DATA LITERACY ESSENTIALS

HWS 2021/22

Daten organisieren

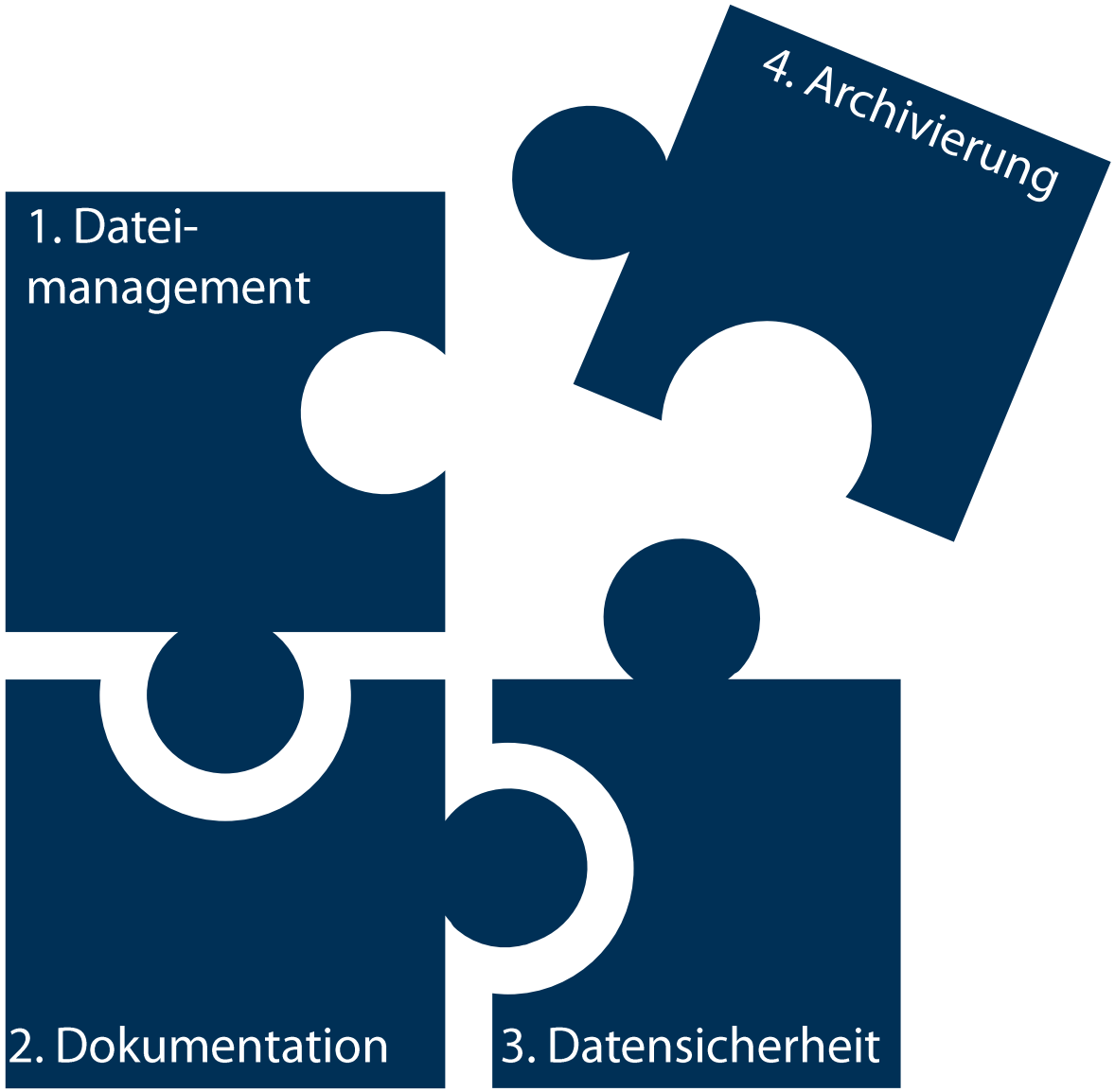
Lorena Steeb, Irene Schumm

(FDZ der UB Mannheim)

21.10.2021

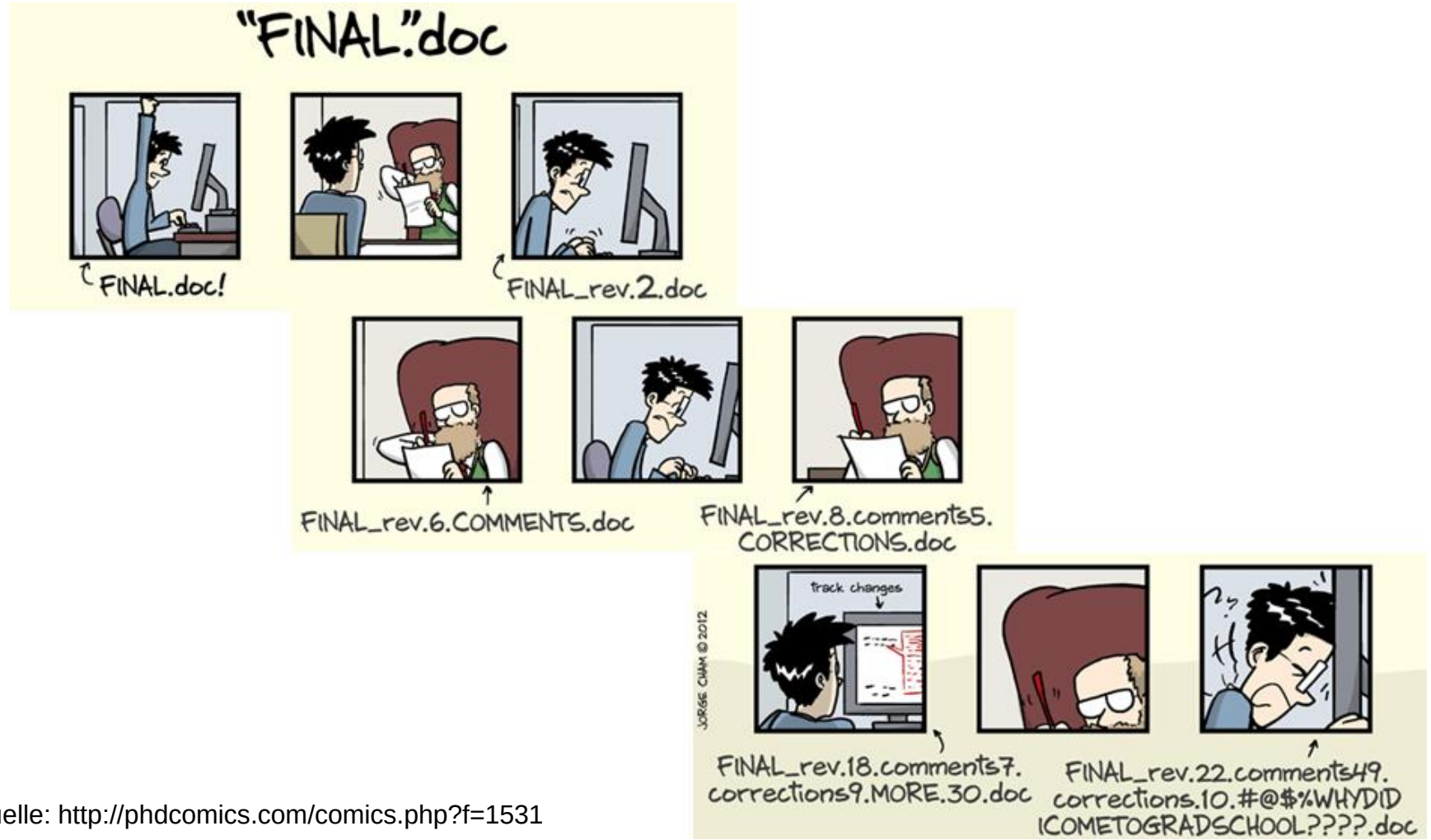
Agenda

HWS 2021/22



Dateimanagement

HWS 2021/22



Quelle: http://phdcomics.com/comics.php?f=1531

Dateimanagement – Bad Practices

HWS 2021/22

- survey01.csv

  → Kontext (z.B. identifizierende Projektinformation)

- Im Ordner sind folgende Dateien:
  - Meier_thesis_customer_survey_01.csv
  - Thesis_Meier_customer_survey.csv
  - Meier_thesis_customer_survey01_00.csv

  → Konsistenz (z.B. Dateinamenaufbau, Nummerierung → Sortierung)

- Befragung-aller-Absolventen-2020-Baden-Württemberg-Hochschulen-Thesis-Meier@2021-10-18-v10b.csv

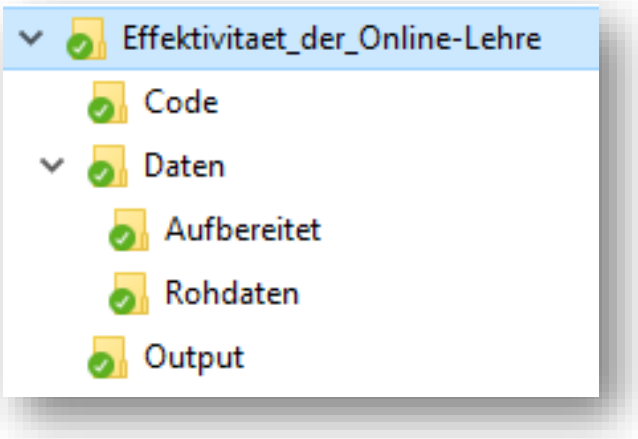
  → Kompatibilität (z.B. Umlaute, Sonderzeichen, Länge Dateinamen)

Dateimanagement – Good Practices

HWS 2021/22

- Datei- und Ordnerbenennung → Konventionen festlegen
  - Kontext: Übergeordnete, identifizierende Informationen (z.B. Titel/Akronym des Projekts oder der Arbeit)
  - Konsistenz → Sortierung: Datum (YYYYMMDD) und/oder Version (v01, v02 – führende Nullen)
  - Kompatibilität: Nicht zu lang (<25 Zeichen) & Keine Sonderzeichen

- Dateiorganisation
  - Für jedes Projekt einen Ordner
  - Unterordner für Daten, Code, Ergebnisse
  - Im Daten-Ordner Rohdaten von aufbereiteten/abgeleiteten Daten trennen
  - Niemals mit der Rohdaten-"Masterdatei" arbeiten – immer nur mit Kopien, Masterdatei am besten schreibgeschützt



→ Mehr Infos bspw. OSF Guides: Best Practices – File Management and Licensing oder Guide der Stanford Univ. Libraries

Dateimanagement – Versionierung&Backup

HWS 2021/22

- Software-Empfehlung für Studierende der Universität Mannheim: Zugang zum Cloud-Dienst Microsoft 365 mit
  - Cloud-Speicher OneDrive (1 TB) → mit Backup-Möglichkeit, Versionstracking, File sharing
  - Office
  - OneNote
  - Teams

→ Für die Arbeit mit personenbezogenen/vertraulichen Daten s. „Datensicherheit“ auf späterer Folie

- Versionierung für umfangreicheren Code: GitHub

Dokumentation

HWS 2021/22

Projekt-/Studienebene (z.B. ReadMe)

Dokumentation: Projektebene - ReadMe

HWS 2021/22

- Gibt Überblick über das Projekt und Material (Daten- und Code-Dateien, …), ermöglicht Verständnis des Gesamtzusammenhangs und erleichtert Reproduzieren der Ergebnisse
- Idealerweise einfaches .txt-Dokument, abgelegt im Projekt-Ordner
- Enthält u.a. folgende Informationen:
  - Projekttitel, Verantwortliche Personen, Projektbeschreibung, ggf. Rechte
  - Informationen über Daten, u.a.
    - Welche Dateien sind vorhanden und was enthalten sie?
    - Wie wurden Daten erhoben und aufbereitet?
  - Informationen über Code u.a.
    - Welche Dateien sind vorhanden und was tun diese? (Aufbereitung, Analyse, Visualisierung…)
    - Welche Software wurde verwendet (Version+Zusatzpakete+Betriebssystem)

→ Umfangreiches ReadMe-Template für die Wirtschafts- und Sozialwissenschaften von Vilhuber et al.

Dokumentation: Projektebene - ReadMe

HWS 2021/22

*ReadMe.txt - Editor

Datei Bearbeiten Format Ansicht Hilfe

```
ReadMe-Datei zum Projekt „Effektivität der Online-Lehre an der Universität Mannheim – Einflussfaktoren auf die Abschlussnote im HWS 2021/22“

(Haupt-)Verantwortliche: Lieschen Müller, lieschen.mueller@stud.uni-mannheim.de

Abstract: Im Rahmen des Projekts wurden die Einflussfaktoren auf die Effektivität der Online-Lehre an der Universität Mannheim untersucht. Hierzu wurden Daten in einer eigenen Befragung unter allen Absolvent*innen des Herbst-/Wintersemesters 2021/22 durchgeführt und die Daten mittels Regressionsanalyse analysiert.

1. DATEN (Ordner Daten)

Datenquelle: Eigene, anonymisierte Erhebung (Befragung unter allen 1.000 Absolvent*innen des HWS 2021/22)
Umfragezeitraum: 01.03.2022-21.03.2022
Vollerhebung (Anschreiben aller Absolvent*innen durch das Studienbüro via E-Mail)

1.1 Unterordner "Rohdaten"
- Schreibgeschützte Datei mit den Rohdaten der Befragung im CSV-Format (survey_effektivitaet-online-lehre_roh_2022-03-21.csv)
- Fragebogen im PDF/A-Format (survey-questions_effektivitaet-online-lehre_2022-03-21.pdf)
- Data Dictionary für den Rohdatensatz im CSV-Format (data-dictionary-rohdaten_effektivitaet-online-lehre_2022-03-21.csv)

1.2 Unterordner "Aufbereitet"
- Datei mit den aufbereiteten Daten der Befragung (survey_effektivitaet-online-lehre_2022-05-10.csv)
- Data Dictionary für den aufbereiteten Datensatz im CSV-Format (data-dictionary-aufbereitet_effektivitaet-online-lehre_2022-03-21.csv)

2. CODE (Ordner Code)

- R-Skript zur Datenaufbereitung (preprocessing_effektivitaet-online-lehre_2022-05-10.r)
  - konsistente Codierung von Missings und "Nicht zutreffend" als NA
  - konsistentes Datumsformat YYYYMMDD

- R-Skript zur Datenanalyse (regression_effektivitaet-online-lehre-2022-05-10.r)
```

Zeile 5, Spalte 268 | 100% | Windows (CRLF) | ANSI

Dokumentation

HWS 2021/22

Projekt-/Studienebene (z.B. ReadMe)

Datenebene
(z.B. codebooks, data dictionaries)

Dokumentation: Daten-Ebene

HWS 2021/22

ID	age	Q1	Q2	…
1	18	2	-	…
2	20	-999	1,4	
3	19	13	2,4	
4	19	xxx	3,6	
5	199	5	2.3	
6	20	20		
7	19	12	2,1	

Dokumentation: Daten-Ebene

HWS 2021/22

- Codebook oder Data Dictionary: dokumentiert die Bedeutung der Variablennamen und Werte im Datensatz
- Idealerweise einfaches .txt- oder .csv-Dokument, abgelegt im Daten-Ordner

Variable	Variablenname	Variablenbeschreibung	Datentyp	Wertebereich	Code für Missings	Frage
ID Teilnehmende/-r	ID	Identifizierende Nummer für jede/-n Teilnehmer/-in	numerisch	00001-99999	NA	Wird automatisch zugewiesen
Alter des/-r Teilnehmende/-n in Jahren	age	Alter des/-r Teilnehmende/-n zum Zeitpunkt der Erhebung	numerisch	10-100	NA	Wie alt sind Sie?
Anzahl Online-Kurse	Q1	Anzahl der belegten und erfolgreich abgelegten Online-Kurse	numerisch	0-50	NA	Wie viele Online-Kurse haben Sie insgesamt belegt und erfolgreich abgeschlossen?
Abschlussnote	Q2	Abschlussnote Studiengang	numerisch	1.0-6.0	NA	Mit welcher Gesamtnote haben Sie Ihren Studiengang abgeschlossen?

Dokumentation

HWS 2021/22

Projekt-/Studienebene (z.B. ReadMe)

Datenebene
(z.B. codebooks, data dictionaries)

Code-Ebene

Dokumentation: Code-Ebene

HWS 2021/22

- Im Code zu Beginn verwendetes Betriebssystem, Software-Version, Zusatzpakete etc. vermerken
- Kommentare im Code verwenden, um Kontext hinzuzufügen oder Entscheidungen zu begründen
- Code sollte möglichst „selbst-dokumentierend“ und „fool proof“ sein, bspw. durch
  - sprechende Variablen- und Funktionsnamen
  - strukturierten Code (z.B. auch aufsplitten in kleinere Skripte)
  - relative Dateipfade für verwendete Daten/Skripte

- Weitere Tipps: https://mitcommlab.mit.edu/broad/commkit/coding-and-comment-style/

Dokumentation

HWS 2021/22

Projekt-/Studienebene (z.B. ReadMe)

Datenebene
(z.B. codebooks, data dictionaries)

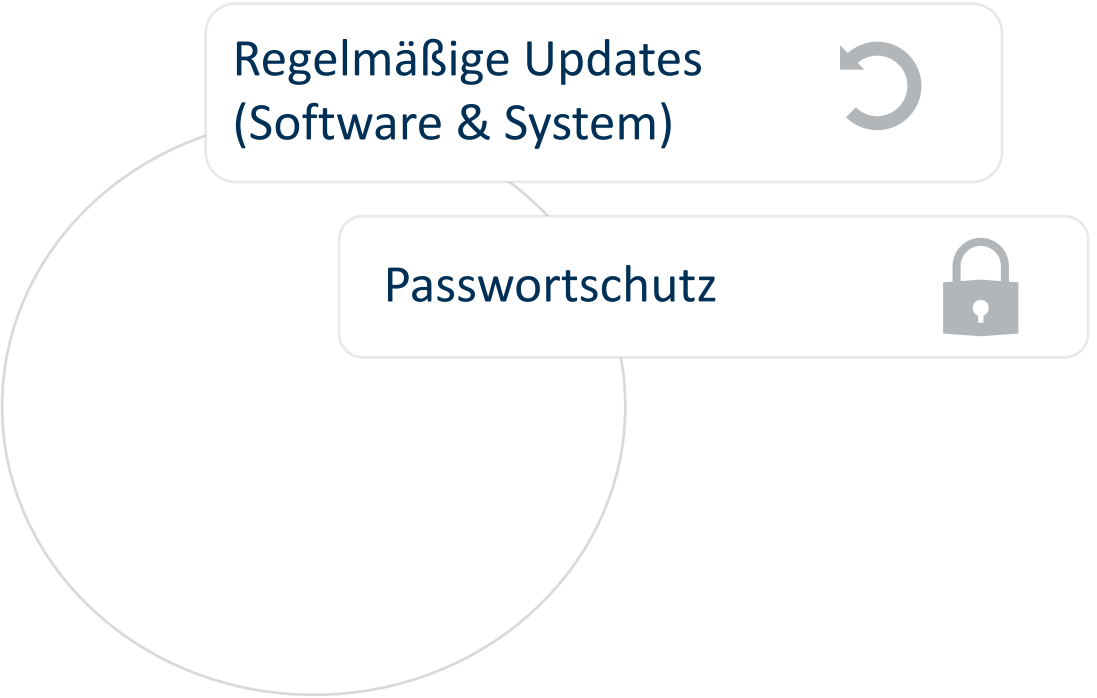
Code-Ebene

→ Dokumentation auf Stand halten!

Datensicherheit

HWS 2021/22

- Unbefugten Zugriff auf Daten verhindern durch:



Datensicherheit
Exkurs: Passwortschutz

HWS 2021/22

Auf einen Blick

Ein sicheres Passwort ...

- ist mindestens 10 Zeichen lang
- enthält Groß-, Kleinbuchstaben, Sonderzeichen & Zahlen
- ist für den persönlichen Gebrauch und darf nicht mit Dritten geteilt werden
- ist für jedes verwendete Benutzerkonto einzigartig

Tipp: Je länger Ihr Passwort ist, desto sicherer ist es!

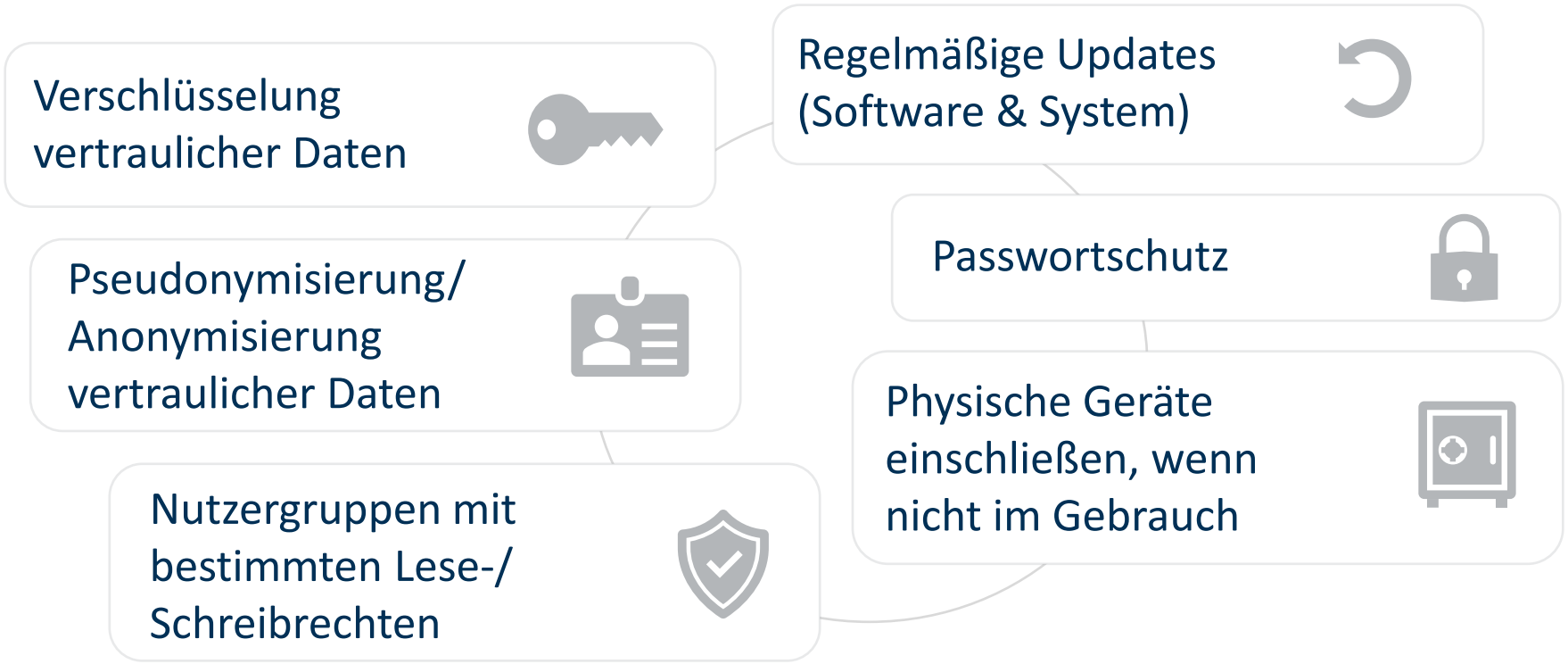
Siehe Informationen der UNIT

Tipp: Passwort Manager verwenden, wie z. B: KeePass

Datensicherheit

HWS 2021/22

- Unbefugten Zugriff auf Daten verhindern durch:



Datensicherheit
Exkurs: Verschlüsselung

HWS 2021/22

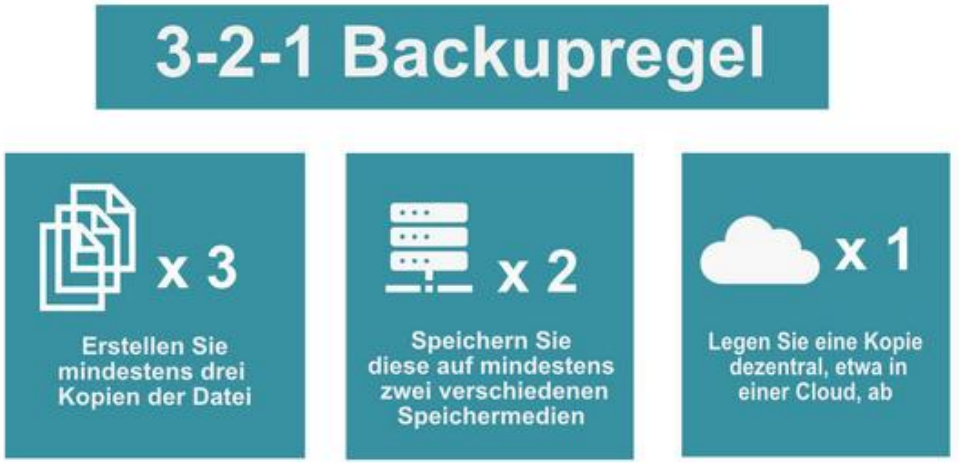
- Computer: Festplatte über Bitlocker
- Externe Speichermedien: Gesamtes Speichermedium mit VeraCrypt; erweiterbarer Ordner mit Cryptomator; einzelne Dateien und Ordner mit verschlüsselten ZIP-Archiven
- E-Mails: Gesamte E-Mail-Kommunikation mit S/MIME; Anhänge mit einem verschlüsselten ZIP-Archiv
- Cloud-Dienste (z. B. OneDrive, Teams & OneNote): Anlegen erweiterbarer verschlüsselter Ordner mit Cryptomator; einzelne Dateien mit verschlüsselten ZIP-Archiven

Siehe Informationen der UNIT

Datensicherheit

HWS 2021/22

- Datenverlust verhindern
  a) Menschliche Fehler: versehentliches verschieben/löschen/überspeichern
  b) Technische Fehler: Hard-/Softwarefehler



3-2-1 Backupregel - I. Lang/Bearbeitung E. Böker / CC BY 4.0

- Das 1x1 der Informationssicherheit der UNIT Mannheim

Archivierung - Und nach dem Projekt?

HWS 2021/22

- Leitlinien zur Sicherung guter wissenschaftlicher Praxis (DFG): Forschungsdaten sollten i.d.R. für einen Zeitraum von **zehn Jahren** zugänglich und nachvollziehbar aufbewahrt werden

- Warum archivieren?
  - Zur Sicherung und Dokumentation
  - Ggf. zur Nachnutzbarkeit durch andere WissenschaftlerInnen

- Welche Daten?
  - Kriterien: Datenqualität, Nachnutzungspotential, Rechtliche Rahmenbedin Einzigartigkeit, Reproduzierbarkeit

- Wo archivieren?
  - Institutionelles/fachliches/fächerübergreifendes Repositorium
  - Forschungsdatenzentrum
  - Archiv

Welcome to MADATA

Welcome to the Research Data Repository of the University of Mannheim.

This service invites all researchers and faculty of the University of Mannheim to submit their research data and to make it accessible through the internet for reference and further investigation.

It is the aim of MADATA to contribute to the quality to academic research by making research data accessible and to provide the basis for transparency and reproducibility of academic research and to satisfy expectations of the academic community, including funding bodies.

Learn more about this repository.

MADATA is indexed in re3data, the registry of research data repositories and evaluated according to following badge:

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
http://doi.org/10.17616/R36W6W
MADATA - Mannheim research data repository

Latest Entries

- A Sociocultural Norm Perspective on Big Five Prediction
  Eck, Jennifer (2021) A Sociocultural Norm Perspective on Big Five Prediction. [Dataset]
- Kategorisierte Lernerkorpusdaten (Dissertation Tassja Weber 2020)
  Weber, Tassja (2019) Kategorisierte Lernerkorpusdaten (Dissertation Tassja Weber 2020). [Dataset]
- Data from the paper: Shared Decision Making during the COVID-19 Pandemic
  Köther, Anja K. and Siebenhaar, Katharina U. and Alpers, Georg W. (2021) Data from the paper: Shared Decision Making during the COVID-19 Pandemic. [Dataset]
- Datensatz zu den Mannheimer Fremdenlisten (1791, 1792, 1807–1818)
  Pister, Sarah (2020) Datensatz zu den Mannheimer Fremdenlisten (1791, 1792, 1807–1818). [Dataset]

https://madata.bib.uni-mannheim.de/

Archivierung - Und nach dem Projekt?

HWS 2021/22

- Empfehlung: Verbreitete Dateiformate verwenden, die offenen Standards folgen und nicht proprietär sind

- Geeignete Dateiformate für die Archivierung (im Auszug):
  - Textformate: PDF/A ***.pdf, *.txt, *.xml,** *.docx, *.odt
  - Spreadsheets: ***.csv,** *.xlsx, *.ods
  - Statistische Umgebungen: ***.R, *.csv,** *.por, *.sas, *.dta
  - Multimedia: ***.wav,** *.mp4, *.mp3
  - Bild (Rastergrafik): ***.tif,** *.png

Quellen

HWS 2021/22

- Forschungsdaten.info. Datenvalidierung: Daten für die Archivierung auswählen und bewerten. https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/datenvalidierung/
- Forschungsdaten.info. Datensicherheit und Backup: Wie Sie Datenverlust entgegen wirken. https://www.forschungsdaten.info/themen/speichern-und-rechnen/datensicherheit-und-backup/
- Forschungsdaten.info. Formate erhalten: Inhalte langfristig sichern. https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/formate-erhalten/
- OSF Guides: Best Practices – File Management and Licensing: https://help.osf.io/hc/en-us/sections/360003624133-File-Management-and-Licensing
- VENKATARAMAN, Shanmugasundaram, & Moura, Paula. (2020). Raw data, backup and versioning: What you need to know to preserve your research data. Zenodo. https://doi.org/10.5281/zenodo.4041557
- VerbundFDB. Datensicherheit und Datensicherung. https://www.forschungsdaten-bildung.de/daten-sichern
- UNIT: Verschlüsselung. https://www.uni-mannheim.de/informationssicherheit/sicherheitstipps/verschluesselung/#c186986
- UNIT: Passwortsicherheit. https://www.uni-mannheim.de/informationssicherheit/sicherheitstipps/passwortsicherheit/

This course was partially developed and held by the BERD – Business, Economic and Related Data initiative.

BERD@BW

funded by

Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

digital@bw

BERD@NFDI funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 460037581