

ELN, LIMS, Repository, Katalog – alles eins, oder doch unterschiedliche Werkzeuge?

Impulsvortrag

Till Biskup (BfR)

Workshop der AG FDM am MRI
Karlsruhe 17.04.2023

Inhaltliche Schwerpunkte

Aus der Ankündigung des Workshops

“ Die digitale Verarbeitung von Forschungsdaten hängt eng mit **geeigneten Software-Lösungen** zusammen, die sowohl die Speicherung der Daten selbst, als auch die Prozesse der Daten-Generierung, Weiterverarbeitung und Publikation bzw. langfristigen Archivierung gewährleisten müssen. Hierbei sind die Institute mit sehr heterogenen Lösungsansätzen konfrontiert, die es erforderlich machen, mehrere Herausforderungen gleichzeitig in den Blick zu nehmen. Einerseits sollen bestehende, gut funktionierende Lösungen integriert werden, **während andererseits eine Fragmentierung in mehrere, parallel betriebene Systeme vermieden werden sollte.**

– Workshop-Ankündigung

- ☛ Heterogene Herausforderungen bedingen oft heterogene Lösungen
- ☛ Geeignete Software-Lösungen ermöglichen, statt einzuschränken

Leitfrage

Aus der Zusammenfassung des Impulsvortrags

❓ Welche Anforderungen müssen sinnvollerweise an eine digitale Infrastruktur gestellt werden, die große Teile des Forschungsdatenlebenszyklus abdeckt?

- ▶ modular
- ▶ flexibel
- ▶ interoperabel
- ▶ einfach bedienbar
- ▶ offensichtliche Vorteile bietend
- ▶ ...

Forschungsdatenlebenszyklus

Digitale Infrastruktur zur Dokumentation jeder einzelnen Station




Forschungsdatenlebenszyklus

Digitale Infrastruktur zur Dokumentation jeder einzelnen Station

Planen

- ▶ Art und Umfang der Daten abschätzen
- ▶ Urheberschaft, Beteiligte, Lizenzen und Schutzrechte klären
- 🔧 digitaler, aktualisierbarer und auslesbarer Datenmanagementplan (z.B. RDMO integriert in Forschungsprojektdatenbank)


Erheben

- ▶ Metadaten *während* der Datenaufnahme erheben
- ▶ Wer hat was mit wem, wann, wie und warum gemacht?
- 🔧 menschengerechtes und maschinenlesbares Metadatenformat (z.B. Infofile,  10.1039/D2DD00131D)


Forschungsdatenlebenszyklus

Digitale Infrastruktur zur Dokumentation jeder einzelnen Station

Auswerten

- ▶ lückenloses Protokoll aller Verarbeitungs- und Analyseschritte
- ▶ (vollständig) reproduzierbare Datenverarbeitung und -analyse
- 🔧 Gesamtsystem zur Datenverarbeitung
(z.B. ASpecD,  10.1002/cmtd.202100097)

Speichern


- ▶ (de)zentraler Speicher mit zentralem Backup
- ▶ Konventionen für Datei- und Verzeichnisnamen oder PIDs
- 🔧 automatischer Datenspeicher mit lokalen PIDs
(z.B. LabInform Datasafe/LOI,  10.26434/chemrxiv-2022-vz360)

ASpecD: J. Popp, T. Biskup. *Chemistry–Methods* 2:e202100097, 2022; LabInform: T. Biskup. *ChemRxiv*, 2022

Forschungsdatenlebenszyklus

Digitale Infrastruktur zur Dokumentation jeder einzelnen Station

Veröffentlichen

- ▶ Beschreibung des zu veröffentlichenden Datenpakets
- ▶ Vollständigkeit: Daten, Dokumentation, Auswertungen, ...
- 🔑 Kuration, automatisiertes Hochladen in Repositorium
(z.B. Checkliste,  10.5281/zenodo.7674307)

Wiederverwenden

- ▶ Überblick über verfügbare Forschungsdaten
- ▶ direkter Link auf Daten, alternativ Kontaktdaten
- 🔑 Repositorien und Kataloge für Forschungsdaten
(z.B. OpenAgrar, CKAN)

Checkliste: C. Odebrecht, T. Biskup. Zenodo, 2023

ELN, LIMS, Repository, Katalog

Alles eins oder doch unterschiedliche Werkzeuge?

- ❓ Wo gehören elektronisches Laborbuch (ELN) und Laborinformations- und -managementsystem (LIMS) in den Forschungsdatenlebenszyklus?
- ❓ Was ist eigentlich ein ELN? Und was ist es *nicht*?
- ❓ Was genau ist ein LIMS, und taugt es auch in der (Grundlagen-)Forschung, außerhalb von Routinelaboren?
- ❓ Was ist der Unterschied zwischen Repository und Katalog?
- ❓ Was ist mit Daten, die (noch) nicht veröffentlicht werden können/sollen?

Definitionen: Elektronisches Laborbuch (ELN)

Mehr als ein Ersatz für das herkömmliche Laborbuch aus Papier?

Softwareprodukt zur Dokumentation der Planung, Durchführung und Auswertung von Labor-Experimenten.

„Elektronische Laborbücher (ELB) sind Softwareprodukte zur Dokumentation der Planung, Durchführung und Auswertung von Labor-Experimenten [. . .]. Der Einsatz von ELB beugt Datenverlust vor und schafft Rechtssicherheit – auch bei Ausscheiden [einer Wissenschaftlerin oder] eines Wissenschaftlers aus der Institution verbleibt eine Version des ELB an der Hochschule.“ [HHU]

Quelle: BfR-FDM-Glossar

- ☞ Eine in vielerlei Hinsicht diskussionswürdige Definition

Was ist ein ELN? Und was ist es *nicht*?

Versuch einer sinnvolleren Definition

essentielle Bestandteile eines ELN

- ▶ Ersatz für das analoge, papierbasierte Laborjournal
- ▶ Zugriff sowohl vom Labor als auch vom Schreibtisch (→ webbasiert)
- ▶ Bilder und Medien
- ▶ strukturierte Inhalte (Schlüssel-Wert-Paare)
- ▶ vorlagenbasierte Erstellung neuer Seiten
- ▶ Anpassbarkeit der Vorlagen über die Nutzerschnittstelle
- ▶ tabellarischer Überblick über Experimente/Messungen mit Möglichkeit der Sortierung und Filterung
- ▶ Export aller Inhalte in ein generisches Format

Was ist ein ELN? Und was ist es *nicht*?

Versuch einer sinnvolleren Definition

optionale Bestandteile eines ELN

- ▶ Inventar für Proben
- ▶ Schnittstelle zu Auswertungsroutinen, um Ergebnisse von Analysen automatisiert ins Laborbuch einzutragen

keine Bestandteile eines ELN

- ▶ Datenspeicher/Repository
- ▶ Forschungsdatenkatalog
- ▶ Management der Messaufbauten^(LIMS?)
- ▶ allgemeines Projektmanagement inkl. Anträge und Publikationen
- ▶ Inventar für Verbrauchsmaterialien^(LIMS?)
- ▶ Workflows für die Datenauswertung

Definitionen: Laborinformations- und -managementsystem (LIMS)

Ein sinnvolles Werkzeug auch außerhalb von Routinelaboren?

Digitale Verwaltung von Daten und Proben in regulierten Umgebungen.

Laborinformations- und -managementsysteme (LIMS) sind EDV-Anwendungen für die Verwaltung von Daten und die Unterstützung von Arbeitsabläufen in probenorientiert arbeitenden Laboren. LIMS unterstützen die Bearbeitung der Proben und die damit verbundenen Arbeitsabläufe. Sie bieten eine transparente Verfolgung der Proben über den gesamten Bearbeitungszyklus der Proben im Labor, gestalten den Laborbetrieb effizient und gewährleisten angemessenes Qualitätsmanagement in regulierten Umgebungen.

Quelle: Wikipedia

Laborinformations- und -managementsystem (LIMS)

Ein sinnvolles Werkzeug auch außerhalb von Routinelaboren?

Eigenschaften konventioneller LIMS

- ▶ Entworfen für regulierte Umgebungen mit bekannten, starren Abläufen
- ▶ Fokus auf Qualitätskontrolle und lückenloser Nachverfolgbarkeit
- ▶ Integration von Messaufbauten und deren Daten
- ▶ Hochspezifisch, Anpassung meist nur durch (externe) Programmierung

LIMS im Forschungskontext

- ▶ Management der Messaufbauten (Belegung, Logbuch, Dokumentation, ...)
- ▶ Inventar für Verbrauchsmaterialien (und Proben)
- ☞ Fokus auf Flexibilität und Anpassbarkeit (durch die Nutzer)

ELN und LIMS im Forschungsdatenlebenszyklus

Auf der Prozess-Seite und nur bei wenigen Stationen

Elektronisches Laborbuch (ELN)

Erheben

- Wer hat was mit wem, wann, wie und warum gemacht?

Laborinformations- und -managementsystem (LIMS)

Planen

- Verfügbarkeit von Geräten
- Inventar von Verbrauchsmaterial (und Proben)

Erheben

- Nachverfolgbarkeit des Untersuchungsmaterials (Proben)

Definitionen: Repository

Ein Ort für die langfristige zugängliche Ablage von Forschungsdaten

Publikationsplattform für Forschungsdaten.

Repositorien sind Publikationsplattformen für Forschungsdaten. Dieser IT-Dienst speichert die Forschungsdaten in der Regel langfristig, dokumentiert die Forschungsdaten mit Metadaten, regelt den Zugang (inkl. Lizenz) zu den Forschungsdaten und vergibt einen PID. Die dort publizierten Forschungsdaten sind meist über eine Metadatensuche und -filterung für Nutzerinnen und Nutzer auffindbar und erschließbar (Datenkatalog).

Quelle: BfR-FDM-Glossar

vgl. dazu <https://www.forschungsdaten.org/index.php/Repositoryum> (besucht am 09.03.2023)

Definitionen: (Daten)Katalog

Überblick über vorhandene Forschungsdaten

Werkzeug zum Auffinden und Erschließen von Forschungsdaten.

Forschungsdaten können mit Hilfe eines Datenkatalogs gesucht, gefunden und erschlossen werden (vgl. FAIR). Ein Datenkatalog enthält vergleichbar zu einem Bibliothekskatalog verschiedene Metadaten, die die Grundlage für die Suche und Filterung darstellen, aber nicht notwendigerweise die Forschungsdaten selbst – im Falle der Bibliothek die Bücher. Solche grundständigen Katalogfunktionen bieten typischerweise auch Repositorien. Ein Katalog (als Sammlung von Metadaten zu bestimmten Objekten) erweist sich als sinnvoll, wenn die Menge der Objekte eine gewisse Schwelle überschreitet, die ein Auffinden und Abrufen (*retrieval*) über die einzelnen Objekte selbst unmöglich macht [vgl. Haynes 2018].

Quelle: BfR-FDM-Glossar

Repository und (Daten)Katalog

Wesentliche Unterschiede

Repository



- ▶ tatsächlicher Ablageort für die Daten
- ▶ weist jedem Datum eine eindeutige Adresse zu
 - idealerweise eine dauerhafte Kennung (*persistent identifier*, PID)
- ▶ kümmert sich um die langfristige Speicherung (Archivierung)
 - Maßnahmen gegen Datenverlust und Datenkorruption (z.B. Prüfsummen)

(Daten)Katalog

- ▶ enthält die Metadaten zu den (in einem Repository liegenden) Daten
 - Metadaten enthalten i.d.R. Verweis auf Ablageort der Daten
- ▶ erlaubt (komplexe) Suche und Filterung der Suchergebnisse

Repository und (Daten)Katalog

Ein paar Gedanken aus der eigenen langjährigen Forschererfahrung

- ▶ Repository und Katalog werden in der Praxis häufig nicht getrennt
 - Repositorien stellen Katalogfunktionen (Suche, Filterung) bereit
 - Kataloge enthalten (mitunter) Daten, nicht nur Metadaten
- ▶ Repository und Katalog werden oft mit Veröffentlichung gleichgesetzt
 - i.d.R. wird nur ein Bruchteil der Daten veröffentlicht
 - Veröffentlichung ist (zurecht) oft ein Prozess mit langwierigen Vorarbeiten
 - *Open Science* ist ein netter Gedanke, aber oft abschreckend/unrealistisch
- ▶ Es fehlt meist an lokalen Repositorien und Katalogen
 - Überblick über und Zugriff auf die eigenen Forschungsdaten ist zentral für  Planung und  Auswertung
 - Voraussetzung: lokale PIDs, Repositorien und Kataloge für alle lokal vorhandenen Daten (unabhängig von Veröffentlichung)











Weitere Aspekte einer digitalen Forschungs-Infrastruktur

Jenseits von ELN, LIMS, Repositorien und Katalogen

- ▶ Metadaten *während* der Datenerhebung
- ▶ Gesamtsystem zur wissenschaftlichen Datenauswertung
- ▶ Repositorium für „warme“ Forschungsdaten
- ▶ lokale PIDs
- ▶ lokales Wissensmanagement
- ▶ Planungswerkzeuge für Projekte, Publikationen etc.
- ▶ Versionsverwaltung für Dokumente und Software

Aspekte einer digitalen Forschungs-Infrastruktur

Eine Reihe modularer Werkzeuge

- ▶ Infofile 
 - Metadaten während der Datenerhebung
 -  Paulus und Biskup, *Digital Discovery* **2**:234–244, 2023.
- ▶ ASpecD 
 - Gesamtsystem zur wissenschaftlichen Datenauswertung
 -  Popp und Biskup, *Chemistry–Methods* **2**:e202100097, 2022.
- ▶ LabInform ELN 
 - modulares ELN auf Basis von DokuWiki
 -  Schröder und Biskup, *ChemRxiv*, 2023.
- ▶ LabInform   
 - modulares LIMS mit Repository (Datasafe), PIDs (LOI), Wiki
 -  Biskup, *ChemRxiv*, 2022.

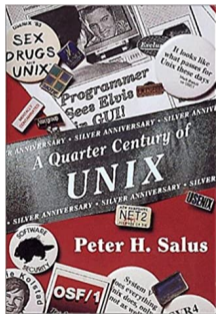
Zurück zur Leitfrage: sinnvolle Anforderungen an digitale Infrastruktur

Eine jahrzehntealte – und bewährte – Antwort

- ❓ Welche Anforderungen müssen sinnvollerweise an eine digitale Infrastruktur gestellt werden, die große Teile des Forschungsdatenlebenszyklus abdeckt?

The Unix Philosophy

- ▶ 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.



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

Zusammenfassung

Eine Reihe provokativer Thesen zur Diskussion

- 📣 Es gibt keine „eierlegende Wollmilchsau“.
Systeme, die das versuchen zu sein, sind zum Scheitern verurteilt.
- 📣 Ein unstrukturiertes papierbasiertes Laborbuch zu digitalisieren hilft wenig.
Struktur ist wichtiger für Informationen als Digitalität.
- 📣 Struktur entsteht aus der intellektuellen Durchdringung von Abläufen.
Das setzt eigene Forschungserfahrung und analytisches Denken voraus.
- 📣 Es gibt keine schlüsselfertigen Lösungen „von der Stange“.
Jedes System muss an die spezifischen Bedarfe angepasst werden.
- 📣 Nur Systeme, die hinreichend einfach nutzbar sind und deren Verwendung offensichtliche Vorteile bietet, werden genutzt werden.
- 📣 Die Heterogenität der Anforderungen erfordert ggf. den parallelen Einsatz unterschiedlicher Systeme im selben Haus.

Danke für Ihre Aufmerksamkeit

Till Biskup (BfR)

Bundesinstitut für Risikobewertung
Max-Dohrn-Straße 8-10 • 10509 Berlin
Telefon 030 - 184 12 - 0 • Fax 030 - 184 12 - 99 0 99
bfr@bfr.bund.de • www.bfr.bund.de