Maastricht UMC+
DataHub

FAIR data management and Disqoverability

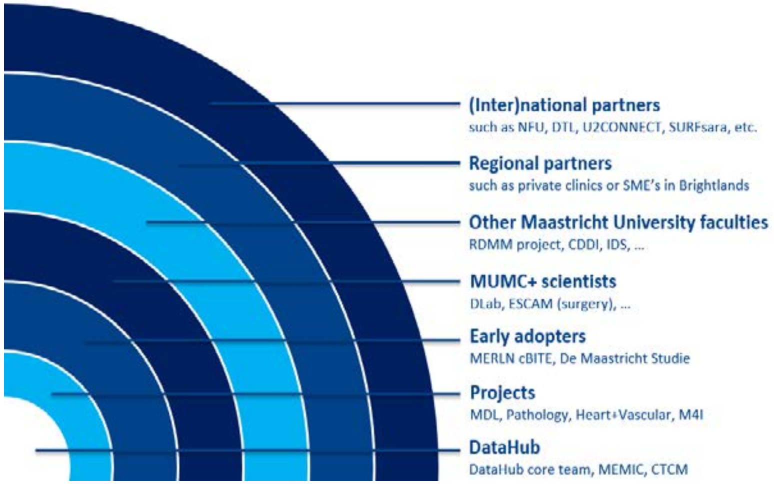
iRODS UGM 2018

Maastricht UMC+
DataHub

Maarten Coonen
Data Architect
DataHub Maastricht

m.coonen@maastrichtuniversity.nl
https://datahub.mumc.maastrichtuniversity.nl
Peter Debyelaan 15, 6229 HX Maastricht,
The Netherlands (route 11 MUMC+, 2nd floor)

DataHub Maastricht



(Inter)national partners
such as NFU, DTL, U2CONNECT, SURFsara, etc.

Regional partners
such as private clinics or SME's in Brightlands

Other Maastricht University faculties
RDMM project, CDDI, IDS, ...

MUMC+ scientists
DLab, ESCAM (surgery), ...

Early adopters
MERLN cBITE, De Maastricht Studie

Projects
MDL, Pathology, Heart+Vascular, M4I

DataHub
DataHub core team, MEMIC, CTCM

Community at Maastricht UMC+

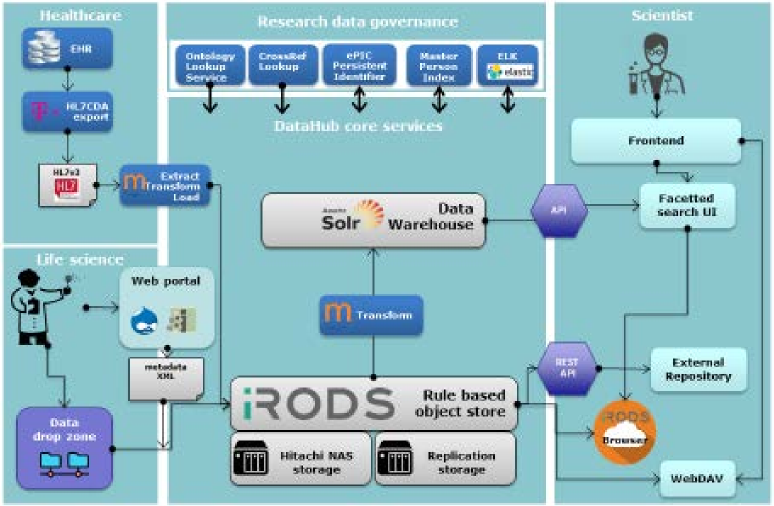
Characteristics

- Service organization
- For hospital and university
- Data broker
- Scope = data management (not data science)
  - Consultancy and Legislation (GDPR)
  - Data management planning
  - (Meta)data modelling
- Decentral data stewards

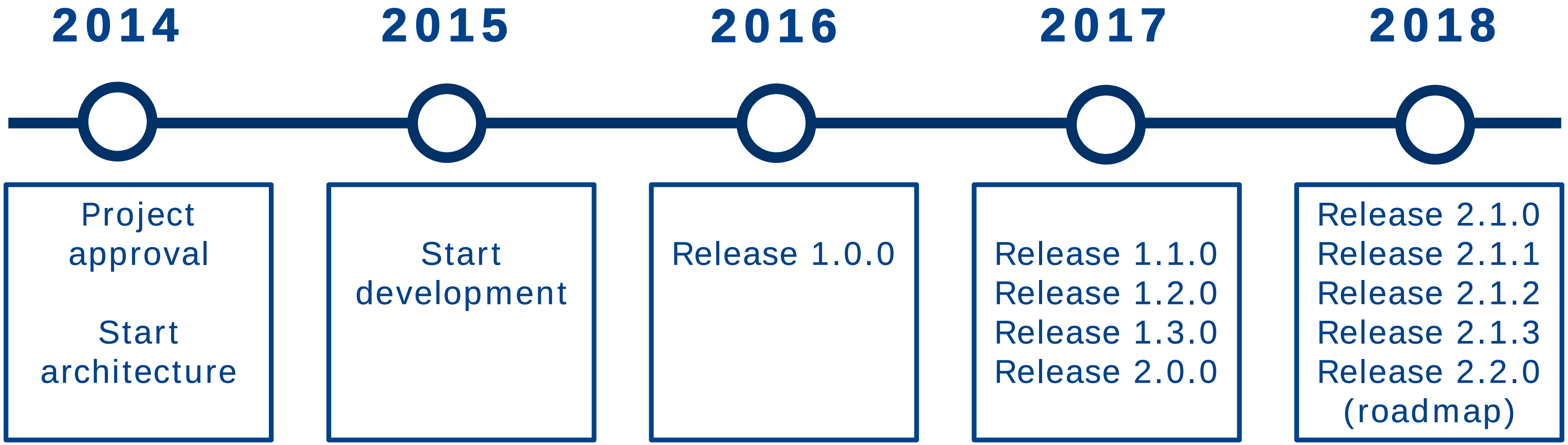
Paul van Schayck @UGM2017

DataHub is more than iRODS alone:

+ Web portal
+ Metadata entry
+ Ontology Lookup Service
+ Pseudonimysation
+ Search Index (Solr)
+ And other (dockerized) microservices



DataHub (iRODS) milestones



2014

Project approval

Start architecture

2015

Start development

2016

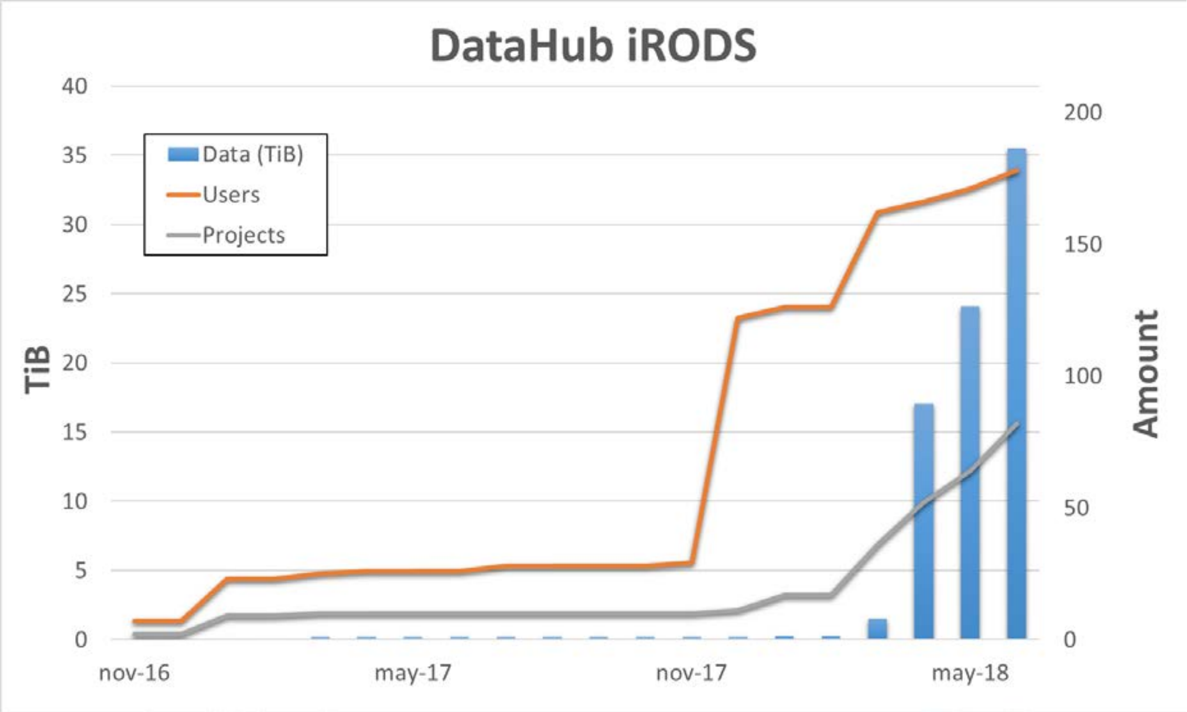
Release 1.0.0

2017

Release 1.1.0
Release 1.2.0
Release 1.3.0
Release 2.0.0

2018

Release 2.1.0
Release 2.1.1
Release 2.1.2
Release 2.1.3
Release 2.2.0
(roadmap)



Findable Accessible Interoperable Reusable

Our FAIR mission

iRODS

DataHub strives

- to be FAIR across research disciplines;
- share data in regulated fashion between organizations;
- to hold data sets that are both human and machine readable.

DataHub implementation	F	A	I	R
Each data set in **iRODS** has a unique and **persistent identifier (PID)**	F1 F3			
Metadata structuring and ontology enrichment using **EBI-OLS**	F2		I1,I2,I3	R1,R1.3
Metadata registered in **iRODS** and indexed in **DISQOVER**	F4			
Metadata retrievable by their PID using **HTTP** landing page		A1,A1.1,A1.2		
Metadata accessible, even when data is deleted or protected by authorization in **iRODS**		A2		

Gaps: data license (R1.1), extended metadata about provenance (R1.2)

Maastricht UMC+
DataHub

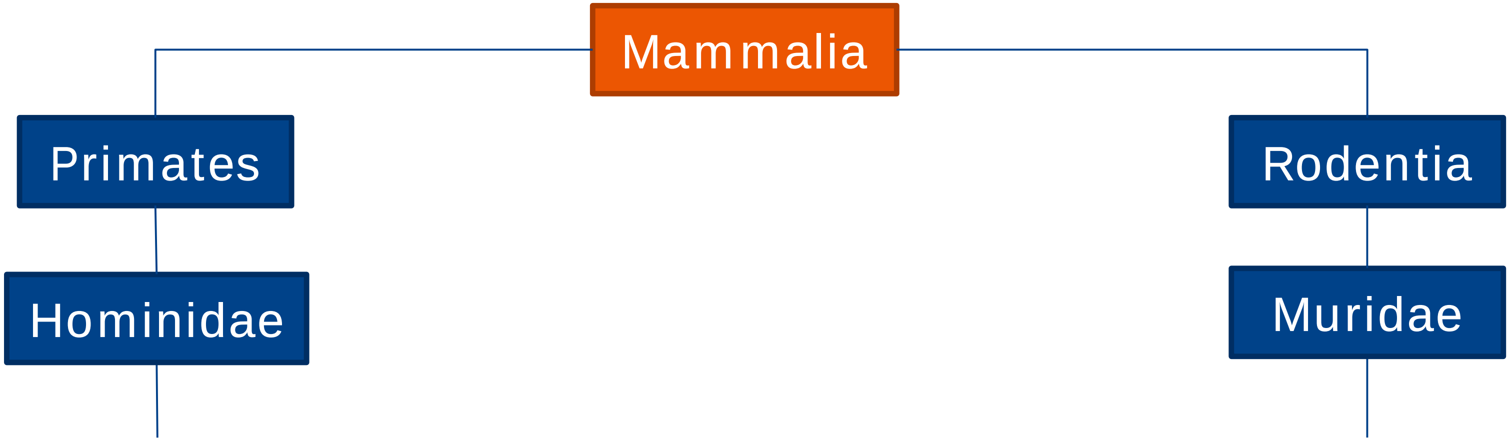
Sources
https://www.dtls.nl/fair-data/fair-principles-explained/
http://doi.org/10.1038/sdata.2016.18

Maastricht University

Data sets that are both human and ***machine*** *readable*

Ontologies enable machine-readability

Find all information regarding mammals



Mammalia

Primates

Hominidae

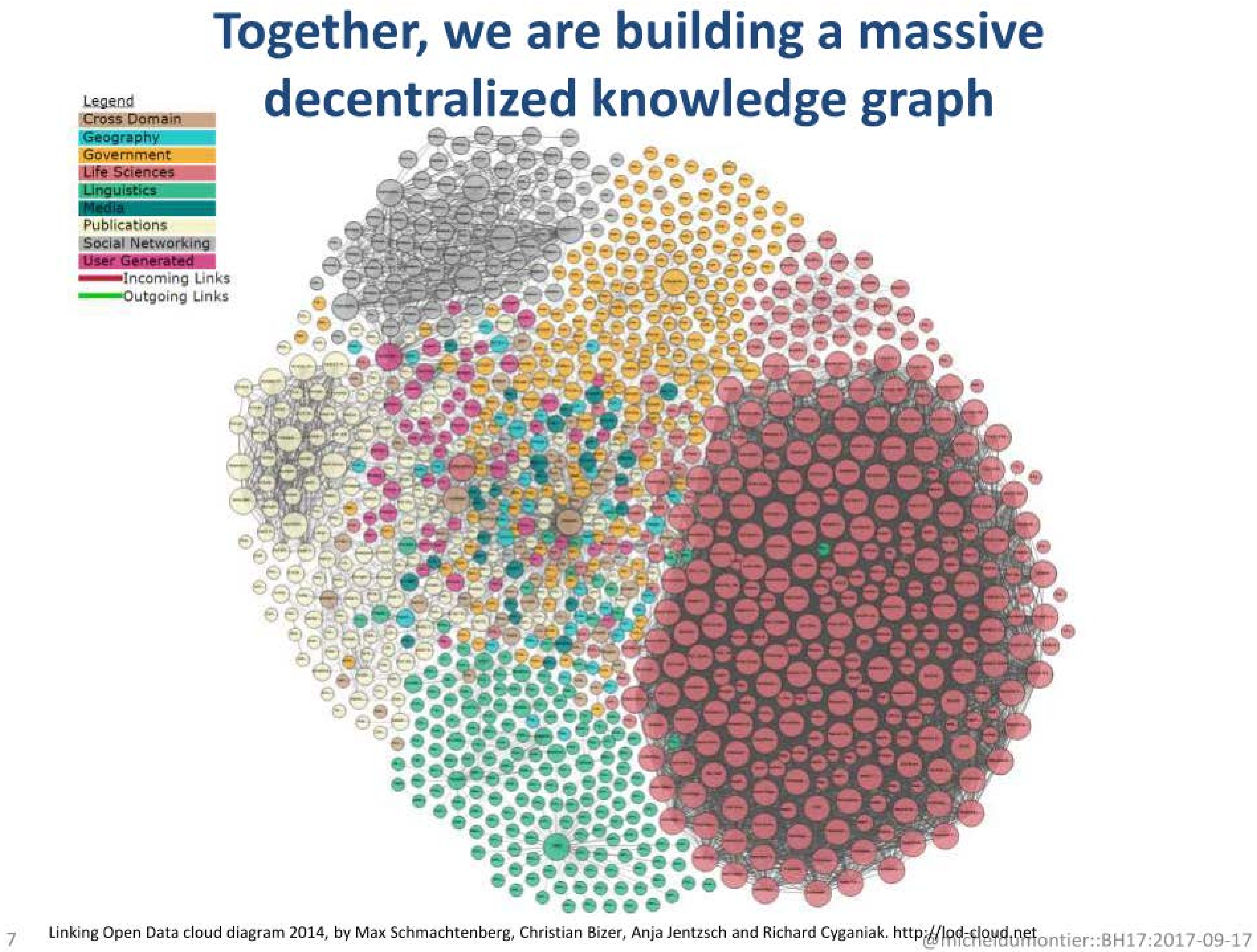
Rodentia

Muridae

Homo sapiens

Mus musculus

The Linked Data Cloud



Together, we are building a massive decentralized knowledge graph

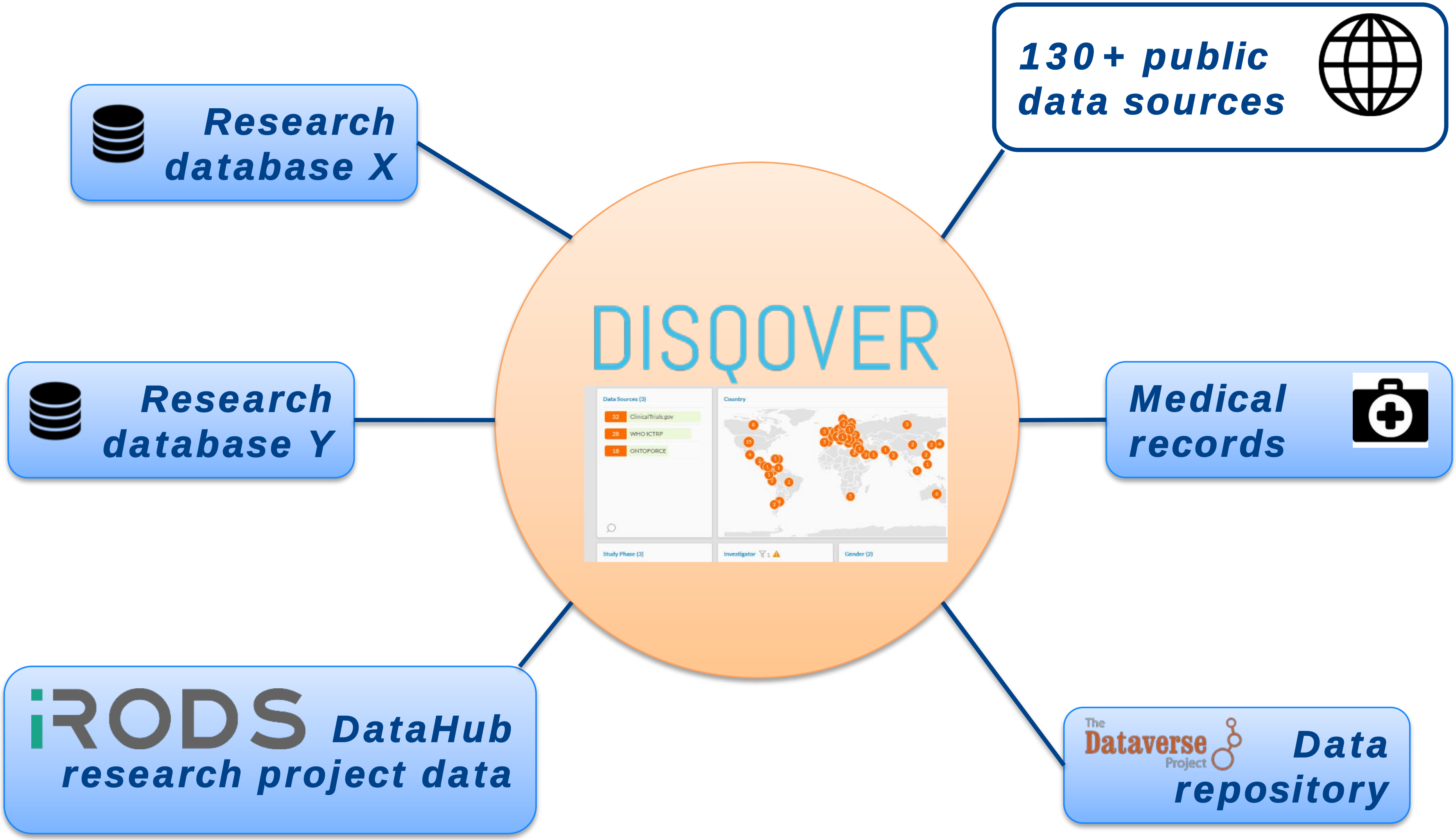
7

Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net

@micheldumontier::BH17:2017-09-17

Source: https://www.slideshare.net/micheldumontier/advancing-biomedical-knowledge-reuse-with-fair

DISQOVER in the Linked Data cloud



Legend on-premises data | Remote federated data | on-premises Linked data

ONTOFORCE DISQOVER

ONTOFORCE

http://www.ontoforce.com

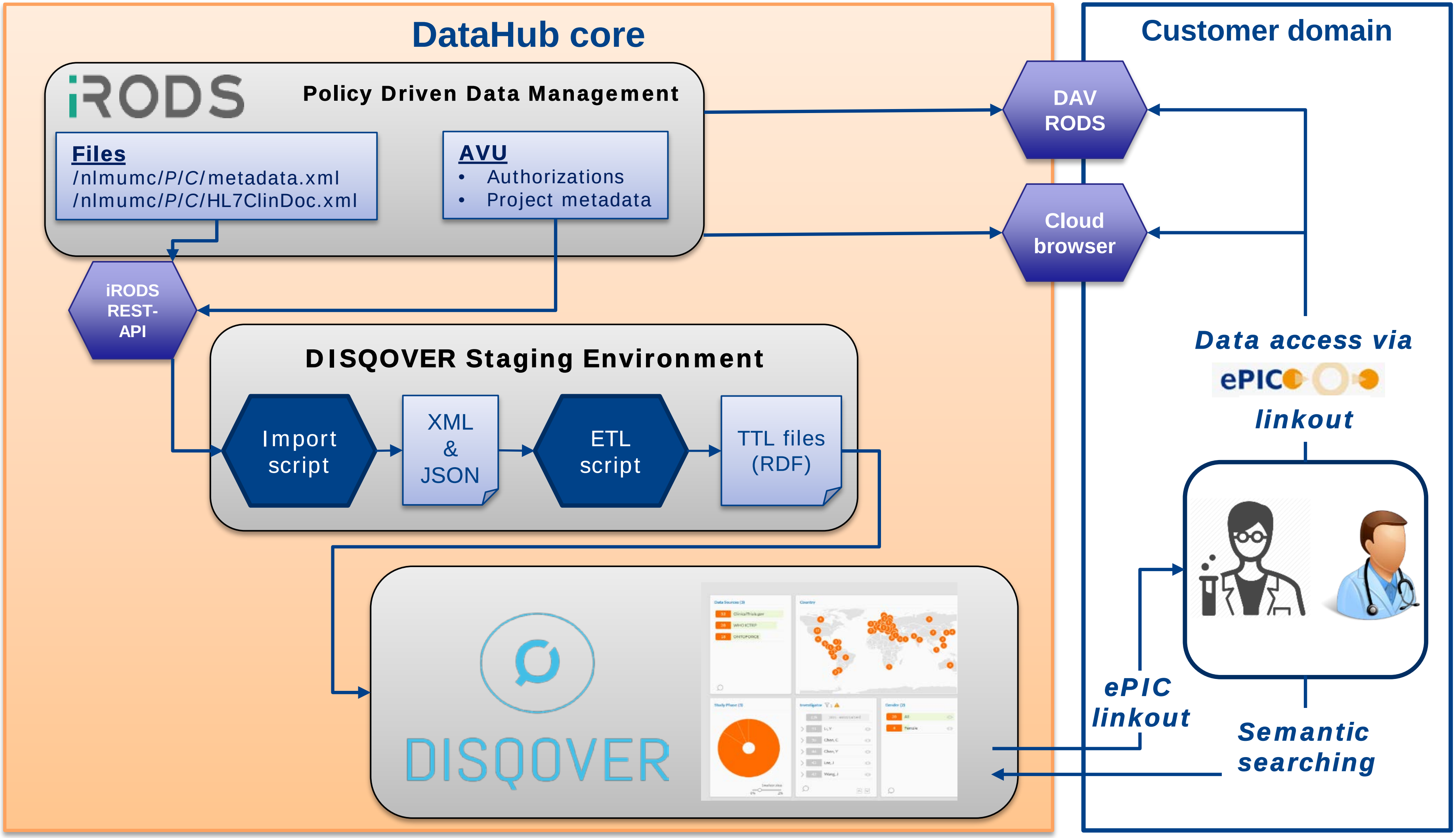
Characteristics

- **Semantic search** application on linked data
- User-friendly **interface and visualizations**
- End-user does not need SPARQL expertise
- Use of **dynamic filters / facets** to construct the search query
- Aggregates linked data from **public** and **private** (local) data sources

Public data sources

- PubMed
- NCBI Gene
- ChEMBL
- ClinicalTrials.gov
- ORCiD
- MesH
- DailyMed
- *and many more (130+)*

iRODS – DISQOVER workflow



DataHub core

iRODS Policy Driven Data Management

Files
/nlmumc/P/C/metadata.xml
/nlmumc/P/C/HL7ClinDoc.xml

AVU
- Authorizations
- Project metadata

iRODS REST-API

DISQOVER Staging Environment

Import script → XML & JSON → ETL script → TTL files (RDF)

DISQOVER

DAV RODS

Cloud browser

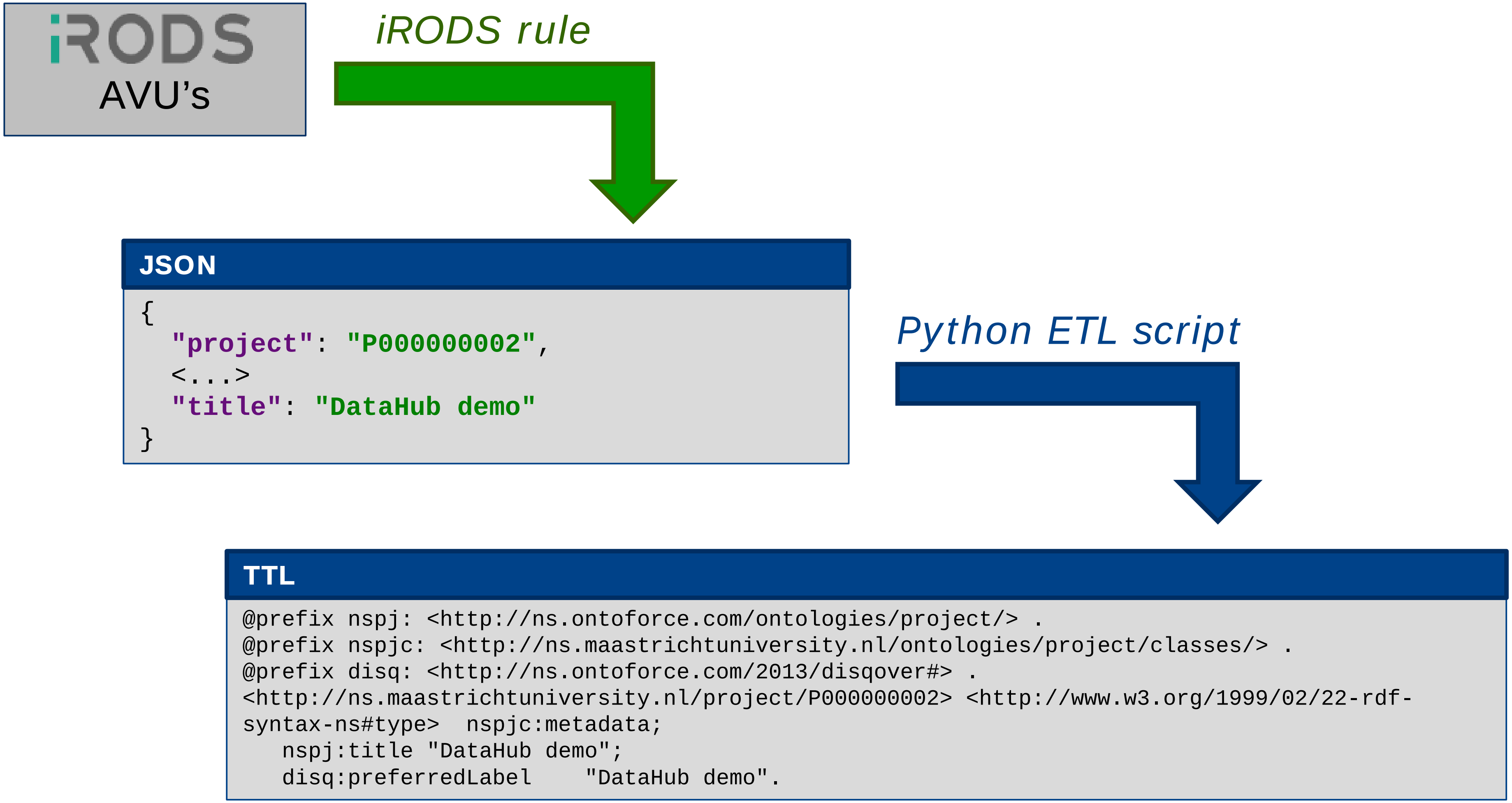
Customer domain

Data access via ePIC linkout

ePIC linkout

Semantic searching

Converting iRODS AVU's to RDF



iRODS

AVU's

iRODS rule

JSON

```
{
  "project": "P000000002",
  <...>
  "title": "DataHub demo"
}
```

Python ETL script

TTL

```
@prefix nspj: <http://ns.ontoforce.com/ontologies/project/> .
@prefix nspjc: <http://ns.maastrichtuniversity.nl/ontologies/project/classes/> .
@prefix disq: <http://ns.ontoforce.com/2013/disqover#> .
<http://ns.maastrichtuniversity.nl/project/P000000002> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  nspjc:metadata;
   nspj:title "DataHub demo";
   disq:preferredLabel    "DataHub demo".
```

Converting XML-file to RDF

iRODS

metadata.xml

REST GET /fileContents/

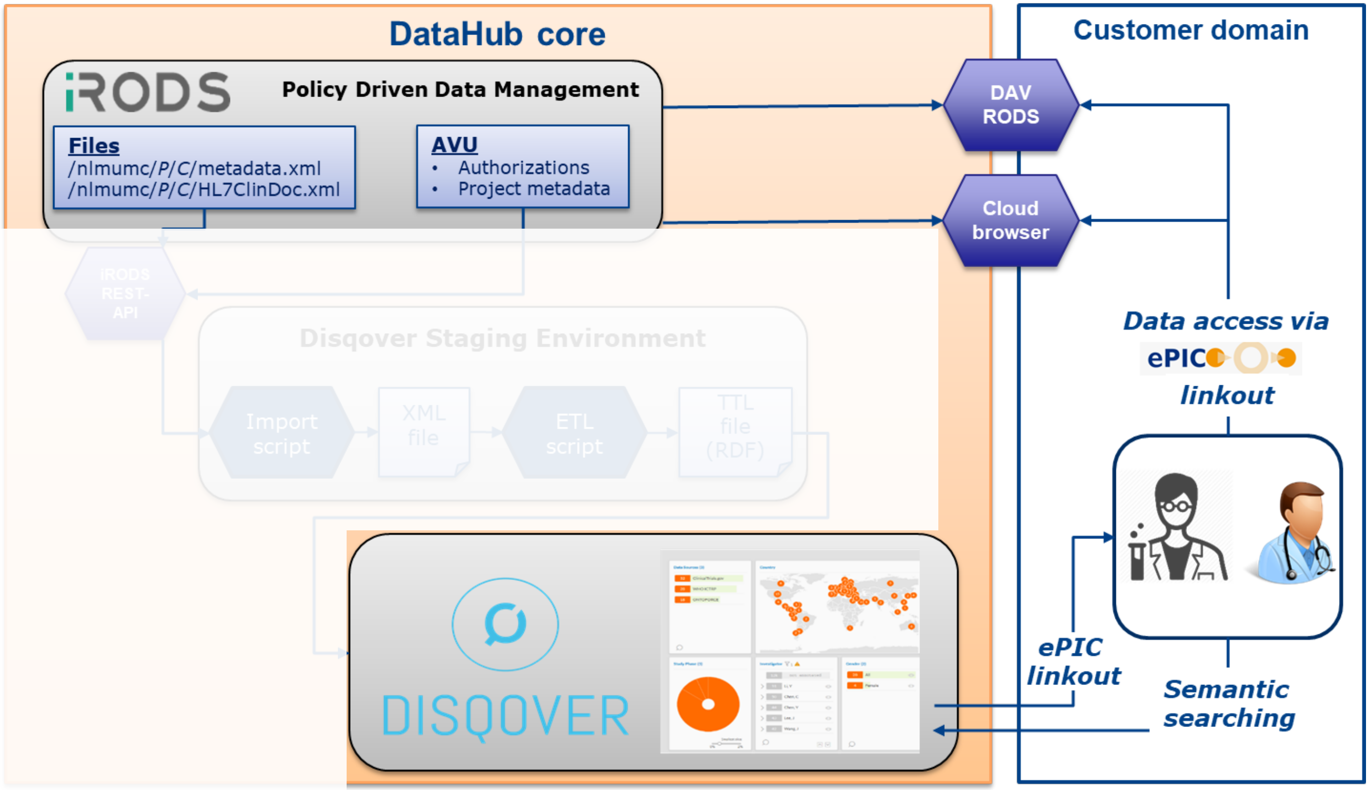
metadata.xml

```xml
<?xml version='1.0' encoding='UTF-8'?>
<metadata>
  <project>P000000002</project>
  <title>ATGL and CGI-58 Western Blot</title>
  <description>CGI-58 is involved in the regulation of energy metabolism in
skeletal muscle. This investigation consists of various Western Blots targeted at
both ATGL and CGI-58 in human myoblasts.</description>
  <date>2010-05-11</date>
<organism id="ncbitaxon:http://purl.obolibrary.org/obo/NCBITaxon_9606">Homo
sapiens</organism>
```

Python ETL script

TTL

```
@prefix ns: <http://ns.ontoforce.com/ontologies/collection/> .
@prefix nst: <http://ns.maastrichtuniversity.nl/ontologies/collection/classes/> .
@prefix nstp: <http://ns.ontoforce.com/ontologies/person/classes/> .
@prefix disq: <http://ns.ontoforce.com/2013/disqover#> .
@prefix nsp: <http://ns.ontoforce.com/ontologies/person/> .
@prefix org: <http://ns.ontoforce.com/organization/> .
<http://ns.maastrichtuniversity.nl/collection/P000000002-C000000001>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  nst:metadata;
   ns:project <http://ns.maastrichtuniversity.nl/project/P000000002>;
   disq:preferredLabel    "ATGL and CGI-58 Western Blot";
   ns:description "CGI-58 is involved in the regulation of energy metabolism in skeletal
muscle. This investigation consists of various Western Blots targeted at both ATGL and CGI-58
in human myoblasts.";
   ns:date    "2010-05-11";
   ns:organism    <http://purl.obolibrary.org/obo/NCBITaxon_9606>.
```



DataHub core

iRODS

Policy Driven Data Management

Files
/nlmumc/*P*/*C*/metadata.xml
/nlmumc/*P*/*C*/HL7ClinDoc.xml

AVU
- Authorizations
- Project metadata

iRODS REST-API

Disqover Staging Environment

Import script

XML file

ETL script

TTL file (RDF)

DISQOVER

Customer domain

DAV RODS

Cloud browser

Data access via ePIC linkout

ePIC linkout

Semantic searching

Screencast

The DataHub team

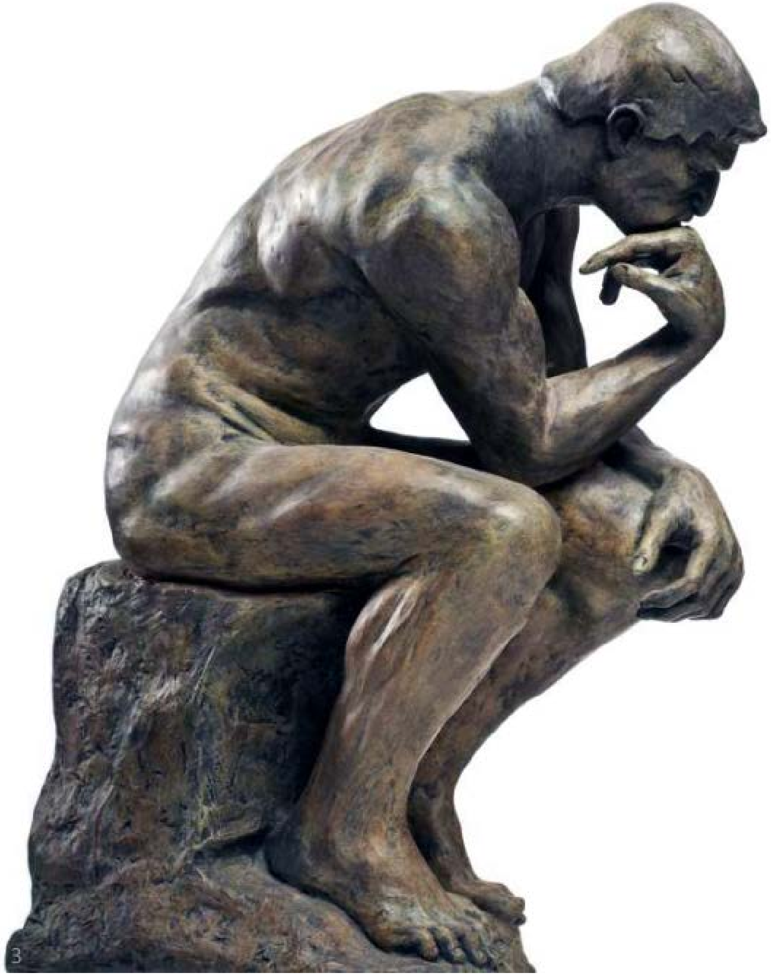
Maastricht UMC+
DataHub

Maarten Coonen
Data Architect
DataHub Maastricht

m.coonen@maastrichtuniversity.nl
https://datahub.mumc.maastrichtuniversity.nl
Peter Debyelaan 15, 6229 HX Maastricht,
The Netherlands (route 11 MUMC+, 2nd floor)

Backup slides

Machines that reason over data



Prof. Dr. Michel Dumontier, Maastricht

How can we **automatically** find the **evidence** that support or dispute a **hypothesis** using the totality of available **data, tools and scientific knowledge**?

Source: https://www.slideshare.net/micheldumontier/developing-and-assessing-fair-digital-resources

FAIR data principles

Set of 15 principles that form a guideline for proper research data management and data stewardship.

Gaining more and more interest of researchers, publishers, funding and government agencies worldwide.

Researchers
Data Scientists

Software vendors

Findable Accessible Interoperable Reusable

Publishers

- *Elsevier*
- *Springer*
- *etc.*

Government
University policy
Research institutes

Funding agencies

- *H2020*
- *NWO*
- *etc.*

Maastricht UMC+
DataHub

Sources
https://www.dtls.nl/fair-data/fair-data/
https://www.nature.com/articles/sdata201618.pdf

Maastricht University