Dateinamen & Ordnerstrukturen

📂 Ordnung halten 😎

Dr. Gabriele Schwiertz & Dr. Denis Arnold

2024-11-05

Dinge mit Daten

Ordnung halten?

- Überblick behalten auf dem eigenen Rechner
- Für wen?
  - für mich: erleichtert mir die Arbeit
  - im Projekt: wenn ich mit anderen zusammenarbeite
  - für die Nachnutzung: wenn meine Daten auch in Zukunft benutzt werden sollen
  - für mich in der Zukunft: wenn ich in einem Jahr nochmal drauf gucken will

Dinge mit Daten

Übersicht

- Übung: 3 Gruppen
- Prinzipien der Dateibenennung
- Prinzipien Ordnerstrukturen
- Versionierung
- Back-up-Strategien

Dinge mit Daten

Übung: Dateinamen

- Ladet Euch den jeweiligen Ordner für Eure Gruppe aus Ilias runter
- Entzippt den Ordner und schaut Euch an, was drin ist
- Ihr habt eine Email von Eurem Professor bekommen: emailDringend.png
- Lest die Mail und versucht, den Auftrag auszuführen
- Schickt die fertige ppt-Datei an gabriele.schwiertz@uni-koeln.de
- Wenn ihr fertig seid, notiert, was an der Ordnerstruktur und den Dateinamen gut, bzw. schlecht war

Dinge mit Daten

Übung: Ergebnisse

- Welche Gruppe war am schnellsten?
- Zeigt kurz, wie Ihr die Datei gefunden habt
- Was sind gute bzw. schlechte Praktiken in bezug auf Dateinamen und Ordnerstrukturen?

Dinge mit Daten

Dateinamen

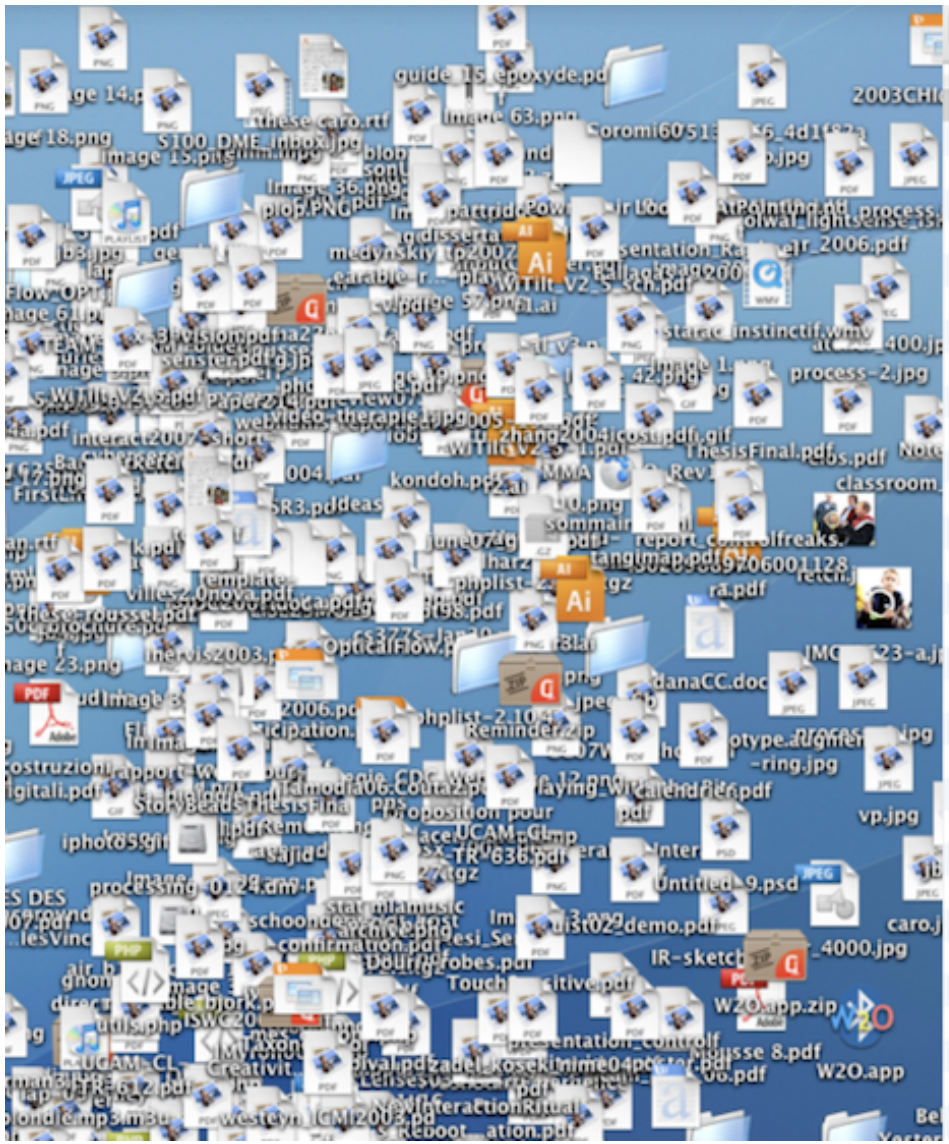
Dinge mit Daten

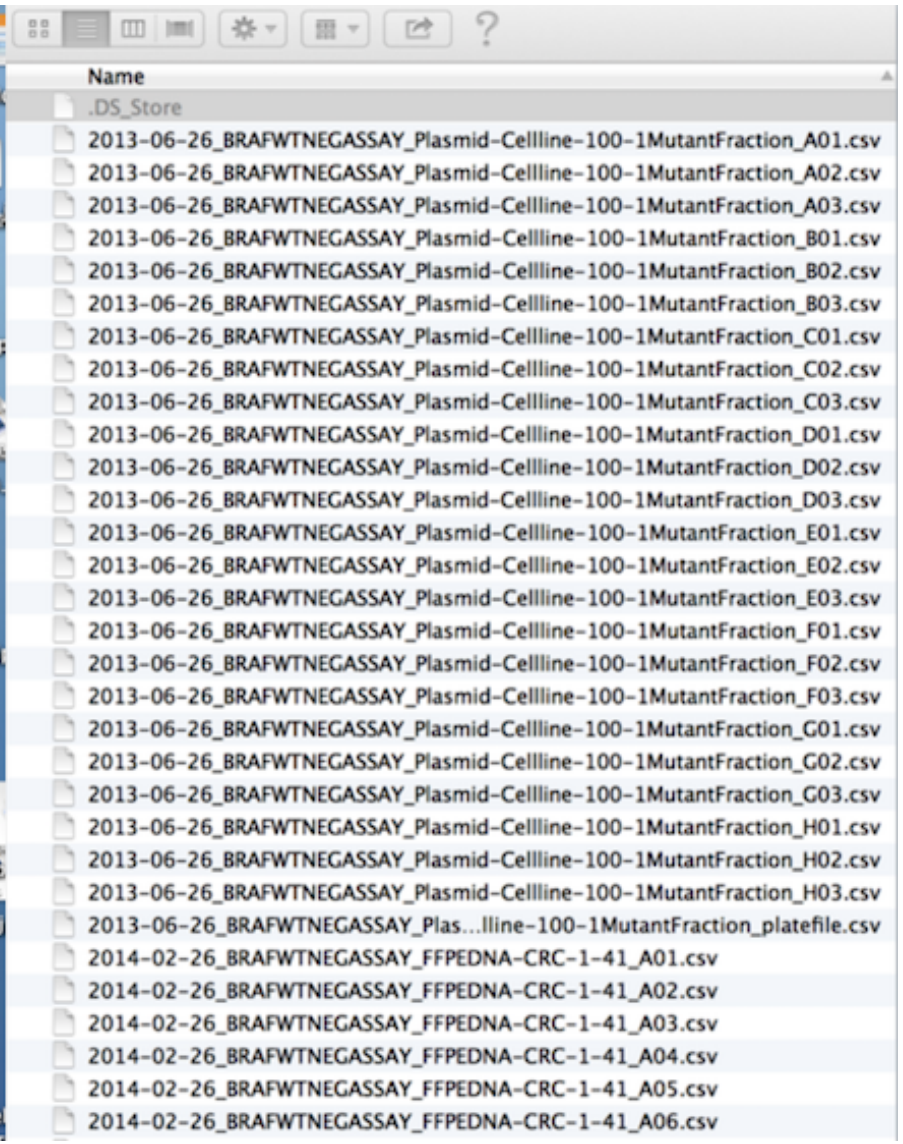
Warum man über Dateinamen nachdenken sollte

- selbst in relativ simplen Projekten (Hausarbeit) gibt es viele Dateien
- sehr viele Dateien
- Dateien verändern sich
- Dateien stehen zueinander in Beziehung

Dinge mit Daten

Es kann kompliziert werden…





Dinge mit Daten

Strategien gegen das Dateienchaos

Dateinamen und Ordnerstrukturen können helfen!

Dateiname und Speicherort (Ordner, wo sich die Datei befindet) sollten direkt Auskunft geben über:

- Was die Datei ist
- Warum es sie gibt
- Wie ihre Beziehung zu anderen Dateien ist
- Je mehr Dinge selbsterklärend sind, desto besser!

Dinge mit Daten

Gute und schlechte Namen

🥲

myabstract.docx

Joe's tolle Datei .txt

figure 2.png

fig1.png

JWD2(%

$Nicht^löschenoderdukkannsteinpacken!!!!.xlsx

😍

2022-10-01_abstract_for_DAE.docx

JoeVerbessertSich.txt

fig01_scatterplot_butterfly_density.png

fig02_boxplot_butterfly_density.png

Dinge mit Daten

1986-04-22_rawdata_challenger.csv

3 Prinzipien für Dateinamen

1. Maschinenlesbar
2. Menschenlesbar
3. Klappt gut mit der Default-Sortierung

Dinge mit Daten

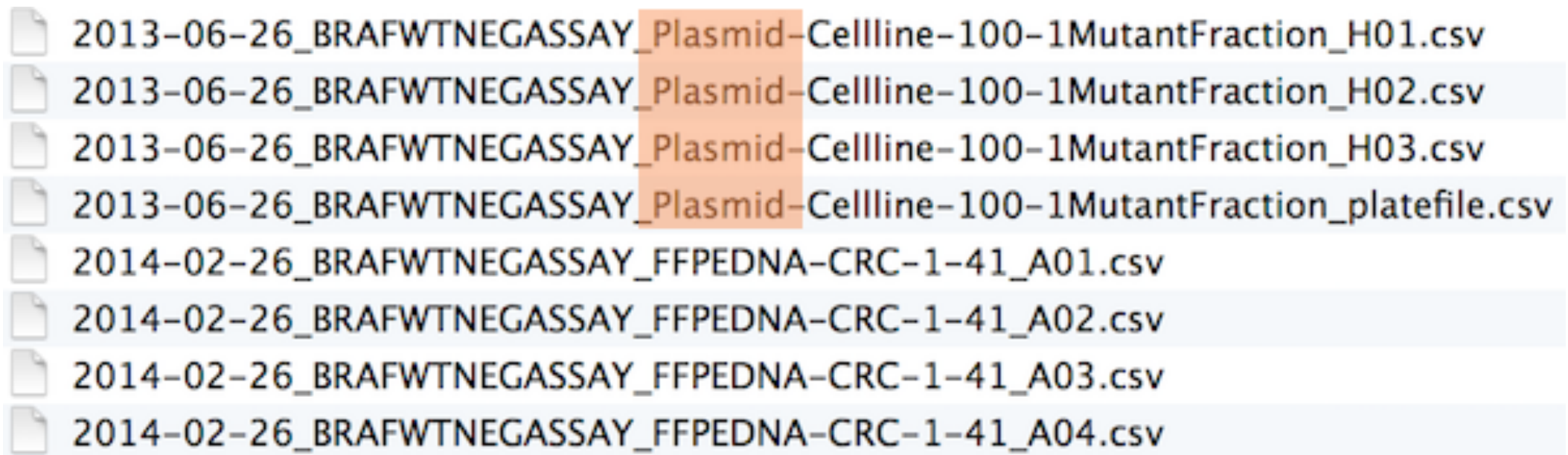
Maschinenlesbar

- kann gut mit regulären Ausdrücken/globbing/filtern bearbeitet werden
  - keine Leerzeichen in Datei oder Ordnernamen
  - keine Interpunktionszeichen (!,?“.)
  - keine Sonderzeichen (ß, ⁎, |, ä, $, の, ぱ, â, ǫ, 🕐)
- leicht in Skripte einzubauen
  - wohlbedachte Nutzung von Trennzeichen (_ oder -)
- Case-Sesitivity beachten: a vs. A?

Dinge mit Daten

Filtern und Suchen mit Globbing/Regulären Ausdrücken

Auszüge aus Dateilisten Filtern:

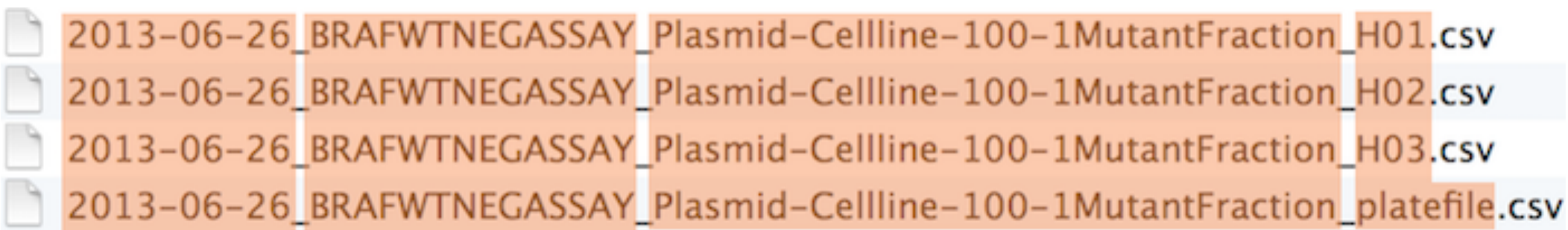


Dinge mit Daten

Trennzeichen mit Bedeutung

Konsistente Nutzung von “-” und “_” macht es möglich, Metadaten aus Dateinamen zu ziehen:

- Der Unterstrich “_” trennt **Metadaten-Einheiten**, auf die man später zugreifen will
- Der Bindestrich “-” trennt Wörter o.ä. **innerhalb** dieser Einheiten



Dinge mit Daten

Trennzeichen mit Bedeutung

Metadaten-Information aus Dateinamen mit sinnvollen Trennzeichen können leicht in Tabellen extrahiert werden (z.B. mit R, der Shell, Python,…)

```
> flist <- list.files(pattern = "Plasmid") %>% head

> stringr::str_split_fixed(flist, "[_\\.]", 5)
     [,1]         [,2]              [,3]                                  [,4]  [,5]
[1,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A01" "csv"
[2,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A02" "csv"
[3,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A03" "csv"
[4,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B01" "csv"
[5,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B02" "csv"
[6,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B03" "csv"
```

date assay sample set well

Dinge mit Daten

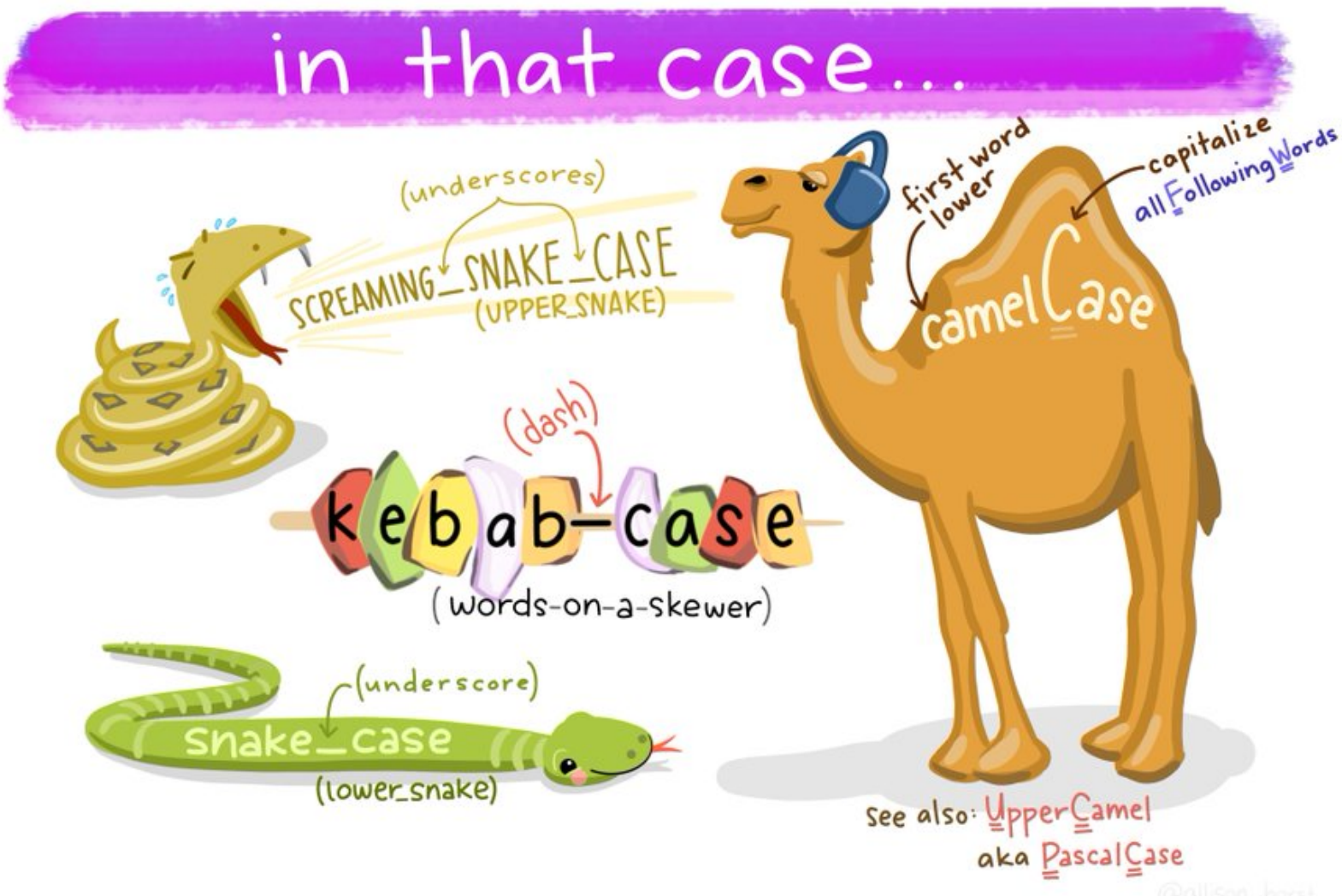
Case-Sensitivity beachten: a vs. A?

- Manche Betriebssysteme machen einen Unterschied zwischen a und A
- Manche Betriebssysteme machen **keinen** Unterschied zwischen a und A
- Daher: Keine Dateinamen, die sich nur in Groß- und Kleinschreibung unterscheiden verwenden!
- Konsistent benennen

test.docx vs. Test.docx vs. TEST.docx

Dinge mit Daten

Konsistente Benennung



Dinge mit Daten

`snake_case`

Pros: Concise when it consists of a few words.
Cons: Redundant as hell when it gets longer.

```
push_something_to_first_queue, pop_what, get_whatever…
```

`PascalCase`

Pros: Seems neat.

```
GetItem, SetItem, Convert, ...
```

Cons: Barely used. (why?)

`camelCase`

Pros: Widely used in the programmer community.
Cons: Looks ugly when a few methods are n-worded.

```
push, reserve, beginBuilding, ...
```

`skewer-case`

Pros: Easy to type.

```
easier-than-capitals, easier-than-underscore, …
```

Cons: Any sane language freaks out when you try it.

`SCREAMING_SNAKE_CASE`

Pros: Can demonstrate your anger with text.
Cons: Makes your eyes deaf.

```
LOOK_AT_THIS, LOOK_AT_THAT, LOOK_HERE_YOU_MORON, …
```

`nocase`

Pros: Looks professional.
Cons: Misleading af.

```
supersexyhippothalamus, bool penisbig, …
```

`fUcKtHeCaSe`

Pros: Can live outside of the law.
Cons: Can be out of a job.

Dinge mit Daten

Nicht zu lang

- Windows MAXPATH 260
- Der Pfad inklusive Dateiname darf nicht zu lang werden
- Das kann bei langen Datei- und Ordnernamen und tiefer Einbettung passieren
- Gleichgewicht zwischen zu vollen Ordnern und einer zu tiefen Einbettung in die Ordnerstruktur

```
"C:
\Users\schwiertz\Documents\DataLiteracy\Beispielpfad\Communiations\2018\Workshop
\Entomology\ButterfliesMothsAndRelatedInsects\Analysis_preliminary
\Scripts_for_Preparation\Python_2_7_1_1\Data_Cleaning_and_Wrangling
\FirstPartButterfliesOnly\myplots\"
```

Dinge mit Daten

Zusammenfassung: Maschinenlesbarkeit

💪 Macht das spätere Durchsuchen einfacher
💪 Macht das spätere Filtern einfacher
💪 Macht das Extrahieren von Informationen einfacher

Wenig Erfahrungen mit globbing und Regulären Ausdrücken? Dann lieber:

☠️ keine Leerzeichen
☠️ keine Interpunktionszeichen
☠️ keine Sonderzeichen

✓ sinnvolle Trennzeichen

✓ **Nur englische Buchstaben, Zahlen, Bindestriche und Unterstriche erlaubt!**

Dinge mit Daten

3 Prinzipien für Dateinamen

1. Maschinenlesbar
2. Menschenlesbar
3. Klappt gut mit der Default-Sortierung

Dinge mit Daten

Menschenlesbarkeit

- Name enthält Information über den Inhalt
- Vgl. das Konzept *slug* in semantischen URLs

https://de.wikipedia.org/?curid=112763

oder

https://de.wikipedia.org/wiki/Sonnenblume

Sonnenblume = slug

Dinge mit Daten

Ein Beispiel

Welche Dateien hättest Du gerne in der Nacht vor der Abgabe?

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

```
01.md
01.r
02.md
02.r
03.md
03.r
04.md
04.r
90.md
90.r
Makefile
figure
helper01.r
helper02.r
helper03.r
helper04.r
tmp.txt
```

Dinge mit Daten

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
```

Dinge mit Daten

Beziehungen zwischen Dateien herstellen

R-Skripte:

01_daten-bereinigen.r
02_vorfiltern.r
03_ergebnisse-explorieren.r
04_statistik.r

Die Plots, die dabei entstehen:

01_daten-bereinigen_sprachen_scatterplot.png
02_vorfiltern_sprachen_barplot.png
02_vorfiltern_sprachen_barplot-facets.png
04_statistik_sprachen-vs-anzahl_boxplot.png

Dinge mit Daten

Zusammenfassung: Menschenlesbarkeit

Es ist leicht herauszufinden, was in der Datei drin ist, einfach durch einen sinnvollen Dateinamen!

Dinge mit Daten

3 Prinzipien für Dateinamen

1. Maschinenlesbar
2. Menschenlesbar
3. Klappt gut mit der Default-Sortierung

Dinge mit Daten

Klappt gut mit der Default-Sortierung

- Zahlen voranstellen
- Datumsangabe gemäß ISO 8601
- Zahlen Nullen voranstellen

Dinge mit Daten

Zahlen voranstellen

Chronologische, bzw. logische Sortierung durch vorangestellte Zahlen

Z.B. R-Skripte von vorher:

01_daten-bereinigen.r

02_vorfiltern.r

03_ergebnisse-explorieren.r

04_statistik.r

Dinge mit Daten

Datumsangabe: YYYY-MM-DD

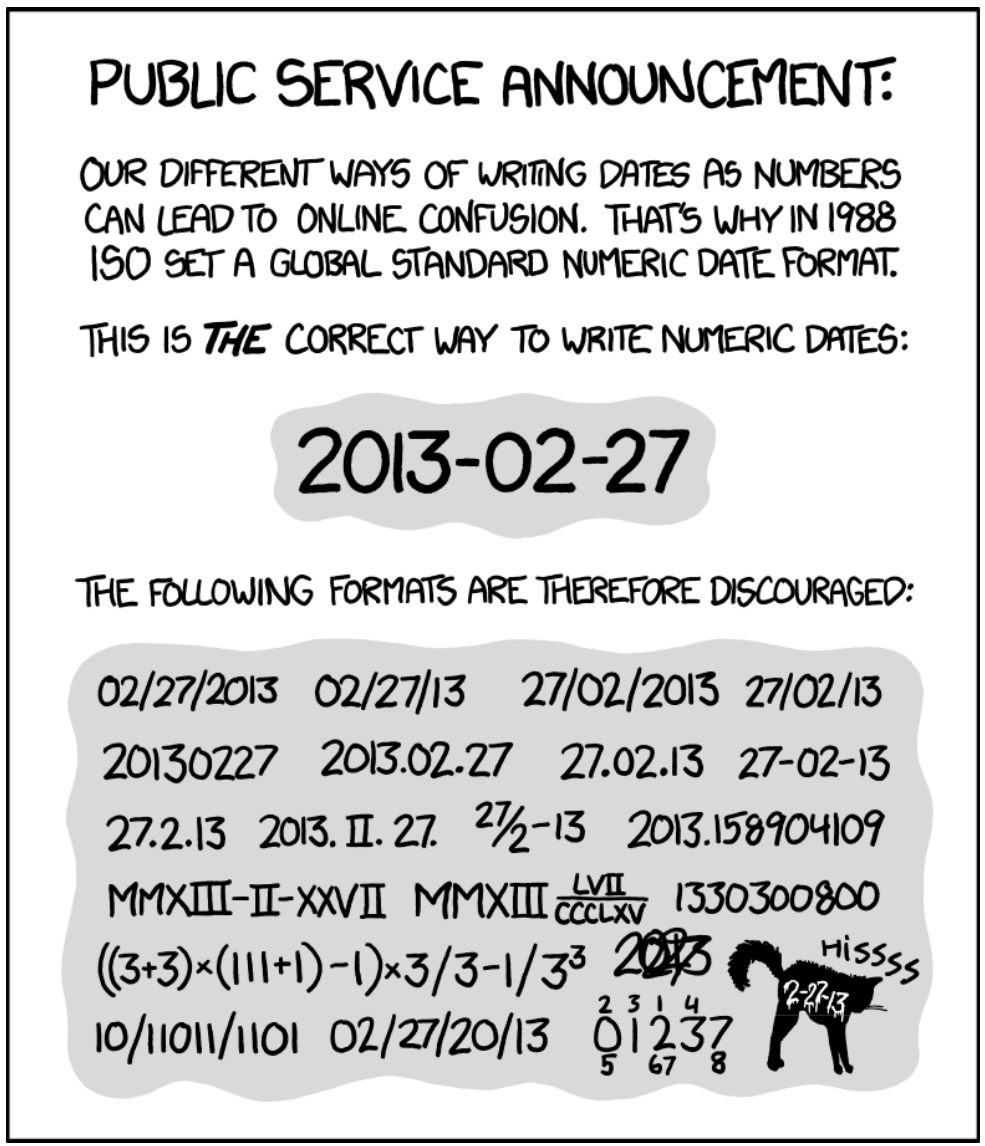
Mithilfe von Datumsangaben werden Dateien automatisch chronologisch sortiert.

1-April-2012.png
1-Jan-2009.png
1-Jan-2012.png
12-Jan-2012.png
2-Jan-2012.png
31-Dec-2009.png

2009-01-01.png
2009-12-31.png
2012-01-01.png
2012-01-03.png
2012-01-12.png
2009-04-01.png

Dinge mit Daten

Datumsangabe: YYYY-MM-DD



Dinge mit Daten

Datumsangabe: YYYY-MM-DD

Dinge mit Daten

Datumsangabe: YYYY-MM-DD

Comprehensive Map of all countries using the Format MM-DD-YYYY

Dinge mit Daten

Zahlen Nullen voranstellen

Ohne Nullen sieht es so aus:

10_plots-fuer-konferenz_final.png
1_data-cleaning.R
2_data-exploration.R
21_boxplot.png
3_data-wrangling.R

Das kann niemand wollen 😟

```
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
```

Dinge mit Daten

Zusammenfassung: Klappt gut mit der Default-Sortierung

- Zahlen voranstellen
- Datumsangabe gemäß ISO 8601
- Zahlen Nullen voranstellen

Dinge mit Daten

3 Prinzipien für Dateinamen

1. Maschinenlesbar
2. Menschenlesbar
3. Klappt gut mit der Default-Sortierung

Dinge mit Daten

Ordnernamen

Dinge mit Daten

Ziele

- Integrität der Daten
- Portabilität
- einfach, später selbst Schritte nochmal nachvollziehen zu können
- einfach, Mitarbeiter:innen mit an Board zu holen

Dinge mit Daten

Ordnerstrukturen und Pfade

- Bei der Benennung: s.o. Dateinamen
- Trennung von Daten, Analyse, Output
- Klarmachen von Beziehungen zwischen den Arbeitsschritten
- nie händisch Rohdaten bearbeiten, später nicht mehr nachvollziehbar
- ein Ordner für die Rohdaten (“raw-data”, read-only)
- weitere Ordner für weitere Bearbeitungsphasen (“data”)
- gute Dokumentation (readme)

Dinge mit Daten

Ordnerpfade

- 📂 Projekt
  - 📄 01_mein-plan.docx
  - 📄 02_datensammlung.csv
  - 📄 03_analyse.R

Pfad, Dateiname, Extension

Die Pfade zu den Dateien sehen so aus:

💻 Windows:

- Projekt\01_meinplan.docx
- Projekt\02_datensammlung.csv
- Projekt\03_analyse.R

🍎 Apple & 🐧 Linux:

- Projekt/01_meinplan.docx
- Projekt/02_datensammlung.csv
- Projekt/03_analyse.R

Dinge mit Daten

ReadMe: Meine Gebrauchsanweisung

- Idealerweise gibt es ReadMe-Dateien für ein Projekt/Ordner
- Einfache Text- oder Markdowndatei, in der steht, was in dem Ordner drin ist und wofür es benutzt wird
- Auch Dateinamen-Konventionen sollten hier erklärt sein
- Hinweise zur Versionierung
- In einem publizierten Datensatz hilft die ReadMe-Datei dem User sich zurechtzufinden

Dinge mit Daten

Ordnerstruktur für ein größeres Projekt:

- Gemeinsam entscheiden, was wo liegt
- Aufschreiben
- Administratives vs. inhaltliches (Reisekostenordner, Analyseordner, Publikation)

Dinge mit Daten

Ordnerstruktur für ein größeres Projekt:

- Auf Ordnungsprinzipien auch in Unterordnern einigen
  - Nach Zweck: Warum werden Daten erhoben (Solar-Studie, Arbeitspaket 7b,…)
  - Nach Kommunikation: Wo werden die Daten diskutiert (Konferenz XY, Workshop Z…)
  - Nach Chronologie: Wie werden die Daten zeitlich ein geordnet (Messzeit, Versuchsreihe, …)
  - Nach Verarbeitungsgrad: Wie stark wurden die Daten bearbeitet? (Rohdaten, Prozessierung, Ergebnis)
  - Nach Ursprung: Wo kommen die Daten her? (Person, Gebiet, Gerät,…)
  - …

Dinge mit Daten

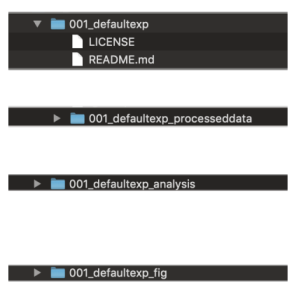
Ordnerstruktur für ein größeres Projekt:

Es gibt templates dafür

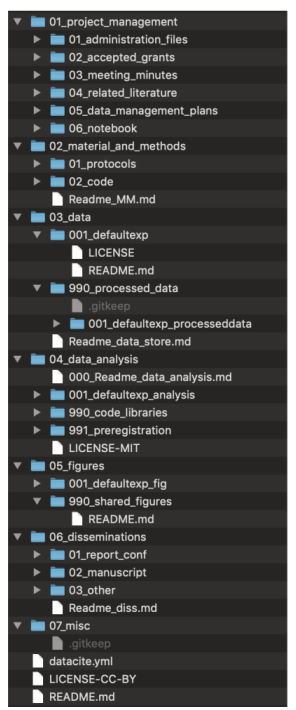
z. B. GinTonic

Experiment level:

add sub-folders for each experiment

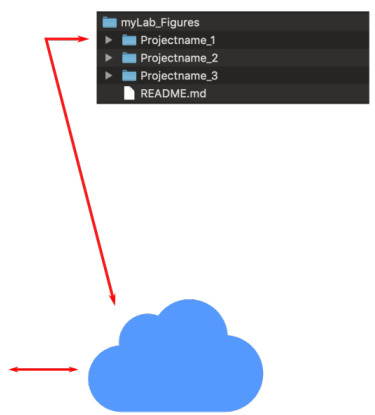


Project level



Laboratory level:

miror sub-folders in other structures

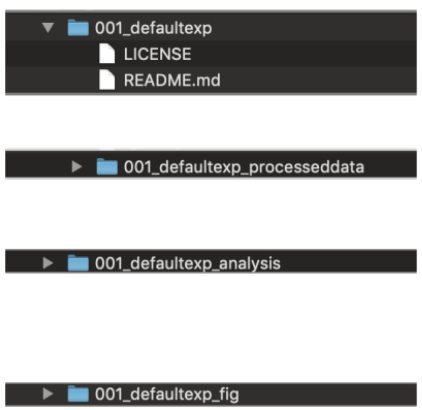


Dinge mit Daten

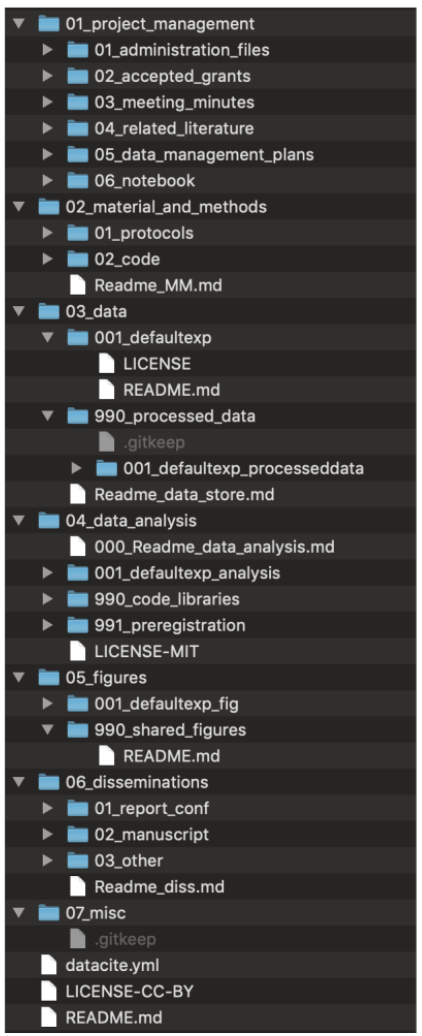
GinTonic

Experiment level:

add sub-folders for each experiment

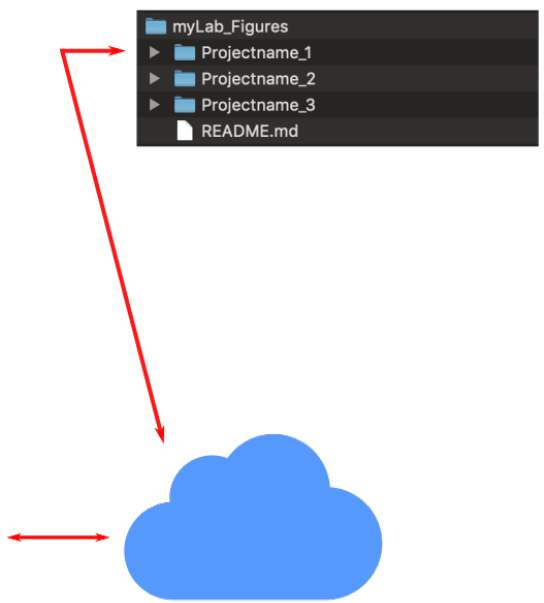


Project level



Laboratory level:

miror sub-folders in other structures

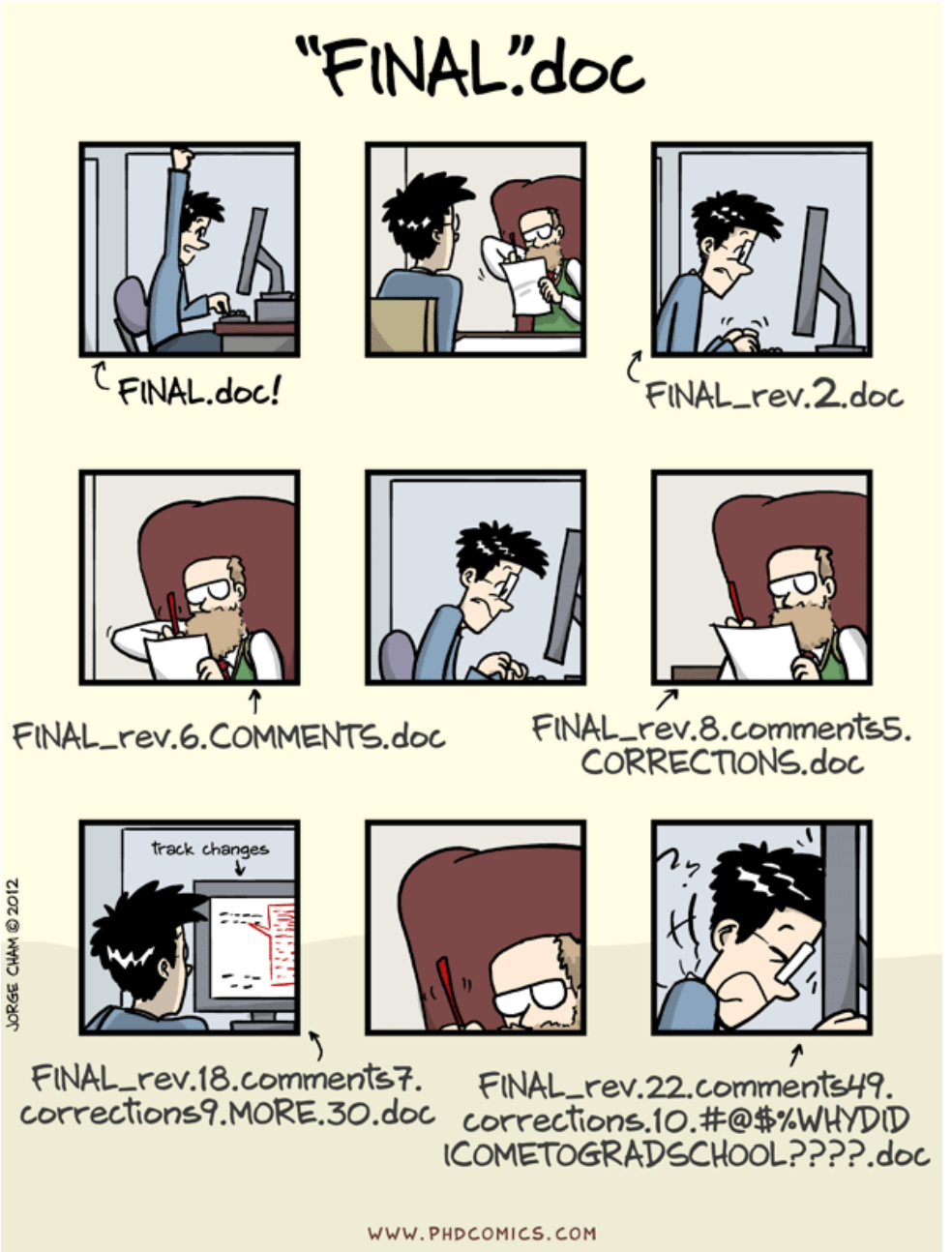


Dinge mit Daten

Versionierung

Dinge mit Daten

- Im Arbeitsprozess entstehen viele Versionen einer Datei - Wie damit umgehen? - Versionierung mit Datum text_2022-04-14.docx - Versionierung mit Nummern text_03.docx - Einen Ordner anlegen für alte Versionen - Professionell: git!



Dinge mit Daten

Backups

Gründe für Datenverlust

- 😫 menschliches Versagen
-  Viren & Malware
- ⌨️ Softwarefehler
- 👮Diebstahl & Verlust (Laptop geklaut, USB-Stick verloren)
- 🔥 Katastrophen (Feuer, Überschwemmung,…)

Dinge mit Daten

Backups

3-2-1-Regel

- Mindestens 3 Kopien der Daten
- Auf mindestens 2 verschiedenen Speichermedien
- 1 davon sollte denzentral/extern liegen (Cloud)

Dinge mit Daten

Zusammenfassung

- Man kann durch geschickte und konsistente Benennung von Dateien und Ordnern Ordnung ins Chaos bringen
- Das sollte eine bewusste Entscheidung sein
- Ihr wisst, worüber man nachdenken sollte
- Ihr wisst, was Ordnerpfade sind und welche Zeichen nicht verwendet werden dürfen
- Ihr wisst, was Versionierung ist (später dazu mehr)
- Ihr wisst, was eine minimale Backup-Strategie ist

Dinge mit Daten

Quellen

Die Folien basieren auf:

https://annakrystalli.me/rrresearchACCE20/filenaming-view.html

https://slides.djnavarro.net/project-structure/#1

https://speakerdeck.com/jennybc/how-to-name-files

https://tu-dresden.de/forschung-transfer/services-fuer-forschende/kontaktstelle-forschungsdaten/news/ordnerstruktur-logik?set_language=en

Bildquellen:

all countries: https://img-9gag-fun.9cache.com/photo/a2mXmGd_700b.jpg

future self: Photo by Amy Shamblen on Unsplash, edited by Gabriele Schwiertz

Dinge mit Daten

Übung: Backup

- Tauscht Euch über Eure Backup-Strategien aus
- Nutzt die Zeit, um ein Backup anzulegen

Dinge mit Daten

Übung: Ordnerstruktur

- Legt einen leeren Ordner an
- Erarbeitet eine bessere neue Ordnerstruktur für Eure
- Tauscht Euch mit der/dem Nachbar*in dazu aus
- Gebt Euch selbst regeln für die Dateibenennung
  - Schreibt sie in einer Datei auf
  - CamelCase oder snake_case
  - Mit Datum am Ende? 2024-11-05
  - Mit Versionsnummer? v2
- HA: räumt alles in die neue Ordnerstruktur um, wenn sie Euch gefällt

Dinge mit Daten