusätzlich in eine Fußnote. Soll eine URL – wie http://de.dariah.eu/ – im Text auftauchen, so setze man sie in spitze Klammern, sie wird dann zum Link, erzeugt jedoch keine Fußnote.

Sonstige Fußnoten können wie im Beispiel inline[4] oder separat[5] gesetzt werden.

```
Sonstige Fußnoten können wie im Beispiel inline^[dann aber ohne Absätze] oder separat

[^bsp]: Separate Fußnoten können durchaus auch aus mehreren Absätzen bestehen.

    Es gilt die übliche vier-Leerzeichen-Einrückregel.
```

Bei separaten Fußnoten kann das Fußnotenkürzel (hier `bsp`) beliebig gewählt werden, die Fußnote kann an einer beliebigen Stelle (in eigenem Absatz) gesetzt werden. Achtung: In Fußnoten sollten URLs nur in der `<>`-Form gesetzt werden, da Fußnoten in Fußnoten nicht unterstützt werden.

1.2 Bibliographie

Für die Bibliographie empfehlen wir, die Literaturverzeichnis-Einträge im BibLaTeX- oder BibTeX-Format in einer Datei mit gleichem Namen wie der Artikel und der Endung `.bib` zu verwalten und sich für die Zitationen an die entsprechenden Pandoc-Konventionen[6] zu halten – in diesem Fall wird das Literaturverzeichnis automatisch einheitlich und entsprechend der Stilvorlagen formatiert.

Wird ein solches automatisches Literaturverzeichnis verwendet, so muss der Artikel mit diesem Kommando enden:

```
\biblio
```

Das Kommando setzt automatisch die entsprechende Überschrift und passt die Formatierungsvorgaben an.

It is easily possible to include references (**???**). To do so we recommend the following:

1. Write or export your bibliography as a BibLaTeX database named alike your article – i. e. for `test.md`, it should be called `test.bib`.
2. End your markdown document with a line that reads `\bibliography`.

[3] https://de.dariah.eu/working-papers
[4] dann aber ohne Absätze
[5] Separate Fußnoten können durchaus auch aus mehreren Absätzen bestehen.
Es gilt die übliche vier-Leerzeichen-Einrückregel.
[6] http://pandoc.org/MANUAL.html#citations

References to your bibliography can be written in a number of ways:

- Easiest way is to write `[@hh2010]` (**???**).
- You can also include prefixes and page references as in `[see @hh2010, p. 1]` (see **???**, p. 1).
- Text references look like (**???**), without the brackets: `@hh2010a`
- Multiple references share brackets (see, e. g., **???**; **???**; **???**)
- „As Hagen (**???**) says" is a common beginning for which the author of a citation can be suppressed

1.3 Titeldaten

Titeldaten und einige Einstellungen gehören in einen Metadatenblock im YAML-Format. Der Block beginnt mit einer Zeile aus drei Bindestrichen `---` und endet mit einer Zeile aus drei Punkten `...`. Metadatenfelder beginnen mit dem Feldnamen am Anfang der Zeile, dann folgt ein Doppelpunkt und ein Leerzeichen und schließlich der Inhalt des Felds.

Einige Felder (z. B. die Autorenliste) kann mehrere Werte aufnehmen. Dazu schreibt man eine YAML-Liste: Die Zeile mit dem Feldnamen endet nach dem Doppelpunkt, darauf folgt ein Listeneintrag pro Zeile, beginnend mit einem Bindestrich. Das Feld `abstract` kann mehrere Absätze umfassen, dazu endet die Zeile mit dem Schlüsselwort mit einem `|` und es folgen die Textabsätze eingerückt. Der Metadatenblock kann also z. B. so aussehen:

```
---
title: DARIAH Working Paper Workflow
subtitle: Spaß mit Pandoc
author:
- Thorsten Vitt
- Mirjam Blümm
lang: de
date: 2016
abstract: |
    Für die Publikation der DARIAH-Working-Papers empfehlen wir einen Workflow
    auf der Basis von Markdown, das mit Pandoc und LuaLatex formatiert wird.

    Dieser Artikel beschreibt die Installation und einige Spezifika der
    Working-Paper-Vorlage; Details zur Markdown-Syntax findet man z.B. auf der
    Pandoc-Homepage.
...
```

Die folgenden Metadatenfelder stehen zur Verfügung:

Feld	Bedeutung
title	Titel des Artikels.
subtitle (optional)	Untertitel.
lang	Sprache, in der der Artikel verfasst ist: `de` oder `en`.
author	Autor des Artikels. Bei mehreren Autoren Liste verwenden.
longauthor (optional)	Autoren mit Fußnotenzeichen für Institute
institute	Institut(e), ggf. mit Fußnotenzeichen (Liste möglich)
date	Veröffentlichungsjahr
abstract	Zusammenfassung
keywords-de	Schlagwörter auf Deutsch (als Liste)
keywords-en	Schlagwörter auf Englisch (als Liste)
wpno	DARIAH-Working-Papers Nr. (wird von der Redaktion eingesetzt)
urn	URN (wird von der Redaktion eingesetzt)

Für Texte, die zuvor als DARIAH-Report veröffentlicht worden sind, sollen die folgenden Metadaten ergänzt werden:

Feld	Bedeutung
report-number	Nummer des Reports, z. B. `1.2.3`
report-date	Veröffentlichungszeitraum, z. B. `Dezember 2015`
report-fkz (optional)	Förderkennzeichen

Für weitere Anmerkungen, die in einem eigenen Matadatenfeld ergänzt werden sollen, stehen folgende Felder zur Verfügung:

Feld	Bedeutung
publish-note	Zusätzliche Angaben (Freitext) vor allem für die Quellenangabe von Erstpublikationen
urn-alt (optional)	URN der Erstveröffentlichung

2 Redaktionsumgebung

2.1 Voraussetzungen

- Pandoc
- LaTeX-Installation mit LuaLaTeX, z.B. aktuelles TeX-Live
- pandoc-citeproc (für das Literaturverzeichnis-Processing)

- Python 3 (für das Convenience-Skript)

2.2 Installation

(Es gibt ein einfaches, experimentelles Installationsscript `install.sh`, das auf Linux und MacOS X funktionieren sollte.)

Die Schrift WeblySleek UI[7] entsprechend dem Betriebssystem installieren. Für übliche Linux-Distributionen ist es ausreichend, die TTF-Dateien in den Ordner `~/.fonts` zu legen.

Einige Dateien aus diesem Verzeichnis müssen über das Dateisystem verteilt werden. Ich empfehle, die entsprechenden Dateien per symbolischem Link zu verlinken statt sie zu kopieren, um für Aktualisierungen gerüstet zu sein:

- **DWP.latex** ist das Pandoc-Template, es muss in das Verzeichnis `~/.pandoc/templates`.
- Die **`*.csl`-Dateien** sind die Styles für die Literaturverwaltung, sie stammen aus dem Zotero Style Repository[8]. `dwp.py` sucht diese Styles ebenfalls im Verzeichnis `~/.pandoc/templates`
- Die Bilder aus dem `img`-Ordner werden für die Titelseite benötigt. Sie werden in einem LaTeX-Baum gesucht.
- `dwp.py` ist ein Script zum einfachen Aufrufen von Pandoc mit entsprechenden Parametern. Es sollte irgendwo in den `$PATH`.

Unter der Annahme, dass dieses Verzeichnis unter `~/projects/dwp-template` liegt:

```
mkdir -p ~/.pandoc/templates
cd ~/.pandoc/templates
ln -s ~/projects/dwp-template .
mkdir -p `kpsexpand '$TEXMFHOME/tex/latex'`
cd `kpsexpand '$TEXMFHOME/tex/latex'`
ln -s ~/projects/dwp-template .
cd /usr/local/bin
sudo ln -s ~/projects/dwp-template .
```

Nach erfolgreicher Installation sollte es in jedem Verzeichnis möglich sein, mit `dwp datei.md` die entsprechende Datei in PDF zu übersetzen.

[7] http://www.dafont.com/weblysleek-ui.font
[8] https://www.zotero.org/styles?q=chicago&format=author-date

2.3 Benutzung

Das beigefügte Skript `dwp.py` kann einfach mit `dwp artikeldatei.md` aufgerufen werden. Es ruft Pandoc mit den richtigen Parametern auf, um `artikeldatei.pdf` zu erzeugen. Das kann dann z. B. so aussehen:

```
pandoc -o article.pdf --latex-engine=lualatex --template=DWP \
       --filter=pandoc-citeproc --bibliography=article.bib \
       --csl=$HOME/.pandoc/templates/chicago-author-date.csl \
       --metadata=link-citations:true \
       article.md
```

`dwp` kann auch mit weiteren Pandoc-Parametern aufgerufen werden, es reicht alle Parameter an Pandoc weiter und lässt dafür ggf. die selbst generierten Versionen weg.

2.4 Troubleshooting

2.4.1 irgendwas mit UTF-8

Eingabedatei ist nicht UTF-8-codiert. Im Texteditor öffnen und als UTF-8 speichern.

2.4.2 PDF-Datei kann nicht erzeugt werden, keine ordentliche Fehlermeldung

1. TeX-Datei erzeugen lassen – das geht entweder auf die entsprechende Rückfrage oder mit `dwp -o article.tex article.md`
2. TeX-Datei manuell mit LuaLaTeX übersetzen lassen: `lualatex article.tex`, Fehlermeldungen prüfen

2.5 Umgang mit Dateien in Office-Formaten

Wenn Dateien in Office-Formaten angeliefert werden empfiehlt sich folgender Workflow:

1. OpenOffice-Dateien nach DOCX konvertieren, die Pandoc-Unterstützung von DOCX ist besser
2. `pandoc -o article.md --extract-media=article_assets article.docx`. Erzeugt aus der Worddatei eine Markdown-Datei namens `article.md`, extrahiert die Bilder und speichert sie im Verzeichnis `article_assets` (wird ggf. angelegt)
3. Markdown-Datei nachbearbeiten

Vorteil: Bilder werden gleich ausgepackt, Grundformatierungen bleiben erhalten