Forschungsdatenmanagement

Notwendige, aber nicht hinreichende Voraussetzung für den wissenschaftlichen Erkenntnisgewinn

04. Forschungsdatenlebenszyklus

Dr. habil. Till Biskup
Physikalische Chemie
Universität Rostock
26.04.2024

Zentrale Aspekte

- Der Forschungsdatenlebenszyklus ermöglicht die Kommunikation mit Forschenden über Forschungsdatenmanagement.
- Die Fokussierung auf Daten und ihre Wiederverwendbarkeit statt Nachvollziehbarkeit ist kontraproduktiv für die Wissenschaft.
- Kernaspekt der wissenschaftlichen Methode ist Nachvollziehbarkeit – und damit Dokumentation auf allen Stufen des Zyklus.
- Die Bedeutung des Begriffs „Forschungsdaten“ ändert sich über die Stationen des Zyklus hinweg.
- Anhand des Forschungsdatenlebenszyklus lassen sich unterschiedliche Verantwortlichkeiten erkennen.

Standortbestimmung

Teil zwei 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

These

Der Forschungsdatenlebenszyklus ist ein Modell der wissenschaftlichen Methode aus Sicht der Forschungsdaten und ermöglicht es, mit Forschenden konstruktiv über ihren Umgang mit Forschungsdaten in den Dialog zu treten – zur Qualitätssicherung der Forschung.

- Forschungsdatenmanagement ist eine notwendige, aber keine hinreichende Voraussetzung wissenschaftlichen Erkenntnisgewinns.
- Wir haben großteils verlernt, richtig mit unseren Daten umzugehen. Quantität statt Qualität und sorgloser Umgang mit Ressourcen wird zunehmend zu einem Problem für die Wissenschaft.
- Alles, was hilft, das Thema ins Bewusstsein zurückzubringen und die Qualität der Forschung zu verbessern, ist begrüßenswert.

Der Forschungsdatenlebenszyklus als Modell

Ein abstraktes Modell – und eine von vielen Ausführungen

Der Forschungsdatenlebenszyklus

Ein paar erste Kommentare zu den einzelnen Stationen

Planen

- Art und Umfang der Daten abschätzen
- Urheberschaft, Beteiligte, Lizenzen und Schutzrechte klären

☛ Datenmanagementplan (nicht nur für Geldgeber!)

☛ einfach zugängliche Werkzeuge zur Projektplanung

Erheben

- Metadaten *während* der Datenaufnahme erheben
- Wer hat was mit wem, wann, wie und warum gemacht?

☛ maschinenlesbar und *menschenschreibbar*

☛ Erhebung kann nicht (vollständig) automatisiert werden

Der Forschungsdatenlebenszyklus

Ein paar erste Kommentare zu den einzelnen Stationen

Auswerten

- lückenloses Protokoll aller Verarbeitungs- und Analyseschritte
- (vollständig) reproduzierbare Datenverarbeitung und -analyse

☛ Protokoll muss *vollständig automatisch* erzeugt werden

☛ Lückenlosigkeit nur mit Gesamtsystem zur Datenverarbeitung

Speichern

- (de)zentraler Speicher mit zentralem Backup
- Konventionen für Datei- und Verzeichnisnamen oder PIDs

☛ Messgeräte laden Daten automatisch in Datenspeicher hoch

☛ lokale PIDs (Hinweis: Pfade im Dateisystem sind nicht dauerhaft)

Der Forschungsdatenlebenszyklus

Ein paar erste Kommentare zu den einzelnen Stationen

Veröffentlichen

- Beschreibung des zu veröffentlichenden Datenpakets
- Vollständigkeit: Daten, Dokumentation, Auswertungen, ...

☛ Datenkuration (um Datenqualität sicherzustellen)

☛ Arbeitsablauf für (automatisiertes) Hochladen in Repositorium

Wiederverwenden

- Überblick über verfügbare Forschungsdaten
- direkter Link auf Daten, alternativ Kontaktdaten

☛ Katalog (lokal) verfügbarer Forschungsdaten

☛ disziplinspezifische Repositorien für Forschungsdaten

Der Forschungsdatenlebenszyklus

Kommentare zur spezifischen Ausgestaltung

- 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).
- Warum Speichern und nicht Archivieren?
  - Archivieren = Speichern für unbegrenzten Zeitraum
  - Archivierung digitaler Artefakte ist ein ungelöstes Problem.
- Warum fehlt „Daten teilen“ (*data sharing*)?
  - hier implizit unter „Veröffentlichen“ und „Wiederverwenden“
  - Daten gewinnbringend mit anderen zu teilen ist sehr aufwendig und oft genug (aus unterschiedlichen Gründen) unmöglich.

Bedeutung des Forschungsdatenlebenszyklus für FDM

- Forschungsdatenmanagement
  - sinnvoller und angemessener Umgang mit Forschungsdaten
  - inkl. der Werkzeuge wie Software, Bibliographien, etc.
- Forschungsdatenlebenszyklus
  - für Forschende intuitives und nachvollziehbares Modell
  - ermöglicht konstruktives Gespräch/Nachdenken über *für die Forschenden* relevante FDM-Aspekte

Forschungsdatenlebenszyklus als Planungsinstrument

- hilft, wesentliche Aspekte anzugehen und im Blick zu behalten
- vergleichbare Entwicklungszyklen: Software, Projekte allgemein
  - Das Rad wurde schon (mehrfach) erfunden . . .

Übersicht

Kritische Betrachtung: Tragfähigkeit und Grenzen

Ein erster Überblick

- Einseitiger Fokus auf Daten: Daten sind kein Selbstzweck.
  - Popper: Daten/Beobachtungen sind nur kritische Argumente.
  - Wissenschaft: Erkenntnisgewinn, nicht Datensammlung
- Weiterverwendung von Daten ist oft schwer oder gar unmöglich.
  - Daten ohne Kontext (d.h. Metadaten) sind wertlos.
  - Ausreichenden Kontext bereitzustellen ist sehr herausfordernd.
- Weiterverwendung von Daten führt (meist) nicht zur Erkenntnis.
  - Erkenntnis entsteht aus Verständnis von Zusammenhängen.
  - Daten sind maximal Grundlage für ein Verständnis.
- In den meisten Fällen ist ein Zyklus das falsche Bild.
  - Weiterverwendung von Daten ist fast nie relevant.
  - Die wissenschaftliche Methode ist viel eher ein zyklisches Vorgehen.
- Manche Aspekte passen nicht in den Forschungsdatenlebenszyklus.

Daten sind kein Selbstzweck

Wissenschaft strebt nach Erkenntnissen statt nur Ergebnissen

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

Was ist der Kern von Wissenschaft?

- Erkenntnisgewinn
- Unabhängigkeit vom Betrachter/Experimentator
- gegründet auf den Erkenntnissen früherer Generationen
- überprüfbar, nachvollziehbar, ggf. reproduzierbar

☛ Wissenschaftler tragen Verantwortung gegenüber jenen, die auf den gewonnenen Erkenntnissen aufbauen.

☛ Wissenschaftler tragen Verantwortung dafür, Erkenntnisse zu erlangen statt nur Ergebnisse zu produzieren.

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

Daten sind kein Selbstzweck

Wissenschaft strebt nach Erkenntnissen statt nur Ergebnissen

Erkenntnis

Aneignung des Sinngehalts von erlebten bzw. erfahrenen Sachverhalten, Zuständen oder Vorgängen, Ergebnis des Vorgangs des Erkennens. Erkenntnis beinhaltet immer eine auf die Erfahrung gestützte Beurteilung und setzt notwendiger Weise ein Subjekt voraus, das erkennt. Neue Erkenntnisse, die von innerer und äußerer Erfahrung unabhängig sind, sind immer Ergebnis einer schöpferischen Phantasie. Bei der Erkenntnis stehen sich Subjekt und Objekt als Erkennendes und Erkanntes gegenüber. Die Erkenntnis führt zu einem Abbild des Objekts im Subjekt. Die grundsätzliche Unvollständigkeit dieses Abbilds ist die Triebkraft hinter dem Erkenntnisgewinn und letztlich der Wissenschaft.

vgl. H. Schmidt: Philosophisches Wörterbuch. Kröner, Stuttgart 1991
vgl. I. Kant: Kritik der reinen Vernunft. Suhrkamp, Stuttgart 1974 (original 1781)
vgl. I. Kant: Metaphysische Anfangsgründe der Naturwissenschaft. Felix Meiner, Hamburg 1997

Problematik des Zyklus-Charakters

Der Zyklus impliziert ein sequentielles Vorgehen, das selten gegeben ist.

- Der Forschungsdatenlebenszyklus ist lediglich ein Modell.
  - Es gibt sehr viele Varianten mit unterschiedlichen Schwerpunkten.
  - „Alle Modelle sind falsch, aber manche sind nützlich.“ (Box)
  - „Cargo Cult Science“ (Feynman): Form statt Wesen
- Stufen des Zyklus folgen nicht unbedingt streng aufeinander.
  - Aspekte je nach Forschungsansatz unterschiedlich gewichtet
  - iterative (letztlich rückgekoppelte) Prozesse
  - viele kleine Schleifen, ggf. auch Überspringen von Stufen

Blick über den Tellerrand: Softwareentwicklung

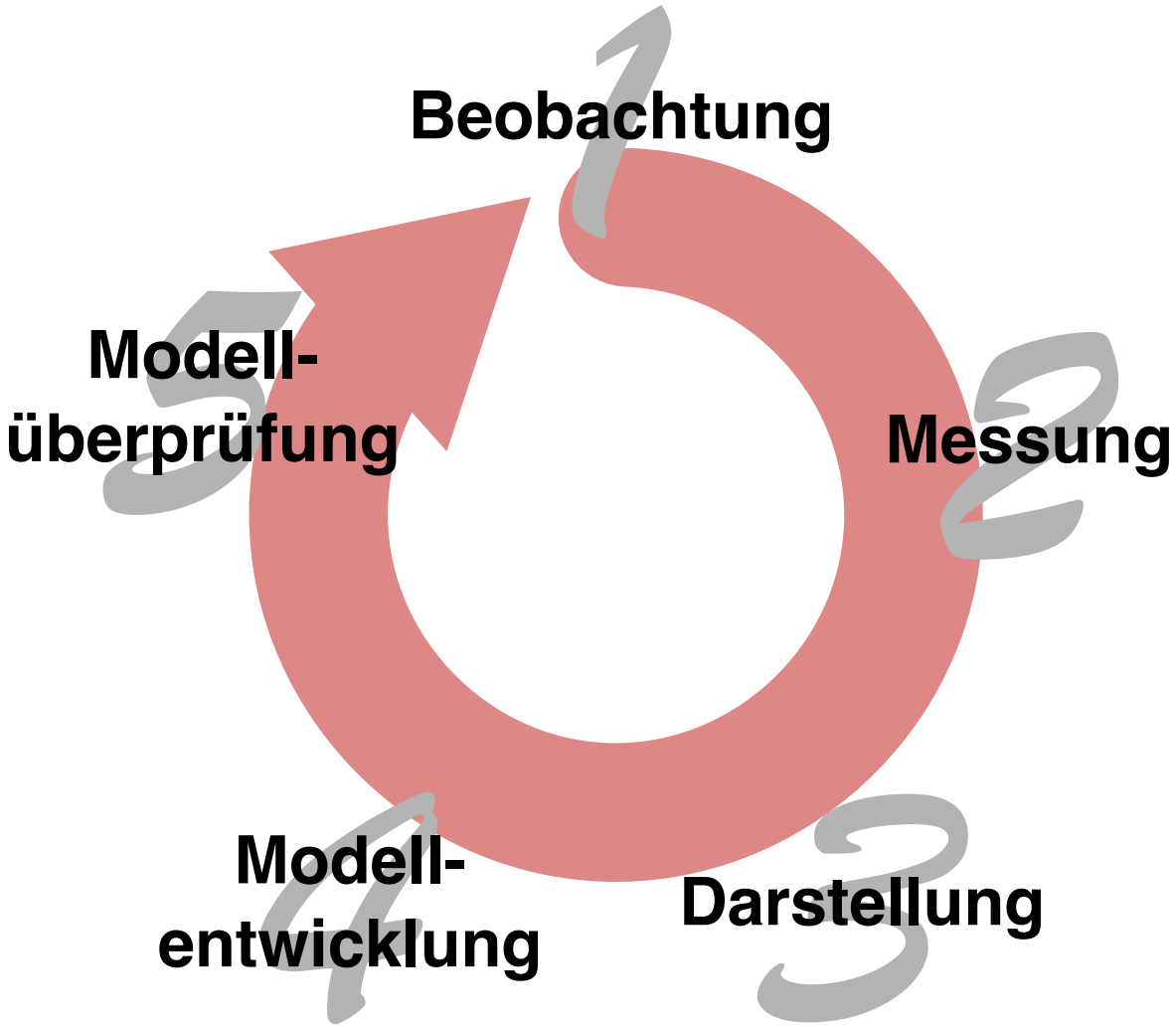
- grob vereinfacht zwei Modelle: formal *vs.* agil
  - von oben herunter *vs.* von unten nach oben
  - Haben beide Stärken und Schwächen: Ziel wichtiger als die Form.

☛ Kernaspekt von Wissenschaft: Lernen von anderen . . .

G. E. P. Box, in: Launer und Wilkinson (Hg.): Robustness in Statistics, Academic Press, New York 1979
R. P. Feynman: Cargo Cult Science. *Engineering and Science* 37(7):10–13, 1974

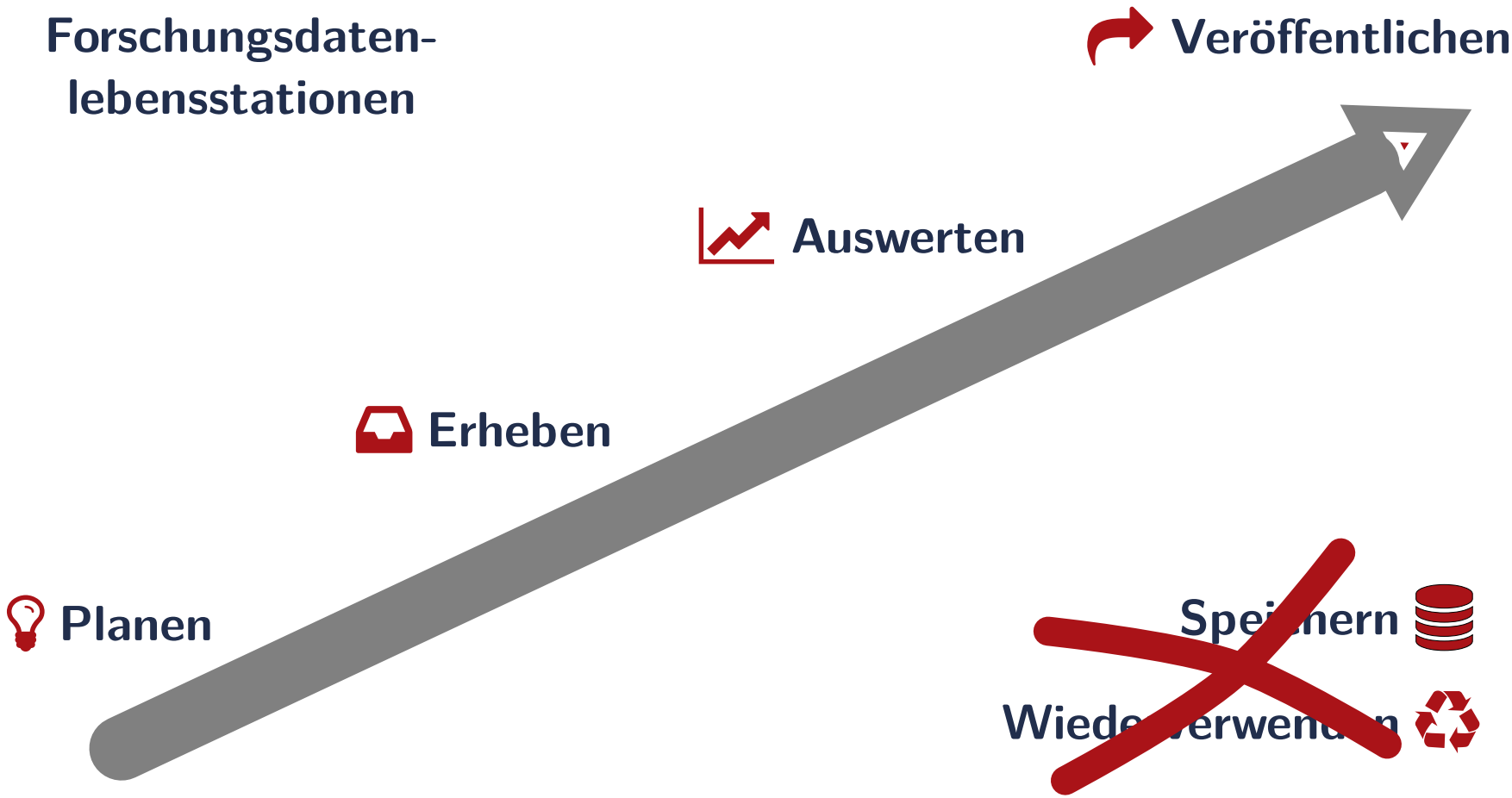
Beispiel für einen tatsächlichen Zyklus

Vorgehen in der experimentellen Wissenschaft



Die Forschungsdatenlebensstationen

Ein treffenderes Bild – da der Zyklus selten geschlossen wird?



Übersicht

Forschungsdaten im Forschungsdatenlebenszyklus

Forschungsdaten verändern ihren Charakter entlang des Zyklus

Erheben

- maximale Datenmenge, ggf. gar nicht komplett speicherbar
- rohe Daten, meist ohne (Vor-)Verarbeitung nicht brauchbar

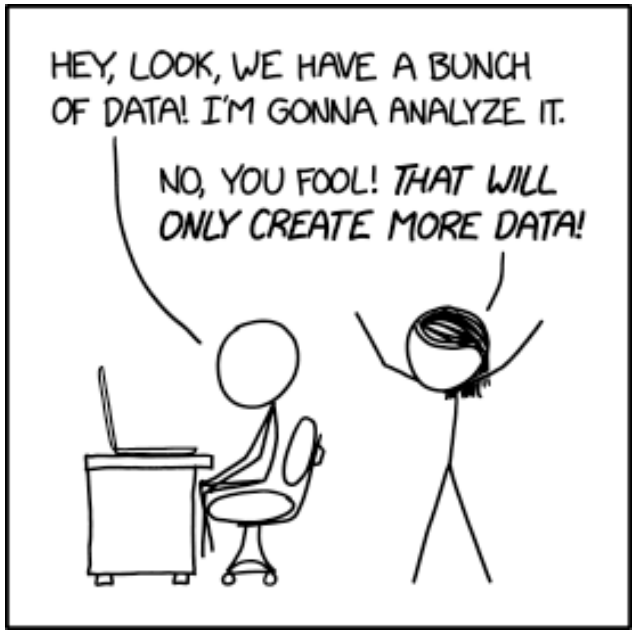
Auswerten

- Reduktion
  - real Daten wegwerfen: ggf. dokumentiert löschen
- Kompression
  - Charakteristika aus Daten extrahieren
- Extraktion
  - Schlussfolgerungen ziehen, idealerweise Erkenntnisse gewinnen

Veröffentlichen

- alle *relevanten* Informationen, nachvollziehbar beschrieben

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.

Forschungsdaten im Forschungsdatenlebenszyklus

Was sind Daten? Was sind Rohdaten? Was ist (wie lange) erhaltenswert?

- Oft einseitiger Fokus auf numerischen Daten
  - Forschungsdaten sind nicht nur numerische Daten.
  - Forschungsdaten liegen nicht nur digital vor.
  - Werkzeuge (z.B. Bibliographien, Fragebögen, Software) sind gleichwertige relevante Forschungsdaten
- Was sind Rohdaten?
  - Viel weniger trivial, als man oft denkt.
  - Widerstandswerte eines Thermometers: eher nein
  - Rohdaten eines Bildsensors (inkl. Artefakte): eher schon
- 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

Forschungssoftware als Forschungsdatum

Abbildung auf den Forschungsdatenlebenszyklus

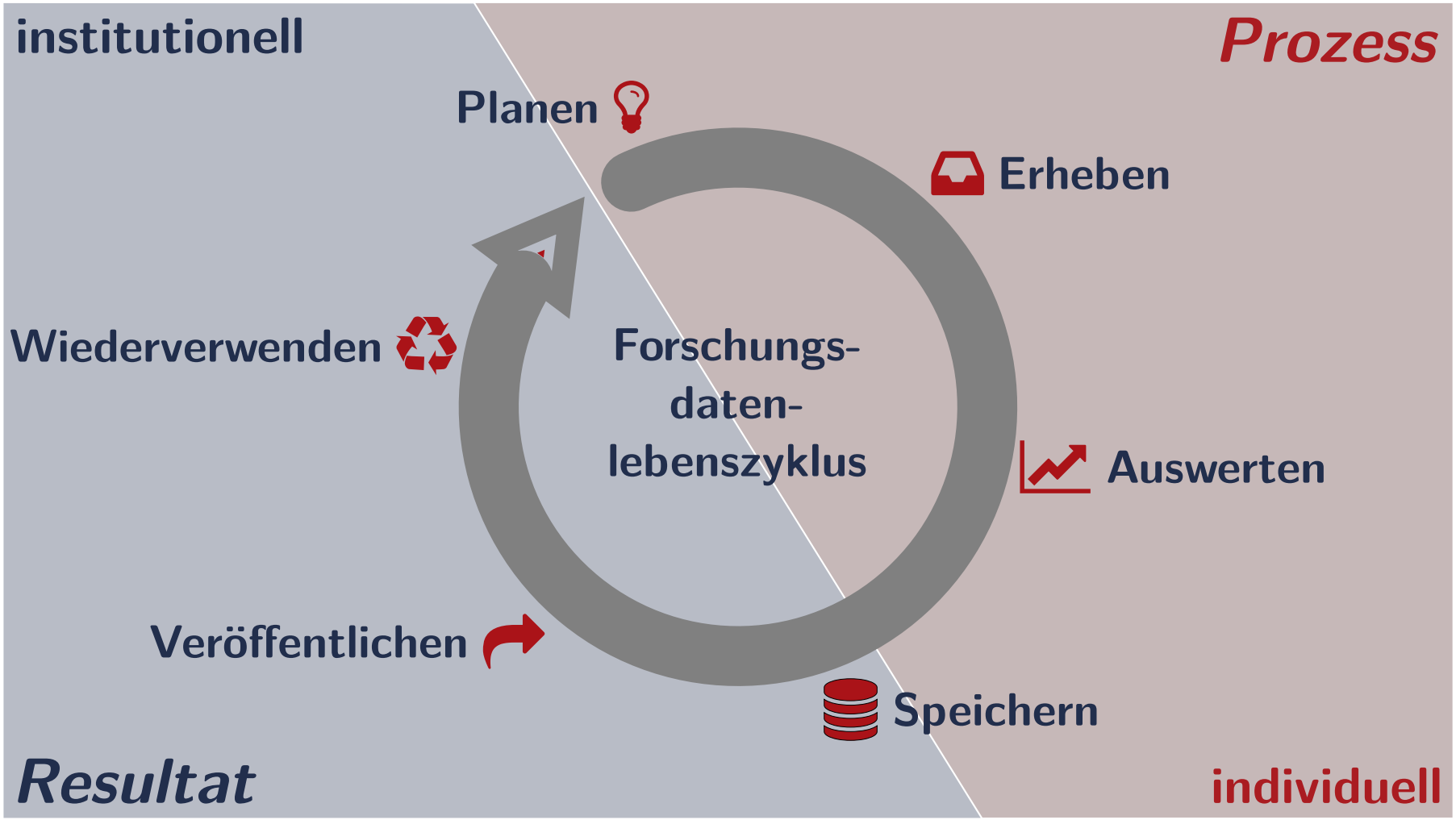
- Planen
  - Softwaremanagementplan – vgl. Softwaretechnik-Literatur
- Erheben
  - Software implementieren
- Auswerten
  - Anwendung der Software zum wissenschaftlichen Erkenntnisgewinn
- Speichern
  - Versionsverwaltungssystem, . . .
- Veröffentlichen
  - GitHub, PyPI etc.; Zenodo; Software-Journale wie JOSS
- Wiederverwenden
  - Voraussetzungen: Dokumentation, Modularität, Qualität, . . .

VL „Wissenschaftliche Softwareentwicklung“: `https://www.till-biskup.de/de/lehre/programmierkonzepte/`

Übersicht

Verantwortlichkeiten . . .

. . . im Forschungsdatenlebenszyklus



Verantwortlichkeiten

Unterscheidung zwischen individuell und institutionell

individuelle Verantwortlichkeiten

- individuelle Forschende (ab B.Sc.)
  - praktisches forschungsnahes Forschungsdatenmanagement
- wissenschaftliche Projektleitung
  - Ermöglichen und Einfordern von Forschungsdatenmanagement

institutionelle Verantwortlichkeiten

- Institutionen
  - Heimatinstitution oder/und übergreifende Institutionen
  - Abstufungen: Institut, Fakultät, zentrale Einrichtungen, Leitung
  - allgemeine Regeln, Basis-Infrastruktur, Ausstattung, Unterstützung
- Fachgesellschaften
  - Schaffung von Standards (statt nur lokaler Konventionen)
  - Bereitstellung u.a. von (kuratierten) Fachrepositorien
  - Langfristigkeit jenseits förderpolitischer Modeerscheinungen

Verantwortlichkeiten

Primär bei den individuellen Forschenden

Leitmotiv

Forschungsdatenmanagement ist primär
die Verantwortung der individuellen Forschenden.

- Prozess-Sicht: Ich bin als forschende Person alleine verantwortlich.
- Resultat-Sicht: Ich brauche i.d.R. Unterstützung durch andere.
- unterschiedliche Rollen innerhalb einer Forschendengruppe
  - Die Leitung muss ein viel größeres Interesse daran haben,
    die institutionellen Aspekte abzudecken,
    als es die real und abhängig Forschenden
    aufgrund ihres i.d.R. kurzen Zeithorizonts haben (können).

☛ Der Forschungsdatenlebenszyklus kann dabei helfen, die richtigen Fragen zu stellen und sich der Verantwortung bewusst zu werden.

Zentrale Aspekte

- Der Forschungsdatenlebenszyklus ermöglicht die Kommunikation mit Forschenden über Forschungsdatenmanagement.
- Die Fokussierung auf Daten und ihre Wiederverwendbarkeit statt Nachvollziehbarkeit ist kontraproduktiv für die Wissenschaft.
- Kernaspekt der wissenschaftlichen Methode ist Nachvollziehbarkeit – und damit Dokumentation auf allen Stufen des Zyklus.
- Die Bedeutung des Begriffs „Forschungsdaten“ ändert sich über die Stationen des Zyklus hinweg.
- Anhand des Forschungsdatenlebenszyklus lassen sich unterschiedliche Verantwortlichkeiten erkennen.