Forschungsdatenmanagement

Notwendige, aber nicht hinreichende Voraussetzung für den wissenschaftlichen Erkenntnisgewinn

07. Auswerten

Dr. habil. Till Biskup
Physikalische Chemie
Universität Rostock
17.05.2024

Zentrale Aspekte

- Datenauswertung umfasst die Vorverarbeitung und Analyse sowie die Zusammenfassung und Bewertung der Ergebnisse.
- Wissenschaftlichkeit und Nachvollziehbarkeit erfordern ein lückenloses Protokoll aller Verarbeitungsschritte.
- Werkzeuge zur Datenauswertung müssen modular, flexibel und automatisierbar sein – und möglichst einfach zu bedienen.
- Zur Auswertung implementierte Software sollte publiziert werden, die zugrundeliegenden Programme/Sprachen frei verfügbar sein.
- Praktisch verantwortlich sind die einzelnen Forschenden, die Projektleitung für die Etablierung der Prozesse.

Der Forschungsdatenlebenszyklus

Modell der wissenschaftlichen Methode aus Sicht der Forschungsdaten

Übersicht

Wissenschaftliche Aspekte

Organisatorische Aspekte

Verantwortung

Anforderungen und notwendige Werkzeuge

Wissenschaftliche Aspekte

Ein erster Überblick

- Vorgehen vom Bekannten zum Unbekannten
  - Wissenschaft findet immer in einem Kontext statt.
  - Reproduzieren bekannter Ergebnisse mit eigenen Auswertungsstrategien und Werkzeugen (Robustheit)
- Systematisches und korrektes Vorgehen
  - korrekter Umgang mit den Auswertungswerkzeugen
  - unter Berücksichtigung des aktuellen Stands der Wissenschaft
- Qualitätskontrolle während der Datenauswertung
  - Konsistenz: Kriterien vorab festgelegt, aus der Erfahrung gewonnen
  - Plausibilität: Sind die Ergebnisse sinnvoll/möglich?
- Metadaten und Dokumentation: Kontext der Datenauswertung
  - Nachvollziehbarkeit ist Kernaspekt von Wissenschaft
  - Nachvollziehbarkeit setzt hinreichende Dokumentation voraus

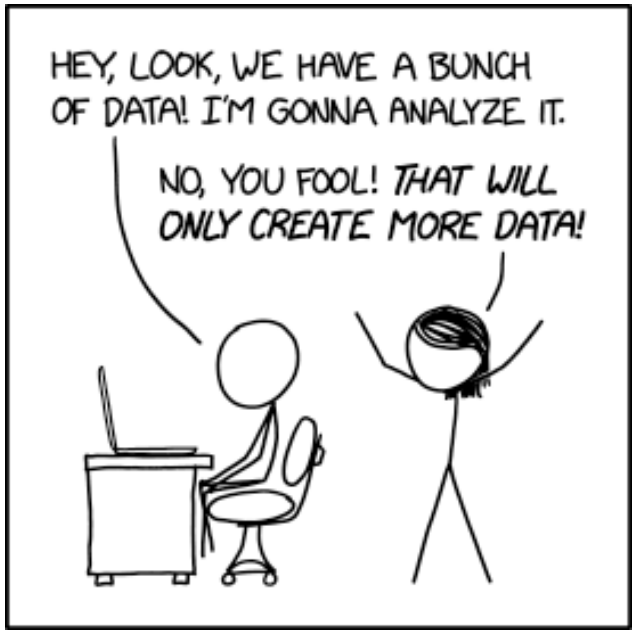
Wissenschaftliche Aspekte

. . . aus dem Wesen der Wissenschaft

- Nachvollziehbarkeit ist ein Kernaspekt von Wissenschaft.
  - erfordert lückenloses Protokoll aller Verarbeitungsschritte von den Rohdaten bis zur fertigen Präsentation (Abbildung, Tabelle)
  - eindeutige Identifizierung der verwendeten Datensätze
  - Bereitstellung oder zumindest hinreichende Beschreibung der verwendeten Werkzeuge
- Wissenschaftliche Fragestellungen sind immer wieder neu.
  - Bausteine einer Datenauswertung sind oft ähnlich/gleich, müssen sich aber der jeweiligen Fragestellung angepasst kombinieren lassen.
- Wissenschaft trachtet nach Erkenntnisgewinn.
  - Datenauswertung ist der entscheidende Prozess der Erkenntnisgewinnung.
  - Daten (und Ergebnisse von Auswertungen) sind keine Erkenntnis.
  - Erkenntnisgewinn lässt sich nicht automatisieren, die Voraussetzungen für den Erkenntnisgewinn mitunter schon.

- Reduktion
  - real Daten wegwerfen: ggf. dokumentiert löschen
  - Selten werden alle erhobenen Daten am Ende auch real verwendet.
  - Verantwortung der Forschenden: Konzentration auf das Wesentliche
  - „das Wesentliche“ immer nur im gegebenen Kontext bestimmbar, erfordert Erfahrung, Wissen und Abstraktionsfähigkeit
- Kompression
  - Charakteristika aus Daten extrahieren
  - Bsp.: Parameter eines Modells zur Beschreibung der Daten
  - Die Realität selbst in all ihrer Komplexität ist nur durch (vereinfachende) Modelle einer Erklärbarkeit zugänglich.
- Extraktion
  - Schlussfolgerungen ziehen, idealerweise Erkenntnisse gewinnen
  - Bsp.: Interpretation der Parameter eines Modells
  - Bsp.: Passgenauigkeit und erklärendes Potential von Modellen

DATA TRAP



"It's important to make sure your analysis destroys as much information as it produces."

☛ Reduktion ist essentieller Teil des Umgangs mit Forschungsdaten.

- mögliche Gründe für das Löschen von Daten
  - offensichtlich fehlerhafte Datenaufnahme
  - fehlender Kontext (Metadaten)

Typischer Ablauf wissenschaftlicher Datenauswertung

- Vorverarbeitung
  - Vorbereitung der Daten für die eigentliche Auswertung
  - Bsp.: Routinekorrekturen, Kalibrierung, ...
  - ggf. anschließend Daten verwerfen, weil ungeeignet
- eigentliche Auswertung
  - Extraktion von Charakteristika (Kompression)
  - Bsp.: Anpassung von parametrisierten Modellen an die Daten
- Zusammenfassung und Bewertung
  - Schlussfolgerungen aus den Ergebnissen der Auswertung ziehen
  - Datenpräsentation in geeigneter Weise

Übersicht

Organisatorische Aspekte

Umsetzung der wissenschaftlichen Aspekte

- Vorgehen vom Bekannten zum Unbekannten
  - relevante vorherige Ergebnisse identifizieren
  - Strategien zum Reproduzieren mit den eigenen Werkzeugen
- Systematisches und korrektes Vorgehen
  - Protokolle für den Umgang mit Strategien/Werkzeugen
  - Forschende entsprechend einlernen
- Qualitätskontrolle während der Datenauswertung
  - Kriterien für die Konsistenzprüfung festlegen
  - Notwendige Kenntnis der wissenschaftlichen Zusammenhänge
- Metadaten und Dokumentation: Kontext der Datenauswertung
  - möglichst wenige Medienbrüche
  - wenn Software beteiligt ist: automatische Protokollerstellung
  - ggf. Formalisierung von Protokollen zur Datenauswertung

Organisatorische Aspekte

Etablierte Abläufe, Werkzeuge, lückenlose Protokollierung

- Protokolle bzw. etablierte (und dokumentierte) Abläufe
  - Viele Bausteine der Datenauswertung sind immer ähnlich.
  - Kontextspezifisch gibt es meist etablierte Verfahren.
  - Dokumentation muss mit der Realität Schritt halten (können).
  - Protokolle/Abläufe müssen bekannt sein
    und korrekt und passend eingesetzt werden.
- Werkzeuge
  - Datenauswertung erfordert (fast) immer Werkzeuge.
  - Werkzeuge müssen bekannt und verfügbar sein.
  - ggf. Werkzeuge selbst entwickeln oder entwickeln lassen
- lückenloses Protokoll aller Verarbeitungsschritte
  - Voraussetzung für die Nachvollziehbarkeit
  - explizite und implizite Parameter jedes Verarbeitungsschritts
  - verwendete Werkzeuge, ggf. deren genaue Bezeichnung und Version
  - eindeutige Identifizierung der verwendeten Daten

Organisatorische Aspekte

Automatisierung, Software, eindeutige Kennungen

- Automatisierung wo immer sinnvoll möglich
  - Automatisierung sorgt für Konsistenz und Reproduzierbarkeit.
  - Formalisierung von Abläufen als Vorstufe zur Automatisierung
  - Automatisierung, um sich auf die wesentlichen Aspekte konzentrieren zu können, die menschliches Denkvermögen erfordern
- Wo immer möglich Einsatz von Software zur Datenauswertung
  - hilft bei der Automatisierung
  - erfordert Kenntnis und Vertrautheit mit dem jeweiligen Werkzeug
  - ggf. eigene (Weiter-)Entwicklung spezifischer Software
  - Bewusstsein für die dadurch entstehenden Abhängigkeiten und Implikationen für die Nachvollziehbarkeit
- eindeutige Kennungen (PID) für
  - Untersuchungsobjekte, von denen Daten erhoben wurden
  - (Roh-)Daten, die ausgewertet wurden
  - eingesetzte Werkzeuge (Software: Versionierung, Versionsnummern)

Übersicht

Verantwortung

Eine erste Übersicht

- wissenschaftliche Projektleitung/Gruppenleitung
  - Ermöglichen und Einfordern von Datenauswertung die den Ansprüchen der Wissenschaft genügt
  - Qualifizierung der Mitarbeitenden
- individuelle Forschende, die die Daten erheben
  - gewissenhaftes Arbeiten gemäß den Ansprüchen der Wissenschaft
  - eigenverantwortliche Qualifikation und Wissenserwerb

Leitmotiv

Forschungsdatenmanagement ist primär
die Verantwortung der individuellen Forschenden.

- ☛ Datenauswertung ist ureigene Aufgabe der Forschenden.
- ☛ Institutionelle Verantwortlichkeiten sind quasi nicht erkennbar.

Verantwortung

Wissenschaftliche Projektleitung/Gruppenleitung

- Vorgehen vom Bekannten zum Unbekannten
  - wissenschaftlichen Kontext bereitstellen und aufzeigen
  - Reproduzieren bekannter Ergebnisse aktiv einfordern
- Systematisches und korrektes Vorgehen
  - Etablierung von Abläufen, Strukturen und Konventionen
  - regelmäßige Überprüfung, ob die Abläufe eingehalten werden und noch angemessen sind bzw. angepasst werden müssen
- Qualitätskontrolle während der Datenauswertung
  - Kriterien zur Qualitätssicherung festlegen
  - für das notwendige Kontextwissen sorgen
  - Qualität der Auswertungen überprüfen (Plausibilität, Konsistenz)
- Metadaten und Dokumentation: Kontext der Datenauswertung
  - Strukturen und Werkzeuge etablieren
  - Einsatz der Werkzeuge motivieren und einfordern
  - Qualität der Dokumentation regelmäßig überprüfen

Verantwortung

Individuelle Forschende, die die Daten auswerten

- Vorgehen vom Bekannten zum Unbekannten
  - eigenständige Recherche des notwendigen konkreten Kontextes
  - gewissenhafte erfolgreiche Reproduktion bekannter Ergebnisse
- Systematisches und korrektes Vorgehen
  - etablierte Abläufe, Strukturen und Konventionen einhalten
  - kritisches Hinterfragen aus der Praxis
  - Strukturen (eigenständig) weiterentwickeln, zur Diskussion stellen
- Qualitätskontrolle während der Datenauswertung
  - Kriterien und Werkzeuge zur Qualitätskontrolle einsetzen
  - Plausibilitätsüberprüfung anhand wissenschaftlicher Kriterien
  - Voraussetzung: Kontextwissen und Verständnis der Fragestellung
- Metadaten und Dokumentation: Kontext der Datenerhebung
  - etablierte Strukturen und Werkzeuge einsetzen
  - kritisch auf Vollständigkeit und Handhabbarkeit hinterfragen, ggf. (eigenständig) weiterentwickeln und zur Diskussion stellen

Übersicht

Anforderungen

Zwei Thesen und ein Leitmotiv

These

Datenauswertung, die nicht vollständig dokumentiert
und nachvollziehbar ist, ist letztlich unwissenschaftlich.

These

Die wenigsten Wissenschaftler sind in der Lage, ihre Auswertungen
von den Daten bis zur Publikation hinreichend nachzuvollziehen.

Leitmotiv

Die Qualität eines Großteils veröffentlichter Forschungsergebnisse
wird den Ansprüchen der Wissenschaft nicht gerecht.

Anforderungen

Kontext, lückenloses Protokoll, nutzbare Werkzeuge

- alle relevanten Metadaten der Datenaufnahme
  - idealerweise maschinenlesbar verfügbar
  - Daten ohne Kontext lassen sich nicht auswerten (und sollten gelöscht werden).
- lückenloses Protokoll der Datenauswertung
  - Antwort auf fünf Fragen: Wer hat was mit wem wann wie gemacht?
- nutzbare Werkzeuge/Systeme
  - hinreichend mächtig, um den Anforderungen zu genügen
  - hinreichend dokumentiert, robust und intuitiv bedienbar

These

Nur Systeme, die hinreichend einfach nutzbar sind und deren Verwendung offensichtliche Vorteile bietet, werden genutzt werden.

Notwendige Werkzeuge

Gesamtsystem zur wissenschaftlichen Datenauswertung

- Gesamtsystem zur wissenschaftlichen Datenauswertung
  - lückenlose, automatische Protokollierung
  - hinreichend intuitiv bedienbar
  - hinreichend mächtig, modular, flexibel, erweiterbar, um den wechselnden Anforderungen und Fragestellungen zu genügen
  - idealerweise keine Medienbrüche:
    von den Rohdaten bis zur fertigen Abbildung/Tabelle
  - Schnittstellen zur restlichen Forschungsinfrastruktur
- quelloffene, frei lizenzierte Software und Programmiersprachen
  - Voraussetzung für die Nachvollziehbarkeit
  - kein Garant für Reproduzierbarkeit

☛ Werkzeuge sind nicht notwendigerweise digital.

☛ Werkzeuge gehören zu den Forschungsdaten im weiteren Sinn und sollten wo immer möglich mit publiziert werden.

Zentrale Aspekte

- Datenauswertung umfasst die Vorverarbeitung und Analyse sowie die Zusammenfassung und Bewertung der Ergebnisse.
- Wissenschaftlichkeit und Nachvollziehbarkeit erfordern ein lückenloses Protokoll aller Verarbeitungsschritte.
- Werkzeuge zur Datenauswertung müssen modular, flexibel und automatisierbar sein – und möglichst einfach zu bedienen.
- Zur Auswertung implementierte Software sollte publiziert werden, die zugrundeliegenden Programme/Sprachen frei verfügbar sein.
- Praktisch verantwortlich sind die einzelnen Forschenden, die Projektleitung für die Etablierung der Prozesse.