Leids Universitair Medisch Centrum

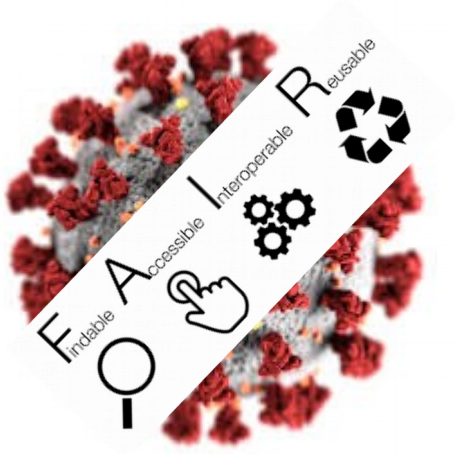
FAIR Data Management to Access Patient Data

Núria Queralt Rosinach, Rajaram Kaliyaperumal, César Bernabé, Qinqin Long, Henk Jan van der Wijk, Barend Mons, and Marco Roos

BioSemantics Group, LUMC, NL

DaMaLOS Workshop, 2 November 2020

N.QUERALT_ROSINACH@LUMC.NL



EUROPEAN JOINT PROGRAMME RARE DISEASES

Introduction

COVID-19 emergency in the Clinics

Clinical Questions

Questions

- What are the criteria that define the different disease trajectories?
- What are the underlaying mechanistic profiles of the different types of groups?

Need to Link Data Across

LUMC Data

Leids Universitair Medisch Centrum

- Clinical
- Lab measurements
- RNA-Seq
- Metabolomics
- Metavision (ICU)

External Knowledge

Introduction

COVID-19 Global Challenge

Questions

- What are the criteria that define the different disease trajectories?
- What are the underlaying mechanistic profiles of the different types of groups?

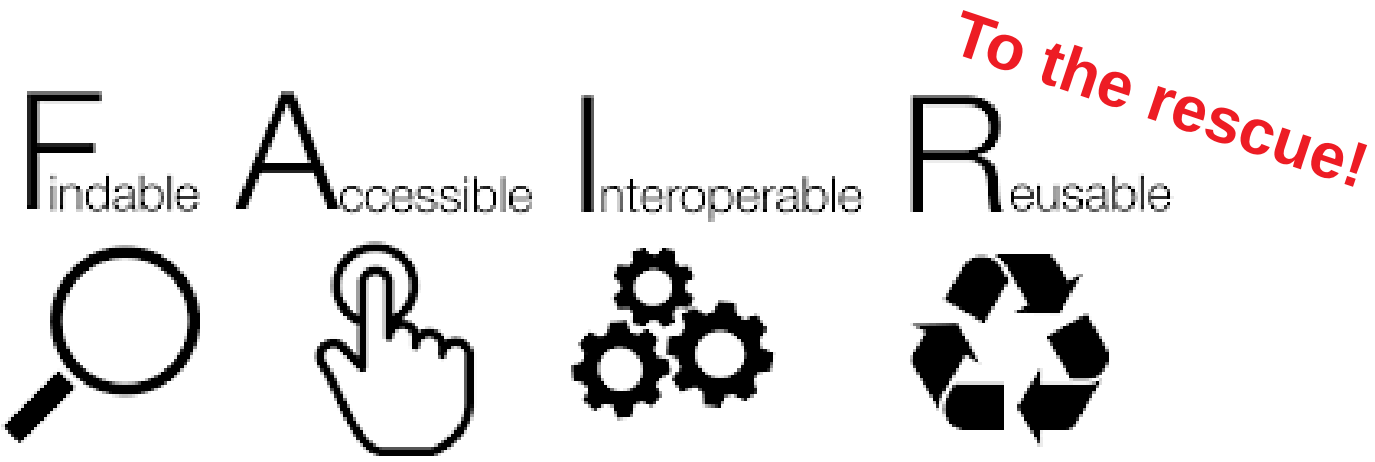
Globally, as of 5:08pm CEST, 29 September 2020, there have been 33,249,563 confirmed cases of COVID-19, including 1,000,040 deaths, reported to WHO.

Research Questions

- How to improve the **access** to patient data in the LUMC for research?
- How to **leverage** patient data with established open biomedical knowledge for research?
- How to **share** patient data in the LUMC with different healthcare institutes?

Hypothesis

Apply the FAIR principles to the LUMC Research Data Management (RDM)



Open Science **Semantic Web**

[1] Wilkinson *et al*. *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data 3, 160018 (2016).

Goals

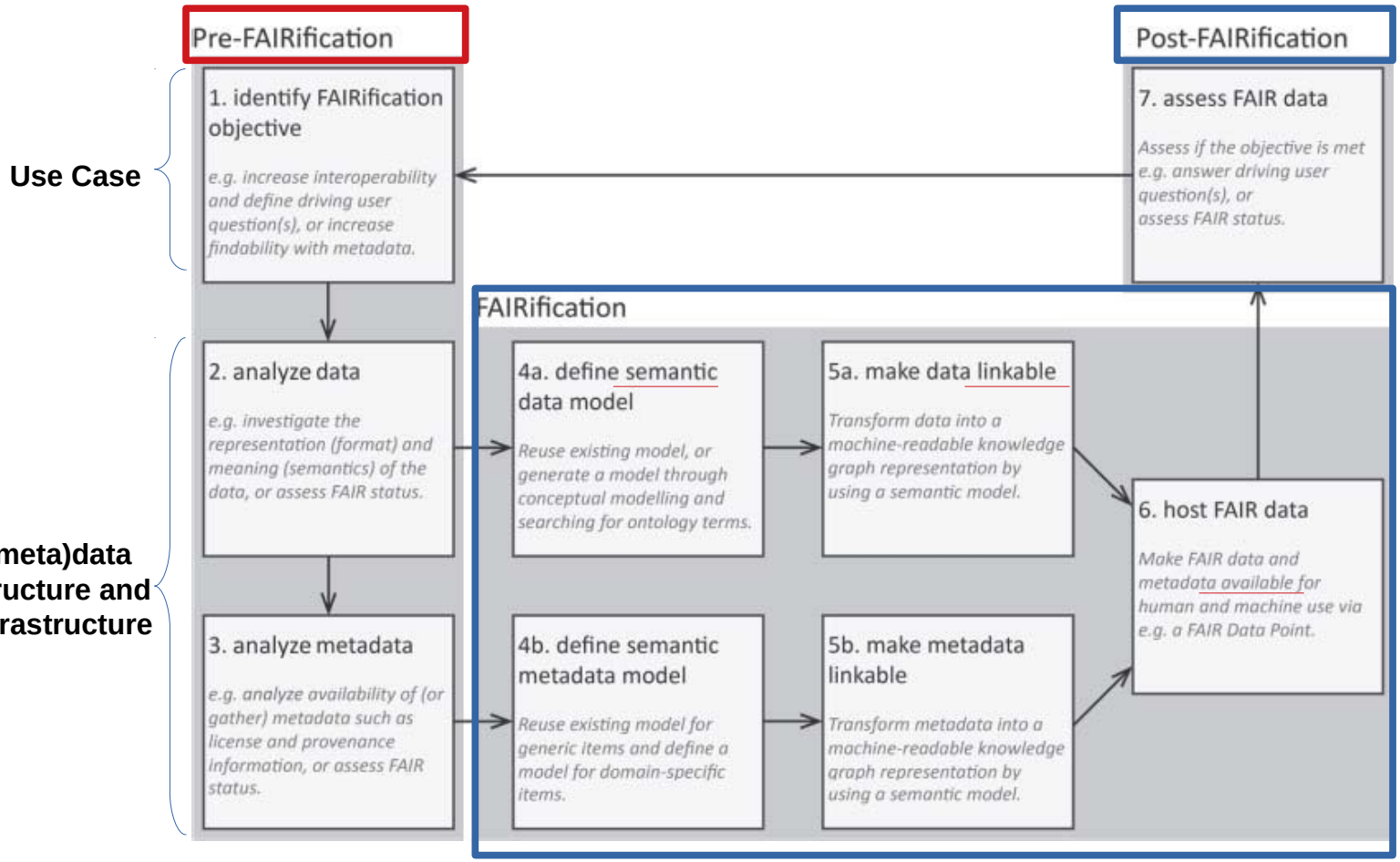
- **Design** a FAIR Research Data Management in the LUMC
- **Implement** the FAIR Research Data Management

Method

- **Beat-COVID team**:
  - Clinical and research groups in the LUMC
  - Multi-disciplinar expertise and **collaborative**
- **Data and Knowledge**:
  - LUMC patient data
  - Established **open** biomedical knowledge
- **FAIRification**:
  - FAIR data
  - FAIR infrastructure
  - **Semantic Web** technologies
- **Evaluation**:
  - FAIR Data Analysis as **Distributed queries**

Method

1- Developing a FAIR Research Data Management Plan



Method

Identifying a Data Management **Goal**

Questions

- What are the criteria that define the different disease trajectories?
- What are the underlaying mechanistic profiles of the different types of groups?

LUMC medical doctors questions **guided** the development of a FAIR RDM plan

Method

LUMC Data **Management**

Different departments

Data Types	Data Storage system	Metadata Publication
Clinical	HiX → Opal	Mica
Lab measurements	Castor → Opal	Mica
RNA-Seq	Castor → Opal	Mica
Metabolomics	Castor → Opal	Mica
ICU	Metavision → Opal	Mica

Opal&Mica (Open Software for Epidemiology)

Method

FAIRness analysis of **LUMC data**

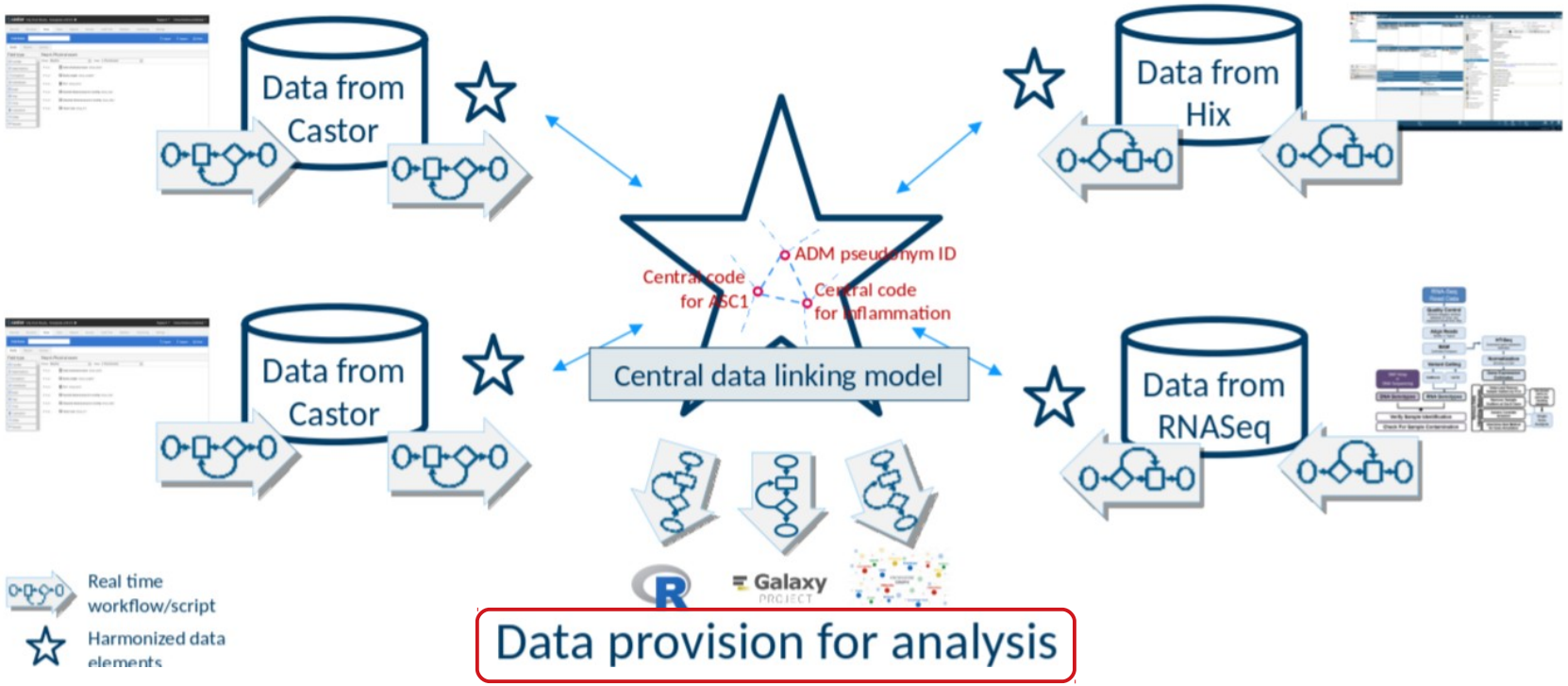
- Data and metadata
- **Representation**: structure and format
- **Meaning**: semantics
- **Existing tools** and databases

Observational clinical **measurements** collected from the laboratories

(**cytokines**, LFA, Glycosylation, Neutralization, NMR, serology, glycocalyx, coagulation, viral load & WGS, cellular, RNA-seq)

Method

FAIR Research Data Management **Plan**

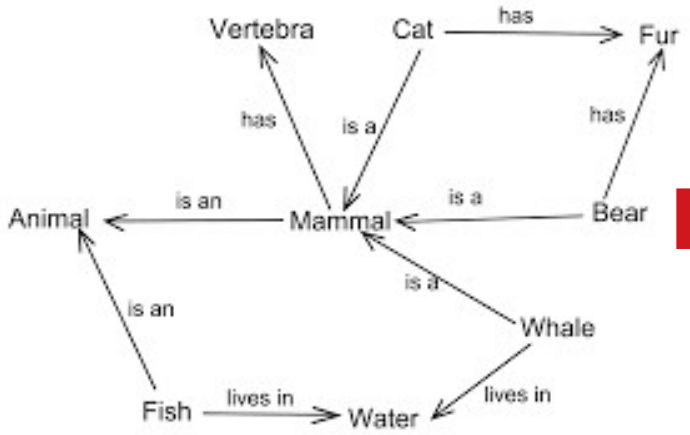


Method

2- FAIRification: Implementing a FAIR RDM Plan with Semantic Web technologies

Improving I in FAIR: *Interoperability*

Semantic Linking Models (*Linked Data*)



Machine-actionable clinical data linkable to LOD

Community

Improving F in FAIR: *Findability*

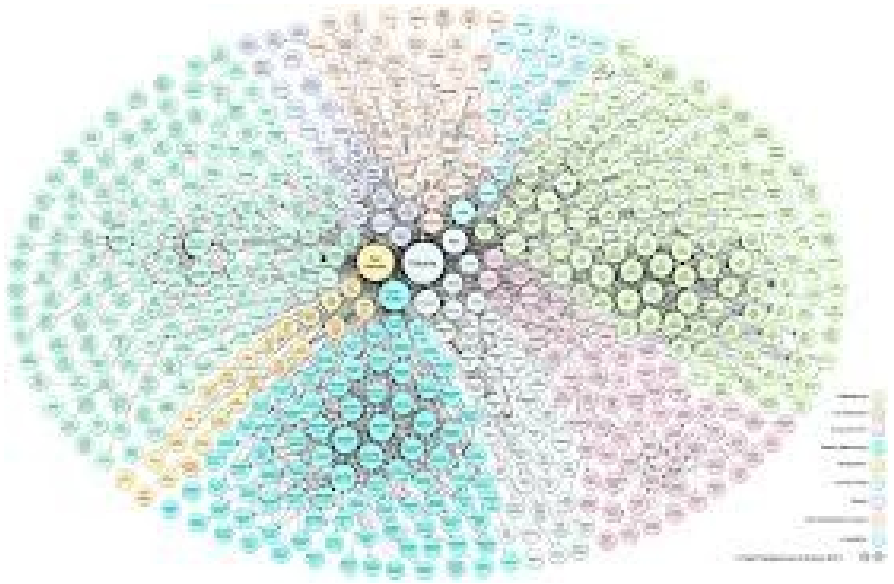
FAIR Data Points *(app&spec DCAT based)*

Machine-actionable clinical metadata

Method

3- Evaluation: FAIR Data Analysis with Semantic Web technologies

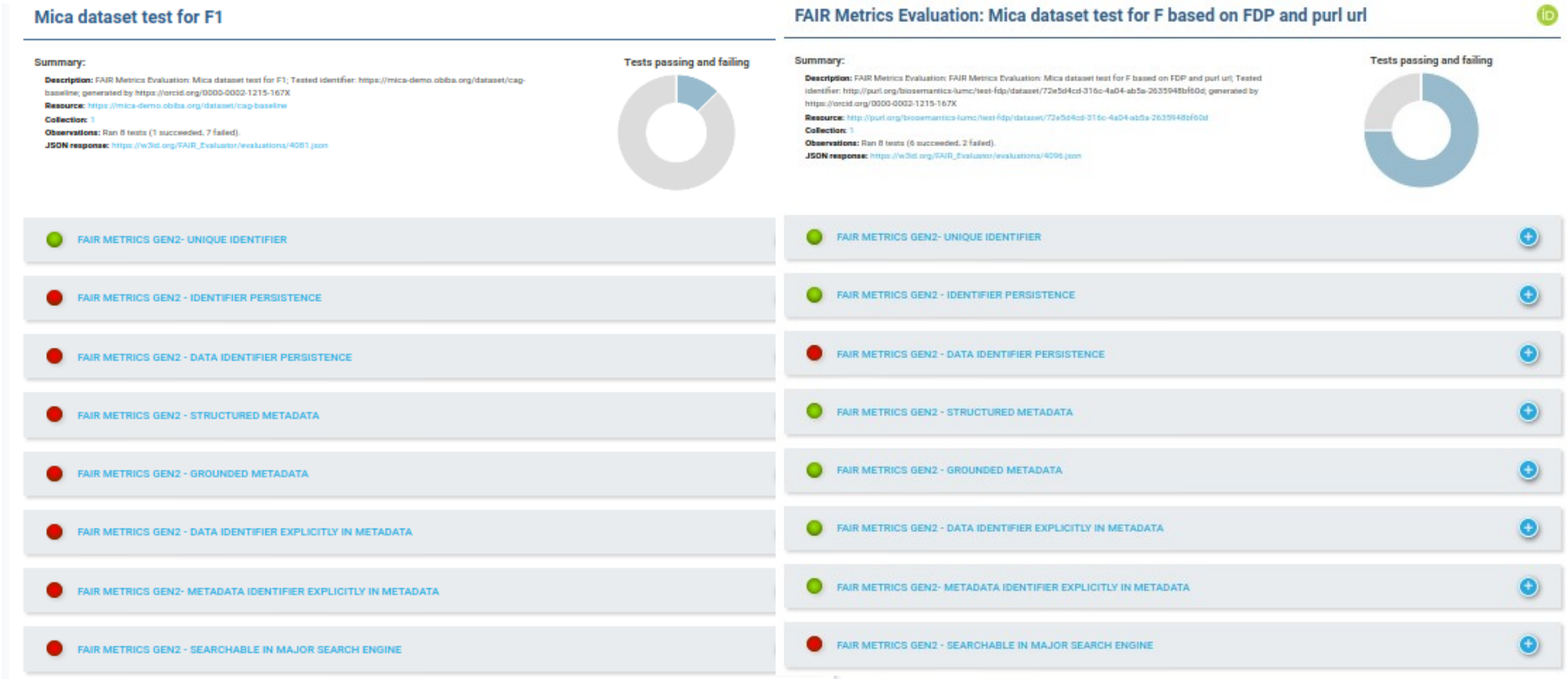
- **Semantic Web** technologies (RDF, OWL, SPARQL)
- Blazegraph **Triple** Store
- **Linked Open Data** (LOD)
- **Queries** over graphs:
    - LUMC data
    - Across distributed graphs: LUMC data + LOD



Results

FAIR status of LUMC data needed **improvement**

- **No semantics** (common identifiers, standards), but Opal&Mica annotation functionality is useful
- **No machine-actionable** metadata



Metadata description improved when the same dataset is described in the FDP

Results

FAIR Research Data Management is a **coordinated** effort

FAIR:
Findable
Accessible
Interoperable
Reusable

Cytokines
Semantic Linking Models
Data providers & users

FAIR Infrastructure
FAIR Data Points
IT & database managers

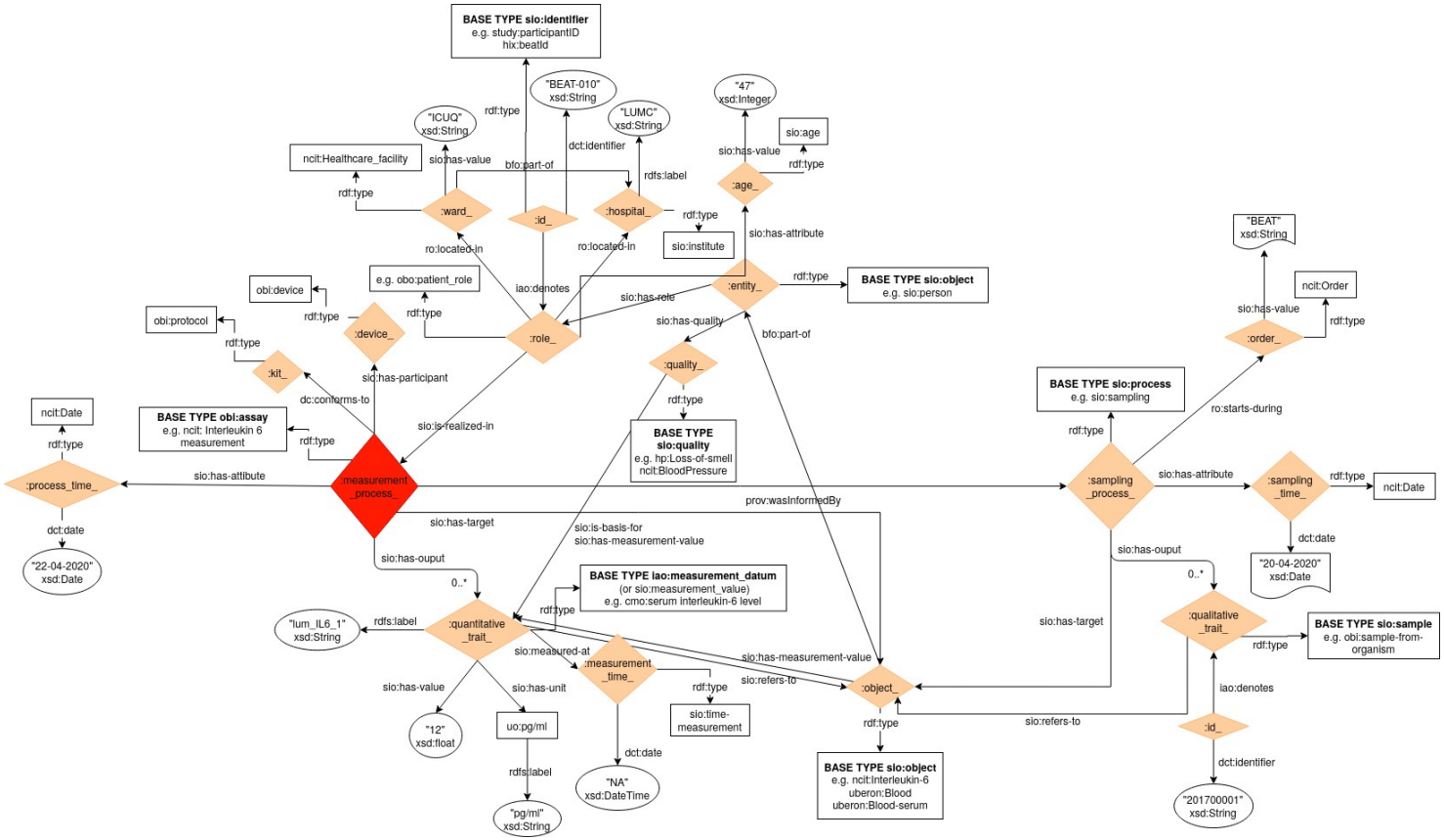
Parallel tracks that sometimes converge

FAIRification is very collaborative

Difficulties: Social, Funding, Governance (patient data privacy)

Results

Semantic models for **interoperability** of clinical measurements: **Cytokines**



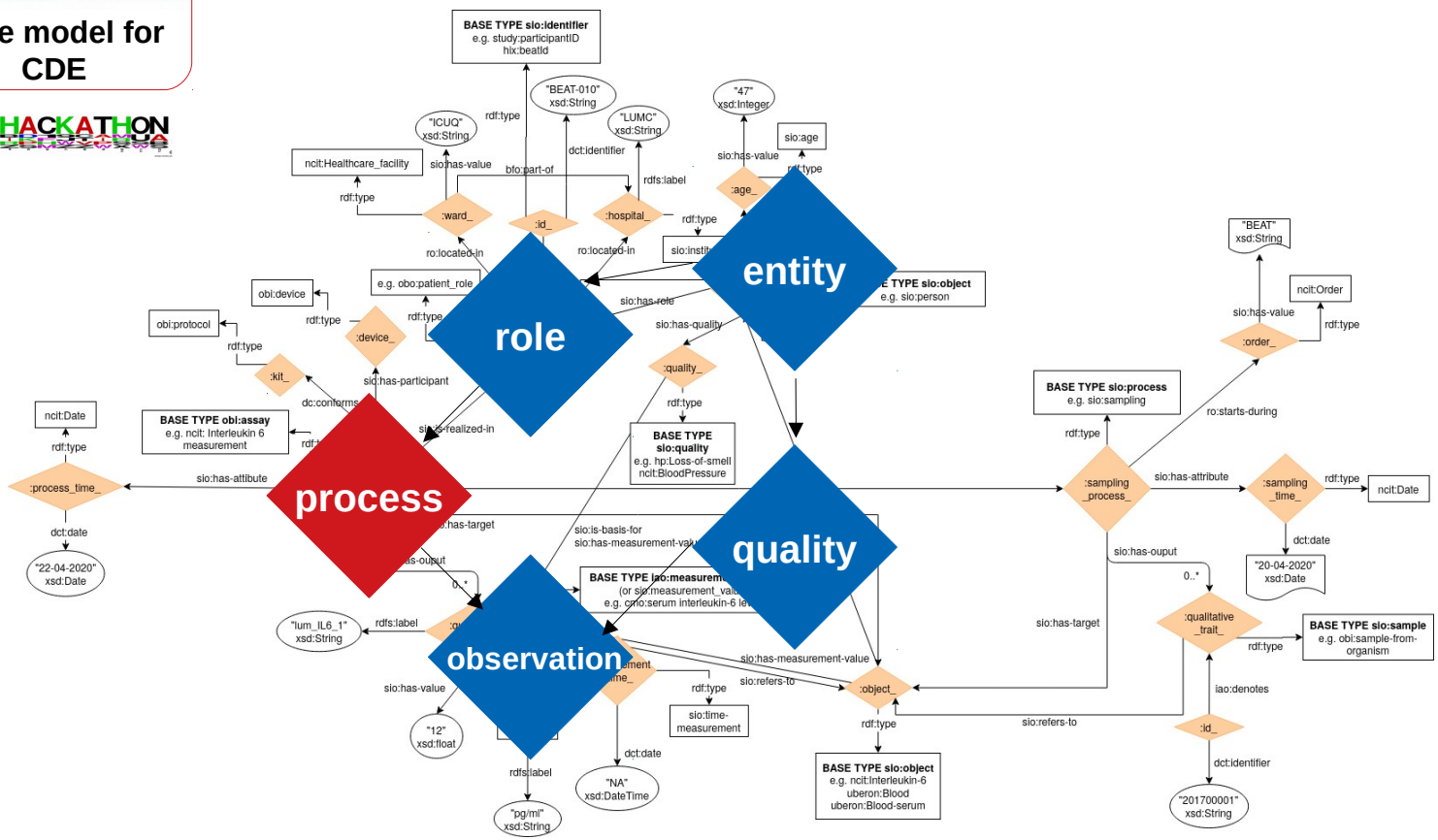
Results

Semantic models for **interoperability** of clinical measurements: **Cytokines**

EUROPEAN JOINT PROGRAMME RARE DISEASES

Core model for CDE

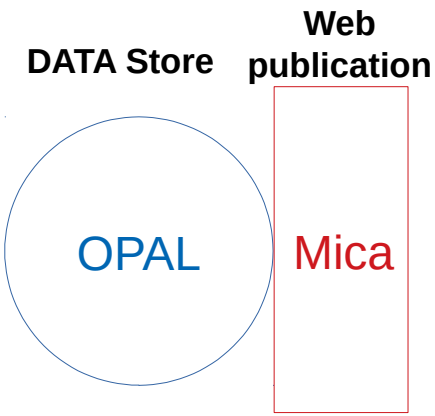
BIOHACKATHON



Reuse of existing models allows interoperability

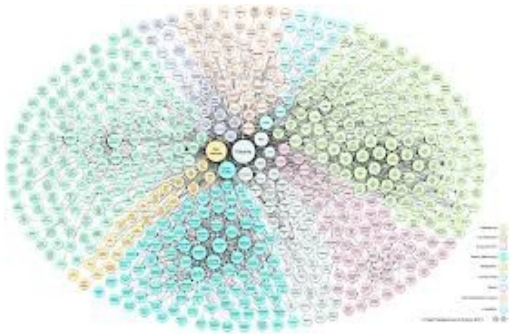
Results

LUMC FAIR Data Points for **findability** of patient data



Data **don't leave** its own repository

FAIR Data Point
(FDPs)



Visiting algorithms

FDPs publish structured metadata for **machines to interpret how to access**

Results

Querying FAIR patient data for medical questions

Distributed Analytics: EU and intercontinental

- FAIR at source
- FDPs: open, secured shared data
- SPARQL queries

- **“count number of patients”**
- **“retrieve LUMC cytokines measurements with protein annotation from UniProt”**

FAIR RDM allow querying patient data with external **open** science knowledge

Discussion and conclusion

- We provide the first **FAIR Research Data Management** plan for FAIRifying health research data in the hospital
- FAIRification **difficulties**:
  - 'Social'
  - Technical
  - Data privacy
- We provide COVID-19 **observational patient data** as FAIR research objects ready to reuse
- **Our first results show that a FAIR Research Data Management plan based on open Science, Semantic Web technologies, and FAIR Data Points is providing data infrastructure in the clinics for FAIR research linkable to established biomedical knowledge for analysis**

Leiden University Medical Center

Acknowledgements

BioSemantics Group, specially:

Marco Roos, Rajaram Kaliyaperumal, César Bernabé, Qinqin Long

LUMC Beat COVID team, specially:

Sesmu Arbous, Jacqueline Janse, Henk Jan van der Wijk, Simone Joosten, Hailiang Mei, Erik Flikkenschild

Colleagues, mentors, and inspiring scientists, specially:

Carole Goble, Mark Wilkinson, Robert Hoehndorf, Barend Mons

Q&A

THANK YOU!