1.2 Ordnerstrukturen und Datenbanken

Daten managen – 1 Datenstrukturierung

Stellen wir uns vor, unser Forschungsteam arbeitet weiter an dem Projekt: Alle Daten werden sorgfältig gespeichert und Dateien tragen zwar sprechende Namen. Jedoch liegen sie kreuz und quer in einem einzigen Ordner. Als ein neuer Kollege später bestimmte Zwischenergebnisse nachvollziehen soll, beginnt das Chaos: Zwar heißen die Dateien eindeutig, aber ohne logische Ordnung ist nicht ersichtlich, welche Versionen zusammengehören, welche inhaltliche Beziehung besteht und an welcher Stelle mit der Analyse angesetzt werden muss. Wertvolle Zeit geht verloren, und wichtige Ergebnisse drohen unterzugehen. Das Team beschließt: Die Ordnerstruktur muss überarbeitet werden.

Eine klare Ordnerstruktur ergänzt unsere gute Dateibenennung: Sie schafft Übersicht, macht Zusammenhänge erkennbar und ermöglicht, dass auch andere – oder man selbst nach längerer Zeit – die Daten schnell verstehen, wiederfinden und nachnutzen können.

Möglichkeiten der Dateistrukturierung

Es gibt neben der klassischen hierarchischen Ordnerstruktur auch die Möglichkeiten Dateien und Daten mit Tags zu organisieren. Im Folgenden wollen wir beides kurz mit Vor- und Nachteilen vorstellen.

Hierarchische Organisation

Organisation in Ordnern mit weiteren Unterordnern (gängige Speicherstruktur von Betriebssystemen)

Vorteile	Nachteile
• Vertraute Art der Organisation • Spiegelt Informationsstruktur gut wieder • Ähnliche Items werden gemeinsam abgespeichert	• Balance zwischen Breite und Tiefe bewahren • Items haben eine mögliche Position • Reorganisation ist schwierig (wenn notwendig)

Tag-basierte Organisation

Jedem Item werden ein oder mehrere Tags (Schlagworte) zugeordnet (z.B. für die Organisation von Literatur in Literaturverwaltungsprogrammen), Schlagwortsystem kann bei Betriebssystemen ergänzt werden (weitere Infos zum Tagging von Dateien hier)

Vorteile	Nachteile
• Items gehören in mehr als eine Kategorie • Kann schneller/einfacher aufgesetzt und erweitert werden • Fusion mit anderen Dateisystemen schneller möglich	• Betriebssysteme haben Schlagwortsystem (meist) nicht direkt implementiert • Item muss direkt mit richtigen/ausreichend Tags versehen werden um aufgefunden werden zu können • Erhöhtes Risiko von Inkonsistenz • Weniger gut zur Repräsentation von Informationsstrukturen geeignet

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt
Finanziert von der Europäischen Union
NextGenerationEU

Merkmale einer guten (hierarchischen) Ordnerstruktur

In einem Computersystem sind Daten meist in einer hierarchischen Anordnung, auch Verzeichnisstruktur genannt, abgelegt. Das Ziel einer guten Verzeichnisstruktur ist es, das Auffinden von Daten zu erleichtern. Die Struktur sollte dafür klar ersichtlich sein, damit sie auch für andere Personen verständlich ist.

Auch hier gibt es einige Regeln, an die man sich für eine gute und verständliche Struktur halten kann:

- Unnötig viele Unterordner in einem Verzeichnis sollten vermieden werden. Das könnte schnell dazu führen, dass man den Überblick verliert.
- Ordner sind nach Themen oder Kategorien geordnet für eine inhaltliche Strukturierung.
- Wenn Ordner viele Dateien enthalten, bietet sich meist eine weitere Strukturierungsebene an
- Ordnernamen sollten dabei auf die Ordnerinhalte durch deskriptive Benennung verweisen.
- Gleiche Bezeichnungen für Unterordner innerhalb eines Asts im Verzeichnisbaum sind nicht zu empfehlen, da sie Verwechslungen ermöglichen.
- Einer Überlappung der Ordnerthemen sollte möglichst aus dem Weg gegangen werden, um die Suche nach bestimmten Daten und Dateien nicht unnötig zu erschweren.
- Übergeordnete Ordner enthalten allgemeinere Elemente, in den Unterordnern werden die Inhalte spezifischer.
- Laufende und abgeschlossene Dokumente sind in unterschiedlichen Ordnern gespeichert.
- Es empfiehlt sich die Verwendung eines „Archiv“-Ordners für vergangene Dateiversionen.
- Rohdaten sind im besten Fall in einem separaten Ordner abgelegt (mehr dazu gleich).

Bestandteile von ReadMe-Dateien

Insbesondere beim kollaborativen Zusammenarbeiten von mehreren Personen, sollten Konventionen beim Umgang mit Dateien und Daten für alle schriftlich festgehalten werden. Es bietet sich eine ReadMe-Datei an, die alle Benennungskonventionen und Ablagestrukturen dokumentiert. Sie sollte in der obersten Ebene der Verzeichnisstruktur gespeichert werden.

Inhalte einer solchen Datei könnten sein:

- Dateibenennungskonventionen
- Ordnerstruktur
- Abkürzungen
- Namenskürzel der beteiligten Personen
- …

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt
Finanziert von der Europäischen Union
NextGenerationEU

Das Ganze könnte also wie folgt aussehen (die Datei ist auch als Beispiel und zur Orientierung abgelegt):

```
DATEIBENENNUNG
- Bericht:      [JJJJMMTT]_BRT_[Stichwort]_[Personenkürzel]_[Versionsnummer].[Dateiendung]
- Präsentation: [JJJJMMTT]_PRS_[Veranstaltung]_[Personenkürzel]_[Versionsnummer].[Dateiendung]
- Literatur:    Lit_[Erstautor]_[Veröffentlichungsjahr]_[Inhaltsstichpunkt].[Dateiendung]

ORDNERSTRUKTUR
- [Projekt]
|-- [Work_Content]
|   |--[Projektaufbau]
|   |--[Ethikantrag]
|   |--[Berichte]
|-- [Background_Info]
|   |--[Literatur]
|   |--[Methoden]
|-- [Templates]
|   |--[LogoBilder]
|   |--[Projektbeschreibung]
|-- [Presentations]
|   |--[JourFixe]
|   |--[Kongresse]

ABKÜRZUNGEN
- Bericht:      BRT
- Literatur:    Lit
- Präsentation: PRS

MITGLIEDERKÜRZEL
- Max Mustermann: MM
- Jean Dupont:    JD
- Fulano de Tal:  FT
```

Abb. 1. Beispiel für eine ReadMe-Datei

Rohdaten

Rohdaten, also unverarbeitete Originaldaten, sollten (wie oben bereits erwähnt) separat gespeichert werden. Durch die Aufbewahrung dieser Daten wird die Nachvollziehbarkeit der Daten gewährleistet. Dies wird auch in der Leitlinie 17 zur Sicherung guter wissenschaftlichen Praxis der DFG (Deutsche Forschungsgemeinschaft) festgehalten. Denn wenn Rohdaten weiterverarbeitet werden, beispielsweise durch Bereinigung, Auswertung oder Anonymisierung, gehen häufig Inhalte verloren. Auch das Komprimieren und Konvertieren von Daten kann zu Informationsverlust führen (siehe für mehr Infos **02 Datenmanipulation**).

Rohdaten können sehr unterschiedlich aussehen: Bei qualitativer Forschung wären es z.B. die Transkripte der Interviews (theoretisch wären die Rohdaten hier sogar die Audiodateien, aber aus Gründen des Datenschutzes (siehe für mehr Infos **04 Datenpflege und -schutz**) sollten diese nicht länger als nötig aufbewahrt werden). Für quantitative Studien sollten die Papierfragebögen (wenn die Studie analog durchgeführt wurde) oder der unbereinigte Datensatz aufbewahrt werden.

Rohdaten sollten nicht herstellerspezifisch (offen) und unkomprimiert gespeichert werden (siehe für mehr Infos **02 Datenmanipulation**).

- Rohdaten: Sind die Daten, die während des Forschungsprojekts generiert werden.
- Verarbeitete Daten: Sind Rohdaten, die so weiterverarbeitet sind, dass sie nützlich und manipulierbar ist (z.B. durch Bereinigung oder das Berechnen weiterer Variablen).
- Analysierte Daten: Zeigen, was die Daten aussagen, wie sie Zusammenhängen und ob sie signifikant sind.
- Finalisierte Daten: Sind so aufgearbeitete Daten, dass sie die Forschungsfrage unterstützen

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt
Finanziert von der Europäischen Union
NextGenerationEU

➔ Wichtig ist, dass diese Abfolge eine vereinfachte Darstellung ist und Daten sich nicht immer linear von Rohdaten zu finalisierten Daten entwickeln. Daten können ergänzt werden oder auch Datensätze mehrmals von verschiedenen Forschenden mit unterschiedlichen Forschungsmethoden ausgewertet werden.

Datenbanken

Wenn mit besonders vielen Daten und Dateien gearbeitet wird oder besondere Anforderungen an die Strukturierung der Daten bestehen, dann kann sich die Verwendung von Datenbanksystemen anbieten. Ziel ist es, große Datenmengen effizient, widerspruchsfrei und dauerhaft zu speichern und den Nutzenden eine bedarfsgerechte Darstellungsform zu bieten. Typischerweise werden relationale Datenbanken verwendet, die Beziehungen der Daten zueinander beschreiben.

Aber warum sollten Daten nicht einfach in einer Textdatei gespeichert werden? Alles ist an einem Ort, man hat die Möglichkeit sich die Inhalte so darzustellen wie man möchte und sogar einfach mit Zusatzinfos zu versehen.

Nehmen wir folgendes Beispiel von Behandlungsdaten:

Maja Mustermann, Blinddarmentzündung, Klinik am Nordring, Musterstraße 1, 01234 Musterhausen
Jos Bleau, Vorhofflimmern, Klinik am Nordring, Musterstraße 1, 01234 Musterhausen
Alexandra Ivanova, Entbindung, Klinik am Nordring, Musterstraße 1, 01234 Musterhausen
Juan Pérez, Pneumonie, Klinik am Nordring, Musterstraße 1, 01234 Musterhausen
Ola Nordmann, Fraktur des Unterarms, Klinik am Nordring, Musterstraße 1, 01234 Musterhausen
Yamada Hanako, Herzinsuffizienz, Klinik am Nordring, Musterstraße 1, 01234 Musterhausen

Könnten wir auf Anhieb folgende Fragen beantworten/Probleme lösen:

- Warum waren Maja und Ola im Krankenhaus?
- Sortiere die Daten alphabetisch nach den Nachnahmen der Patient:innen.
- Was passiert, wenn das Krankenhaus von „Klinik am Nordring“ in „Nordklinik am Ring“ umbenannt wird?

Was wäre, wenn wir nicht nur Informationen zu einer handvoll Personen hätten, sondern über hunderte oder sogar tausende?

Weitere Möglichkeit:

Kalkulationstabellen wie z.B. von Excel erlauben deutlich mehr Funktionalität und Interaktionsmöglichkeit wie Berechnungen, Sortierung, Filterung und grundlegende Analysen. Die Daten sind hier in nummerierte Zeilen und mit Buchstaben beschriftete Spalten organisiert. Die Darstellung ist, ähnlich wie in der Textdatei, leicht verständlich, aber solche Formate bieten neben den gegebenen Vorteilen auch einige Nachteile: Die Datensicherheit ist gering, es gibt keine Einstellungen für mehrere Benutzer und redundante und inkonsistente Daten sind möglich. Auch für die Arbeit mit sich verändernden Daten und Datennutzung über einen langen Zeitraum sind Kalkulationstabellen nicht geeignet.

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt
Finanziert von der Europäischen Union
NextGenerationEU

	A	B	C	D	E	F	G
1	Maja	Mustermann	Blinddarmentzündung	Klinik am Nordring	Musterstraße 1	1234	Musterhausen
2	Jos	Bleau	Vorhofflimmern	Klinik am Nordring	Musterstraße 1	1234	Musterhausen
3	Alexandra	Ivanova	Entbindung	Klinik am Nordring	Musterstraße 1	1234	Musterhausen
4	Juan	Pérez	Pneumonie	Klinik am Nordring	Musterstraße 1	1234	Musterhausen
5	Ola	Nordmann	Fraktur des Unterarms	Klinik am Nordring	Musterstraße 1	1234	Musterhausen
6	Yamada	Hanako	Herzinsuffizienz	Klinik am Nordring	Musterstraße 1	1234	Musterhausen
7							

Abb. 2. Beispiel von Daten in einer Kalkulationstabelle

Also brauchen wir eine bessere Lösung:

Eine Datenbank besteht aus mehreren, miteinander verknüpften Daten, die in einer logischen Beziehung stehen. Sie werden in Tabellen organisiert, in denen Daten in Zeilen und Spalten sortiert sind, und die Spalten jeweils Daten eines bestimmten, benannten Typs speichern. Tabellen werden häufig als „Entitäten“ modelliert, deren Spalten Attribute definieren. Mehrere Tabellen haben hierbei Beziehungen zueinander, die definiert werden müssen.

Key	PersonVorname	PersonNachname	Aufenthaltsgrund	KrankenhausName	KrankenhausAdresse	KrankenhausPLZ	KrankenhausStadt
1	Maja	Mustermann	Blinddarmentzündung	Klinik am Nordring	Musterstraße 1	01234	Musterhausen
2	Jos	Bleau	Vorhofflimmern	Klinik am Nordring	Musterstraße 1	01234	Musterhausen
3	Alexandra	Ivanova	Entbindung	Klinik am Nordring	Musterstraße 1	01234	Musterhausen
4	Juan	Pérez	Pneumonie	Klinik am Nordring	Musterstraße 1	01234	Musterhausen
5	Ola	Nordmann	Fraktur des Unterarms	Klinik am Nordring	Musterstraße 1	01234	Musterhausen
6	Yamada	Hanako	Herzinsuffizienz	Klinik am Nordring	Musterstraße 1	01234	Musterhausen

Bisher sieht das Ganze noch recht ähnlich zu der Kalkulationstabelle aus. Die Angabe der Krankenhausdaten ist hier redundant und könnte in eine weitere Tabelle ausgelagert werden. Dies nennt sich Datenbanknormalisierung. Dieser Prozess sorgt dafür, dass Redundanzen reduziert und die Wahrscheinlichkeit für Inkonsistenzen und Aktualisierungsfehler minimiert werden. Eine normalisierte Datenbank könnte also wie folgt aussehen:

Personentabelle:

PersonKey	PersonVorname	PersonNachname	Aufenthaltsgrund	KrankenhausKey
1	Maja	Mustermann	Blinddarmentzündung	1
2	Jos	Bleau	Vorhofflimmern	1
3	Alexandra	Ivanova	Entbindung	1
4	Juan	Pérez	Pneumonie	1
5	Ola	Nordmann	Fraktur des Unterarms	1
6	Yamada	Hanako	Herzinsuffizienz	1

Krankenhaustabelle:

KrankenhausKey	Name	Adresse	PLZ	Stadt

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt
Finanziert von der Europäischen Union
NextGenerationEU

1	Klinik am Nordring	Musterstraße 1	01234	Musterhausen

Datenbanksysteme bestehen aus einem Datenbankverwaltungssystem oder Datenbankmanagementsystem und einer oder mehreren Datenbanken. Das Datenbankverwaltungssystem ist eine Software, die Nutzenden einen effizienten und gebündelten Zugang zu den Daten bietet. Gängige Beispiele hierfür sind z.B. Oracle, MySQL, Microsoft Access und SAP HANA. Es ermöglicht die Erstellung von Tabellen, Ergänzung von Beziehungen und kontrolliert den Zugang.

Gute Datenbanksysteme bieten folgende Vorteile:

- Auswertung der Dateien nach beliebigen Merkmalen
- Einfache Abfragemöglichkeit und Auswertung, schnelle Bereitstellung der Daten
- Zuweisung verschiedener Nutzungsrechte an die einzelnen Benutzer:innen
- Daten und Anwenderprogramme sind unabhängig voneinander, sodass die Anwender:innen nur logische Datenstrukturen kennen müssen und das Datenbanksystem die organisatorische Verwaltung übernimmt
- Keine Datendopplung und Datenintegrität
- Datensicherheit bei Hardwareausfällen und Fehlern der Anwenderprogramme
- Datenschutz gegen unbefugten Zugriff
- Flexibilität hinsichtlich neuer Anforderungen
- Zulassung von Mehrbenutzerzugriffen
- Einhaltung einheitlicher Standards

Zur Interaktion mit Datenbanken wird meist die Datenbanksprache SQL (Structured Query Language) verwendet. Weitere umfangreiche Informationen zur Nutzung von SQL hier (auf Englisch)

Für einen detaillierteren Einblick, was für Datenbanken es gibt (hierarchisch, Netzwerk-, relational, objektorientiert, objektrelational, dokumentorientiert, Graph-) und wie diese aufgebaut sein können empfehlen wir folgendes Video.

Quellenverzeichnis

Blümm, M., Fritsch, K., Bock, S., Hackenbuchner, J., Arning, U., & Förstner, K. U. (2024). 07_LE_Dateistruktur. FDM@Studium.nrw Blended-Learning-Basiskurs „Forschungsdatenmanagement“ (Version 1.0). Abgerufen am 30.09.2025, von https://landesinitiativefdmnrw.github.io/FDMatStudium/thk/texte/07_LE_Dateistruktur.html

Deutsche Forschungsgemeinschaft. (2025). Guidelines for Safeguarding Good Research Practice. Code of Conduct. https://doi.org/10.5281/zenodo.14281892

Krähwinkel, E., Langner, P., Lipp, R., & Pietsch, A. M. (2022). HeFDI Data OER: Forschungsdatenmanagement – eine Online-Einführung. Zenodo. https://doi.org/10.5281/zenodo.6373596

Mcleod, J. (2022). DATUM for Health: Research data management training for health studies. Abgerufen am 30.09.2025, von https://oercommons.org/courses/datum-for-health-research-data-management-training-for-health-studies

MIT Libraries Data Management Group (2016). Data Management: File Organization. Abgerufen am 30.09.2025, von https://www.twillo.de/edu-

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt
Finanziert von der Europäischen Union
NextGenerationEU

sharing/components/render/01887843-1814-42e6-9ad5-6941a88d5c5a

MIT Libraries Data Management Group (2016). Research Data Management: Strategies for Data Sharing and Storage. Abgerufen am 30.09.2025, von https://www.twillo.de/edu-sharing/components/render/01887843-1814-42e6-9ad5-6941a88d5c5a

Oregon Health and Science University (2012). Introduction to Information and Computer Science: Databases and SQL. Health IT Workforce Curriculum. Abgerufen am 30.09.2025, von https://merlot.org/merlot/viewMaterial.htm?id=849864

Verbund Forschungsdaten Bildung (2024). Daten langfristig sichern. Abgerufen am 30.09.2025, von https://www.forschungsdaten-bildung.de/datenmanagement/sichern-teilen/daten-langfristig-sichern/

Schembera, B. (2024). Schlüsselqualifikation Forschungsdatenmanagement 4/8 (SQ): Dateninfrastrukturen. Abgerufen am 30.09.2025, von https://www.zoerr.de/edu-sharing/components/render/7753ff28-917e-4f10-9d9b-34e39217cb93

Weiterführende Info

Carrano, J. (2022). Tagging and Finding Your Files: Home. Abgerufen am 30.09.2025, von https://libguides.mit.edu/metadataTools/

Holzgraefe, S. (2018) Welches Datenbanksystem ist das Beste für mich und meine Anwendung?. TIB AV-Portal. https://doi.org/10.5446/45868

Software Carpentry (2018). File Organization: Organization. Abgerufen am 30.09.2025, von https://datacarpentry.github.io/rr-organization1/02-file-organization/index.html

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt
Finanziert von der Europäischen Union
NextGenerationEU

Das Material ist im Rahmen des Projekts „DIM.RUHR: Datenkompetenzzentrum für die interprofessionelle Nutzung von Gesundheitsdaten in der Metropole Ruhr“ (Förderkennzeichen: 16DKZ2008[A-F]) entstanden. Dementsprechend spiegelt das Material in Version 1.0 den Stand von 2025 wider.

„Daten managen - 1.2 Datenstrukturierung – Ordnerstrukturen“ von Anne Mainz und Sven Meister.

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt
Finanziert von der Europäischen Union
NextGenerationEU