

Daten organisieren für bessere Datenqualität

Dr. Melanie Röthlisberger, Open Science Services

Wie organisiere ich Daten und vermeide Fehler?

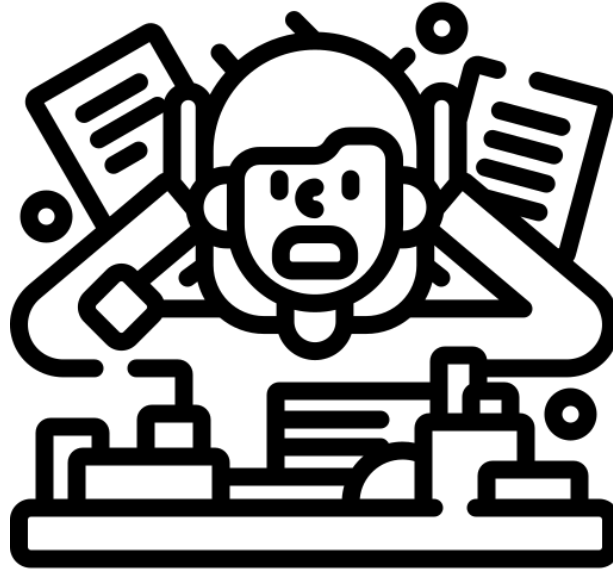


Wie und wo können Fehler entstehen?

Inkonsistente Formatierung
oder Verwendung von
unterschiedlichen Formaten



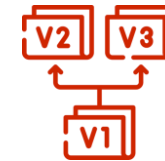
Uneinheitliche
Verwendung von
Variablen und Begriffen



Verwendung von
unterschiedlichen
Tools/Messinstrumenten



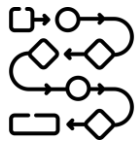
Uneinheitliche
Abläufe bei der
Datensammlung und
Datenorganisation



Verwendung von
unterschiedlichen
Versionen

Wie kann Dateninkonsistenz vermieden werden? (1)

Prozess der Datenorganisation



- Einen klaren Prozess zur Organisation der Daten vorab formulieren.
- Festlegen der Tools, Variablen, Masseinheiten, Zeitpunkte und Orte der Messungen, Methoden zur Datenverarbeitung etc.

Arbeiten im Team



- Regelmässige Treffen und klare Kommunikationskanäle helfen, vorzeitig Inkonsistenz aufzudecken und Missverständnisse und Unsicherheiten zu klären.

Benennung der Variablen



- Festlegen, was unter den für den Datensatz zentralen Begriffen (z.B. Variablen) verstanden wird und diese Begriffe konsequent verwenden.

Wie kann Dateninkonsistenz vermieden werden? (2)

Dateibenennung



- kurze, aber aussagekräftige Namen verwenden
- Sonder- oder Leerzeichen vermeiden
- Dateinamen mit Datum (bspw. JJJJMMTT) beginnen, um die Dateien chronologisch zu ordnen.
- Versionskontrolle miteinbeziehen (bspw. “_v.1.0“, „_v.1.1“, “_v.2.0“ etc.)

Beispiel:

20250203_Seminararbeit_v.1.2

Beispiele für die Versionierung

Name der Datei

LC_Interviewschedule_1.0

LC_Interviewschedule_1.1

LC_Interviewschedule_1.2

LC_Interviewschedule_2.0

Änderungen an der Datei

Ursprüngliches Dokument

Kleine Änderungen vorgenommen

Weitere kleine Änderungen vorgenommen

Größere Änderungen vorgenommen

Wie kann Dateninkonsistenz vermieden werden? (2)

Dateibenennung



- kurze, aber aussagekräftige Namen verwenden
- Sonder- oder Leerzeichen vermeiden
- Dateinamen mit Datum (bspw. JJJJMMTT) beginnen, um die Dateien chronologisch zu ordnen.
- Versionskontrolle miteinbeziehen (bspw. “_v.1.0“, „_v.1.1“, “_v.2.0“ etc.)

Beispiel:

20250203_Seminararbeit_v.1.2

Ordnerstrukturen



- Dateien hierarchisch anlegen und Kategorien festlegen, um die Dateien unterteilen zu können (wie bspw. (Teil-)Projekte, Zeit, Ort, Dateityp).
- Ordnerstrukturen sollten nicht mehr als 3 bis 4 Ebenen tief sein.
- Rohdaten und verarbeitete Daten klar getrennt anlegen.

Beispiel für eine Ordnerstruktur

Example 1

```

└─ Project_name
  └─ README.txt*      # Basic project description
  └─ LICENSE*         # License information clarifying conditions for reuse
  └─ Analysis
    └─ pipeline_01
  └─ Data
    └─ Source/raw     # Data as-collected.
    └─ Derivatives    # Preprocesseed for analysis.
      └─ pipeline_01
  └─ Docs
    └─ Ethics          # Submitted to Ethics Committee
    └─ Procedures      # Standard Operating Procedures (SOPs), protocols, etc.
    └─ Publications    # Research reports
    └─ ⚠ Recruitment
  └─ ⇐ Scripts
```

⚠ Contains no identifying information

⇐ Synchronized with a Git version controlled code repository

* Additional README and LICENSE files can be found at subfolder level (e.g., in the Scripts folder)

Wie kann Dateninkonsistenz vermieden werden? (2)

Dateibenennung



- kurze, aber aussagekräftige Namen verwenden
- Sonder- oder Leerzeichen vermeiden
- Dateinamen mit Datum (bspw. JJJJMMTT) beginnen, um die Dateien chronologisch zu ordnen.
- Versionskontrolle miteinbeziehen (bspw. “_v.1.0“, „_v.1.1“, “_v.2.0“ etc.)

Beispiel:

20250203_Seminararbeit_v.1.2

Ordnerstrukturen



- Dateien hierarchisch anlegen und Kategorien festlegen, um die Dateien unterteilen zu können (wie bspw. (Teil-)Projekte, Zeit, Ort, Dateityp).
- Ordnerstrukturen sollten nicht mehr als 3 bis 4 Ebenen tief sein.
- Rohdaten und verarbeitete Daten klar getrennt anlegen.

Auswahl des Dateiformats



- Daten sollten in einem Format abgespeichert werden, das von verschiedenen Computerprogrammen gelesen werden kann.

Empfohlene Dateiformate für die längerfristige Nutzung

Art der Daten	Geeignet	Akzeptabel	Ungeeignet
Tabellarische Daten mit viel Metadaten	.csv / .tsv / .hdf5	.txt / .html / .tex / .por	
Tabellarische Daten mit einem Minimum an Metadaten	.csv / .tsv / .tab / .ods / SQL	.xml if appropriate DTD / .xlsx	.xls / .xlsb
Textdaten	.pdf / .txt / .odt / .odm / .tex / .md / .htm / .xml	.pptx / .pdf with embedded forms / .rtf	.doc / .ppt
Code	.m / .R / .py / .iypnb / .rstudio / .rmd / NetCDF	.sdd	.mat / .rdata
Bilddateien	.tif (uncompressed) / .png / .svg / .jpeg	.jpg / .jp2 / .tif (compressed) / .tiff / .pdf / .gif / .bmp	.indd / .ait / .psd
Audiodateien	.flac / .wav / .ogg	.mp3 / .mp4 / .aif	
Videodateien	.mp4 / .mj2 / .avi / .mkv	.ogm / .webm	.wmv / .mov
Räumliche (Geo) Daten	NetCDF, tabular GIS attribute data, .shp / .shx / .dbf / .prj / .sbx / .sbn / PostGIS / .tif / .tfw / GeoJSON	.mdb / .mif	
CAD / Vektor und Raster Daten	.x3d / .x3dv / .x3db / PDF3D .pdf	.dwg / .dxf	
Generische Daten	.xml / .json / .rdf		

Wie kann Dateninkonsistenz frühzeitig aufgedeckt werden?



Überprüfung der Datenerhebung durch eine unabhängige Person
(bspw. Überprüfung der Annotierungen bei qualitativen Textdaten
oder Überprüfung der verwendeten Variablen)

Automatische Überprüfung der Werte durch spezifische
sogenannte «ChecksumTools» wie bspw. MD5summer
(siehe Anleitung).



2

Abschluss



Fragen?

Kontakt

melanie.roethlisberger@ub.uzh.ch

Feedback Coffee Lecture
Frühjahrssemester 2026



<https://shorturl.at/Fbss6>

Weiterführendes

Weiterführende Informationen und Links

Forschungsdatenmanagement an der UZH: <https://www.openscience.uzh.ch/de/forschungsdatenmanagement.html>

Weitere Coffee Lectures

<https://veranstaltungen.ub.uzh.ch/de/page/coffee-lectures>



Alle Logos und Symbole von Organisationen sind urheberrechtlich geschützt. Sofern ansonsten nicht anders angegeben, ist dieser Foliensatz lizenziert mit einer [CC-BY-4.0 International](https://creativecommons.org/licenses/by/4.0/) Lizenz. Adaptiert von: Regula Zwicky. 2025. Teaching Tool «Datenmanagement im Studium: Datenorganisation verbessert Datenqualität». Universität Zürich. Folien 6, 8, und 10 wurden hinzugefügt.

Zitieren als: Melanie Röthlisberger & Regula Zwicky. 2026. Daten organisieren verbessert Datenqualität. Coffee Lectures FS26. Universitätsbibliothek, Universität Zürich.