Collecting FAIR Data for Biomedical Research

Dr Katy Wolstencroft
LIACS, Leiden University LUMC Biosemantics

Universiteit Leiden
The Netherlands

FAIR Initiatives

NL-BIOIMAGING AM

OME

FAIRDOM HUB

FAIR by design and FAIR by increment

FAIRifying Existing Data

Hebon

Onderzoek naar erfelijke borst- en eierstokkanker

BENEFIT FOR ALL

Who Benefits from FAIR Data?

- More work for the scientists making FAIR data
  - Not seen as part of the scientific process
- Others can reuse data and benefit straight away
  - Time and cost of generating data must be 'worth it'

Need incentives for individual scientists and for projects → Data citation and credit for further funding

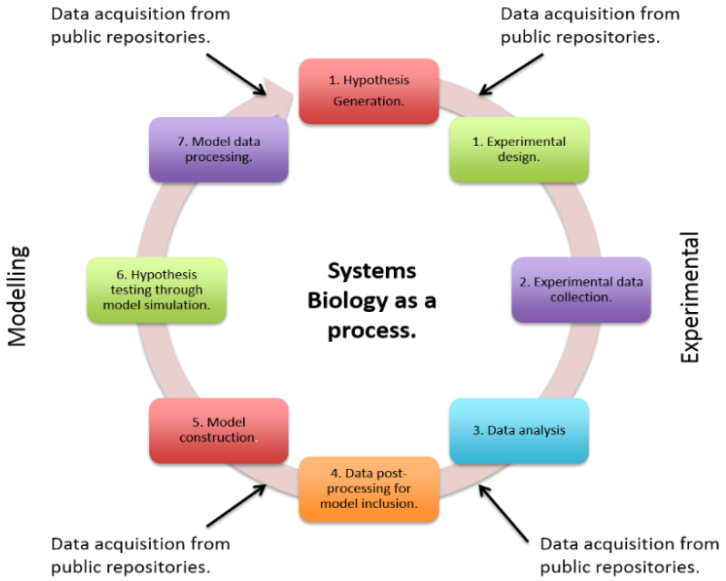
- My data is privacy sensitive, I can't share
  - Are there benefits from integrating public knowledge
  - Do you need to collaborate to make it useful?

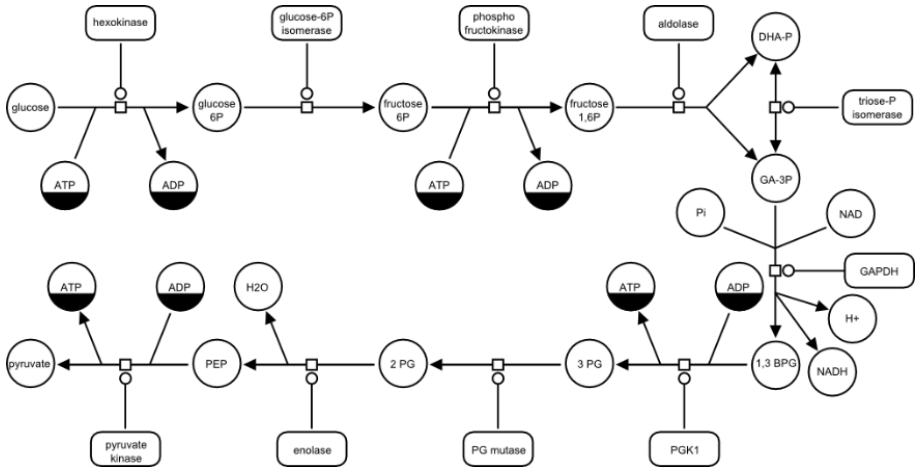
FAIR by Increment and by Design
FAIRDOM Systems Biology

FAIRDOM

Findable
Accessible
Interoperable
Reusable

Data
Operations
Models

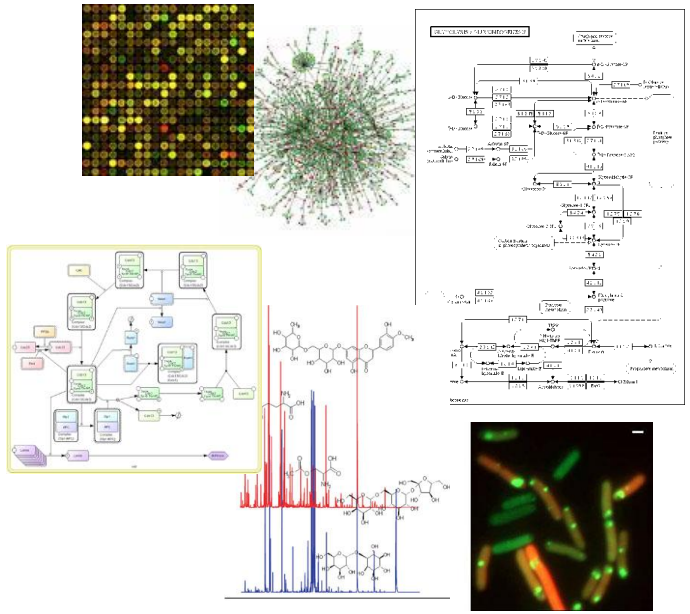




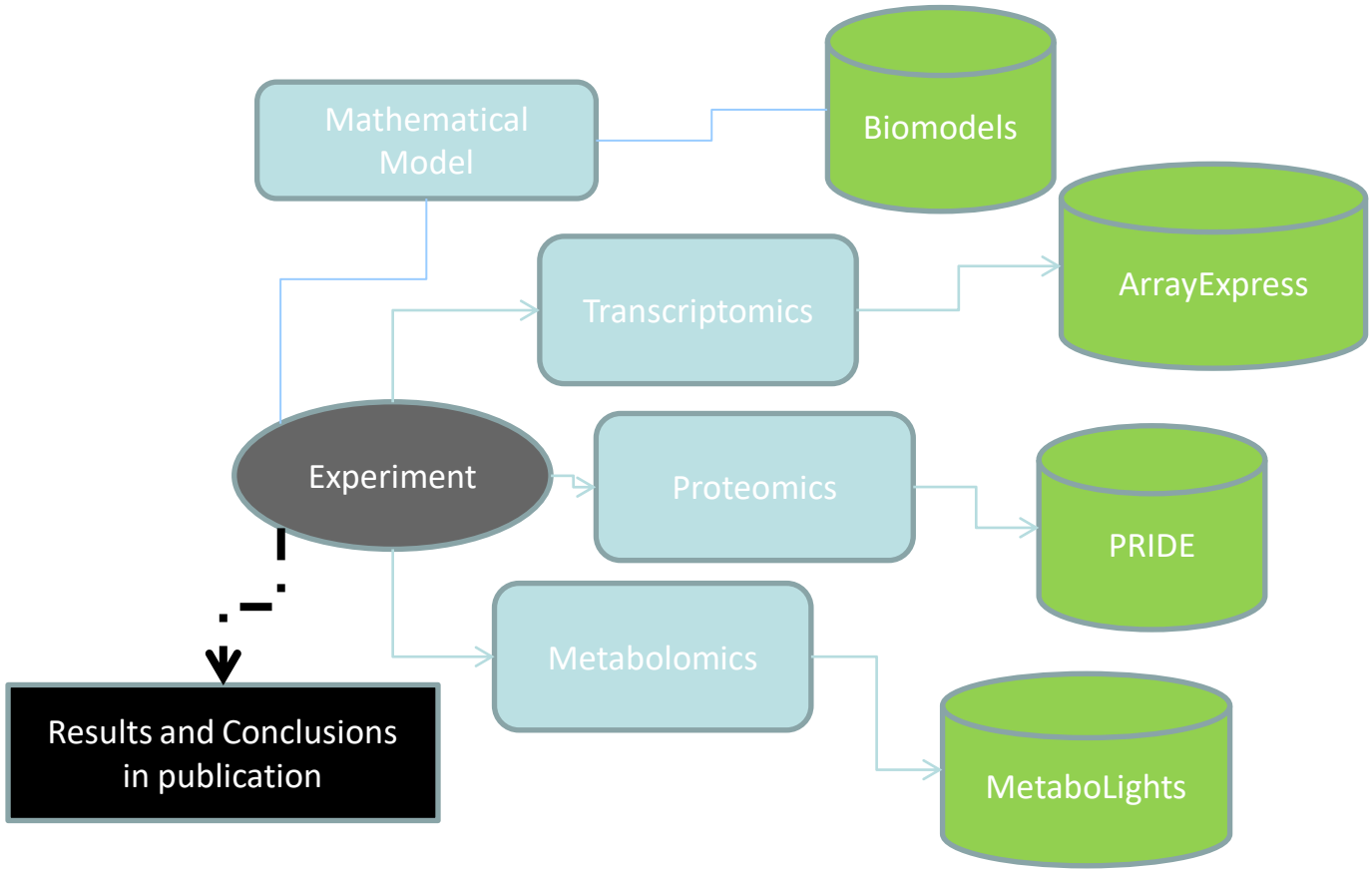
- Model simulations predict behaviour in different conditions
- Experimental measurements for enzyme reaction rates, metabolite concentrations, expression of enzymes etc
- ODE, PDEs, agent models, stochastic models
  Matlab, R, Mathematica

Challenges For Systems Biology: Heterogeneous data and other Research Assets

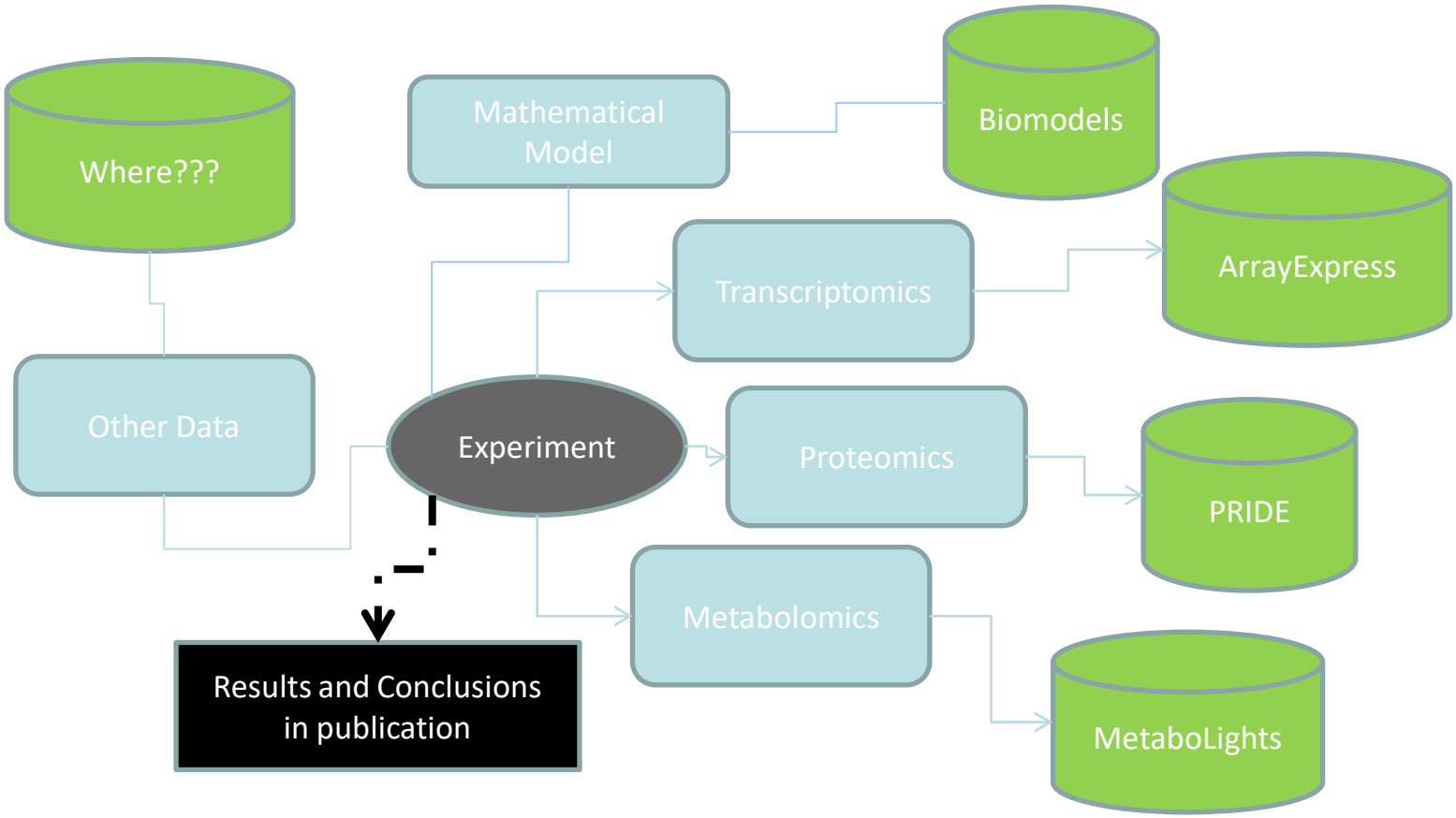
- Multiple omics
  - genomics, transcriptomics
  - proteomics, metabolomics
  - fluxomics, reactomics
- Images
- Molecular biology
- Reaction Kinetics
- Models
  - Metabolic, gene network, kinetic
- Relationships between data sets/experiments
  - Procedures, experiments, data, results and models
- Analysis of data



Systems Biology Experiments and Data Repositories



Systems Biology Experiments and Data Repositories



Aims of the FAIRDOM Platform

- Share and exchange data, models and other research assets
  - Systems biology, synthetic biology
- Share in the context of the whole experiment
  - Linking and integrating datasets
  - Linking models and data
  - Understanding the relationships between them
- **Provide a stewardship environment for research assets**
  - **Support project whilst they run, not just as a repository afterwards**

Making Results Reusable

- Common standards
  - Identifiers, metadata descriptions, exchange formats, vocabularies
- Common repositories
  - Enables search, aggregation and meta-analysis
- Promote good practices

- Infrastructure and tools
  - Lowers the barrier for participation

Systems Biology Standards landscape

	Data	Models	Simulation	Results
Minimal Information Models *checklists*	MIBBI	MIRIAM	MIASE	
Standard Formats *markup*	MAGE-TAB	SBML, cellML	SED ML	SBRML
Ontologies *Controlled vocabularies* *meaning*	BioPortal, OBO, GO the Gene Ontology, MGED	SBO, OPB	KISAO	TEDDY

[Nicolas Le Novere]

Type of data/model	What was measured	Values in the datasets	Biological sample and treatments applied
Microarray, growth curve, enzyme activity…	Gene expression, OD, metabolite concentration…	Units, time series, repeats…	

Common elements

Enzyme reactions

reactions catalyzed, substrates, products, inhibitions

CheBI ids

Microarray

QC methods
normalisation

MGED/EFO

Proteomics

instruments

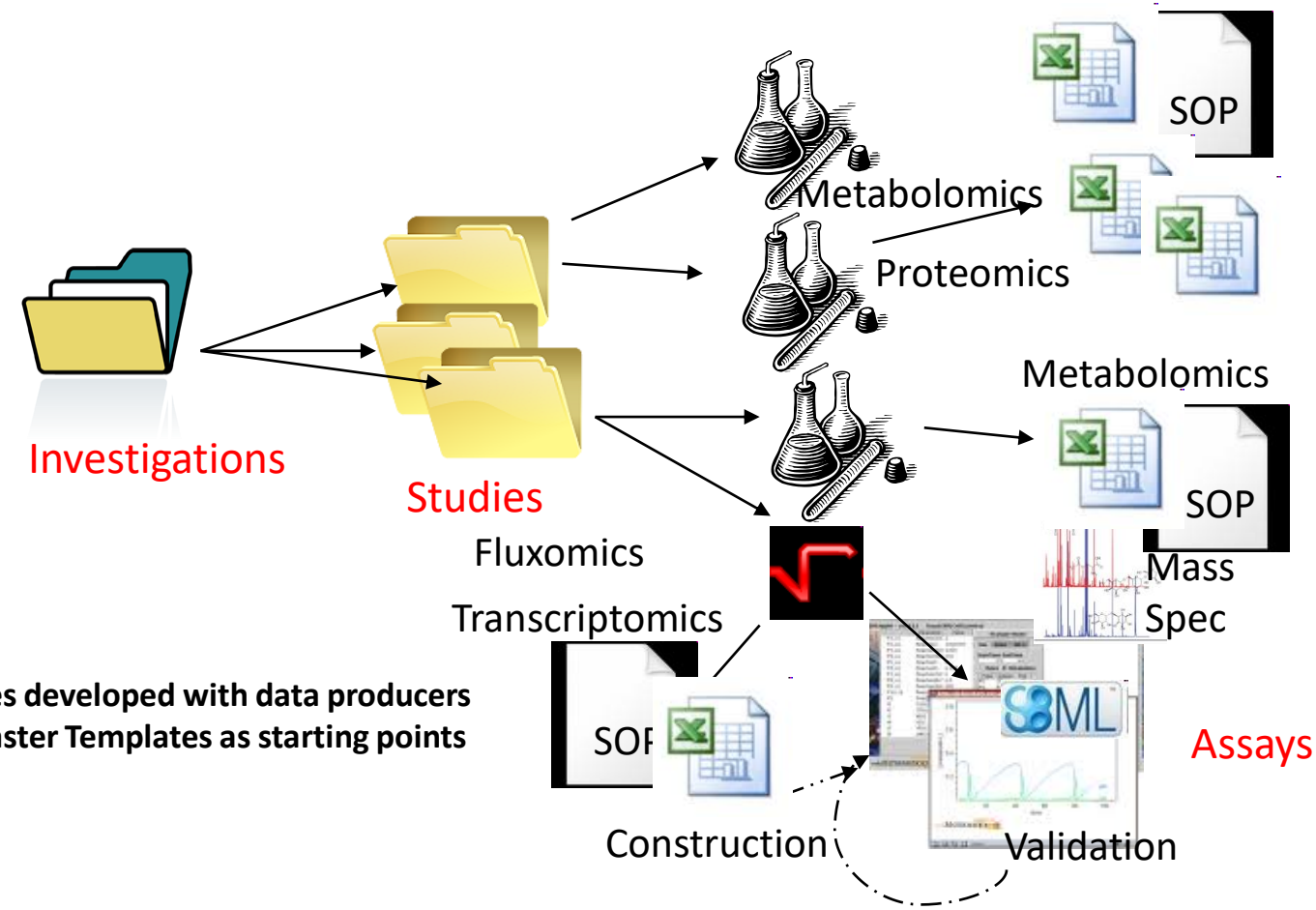
PSI-MS

Data type specific elements

Just Enough Results Model – Application Ontology and annotation vocabulary for FAIRDOM Metadata

Based on MIMs and ISA

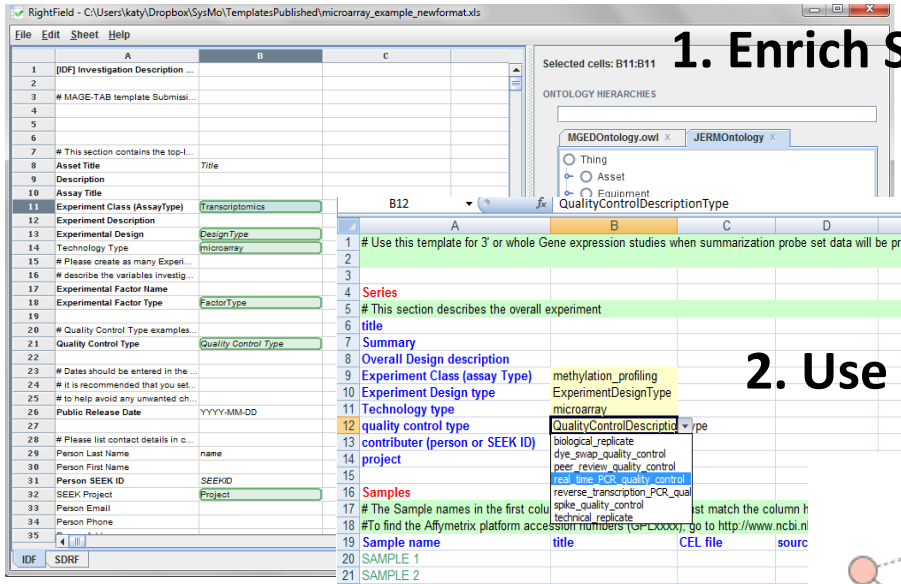
Aggregating Experiments: ISA



- **Templates developed with data producers**
- **JERM Master Templates as starting points**

RightField

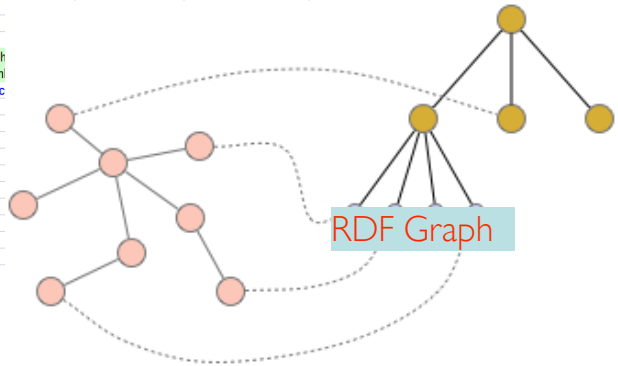
Semantic Annotation by Stealth



1. Enrich Spread sheet Template

2. Use in Excel or OpenOffice

3. Extract and Process in SEEK



FAIRDOMHub.org: Facilitating FAIR Collaboration and Sharing in Systems Biology

FAIRDOM

- SOPs
- Data
- Models
- Samples

FAIRDOM HUB

isa

Write API

Other tools

JWS ONLINE

Model tools

jbeir public

SABIO-RK

External Databases

ChEBI

PRIDE

PubMed

zenodo

Publishing services

Secure Stores

National Stores

Local Stores

openBIS Biology Information System

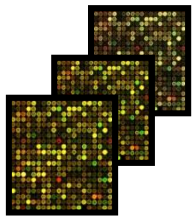
LabCollector

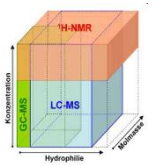
RightField

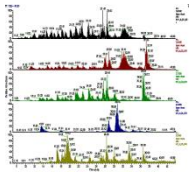
JERM

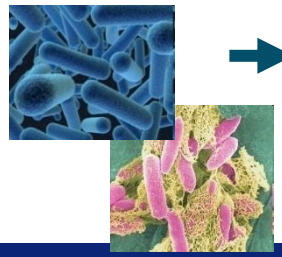
Omics data

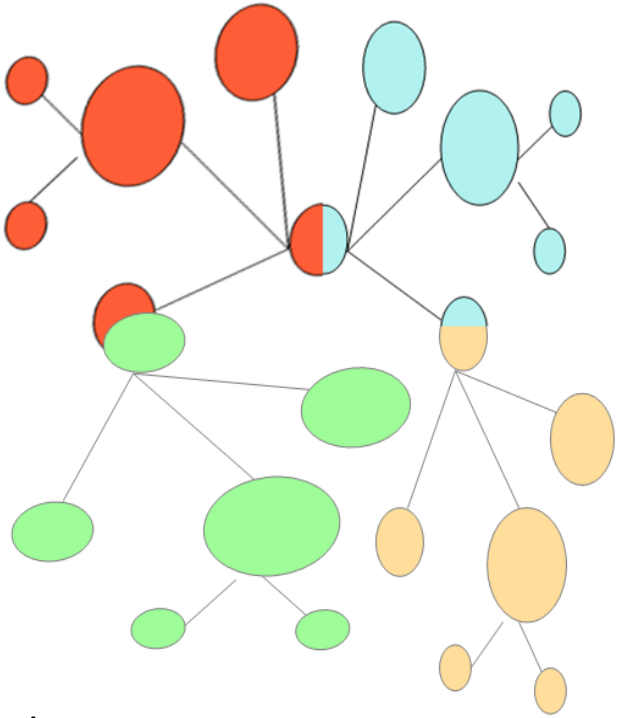
standard based templates









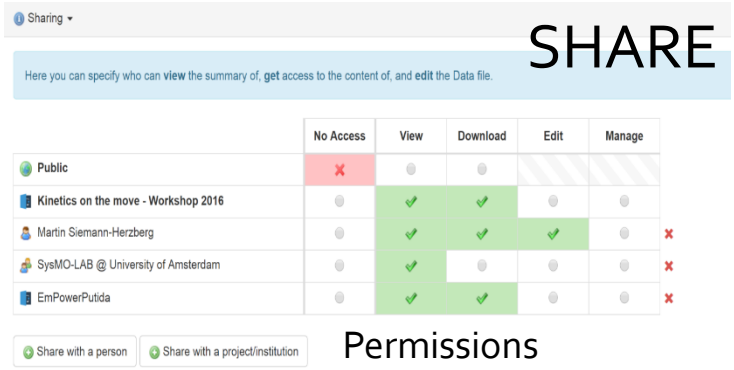


RDF /Linked Data
Semantic Linking and sharing →
'by stealth'

Project Commons
Organise > Share > Disseminate

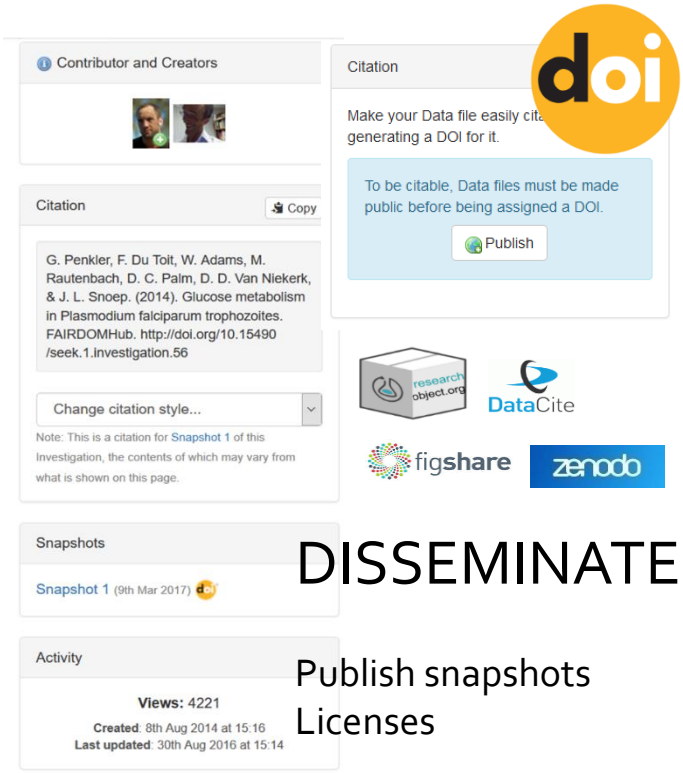
Self managed spaces
Yellow pages

Single entry point to assets



SHARE

Permissions



DISSEMINATE

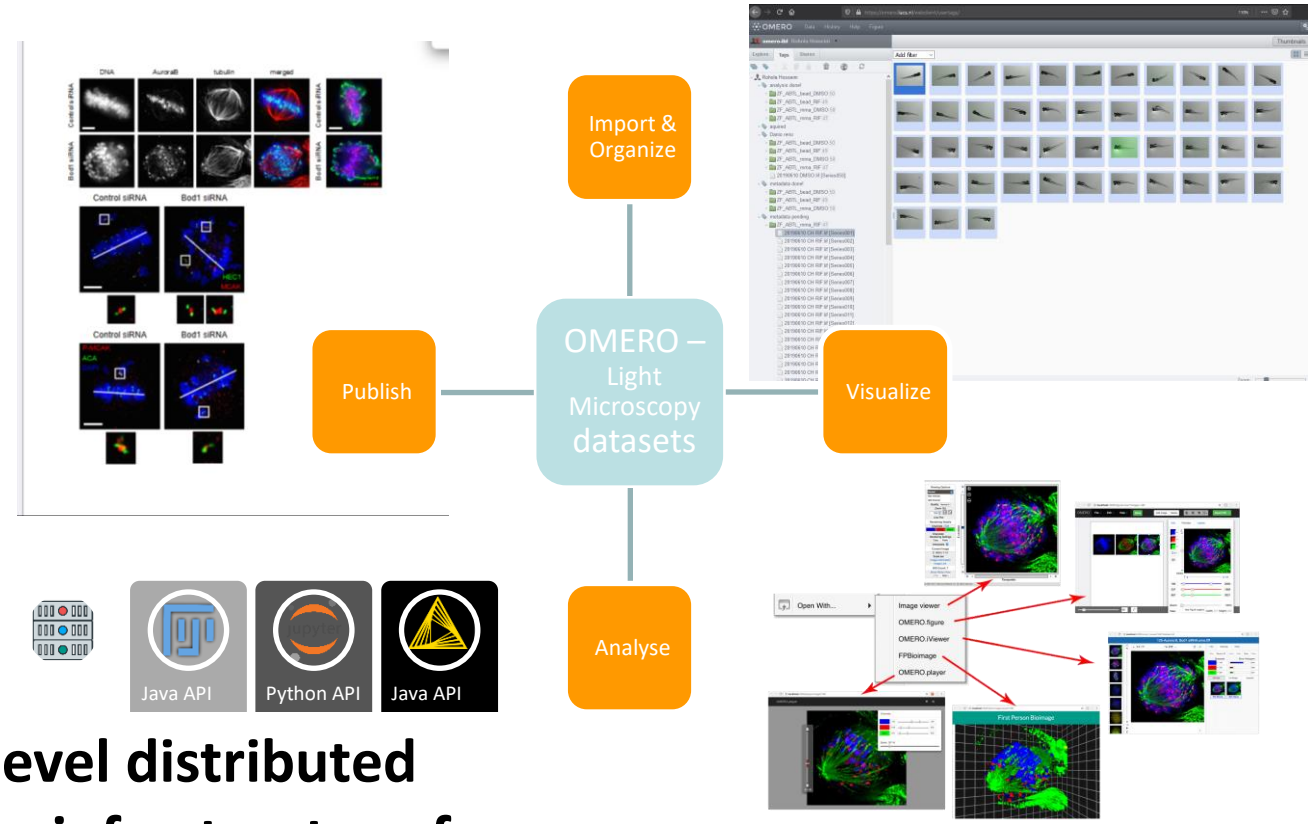
Publish snapshots
Licenses

FAIRDOMHub, Local FAIRDOM instances, as a component in the ELIXIR Converge toolkit

WorkflowHub

Leiden FAIR Cell Observatory:

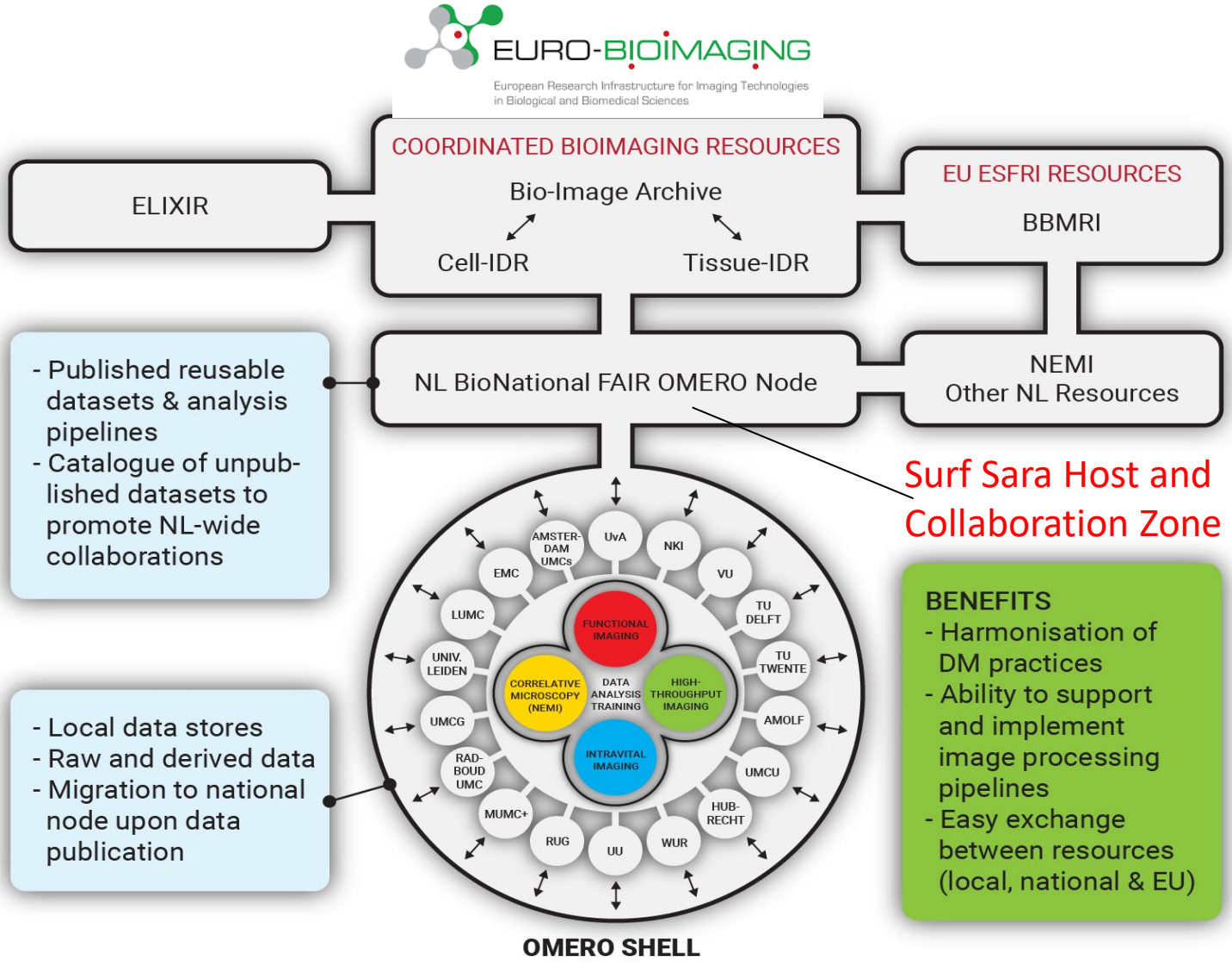
No Reinvention
Common Standards
Open Microscopy
Environment



Aim to provide top-level distributed advanced microscopy infrastructure for life sciences research in the Netherlands

Part of EU Bioimaging

National Roll-Out



DM Pilot@ UL and Beyond

ELIXIR

Workflows for IDR/Bioimage Archive Deposit

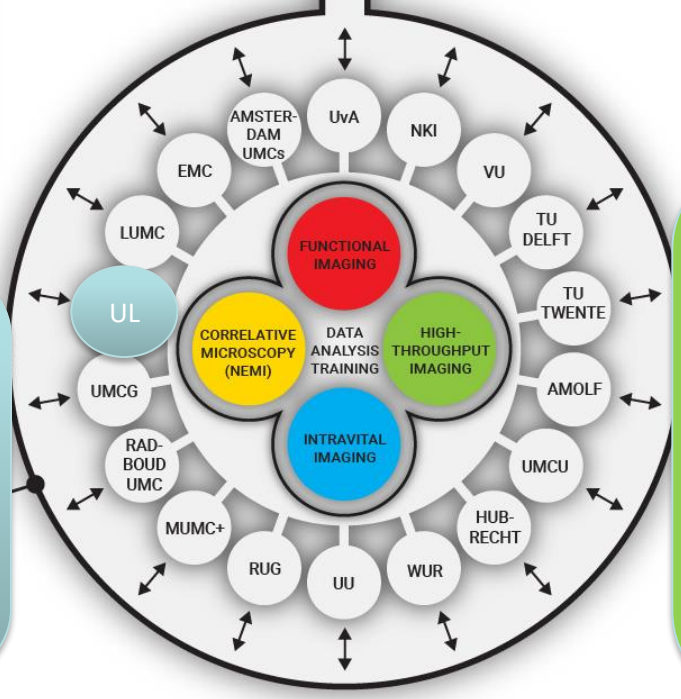
EU ESFRI RESOURCES

BBMRI

- **Faculty wide roll out Nov 2020**
- Serves 4 institutes
- 50 microscopes and 150 users
- Links to ELNs

"National" node @ LIACS

NEMI/NL-BI Team



OMERO SHELL

- **Data management**
- **Metadata standards**
- **Image analysis methods**
- **Image analysis processes**

Benefits

- Images from all vendors
- FAIRer Data
- Data management cheaper
- Data archiving easier
- OMERO worldwide dev community > 3000 instances

FAIR by Increment Requires:

- No reinvention
- Fitting in with common standards
- Fitting in with common practice
- Respecting the requirements of individual scientists as well as PIs/consortia – scientists remain in control
- Incremental development – solution now and innovate with pioneers
- Ensure data complies with new requirements for sharing and archiving data

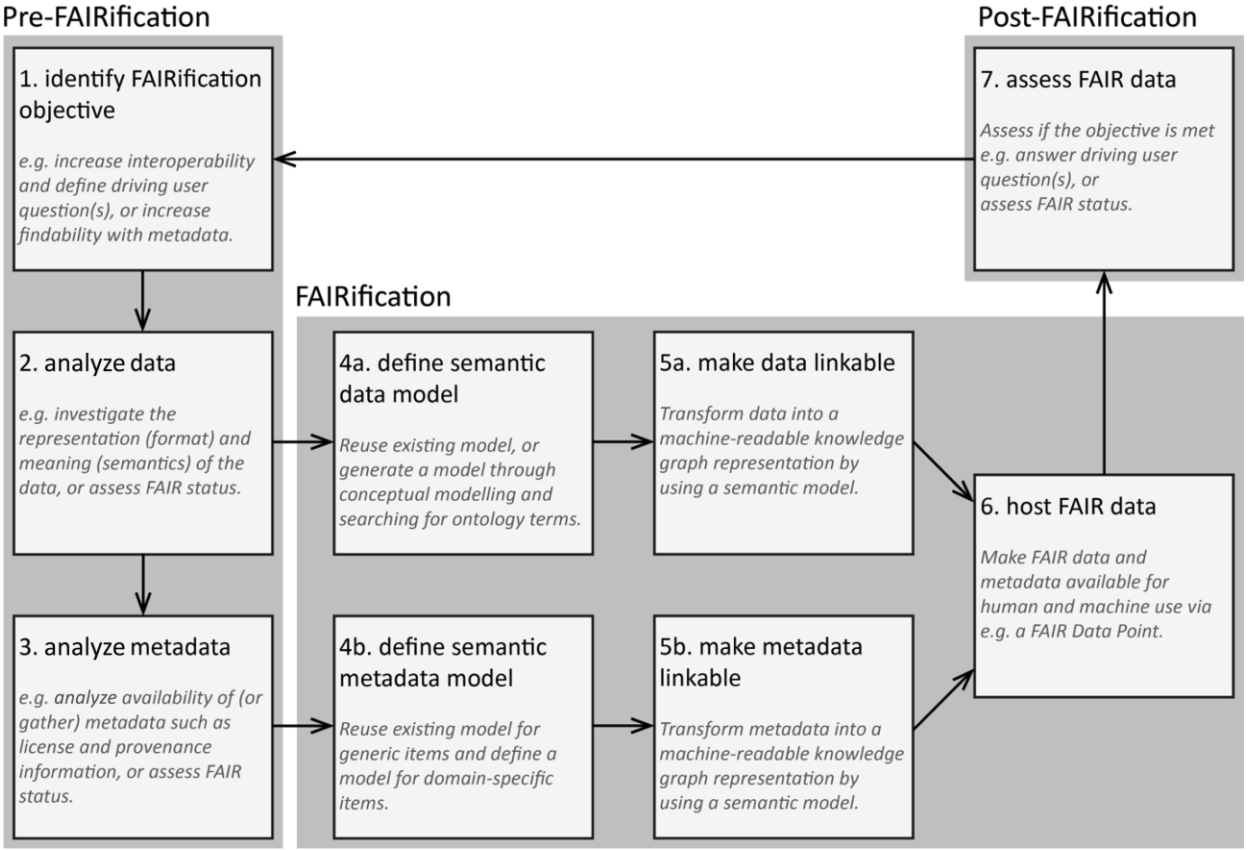
Iterations

FAIRification of Existing Data: Changing Practices →FAIR by Design

Universiteit Leiden
The Netherlands

FAIRification of Existing Data

A Generic Workflow for the Data FAIRification Process, Jacobsen et al, 2020 Data Intelligence 2 (1-2): 56–65



- Rare disease registries
- Rare disease datasets
- Covid data integration

FAIRifying Existing Data

HEBON: Prof Peter Devilee LUMC, Tushar Mandloi
Leiden Centre for Computational Oncology

- Hereditary Breast and Ovarian Cancer research Netherlands
- Nationwide survey of families where breast and ovarian cancer is common.

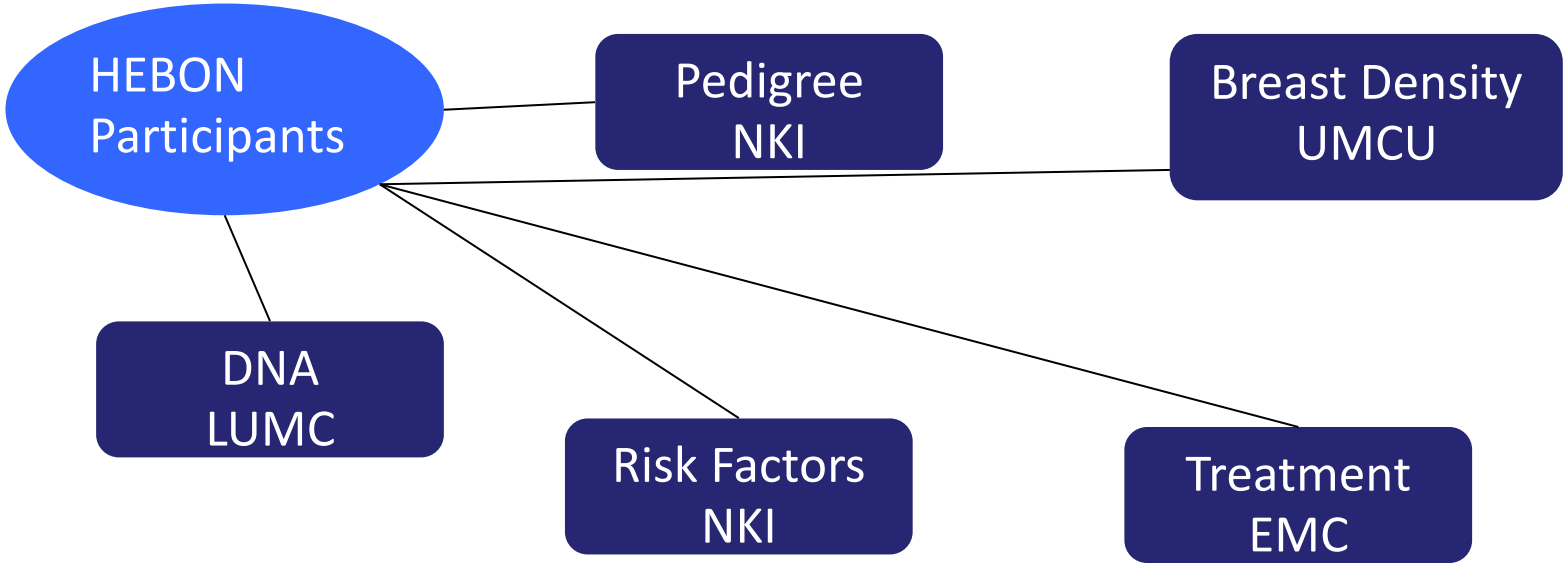
HEBON FAIRification Objectives

- Data collected and managed in different nodes
- Research projects request access to integrated data sets, which are hard for HEBON network to generate

- Data has been collected over decades
- Data needs to be compared to current knowledge and updated

- **Genomic Association** of BRCA mutation-status with pedigree information and risk factors, and to compute polygenic risk scores

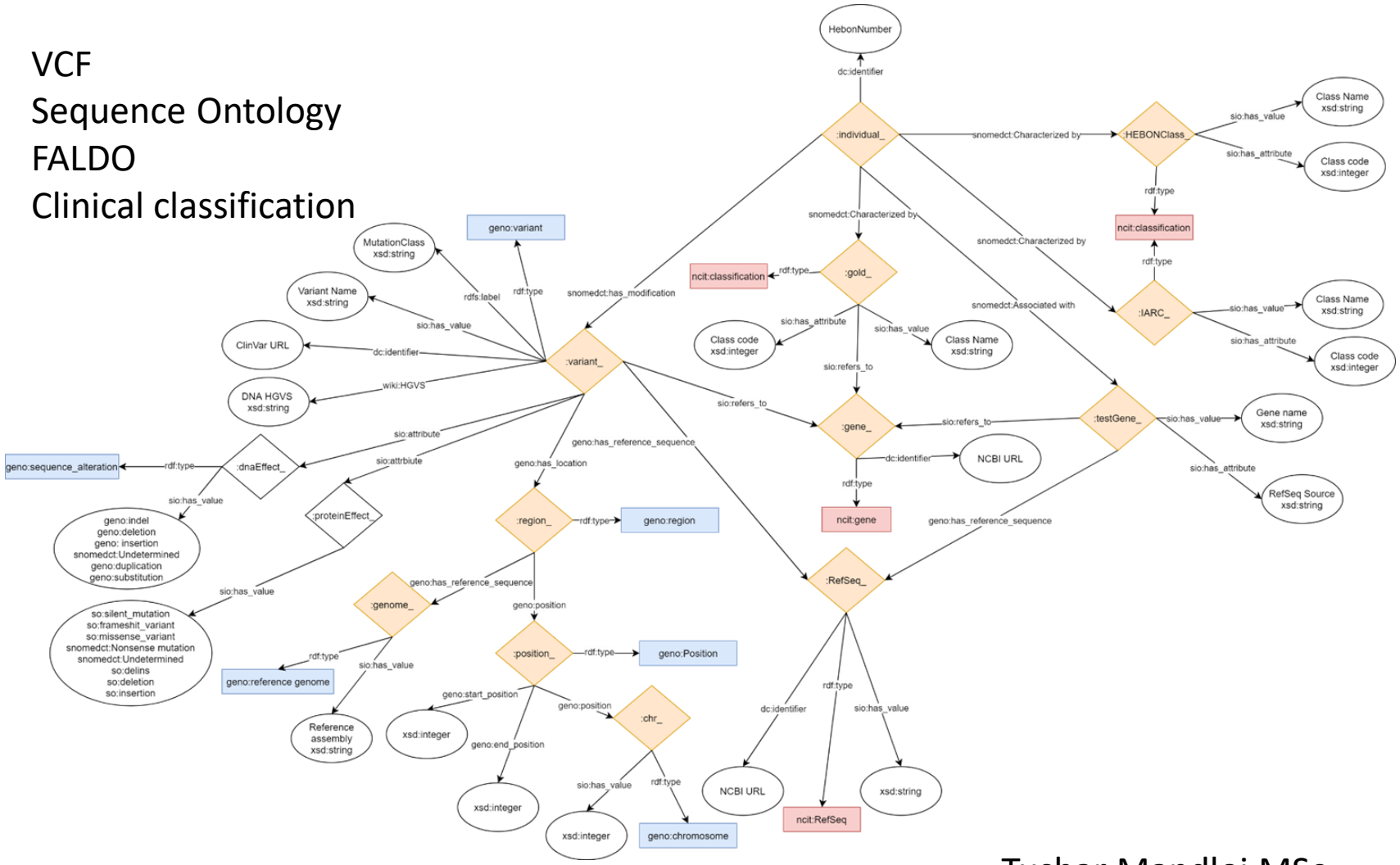
HEBON Databases



Genomic Association of BRCA mutation-status with pedigree information and risk factors, and to compute polygenic risk scores

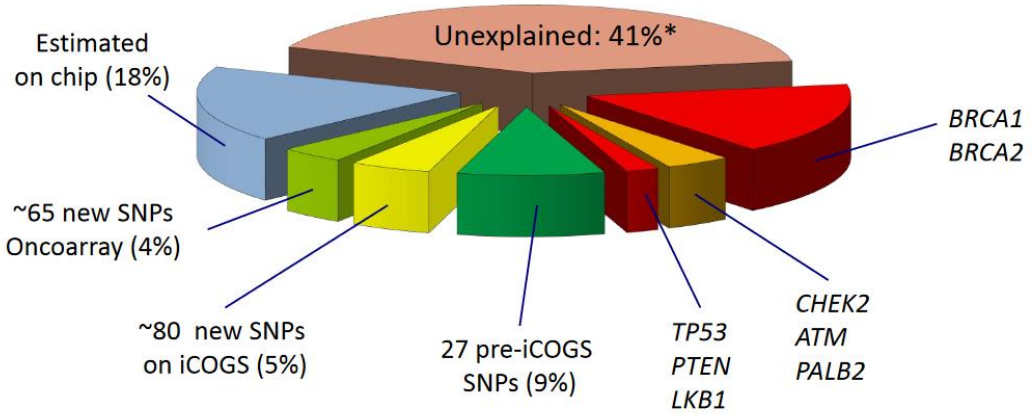
Hebon FAIR Semantic Model Variants

- VCF
- Sequence Ontology
- FALDO
- Clinical classification



Tushar Mandloi MSc
bioinformatics thesis 2021

HEBON DNA @LUMC



* For overall breast cancer in Europeans (Lower for ER-negative disease, early onset disease, and breast cancer in non-Europeans)

- 25 years of research
- New genes
- New pathogenic variants
- New risk variants

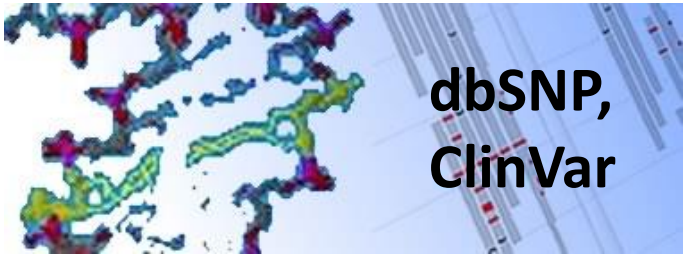
Compile all test-results per individual → Run queries → "Gold" classification BRCA1 and BRCA2 → Update query → Add to central Hebonnrs table

1 = Carrier of one single pathogenic BRCA1-variant
2 = Carrier of two or more pathogenic variants in BRCA1
3 = Carrier of VUS in BRCA1 and not a pathogenic BRCA1-variant
4 = Non-carrier of the pathogenic BRCA1-variant segregating in the family
5 = Non-carrier of the BRCA1-VUS segregating in the family
9 = No class IARC3-5 variants known

HEBON Data can now query public resources

FAIR HEBON DNA

HEBON Results
And provenance

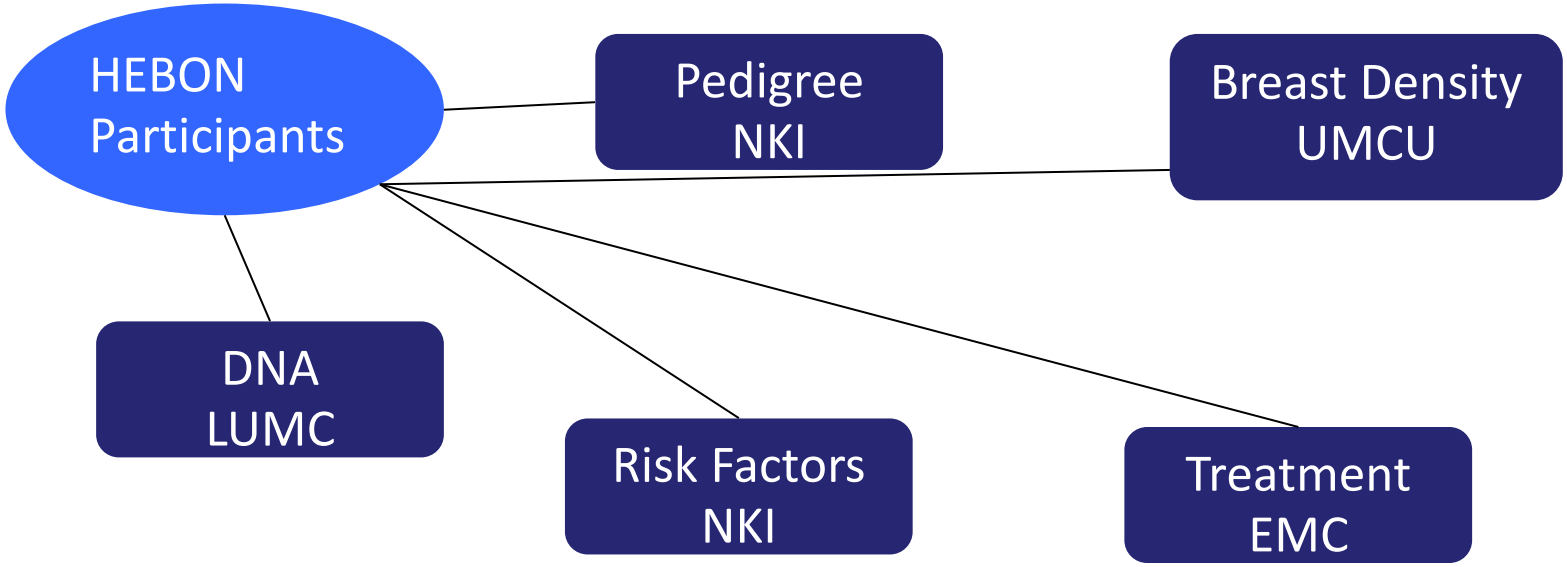


HGMD®
The Human Gene Mutation Database
Cardiff

LOVD 3
Leiden Open Variation Database

COSMIC
Catalogue Of Somatic Mutations In Cancer

FAIR HEBON Databases



Genomic Association of BRCA mutation-status and other variations with pedigree information and risk factors, and to compute polygenic risk scores

Promoting and Enabling FAIR Data

- Have clear FAIRification goals
- Make it as easy as possible
- Make sure there are incentives for individuals, Pis and institutions
- Added value of FAIR data is a large incentive

Acknowledgements

- FAIRDOM Consortium
  - Carole Goble, Stuart Owen
- Cell Observatory OMERO team
  - Rohola Hosseini
- Benefit Consortium
  - Linda Breeman
  - Erik Flikkenschild
- HEBON Network
  - Peter Devilee, Tushar Mandloi
- LIACS semantic systems bioinformatics
- Biosemantics group