Forschungsdatenmanagement

Notwendige, aber nicht hinreichende Voraussetzung für den wissenschaftlichen Erkenntnisgewinn

12. Prinzipien

Dr. habil. Till Biskup
Physikalische Chemie
Universität Rostock
Sommersemester 2024

Zentrale Aspekte

- Automatisierung ist eine Frage der Ökonomie und ermöglicht die Fokussierung auf die eigentliche Wissenschaft.
- Automatisierung bedeutet nicht automatisch Digitalisierung, auch wenn beide häufig miteinander einher gehen.
- Klare Abläufe sorgen für Konsistenz und Routine und erleichtern die Nutzung.
- Werkzeuge und Lösungen für das Forschungsdatenmanagement müssen robust und resilient sein, um akzeptiert zu werden.
- Nur Systeme, die hinreichend einfach nutzbar sind und deren Verwendung offensichtliche Vorteile bietet, werden genutzt werden.

❓ Welchen Prinzipien sollte Forschungsdatenmanagement und sollten seine Bausteine folgen?

Prinzipien

- eher abstrakte Handlungsanweisungen, um brauchbare Werkzeuge für ein individuelles, funktionierendes Forschungsdatenmanagement zu entwickeln (Abgrenzung zu Eigenschaften)
- Werkzeuge, die den Prinzipien folgen bzw. die Prinzipien umsetzbar machen, werden die genannten Eigenschaften aufweisen.
- Prinzipien sind von den konkreten Werkzeugen unabhängig und weisen über sie hinaus.

☛ Vier Prinzipien werden nachfolgend eingehender betrachtet.

Übersicht

Automatisierung

Robustheit

Resilienz

Attraktivität

Automat

von gr. *autómatos*: sich selbst bewegend; Apparat, der einen einmal eingeleiteten technischen Vorgang ohne weiteres menschliches Zutun steuert oder regelt.

- entscheidender Aspekt: „ohne menschliches Zutun“
- Nur was keiner menschlichen Interaktion bedarf, kann/sollte automatisiert werden.
- Automatisierung ermöglicht die Konzentration auf die Dinge, die tatsächlich menschlicher Interaktion bedürfen – weil es um Verständnis und/oder Entscheidungen geht.

☛ Nicht alles (vermeintlich) technisch mögliche ist auch sinnvoll.

Automatisierung

Motor des (gesellschaftlichen) Fortschritts

> *It is a profoundly erroneous truism, repeated by all copybooks and by eminent people when they are making speeches, that we should cultivate the habit of thinking of what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them. Operations of thought are like cavalry charges in a battle—they are strictly limited in number, they require fresh horses, and must only be made at decisive moments.*
>
> – Alfred North Whitehead

- ☛ *kein* Plädoyer, Dinge nicht zu durchdenken (im Gegenteil)
- ☛ wiederkehrende Abläufe und Zusammenhänge *einmal* durchdenken und formalisieren, um nachfolgend Denkleistung einzusparen

Whitehead: *An Introduction to Mathematics*, Dover, Mineola 2017 (orig. 1911), S. 34

Gründe für Automatisierung

- begrenzte Ressourcen zur Zielerreichung
- Beschleunigung von Abläufen

Voraussetzungen der Automatisierung

- zu automatisierende Abläufe intellektuell durchdrungen

Folgen der Automatisierung

- Ermöglichung von mangels Ressourcen bislang unmöglichen Dingen
- Konsistenz (aber: nicht zwangsläufig Korrektheit)
- Beschleunigung von Abläufen
- Entfernung der Durchführenden von den eigentlichen Vorgängen
  - Verlust des Überblicks und des Verständnisses, was wirklich passiert
  - sorgt im Fehlerfall/bei unerwarteten Resultaten oft für Hilflosigkeit

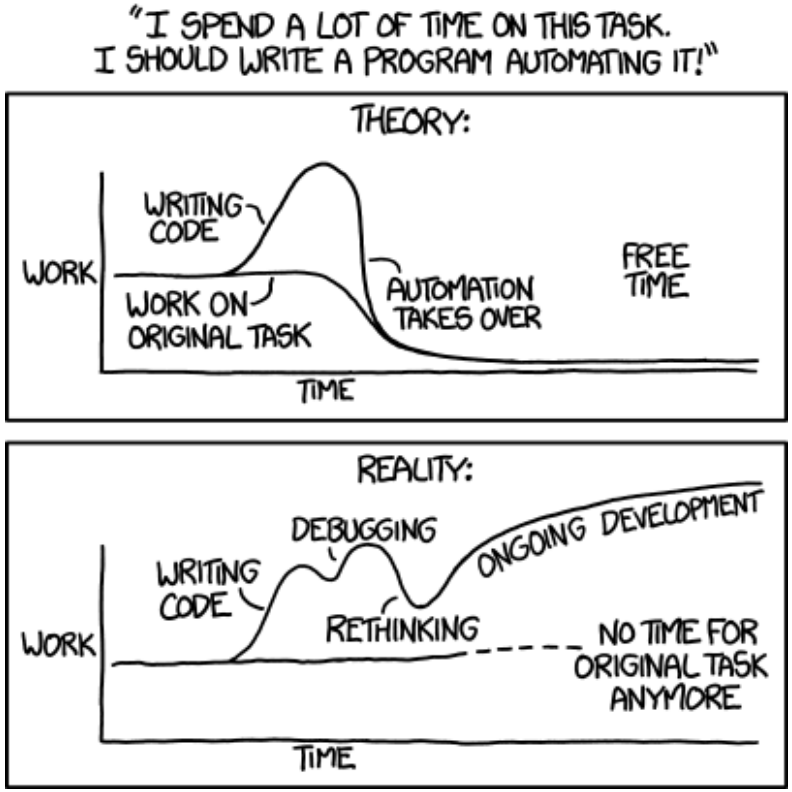
Automatisierung bedeutet nicht automatisch Digitalisierung

- Automaten waren mechanisch, sind es mehrheitlich immer noch
  - Programmierbarkeit von Automaten nimmt mittlerweile zu
- Folgen der Programmierbarkeit
  - größere Flexibilität
  - potentiell größere Fehleranfälligkeit und Angreifbarkeit

Formalisierung von Abläufen

- meist notwendiger „Vorläufer“ der Automatisierung
- Bsp.: Metadatenerhebung während der Datenaufnahme
  - durch Formular strukturierbar
  - Nachdenken entfällt, mehr oder weniger mechanischer Prozess
- Automatisierung i.e.S. ohne menschliche Interaktion
  - sorgt so für Konsistenz und oft genug für Existenz

Automation



"'Automating' comes from the roots 'auto-' meaning 'self-', and 'mating', meaning 'screwing'."

© Randall Munroe, CC-By-NC 2.5, https://xkcd.com/1319/

Übersicht

Unterschiedliche Aspekte

- Toleranz gegenüber Fehlbedienung/falschen Nutzereingaben
- Langlebigkeit/Langfristigkeit
- numerische Stabilität von Lösungen
- Idempotenz
- Gegenteil von Zerbrechlichkeit bei (Software-)Systemen

Strategien, um Robustheit zu erzeugen

- modulare Systeme
  - wohldefinierte und stabile Schnittstellen
- Verweise/Verknüpfungen
  - dauerhafte und eindeutige Kennungen (PID, UID)

Robustheit

Toleranz gegenüber Fehlbedienung und Langlebigkeit

- Toleranz gegenüber Fehlbedienung/falschen Nutzereingaben
  - hilfreiche Fehlermeldungen
  - Überprüfung von Nutzereingaben: Wertebereich und Datentyp
  - Formate, die robust gegenüber Leerraum etc. sind
  - Formate, die sich automatisch validieren lassen (Bsp.: XML)
  - Bsp. sequenziellen Datenauswertung: kein Programmabbruch mit Verlust aller Ergebnisse nach 20 von 32 Schritten...
- Langlebigkeit/Langfristigkeit
  - Wissenschaft ist generationenübergreifend...
  - Bsp.: (möglichst) Verzicht auf proprietäre Formate und Software (siehe den letzten Teil der Unix-Philosophie)
  - Kriterien für die Auswahl von Werkzeugen: Wartbarkeit, langfristige(re) Verfügbarkeit, aktuelle Wartung/Pflege

Robustheit

numerische Stabilität, Idempotenz, (Nicht-)Zerbrechlichkeit

- (numerische) Stabilität von Lösungen
  - Viele Algorithmen sind numerisch, nicht analytisch.
  - Algorithmen sind unterschiedlich numerisch stabil.
  - setzt hinreichendes Verständnis der verwendeten Algorithmen, ihrer Grenzen und ihrer Eignung für die konkrete Fragestellung voraus
- Idempotenz
  - mehrfache Ausführung führt zum gleichen Zustand des Systems wie die einmalige Ausführung
  - wichtig für die Fehlertoleranz mehrschrittiger Abläufe, wenn sie in einem Durchlauf nicht erfolgreich waren
- Gegenteil von Zerbrechlichkeit bei (Software-)Systemen
  - Zerbrechlichkeit: Änderungen an einer Stelle führen zu Problemen an ganz anderen, davon eigentlich unabhängigen Stellen (Kopplung)
  - Robustheit sorgt für wartbare und erweiterbare Systeme

Übersicht

Automatisierung

Robustheit

Resilienz

Attraktivität

Resilienz

Eigenschaft komplexer Systeme, trotz massiver interner oder externer Störungen wieder in den Ausgangszustand zurückzukehren.
Bei technischen Systemen die Fähigkeit, ihre wesentlichen Aufgaben auch bei Störungen und Teilausfällen weiter zu erfüllen.

- Forschungsdatenmanagement erfordert ein komplexes System
  - *intrinsische* Komplexität aus dem Anspruch der Wissenschaft
- komplexe Systeme sind fehleranfällig
  - oft ist die Komplexität nicht bewusst/bekannt
- Ausfallzeiten können teuer werden
  - verantwortungsvoller Ressourceneinsatz

- Redundanz
  - Vorhalten mehrerer gleicher Systeme
  - eines kann für das andere einspringen
- Backup
  - Möglichkeit, im Fehlerfall Informationen wiederherzustellen
  - auch wenn sie auf dem fehlerhaften System nicht mehr existieren
- technisch komplett unabhängige Verfahren
  - Bsp.: Papier statt digital
  - Voraussetzung: Existenz der notwendigen Werkzeuge und Kompetenz der Beteiligten/Betroffenen zu deren Einsatz
- automatisierte Konfektionierung/Konfigurierung von Systemem
  - ermöglicht automatisierte Konfiguration von Ersatzsystemen
  - reduziert so den Stresspegel im Fehlerfall ganz erheblich
  - Voraussetzung: keine manuelle Nacharbeit an installierten Systemen

Resilienz

Ein prototypisches Beispiel: Internet (bzw. ARPANET)

- Zielstellung
  - resilientes, funktionierendes Kommunikationsnetz, auch beim Ausfall sehr vieler Verbindungen
- Kernkomponenten
  - Netzwerk vieler gleichwertiger untereinander verbundener Knoten statt Punkt-zu-Punkt-Verbindungen wie beim Telefon
  - Datenpakete mit Metadaten: Absender, Empfänger, Reihenfolge
  - Transportprotokoll mit Rückmeldung bei Empfang
  - mehrere Verbindungen pro einzelnem Knoten
- Ergebnis
  - Datenpakete werden solange weitergeleitet, bis sie ankommen
  - einzelne Datenpakete können unterschiedliche Routen nehmen
  - Reihenfolge der Ankunft am Ziel ist nicht entscheidend

Resilienz

Ein paar Beispiele für das Forschungsdatenmanagement

- Datenaufnahme: Speicher
  - unabhängig vom (lokalen) Netzwerk
  - automatische Synchronisation der Daten, wenn verfügbar
  - bestenfalls periodische Überprüfung und automatische Synchronisation, wenn wieder verfügbar
- Datenauswertung: Algorithmen
  - Verwendung beschleunigter Algorithmen nur bei Verfügbarkeit
  - automatische Überprüfung der Verfügbarkeit
  - Rückgriff auf immer vorhandene Varianten
- Datenablage: Nachvollziehbarkeit
  - Verzeichnisse statt (ausschließlich) Datenbank
  - automatisierte Ablage redundanter Informationen (z.B. Metadaten mit Beschreibung)
  - Redundanz bei automatischer Synchronisation unproblematisch

Listing 1: Resilienter Import eines Pakets nur bei dessen Verfügbarkeit

```
try:
    from optimparallel import minimize_parallel as minimize
except ImportError:
    from scipy.optimize import minimize
```

Anmerkungen

- `try..except` fängt Fehler ab
  - wichtig: spezifische Fehler abfangen (hier: `ImportError`)
- Paket immer unter identischem Namen verfügbar: `minimize`
  - Vorteil u.a. des Python-Import-Mechanismus
  - Beispiel für saubere Schnittstelle
- Resilienz: Rückgriff auf Standard-Bibliothek
  - ermöglicht die Verwendung experimenteller(er) Bibliotheken
  - Code funktioniert tendenziell langfristig(er)

Übersicht

Attraktivität

Voraussetzung für die Nutzung eines Systems

These

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

- Forschungsdatenmanagement ist immer Mehraufwand, auch wenn man die Ansicht vertreten kann, dass schlechtes oder nicht ausreichendes Forschungsdatenmanagement zu unwissenschaftlichem Arbeiten führt.
- Es gibt unterschiedliche Motivationen: (äußerer) Zwang ist vermutlich mit die schlechteste Motivation; am Besten ist die Einsicht in die Notwendigkeit gepaart mit dem Wissen, dass gute Routinen später nicht nur sehr viel Arbeit sparen, sondern manche Dinge überhaupt erst möglich machen.

Attraktivität

Aspekte, die die Attraktivität eines Systems beeinflussen (können)

- intuitiv
  - logische/gewohnte Anordnung von Elementen
  - offensichtliche Bedienung, zumindest grundlegend
- automatisiert
  - kommt der menschlichen Faulheit entgegen
  - hilft dabei, sich auf die relevanten Aspekte zu konzentrieren
- robust
  - tolerant gegenüber Fehlbedienung
- ermöglichend
  - hilft, Ansprüchen (z.B. Wissenschaftlichkeit) gerecht zu werden
- wirkmächtige Abstraktionen bietend
  - Voraussetzung: Abstraktionsvermögen, analytisches Denken, intellektuelle Durchdringung der Fragestellung
  - wirkmächtige Abstraktionen können, müssen aber nicht intuitiv sein

Zentrale Aspekte

- Automatisierung ist eine Frage der Ökonomie und ermöglicht die Fokussierung auf die eigentliche Wissenschaft.
- Automatisierung bedeutet nicht automatisch Digitalisierung, auch wenn beide häufig miteinander einher gehen.
- Klare Abläufe sorgen für Konsistenz und Routine und erleichtern die Nutzung.
- Werkzeuge und Lösungen für das Forschungsdatenmanagement müssen robust und resilient sein, um akzeptiert zu werden.
- Nur Systeme, die hinreichend einfach nutzbar sind und deren Verwendung offensichtliche Vorteile bietet, werden genutzt werden.