Forschungsdatenmanagement

Notwendige, aber nicht hinreichende Voraussetzung für den wissenschaftlichen Erkenntnisgewinn

21. Modulare digitale (dezentrale) Infrastruktur

Dr. habil. Till Biskup
Physikalische Chemie
Universität Rostock
19.07.2024

Zentrale Aspekte

- Forschungsdatenmanagement umfasst viele weitere Aspekte neben Metadaten, Protokoll der Verarbeitung und ELN.
- Essentiell sind lokale PIDs, lokale Repositorien, Versionsverwaltung und Werkzeuge zum Wissens- und Projektmanagement.
- Verantwortung für Inhalte, Qualitätssicherung und Strukturen liegt bei den Forschenden, nicht bei Institutionen.
- Schlüssel zum Erfolg einer digitalen Infrastruktur sind Modularität, Dezentralität und bewährte Komponenten.
- Nur dezentrale Lösungen sind erfolgsversprechend.
  Die notwendige Infrastruktur ist von Gruppen handhabbar.

Übersicht

Problemstellung, Herausforderungen und Anforderungen

Konzept und Umsetzung

Anwendung auf die eigene Praxis

Vorteile gegenüber Alternativen

Problemstellung

- Alle vorher vorgestellten Aspekte (Metadaten während der Datenaufnahme, nachvollziehbare Datenauswertung, ELN) reichen nicht aus, um wirklich nachvollziehbare Forschung zu machen.

Herausforderungen

- Wissenschaft, insbesondere Grundlagenforschung,
  ist viel zu divers und unvorhersehbar,
  als dass sie sich in ein systemisches Korsett zwängen ließe.
- Die Komplexität der Fragestellung und der Abläufe erforderte eigentlich professionelle Programmierer und IT-Administratoren – beides ist aber im akademischen Bereich quasi nicht zu haben (zumal der öffentliche Dienst auch viel zu schlecht bezahlt).

Anforderungen

- Wissenschaft ermöglichen, nicht behindern (kein zu enges Korsett)
- Weitgehend digitale Abläufe, Ausnutzung der Möglichkeiten, die sich durch eine zunehmende Digitalisierung auch in den Wissenschaften und der Grundlagenforschung bieten.
- Modulares System, das sich sowohl schrittweise einführen als auch von Einzelpersonen beherrschen und intellektuell ausreichend durchdringen lässt, um nebenher wartbar zu sein.
- Interoperabilität mit der jeweiligen institutionellen IT-Infrastruktur.
- Keine Abhängigkeit von einzelnen Herstellern (*vendor lock-in*), Möglichkeit der hinreichenden Kapselung und Mitnahme von Daten und Inhalten (Eigenständigkeit von Forschenden).

Kontext

- wissenschaftlicher Kernaspekt
  - Nachvollziehbarkeit
- Stationen im Forschungsdatenlebenszyklus
  - Planung, Datenerhebung Speicherung, Veröffentlichung
- Bedeutung
  - Nachvollziehbarkeit der Datenerhebung
  - Ausgangspunkt für weitere Planungen

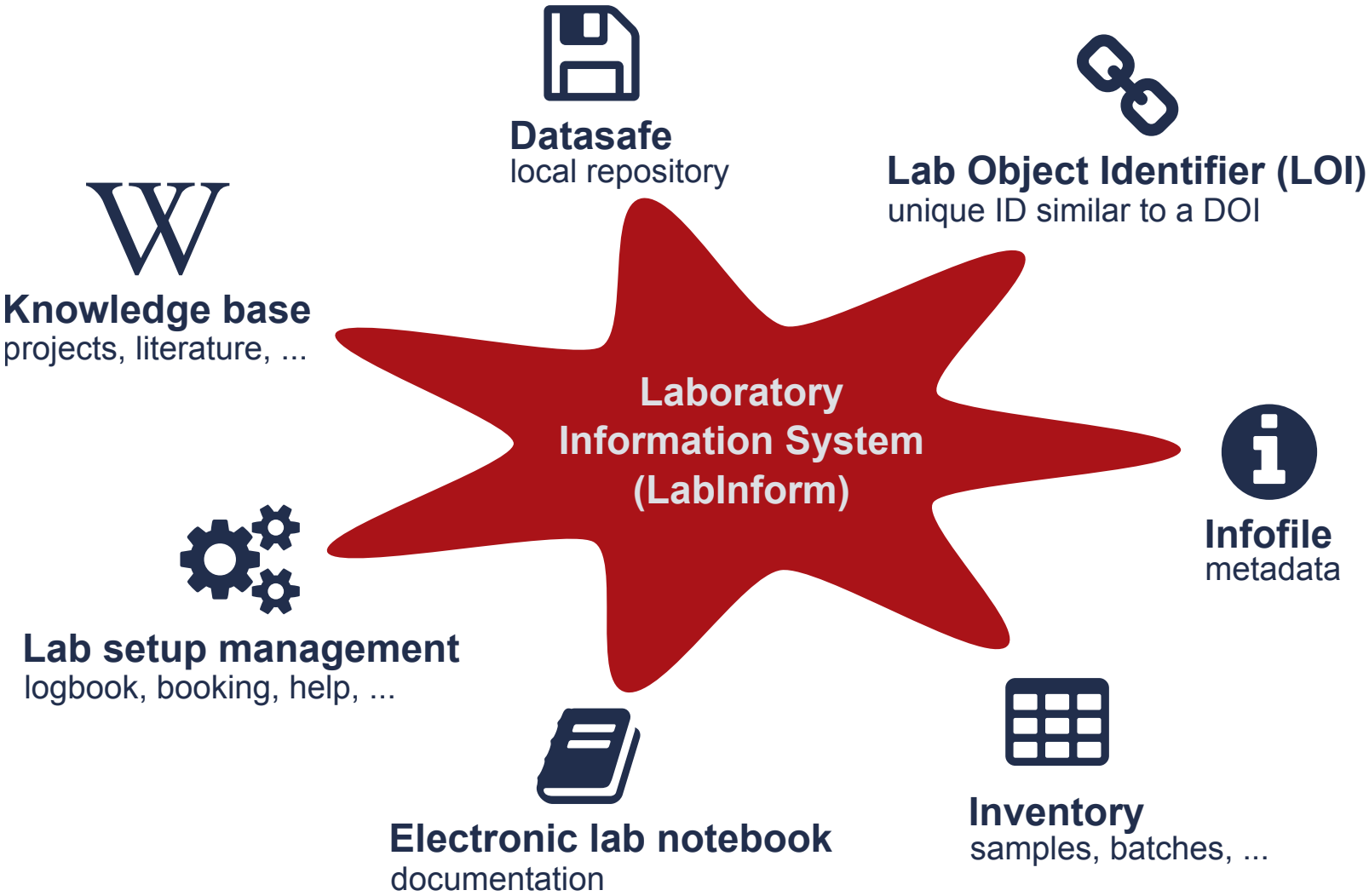
Beitrag

- digitale, strukturierte Unterstützung der Forschung
  - intuitiv aufgrund intellektueller Durchdringung der Abläufe
- flexibles, von den Nutzenden weitestgehend anpassbares System

Übersicht

Konzept und Umsetzung: LabInform

Eine erste Übersicht über zentrale Komponenten



vgl. T. Biskup, *ChemRxiv* 2022, doi:10.26434/chemrxiv-2022-vz360

Konzept und Umsetzung: Datasafe

Lokales Repositorium für „warme“ Forschungsdaten

Problemstellung

- (Roh-)Daten müssen auffindbar und zugreifbar abgelegt werden. Zielgruppe sind die Mitglieder der eigenen Arbeitsgruppe.
- Eine Veröffentlichung ist meist weder sinnvoll noch möglich, da es sich um „warme“, weitgehend unkuratierte Daten handelt.

Herausforderungen und Anforderungen

- Zugriff über eindeutige und dauerhafte Kennung (UID/PID)
- Ablage perspektivisch auf einem Netzwerklaufwerk
- einfacher Zugriff bei Datenerhebung (*push*) und Auswertung (*pull*)
- Speicherung von Daten und zugehörigen Metadaten
- Überprüfung der Unversehrtheit von Daten und Metadaten
- plattformunabhängig

Lösung: Kernaspekte des LabInform Datasafe

- Client-Server-Architektur
  - lokal genauso nutzbar wie über ein Netzwerk
  - plattformunabhängig: Client in Python implementiert
- Prüfsummen (*hashes*) zur Überprüfung der Unversehrtheit
  - zwei Prüfsummen: Metadaten; Metadaten & Daten
  - Metadaten ändern sich häufiger/schneller als Daten
  - Client erzeugt Prüfsummen vor dem Hochladen, Server überprüft
  - Client überprüft Prüfsummen nach dem Herunterladen
- Zugriff über eindeutige, dauerhafte Kennungen (LOI)
  - können *vor* dem Hochladen registriert (und verwendet) werden
- Datenablage intern (*backend*) in einer Verzeichnishierarchie
  - robust und resilient, einfach zu sichern (Backup)
  - Datenbank nur zur schnelleren Suche, automatisch generierbar

`https://datasafe.docs.labinform.de/`

Konzept und Umsetzung: LOI

Lokale dauerhafte und eindeutige Kennungen (PID/UID)

Problemstellung

- eindeutige Identifizierung von (Roh-)Daten, Proben, ...
  – letztlich allen relevanten „Objekten“ der Forschung
- Zugriff auf Informationen unabhängig von ihrem Speicherort
  über eine eindeutige und möglichst dauerhafte Kennung

Herausforderungen und Anforderungen

- Pfade im Dateisystem sind meist nicht ausreichend langzeitstabil.
- Nicht alle Informationen sind Dateien auf einer Festplatte.
- Kennungen sollten möglichst sprechend und ableitbar sein:
  eine kryptographische Prüfsumme ist eindeutig, aber unleserlich.
- lokale Vergabe ohne Abhängigkeit von zentralen Stellen
- sollte in Kooperationen über Gruppengrenzen hinweg funktionieren

Lösung: Kernaspekte der LabInform LOIs

- Schema vergleichbar dem *Digital Object Identifier* (DOI)
  - aber: keine zentrale Instanz und keine Kosten
- zweigeteilt: Herausgeber und eindeutige Kennung
  - erlaubt vollkommene Freiheit bei der Organisation der Kennung
- hierarchische Gliederung der Kennung
  - „sprechend“, manuell ableitbar, „*hackable*“
  - Hierarchieebenen komplett im Belieben der Herausgebenden
  - Datasafe: 1:1-Abbildung der Verzeichnishierarchie auf LOIs

Listing 1: Beispiele für LOIs

```
1 loi:42.<publisher>/<ID>              # allgemeines Schema

3 loi:42.1001/ds/sa/1/trepr/2          # Datensatz im Datensafe (Probe #1, Messung #2)

5 loi:42.1001/lb/sa/1                  # Informationen zur Probe (im ELN-Inventar)
```

im Prinzip: `https://datasafe.docs.labinform.de/api/datasafe.loi.html`

Problemstellung

- Wissenschaft findet immer in einem Kontext statt, der erarbeitet und strukturiert werden muss.
- Erfolgreiche Projekte erfordern ein Mindestmaß an Planung und einfache Übersichten über alle relevanten Informationen.

Herausforderungen und Anforderungen

- Strukturierte Ablage von Kontextwissen erfordert hinreichend mächtige Textauszeichnung, Referenzen und Medien (Bilder etc.).
- Wissens- und Projektmanagement erfordert flexible Formate, da sich Informationen häufig ändern und aktuell bleiben müssen.
- (hierarchische) Strukturierung, Durchsuchbarkeit und einfacher ortsunabhängiger Zugriff (schreibend und lesend)

Lösung: Kernaspekte der LabInform Knowledge Base

- wiki-basiertes System (DokuWiki): hierarchisch gegliedert, plattformunabhängig, browserbasiert, robust und resilient
- einfache, aber mächtige und flexibel erweiterbare Syntax zur Textauszeichnung inkl. Bibliographien und Medieneinbindung

Lösung: Kernaspekte des LabInform Projektmanagements

- Formulare und strukturierte Vorlagen zur Erstellung von Einträgen
- strukturierte Schlüssel–Wert-Paare zur Aggregation von Übersichtstabellen und zur Kategorisierung
- in Verbindung mit dem LabInform ELN Querverweise auf zu einem Projekt gehörendes Probeninventar und Messungen

im Prinzip: `https://demo.wiki.labinform.de/`

Konzept und Umsetzung: Versionsverwaltung

Für Software und Dokumente aller Art

Problemstellung

- inkrementelles Arbeiten an Texten mit mehreren Personen und/oder von unterschiedlichen Rechnern aus
- Entwicklung von Software, die länger als zwei Wochen verwendet werden soll und deren Entwicklung nachvollziehbar sein muss

Herausforderungen und Anforderungen

- dezentral: muss ohne Netzwerkverbindung lokal funktionieren
- plattformunabhängig, verbreitet, langfristig verfügbar
- Serverkomponente für den einfachen Abgleich einzelner Instanzen
- Web-Oberfläche für die einfache(re) Verwaltung komplexer Abläufe

☛ Es gibt etablierte Lösungen aus der Softwareentwicklung!

Lösung: Git als Versionsverwaltung

- verteiltes Versionsverwaltungssystem
  - Alle Operationen sind rein lokal und damit netzwerkunabhängig.
  - Abgleich mit beliebigen anderen Instanzen ist einfach möglich.
- weit verbreitet, robust, quelloffen, *de facto*-Standard
  - seit 2004 aktiv entwickelt, abwärtskompatibel zu SVN

Lösung: zentrale Git-Instanz mit Web-Oberfläche

- Installation auf einem erreichbaren, permanent laufenden Server
  - Abgleich zwischen Instanzen unabhängig davon möglich, ob der jeweilige Arbeitsrechner gerade angeschaltet ist
- Web-Oberfläche zur Verwaltung
  - vereinfacht viele Prozesse, verschafft einfachen Überblick
  - Lösungen: Gitea bzw. Forgejo, alternativ GitLab

Git: `https://git-scm.com/`; **Web-UIs**: `https://gitlab.org/`, `https://gitea.io/`, `https://forgejo.org/`

Übersicht

Anwendung auf die eigene Praxis

Ein paar Ideen

- Datasafe: Repositorium für „warme“ Forschungsdaten
  - hierarchische Verzeichnisstruktur zur Ablage von Messdaten überlegen (Bestandteile: Methode, Probe, Messung)
- LOI: lokale dauerhafte und eindeutige Kennungen
  - Schema(ta) für Kennung überlegen (vgl. Datasafe)
  - Tipp: mit Messdaten anfangen
- Knowledge Base: Wissens- und Projektmanagement
  - DokuWiki mit geeigneten Plugins (lokal) installieren/ausprobieren
  - eigene Strukturen für Wissens- und Projektmanagement etablieren
- Versionsverwaltung
  - Git lokal installieren
  - einführende Kapitel des Git-Buches lesen
  - Git in der täglichen Arbeit verwenden (*learning by doing*)
  - Git-Weboberfläche lokal installieren oder vom RZ nutzen

SOLVed-IT
Small Organisation Linux-based Virtualised IT

IT stack for small scientific work groups fostering digital methods and processes

Die Idee

- digitaler Softwarestack für wissenschaftliche Arbeitsgruppen
- Bereitstellung auf lokaler Infrastruktur
- weitgehend automatisiertes Setup

Kernkomponenten

- LabInform ELN
- LabInform Datasafe
- LabInform
- Git Web-UI

☛ dezentrale robuste Infrastruktur, die von einer wissenschaftlichen Arbeitsgruppe ohne externe Hilfe administriert werden kann

https://www.solved-it.org/

Übersicht

Mögliche Alternativen

- keine digitale Infrastruktur (oft *Status quo*)
- zentral bereitgestellte einheitliche Systeme

These

Zentrale (einheitliche) Lösungen sind zum Scheitern verurteilt,
weil sie die komplexe Realität nicht berücksichtigen (können).

- Wissenschaft ist sowohl viel zu komplex als auch viel zu individuell, als dass generelle Lösungen ohne tiefgreifende Anpassungen an die realen Gegebenheiten nutzbar wären.
- Funktionierende Werkzeuge und Lösungen müssen dezentral, modular, flexibel und erweiterbar sein.

Vorteile gegenüber Alternativen

Vorteile einer dezentralen, von Forschenden handhabbaren IT-Infrastruktur

- Flexibilität, Modularität, Anpassbarkeit, Erweiterbarkeit
  - Voraussetzung für die gewinnbringende Nutzung
  - Anforderungen müssen von den Forschenden formuliert werden.
- schrittweise Einführbarkeit priorisiert nach lokalem Bedarf
  - nicht alles kann und muss gleichzeitig implementiert werden
- Qualität der eingesetzten Bausteine
  - Verwendung robuster, bewährter Lösungen (DokuWiki, Git, . . .)
  - Eigenentwicklungen berücksichtigen explizit die Anforderungen der Softwareentwicklung: Codequalität und langfristige Wartbarkeit.
- geringere Angriffsfläche als zentral administrierte Lösungen
  - Sind die Rechenzentren der Universitäten (perspektivisch) personell und hinsichtlich der vorhandenen Kompetenzen dafür aufgestellt, kritische Infrastrukturen ausreichend sicher zu betreiben?

Zentrale Aspekte

- Forschungsdatenmanagement umfasst viele weitere Aspekte neben Metadaten, Protokoll der Verarbeitung und ELN.
- Essentiell sind lokale PIDs, lokale Repositorien, Versionsverwaltung und Werkzeuge zum Wissens- und Projektmanagement.
- Verantwortung für Inhalte, Qualitätssicherung und Strukturen liegt bei den Forschenden, nicht bei Institutionen.
- Schlüssel zum Erfolg einer digitalen Infrastruktur sind Modularität, Dezentralität und bewährte Komponenten.
- Nur dezentrale Lösungen sind erfolgsversprechend.
  Die notwendige Infrastruktur ist von Gruppen handhabbar.