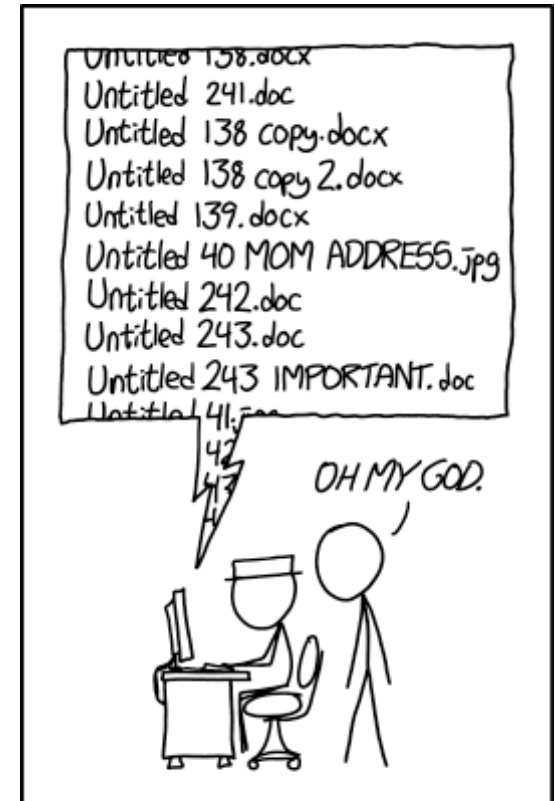


Praxislabor Digitale Geisteswissenschaften

Tidy Data - Basics

Agnes Brauer
a.brauer@ub.uni-frankfurt.de



PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

[\(https://xkcd.com/1459/\)](https://xkcd.com/1459/)

Warum sollte ich mir über meine Daten(formate) Gedanken machen?

- Sauber aufbereitete Daten ermöglichen:
 - Reproduzierbarkeit
 - Nachnutzung (für andere, aber auch für sich selbst)
 - Skalierbarkeit

„Goldene“ Regeln im Umgang mit Daten(-formaten)

Regel 1: Es ist gut, Dateien sinnvoll zu benennen und bei „seinem“ Schema zu bleiben

Bad practice

- Unbenannt.docx
- Agnes' Dateiname mit Leerzeichen und schönen Umlauten.xlsx
- Abbildung 1.png
- Abb 3.png
- Protokoll*.txt



Good practice

- 2022-05-24_Abstract_DSH.docx
- Agnes_Dateiname_macht_Fortschritte.xlsx
- Abb01_Umfrage-DH-Workshop.png
- 2022-05-24_Protokoll_Ergaenzungen-Agnes.txt



„Goldene“ Regeln im Umgang mit Daten(-formaten)

- sinnvolle Dateinamen sind:
 - maschinenlesbar und
 - menschenlesbar und
 - sinnvoll sortierbar
 - mit regulären Ausdrücken und Wildcards gut auffindbar und damit gruppierbar

„Goldene“ Regeln im Umgang mit Daten(-formaten)

- sinnvolle Dateinamen vermeiden daher:
 - Leerzeichen
 - Sonderzeichen
 - Satzzeichen

„Goldene“ Regeln im Umgang mit Daten(-formaten)

- sinnvolle Dateinamen verwenden:
 - Trennzeichen zur Kurzinfo über den Inhalt einer Datei, z.B.
 - „_“ für Metadaten
 - „-“ als Worttrenner für bessere Lesbarkeit
 - Numerische Zeichen am Anfang und führende Nullen

Beispiele

Embrace the *slug*



```
01_marshal-data.r  
02_pre-dea-filtering.r  
03_dea-with-limma-voom.r  
04_explore-dea-results.r  
90_limma-model-term-name-fiasco.r  
helper01_load-counts.r  
helper02_load-exp-des.r  
helper03_load-focus-statinf.r  
helper04_extract-and-tidy.r
```

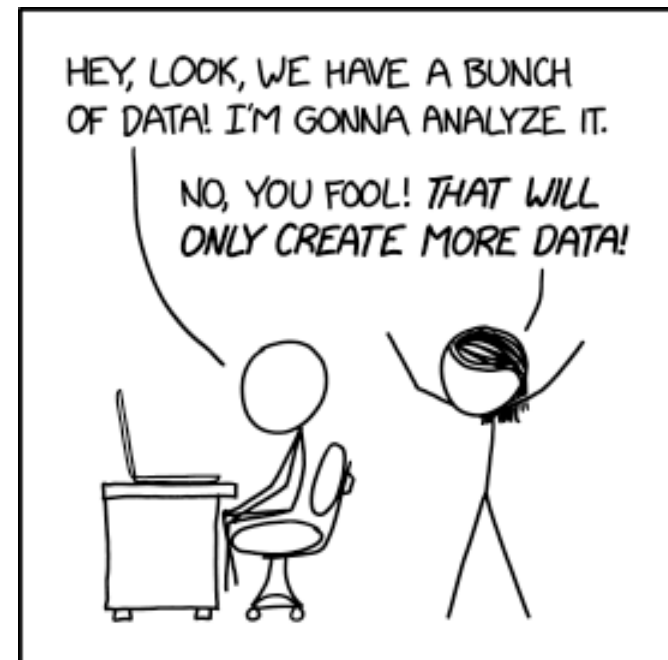
Quelle: <https://datacarpentry.org/rr-organization1/01-file-naming/index.html>

„Goldene“ Regeln im Umgang mit Daten(-formaten)

Regel 2: Es ist gut, für jedes Projekt eigene Ordner mit sinnvollen Unterordnern anzulegen

Ein Projektordner kann z.B. immer so aufgebaut sein:

- 📁 Abbildungen
- 📁 Auswertung
- 📁 Daten
- 📁 Paper
- 📁 Skripte
- 📄 README.md



Quelle: <https://xkcd.com/2582/>

„Goldene“ Regeln im Umgang mit Daten(-formaten)

Regel 3: Noch besser ist es, die eigenen Daten/Dateien mit Git zu versionieren → Workshop am 6. und 13. Juli!

Wozu Git?

- Versionierung
- Zusammenarbeit
- Wiederherstellung zurückliegender Versionen
- Nachvollziehbarkeit
- Nachhaltigkeit

„Goldene“ Regeln im Umgang mit Daten(-formaten)

Regel 4: Es ist gut, Plain Text-Formate zu nutzen...

Beispiele für Plain Text-Formate:



Vorteile:

- plattformunabhängig
- langfristig verfügbar
- unabhängig von proprietärer Software
- flexibler Datenaustausch

Ikons:
<https://commons.wikimedia.org/wiki/File:Text-tsv-text.svg>
<https://commons.wikimedia.org/wiki/File:Text-csv-text.svg>
<https://commons.wikimedia.org/wiki/File:Text-txt.svg>

„Goldene“ Regeln im Umgang mit Daten(-formaten)

... und maschinenlesbare Notationen für Formatierungen zu nutzen, z.B. Markdown

```
---  
tags: Praxislabor  
---
```

```
# Praxislabor Digitale Geisteswissenschaften - Sommersemester 22
```

```
Alle Materialien sind im [moodle-Kurs](https://moodle.studiumdigitale.uni-frankfurt.de/moodle/course/view.php?id=2817) verfügbar.
```



Praxislabor Digitale Geisteswissenschaften - Sommersemester 22

Alle Materialien sind im [moodle-Kurs](#) verfügbar.



“One rule of thumb is if you can’t find it by Ctrl+F/Command+F it isn’t machine readable.”

Quelle: <https://librarycarpentry.org/lc-overview/06-file-naming-formatting/index.html>

„Goldene“ Regeln im Umgang mit Daten(-formaten)

Regel 4: Ausgangsdaten werden nicht geändert

→ Alle Operationen erfolgen auf einer Kopie des Datensets!



“Keep raw data raw”

Vgl.: <https://dataoneorg.github.io/Education/bestpractices/preserve-information-keep>

Organisation von tabellarischen Daten

- Sinnvolle Organisation der Daten (Tabellendesign) ist die Grundlage datengetriebenen Arbeitens
- Grundsätze
 - Variablen(-namen) stehen in Spalten, z.B. Datum, Dauer etc.
 - Jede Zeile ist eine Beobachtung / Erhebung
 - Jede Zelle enthält die Daten / Werte der Beobachtung
 - Erhobene und bereinigte Daten werden als textbasierte Datei, z.B. *.csv, exportiert

Organisation von tabellarischen Daten

Übung:

- Laden Sie bitte [diese Datei](#) herunter
- Was könnte an dieser Art der Datenerhebung nicht ganz optimal sein? Was könnte man besser realisieren? Fassen Sie bitte ihre Beobachtungen [im Pad](#) zusammen.

Organisation von tabellarischen Daten

Don'ts



- mehrere Tabellen in einem Datenblatt
- Verwendung von Tabs bei der Datenerfassung
- Nullen werden nicht erfasst („Zero observations are real data!“)
- ungeeignete Nullwerte werden verwendet (z.B. “0”, “999”, “None”) → sie werden nicht als solche interpretiert
- Formatierung enthält Information (z.B. farbliche Hervorhebungen)
- Formatierung zur “Aufhübschung” der Tabelle (z.B. Zellen verbinden → nicht verwertbar für Statistikprogramme)
- Kommertare oder Einheiten in Zellen
- Mehrere Werte pro Zelle
- Datenfeldnamen beginnen mit Zahlen, enthalten Leer- oder Sonderzeichen, z.B.:
“Maximale Temperatur in C°”
- Metadaten in den Daten, z.B. “Legenden” zur Erklärung von Spaltenbezeichnern etc.

Organisation von tabellarischen Daten

Dos



- eine Tabelle pro Datei
- eine Null wird immer eingetragen
- Nullwerte werden angemessen erfasst, z.B. "NA"
- Formatierungen sind für's Auge; zum Speichern von Informationen finden wir andere Lösungen
- Metadaten und ggf. Kommentare gehören in eine separate Datei
- eine Zelle → ein Wert
- Datenfeldnamen / Bezeichner werden besonnen gewählt und konsistent gehalten, d.h. Verwendung von "_" und/oder Binnengroßschreibung, z.B.: "Max_Temp_C"
- Sonderzeichen werden vermieden

Organisation von tabellarischen Daten

Datumsangaben in Tabellen(kalkulationsprogrammen)

Übung: Schauen Sie sich bitte in der Übungstabelle die Datumsangaben für 2015 an. Extrahieren Sie Tag, Monat und Jahr jeweils in eine neue Spalte mit den entsprechenden Funktionen:

=TAG(A2)

=MONAT(A2)

=JAHR(A2)

Welche Beobachtung können Sie machen?

s. auch: <https://uc3.cdlib.org/2014/04/09/abandon-all-hope-ye-who-enter-dates-in-excel/>

Organisation von tabellarischen Daten

Datumsangaben und Excel

Excel speichert Datumsangaben intern als Zahlen!

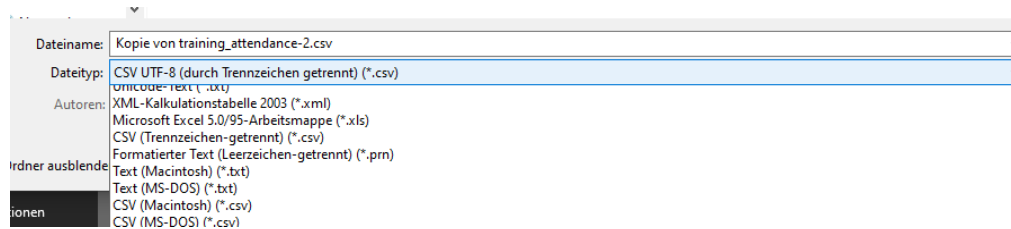
- Vorteil: Rechnen mit Datumsangaben ist sehr einfach!
- Nachteil: Interoperabilität!

s. auch: <https://uc3.cdlib.org/2014/04/09/abandon-all-hope-ye-who-enter-dates-in-excel/>

Organisation von tabellarischen Daten

„Awareness“-Übung

- Speichern Sie die Übungsdatei als csv-Datei ab (Datenblatt „Dates“)



- Öffnen Sie die Datei in einem Texteditor ...
- und dann wieder in Excel

Was stellen Sie fest?

Organisation von tabellarischen Daten

Alternativen

- einzelne Spalten für Tag, Monat, Jahr (s.o.)
- Noch besser: Ein einzelner String im ISO-Format 8601: **YYYYMMDDhhmmss**
Bsp: **20150324172535** für 24. März, 2015 17:25:35



<https://xkcd.com/1179/>

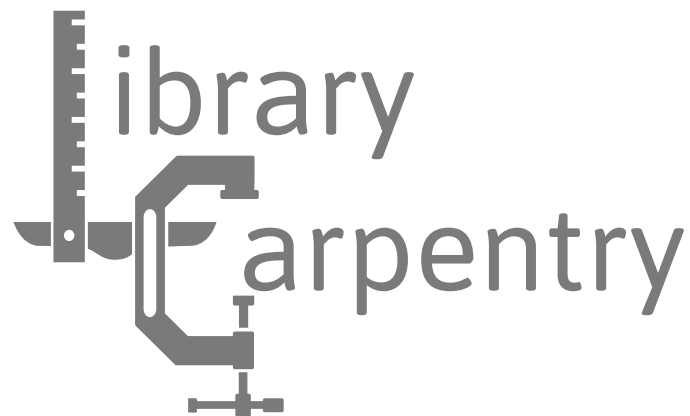
Tipps zur Qualitätssicherung

- Schreibschutz für Raw data verwenden
- Funktion „Datenüberprüfung“ nutzen
- Sortierfunktionen nutzen, um Inkonsistenzen zu identifizieren
- Erfasste Daten im Plain text-Format speichern und weiterverarbeiten

Quellen

- Data Carpentry Reproducible Research Committee. 2016. "File organization for reproducible research", <https://datacarpentry.org/rr-organization1/01-file-naming/index.html>
- Library Carpentry: Tidy data for librarians, <https://librarycarpentry.org/lc-spreadsheets/>
- Library Carpentry: Workshop Overview. September 2019, <https://librarycarpentry.org/lc-overview/06-file-naming-formatting/index.html>
- Hadley Wickham, Tidy Data, Vol. 59, Issue 10, Sep 2014, Journal of Statistical Software, <http://www.jstatsoft.org/v59/i10>

Danke für Ihre Aufmerksamkeit!



Workshop konzipiert in Anlehnung an "[Library Carpentry: Tidy Data Lessons for Librarians.](#)"