Designing an institutional research data management infrastructure for the life sciences

Paul van Schayck
PhD student, data steward
Maastricht University Medical Center+
p.vanschayck@maastrichtuniversity.nl
https://datahub.mumc.maastrichtuniversity.nl

Peter Debyelaan 15, 6229 HX Maastricht
P.O. Box 616, 6200 MD Maastricht
The Netherlands
Designing an institutional research data management infrastructure for the life sciences

Life Sciences Faculty

- Independent research groups
- Heterogeneous (meta)data
- Right incentives

Academic Hospital

- Patient privacy
- Electronic Health Records
- Bridging organisations

providing Research Data Management services for
Life science background

*Life science* depends more and more on the collection and analysis of *comprehensive datasets*.

‘Small Science’. Life science is performed in small temporary project groups.

*Open Science*. There is an urgent call for more open, transparent and reproducible science.
<table>
<thead>
<tr>
<th>DataHub characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FAIR</strong>-inspired from start.</td>
</tr>
<tr>
<td><strong>Open-source</strong> where possible.</td>
</tr>
<tr>
<td>(Meta)data <strong>structuring</strong> + ontology <strong>enrichment</strong>.</td>
</tr>
<tr>
<td><strong>Project data structuring</strong>; Hierarchical organisation in projects and datasets.</td>
</tr>
<tr>
<td><strong>Faceted search</strong>, Lucene &amp; ontology-powered, authorization controlled</td>
</tr>
<tr>
<td><strong>High volume</strong>; The infrastructure has been designed and tested with petabyte scale and high throughput in mind.</td>
</tr>
</tbody>
</table>
Designing an institutional research data management infrastructure for the life sciences

**DataHub core services**

- **Life science**
  - Web portal
  - metadata XML
  - Data drop zone

- **EHR**
  - HL7CDA export
  - HL7v3

- **Healthcare**
  - Frontend
  - Facetted search UI
  - EHR
  - HL7CDA export

- **Scientist**
  - Frontend
  - External Repository
  - REST API
  - WebDAV

- **DataHub 2.0.0**
  - Data Warehouse
  - Rule based object store
  - iRODS
  - Rule based object store
  - Hitachi NAS storage
  - Replication storage
  - REST API
  - ELK

- **Research data governance**
  - Ontology Lookup Service
  - CrossRef Lookup
  - ePIC Persistent Identifier
  - Master Person Index
  - API

- **iRODS Browser**
  - CrossRef Lookup
  - ePIC Persistent Identifier
  - Master Person Index

- **DataHub 2.0.0**
  - Frontend
  - Facetted search UI
  - EHR
  - HL7CDA export
  - HL7v3

- **Life science**
  - Web portal
  - metadata XML
  - Data drop zone
Authentication (federated)

Providing federated authentication in two methods: proxy-user and temporary password

Outstanding issue:
• Automated handling of user provisioning/expiration
Ingesting high volume data

SMB/CIFS network share connected as iRODS mounted collection is ingested into iRODS using msiCollRsync

Advantage:
• No extra (client) software for users
• SMB/CIFS performs very well

Disadvantage:
• Not compatible with federated authentication
• msiCollRsync not performing (yet)
Project collection structure

/nlumuc/projects/

P000000001

C000000001

P000000002

C000000002

P000000003

C000000003

Providing a generic project collection hierarchy with no assumptions

- Unidentifiable collection names
- Virtual collections?
- Title AVUs on Project and Collections

Designing an institutional research data management infrastructure for the life sciences
Designing an institutional research data management infrastructure for the life sciences

Project authorization

- Project: P000000001
- Project collection: C000000001

Can assign

- own
  - manager
  - contributor
  - reader

Can create

- inherited
- read only
  - Open phase
  - Closed phase

Disadvantages:
- Only on project level
- Too simplistic?

Note: iRODS groups are organizational units (departments)

Keeping data authorization in iRODS using the rule engine to enforce policies
Metadata modeling: being FAIR

Helping users early with annotating data FAIR

Project -> Investigation -> Sample -> Assay (PISA)

- Inspired by ISA tools, compatible with HCLS
- Implemented Project and Investigation level
- Descriptive metadata stored in file (!), AVUs for system metadata
Metadata indexing

Providing a user friendly facetted search interface for data findability

- Indexed in SOLR:
  - All metadata
  - Semantics (OLS)
  - References (CrossRef)
  - Authorization on data (iRODS)
- Rebuild on demand
Metadata: making use of semantics

Autocomplete for ontology terms

Ontology derived facetted search

Designing an institutional research data management infrastructure for the life sciences
DTAP: deployment for development

Challenge
- Interactions with external services (AD, NAS storage)

Highlights
- 16 interacting containers for full environment
- Runnable from laptop
Designing an institutional research data management infrastructure for the life sciences

DTAP: deployment for acceptation/production

Challenge

• Differences in deployments and some environments
Today's challenge in the data life cycle

Active data
- Phases
  - Create
  - Process
  - Analyse
- Highly specific RDM solutions

Preserved data
- Phases:
  - Archive
  - Access
  - Re-use
- Generic repositories
- Domain specific repositories

BRIDGE THE GAP!

Designing an institutional research data management infrastructure for the life sciences
Lessons learned

1. Dual position of staff. Decentralize data stewards
2. Micro Service approach
3. Remote Procedure Calls for rules
4. Funding for long term storage is hard…
5. Open Source re-useable parts

https://github.com/MaastrichtUniversity