

(Halb-) Automatisiert Codebooks erstellen

Reproduzierbarkeit und Nachnutzung fördern mit dem codebook Package

Jürgen Schneider

16 February 2023

Vorab

Folien, Übungen, ...

bit.ly/codebook-workshop

ein paar Fragen

cbk-wrk.formr.org

Disclaimer



**RUBEN ARSLAN
CREATING THE
CODEBOOK
PACKAGE**



**ME GIVING
A WORKSHOP
ABOUT IT**

Heutiger Workshop

- Warum ein Codebook?
 - Reproduzierbarkeit
 - Nachnutzung
 - Eigenschaften eines Codebooks
- Hands-on

Reproduzierbarkeit

Reproduzierbarkeit vs. Replizierbarkeit

Reproduzierbarkeit und Replizierbarkeit zwischen Disziplinen unterschiedlich verwendet (Barba, 2018)

...

Reproduzierbarkeit

“obtaining consistent results using the **same input data; computational steps, methods, and code**; and conditions of analysis” (NAS, 2019, p. 46)

= “computational reproducibility”

Replizierbarkeit

“obtaining consistent results **across studies** aimed at answering the same scientific question, each of which has obtained its **own data**” (NAS, 2019, p. 46)

Reproduzierbarkeit

Warum?

- Prüfen von “scientific claims” (Transparenz → Lesende beurteilen selbst)
 - Robustness: Alternative Auswertungen durch Reviewer*innen / Lesende
 - Selective reporting (z. B. erkennbar welche Variablen nicht berichtet wurden)
 - P-hacking (z. B. erkennbar welche Ausprägungen Variablen hatten vor Gruppierung)
- Fördert auch Replikationen (Errington et al., 2021)
- Eigene Nachvollziehbarkeit des Datensatzes & der Analysen

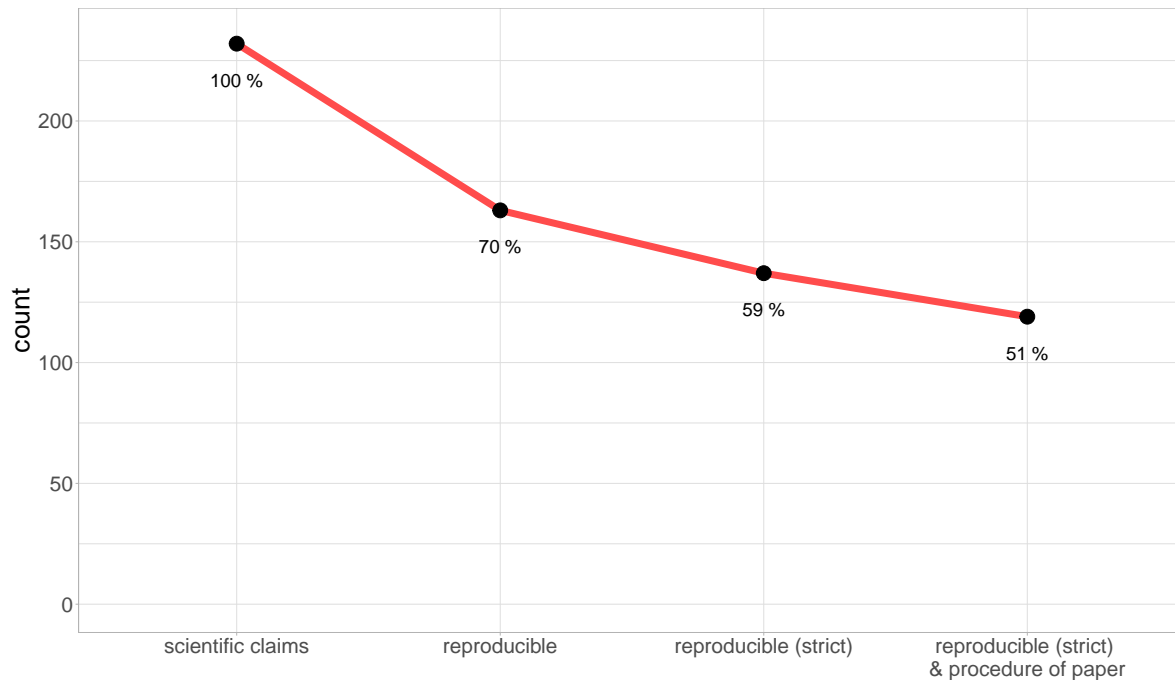
Reproduzierbarkeit

Optimierungspotential

Artner et al. (2021): 232 scientific claims aus 46 Zeitschriftenartikel

```
artner <- data.frame(what = factor(c("scientific claims",  
                                   "reproducible",  
                                   "reproducible (strict)",  
                                   "reproducible (strict) \n& procedure of paper"),  
                    levels = c("scientific claims",  
                               "reproducible",  
                               "reproducible (strict)",  
                               "reproducible (strict) \n& procedure of paper")),  
                  count = c(232, 163, 137, 119),  
                  percent = c("100 %", "70 %", "59 %", "51 %"))  
  
ggplot(artner, aes(x=what, y=count)) +  
  stat_summary(fun=mean, colour="#ff4c4c", geom="line", aes(group = 1), size = 3) +  
  geom_point(size = 6) +  
  annotate("text", x = artner$what, y = artner$count - 15, label = artner$percent, size =  
  scale_y_continuous(limits = c(0,235)) +  
  xlab("") +  
  theme_light() +  
  theme(text = element_text(size = 24))
```

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.



Reproduzierbarkeit

Optimierungspotential

Artner et al. (2021): **232** scientific claims aus **46** Zeitschriftenartikel

Nachnutzung

Definition

Archivierung

- langfristige Speichern von Forschungsdaten typischerweise auf einschlägigen **Repositorien**
- **nicht** primär der Zweck der Nachnutzung (sondern beispielsweise der Transparenz)

Bereitstellung

- Aufbereitung & Kontextualisierung von Forschungsdaten
- Speicherung auf **Forschungsdatenzentren**
- **primär** der Zweck der Nachnutzung

Anforderung bei Drittmittel (BMBF, 2021; Commission, 2017; DFG, 2022; DFG, 2015)

. . .

Nachnutzung

- Reanalyse, Replikation, Sekundärdatenanalyse
- Forschungssynthesen
- Einsatz in der Lehre

(DGfE, 2017)

Nachnutzung

Warum

- Ressourcenschonung
- Analysepotential oftmals nicht ausgeschöpft
- Forschungssynthesen (z.B. meta-analytic SEM)

Wie hilft ein Codebook?

FAIR Data herstellen



FAIR Data herstellen

Findable

Accessible

Interoperable

Reusable

FAIR Data herstellen

Findable

F2. Daten werden mit umfangreichen Metadaten beschrieben *“Forscherinnen und Forscher sollten jeden Datensatz sorgfältig und möglichst vollständig mit Metadaten beschreiben.”*

Accessible

Interoperable

Reusable

R1. (Meta)Daten sind detailliert beschrieben und enthalten präzise, relevante Attribute *“Eine gute Beschreibung von Daten und Metadaten sorgt dafür, dass die Daten für die zukünftige Forschung wiederverwendet werden können”*

(Wilkinson et al., 2016)

💡 TIB-Blog: [Rolle der Wissenschaftler*innen](#)(Kraft, 2017)

Codebook

- Ermöglicht Metadaten → Erklärt Variablen- und Wertelabels
- Überblick über Merkmale der Items / Skalen → Nachvollziehbarkeit der Daten (auch für eigene Arbeitsgruppe) → Data Cleaning & Qualitätskontrolle (Codierfehler, unerwartete Verteilung, ...)
- nicht zu unterschätzen: Standardisierung

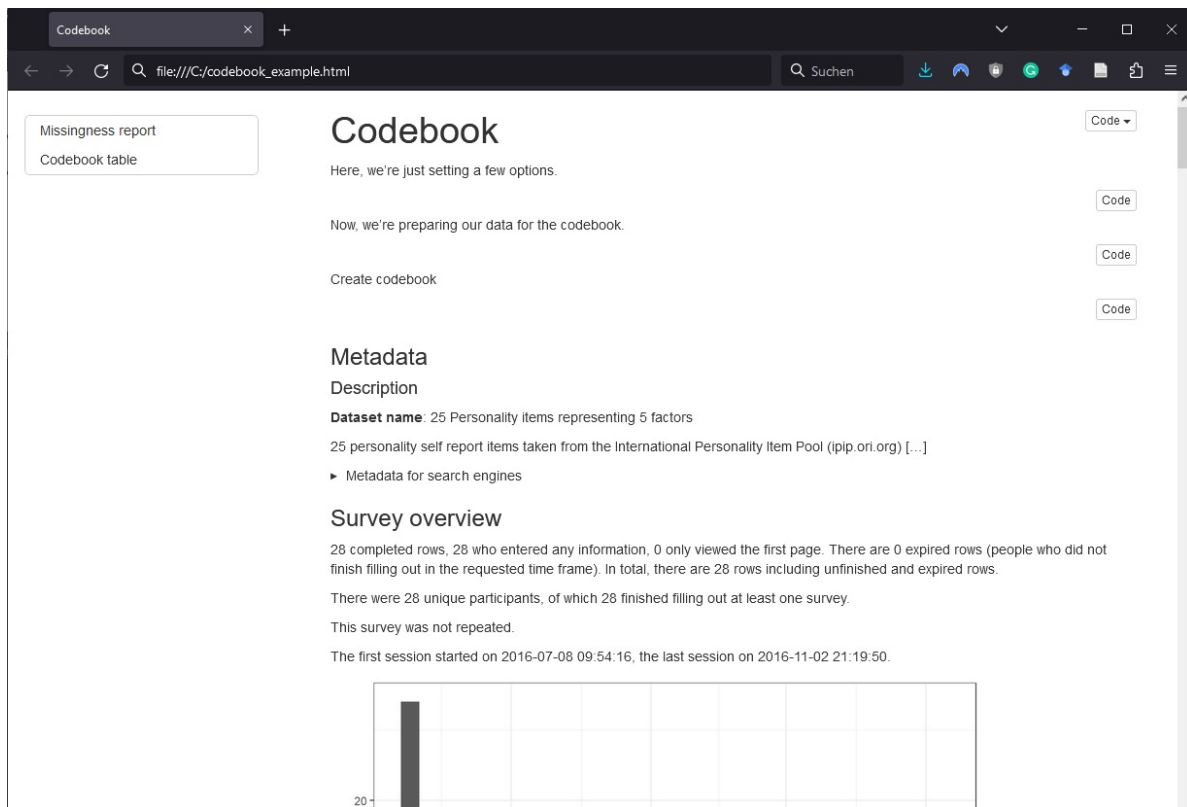
Das codebook package

codebook package

- stellt R Markdown Template & R-Funktionen bereit
- exportierbar in HTML, PDF, ...
- Idealerweise in RStudio verwenden


```
← → ↩ 🔍 Knit on Save ABC 🔍 Knit ⚙️
Source Visual
1 ---
2 title: "codebook"
3 output:
4   html_document:
5     toc: true
6     toc_depth: 4
7     toc_float: true
8     code_folding: 'hide'
9     self_contained: true
10  pdf_document:
11    toc: yes
12    toc_depth: 4
13    latex_engine: xelatex
14 ---
15
16 Here, we're just setting a few options.
17
18 ```{r setup}
19 knitr::opts_chunk$set(
20   warning = FALSE, # show warnings during codebook generation
21   message = FALSE, # show messages during codebook generation
22   error = TRUE, # do not interrupt codebook generation in case of errors,
23                 # usually better for debugging
24   echo = TRUE # show R code
25 )
26 ggplot2::theme_set(ggplot2::theme_bw())
27 ```
28
29
30 Now, we're preparing our data for the codebook.
31
32 ```{r prepare_codebook}
33 library(codebook)
34 codebook_data <- codebook::bfi
35 # to import an SPSS file from the same folder uncomment and edit the line below
36 # codebook_data <- rio::import("mydata.sav")
37 # for Stata
38 # codebook_data <- rio::import("mydata.dta")
39 # for CSV
```





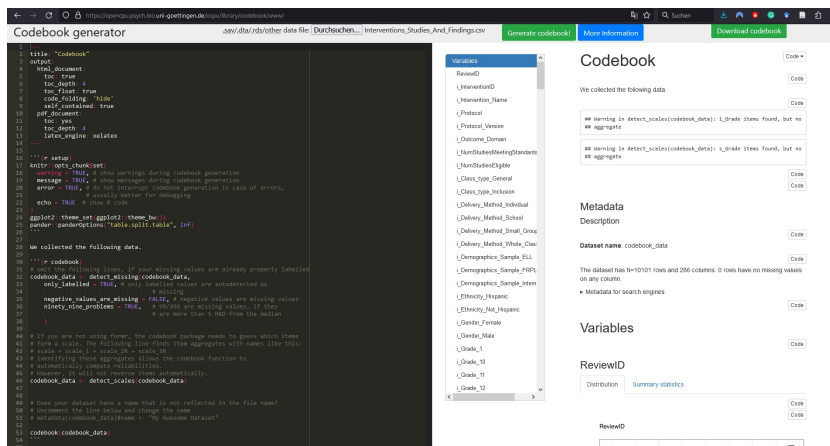
codebook package

Das exportierte Codebook:

- enthält Metadaten der Daten (Variablenlabels, Wertelabels, Beschreibung, ...)
- enthält Datenzusammenfassungen auf Item- & Skalenebene
- ist maschinenlesbar ([Beispiel](#))

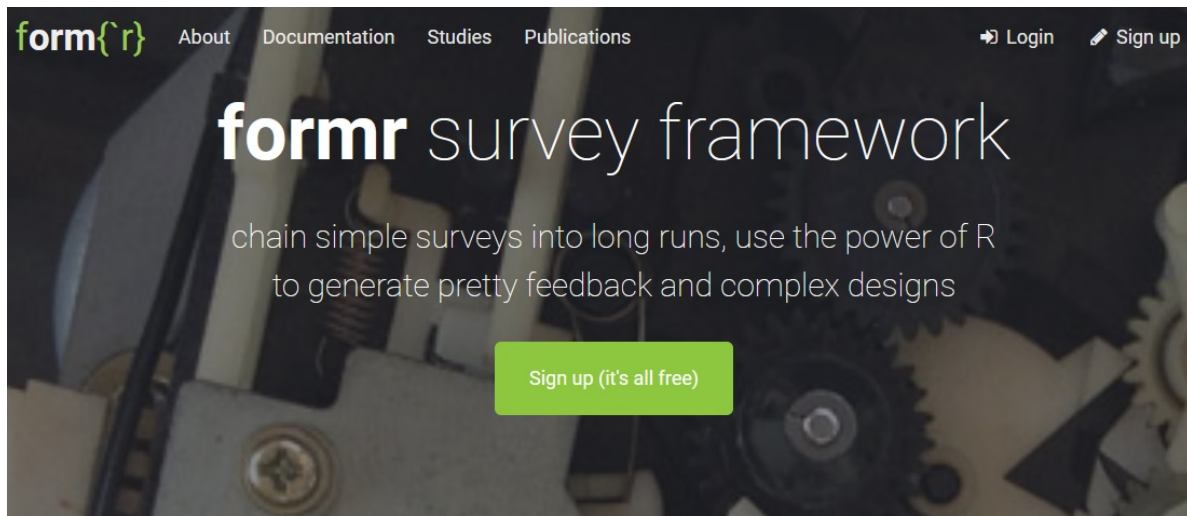
codebook package

Keine Lust auf R? Es gibt auch ein **Web-Interface**: codebook.formr.org



codebook package

Automatismus besonders gut mit **Datensätzen, die Metadaten beinhalten**.



+ {codebook} =

“mind blown”

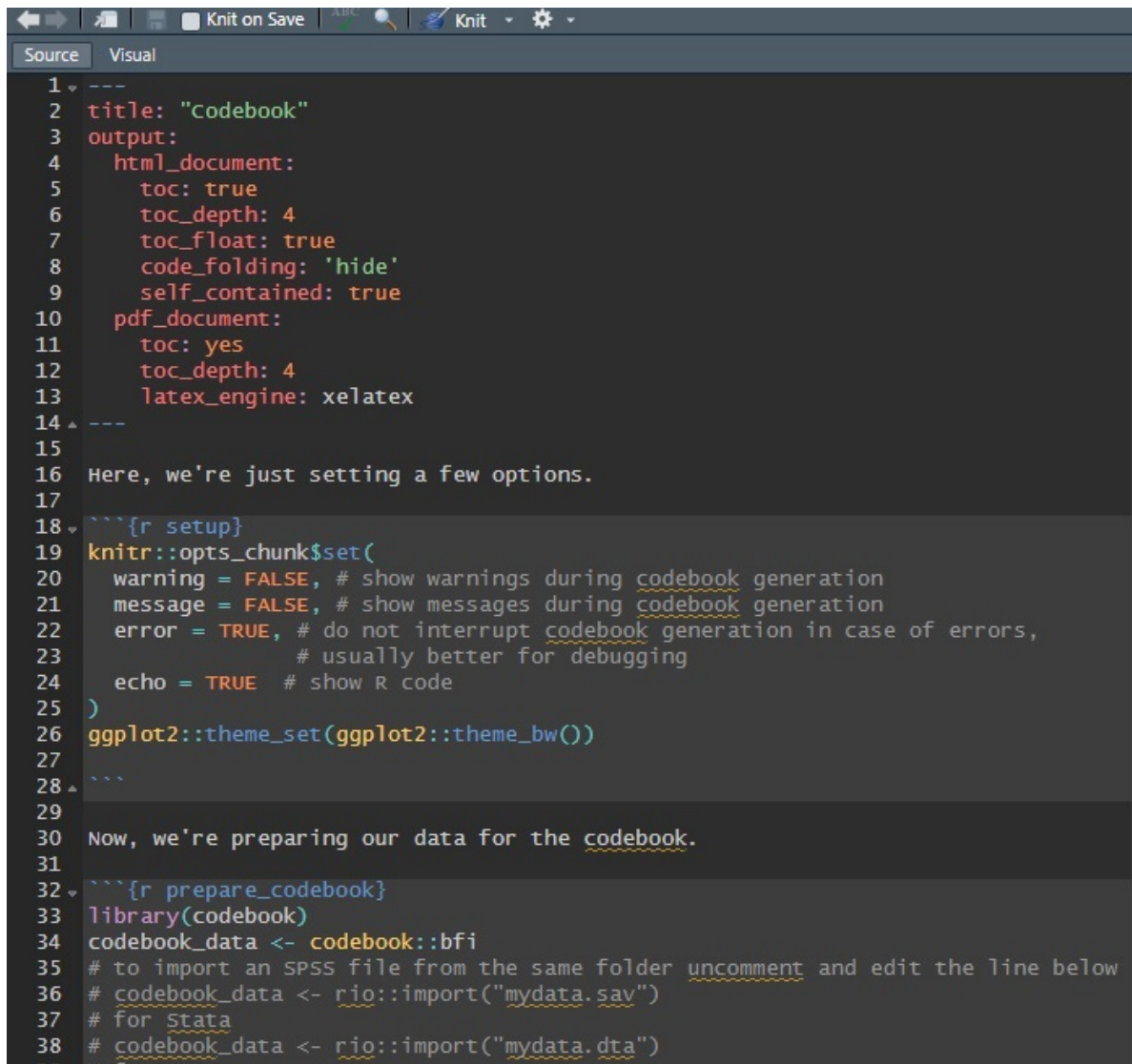
(Arslan, 2019)

(Arslan et al., 2020)

codebook package

R Markdown in aller Kürze

- R Markdown: R package und Dateiformat, um Dokumente aus R heraus zu generieren
- Ist die Basis für codebook-Dateien
- Hauptsächlich durch 3 Elemente aufgebaut

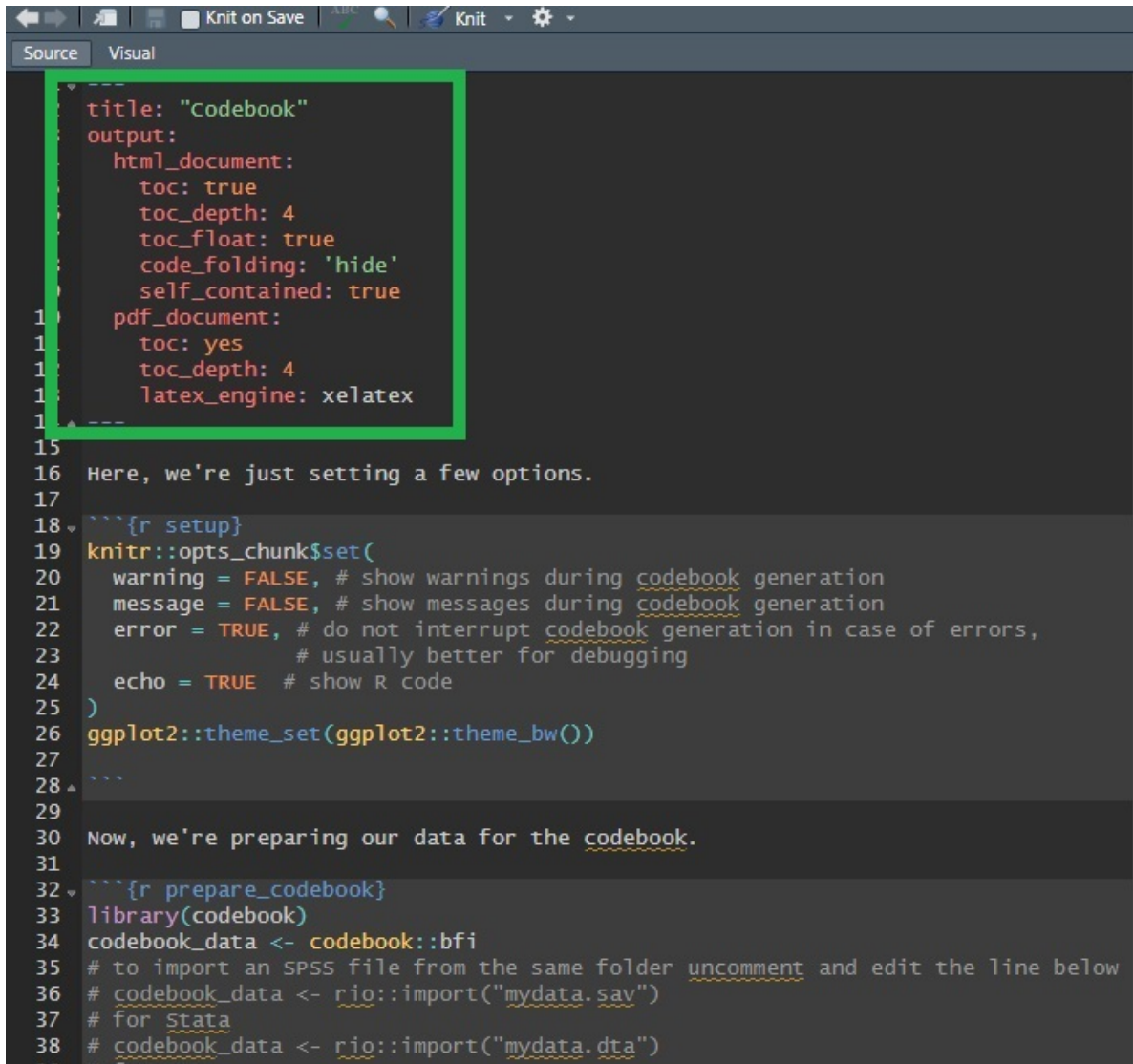


```
1 ---
2 title: "Codebook"
3 output:
4   html_document:
5     toc: true
6     toc_depth: 4
7     toc_float: true
8     code_folding: 'hide'
9     self_contained: true
10  pdf_document:
11    toc: yes
12    toc_depth: 4
13    latex_engine: xelatex
14 ---
15
16 Here, we're just setting a few options.
17
18 ```{r setup}
19 knitr::opts_chunk$set(
20   warning = FALSE, # show warnings during codebook generation
21   message = FALSE, # show messages during codebook generation
22   error = TRUE, # do not interrupt codebook generation in case of errors,
23               # usually better for debugging
24   echo = TRUE # show R code
25 )
26 ggplot2::theme_set(ggplot2::theme_bw())
27
28 ```
29
30 Now, we're preparing our data for the codebook.
31
32 ```{r prepare_codebook}
33 library(codebook)
34 codebook_data <- codebook::bfi
35 # to import an SPSS file from the same folder uncomment and edit the line below
36 # codebook_data <- rio::import("mydata.sav")
37 # for Stata
38 # codebook_data <- rio::import("mydata.dta")
39 # for CSV
```

codebook package

R Markdown in aller Kürze

- R Markdown: R package und Dateiformat, um Dokumente aus R heraus zu generieren
- Ist die Basis für codebook-Dateien
- Hauptsächlich durch 3 Elemente aufgebaut

A screenshot of an R Markdown source editor. The top toolbar includes icons for navigation and a 'Knit' button. Below the toolbar, there are two tabs: 'Source' and 'Visual'. The 'Source' tab is active, showing a dark-themed code editor. A green rectangular box highlights the YAML header section at the top of the document. The code is as follows:

```
title: "Codebook"
output:
  html_document:
    toc: true
    toc_depth: 4
    toc_float: true
    code_folding: 'hide'
    self_contained: true
  pdf_document:
    toc: yes
    toc_depth: 4
    latex_engine: xelatex
---
```

```
15
16 Here, we're just setting a few options.
17
18 ```{r setup}
19 knitr::opts_chunk$set(
20   warning = FALSE, # show warnings during codebook generation
21   message = FALSE, # show messages during codebook generation
22   error = TRUE, # do not interrupt codebook generation in case of errors,
23                 # usually better for debugging
24   echo = TRUE # show R code
25 )
26 ggplot2::theme_set(ggplot2::theme_bw())
27
28 ```
29
30 Now, we're preparing our data for the codebook.
31
32 ```{r prepare_codebook}
33 library(codebook)
34 codebook_data <- codebook::bfi
35 # to import an SPSS file from the same folder uncomment and edit the line below
36 # codebook_data <- rio::import("mydata.sav")
37 # for Stata
38 # codebook_data <- rio::import("mydata.dta")
39 # for CSV
```

YAML

Definition dokumentübergreifender Eigenschaften

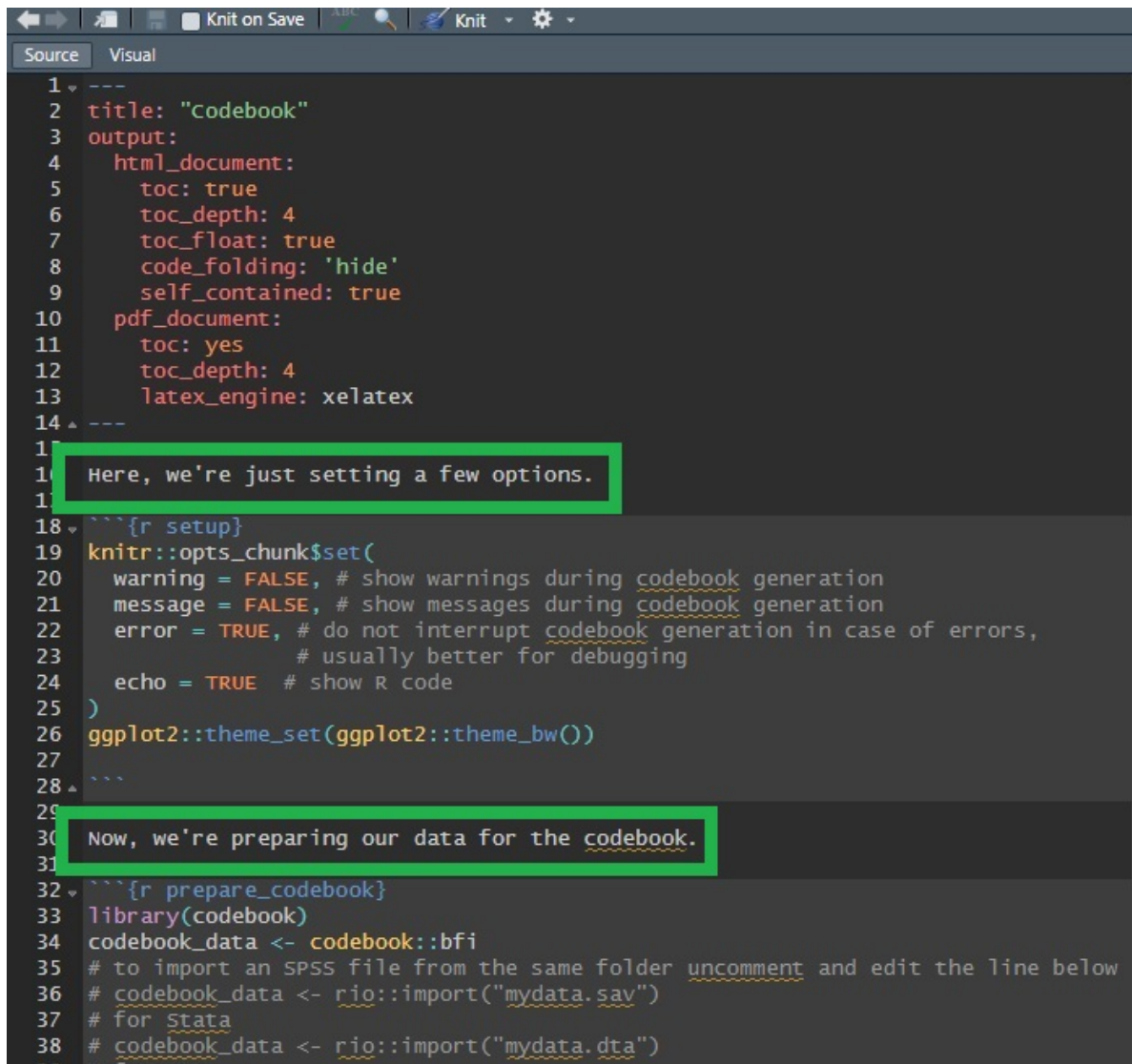
- Titel, Autor*in, ... des Dokuments

- Welches Outputformat?
- Welches Theme?
- Welcher Zitationsstil
- ...

codebook package

R Markdown in aller Kürze

- R Markdown: R package und Dateiformat, um Dokumente aus R heraus zu generieren
- Ist die Basis für codebook-Dateien
- Hauptsächlich durch 3 Elemente aufgebaut



```
1 ---
2 title: "Codebook"
3 output:
4   html_document:
5     toc: true
6     toc_depth: 4
7     toc_float: true
8     code_folding: 'hide'
9     self_contained: true
10  pdf_document:
11    toc: yes
12    toc_depth: 4
13    latex_engine: xelatex
14 ---
15
16 Here, we're just setting a few options.
17
18 ```{r setup}
19 knitr::opts_chunk$set(
20   warning = FALSE, # show warnings during codebook generation
21   message = FALSE, # show messages during codebook generation
22   error = TRUE, # do not interrupt codebook generation in case of errors,
23                 # usually better for debugging
24   echo = TRUE # show R code
25 )
26 ggplot2::theme_set(ggplot2::theme_bw())
27
28 ```
29
30 Now, we're preparing our data for the codebook.
31
32 ```{r prepare_codebook}
33 library(codebook)
34 codebook_data <- codebook::bfi
35 # to import an SPSS file from the same folder uncomment and edit the line below
36 # codebook_data <- rio::import("mydata.sav")
37 # for Stata
38 # codebook_data <- rio::import("mydata.dta")
39 # for CSV
```

Text/Bilder/Videos/...

Formatierbarer Text

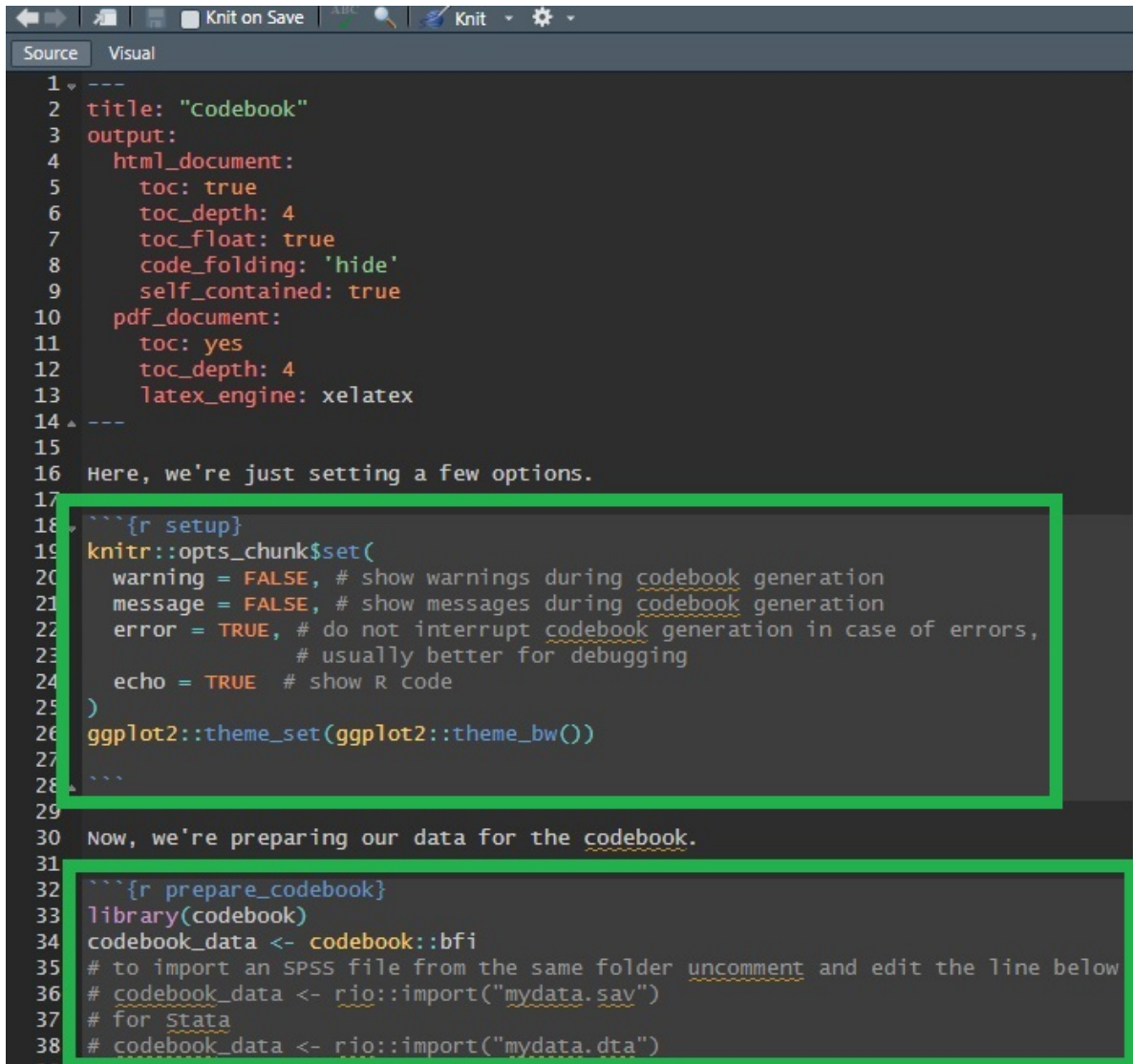
- Fett/kursiv
- Einbinden von Links
- Einbinden von Bildern/Videos
- ...

[R Markdown Cheat Sheet \(Link\)](#)

codebook package

R Markdown in aller Kürze

- R Markdown: R package und Dateiformat, um Dokumente aus R heraus zu generieren
- Ist die Basis für codebook-Dateien
- Hauptsächlich durch 3 Elemente aufgebaut



```
1 ---
2 title: "Codebook"
3 output:
4   html_document:
5     toc: true
6     toc_depth: 4
7     toc_float: true
8     code_folding: 'hide'
9     self_contained: true
10  pdf_document:
11    toc: yes
12    toc_depth: 4
13    latex_engine: xelatex
14 ---
15
16 Here, we're just setting a few options.
17
18 ```{r setup}
19 knitr::opts_chunk$set(
20   warning = FALSE, # show warnings during codebook generation
21   message = FALSE, # show messages during codebook generation
22   error = TRUE, # do not interrupt codebook generation in case of errors,
23               # usually better for debugging
24   echo = TRUE # show R code
25 )
26 ggplot2::theme_set(ggplot2::theme_bw())
27 ```
28
29
30 Now, we're preparing our data for the codebook.
31
32 ```{r prepare_codebook}
33 library(codebook)
34 codebook_data <- codebook::bfi
35 # to import an SPSS file from the same folder uncomment and edit the line below
36 # codebook_data <- rio::import("mydata.sav")
37 # for Stata
38 # codebook_data <- rio::import("mydata.dta")
39 ```
```

R Code

R Code in Chunks oder “in line”.

- Ermöglicht Input von Code

- Führt Code aus und zeigt Output an
- Zahlreiche Einstellungsmöglichkeiten, z. B. ob Output angezeigt werden soll
- Code Chunks durch ```{r}``` (“backticks”) begonnen und ````` beendet
- Inline Code durch ``r 2+2`` ergibt: 4
- ...

Hands on

Wenn fertig mit Aufgabe in Kleingruppen: pollev.com/js123

Arslan, R. C. (2019). How to Automatically Document Data With the codebook Package to Facilitate Data Reuse. *Advances in Methods and Practices in Psychological Science*, 2(2), 169–187. <https://doi.org/10.1177/2515245919838783>

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). Formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods*, 52(1), 376–387. <https://doi.org/10.3758/s13428-019-01236-y>

Artner, R., Verliefe, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, 26(5), 527–546. <https://doi.org/10.1037/met0000365>

Barba, L. A. (2018). *Terminologies for Reproducible Research* (No. arXiv:1802.03311). arXiv. <https://arxiv.org/abs/1802.03311>

BMBF. (2021). Bekanntmachung. Richtlinie zur Förderung von Projekten zum Thema Nachnutzung und Management von Forschungsdaten an Fachhochschulen. In *Bundesministerium für Bildung und Forschung - BMBF*. <https://www.bmbf.de/bmbf/shareddocs/bekanntmachungen/de/08-17-Bekanntmachung-Fachhochschulen.html>.

Commission, E. (2017). *H2020 Programme. Annotated Model Grant Agreement*.

DFG. (2022). Konkretisierung der Anforderungen zum Umgang mit Forschungsdaten in Förderanträgen. *Information für die Wissenschaft*, 25.

DFG. (2015). *Leitlinien zum Umgang mit Forschungsdaten*.

DGfE. (2017). *Stellungnahme der DGfE zur Archivierung, Bereitstellung und Nachnutzung qualitativer Forschungsdaten in der Erziehungswissenschaft*.

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *eLife*, 10, e67995. <https://doi.org/10.7554/eLife.67995>

Kraft, A. (2017). Die FAIR Data Prinzipien für Forschungsdaten. In *TIB-Blog*.

NAS (Ed.). (2019). *Reproducibility and replicability in science*. National Academies Press.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. <https://doi.org/10.1038/sdata.2016.18>

Credit

Foto title page by Max Harlynking on Unsplash

Comic “Publications and Data” by Auke Herrema, Ms. Gerry, CC-BY

Icons by Font Awesome CC-BY 4.0