Forschungsdatenmanagement

Notwendige, aber nicht hinreichende Voraussetzung für den wissenschaftlichen Erkenntnisgewinn

08. Speichern

Dr. habil. Till Biskup
Physikalische Chemie
Universität Rostock
17.05.2024

Zentrale Aspekte

- Rohdaten sollten wo immer möglich aufbewahrt werden, proprietäre Datenformate *zusätzlich* in einem offenen Format.
- Die begründete Löschung von (Roh-)Daten muss dokumentiert werden und darf nicht zulasten der Repräsentativität gehen.
- Daten müssen auffindbar (PIDs) strukturiert abgelegt werden. Das umfasst Roh- und abgeleitete Daten und Werkzeuge.
- Gespeicherte Daten müssen adressatengerecht dokumentiert sein und sollten gegen Verlust und Veränderung abgesichert werden.
- Projektleitung und Forschende sind für die Dokumentation, Institutionen für die langfristige Speicherung verantwortlich.

Der Forschungsdatenlebenszyklus

Modell der wissenschaftlichen Methode aus Sicht der Forschungsdaten

Speichern

Ein paar Gedanken zur Einleitung

- Warum Speichern erst nach Auswerten?
  Müssen Daten nicht direkt bei der Erhebung gespeichert werden?
  - Speichern für alle Stufen des Forschungsdatenlebenszyklus relevant
  - hier: „Speichern“ = längerfristige Speicherung der Forschungsdaten
  - unabhängig, ob sie „nur“ erhoben oder auch analysiert wurden
  - unabhängig, ob Rohdaten, abgeleitete Daten oder Ergebnisse von Analysen (z.B. Abbildungen, Tabellen).
- Was ist (wie lange) erhaltenswert?
  - DFG: zehn Jahre Aufbewahrungsfrist (für was genau?)
  - Aufenthaltsdauer von Forschenden an einem Ort: i.d.R. 5 ± 2 Jahre
  - Daten können über die Zeit an Wert gewinnen – weil sie (für individuelle Forschende) Kontext liefern.
  - (nachvollziehbar) verarbeitete Daten oft wertvoller als Rohdaten
- Zielgruppe: Zukunfts-Ich und direkte Kolleginnen und Kollegen
  - Gegensatz zu Veröffentlichung mit wesentlich breiterer Zielgruppe

Übersicht

Wissenschaftliche Aspekte

Organisatorische Aspekte

Verantwortung

Anforderungen und notwendige Werkzeuge

- Charakter der Daten verändert sich im Verlauf (des Zyklus)
  - viel heterogener als einfach nur (digitale) numerische Daten
  - Jede Art von Daten hat ggf. andere Ansprüche an die Speicherung.
- relevante Arten von Forschungsdaten
  - Rohdaten
  - abgeleitete und ausgewertete Daten
  - physische Artefakte
  - „Rezepte“ zur Datenverarbeitung und entsprechende Protokolle erfolgter Verarbeitungen
  - Werkzeuge (Software etc.)
  - Ergebnisse der Auswertung (Abbildungen, Tabellen, Berichte, etc.)
- drei Aspekte unter dem Oberbegriff „Datenkuration“
  - Auswahl – Reduktion, Kompression, Extraktion
  - Nachvollziehbarkeit
  - Transparenz

> *If I have seen further*
> *it is by standing on y^e shoulders of giants.*
>
> – Sir Isaac Newton

- Was macht einen „Riesen“ aus?
  - Nicht die Menge an Daten, sondern die Qualität der Erkenntnisse.
- Was sind „relevante“ Daten?
  - nichttrivial, muss immer wieder neu verhandelt werden
  - essentielle Verantwortung der Forschenden
- Auswahl *relevanter* Daten
  - Aspekte der Datenauswertung: Reduktion, Kompression, Extraktion
  - klare Regeln, welche Daten gelöscht werden können (*opt-in*)
  - dokumentieren, welche Daten warum nicht gespeichert wurden

Sir Isaac Newton: Brief an Robert Hooke, 5. Februar 1676

- Nachvollziehbarkeit
  - Metadaten auf abstrakterem Niveau:
    für Datensammlungen, aggregiert, pro (Unter-)Projekt
  - verständlich: Quellcode ist keine nachvollziehbare Dokumentation
  - Zielgruppe: Zukunfts-Ich und direkte Kolleginnen und Kollegen:
    macht hinreichende Dokumentation deutlich einfacher
- Transparenz
  - gelöschte oder nicht berücksichtigte Daten
  - nicht berücksichtigte oder verworfene Auswertungen
  - Wege der Entscheidungsfindungen:
    Fokus auf dem Warum (nicht), nicht nur dem Was Wie
- ☛ Parallele zu „clean code“ in der Softwareentwicklung
  - Quellcode kann idealerweise selbstdokumentierend sein, aber nur
    Was und Wie lässt sich in Code ausdrücken, nicht Warum (nicht).

Übersicht

- Unterschied zwischen Speichern und Archivieren
  - Archivieren: dauerhaftes Speichern ohne geplante zeitliche Begrenzung (für digitale Artefakte ein ungelöstes Problem)
  - Speichern: sichern für einen bestimmten Zeitraum

Datenarchivierung

revisionssichere langfristige Aufbewahrung von Daten gemäß gesetzlicher Ansprüche. Revisionssicherheit beinhaltet Kriterien wie Richtigkeit, Vollständigkeit, Schutz vor Veränderung und Verfälschung, Zugriffsbeschränkung, Einhaltung von Aufbewahrungsfristen und Dokumentation des Gesamtverfahrens. Datenarchivierung ist i.d.R. von Einzelpersonen und ohne unterstützende technische und qualifizierte personelle Infrastruktur nicht leistbar – und für digitale Artefakte ein nach wie vor weitgehend ungelöstes Problem.

Datenformate

- Kriterien für die Wahl des Datenformats
  - langzeit-stabil, offene Standards, nicht-proprietär
- Konversion proprietärer in langzeit-stabile Formate
  - Wichtig: (fast) keine Konversion ist verlustfrei
  - umsichtig planen, i.d.R. Originalformate mit aufbewahren
  - proprietäres Format klar benennen: Dateiendung hilft meist nicht (!)
  - weitere Informationen über das Format (soweit vorhanden) für das spätere Reverse Engineering mit ablegen

Versionierung

- Daten, Auswertungen und Darstellungen verändern sich
  - Parallele zur Softwareentwicklung: es existieren bewährte Lösungen
- Problem mit binären Daten: meist wenig geeignet für platzsparende Versionierung durch Speicherung der Unterschiede

Organisation der Datenablage

- Schemata für Datei- und Verzeichnisnamen
  - Kriterien: modular, flexibel, erweiterbar, „open–closed“
  - wenn Datum, dann YYYY-MM-DD, *nicht* (!) DD-MM-YYYY
  - nicht alle Informationen in den Datei- und Verzeichnisnamen
  - relevante Metadaten neben den Daten (durchsuchbar) ablegen
- Möglichkeit: Aufteilung nach Projekten
  - Verweise bei Daten, die für mehrere Projekte relevant sind
  - ggf. eine einheitliche Struktur, die an verschiedenen Stellen existiert (Dateisystem, Wiki, . . .)
- Alternative: lokales „Repositorium“
  - kann (erstmal) eine Verzeichnishierarchie sein
  - pro Projekt eine Liste mit PIDs der betreffenden Daten
  - komplett getrennt von thematischen Zusammenhängen

Organisation der Dokumentation

- Dokumentation i.d.R. getrennt von den eigentlichen Daten
  - egal, welcher Art die Daten sind
  - einfachster Fall: Datei neben den eigentlichen Daten
  - Alternativen: Verweise oder parallele Verzeichnishierarchien
- Werkzeuge: so niederschwellig wie möglich
  - nur so wird sie umgesetzt
  - Dokumentation essentiell für Nachvollziehbarkeit und Transparenz

dauerhafte (lokal) eindeutige Kennungen (PIDs)

- ein(ein)deutiger Verweis auf die Daten möglich
  - Beziehung zwischen Daten und ihrer Darstellung
  - bekannte (globale) Beispiele: DOI, ISBN
- Pfade im Dateisystem sind nicht langzeitstabil

Datensicherung vor Verlust/Veränderung

- Backup-Strategien
  1. lokale Ausfallsicherheit (RAID)
  2. Netzwerkspeicher (vor Ort)
  3. (institutionelle) professionelle Backup-Lösungen (örtlich getrennt)
- Prüfsummen
  - mindestens Erkennbarkeit der Datenveränderung
  - für Daten und Metadaten getrennt (unterschiedliche Änderungen)

Datensicherung vor unerlaubtem Zugriff

- Zugriffsschutz
  - Passwörter etc.
  - physischer Zugriffsschutz
- Verschlüsselung
  - erlaubt Speicherung auf öffentlichen Netzlaufwerken (*cloud*)

Übersicht

Verantwortung

Ein erster Überblick

- primäre Verantwortung
  - wissenschaftliche Projektleitung/Gruppenleitung
  - individuelle Forschende
- Unterstützung
  - Institution
  - Fachgesellschaften, Fördermittelgeber etc.

Leitmotiv

Forschungsdatenmanagement ist primär
die Verantwortung der individuellen Forschenden.

- ☛ Datenspeicherung zielt auf das Zukunfts-Ich und die eigene Gruppe.
- ☛ Institutionen/Fachgesellschaften sind lediglich unterstützend tätig.

wissenschaftliche Projektleitung/Gruppenleitung

- Auswahl der Daten (die dokumentiert gelöscht werden)
- Etablieren von Konventionen für
  - Datei- und Verzeichnisnamen
  - Kriterien für die Datenauswahl
  - Schemata und Formate für die Dokumentation
    (auf unterschiedlichen Ebenen: Daten, Verarbeitung, Projekt, . . .)

individuelle Forschende

- (Vor-)Auswahl der Daten inkl. dokumentierter Löschung
- Dokumentation der erhobenen und verarbeiteten Daten (Kuration)
- Einhalten von Konventionen für Datei- und Verzeichnisnamen
- Konvertierung proprietärer in offene, langzeit-stabile Formate

Institution

- Bereitstellung technischer Lösungen zur langfristigen Speicherung
- Beratung bei technischen Lösungen (Bibliothek, IT, Justiziariat)
- zeitlicher Horizont jenseits üblicher Projektlaufzeiten (3–5 Jahre)

Fachgesellschaften

- Etablierung von Standards für offene Formate
- Einwirken auf die Hersteller von Messgeräten
  - Bereitstellung des (verlustfreien) Export in offene Formate
  - alternativ Spezifikation des Dateiformats offenlegen

Fördermittelgeber

- Ausarbeitung *fachspezifischer* Leitlinien
- Bereistellung finanzieller Mittel für Werkzeuge und Lösungen

Übersicht

Anforderungen

Lesbarkeit, Auffindbarkeit, Transparenz

- Lesbarkeit der Daten/Informationen über lange Zeit
  - mindestens 10 Jahre (ungeachtet von Löschfristen)
  - idealerweise ohne technische Hilfsmittel
  - möglichst „selbstbeschreibend“
  - etablierte Konventionen (oder Standards) wo immer möglich (erhöht die Chance auf längerfristige Lesbarkeit)
- Zuordnung von Daten zu Projekten/Publikationen und Auffinden von Daten zu Projekten/Publikationen
  - unabhängig von einer momentan existierenden Verzeichnisstruktur
  - bidirektionale Beziehung bzw. Verweise: Daten $\Leftrightarrow$ Projekt
- für Dritte nachvollziehbare Datenablage
  - Daten überdauern i.d.R. länger als Forschende an einem Ort
- begründete Löschung von (Roh-)Daten muss dokumentiert werden
  - einfach zugängliche Dokumentationswerkzeuge
  - minimaler zusätzlicher Aufwand, offensichtliche Vorteile

Anforderungen

Datensicherung, Nachvollziehbarkeit, Transparenz, Dokumentation

- Daten sollten gegen Verlust gesichert abgelegt werden.
  - Anforderungen der Datensicherung aus der IT (Backup-Strategie)
- Nachvollziehbarkeit und Transparenz
  - Andere als die ursprünglichen Erhebenden müssen die Daten auswerten und ggf. gewinnbringend weiterverwenden können.
  - Dokumentation der Daten (durch Metadaten)
  - hinreichender Kontext von Datenaufnahme und -Verarbeitung
  - Zielgruppe: Personen mit vergleichbarem Kenntnisstand
  - Nachvollziehbarkeit $\neq$ Reproduzierbar- oder Replizierbarkeit
  - Dokumentation der begründeten Löschung von Daten
- Dokumentation auf unterschiedlichen Ebenen:
  - Datenaufnahme
  - Datenverarbeitung
  - konzeptionelle Dokumentation des (Projekt-)Zusammenhangs: warum, nicht nur was und wie – und was warum nicht

Notwendige Werkzeuge

Offene Formate, lokale IT, Repositorium für „warme Forschungsdaten“

- Offene, nicht-proprietäre, langzeit-stabile Datenformate
  - müssen alle verfügbaren Informationen speichern können
  - *verlustfreie* Konvertierung proprietärer Datenformate
  - Routinen zur möglichst automatischen Datenkonvertierung
- lokale Infrastruktur zur Speicherung primärer Forschungsdaten
  - primär auf lokaler Ebene (Arbeitskreis)
  - inkl. entsprechender Sicherung gegen unbeabsichtigten Verlust
  - modular, Anbindung an institutionelle Infrastruktur möglich
- Repositorium für „warme Forschungsdaten“
  - Repositorien werden meist als Ablage- bzw. Veröffentlichungsorte für Daten „abgeschlossener“ Projekte verstanden.
  - systematischer Ablageort für Daten, die *während* der Forschung anfallen und (erstmal) nicht veröffentlicht werden sollen/können
  - automatisches Hochladen durch Messgeräte (inkl. Metadaten)
  - funktional getrennt von ELN etc. (!)

Brandt, N. *et al.*, *Data Sci. J.* **20**:8, 2021

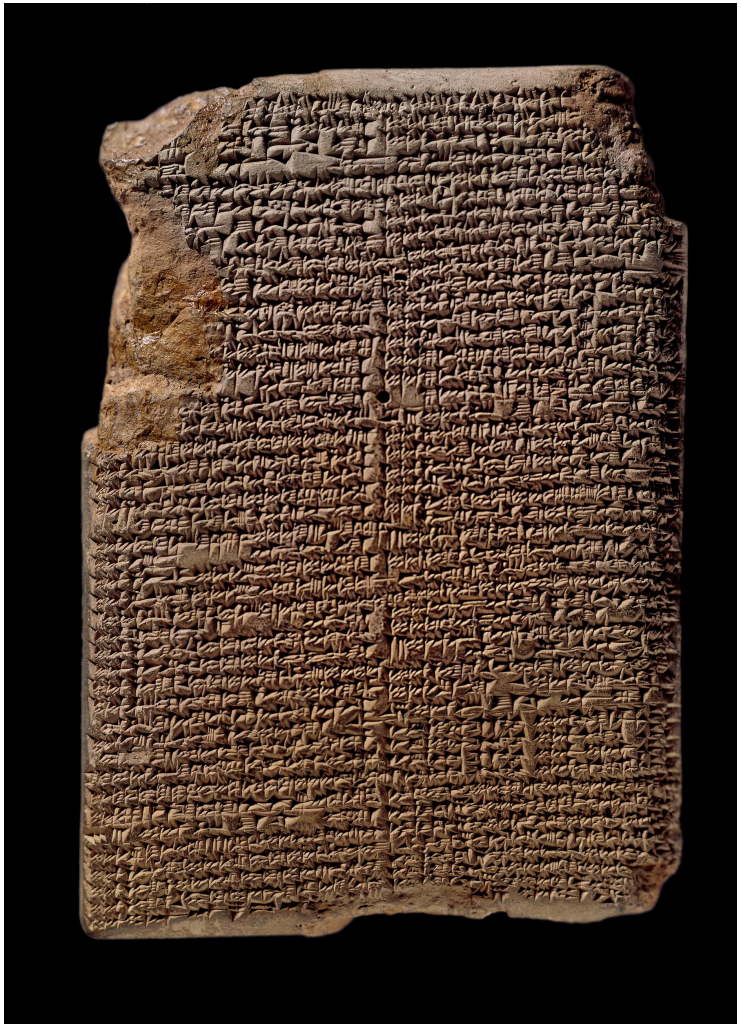
Notwendige Werkzeuge

Lokale PIDs, Dokumentationsschemata, dokumentierte Löschung

- lokale PIDs für das wiederauffindbare Ablegen der Daten
  - Pfade im Dateisystem sind i.d.R. nicht stabil
  - *lokal*: muss umsetzbar sein – und *eindeutig* im eigenen Kontext
- Schemata zur Dokumentation gespeicherter Daten(sammlungen)
  - Informationen über die Metadaten zur Datenerhebung hinaus
  - hinreichende Informationen zur Datenverarbeitung
  - ggf. „README“ o.ä. auf konzeptioneller Ebene
  - Parallele aus der Softwareentwicklung: der Code erklärt das wie, aber nicht das warum (und schon gar nicht das was warum nicht)
- Abläufe zur dokumentierten begründeten Löschung von Daten
  - i.d.R. Konventionen auf unterster Ebene
  - möglichst geringer administrativer Aufwand
  - Praxis: Daten werden nicht aktiv gelöscht, sondern irgendwann „entsorgt“, weil niemand mehr weiß, was es ist.

Forschungsdaten, die die Zeiten überdauert haben

Zwei unterschiedliche (prä-)historische Beispiele



Links: MUL.APIN (1250–1000 v.Chr.; Tafel 700 v.Chr.), British Museum via Wikipedia, CC BY-SA 4.0
Rechts: Himmelsscheibe von Nebra (2100–1700 v. Chr.), Dbachmann via Wikipedia, CC BY-SA 3.0

Zentrale Aspekte

- Rohdaten sollten wo immer möglich aufbewahrt werden, proprietäre Datenformate *zusätzlich* in einem offenen Format.
- Die begründete Löschung von (Roh-)Daten muss dokumentiert werden und darf nicht zulasten der Repräsentativität gehen.
- Daten müssen auffindbar (PIDs) strukturiert abgelegt werden. Das umfasst Roh- und abgeleitete Daten und Werkzeuge.
- Gespeicherte Daten müssen adressatengerecht dokumentiert sein und sollten gegen Verlust und Veränderung abgesichert werden.
- Projektleitung und Forschende sind für die Dokumentation, Institutionen für die langfristige Speicherung verantwortlich.