Forschungsdatenmanagement

Notwendige, aber nicht hinreichende Voraussetzung für den wissenschaftlichen Erkenntnisgewinn

11. Eigenschaften und Konzepte

Dr. habil. Till Biskup
Physikalische Chemie
Universität Rostock
31.05.2024

Zentrale Aspekte

- Forschungsdatenmanagement ist die Aufgabe der Forschenden und muss mit vorhandenen Mitteln umsetzbar sein.
- Ein funktionierendes individuelles Forschungsdatenmanagement muss dezentral, modular, flexibel und erweiterbar sein.
- Ziel ist die intellektuelle Beherrschbarkeit komplexer Zusammenhänge und zunehmender Datenmengen.
- Es gibt keine „eierlegende Wollmilchsau“.
  Systeme, die das versuchen zu sein, sind zum Scheitern verurteilt.
- Es gibt keine schlüsselfertigen Lösungen „von der Stange“.
  Jedes System muss an die spezifischen Bedarfe angepasst werden.

Standortbestimmung

Teil drei von fünf

1. Ausgangspunkt und Motivation
   - Wissenschaft; Datenmenge und Digitalität
2. Aspekte eines forschungsnahen Forschungsdatenmanagements
   - Forschungsdatenlebenszyklus
3. Bausteine eines funktionierenden, individuellen Forschungsdatenmanagements
   - Eigenschaften; Prinzipien; notwendige Kompetenzen; Werkzeuge
4. Hindernisse und Probleme
   - Hindernisse und mögliche Lösungen; Antimuster
5. Funktionierende Lösungen
   - Bewährte Verfahren (aus eigener Anschauung)

Übersicht

Fokus: funktionierendes, individuelles Forschungsdatenmanagement

Ziel: intellektuelle Beherrschbarkeit

Eigenschaften

Konzepte

Leitmotiv

Forschungsdatenmanagement ist primär
die Verantwortung der individuellen Forschenden.

- Fokus der Vorlesung: individuelle Forschende, die sich nicht auf externe Infrastruktur verlassen können (und wollen).
- Ungeachtet dessen, dass es Dinge gibt, die ich nicht selbst in der Hand habe: Bevor ich mit dem Finger auf andere zeige, gibt es eigentlich immer sehr viel, was ich (erst einmal) selbst tun kann.
- Die hier vorgestellten Strategien und Konzepte entspringen der eigenen langjährigen Erfahrung ohne wirkliche Ressourcen (weder technisch noch personell).

Leitfrage

- Was kann ich als individuelle(r) Forschende(r) im Hier und Jetzt tun, um zu einer Verbesserung der Wissenschaftlichkeit (meiner eigenen Forschung) beizutragen?

- Ziel: Sicherstellung/Verbesserung der Wissenschaftlichkeit
  - Es geht nicht um Häkchen und Metriken.
  - Extrinsische Anforderungen nutzen, um sich selbst die richtigen (wichtigen) Fragen zu stellen – und zu beantworten.
- persönliche Verantwortung
  - Es geht um *meine* eigene Forschung.
  - Es geht um das, was *ich* tun kann.

Umsetzung von Forschungsdatenmanagement

Fahrplan für das persönliche Umfeld

> *Do what you can, with what you've got, where you are.*
>
> – Theodore Roosevelt

- Bewusstsein für die Relevanz von FDM schärfen
  - Es geht um die Wissenschaft(lichkeit) – nicht weniger.
  - FDM ist Voraussetzung, aber kein Garant für Wissenschaftlichkeit.
- notwendiges Wissen aneignen
  - FDM ist nicht neu: Es gibt genug gute (alte) Literatur dazu.
  - Auf diejenigen hören, die langjährige praktische Erfahrung haben.
- notwendige Kompetenzen aneignen
  - Nur kompetent eingesetzte Werkzeuge werden zu Lösungen.
- FDM im persönlichen Kontext umsetzen
  - Schrittweise vorgehen: von einfach zu komplex, von bekannt und bewärt zu unbekannt und unerprobt.

Theodore Roosevelt: An Autobiography, Macmillan, New York 1913, p. 364

Thema: „Bausteine eines funktionierenden individuellen FDM“

- Welche Kriterien müssen Bausteine eines funktionierenden individuellen Forschungsdatenmanagements erfüllen?
- Welche Kompetenzen müssen die Nutzenden dieser Bausteine mitbringen bzw. sich erarbeiten?

Zielstellung

- Befähigung, Anforderungen selbst zu formulieren
- Überblick über hilfreiche Werkzeuge,
  deren Nutzung in Eigenregie erlernt werden muss

☛ Die Fähigkeit, Ansprüche und Anforderungen zu formulieren, stellt sicher, dass die *richtigen* Lösungen entwickelt werden.

Fünf Aspekte

Ein erster Überblick und Unterscheidung

- Eigenschaften
  - konkrete Charakteristika einzelner Werkzeuge oder Bausteine, die zu deren Auswahl verwendet werden können
- Konzepte
  - abstrakte Bausteine, die sich in Werkzeugen integrieren lassen, die genannten Eigenschaften aufweisen und den Prinzipien folgen
- Prinzipien
  - abstrakte Handlungsanweisungen zur Entwicklung von Bausteinen
  - unabhängig von den konkreten Werkzeugen
- Kompetenzen
  - notwendige persönliche Fähigkeiten für FDM
  - Grundausstattung jeder Wissenschaftlerin/jedes Wissenschaftlers
- Werkzeuge
  - konkrete Bausteine für das Forschungsdatenmanagement
  - keine Lösungen

Übersicht

Fokus: funktionierendes, individuelles Forschungsdatenmanagement

Ziel: intellektuelle Beherrschbarkeit

Eigenschaften

Konzepte

These

Die Anforderungen an eine Vorgehensweise, die den Ansprüchen der Wissenschaft genügt, insbesondere eine hinreichende Nachvollziehbarkeit gewährleistet, sind so groß, dass jeglicher Versuch, diesen Ansprüchen gerecht zu werden, zu einer erheblichen Komplexität führt.

zwei Arten von Komplexität

- inhärente, unvermeidliche Komplexität (*essential complexity*)
- unnötige, vermeidbare Komplexität (*accidental complexity*)

☛ Ziel: unnötige Komplexität möglichst vermeiden, inhärente Komplexität handhabbar gestalten

Fred Brooks: *The Mythical Man Month*. Addison Wesley Longman, Boston 1995.

Intellektuelle Beherrschbarkeit (*intellectual manageability*)

> *[. . .] the only problems we can really solve in a satisfactory manner are those that finally admit a nicely factored solution. [. . .] By the time that we are sufficiently modest to try factored solutions only, because the other efforts escape our intellectual grip, we shall do our utmost best to avoid all those interfaces impairing our ability to factor the system in a helpful way.*
>
> – E. W. Dijkstra

- komplexe Problem in lösbar(er)e Teilprobleme zerlegen
- entsprechende Schnittstellen zwischen diesen Teilen schaffen

E. W. Dijkstra, *Commun. ACM* 15:859–865, 1972

Parallele: (Entwurfs-)Muster

Beschreibung eines wiederkehrenden Problems und seiner möglichen Lösung

Parallele zu Mustern (nicht nur) in der Softwareentwicklung

- Muster beschreiben sowohl ein wiederkehrendes Problem als auch eine abstrakte (mögliche) Lösung.
- Muster müssen immer in einem konkreten Kontext implementiert werden (Muster selbst sind keine Lösung).
- Muster können zu wenig und zu viel verwendet werden.
- Muster beschreiben in der Praxis *bewährte* Verfahren.
- Muster beinhalten notwendigerweise eine Kosten–Nutzen-Abwägung.

Ursprung von Mustern

1977 Christopher Alexander (Architekturtheoretiker)

1995 Gamma *et al.*: Entwurfsmuster (Software)

Übersicht

Fokus: funktionierendes, individuelles Forschungsdatenmanagement

Ziel: intellektuelle Beherrschbarkeit

Eigenschaften

Konzepte

These

Funktionierende Werkzeuge und Lösungen müssen dezentral, modular, flexibel und erweiterbar sein.

- Dezentral: Nur Forschende selbst können im realen Forschungsalltag die Komplexität der relevanten Abläufe erfassen.
- Modular: Die Aufgabe ist zu groß, als dass sie von einem Team auf einmal angegangen werden kann. Modularität ermöglicht Priorisierung und überhaupt erst die (schrittweise) Umsetzung.
- Flexibel: Ansprüche ändern sich ständig, auch aufgrund der fortschreitenden intellektuellen Durchdringung der Fragestellung.
- Erweiterbar: Anforderungen entwickeln sich ebenso mit der fortschreitenden intellektuellen Durchdringung der Fragestellung.

Dezentral

Ausmaß der Komplexität nur vor Ort wirklich erkennbar

These

Zentralisierung scheitert meist an der Komplexität der Realität und der Unmöglichkeit, ohne die praktische Erfahrung mit realen Anwendungsfällen ein relevantes funktionierendes System zu erstellen.

- dezentral bedeutet *nicht* ...
  - (ausschließlich) lokal: automatisierte Synchronisierung mit zentralen Services sollte immer möglich sein
  - das Rad neu zu erfinden (*not invented here*-Syndrom) oder auf Synergien und Erfahrungen anderer zu verzichten
  - dass jede/r untereinander inkompatible Insellösungen erarbeitet
- Beispiel: git als *verteiltes* Versionsverwaltungssystem
  - hat serverbasierte Lösungen weitgehend abgelöst

Dezentral

Gründe für dezentrale Werkzeuge

- Unabhängigkeit von zentraler Infrastruktur
  - Arbeiten ohne Internet-/Netzwerkzugriff möglich
- Nähe zur eigentlichen Problemdomäne
  - zentrale Werkzeuge sind meist zu unspezifisch
  - funktionierende Werkzeuge müssen an die spezifischen lokalen Gegebenheiten anpassbar sein
- geringere Angriffsfläche
  - zentrale Dienste sind mitunter ein wertvolles Ziel
  - verteilte Systeme sind potentiell datenschutzfreundlicher
- tendenziell geringere Komplexität
  - Mandantenfähigkeit und Authentifizierung ggf. verzichtbar
  - Fokussierung auf den konkreten Anwendungsfall statt Berücksichtigung aller Eventualitäten
  - kein Widerspruch zur Modularität und Flexibilität

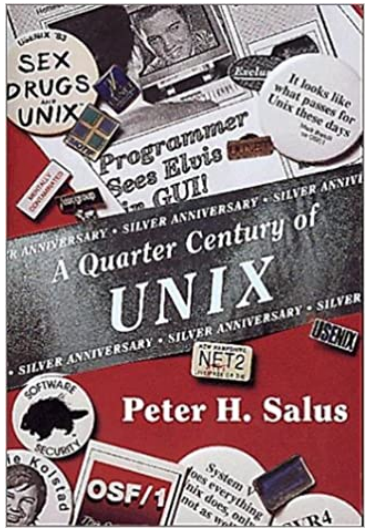
Modular

Einzige Chance für intellektuelle Beherrschbarkeit

- einzige Chance für intellektuelle Beherrschbarkeit der Bausteine
  - Anforderungen viel zu komplex für ein monolithisches System
  - Die „eierlegende Wollmilchsau“ ist immer zum Scheitern verurteilt.
- mit aktuellen Ressourcen umsetzbar
  - Forschungsdatenmanagement und IT spielen nach wie vor eine untergeordnete Rolle in der akademischen Wissenschaft.
  - meist keine Möglichkeit, sich auf zentrale Dienste zu verlassen (abgesehen vom generellen Vorteil dezentraler Systeme, s.o.)
- Voraussetzung: Schnittstellen
  - müssen explizit und hinreichend detailliert beschrieben sein
- Voraussetzung für Flexibilität und Erweiterbarkeit
  - beides intrinsische Anforderungen der Wissenschaft (s.u.)
- miteinander lose verbundene Einheiten statt ein Gesamtsystem
- jahrzehntelang bewährtes Beispiel: Unix – vgl. Unix-Philosophie

Die Unix-Philosophie

- Write programs that do one thing and do it well.
- Write programs to work together.
- Write programs that handle text streams, because that is a universal interface.



☛ Gilt für die Softwareentwicklung genauso wie für die analoge Welt.

☛ Modularität und Interoperabilität sind entscheidend, um mit begrenzten Ressourcen Werkzeuge zu entwickeln und zu etablieren.

Peter H. Salus (1994): A Quarter Century of UNIX. Reading (MA), Addison-Wesley; S: 53.

Flexibel

Intrinsische Anforderung der Wissenschaft

- intrinsische Anforderung der Wissenschaft
  - Auch wenn viele Aufgaben ähnlich sind,
    ist doch jede im Detail wieder speziell und spezifisch.
  - Ansprüche entwickeln sich mit fortscheitender
    intellektueller Durchdringung der Fragestellung.
- Voraussetzung: Modularität
  - sauber definierte, dokumentierte und verwendete Schnittstellen
  - erlaubt die Kombination vorhandener Bausteine in ganz neuer,
    ungeahnter Weise (Beispiel: LEGO®)
- Balance zwischen Freiheit und Komplexität
  - Ultimative Freiheit führt zu unbeherrschbarer Komplexität.
  - einfache Aufgaben leicht und intuitiv machen,
    schwere ermöglichen und nicht verhindern
  - komplexe Anwendungsfälle gründlich durchdenken,
    nicht sinnvolle Kombinationen von Parametern ggf. verhindern

Erweiterbar

Einzige Chance, komplexe Systeme mit begrenzten Ressourcen aufzubauen

- intrinsische Anforderung der Wissenschaft
  - Anforderungen entwickeln sich mit fortschreitender intellektueller Durchdringung der Fragestellung.
- Freiheit, nur das gerade konkret notwendige zu implementieren
  - Niemand hat eine Kristallkugel: Bedarfe schwer vorhersehbar
  - Eventualitäten zu implementieren, obwohl man sie jetzt nicht braucht, führt meist zu vermeidbarer Komplexität (YAGNI).
  - Fokussierung auf die aktuellen Notwendigkeiten
  - einzige Chance, mit begrenzten Ressourcen ein komplexes System schrittweise aufzubauen
- Voraussetzung: Modularität
  - sauber definierte, dokumentierte und verwendete Schnittstellen
  - Erweiterbarkeit möglichst ohne Bruch der Abwärtskompatibilität

YAGNI = *You ain't gonna need it*

- Herausforderung: Systemarchitektur
  - das große Bild nicht aus dem Blick verlieren
  - Möglichkeiten nicht durch ungeschickte Entscheidungen verbauen
  - Architektur: Summe der getroffenen Entscheidungen
  - gute Architekturen minimieren die Zahl getroffener Entscheidungen
- Balance zwischen Generalisierbarkeit und Beherrschbarkeit
  - Ultimative Generalisierung führt zu unbeherrschbarer Komplexität.
  - intellektuelle Durchdringung der Problemstellung
    hilft bei der Einschränkung auf relevante Aspekte
  - Voraussetzung: Erfahrung sowohl in der Fachwissenschaft als auch
    in Entwurf und Implementierung entsprechender Systeme
  - Generalisierbarkeit ist keine Ausrede,
    auf tragfähige und fruchtbare Abstraktionen zu verzichten.

Übersicht

Fokus: funktionierendes, individuelles Forschungsdatenmanagement

Ziel: intellektuelle Beherrschbarkeit

Eigenschaften

Konzepte

Konzepte

Konzepte sind keine Werkzeuge und keine Prinzipien,
sondern abstrakte Bausteine, die sich in Werkzeugen integrieren lassen.

- Welche grundlegenden Konzepte gibt es, um ein funktionierendes individuelles Forschungsdatenmanagement aufzubauen?

Übersicht über die hier behandelten Konzepte

- Metadaten
- Schnittstellen
- Verweise (Links)
- eindeutige/dauerhafte Kennungen (UID/PID)

Metadaten

Arten und Ebenen, Ziele, technische Aspekte

- unterschiedliche Arten/Ebenen von Metadaten
  - Datenerhebung:
    untersuchtes Objekt, Instrumentierung, Untersuchung
  - Datenverarbeitung und -Auswertung
  - Aggregation von Daten für eine Veröffentlichung oder ein Projekt
  - Veröffentlichung
- Ziele von Metadaten
  - strukturierte Ablage aller relevanten Informationen
  - Möglichkeit, die durch die Metadaten beschriebenen Daten
    zu klassifizieren und zu sortieren
- technische Aspekte
  - Schlüssel–Wert-Paare (gern auch hierarchisch)
  - idealerweise maschinenverarbeitbar
  - Metadaten sind vollkommen unabhängig vom Medium
    und haben *per se* nichts mit Digitalität zu tun.

Schema

formales Modell der Struktur von Daten bzw. Informationen

- Kriterien
  - deckt alle relevanten Informationen ab
  - möglichst intuitiv aufgebaut (z.B. Reihenfolge)
  - modular und flexibel erweiterbar (Open–Closed-Prinzip)
- Nicht das Schema mit dem Format für die Umsetzung verwechseln!

Open–Closed-Prinzip

Offenheit für Erweiterungen bei gleichzeitiger Abgeschlossenheit gegenüber (inkompatiblen) Abänderungen

Ontologie

Darstellung der Eigenschaften eines Fachgebiets und ihre Beziehungen zueinander, indem eine Reihe von Konzepten und Kategorien definiert wird, die das Fachgebiet repräsentieren.

Kontrolliertes Vokabular

Sammlung von Begriffen mit dem Ziel, die Beschreibung von Objekten zu vereinheitlichen. Innerhalb des kontrollierten Vokabulars sind die Begriffe eindeutig identifiziert.

- Ontologien
  - große Vorteile bei korrektem Einsatz:
    Möglichkeit, automatisiert Schlüsse zu ziehen.
  - Versprechen: echtes „Verständnis“ durch Maschinen
  - nette Idee, aber in der Praxis fast nicht umsetzbar
  - Voraussetzungen für die Erstellung:
    umfassende Kenntnis des Fachgebiets und der Zusammenhänge
  - bis auf wenige Spezialfälle so komplex und so weit weg vom Erfahrungshorizont der Wissenschaftlerinnen und Wissenschaftler, dass ein praktischer Einsatz quasi ausgeschlossen ist
- kontrollierte Vokabulare
  - lokalere bzw. einfachere Alternativen zu Ontologien
  - deutlich weniger formal
  - keine Beziehungen von Begriffen untereinander
- noch einfacher: Festlegung von Datentypen in einem Schema

Schnittstellen

Essentielle Voraussetzung für die Modularität

- Trennung von Zuständigkeiten (*separation of concerns*)
  - ggf. auch organisatorisch/personell getrennte Zuständigkeiten
  - Was hinter einer Schnittstelle passiert, ist „privat“.
- Bedeutung
  - essentielle Voraussetzung für die Modularität
    (und damit intellektuell beherrschbare Systeme, s.o.)
- Voraussetzung
  - explizite Spezifikation (und Dokumentation)
  - konsequente Verwendung der Schnittstellen
- Herausforderung
  - Schnittstellen sollten gut entworfen werden
  - einmal eingeführt nur selten, mit Vorwarnung und begründet ändern
  - Voraussetzung: praktische Erfahrung sowohl in der Fachdomäne
    als auch in Systemarchitektur

These

Die „eierlegende Wollmilchsau“ ist grundsätzlich unerreichbar und jeder Versuch, sie zu implementieren, zum Scheitern verurteilt.

- Grund für die Entwicklung umfassender monolithischer Systeme
  - viele miteinander zusammenhängende Anforderungen
  - Arbeiten möglichst ohne Medienbrüche
  - einzelne Bereiche sollen automatisiert miteinander interagieren
- Lösung für das Problem der „eierlegenden Wollmilchsau“
  - Verweise (Links) zwischen Subsystemen:
    lose Kopplung über klar definierte Schnittstellen
  - erhält die Modularität und damit intellektuelle Beherrschbarkeit
  - verringert zu enge Kopplung und Abhängigkeiten (Resilienz)

Verweise (Links)

Beispiele, Voraussetzungen, Vorteile

- Das Konzept des Verweises ist unabhängig von der Digitalität:
  - Schnitzeljagd
  - Findbücher, Zettelkataloge etc. in Bibliotheken
  - Referenzen und Zitationen
- bekannte Beispiele digitaler Verweise
  - Hypertext – durch das WWW bekannt gemacht
  - Links im Dateisystem
- Voraussetzungen
  - definierte Protokolle für den Informationstransport
  - wohldefinierte Schnittstellen (s.o.)
  - eindeutige und dauerhafte Kennungen (s.u.)
- Vorteile
  - Entkopplung von (Teil-)Systemen
  - Verweise unabhängig vom physischen Aufbewahrungsort änderbar
  - Verweise von mehreren Orten aus auf eine Instanz

Eindeutige/dauerhafte Kennungen

Allgemeine Aspekte

- zwei unterschiedliche Eigenschaften
  - (im jeweiligen Kontext) eindeutig (*unique identifier*, UID)
  - dauerhaft (*persistent identifier*, PID)
- Vorteile
  - eindeutiger Verweis auf eine Ressource möglich
  - Entkopplung vom realen (physischen) Ort
  - unabhängig von der Organisation der Ressourcen
- grundsätzliche Möglichkeiten für Schemata
  - „sprechend“: z.B. hierarchisch, manuell erzeugbar
  - eindeutig, aber nicht menschenlesbar (z.B. Hash)
- bekannte (globale) Vertreter
  - DOI, ISBN, ORCID, ...
  - eigentlich auch URLs (vgl. Berners-Lee)

Tim Berners-Lee (1998): „Cool URIs don't change“, `https://www.w3.org/Provider/Style/URI.html`

Eindeutige/dauerhafte Kennungen

Ein paar Anmerkungen

- Pfade im Dateisystem
  - lokal ggf. eindeutig
  - nicht wirklich dauerhaft
  - können als Ausgangspunkt dienen
- Umgang mit Schemata
  - wenn nicht lesbar: Zuordnungstabelle gut sichern
  - wenn lesbar: Zuordnungstabelle i.d.R. generierbar
- Dauerhaftigkeit ist nur institutionell zu gewährleisten
  - auch dann abhängig von der Institution
  - bei digitalen Ressourcen oft nach wie vor fehlendes Bewusstsein
- ein paar mögliche Anwendungsfälle
  - Proben
  - Datensätze
  - Instrumente
  - Protokolle von Datenauswertungen

Zentrale Aspekte

- Forschungsdatenmanagement ist die Aufgabe der Forschenden und muss mit vorhandenen Mitteln umsetzbar sein.
- Ein funktionierendes individuelles Forschungsdatenmanagement muss dezentral, modular, flexibel und erweiterbar sein.
- Ziel ist die intellektuelle Beherrschbarkeit komplexer Zusammenhänge und zunehmender Datenmengen.
- Es gibt keine „eierlegende Wollmilchsau“.
  Systeme, die das versuchen zu sein, sind zum Scheitern verurteilt.
- Es gibt keine schlüsselfertigen Lösungen „von der Stange“.
  Jedes System muss an die spezifischen Bedarfe angepasst werden.