

# Data First !

## FAIR Principles Implementation at Roche A Pharma Perspective

Martin Romacker & Nick Perry  
Product Managers  
Enterprise Data  
Digital Integrations Generating Insights (DIGI)  
Roche Informatics, Basel

06 April 2023, OntoCommons Workshop, Berlin

# Table of contents

1. Data Economics
2. FAIR & Roche Data Commons
3. FAIR is simple and beautiful
4. FAIR is complex and ugly
5. Pharma Interoperability Hub
6. Conclusions
7. Acknowledgements



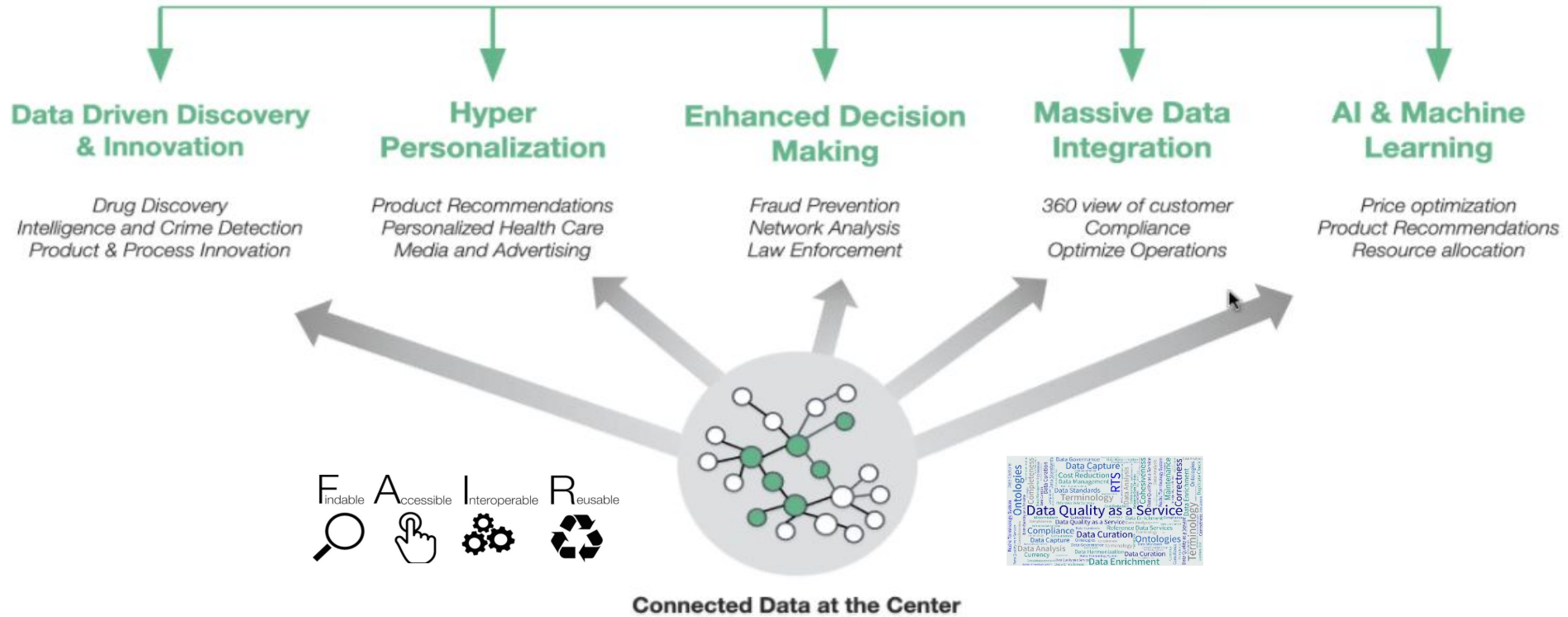
# Data Economics

# Megatrends in Digitilization

FAIR plus Q Data

## Harnessing Connections Drives Business Value

### Digital Transformation Megatrends



**Data Standards: Terminology, Metadata, Dataset Models & Ontology (FAIR+Q Data)**

# Planned and unplanned Costs in Data Management

Business Case for Prospective FAIRification and High Data Quality

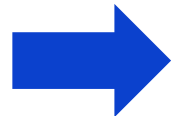


## Planned/ Visible Costs

- FTEs creating Data Asset
- Material procurement (sample, reagent, compounds etc.)
- Infrastructure

## Unplanned/ Invisible Costs

- Business Analysis
- ETL processes/ Data Cleansing
- Searching & accessing
- Data Curation/ Semantic Data Integration
- IT Infrastructure supporting unplanned activities



Backcharge the costs for processing to the data producers



# Standards in Pharma Industry

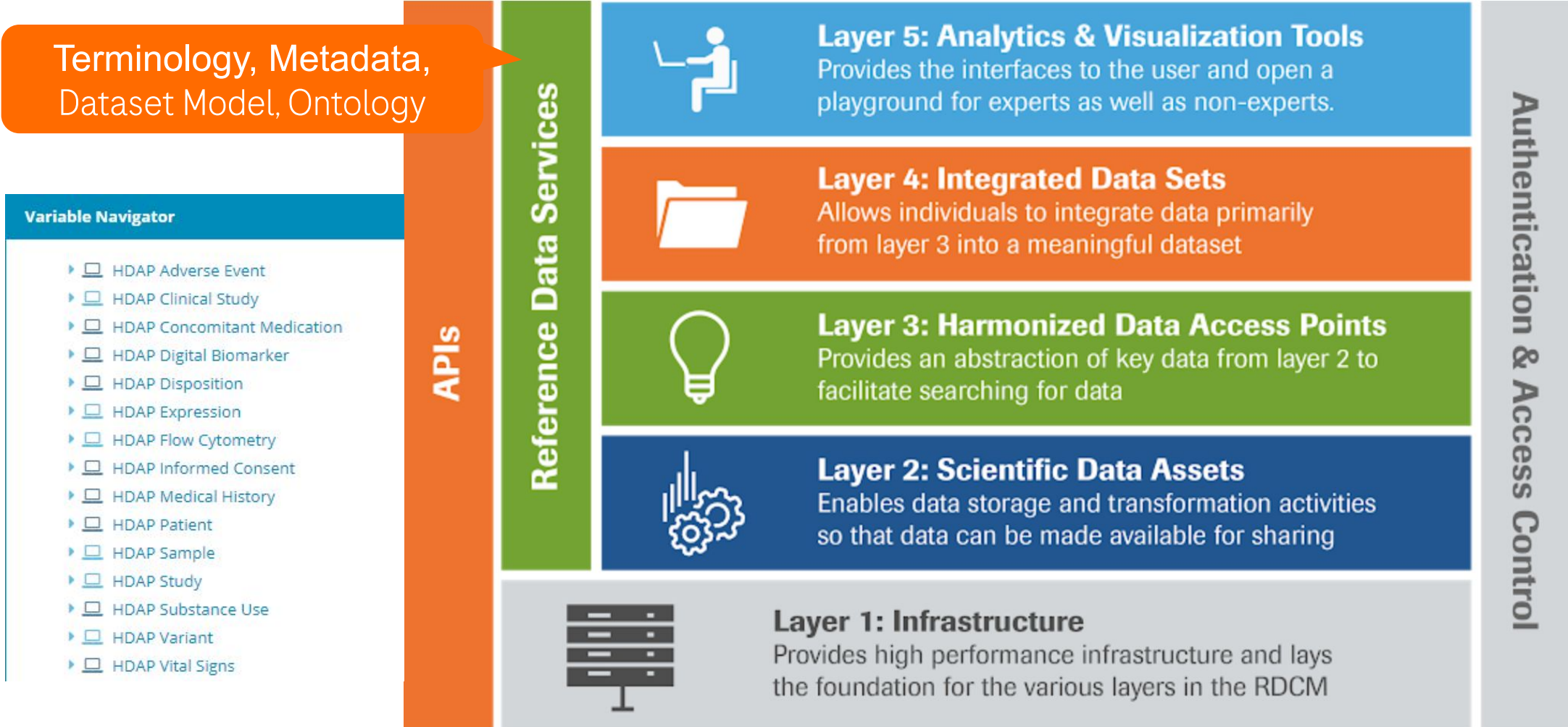
# FAIRification at Scale using Community Standards

Vision

*An open public-private semantic infrastructure of fully standardized FAIR applications, services & data*

# Roche Data Commons

From Application-Centric to Information-Centric





# Roche Data Commons

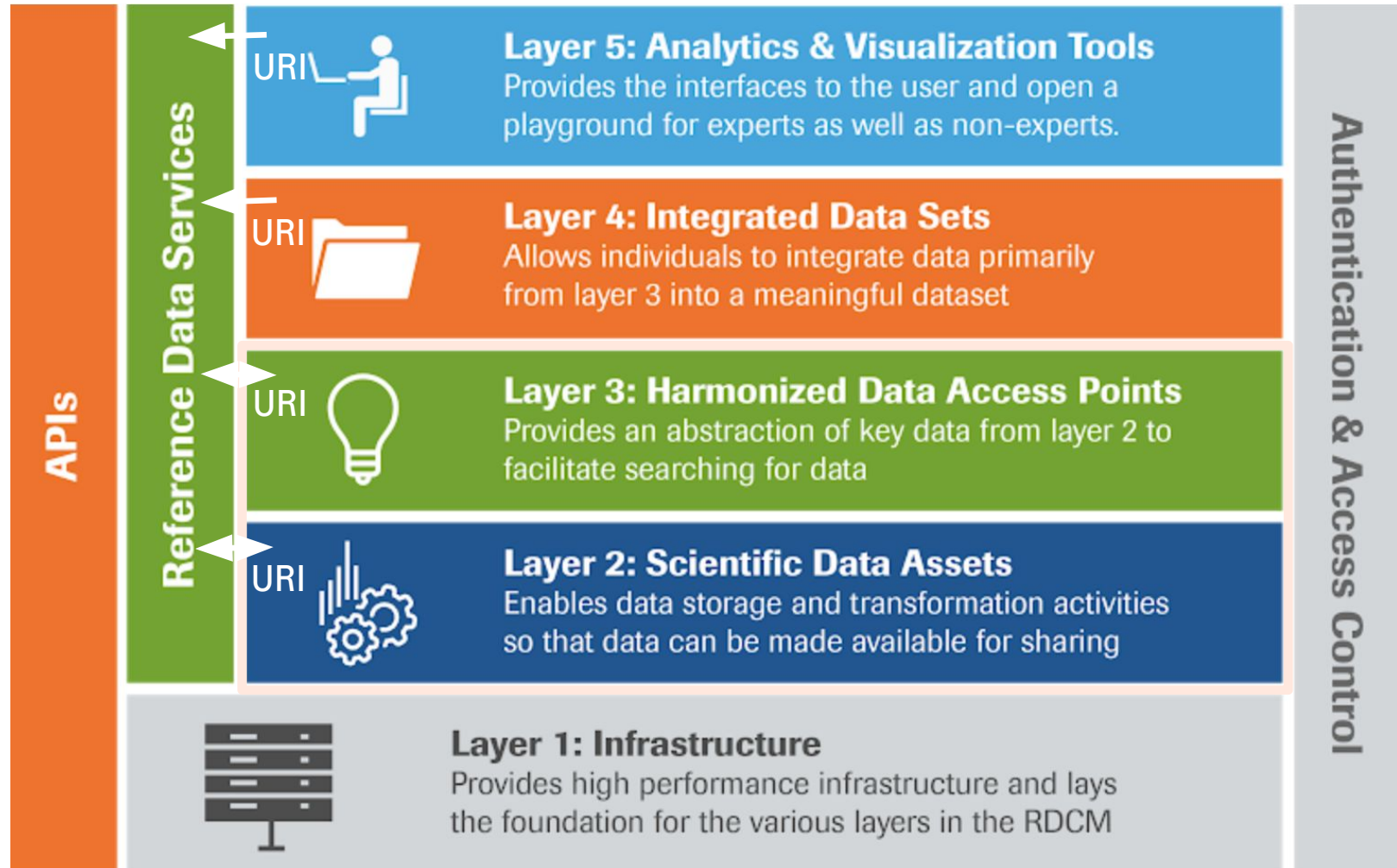
FAIR by Design - Reference by Global Unique Persistent & Resolvable Identifiers (GUPRI)

HDAPs organize data in Information Types

Interoperability (URIs):  
semantic data dictionary  
semantic models

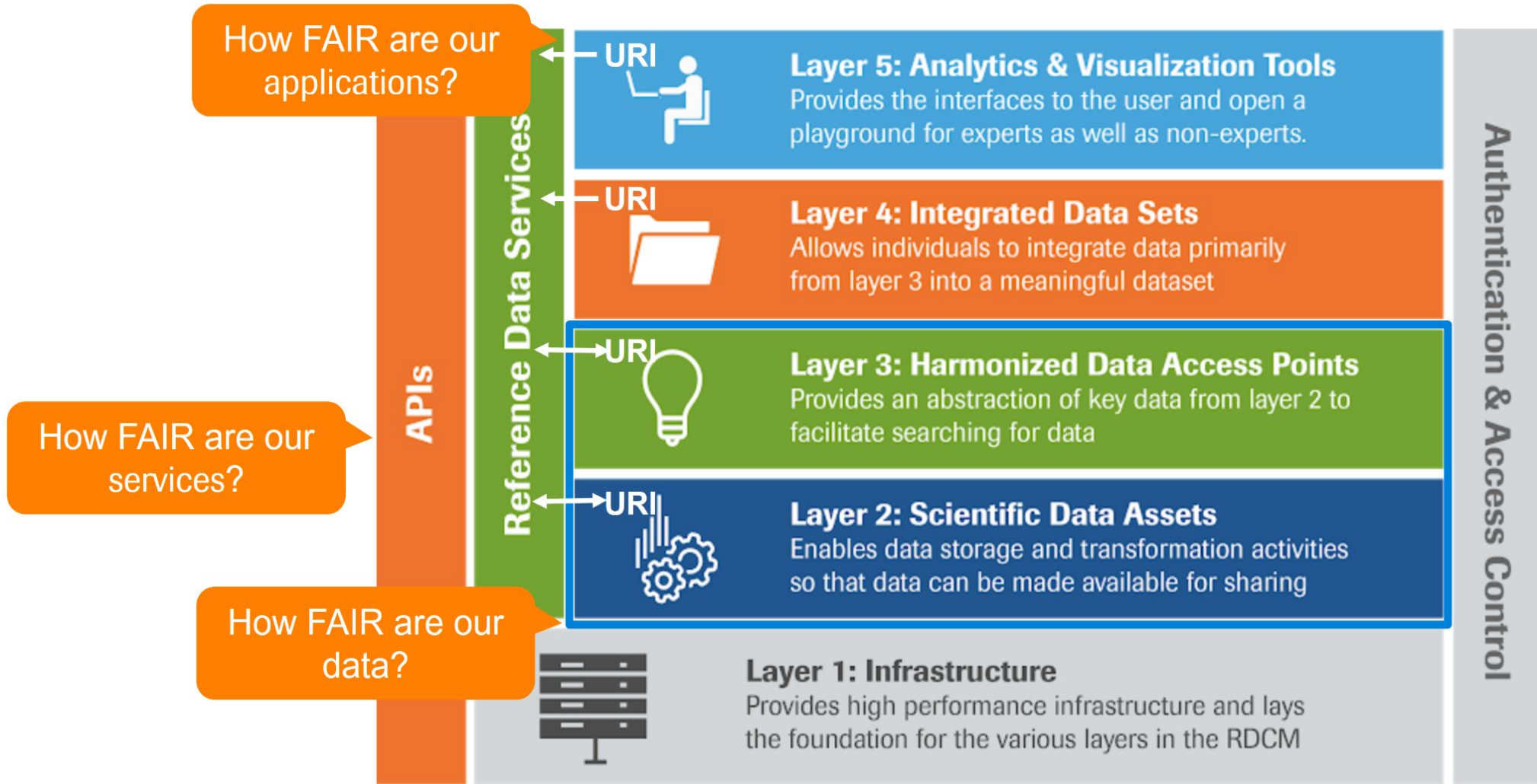
Data FAIRification only in layer 2 & 3

No more transformation between layer 3 & 4,5



# Roche Data Commons

Semantic Infrastructure of FAIR Applications, Services & Data



# FAIR Assessment

FAIR Maturity



The screenshot shows the homepage of the FAIR Toolkit website. At the top left, there is a logo consisting of a blue triangle with the word 'FAIR' and a yellow rectangle below it with the word 'Toolkit'. To the right of this is the 'Pistoia Alliance' logo, which includes a stylized blue icon of a building or structure. Below the logo is a navigation menu with the following items: 'Home', 'FAIR & Life Science Industry' (with a dropdown arrow), 'Use Cases', 'Methods', and 'About' (with a dropdown arrow). On the far right of the navigation bar is a green button labeled 'Contact Us'. The main content area has a dark blue background with a bokeh effect of light spots. A red speech bubble with a white border points to the 'FAIR Maturity' text. The main heading reads 'The FAIR Toolkit for Life Science Industry'. Below this, a paragraph states: 'Use cases and methods have been collated by data science professionals from leading companies in the pharmaceutical, agrifood and biotechnology sectors'.

[FAIR Toolkit](#)

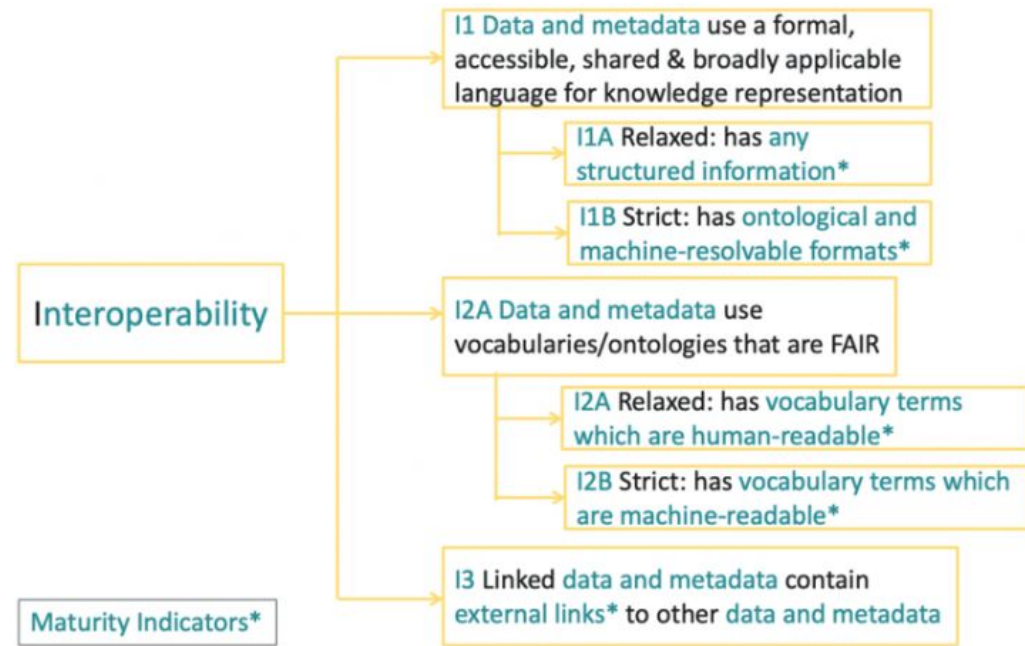
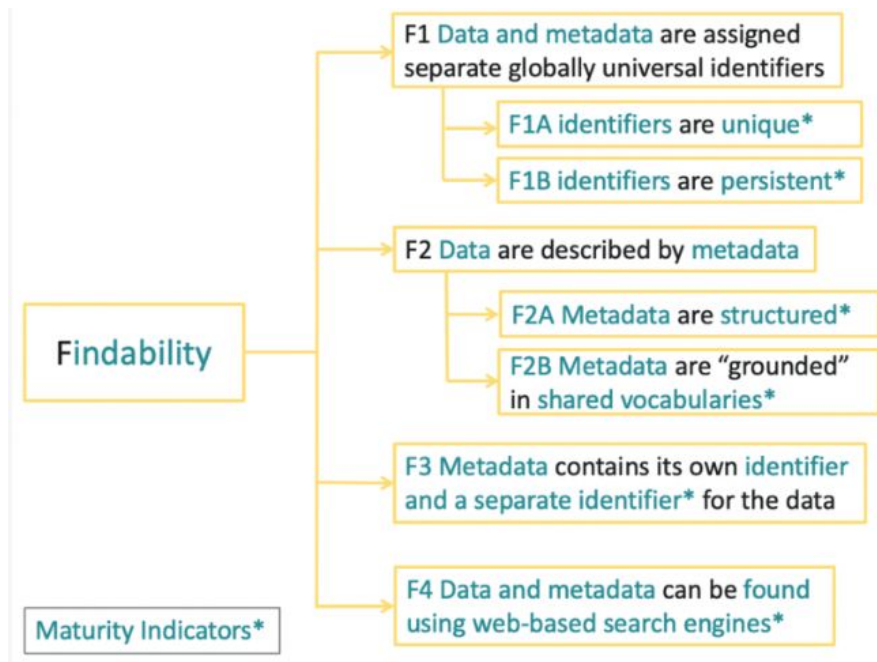
# FAIR Assessment

FAIR Maturity



<https://fairtoolkit.pistoiaalliance.org/>

Home FAIR & Life Science Industry Use Cases **Methods** About



FAIR is about data \*and\* metadata



# Standards in Pharma Industry

# FAIR scientific data management

FAIR guiding principles

F

A

I

R



Ability for scientist/data consumer to find, access and understand the data  
*(without the presence of the data owner)*

Ability for a machine to automatically find and semantically use the data  
*(machine actionable)*

by Olivier Roche (pREDi)

# FAIR data architecture

## Terminology & Concepts

Subject ID	Sex	Age Value	Age Unit	Substance Name	Dose Value	Dose Unit	Dosage Form	Route	Dosage Frequency
S32345	<a href="#">male</a>	230	day	bevacizumab	50	mg/ml	tablet	oral	daily
S93784	female		year	oseltamivir	100	mg/l	solution	intravenous	weekly
S11324	male	88	year	trastuzumab	35	g/l	solution	subcutaneous	hourly
S92833	unknown	1	year	MabThera	50	umol/L	solution	intraocular	hourly
S33021	female	11	month	Polivy	12.5	mg/ml	inhalant	intrabronchial	weekly

Concept

ROX1380015111414

# FAIR data architecture

## Terminology & Concepts

Subject ID	Sex	Age Value	Age Unit	Substance Name	Dose Value	Dose Unit	Dosage Form	Route	Dosage Frequency
S32345	<a href="#">male</a>	230	day	bevacizumab	50	mg/ml	tablet	oral	daily
S93784	female	11	year	oseltamivir	100	mg/l	solution	intravenous	weekly
S11324	male	88	year	trastuzumab	35	g/l	solution	subcutaneous	hourly
S92833	unknown	1	year	MabThera	50	umol/L	solution	intraocular	hourly
S33021	female	11	month	Polivy	12.5	mg/ml	inhalant	intrabronchial	weekly

[Sex Terminology](#)

ROX37210752443769243



# FAIR data architecture

## Variable & Schema

Subject ID	<u>Sex</u>	Age Value	Age Unit	Substance Name	Dose Value	Dose Unit	Dosage Form	Route	Dosage Frequency
S32345	<a href="#">male</a>	50	year	ROX37507104443803061	50	mg/ml	tablet	oral	daily
S93784	female	11	year	oseltamivir	100	mg/l	solution	intravenous	weekly
S11324	male	88	year	trastuzumab	35	g/l	solution	subcutaneous	hourly
S92833	unknown	1	year	MabThera	50	umol/L	solution	intraocular	hourly
S33021	female	11	month	Polivy	12.5	mg/ml	inhalant	intrabronchial	weekly

Variable Concept

ROX37507104443803061

[Sex Terminology](#)

# FAIR data architecture

## Variable & Schema

Subject ID	<u>Sex</u>	Age Value	Age Unit	Substance Name	Dose Value	Dose Unit	Dosage Form	Route	Dosage Frequency
S32345	<a href="#">male</a>	230	day	bevacizumab	50	mg/ml	tablet	oral	daily
S93784	female	11	year	oseltamivir	100	mg/l	solution	intravenous	weekly
S11324	male	88	year	trastuzumab	35	g/l	solution	subcutaneous	hourly
S92833	unknown	1	year	MabThera	50	umol/L	solution	intraocular	hourly
S33021	female	11	month	Polivy	12.5	mg/ml	inhalant	intrabronchial	weekly

[Subject Schema](#)

ROX37643616443820633

# FAIR data architecture

## Simple Graph Generation

Subject ID	<u>Sex</u>	Age Value	Age Unit	Substance Name	Dose Value	Dose Unit	Dosage Form	Route	Dosage Frequency
S32345	<a href="#">male</a>	230	day	bevacizumab	50	mg/ml	tablet	oral	daily
<b>Object</b>	female	11	year	oseltamivir	100	mg/l	solution	intravenous	weekly
S11324	male	88	year	trastuzumab	35	g/l	solution	subcutaneous	hourly
S92833	unknown	1	year	MabThera	50	umol/L	solution	intraocular	hourly
S33021	female	11	month	Polivy	12.5	mg/ml	inhalant	intrabronchial	weekly

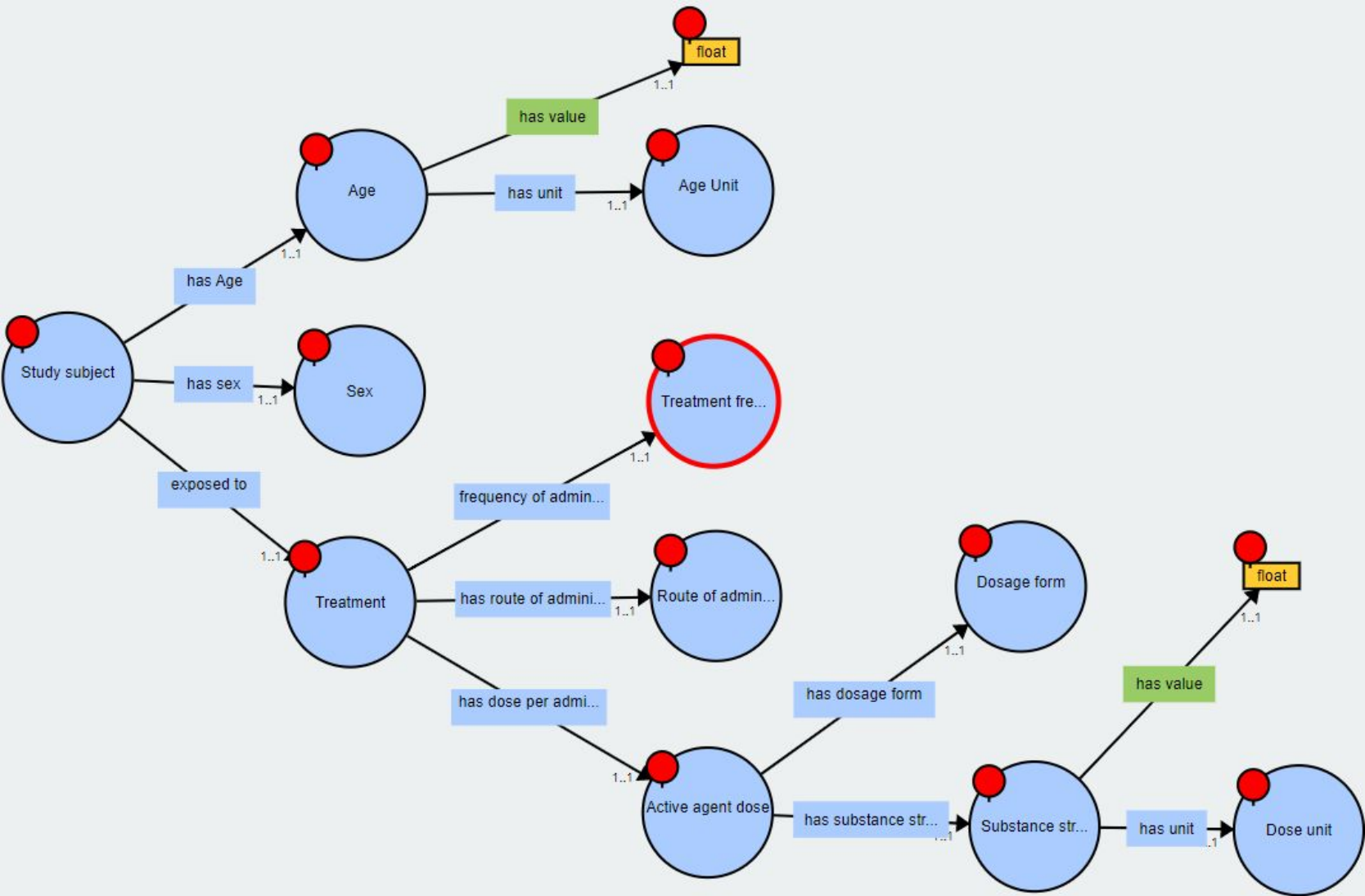
S32345 - hasSex - male

S32345 - ageValue - 230

S32345 - ageUnit - day

# FAIR data architecture

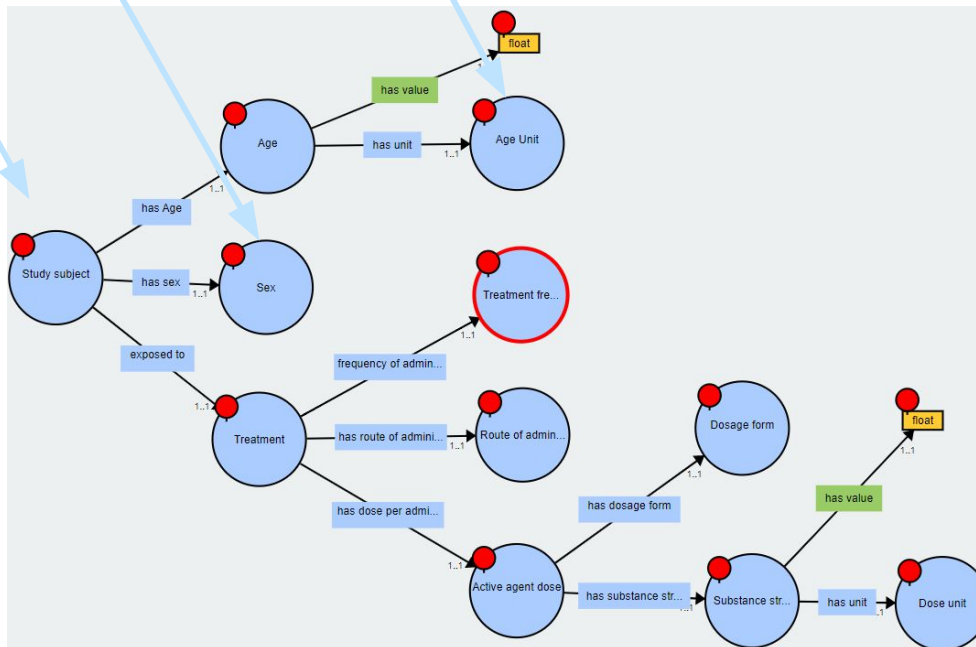
## Semantic Graph Generation



# FAIR data architecture

## Semantic Graph Generation

Subject ID	<u>Sex</u>	Age Value	Age Unit	Substance Name	Dose Value	Dose Unit	Dosage Form	Route	Dosage Frequency
S32345	<a href="#">male</a>	230	day	bevacizumab	50	mg/ml	tablet	oral	daily



**FAIR is ugly and complex**

# FAIRsharing Catalog of Biomedical Resources

Proliferation and Fragmentation of Standards

The screenshot displays the FAIRsharing.org website. At the top left is the logo "FAIRsharing.org standards, databases, policies". A search bar contains the text "Search all of FAIRsharing". Navigation buttons include "Standards", "Databases", "Policies", "Collections", "Add/Claim Content", "Stats", and "Log in or Register".

The "Standards" section features a dark blue header with the text "The standards in FAIRsharing are manually curated from a variety of sources, including BioPortal, MIBBI and the Equator Network." Below this text are three icons: a document with a checkmark, a tree diagram, and a grid with an arrow. An orange callout box with white text says "Manually done- no smart interfaces".

A search bar for "Search Standards" is located below the header, with buttons for "Search", "Reset", and "Advanced".

The main content area shows "Showing records 1 - 50 of 1299." and a pagination bar with page numbers 1 through 26. An orange callout box with white text says "1299 entries for Standards".

Below the pagination is a table of search results. The table has columns: Registry, Name, Abbreviation, Type, Subject, Related Database, Related Standard, Related Policy, In Collection/Recommendation, and Status. Two rows are visible:

Registry	Name	Abbreviation	Type	Subject	Related Database	Related Standard	Related Policy	In Collection/Recommendation	Status
	ABA Adult Mouse Brain	ABA	Standard	None	None	None	None	None	
	Access to Biological Collection Data	ABCD	Standard	Biodiversity  Biology  Life Sciences	None	AI	GBIF Atlas of Living Australia IPT - GBIF Australia	ABCD EFG ABCDDNA TDWG Biodiversity Information Standards	

On the left side of the results, there are options for "View as Table" and "View as Grid", a "Sort by" dropdown menu set to "Name", and a "Recommended Records" section with a red "Recommended" button. Below that is an "Associated Publication?" section.

# Interoperability - Standards

## Clinical Data vs FHIR

```

"name": [
  {
    "extension": [
      {
        "url": "http://hl7.org/fhir/StructureDefinition/humanname-assembly-order",
        "valueCode": "NL1"
      }
    ],
    "use": "official",
    "family": "001",
    "_family": {
      "extension": [
        {
          "url": "http://hl7.org/fhir/StructureDefinition/humanname-own-name",
          "valueString": "001"
        }
      ]
    }
  }
],
"gender": "male",
"birthDate": "1975-06-12",
"deceasedBoolean": false

```

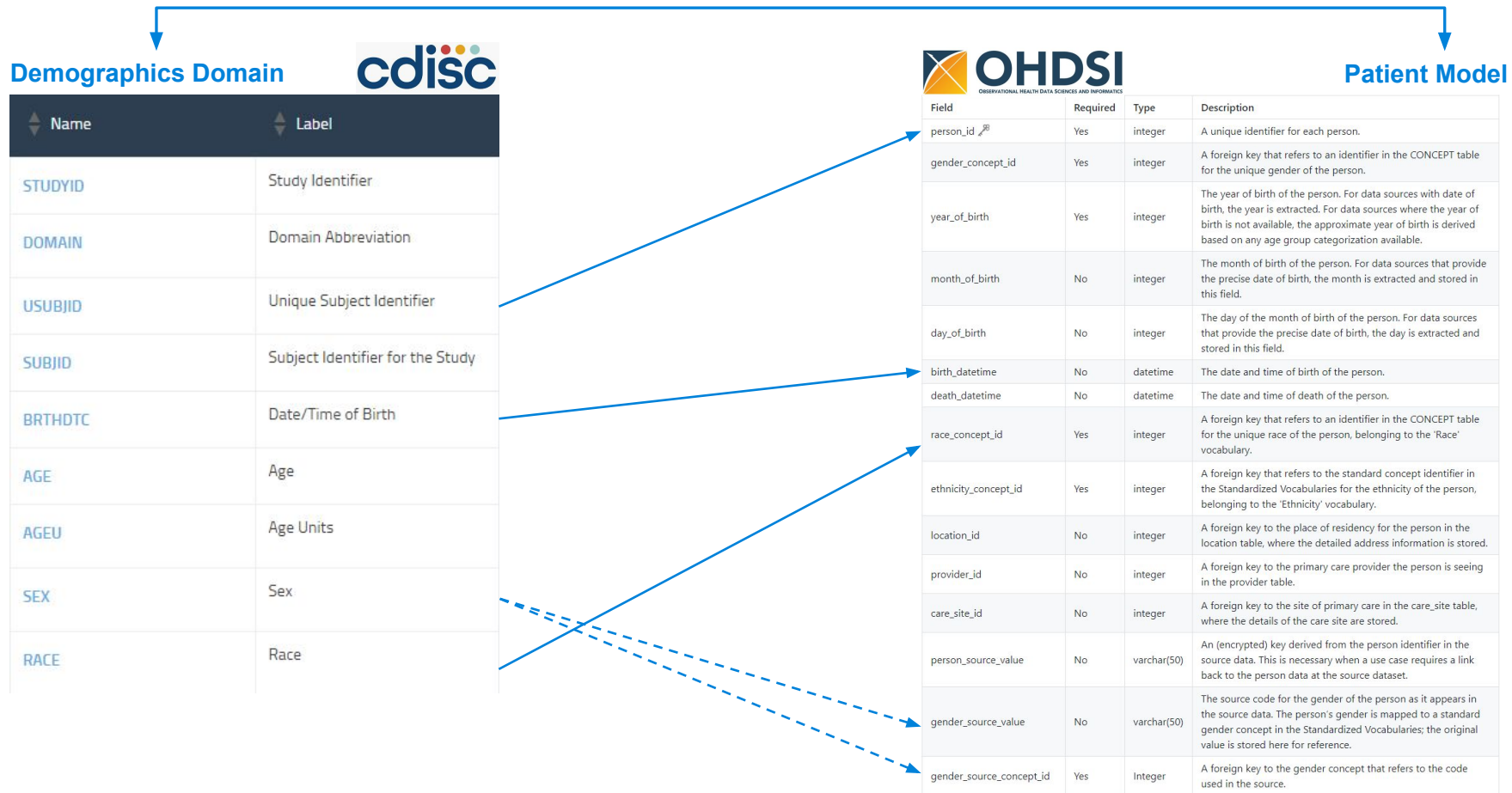
SurvivalStatus
Alive
Alive
Death
Alive
Alive
Alive
Alive
Alive
Alive
Death
Alive
Alive
Death
Death
Alive
Alive
Death
Death
Alive
Alive
Death

IF deceasedBoolean = "false" THEN SurvivalStatus = "alive"



# Data Standards & Interoperability Challenges

CDISC vs OMOP/ OHDSI



Creation of insights & analytics blocked: different model, variables and values

# Data Standards & Interoperability Challenges

CDISC vs OMOP/ OHDSI vs DICOM

**DICOM Standard Browser** by Innolitics

Demograp

Domain Name	CR Image	CIOD	
	▼ Patient	M	Module - Patient
STUDYID	▶ (0008,1120) Referenced Patient Sequence	3	Sequence
DOMAIN	(0010,0010) Patient's Name	2	Person Name
	(0010,0020) Patient ID	2	Long String
USUBJID	(0010,0021) Issuer of Patient ID	3	Long String
SUBJID	(0010,0022) Type of Patient ID	3	Code String
	▶ (0010,0024) Issuer of Patient ID Qualifiers Sequence	3	Sequence
BRTHDTC	▶ (0010,0026) Source Patient Group Identification Sequence	3	Sequence
AGE	▶ (0010,0027) Group of Patients Identification Sequence	3	Sequence
AGEU	(0010,0030) Patient's Birth Date	2	Date
	(0010,0032) Patient's Birth Time	3	Time
SEX	(0010,0033) Patient's Birth Date in Alternative Calendar	3	Long String
	(0010,0034) Patient's Death Date in Alternative Calendar	3	Long String
RACE	(0010,0035) Patient's Alternative Calendar	1C	Code String
	(0010,0040) Patient's Sex	2	Code String
	(0010,0200) Quality Control Subject	3	Code String
	(0010,0212) Strain Description	3	Unlimited Characters

**Creation of insights & analytics blocked: different model, variables and values**

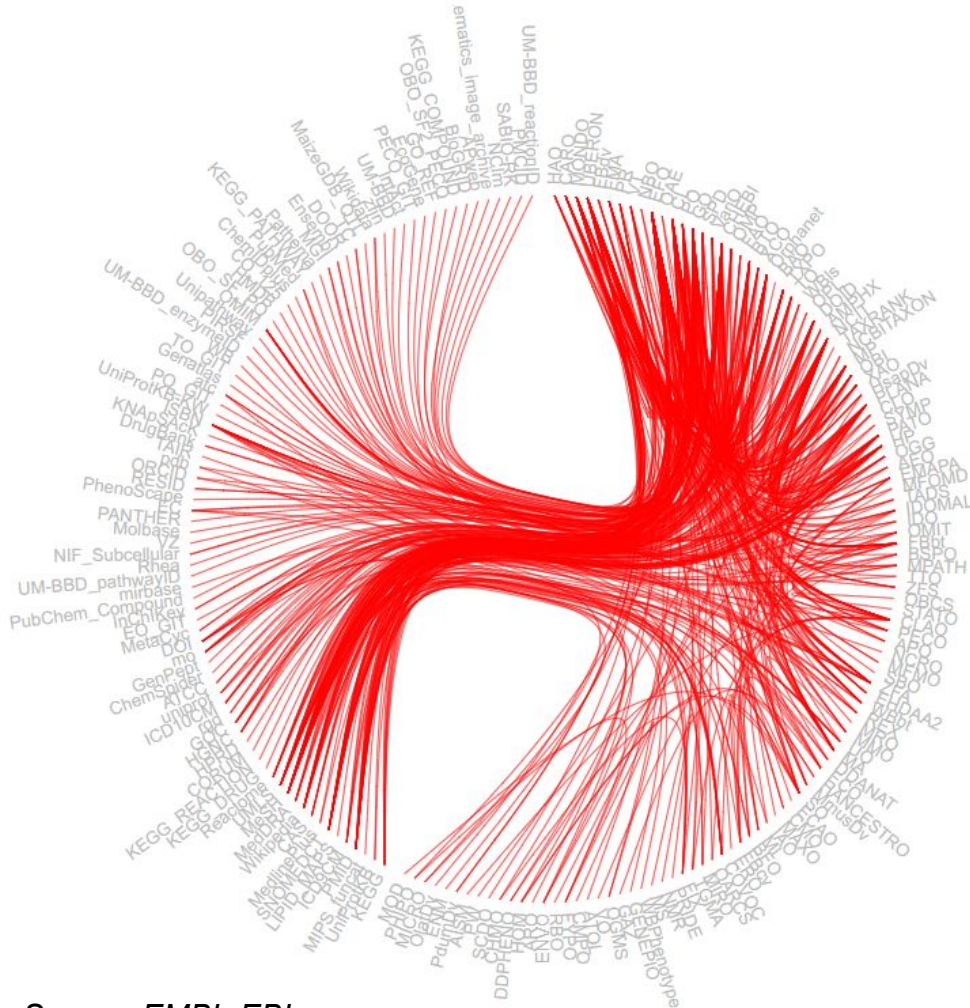
# EMBL-EBI Ontology Xref Service

Creating referential identity by ontology mapping

Welcome to the EMBL-EBI Ontology Xref Service (OxO).

OxO is a service for finding mappings (or cross-references) between terms from ontologies, vocabularies and coding standards. OxO imports mappings from a variety of sources including the [Ontology Lookup Service](#) and a subset of mappings provided by the [UMLS](#). We're still developing the service so please [get in touch](#) if you have any feedback.

1. Allocating significant resources to inflate a problem
2. Allocating significant resources to reduce a problem (loss of information & interoperability)



# Interoperability for Ontology Mappings

RDF standard for a FAIR representation of OM

## A Simple Standard for Sharing Ontology Mappings (SSSOM)

About SSSOM, A Simple Standard for Sharing Ontological Mappings

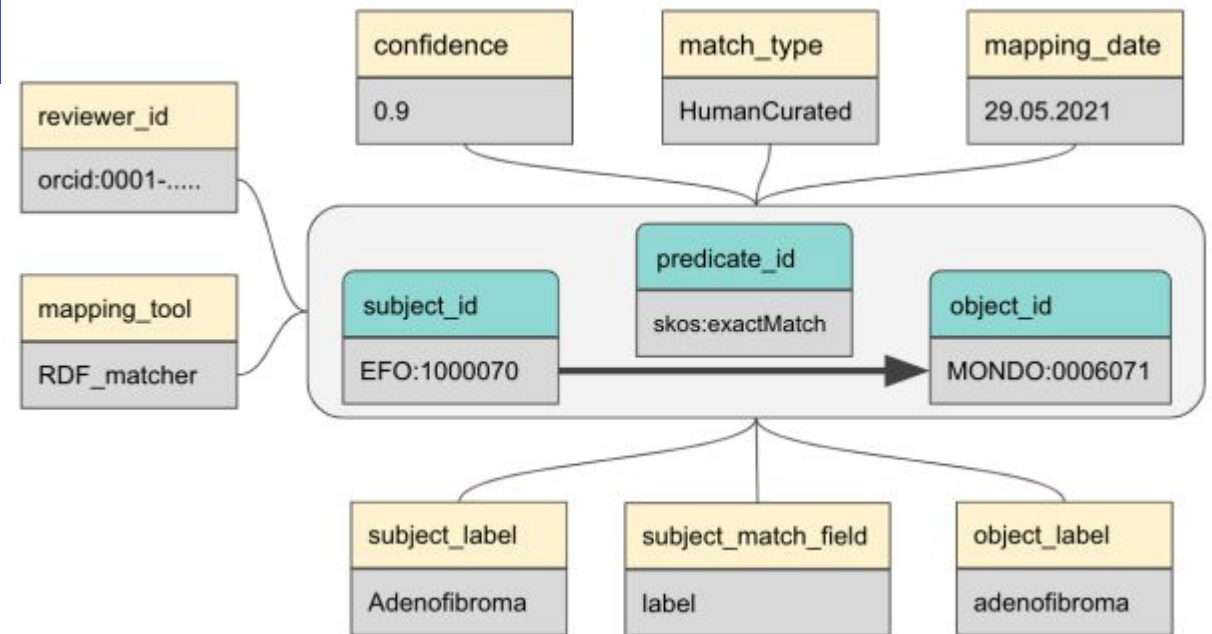
SSSOM is a simple metadata standard for describing semantic mappings:

1. Introducing a machine-readable and extensible vocabulary to describe metadata of mappings.
2. Defining an easy to use table-based format that can be integrated into existing data science pipelines without the need to parse or query ontologies, and that integrates seamlessly with Linked Data standards.
3. Implementing open and community-driven collaborative workflows designed to evolve the standard continuously to address changing requirements and mapping practices.
4. Providing reference tools and software libraries for working with the standard.

A SSSOM mapping comprises three major components:

1. The **mapping** itself, that is, a triple `<subject, predicate, object>` that reflects a correspondence of a `subject` entity, for example a class in an ontology, to an `object` entity, for example an identifier in some database, via a semantic mapping `predicate`, such as `skos:exactMatch`.
2. A **mapping justification**, which the process or activity that led us to consider the mapping to be correct or reasonable (typical examples: labels match exactly; two classes are logically equivalent; a domain expert determined that two terms reflect the same real world concept).
3. **Provenance metadata**, including information about `author` and `mapping_tool`.

[Reference: SSSOM](#)

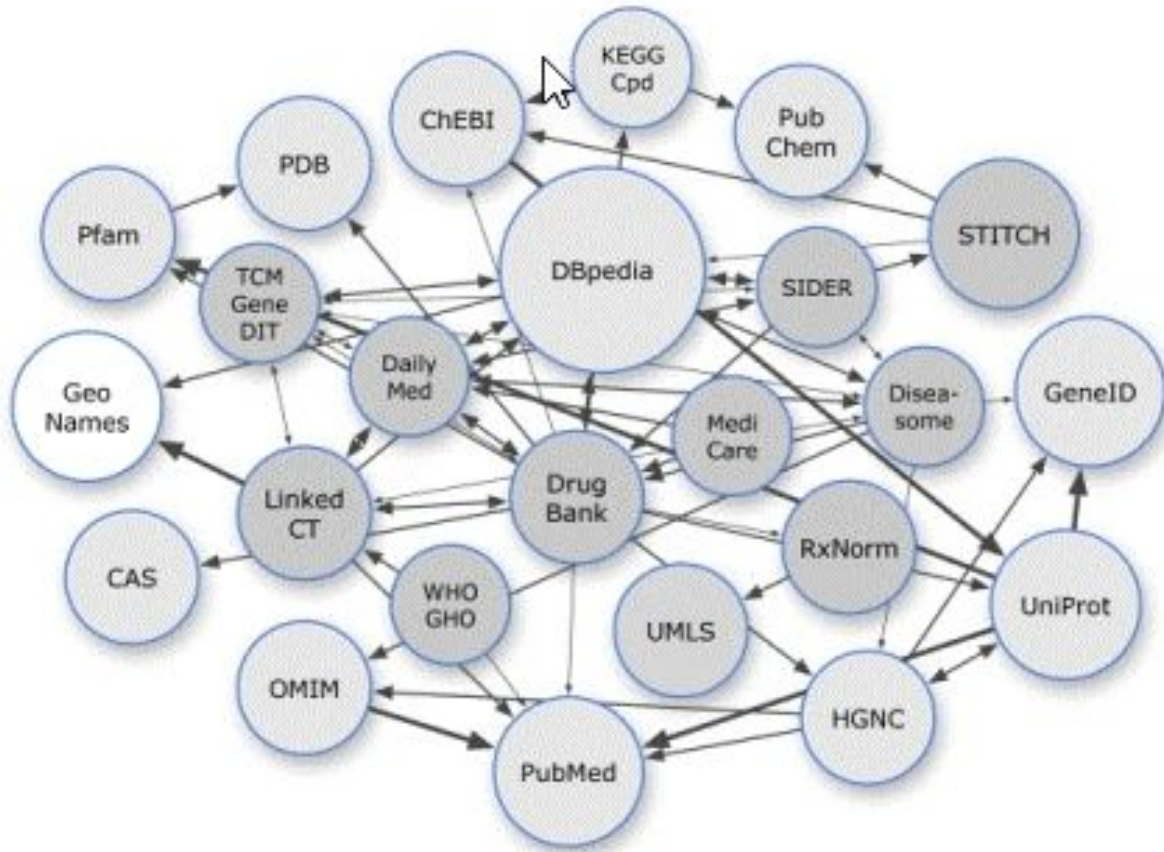


Not fully FAIR (dct:creator & dct:created)  
No guidelines on property labels

[Linked Open Vocabularies](#)

# Linked Open Data Cloud

The Linked Data Illusion



Data in the Linked Data Cloud is not linked  
 Linkage comes with referential identity  
 Referential identity comes with interoperability  
 LODD is not FAIR



# Pharma Interoperability Hub & Knowledge Graphs

# Pharma Interoperability Hub & Data Harmonization Service

FAIR by Design to support FAIRification at Scale



Product Line “Scientific Interoperability Hub” offering three products: terminology management, semantic dataset definition & conceptual modeling (purpose-driven ontologies)



Products are FAIR by design supporting FAIRification at scale for the entire Roche organization. Data Harmonization Service ensures semantic interoperability & high data quality.



Products serve as reference data for standardized terminologies, metadata & conceptual models semantically linking internal and external data assets for data acquisition and data integration.



Supporting more than 100 productive applications across all Roche functions and sites. The Data Harmonization Services guarantees currency and ongoing support.

# Reference Data Services for Data Management

## Terminology Management - Contextualize Concepts (FAIR)

Roche Terminology System v2.52.0 [PRD]

FAIR Metrics

Terminology Application Variable Curation Administration Information

-> More than 110 productive applications

martin.romacker@roche.com Logout

The screenshot displays the Roche Terminology System interface. On the left is the **Terminology Navigator** showing a tree of categories, with **Non Small Cell Lung Cancer** selected. A red callout bubble labeled **Master Terminology** points to this section. In the center is the **Application Navigator** showing a list of applications, with **EpiCX** selected. A red callout bubble labeled **Application** points to this section. On the right is the **Concept Entity Properties** panel for **Non Small Cell Lung Cancer**. It includes fields for **Label** (Non Small Cell Lung Cancer), **Terminology** (Indication), **Status** (Active), and **Identifier** (ROX1305277804386). A red callout bubble labeled **Concept** points to this panel. Below the properties is a **Definition** field containing the text: "A group of at least three distinct histological types of lung cancer, including squamous cell carcinoma, adenocarcinoma, and large cell...". At the bottom is a table of **Application Terminology** with columns for Label, Language, Source, Label Type, and Lexical Type. A red callout bubble labeled **Application Terminology** points to this table.

Label	Language	Source	Label Type	Lexical Type
Non Small Cell Lung Cancer	en	Roche	Synonym	prefLabel
Cancer, lung, non small cell	en	PIP	Synonym	altLabel
Cancer, non small cell lung	en	Roche	Synonym	altLabel
Carcinoma, Non Small Cell Lung	en	Roche	Synonym	altLabel
Carcinoma, non small cell lung	en	Roche	Synonym	altLabel
Carcinoma, non small cell lung cancer	en	Roche	Synonym	altLabel
Carcinoma, Non-Small-Cell Lung	en	Roche	Synonym	altLabel
Non small cell lung cancer	en	ADIS, TPP	Synonym	altLabel
Non small cell lung cancer (NSCLC)	en	Roche	AcroDefinition	altLabel



# Reference Data Services for Data Management

Metadata Registry/ Dataset Models – Metadata Harmonization (FAIR)

Roche Terminology System v2.52.0 [PRD]

FAIR Metrics

[martin.romacker@roche.com](mailto:martin.romacker@roche.com) [Logout](#)

The screenshot displays the Roche Terminology System interface with several key components:

- Search:** A search bar containing the text "Country" and a scope dropdown set to "SP Variable".
- Terminology Navigator:** A tree view on the left showing a list of "DM variable"s, with "Country" highlighted.
- Variable Navigator:** A tree view in the center showing a hierarchy of "DM Domain"s, with "Country" selected under "HDAP Subject".
- Variable Entity Property:** A detailed view on the right for the "Country" variable, including:
  - General information:** A table of properties such as Variable name (Country), Value Domain type (Application Terminology), App Terminology (Country Code (Alpha 3)), Variable Multiplicity (single-valued), Variable Policy (Required Variable), Curation Policy, and Variable Context.
  - Definition:** A text box containing "Country of the investigational site in which the subject participated in the trial (GDSR)."
  - Comment:** A text box containing "ISO 3166 format."
  - Concept Reference:** A table with columns for Concept and Link, showing "Country" as a concept.

Four red callout boxes with white text are overlaid on the interface to identify specific areas:

- Application:** Points to the "Application Terminology" value domain type in the General information section.
- Data Dictionary:** Points to the "Country" variable in the Terminology Navigator.
- Variable:** Points to the "Country" variable in the Variable Navigator.
- Variable Properties:** Points to the "Country" variable name in the General information section.

# Reference Data Services for Data Management

Conceptual Model - Purpose-build FAIR Ontologies

### Model Navigator

- ▶ HGDI
- ▶ Home Cage Analysis
- ▶ I2O Knowledge Base
  - ▣ I2O KB Core Model
    - ▶ \* Pathway
    - ▶ \* Genetic variation
      - ▣ Reference SNP Cluster Identifier
      - ▣ minor allele frequency
    - ▶ \* Gene-disease association
    - ▶ \* Gene variant-disease association
      - ▣ associated gene variant
    - ▶ \* Drug
      - ▣ target molecule
      - ▣ treatment indication
    - ▶ \* Tissue
    - ▶ \* Biomolecule
    - ▶ \* Expression group
    - ▶ \* Cell
    - ▶ \* Disease
    - ▶ \* Gene variant-disease association evidence
      - ▣ **associated disease**
      - ▣ associated gene variant
      - ▣ p-value
      - ▣ odds ratio
      - ▣ odds ratio upper 95% confidence interval
      - ▣ odds ratio lower 95% confidence interval

### Property Entity View

**Model Global Properties**

Master Concept Identifier: ROX38009088443943245

Preferred Label Identifier: associated disease

Local Technical Key:

Preferred Reference URI:

Edit

↻

**Local Usage Properties**

Used at class: Gene variant-disease association evidence

Target class: Disease

Data type:

Multiplicity: 1..1

**Definition**

Disease that is part of an association with one or multiple other concepts.

**Comment**

Model

GUPRI (ROX ID)

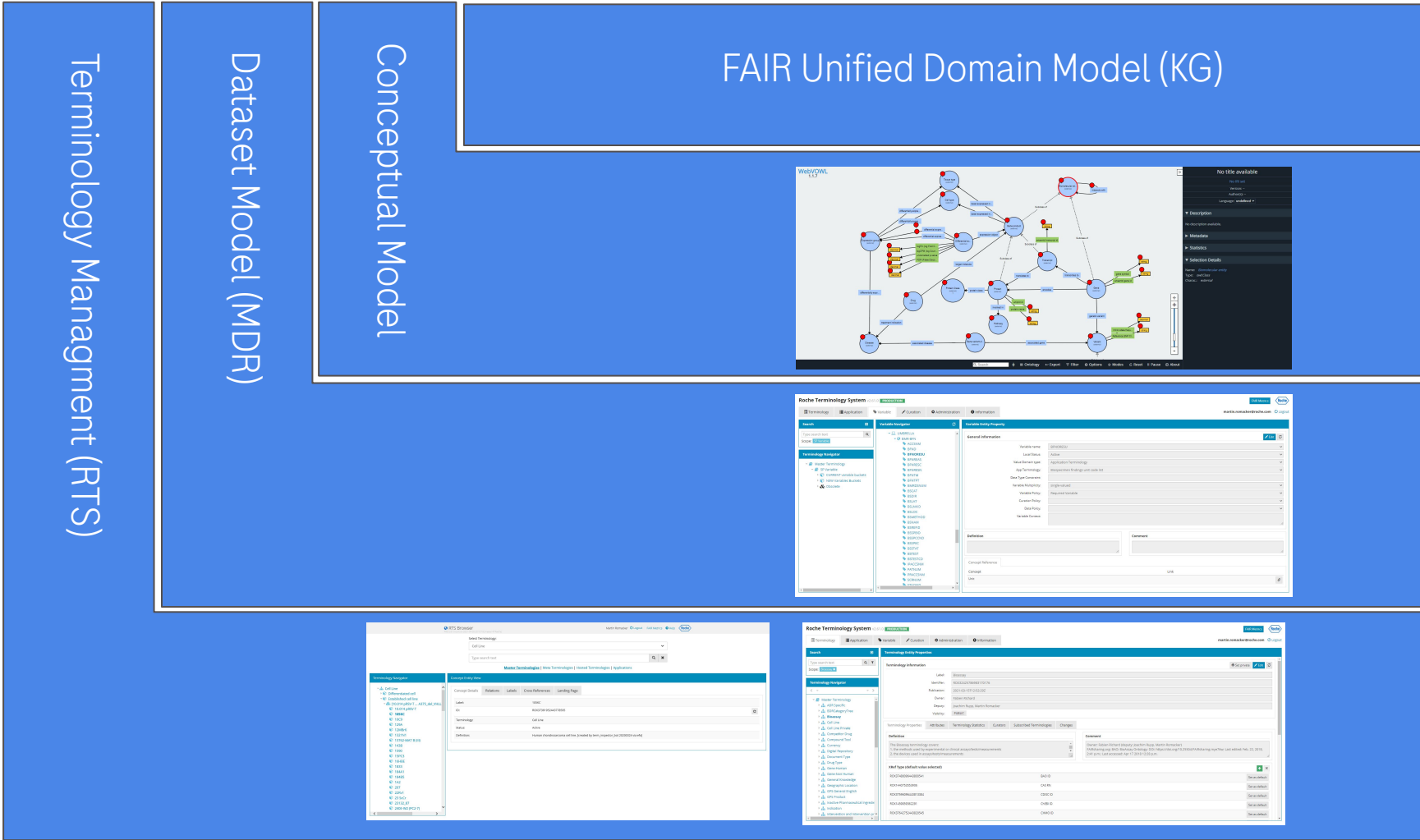
Class

Property

Target Class

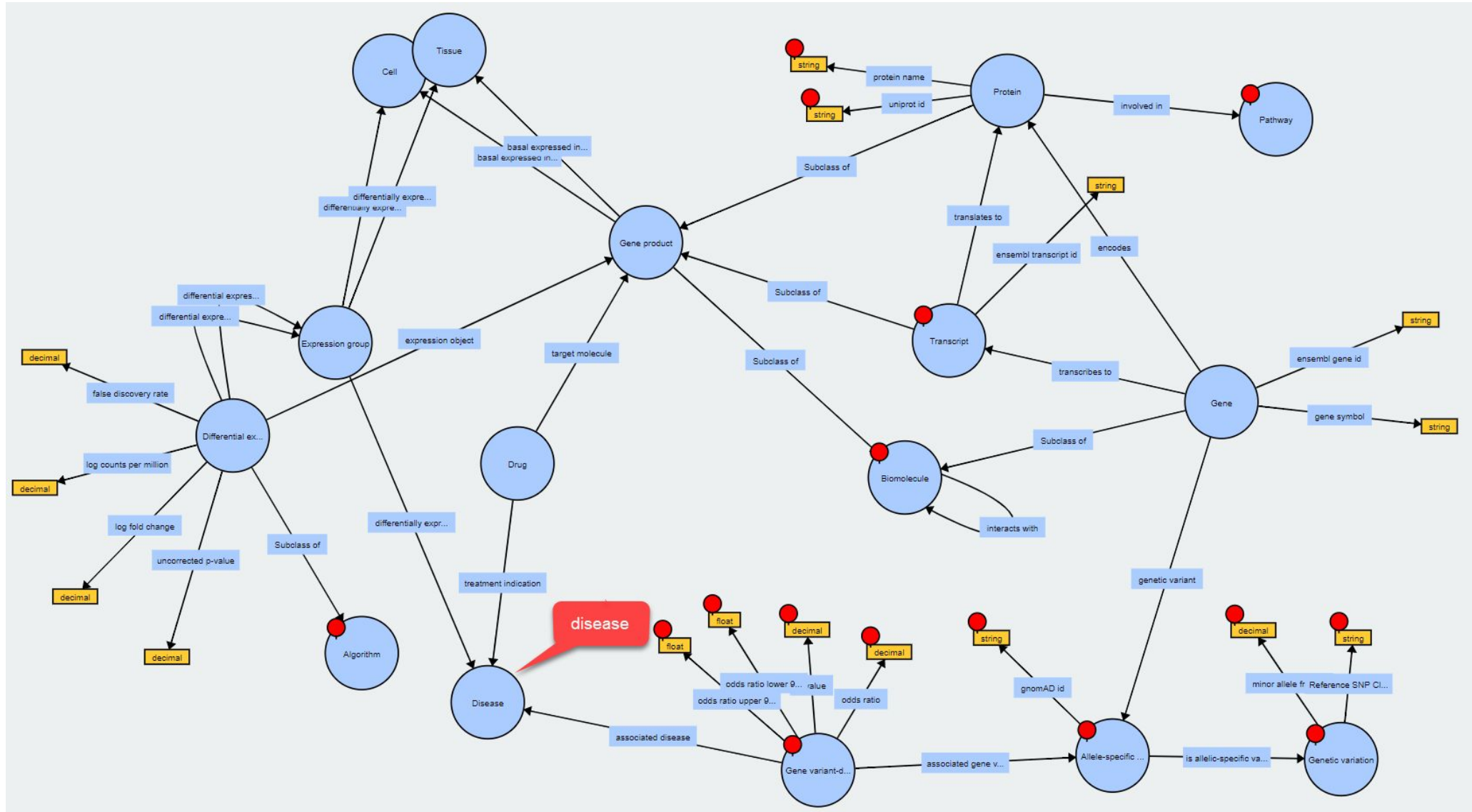
# Semantic Interoperability Hub - Capability Stack

Data Management Value Chain - From Terminologies to a Unified Domain Model



# Infectious Disease Ontology

Instantiation of a Knowledge Graph



# FAIR Data Integration

## Federation of Knowledge Graphs (Zero Integration)



I2O Knowledge Graph

Competitor Information Knowledge Graph

# Conclusions

## Conclusions

- Successful and value-generating Digitilization requires true machine-actionable data, machine-readability is not sufficient. Application of FAIR principles is mandatory.
- FAIR data principles intrinsically tie Data Management to Semantic Technologies. (usage of terminologies, dataset definitions & ontologies )
- Transformationless data integration based on fully harmonized and standardized machine-actionable data assets (FAIR by design) results in fully linked data ecosystem to produce more reliable insights in less time at lower costs.
- Data Management Value Chain: new architectural approaches around data and information. Semantic Interoperability of terminologies, dataset definitions and ontologies is key to make our data assets machine-actionable.
- It's all about Semantics.

# Acknowledgements





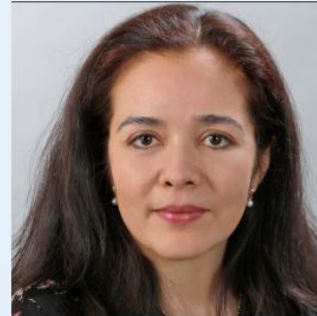
**Joachim Rupp**

RTS Functional Manager, Basel



**Fabien Richard**

Terminology Specialist, Basel



**Silvia Jimenez**

Terminology Specialist, Basel

Dataset portal team:

- Hugo de Schepper
- Oliver Steiner
- Roy Weiler



**Felix Schwagereit**

Scientific Technical Manager,  
Basel



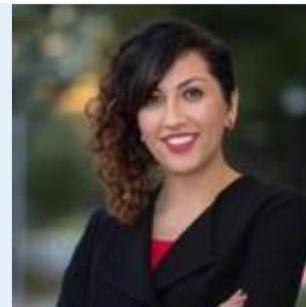
**Pratishtha Duhan**

Business Manager, SSF



**Rama Balakrishnan**

Biomedical Ontology Specialist,  
SSF



**Shima Dastgheib**

Semantic Integrator, SSF

# Roche Terminology System

Dev and Ops Team, Curation Team

## RTS Dev and OPS Team



**Michal Bielak**



**Michal Paradowski**



**Adam Zawada**



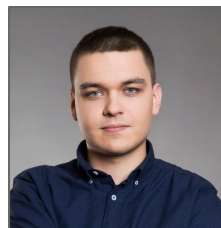
**Konrad Borowka**



**Robert Trypuz**



**Agnieszka  
Bananszynska-  
Krolikiewicz**



**Majewski  
Krzysztof**

### Additional Members:

- Michal Kolacki
- Michal Openchowski
- Adam Sedra
- Tomasz Gil
- Piotr Bablok
- Pawel Nowicki

## Molecular Connections Team



**Arathi Raghunath**  
Technical & Project  
Lead for Roche



**Krishna K Chinnaiah**  
Business & Account  
Manager for Roche

### Curation supported by:

- Ananda Kembathahally Mahadevaiah
- Bharat Bhat
- Farheen Shaikh
- Nethravathy Nagaraju
- Priyadarsini Panda
- Shruthi Shankar
- Vanitha Sharath

## Rancho Biosciences Team



**Erfan Younesi**  
Sr. Data Curator



**Svetlana Koltsova**  
Sr. Data Curator



**Maxim Papin**  
Sr. Data Curator

**Doing now what patients need next**