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

Maastricht UMC+

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

Maastricht UMC+
DataHub

## providing Research Data Management services for

### Life Sciences Faculty

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

### Academic Hospital

- Patient privacy
- Electronic Health Records
- Bridging organisations

# 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.

# DataHub characteristics

**FAIR**-inspired from start.

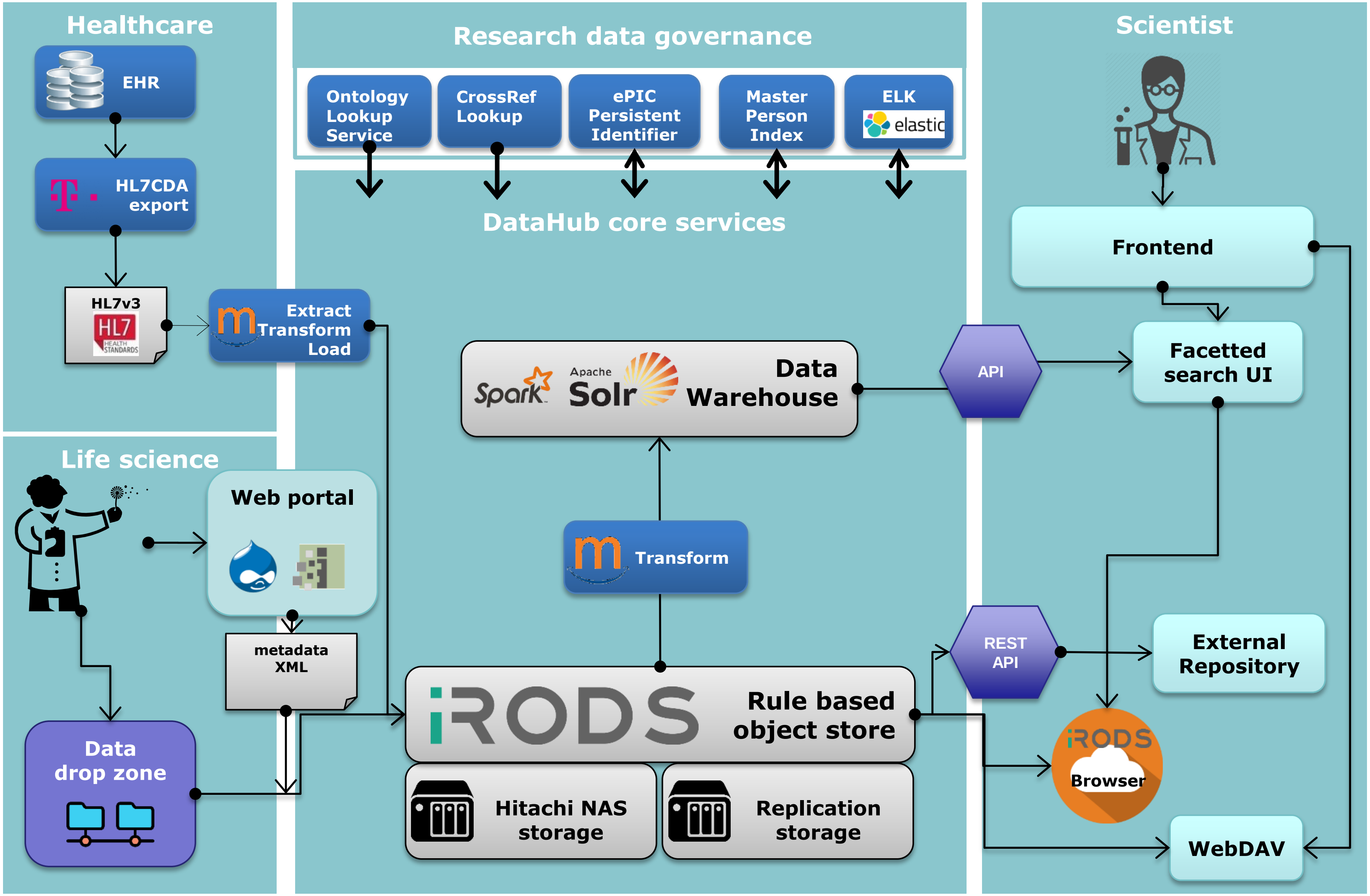
**Open-source** where possible.

(Meta)data **structuring** + ontology **enrichment.**

**Project data structuring;** Hierarchical organisation in projects and datasets.

**Faceted search**, Lucene & ontology-powered, authorization controlled

**High volume;** The infrastructure has been designed and tested with petabyte scale and high throughput in mind.



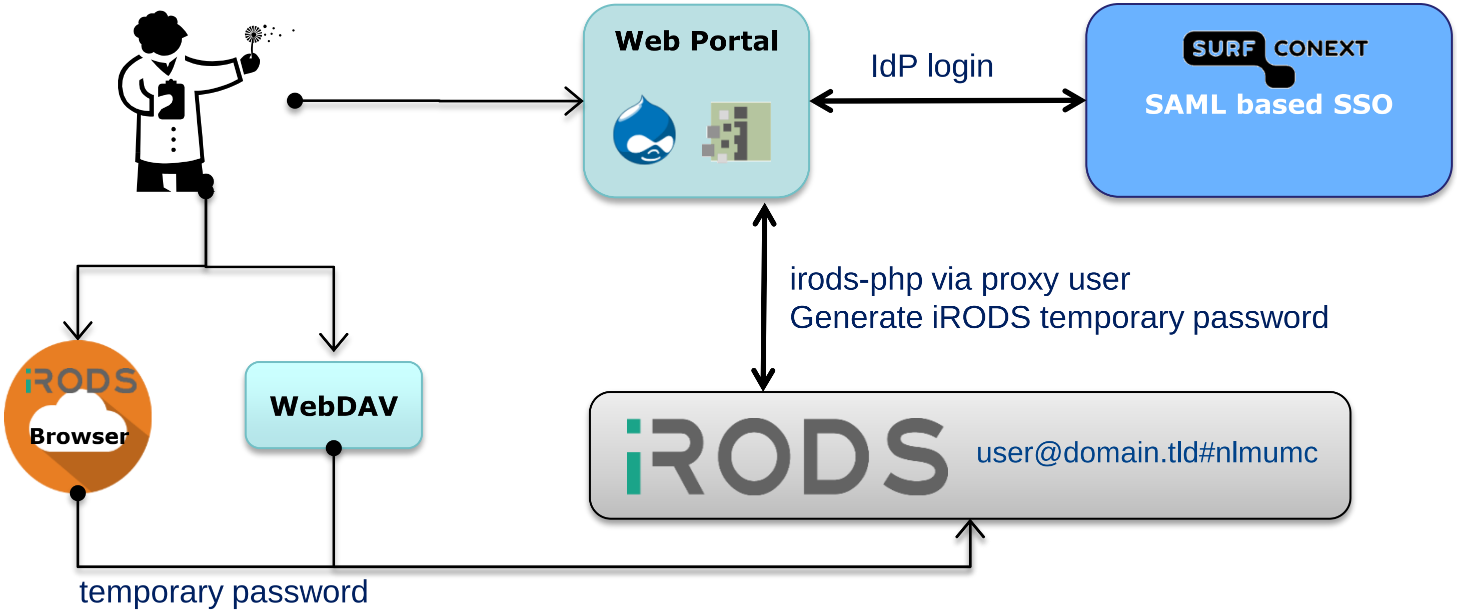
Designing an institutional research data management infrastructure for the life sciences

Maastricht UMC+ DataHub

# DataHub 2.0.0

Maastricht University

# Authentication (federated)

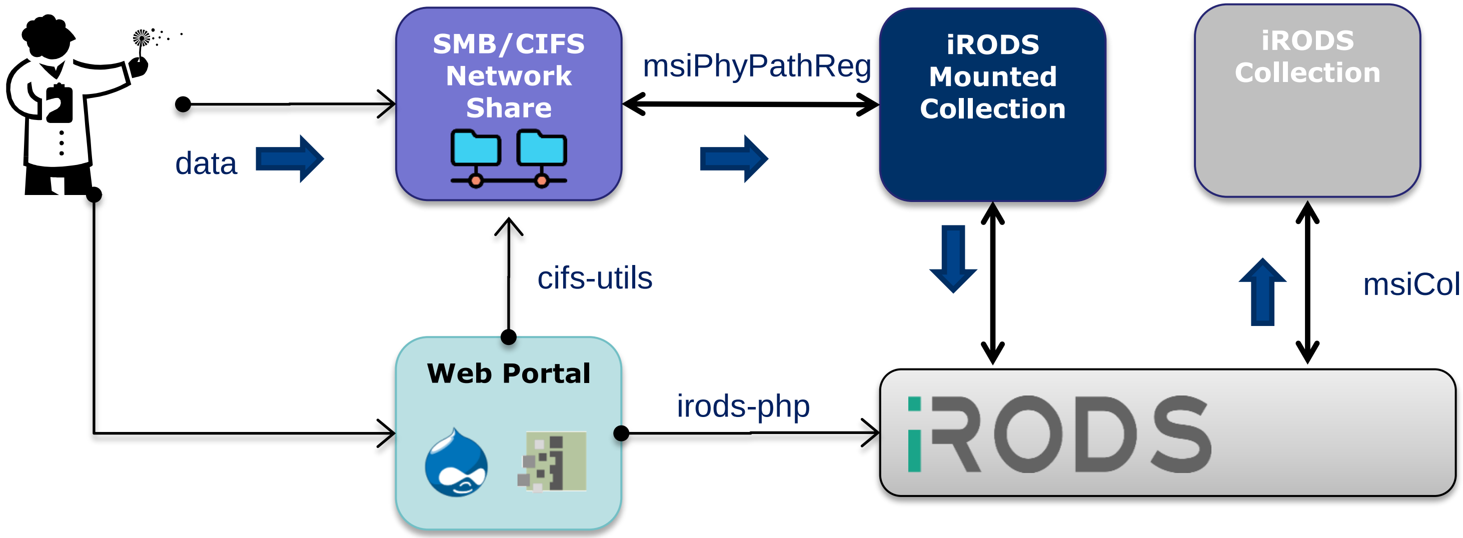


**Providing federated authentication in two methods: proxy-user and temporary password**

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**

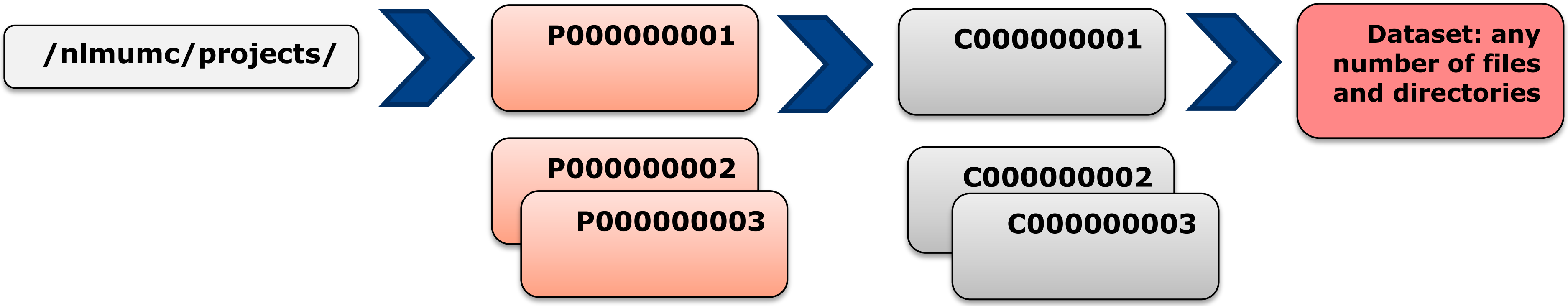
Advantage:

- No extra (client) software for users
- SMB/CIFS performs very well

Disadvantage:

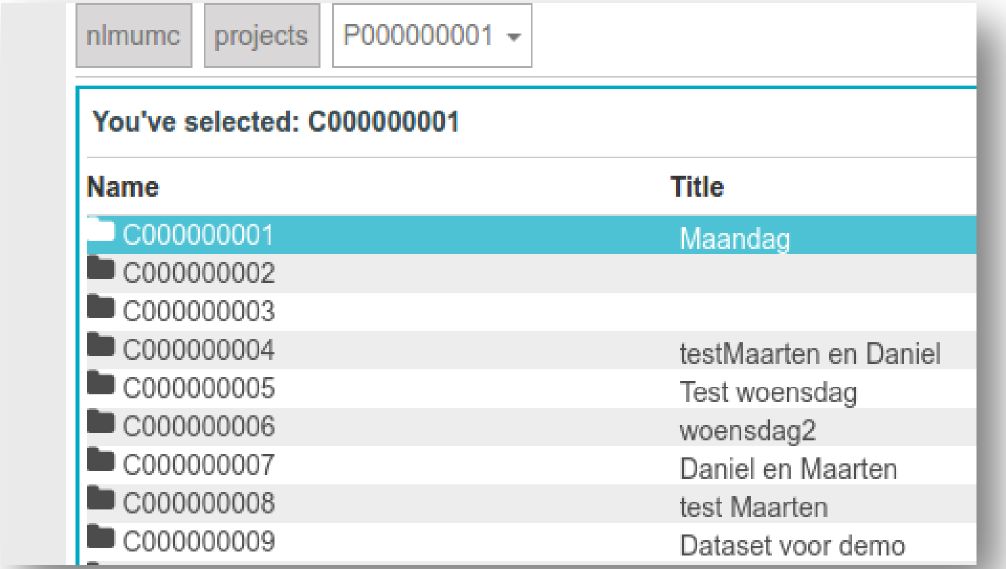
- Not compatible with federated authentication
- msiCollRsync not performing (yet)

# Project collection structure



**Providing a generic project collection hierarchy with no assumptions**

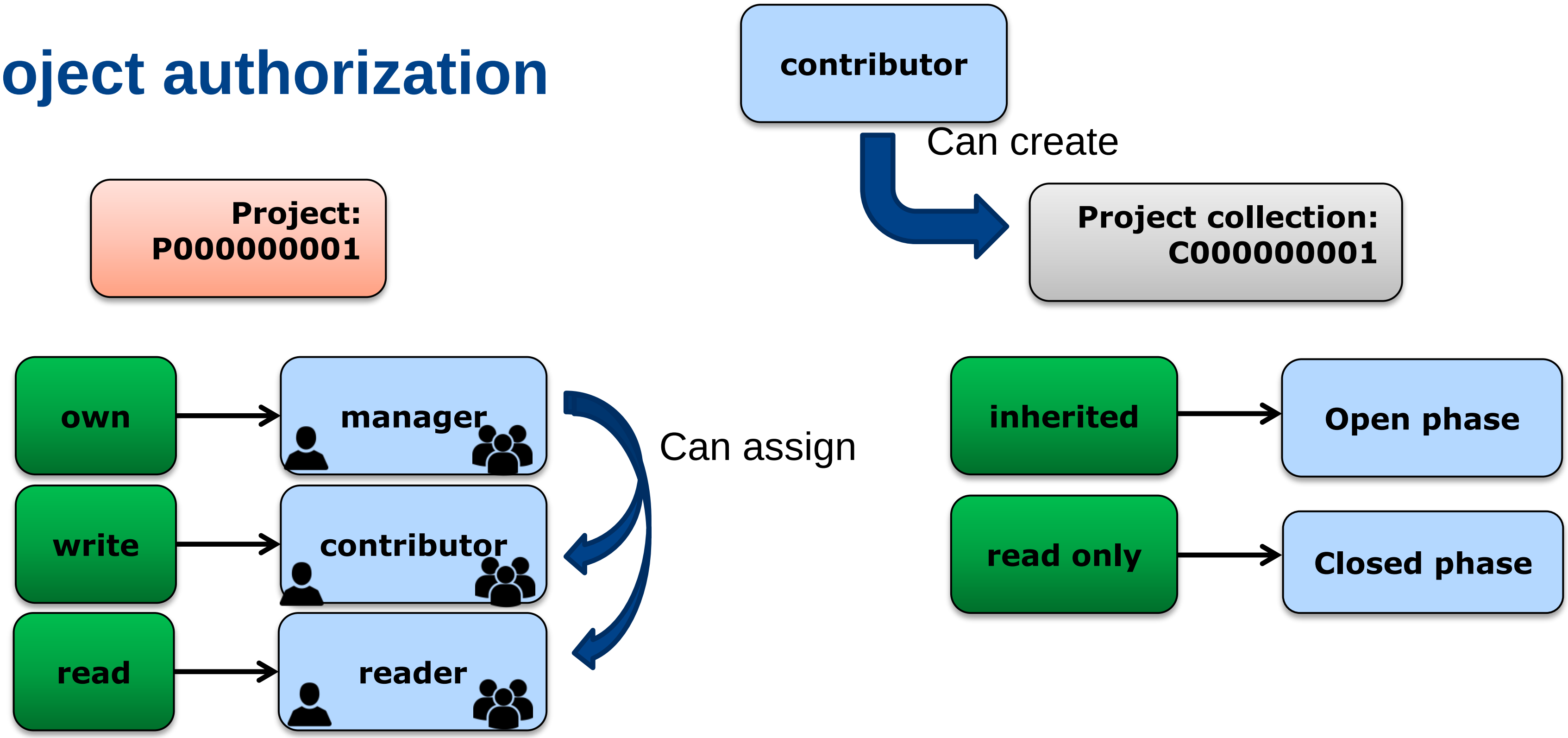
- Unidentifiable collection names
- Virtual collections?
- Title AVUs on Project and Collections



**You've selected: C000000001**

| Name | Title |
|---|---|
| C000000001 | Maandag |
| C000000002 | |
| C000000003 | |
| C000000004 | testMaarten en Daniel |
| C000000005 | Test woensdag |
| C000000006 | woensdag2 |
| C000000007 | Daniel en Maarten |
| C000000008 | test Maarten |
| C000000009 | Dataset voor demo |

# Project authorization



contributor

Can create

Project collection:
C000000001

Project:
P000000001

own → manager

write → contributor

read → reader

Can assign

inherited → Open phase

read only → Closed phase

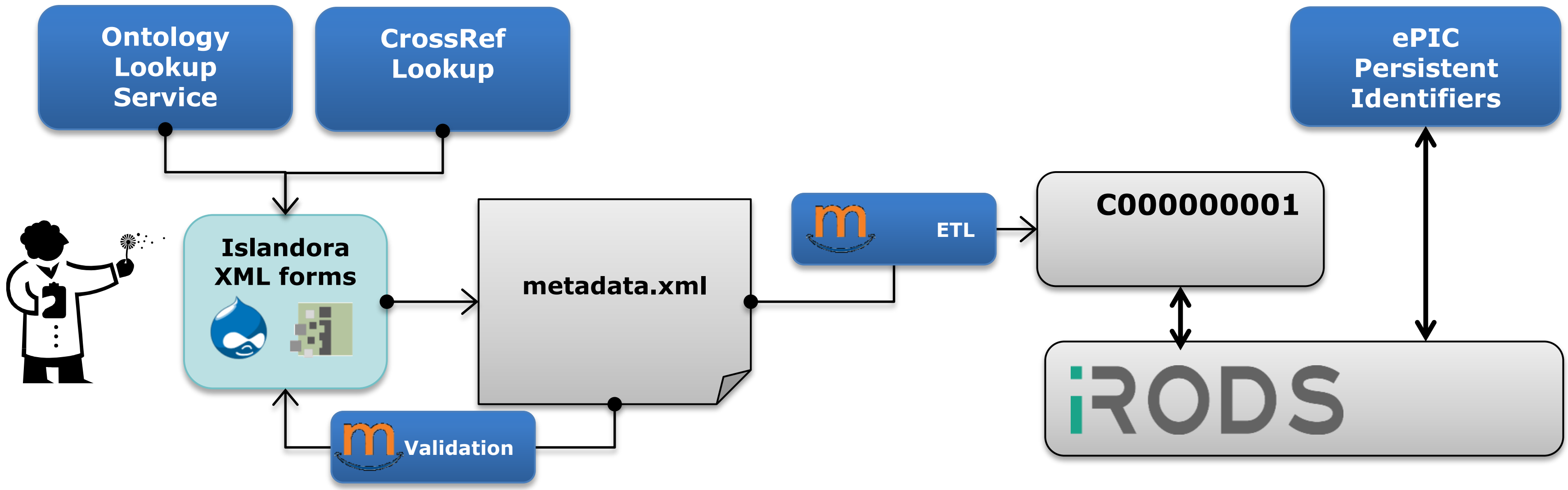
**Keeping data authorization in iRODS using the rule engine to enforce policies**

Disadvantages:

- Only on project level
- Too simplistic?

Note: iRODS groups are organizational units (departments)

# Metadata modeling: being FAIR

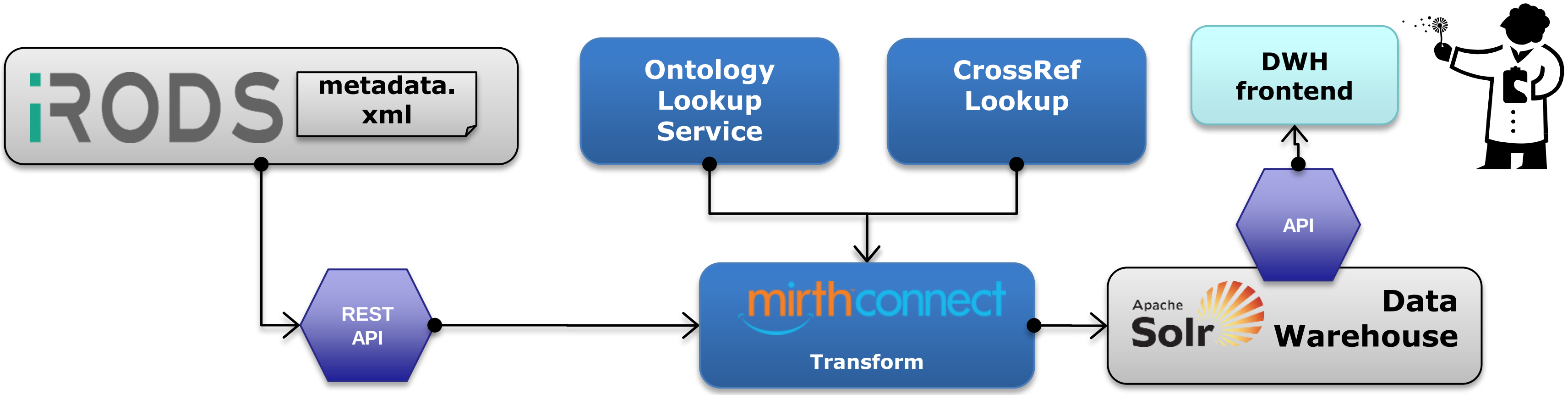


## Helping users early with annotating data FAIR

### 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

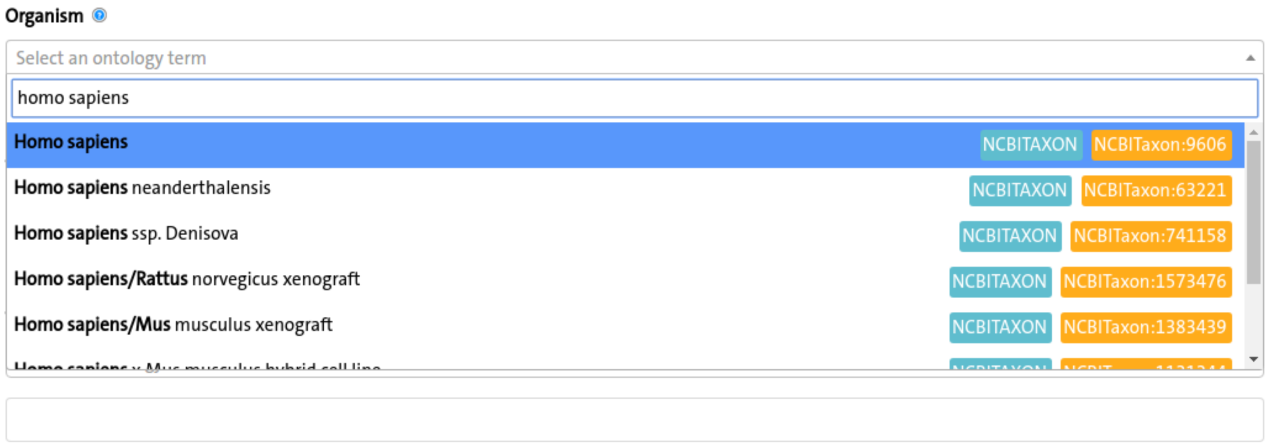
# Metadata indexing



**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



**Organism**

Select an ontology term

homo sapiens

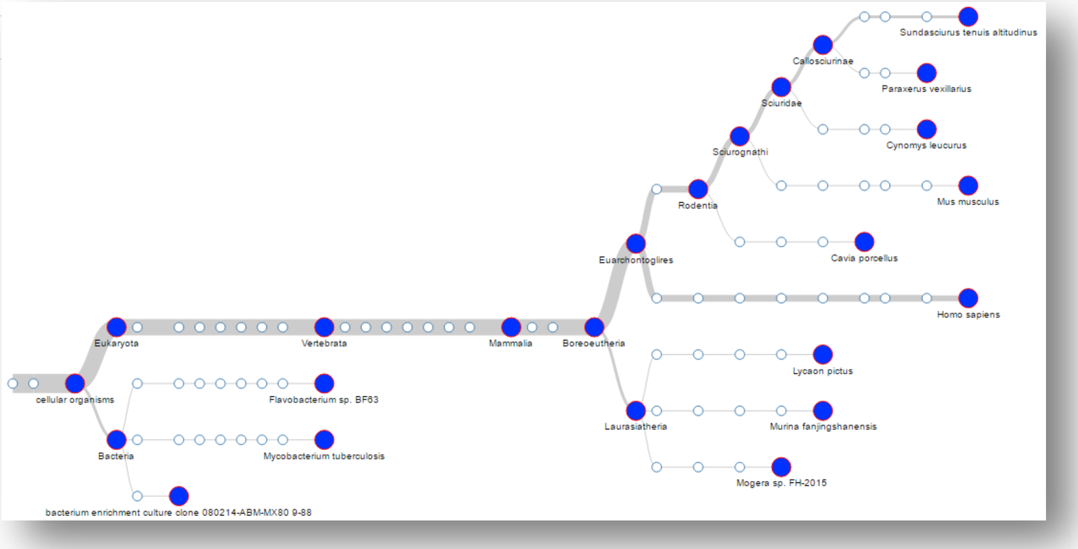
| | | |
|---|---|---|
| **Homo sapiens** | NCBITAXON | NCBITaxon:9606 |
| **Homo sapiens** neanderthalensis | NCBITAXON | NCBITaxon:63221 |
| **Homo sapiens** ssp. Denisova | NCBITAXON | NCBITaxon:741158 |
| **Homo sapiens/Rattus** norvegicus xenograft | NCBITAXON | NCBITaxon:1573476 |
| **Homo sapiens/Mus** musculus xenograft | NCBITAXON | NCBITaxon:1383439 |
| **Homo sapiens** x Mus musculus hybrid cell line | NCBITAXON | NCBITaxon:1131344 |

**Related publications (DOIs)**

https://doi.org/...

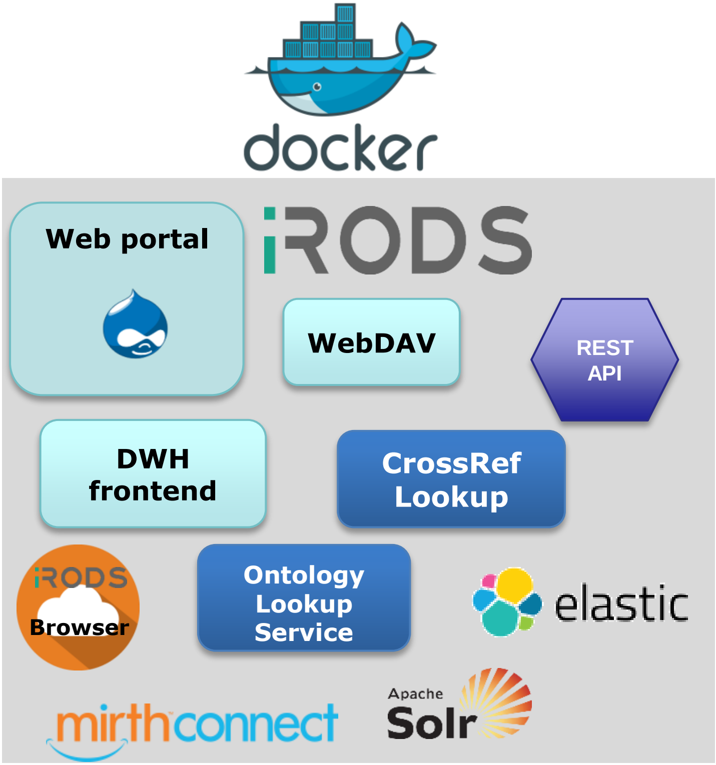
Add article

## Autocomplete for ontology terms

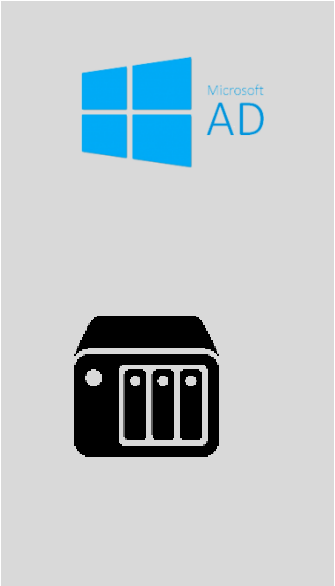


## Ontology derived facetted search

# DTAP: deployment for development



External



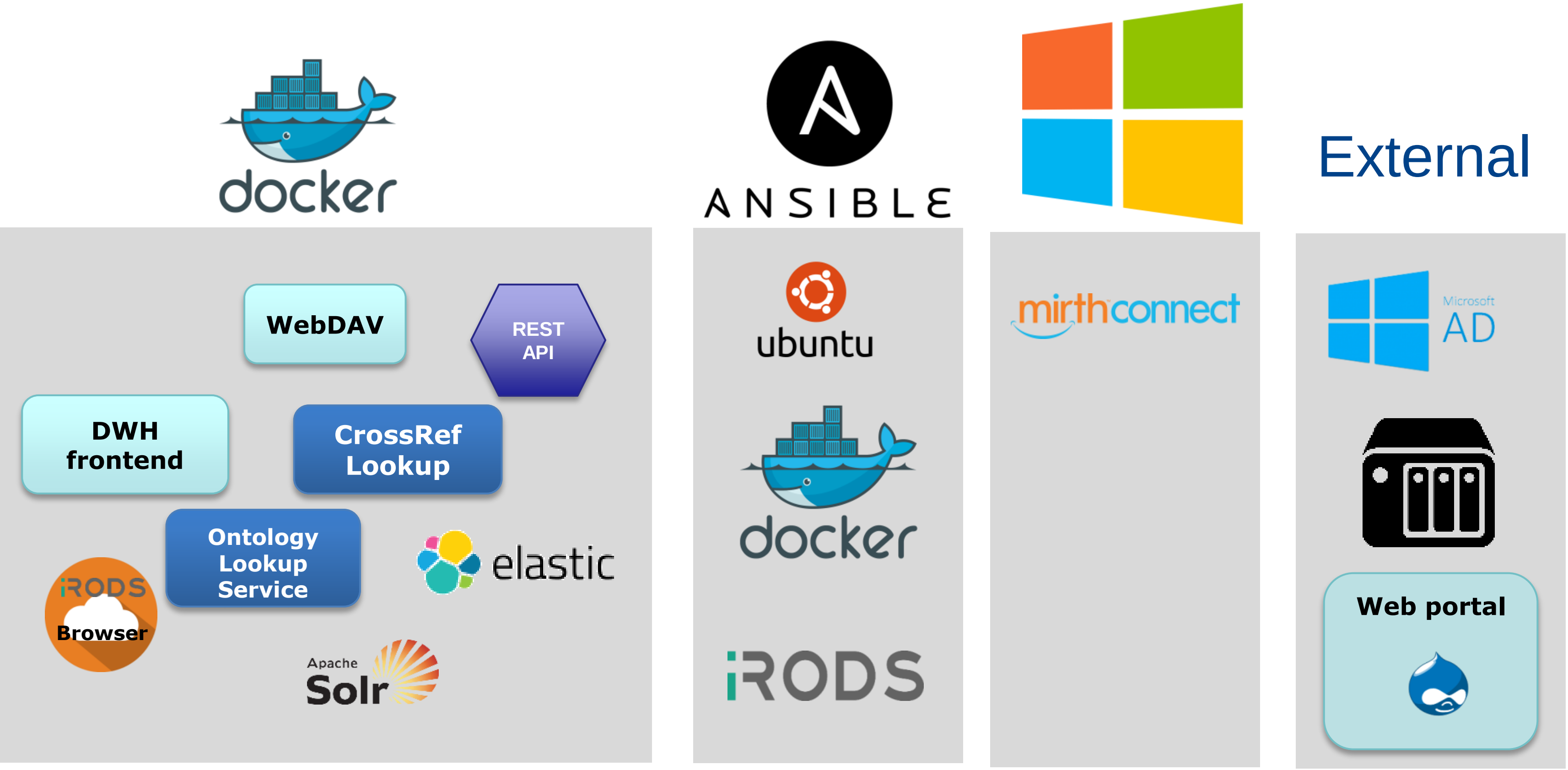
Challenge

- Interactions with external services (AD, NAS storage)

Highlights

- 16 interacting containers for full environment
- Runnable from laptop

# DTAP: deployment for acceptation/production



External

WebDAV

REST API

DWH frontend

CrossRef Lookup

Ontology Lookup Service

Web portal

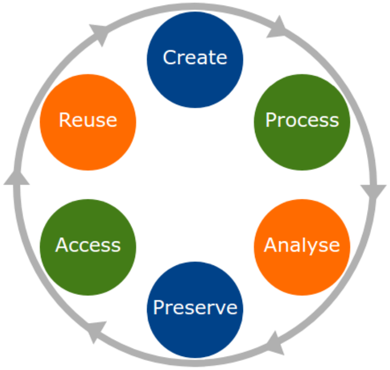
## Challenge

- Differences in deployments and some environments

# Todays challenge in the data life cycle

## Active data

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



FAIR

**BRIDGE THE GAP!**

iRODS

## Preserved data

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

figshare

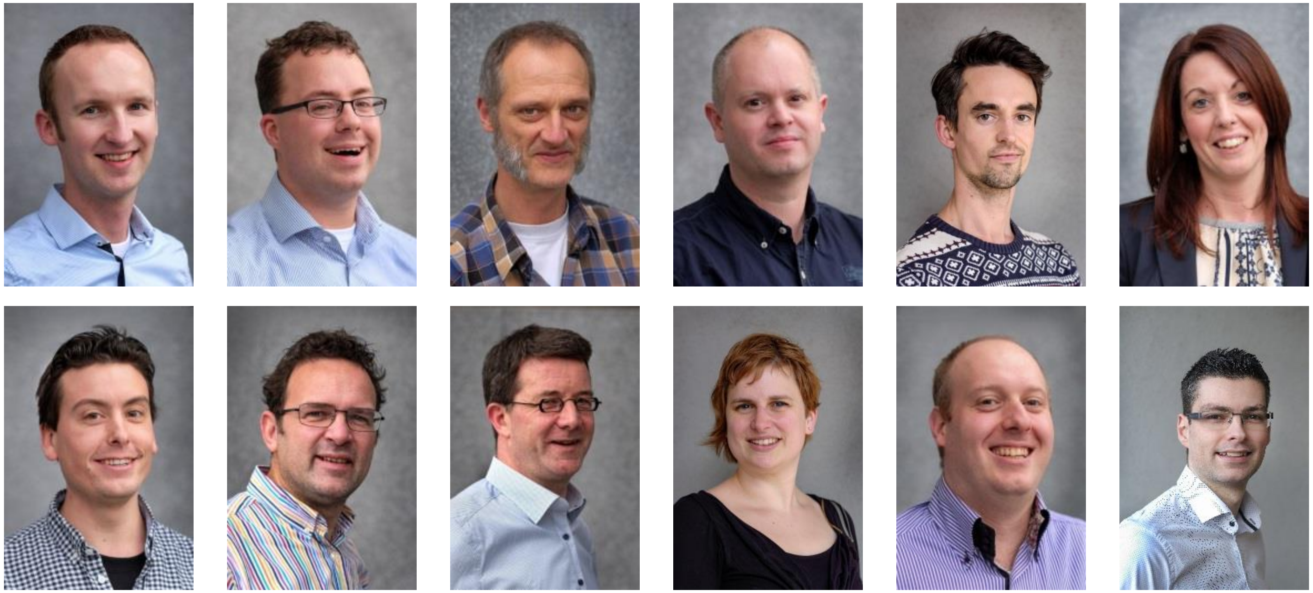
The Dataverse Project

zenodo

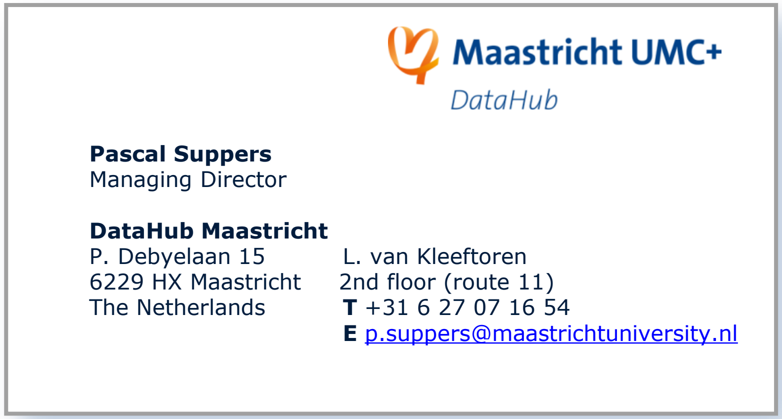
# Lessons learned

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



# Questions?



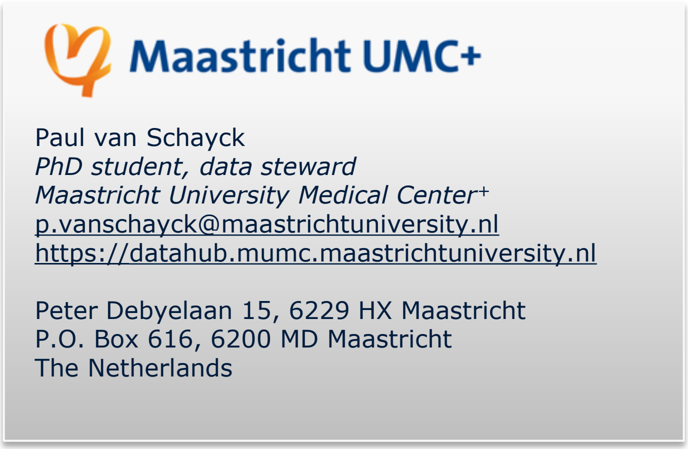
Maastricht UMC+

DataHub

**Pascal Suppers**
Managing Director

**DataHub Maastricht**
P. Debyelaan 15
6229 HX Maastricht
The Netherlands

L. van Kleeftoren
2nd floor (route 11)
**T** +31 6 27 07 16 54
**E** p.suppers@maastrichtuniversity.nl



Maastricht UMC+

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