

Forschungsdatenmanagement

Notwendige, aber nicht hinreichende Voraussetzung
für den wissenschaftlichen Erkenntnisgewinn

14. Werkzeuge

Dr. habil. Till Biskup

Physikalische Chemie

Universität Rostock

14.06.2024





- 🔑 Werkzeuge nicht mit Lösungen verwechseln!
Lösungen erfordern den kompetenten Einsatz von Werkzeugen.
- 🔑 Werkzeuge reduzieren nicht die intrinsische Komplexität.
Sie erleichtern bestenfalls den Umgang mit komplexen Abläufen.
- 🔑 Entscheidend ist zu verstehen, *was* ich erreichen möchte.
Die Umsetzung ist ein zweitrangiger technischer Aspekt.
- 🔑 Werkzeuge sollten mindestens plattformunabhängig, robust, langfristig verfügbar, dezentral, modular und bewährt sein.
- 🔑 Grundausstattung ist ein Satz allgemeiner(er) Werkzeuge, die je nach Bedarf passend ergänzt und erweitert werden.

Werkzeuge sind keine Lösungen

Essentielle Eigenschaften

Arten von Werkzeugen

Beispiele konkreter Werkzeuge

These

Meist werden Werkzeuge mit der Lösung verwechselt.
Lösungen erfordern den kompetenten Einsatz von Werkzeugen.

- ▶ Die meisten politisch als Lösungen gehandelten Aspekte sind bestenfalls Werkzeuge, die Partikularinteressen einer verschwindenden Minderheit zufriedenstellen (können).
- ▶ Lösungen basieren immer auf dem kompetenten Einsatz und der Beherrschung von Werkzeugen, die den individuellen Gegebenheiten angepasst sind und entsprechend modular, flexibel und erweiterbar sein müssen.
- ▶ Lösungen setzen ein *Verständnis* der Problemstellung voraus.

“ *Every maker starts their journey with a basic set of good-quality tools. [...] These tools will be lovingly chosen, will be built to last, will perform specific jobs with little overlap with other tools, and, perhaps most importantly, will feel right in the budding woodworker's hands.*

Then begins a process of learning and adaptation. Each tool will have its own personality and quirks, and will need its own special handling. Each must be sharpened in a unique way, or held just so.

[...]

– Thomas und Hunt

“ *Tools amplify your talent.*

The better your tools, and the better you know how to use them, the more productive you can be.

Start with a basic set of generally applicable tools.

As you gain experience, and as you come across special requirements, you'll add to this basic set.

Like the maker, expect to add to your toolbox regularly.

Always be on the lookout for better ways of doing things.

If you come across a situation where you feel your current tools can't cut it, make a note to look for something different or more powerful that would have helped.

Let need drive your acquisitions.

– Thomas und Hunt

- ▶ Ausgangspunkt: grundlegender Satz hochwertiger Werkzeuge
 - bewusst gewählt, müssen sich „richtig“ anfühlen (individuell!)
 - auf dauerhafte Benutzung ausgelegt
- ▶ nächster Schritt: Einarbeitung und Kompetenzerwerb
 - Jedes Werkzeug hat seine Eigenheiten.
 - Wirksamkeit korreliert mit Vertrautheit und korrektem Gebrauch
- ▶ Werkzeuge verstärken *vorhandenes* Talent.
 - Werkzeuge sind nur so gut wie ihre Nutzung.
 - Werkzeuge können fehlendes Talent nicht ersetzen.
- ▶ regelmäßig Ausschau halten nach neuen Werkzeugen
 - Frage der Erfahrung und der speziellen Notwendigkeiten
 - eigene Prozesse kritisch hinterfragen: Verbesserungspotential?
- ▶ Der reale Bedarf entscheidet über neue Werkzeuge.
 - Voraussetzung: Bewusstsein und kritische Reflexion
 - Prioritätensetzung und sinnvoller Ressourceneinsatz

“ *There is no single development, in either technology or management technique, which by itself promises even one order-of-magnitude improvement within a decade in productivity, in reliability, in simplicity.*

– Frederick P. Brooks

- ▶ essentieller Unterschied: inhärente und unnötige Komplexität
 - inhärente Komplexität kann nicht reduziert, nur intellektuell beherrschbar(er) gemacht werden
- ▶ Produktivität von Personen bis ca. Faktor 10 unterschiedlich
 - abhängig vom kompetenten Einsatz von Werkzeugen und Strategien
 - meist relativ enge Rahmenbedingungen
 - meist Kombination diverser Werkzeuge und Strategien

These

Entscheidend ist die intellektuelle Durchdringung der Fragestellung.
Die konkreter Umsetzung ist ein zweitrangiger technischer Aspekt.

- ▶ Entscheidend ist zu verstehen, *was* ich erreichen möchte.
 - setzt Beschäftigung mit der Fragestellung und Geduld voraus
 - gute Lösungen sind i.d.R. Ergebnis eines langen Prozesses
 - gilt genauso für die Wahl der richtigen (Art der) Werkzeuge
- ▶ Die Umsetzung ist ein zweitrangiger technischer Aspekt.
 - hat ganz eigene Bedingungen und Randparameter
 - idealerweise mit klarer Spezifikation relativ einfach
- ▶ Arten von Werkzeugen \neq konkrete Implementierungen (!)
 - Implementierungen sind idealerweise austauschbar

Was

- ▶ Schema
- ▶ Semantik: Bedeutung
- ▶ Spezifikation

Wie

- ▶ Format
- ▶ Syntax: (äußere) Form
- ▶ Umsetzung

- ▶ Schemata sind vom konkreten Format (und Medium) unabhängig.
 - Ein Schema lässt sich vielfältig umsetzen.
 - Unterschiedliche Realisierungen können für Resilienz sorgen.
- ▶ Formate sind oft genug *abstrakte* (mächtige) Werkzeuge
 - Schema muss mit ihrer Hilfe abgebildet werden
 - Bsp.: XML, JSON, YAML, HDF5, ...
- ▶ Es gibt Formate, um Schemata zu beschreiben.
 - erlaubt die automatische Validierung gegen ein Schema
 - Bsp.: XSD

Werkzeuge sind keine Lösungen

Essentielle Eigenschaften

Arten von Werkzeugen

Beispiele konkreter Werkzeuge

These

Werkzeuge müssen weitestgehend unter Kontrolle der individuellen Forschenden sein. Bis auf die basale Infrastruktur (Schreibtisch, Strom, Internet) kann man sich seriös auf nichts und niemanden verlassen.

- ▶ „Bausteine eines funktionierenden, individuellen Forschungsdatenmanagements“
 - entspringt der eigenen langjährigen Erfahrung
 - positiver Aspekt: Gestaltungsfreiraum, Eigenverantwortlichkeit
 - negativer Aspekt: man ist für alles selbst verantwortlich
- 👉 Es gibt genug, was jede(r) selbst tun kann.
- 👉 Forschungsdatenmanagement ist primär die Aufgabe der individuellen Forschenden.

Standard

von einem oft internationalen und anerkannten Gremium bzw. einer Institution definierte (verbindliche) Festlegung

Konvention

innerhalb einer Gruppe oder einem (lokalen) Kontext getroffene (temporäre) Festlegung

- ▶ Ziel von Konventionen
 - Vereinheitlichung
 - Befreiung von der Notwendigkeit, jedesmal nachdenken zu müssen
- 👉 Konventionen lassen sich *immer* vereinbaren und umsetzen.

These

Grundlegende Werkzeuge müssen ohne größere Einarbeitungszeit funktionieren. Die Bereitschaft zur Einarbeitung ist fast nicht gegeben.

- ▶ Voraussetzung für kompetente und souveräne Nutzung
 - Einarbeitung, intellektuelle Auseinandersetzung
 - setzt Zeitinvestition (und Zeit am Stück) voraus
- ▶ intrinsische Komplexität wissenschaftlicher Fragestellungen
 - lässt sich durch Werkzeuge nicht reduzieren
 - kann nur intellektuell beherrschbar gemacht werden
- ▶ (Teil-)Kompensation durch Formalisierung und Automatisierung
 - reduziert im Idealfall Fehlbedienungen
 - sorgt für Entfernung der Nutzenden von der Fragestellung

- ▶ robust
 - viele verschiedene Aspekte
 - u.a. Toleranz gegenüber (versehentlicher) Fehlbedienung
- ▶ resilient
 - Forschungsdatenmanagement ist ein komplexer Prozess
 - komplexe Prozesse laufen selten komplett fehlerfrei
- ▶ modular
 - klar definierte Schnittstellen
 - einzelne Komponenten austauschbar
- ▶ dezentral
 - unter Kontrolle der individuellen Forschenden
 - trägt zur Robustheit bei
- ▶ attraktiv
 - Nur hinreichend einfach bedienbare Systeme mit öffentlichen Vorteilen ihrer Nutzung werden genutzt werden.

plattformunabhängig

- ▶ Software/digitale Werkzeuge
 - Unterstützung aller gängigen Betriebssysteme
- ▶ Hardware
 - Kompatibilität untereinander
 - sichert die Austauschbarkeit

keine Bindung an einen Anbieter (*vendor lockin*)

- ▶ Wissenschaft lebt von ihrer Unabhängigkeit
 - Anbieter haben i.d.R. wirtschaftliche Interessen
- ▶ Voraussetzung
 - offengelegte und hinreichend dokumentierte Schnittstellen und Austauschformate

langfristig wartbar

▶ Software

- muss aktiv gewartet werden, um in 1, 2, 5 Jahren noch zu laufen
- direktes Umfeld ändert sich schnell und beeinträchtigt Lauffähigkeit
- IT-Sicherheit erzwingt extrinsisch Änderungen des Umfelds

▶ Voraussetzungen

- hinreichende, auch konzeptionelle und technische Dokumentation

aktiv gewartet

▶ Werkzeuge sind Mittel zum Zweck

- ohne aktive Wartung ist ihre Funktionalität nicht sicherzustellen

▶ ggf. auch lokal umsetzbar, aber eigentlich immer ein Sonderfall

- Voraussetzung: ausreichend Kompetenz, Personal, Dokumentation

☞ Wartbarkeit und aktive Wartung unterscheiden sich deutlich.

- ❓ Ich habe hier ein ganz tolles Paket gefunden, das genau das tut, was ich brauche. Soll ich es für meine Auswertung verwenden?

Kriterien zur Evaluation des Wartungszustands von Software

- ▶ Changelog, Veröffentlichungshistorie, letzter Commit
 - ▶ Veröffentlichungsplattform
 - ▶ Bewertungen
 - ▶ Dokumentation
 - ▶ Zahl der Entwickler
 - ▶ ggf. Codequalität
 - ▶ Zahl der Suchmaschinentreffer
- 👉 Rechtzeitig recherchieren erspart später Kopfschmerzen.

bewährt

- ▶ Werkzeuge sind Mittel zum Zweck
 - sollten sinnvoller Weise funktionieren
 - Werkzeugentwicklung ist kein sinnvoller Nebenjob
- ▶ Hilfe von anderen Nutzenden
 - nur bei ausreichender Verwendung des Werkzeugs
 - einfaches Kriterium: Zahl der (sinnvollen) Suchmaschinentreffer

langfristige Perspektive

- ▶ Wissenschaft denkt in Generationen
 - Perspektive: > 10 Jahre
- ▶ komplexe Werkzeuge
 - kompetente Nutzung setzt Einarbeitung voraus
 - Zeit für die Einarbeitung muss gut investiert sein

Werkzeuge sind keine Lösungen

Essentielle Eigenschaften

Arten von Werkzeugen

Beispiele konkreter Werkzeuge

- ▶ Kernaspekt der Wissenschaft: Nachvollziehbarkeit
 - ▶ Wesen des Forschungsdatenmanagements
 - der Wissenschaft angemessener Umgang mit Forschungsdaten
 - ▶ Ziel des Forschungsdatenmanagements
 - Nachvollziehbarkeit sicherstellen
 - von der Planung über die Datenerhebung und Datenverarbeitung bis zu Veröffentlichung und Wiederverwendung von Forschungsdaten
 - ▶ Aspekte des Forschungsdatenmanagements
 - Dokumentation
 - sortierte Ablage und langfristige Bereitstellung von Forschungsdaten
 - eindeutige Verweise auf Forschungsdaten/Informationen
 - Automatisierung von Abläufen
- ☞ Jeder Aspekt erfordert andere Arten von Werkzeugen.

Planen

- Daten- und Softwaremanagementplan (DMP, SMP)
- Projektmanagementwerkzeuge

Erheben

- Schema und Format für Metadaten(erhebung)
- Strukturen und Protokolle für die Datenerhebung („SOP“)
- Konventionen für die strukturierte Datenablage
- Mechanismen zur Qualitätskontrolle

Auswerten

- Gesamtsystem zur wissenschaftlichen Datenauswertung
- quelloffene, frei lizenzierte Software und Programmiersprachen

Speichern

- Datenformate, Speichermedien, lokale Speicherinfrastruktur
- lokales Repositorium für „warme Forschungsdaten“
- lokale PIDs für das wiederauffindbare Ablegen von Daten

Veröffentlichen

- Zitierfähigkeit: Zitationsformate und eindeutige, stabile Kennungen
- Mechanismen zur Qualitätskontrolle
- Transparenz und Nachvollziehbarkeit:
schriftliche Kontextualisierung und Metadaten
- langfristige Verfügbarkeit: u.a. Datenformate, Speichermedien
- Auffindbarkeit: Suchmaschinen, Kataloge, Überblicksliteratur

Wiederverwenden

- (automatisierte) Qualitätskontrolle
- Relevanz: automatische Vorauswahl, hinreichende Metadaten
- Repräsentativität: Kataloge, Repositorien
- Rückführbarkeit: Zitierfähigkeit, dauerhafte Kennungen

 kein Anspruch auf Vollständigkeit

 nicht alles explizit im Forschungsdatenlebenszyklus abbildbar

Werkzeuge sind keine Lösungen

Essentielle Eigenschaften

Arten von Werkzeugen

Beispiele konkreter Werkzeuge

- ▶ Dokumentation
 - Metadaten (zur Datenerhebung, Datenverarbeitung, ...)
 - Laborbuch (analog und/oder digital)
 - schriftliche Kontextualisierung:
Berichte, Abschlussarbeiten, Veröffentlichungen, ...
- ▶ Ablage und Bereitstellung von Daten
 - modulare Verzeichnishierarchie
 - lokales Repositorium für „warme“ Forschungsdaten
 - technische Infrastruktur: lokal – NAS – Backup
- ▶ eindeutige Verweise auf Forschungsdaten/Informationen
 - Nummerierungsschemata für Proben, Ansätze, (Mess-)Daten, ...
 - lokale PIDs/UIDs
- ▶ Automatisierung
 - Skriptsprachen (bash, ...)
 - Programmiersprachen (Python, MATLAB, ...)

- ▶ Metadaten
 - entscheidend: Schema („Formular“)
 - technisch: Schlüssel-Wert-Paare
 - Schemaentwicklung: Welche Informationen sind relevant?
 - Daten und Metadaten nicht trennen
- ▶ Laborbuch
 - entscheidend: Struktur
 - Ablaufprotokoll der Arbeiten im Labor
 - Beschreibung der (relevanten) Beobachtungen
 - Daten und Auswertungen i.d.R. *nicht* im Laborbuch
- ▶ schriftliche Kontextualisierung
 - entscheidend: verbale Beschreibung der Ergebnisse
 - Struktur ist fast immer gleich
(egal ob Protokoll, Abschlussarbeit, Veröffentlichung, ...)
 - Werkzeuge zur (automatischen) Verarbeitung von Referenzen

- ▶ Datenformate
 - Daten und Metadaten gemeinsam ablegen
 - offen, nicht-proprietär, dokumentiert
 - Importroutinen einfach für viele Programmiersprachen erstellbar
- ▶ Dateinamen
 - kompakt, nicht zu lang
 - eindeutig zuzuordnen
 - relevante weitere Informationen in den Metadaten
 - Datum: YYYY-MM-DD, *nicht* DD-MM-YYYY
- ▶ modulare Verzeichnishierarchie
 - Hierarchie muss erweiterbar sein
 - mögliche Ebenen: Methode, Datum, Probe, Messung, ...
 - so intuitiv, dass andere sich zurechtfinden

- ▶ lokales Repository für „warme“ Forschungsdaten
 - ein Ort für die (Roh-)Daten
 - kann erstmal eine Verzeichnisstruktur sein
 - eindeutige und dauerhafte Verweise (PID/UID) auf Daten
 - Sicherung gegen versehentliche Veränderung
 - *nicht* (nur) auf individuellen Arbeitsrechnern
 - ggf. mit automatischer Synchronisation für Offline-Arbeit
- ▶ Kataloge
 - gleichförmige Beschreibung: automatisch erzeugte Berichte
 - relevante Metadaten, Verschlagwortung
- ▶ technische Infrastruktur
 - lokal: redundante Festplatten (RAID)
 - automatische Synchronisierung auf dezentrales NAS
 - automatisches Backup auf zentrale Speichermedien der Institution

- ▶ zwei Aufgaben
 - Planungswerkzeug
 - Dokumentation und Referenz
- ▶ Voraussetzungen für den Erfolg
 - klare Struktur
 - Antworten auf alle relevanten Fragen
 - „lebendiges Dokument“: muss die Realität wiedergeben und aktuell gehalten werden, sonst ist es irrelevant (oder falsch)
- ▶ Anforderungen an das Format
 - Änderungen/Aktualisierungen einfach möglich
 - ggf. Versionierung zur Nachvollziehbarkeit der Entwicklung
 - einfacher Zugriff von allen relevanten Orten aus
- ☛ Der Inhalt ist entscheidend – Form und Name sind egal.
- ☛ Planung und Dokumentation sind keine lästige Pflicht, sondern essentielle Voraussetzung für Wissenschaftlichkeit.

- ▶ Versionsverwaltungssystem
 - für alle Arten von Dokumenten und Code
 - muss dezentral/lokal funktionieren
 - automatische Synchronisierung von Instanzen untereinander
 - kann essenziell sein für die Nachvollziehbarkeit
- ▶ Wiki
 - strukturierte Ablage von Informationen („Wissensdatenbank“)
 - einfach editierbar, niederschwellig
 - geringe technische Voraussetzungen (Webbrowser)
 - gut geeignet für veränderliche Informationen
 - Nachteil: setzt i.d.R. Netzwerkverbindung voraus
- ▶ Checklisten etc.
 - Formalisierung wiederkehrender Abläufe
 - Vorstufe für eine Automatisierung
 - können (und werden oft) Listen auf Papier sein



- 🔑 Werkzeuge nicht mit Lösungen verwechseln!
Lösungen erfordern den kompetenten Einsatz von Werkzeugen.
- 🔑 Werkzeuge reduzieren nicht die intrinsische Komplexität.
Sie erleichtern bestenfalls den Umgang mit komplexen Abläufen.
- 🔑 Entscheidend ist zu verstehen, *was* ich erreichen möchte.
Die Umsetzung ist ein zweitrangiger technischer Aspekt.
- 🔑 Werkzeuge sollten mindestens plattformunabhängig, robust, langfristig verfügbar, dezentral, modular und bewährt sein.
- 🔑 Grundausstattung ist ein Satz allgemeiner(er) Werkzeuge, die je nach Bedarf passend ergänzt und erweitert werden.