

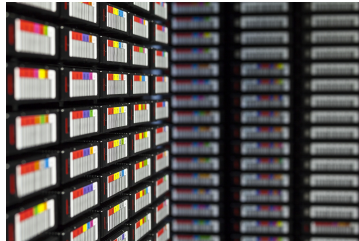


Anforderungen der Ingenieurwissenschaften an ein institutionelles Forschungsdatenrepositorium der Universität Stuttgart

Dorothea Iglezak^{*} und Björn Schembera⁺

^{*}Universitätsbibliothek Stuttgart, ⁺Hochleistungsrechenzentrum Stuttgart

8. Workshops der DINI-/nestor-AG Forschungsdaten "Forschungsdatenrepositorien"
Stuttgart, 27. November 2017





Outline

Einführung und Motivation

Anforderungen aus der Bedarfsanalyse

- Große Datenmengen
- Fachspezifische Metadaten
- Code und Daten
- Förderung der Veröffentlichungsbereitschaft
 - Lokales Datenmanagement als Voraussetzung
 - Datenqualität
- Organisatorische Anforderungen

Ein ideales Repositorium für die Ingenieurwissenschaften



Einführung und Motivation

Einführung - Projekt DIPL-ING

- ▶ DIPL-ING: **D**atenmanagement in **I**nfrastrukturen, **P**rozessen und **L**ebenszyklen in den **I**ngenieurwissenschaften
- ▶ BMBF-Projekt 4/2017 - 3/2019
 - ▶ Projektpartner HLRS, UB, IAG, ITT, TIK
- ▶ Ziel: Konzepte für FDM in den Ingenieurwissenschaften
- ▶ Stand: Bedarfsanalyse, Metadaten
- ▶ kein institutionelles Repositorium



Einführung - Daten in den Ingenieurwissenschaften

- ▶ Beispiel Datenproduktion auf CRAY XC40@HLRS:
 - ▶ Jedes Simulationsprojekt erzeugt
 - ▶ ca. 1000 – 10000 Dateien
 - ▶ jeweils 10MB – 100GB gross
 - ▶ Σ mehrere 100TB/Projekt
 - ▶ Daten werden auf 13PB Lustre Dateisystem geschrieben
 - ▶ Dieses ist "überbucht": Auslagerung auf auf 20PB Bandarchivsystem möglich
- ▶ Beispiel: Thermodynamik-Simulation
 - ▶ Trajektorien molekularer Systeme
 - ▶ Berechnung der Trajektorien auf Compute-Knoten
 - ▶ Ausgabedateien werden auf paralleles Dateisystem geschrieben
 - ▶ Wissen wird durch die Auswertung der Ausgabedateien gewonnen

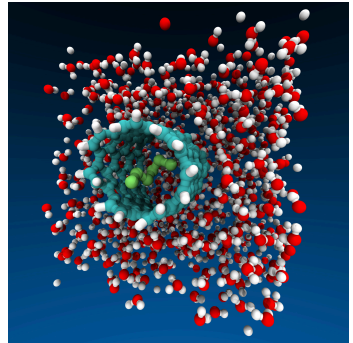


Figure: Visualisierung einer Simulation von Nanoröhren. Trajektorien der Moleküle können basierend auf Computermodellen berechnet werden. (Gernot Bauer/ITT)



Anforderungen aus der Bedarfsanalyse

Große Datenmengen - Problem

Problem/Status Quo

- ▶ Datenmenge wächst stetig
- ▶ Einzelne Dateien können mehrere 100 GB groß sein

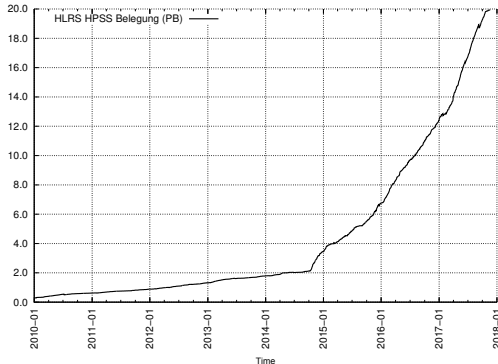
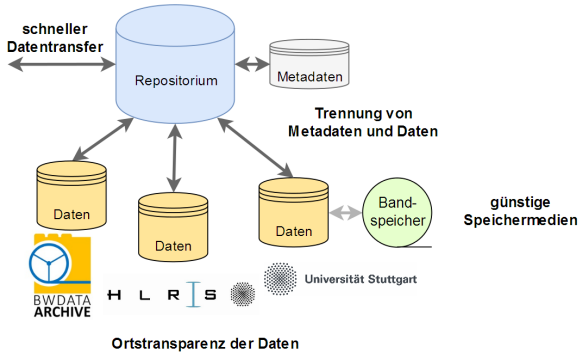


Figure: Belegung des HLRS Bandspeichersystems zur Datenauslagerung (HSM), 2010-2017

Große Datenmengen - Anforderungen

Anforderungen

- ▶ Trennung von Daten und Metadaten
- ▶ Orts- und Verteilungstransparenz
- ▶ Backendspeicher als Tape (viel Platz!)
- ▶ Schnelle Transfermöglichkeiten (z.B. 50TB via 1GB/s (zu) langsam)

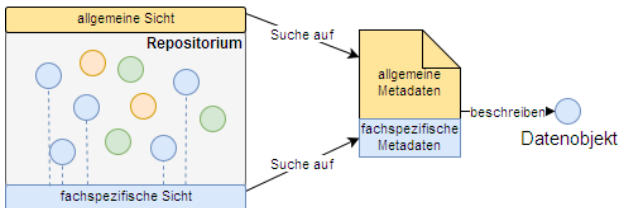


Metadaten

Problem/Status Quo

- ▶ Institutionelles Repository braucht allgemeinen Metadatenkanon
- ▶ Fachspezifische Metadaten sind wichtige Suchkriterien

Anforderungen

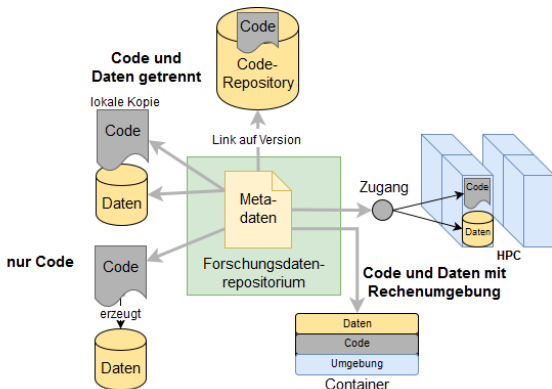


Code und Daten

Problem/Status Quo

- ▶ Code ist Grundlage der Datenerzeugung
- ▶ Abhängigkeit von der Rechenumgebung
- ▶ Ablage in Repositorien oder in institutsinternen Dateisystemen
- ▶ Viel technisches Verständnis zum Reproduzieren von Ergebnissen notwendig

Anforderungen



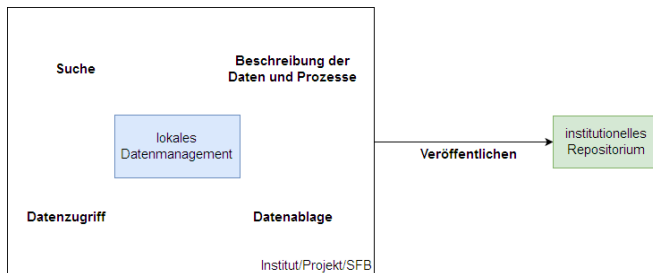
Lokales Datenmanagement als Voraussetzung

Problem/Status Quo

- ▶ geringe Veröffentlichungsbereitschaft
- ▶ lokales Datenmanagement vordringlicheres Problem

Anforderungen

- ▶ zuerst lokales Datenmanagement lösen

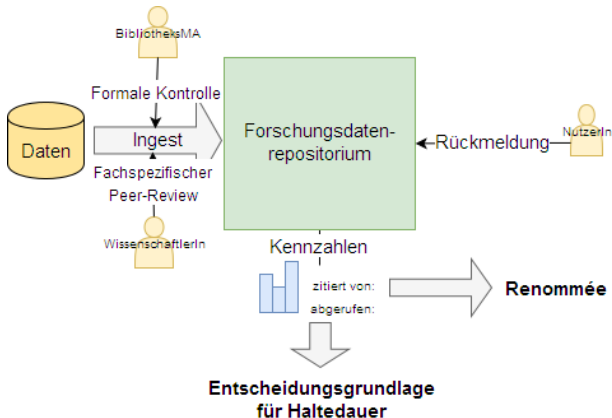


Datenqualität

Problem/Status Quo

- ▶ Anerkennung für die wissenschaftliche Arbeit als Anreiz für Veröffentlichung
- ▶ Qualitätseinschätzung für Daten als Entscheidungsgrundlage für Haltedauer notwendig

Anforderungen





Organisatorische Anforderungen

Problem/Status Quo

- ▶ Handwerkliches Wissen zum Datenmanagement fehlt (Aufbereitung, Pflege)
- ▶ Richtlinien stehen Praxis entgegen

Empfehlung 7: Sicherung und Aufbewahrung von Primärdaten

Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, zehn Jahre lang aufbewahrt werden.

Figure: 7. Empfehlung der DFG zur Sicherung guter wissenschaftlicher Praxis.

Anforderungen

- ▶ Integration in Ingenieurausbildung
- ▶ Scientific Data Officer /Datenkurator

Ein ideales Repository für die Ingenieurwissenschaften

