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

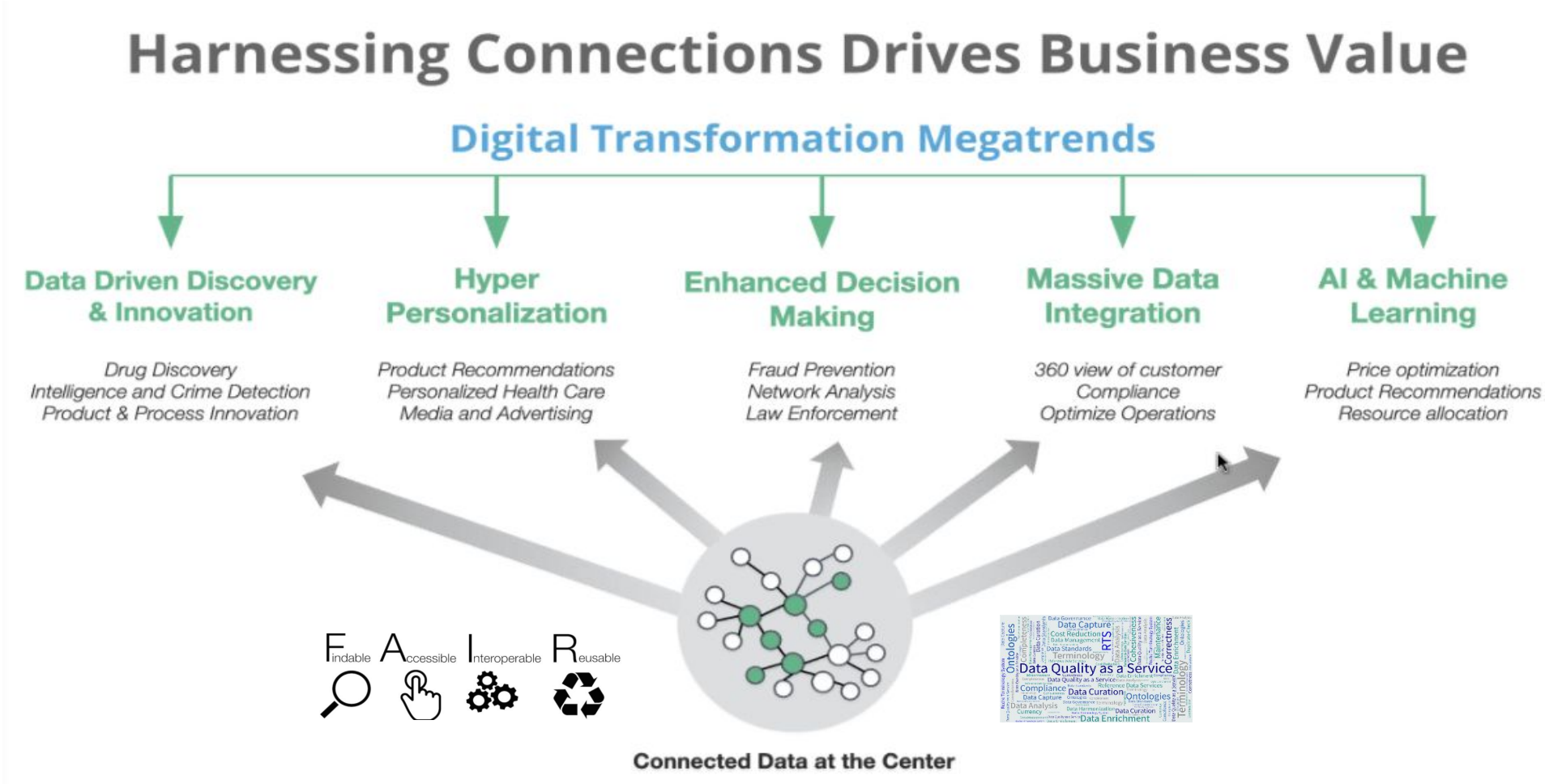
06 April 2023, OntoCommons Workshop, Berlin

Table of contents

Data Economics

Megatrends in Digitilization

FAIR plus Q Data



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

The Semantic Web is Dead - Long Live the Semantic Web!

Source: Rik van Bruggen, Neo4J

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

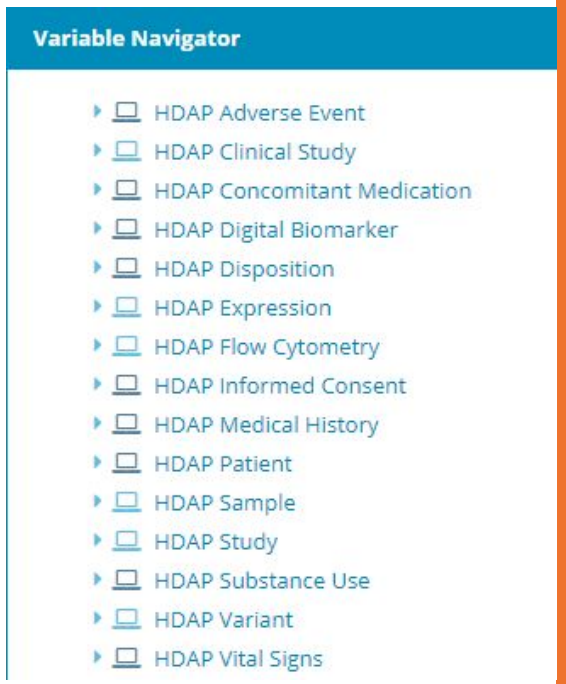
Vision

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

Roche Data Commons

From Application-Centric to Information-Centric

Terminology, Metadata,
Dataset Model, Ontology



APIs

Reference Data Services

Layer 5: Analytics & Visualization Tools
Provides the interfaces to the user and open a playground for experts as well as non-experts.

Layer 4: Integrated Data Sets
Allows individuals to integrate data primarily from layer 3 into a meaningful dataset

Layer 3: Harmonized Data Access Points
Provides an abstraction of key data from layer 2 to facilitate searching for data

Layer 2: Scientific Data Assets
Enables data storage and transformation activities so that data can be made available for sharing

Layer 1: Infrastructure
Provides high performance infrastructure and lays the foundation for the various layers in the RDCM

Authentication & Access Control

Roche Data Commons

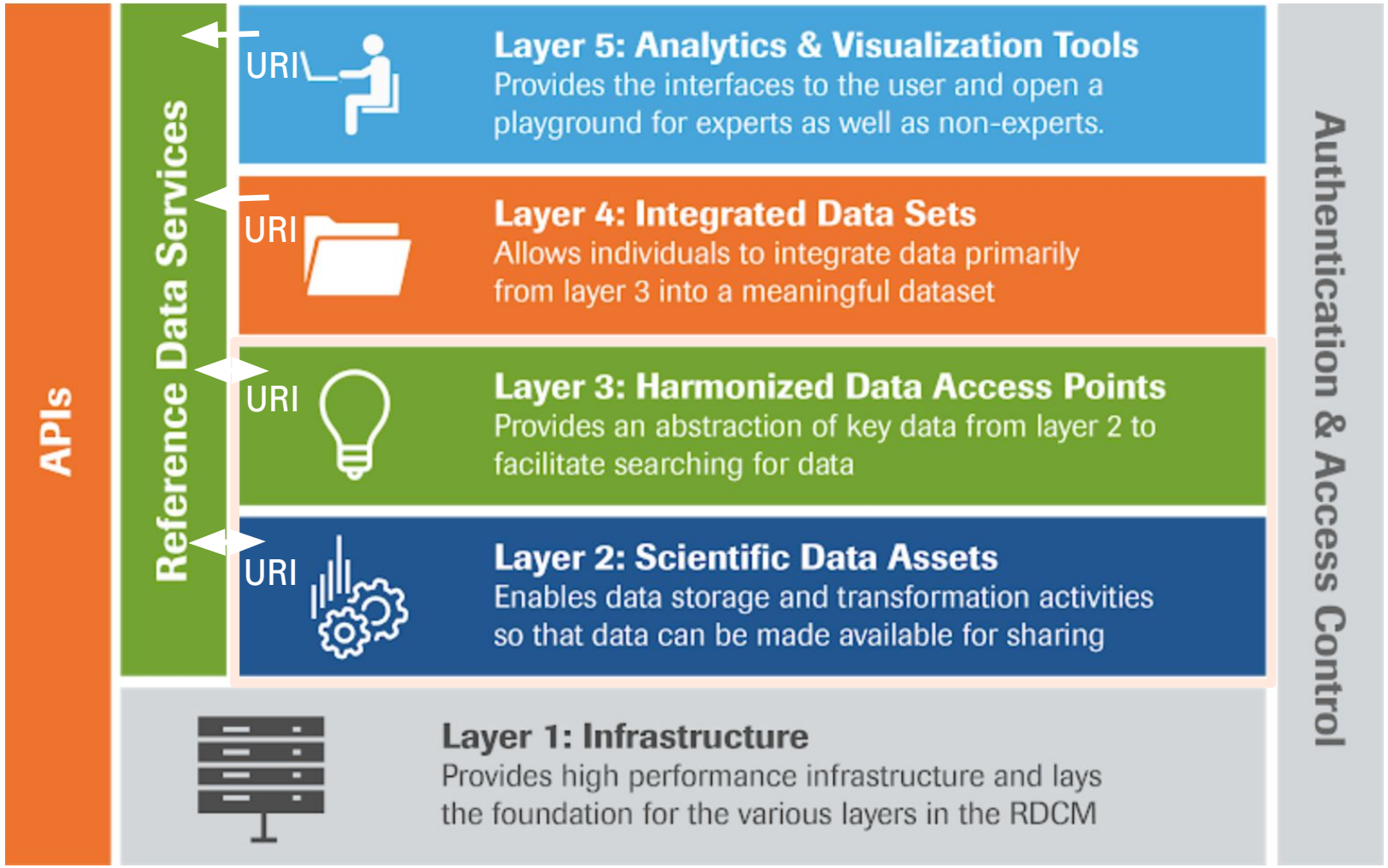
FAIR by Design - Reference by Global Unique Persistent & Resolvable Identifiers (GUPRI)

HDAPs organize data in Information Types

Interoperability (URIs):
semantic data dictionary
semantic models

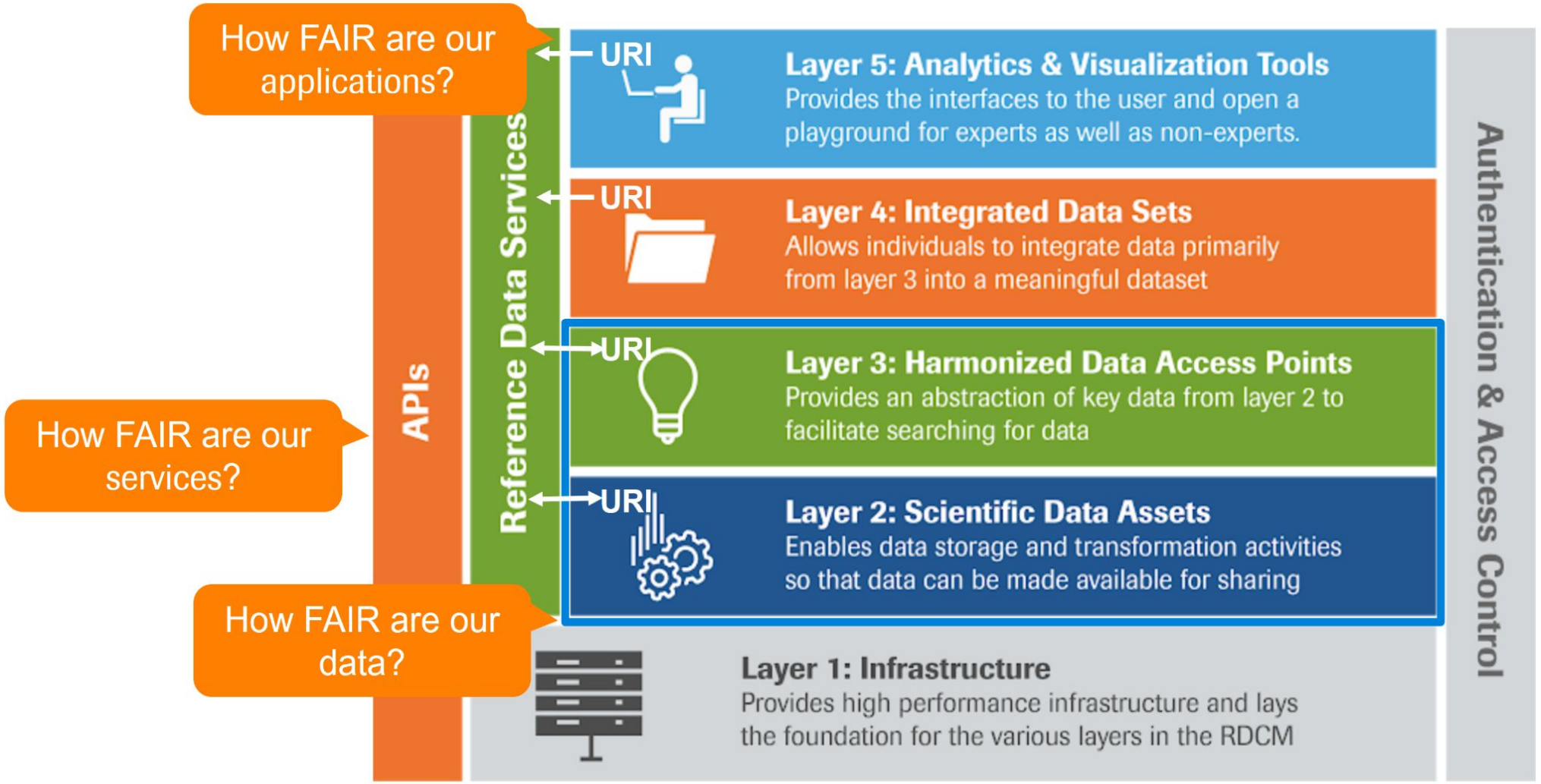
Data FAIRification only in layer 2 & 3

No more transformation between layer 3 & 4,5



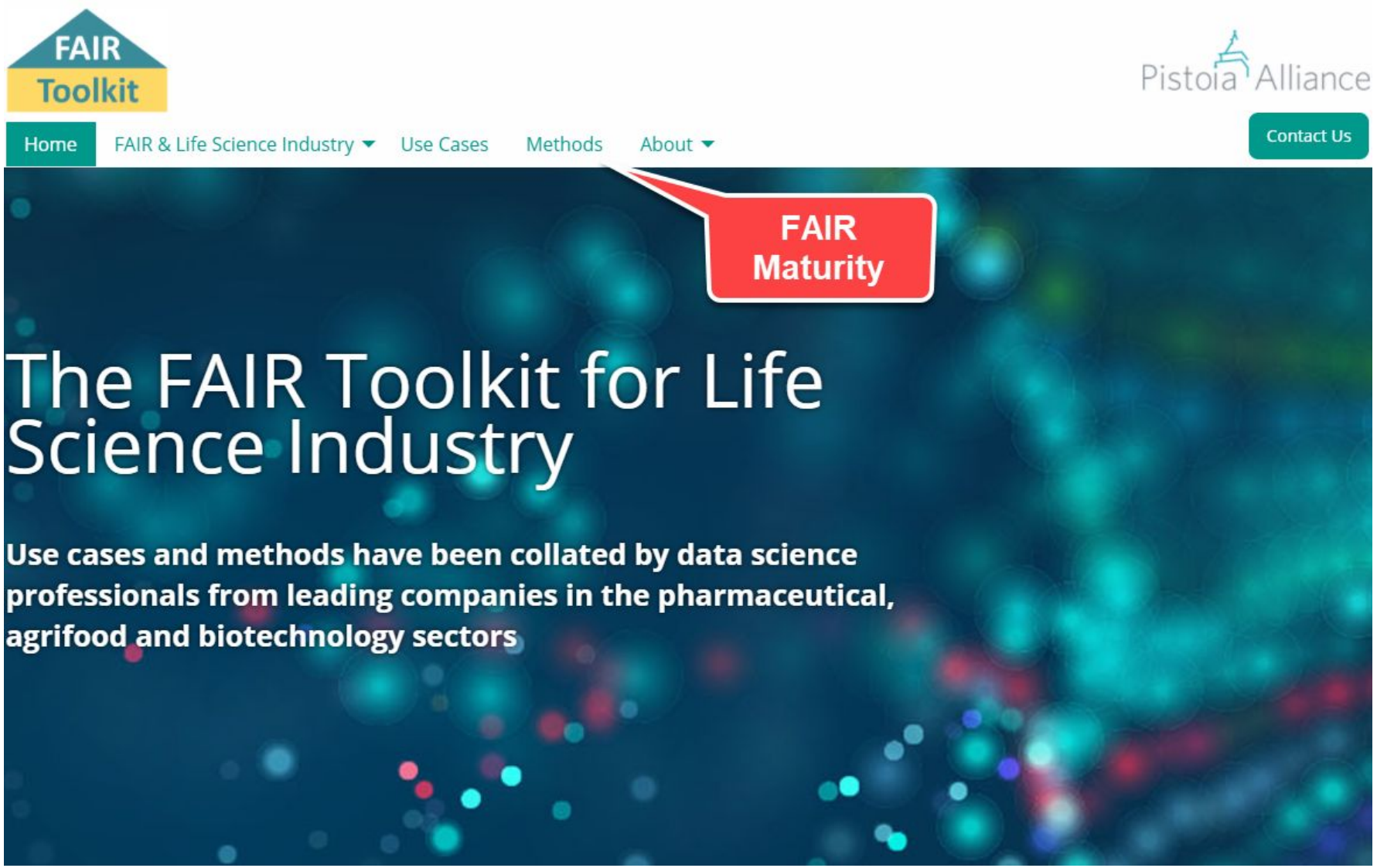
Roche Data Commons

Semantic Infrastructure of FAIR Applications, Services & Data



FAIR Assessment

FAIR Maturity



FAIR Toolkit

FAIR Assessment

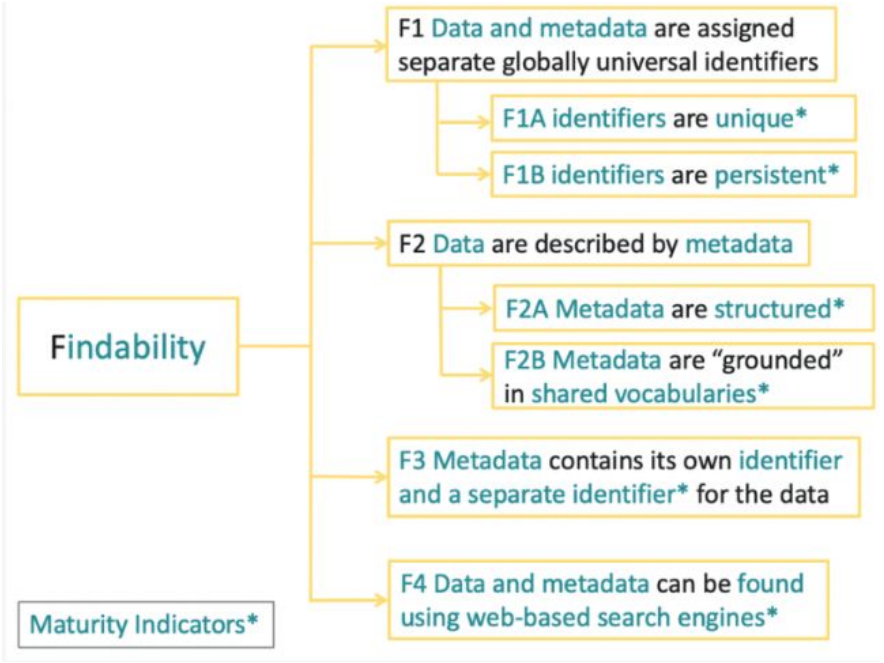
FAIR Maturity

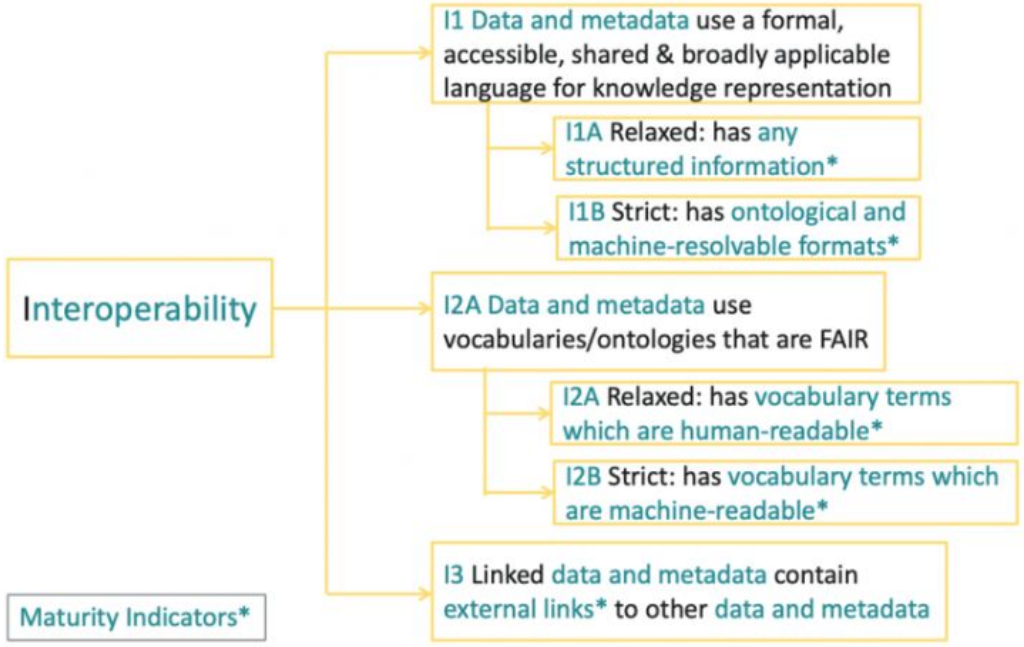
FAIR Toolkit

Pistoia Alliance

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

Concept

Subject ID	Sex	Age Value	Age Unit	Substance Name	Dose Value	Dose Unit	Dosage Form	Route	Dosage Frequency
S32345	male	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

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	male	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

Variable Concept

Subject ID	Sex	Age Value	Age Unit	Substance Name	Dose Value	Dose Unit	Dosage Form	Route	Dosage Frequency
S32345	male			b	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

ROX37507104443803061

Sex Terminology

FAIR data architecture

Variable & Schema

Subject ID	Sex	Age Value	Age Unit	Substance Name	Dose Value	Dose Unit	Dosage Form	Route	Dosage Frequency
S32345	male	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	
S33021	female	11	month	Polivy	12.5	mg/ml	inhalant	intrabronchial	weekly

Subject Schema

ROX37643616443820633

FAIR data architecture

Simple Graph Generation

Subject

Predicate

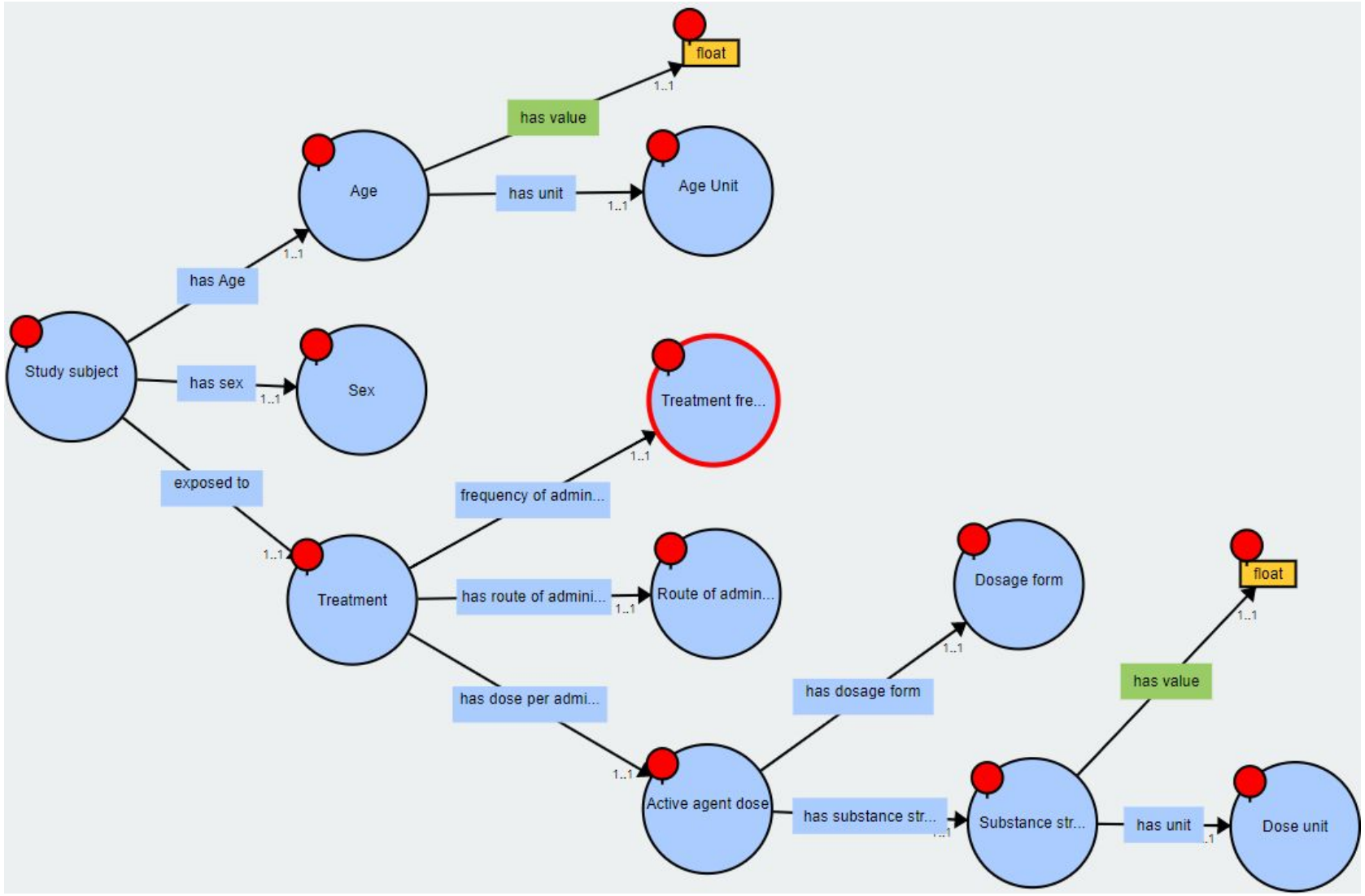
Subject ID	Sex	Age Value	Age Unit	Substance Name	Dose Value	Dose Unit	Dosage Form	Route	Dosage Frequency
S32345	male	230	day	bevacizumab	50	mg/ml	tablet	oral	daily
	female	11	year	oseltamivir	100	mg/l	solution	intravenous	weekly
S11324	male	88	year	trastuzumab	35	g/l	solution	subcutaneous	hourly
S92833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

Object

S32345 - hasSex - male
S32345 - ageValue - 230
S32345 - ageUnit - day

FAIR data architecture

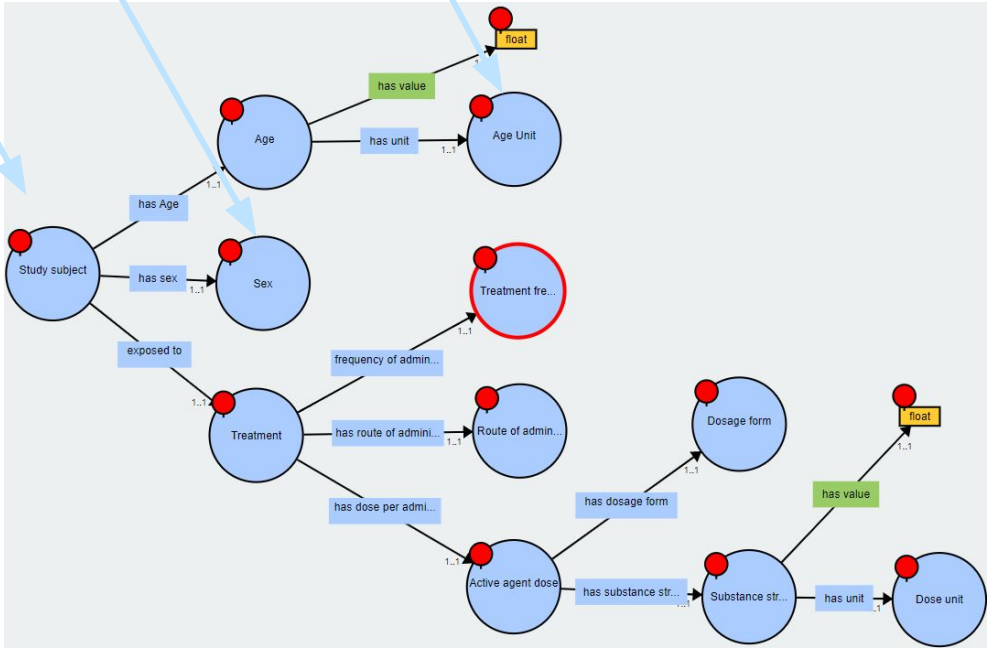
Semantic Graph Generation



FAIR data architecture

Semantic Graph Generation

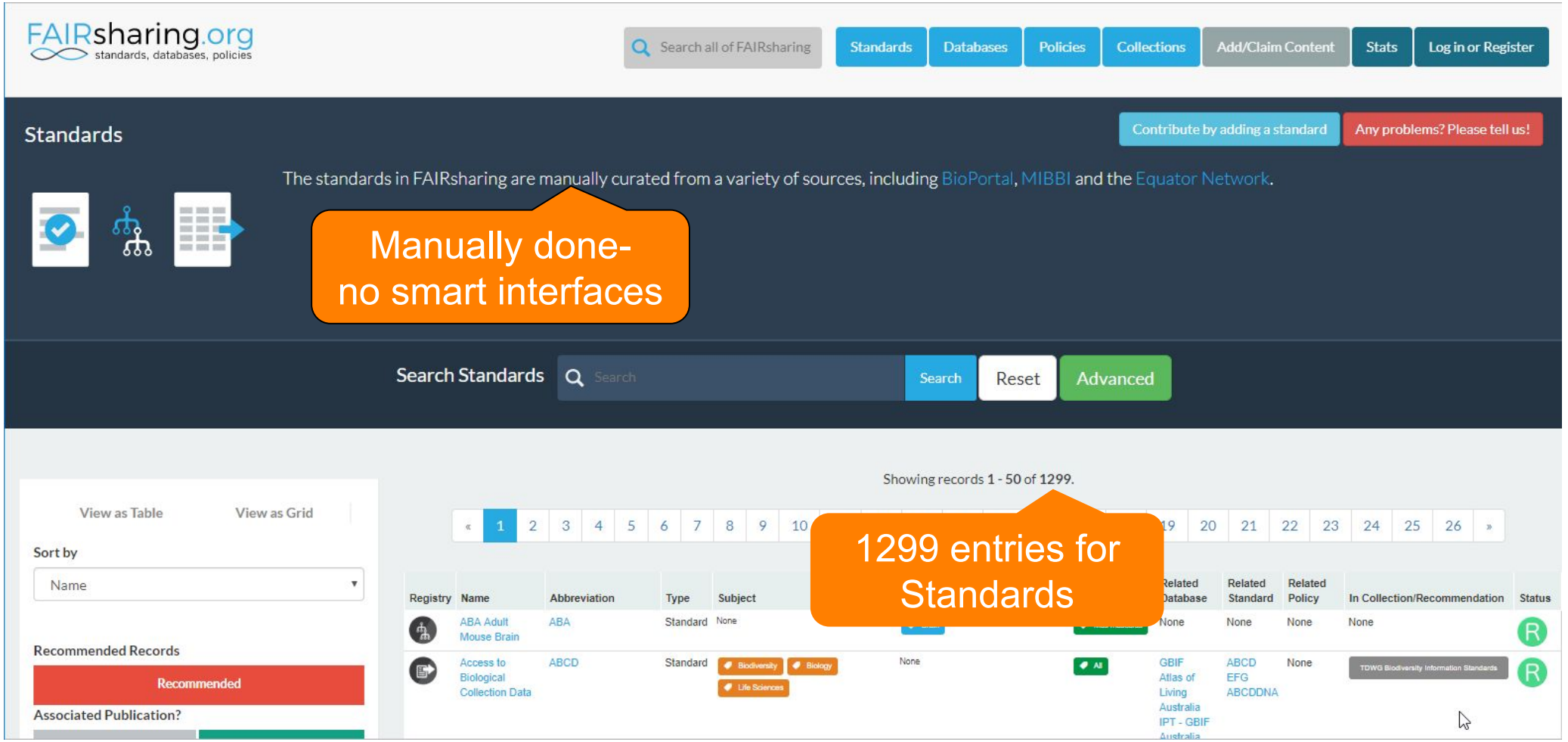
Subject ID	Sex	Age Value	Age Unit	Substance Name	Dose Value	Dose Unit	Dosage Form	Route	Dosage Frequency
S32345	male	230	day	bevacizumab	50	mg/ml	tablet	oral	daily



FAIR is ugly and complex

FAIRsharing Catalog of Biomedical Resources

Proliferation and Fragmentation of Standards



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

Demographics Domain (cdisc)

Name	Label
STUDYID	Study Identifier
DOMAIN	Domain Abbreviation
USUBJID	Unique Subject Identifier
SUBJID	Subject Identifier for the Study
BRTHDTC	Date/Time of Birth
AGE	Age
AGEU	Age Units
SEX	Sex
RACE	Race

Patient Model (OHDSI)

Field	Required	Type	Description
person_id	Yes	integer	A unique identifier for each person.
gender_concept_id	Yes	integer	A foreign key that refers to an identifier in the CONCEPT table for the unique gender of the person.
year_of_birth	Yes	integer	The year of birth of the person. For data sources with date of birth, the year is extracted. For data sources where the year of birth is not available, the approximate year of birth is derived based on any age group categorization available.
month_of_birth	No	integer	The month of birth of the person. For data sources that provide the precise date of birth, the month is extracted and stored in this field.
day_of_birth	No	integer	The day of the month of birth of the person. For data sources that provide the precise date of birth, the day is extracted and stored in this field.
birth_datetime	No	datetime	The date and time of birth of the person.
death_datetime	No	datetime	The date and time of death of the person.
race_concept_id	Yes	integer	A foreign key that refers to an identifier in the CONCEPT table for the unique race of the person, belonging to the 'Race' vocabulary.
ethnicity_concept_id	Yes	integer	A foreign key that refers to the standard concept identifier in the Standardized Vocabularies for the ethnicity of the person, belonging to the 'Ethnicity' vocabulary.
location_id	No	integer	A foreign key to the place of residency for the person in the location table, where the detailed address information is stored.
provider_id	No	integer	A foreign key to the primary care provider the person is seeing in the provider table.
care_site_id	No	integer	A foreign key to the site of primary care in the care_site table, where the details of the care site are stored.
person_source_value	No	varchar(50)	An (encrypted) key derived from the person identifier in the source data. This is necessary when a use case requires a link back to the person data at the source dataset.
gender_source_value	No	varchar(50)	The source code for the gender of the person as it appears in the source data. The person's gender is mapped to a standard gender concept in the Standardized Vocabularies; the original value is stored here for reference.
gender_source_concept_id	Yes	Integer	A foreign key to the gender concept that refers to the code used in the source.

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

Data Standards & Interoperability Challenges

CDISC vs OMOP/ OHDSI vs DICOM

Demograp

Domain

Name
STUDYID
DOMAIN
USUBJID
SUBJID
BRTHDTC
AGE
AGEU
SEX
RACE

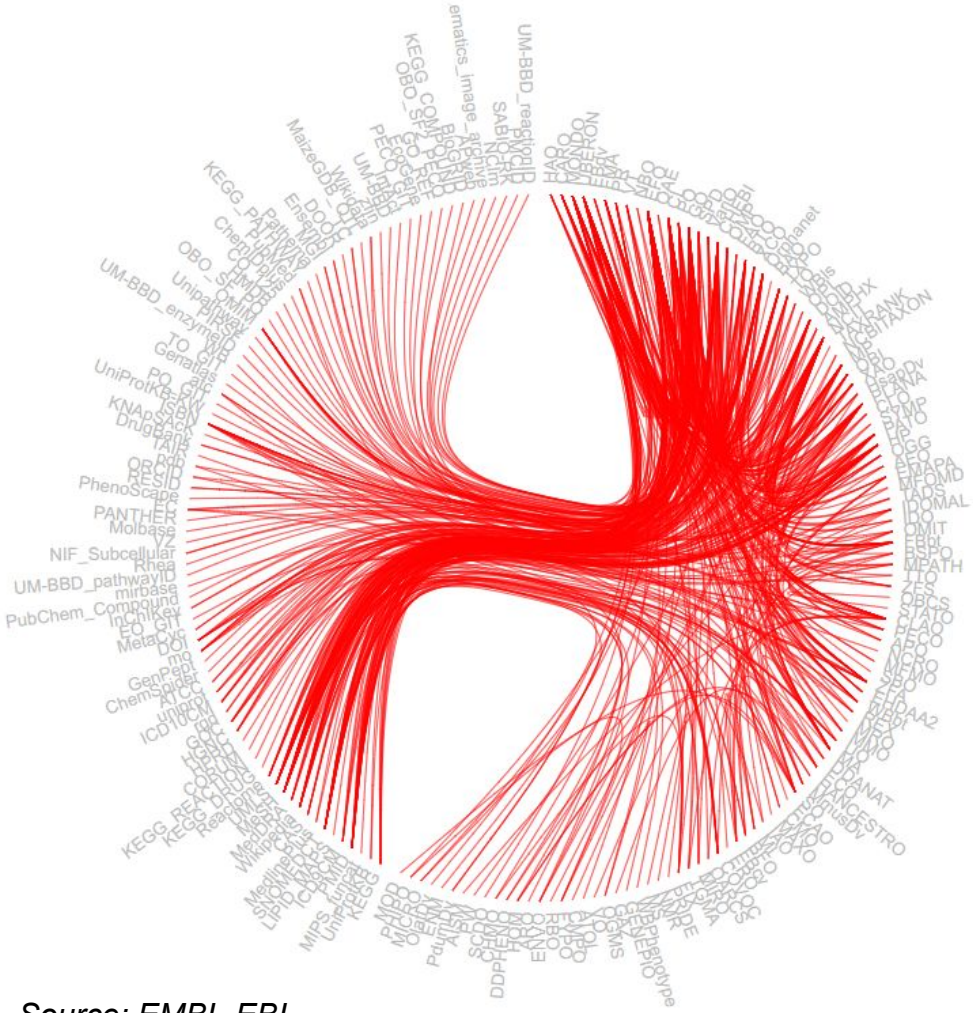
DICOM Standard Browser by Innolitics

▼ CR Image			CIOD
▼ Patient		M	Module - Patient
▶	(0008,1120) Referenced Patient Sequence	3	Sequence
	(0010,0010) Patient's Name	2	Person Name
	(0010,0020) Patient ID	2	Long String
	(0010,0021) Issuer of Patient ID	3	Long String
	(0010,0022) Type of Patient ID	3	Code String
▶	(0010,0024) Issuer of Patient ID Qualifiers Sequence	3	Sequence
▶	(0010,0026) Source Patient Group Identification Sequence	3	Sequence
▶	(0010,0027) Group of Patients Identification Sequence	3	Sequence
	(0010,0030) Patient's Birth Date	2	Date
	(0010,0032) Patient's Birth Time	3	Time
	(0010,0033) Patient's Birth Date in Alternative Calendar	3	Long String
	(0010,0034) Patient's Death Date in Alternative Calendar	3	Long String
	(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



Source: EMBL-EBI OxO

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

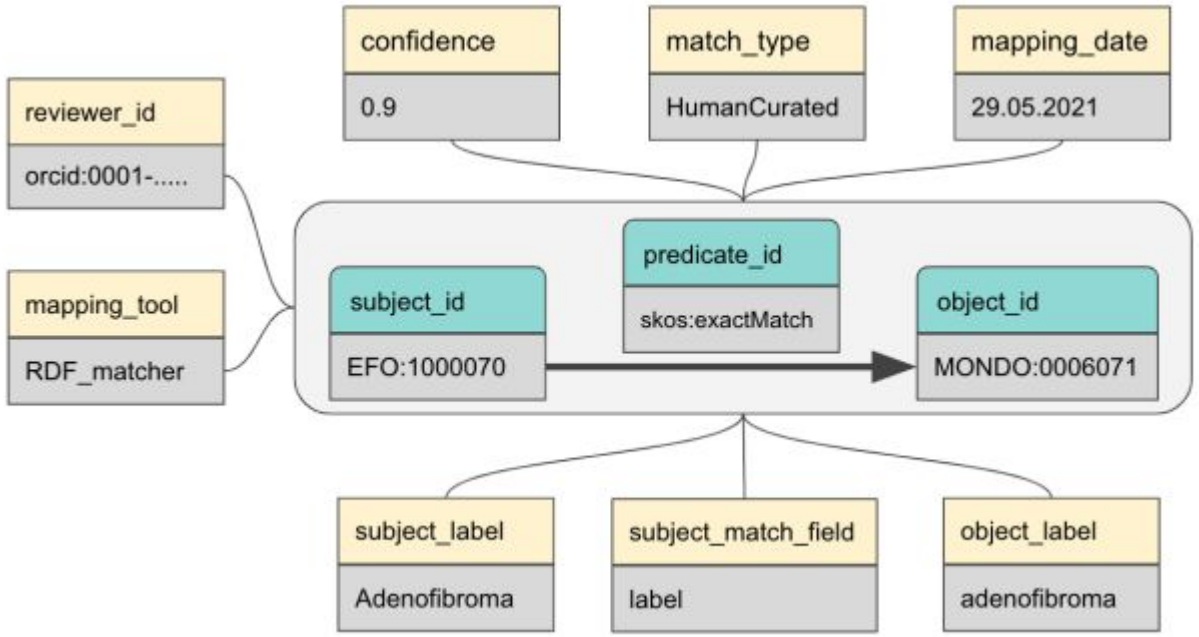
SSSOM is a simple metadata standard for describing semantic mappings:

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

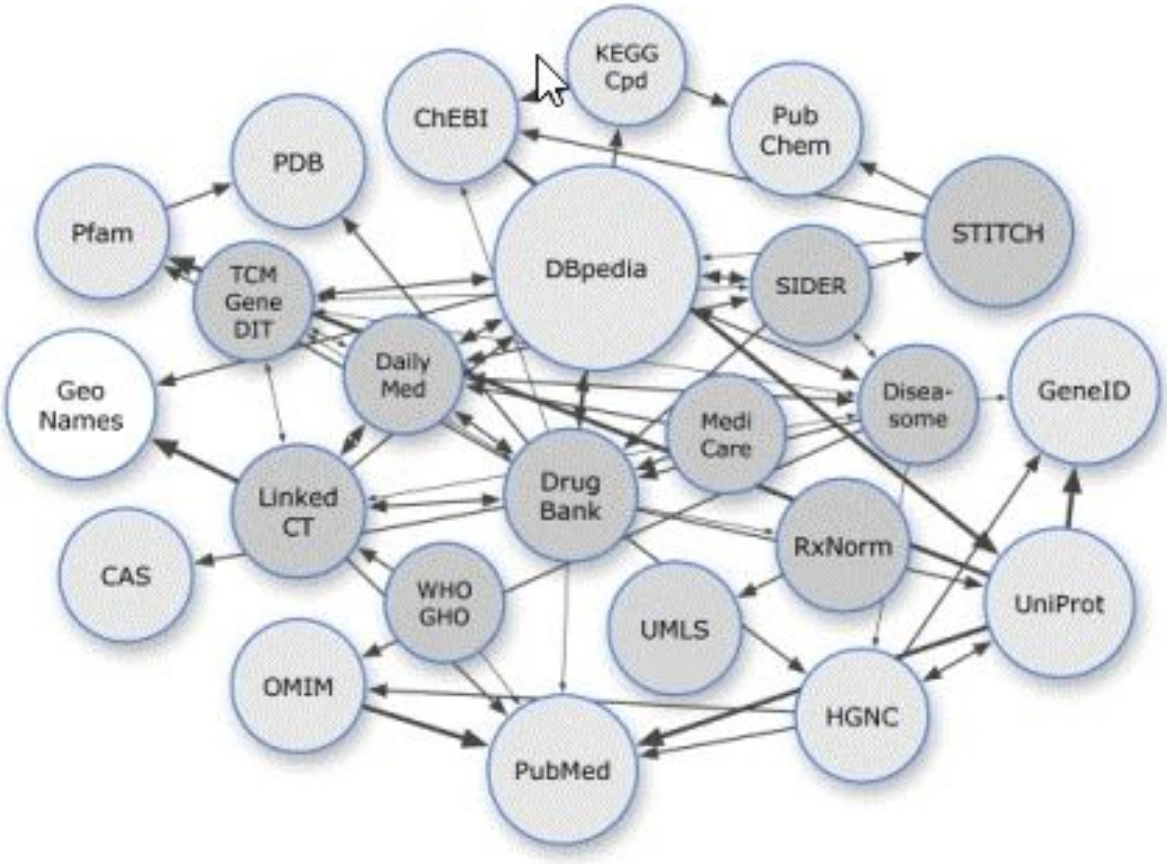


Not fully FAIR (dct:creator & dct:created)
No guidelines on property labels

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)

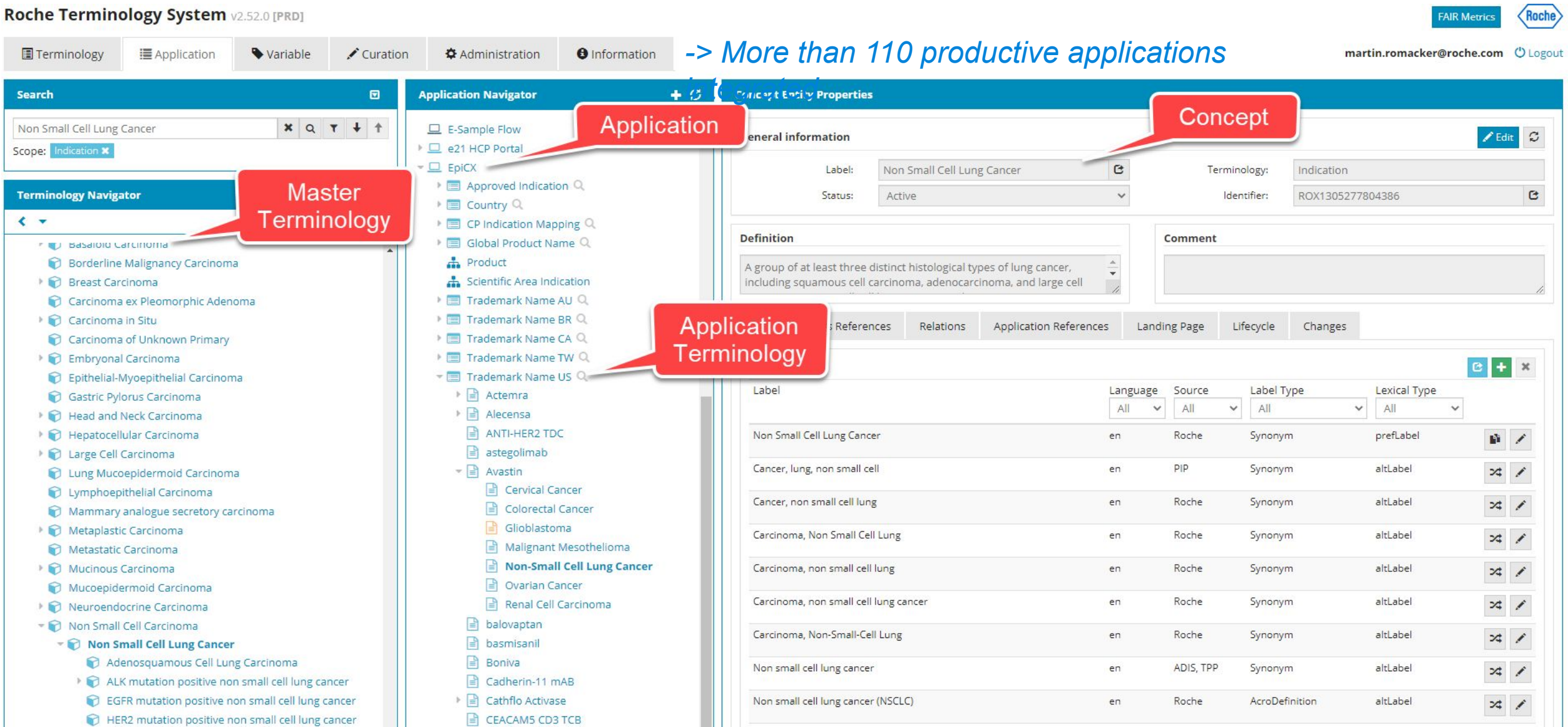
Products are FAIR by design supporting FAIRification at scale for the entire Roche organization. Data Harmonization Service ensures semantic interoperability & high data quality.

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.

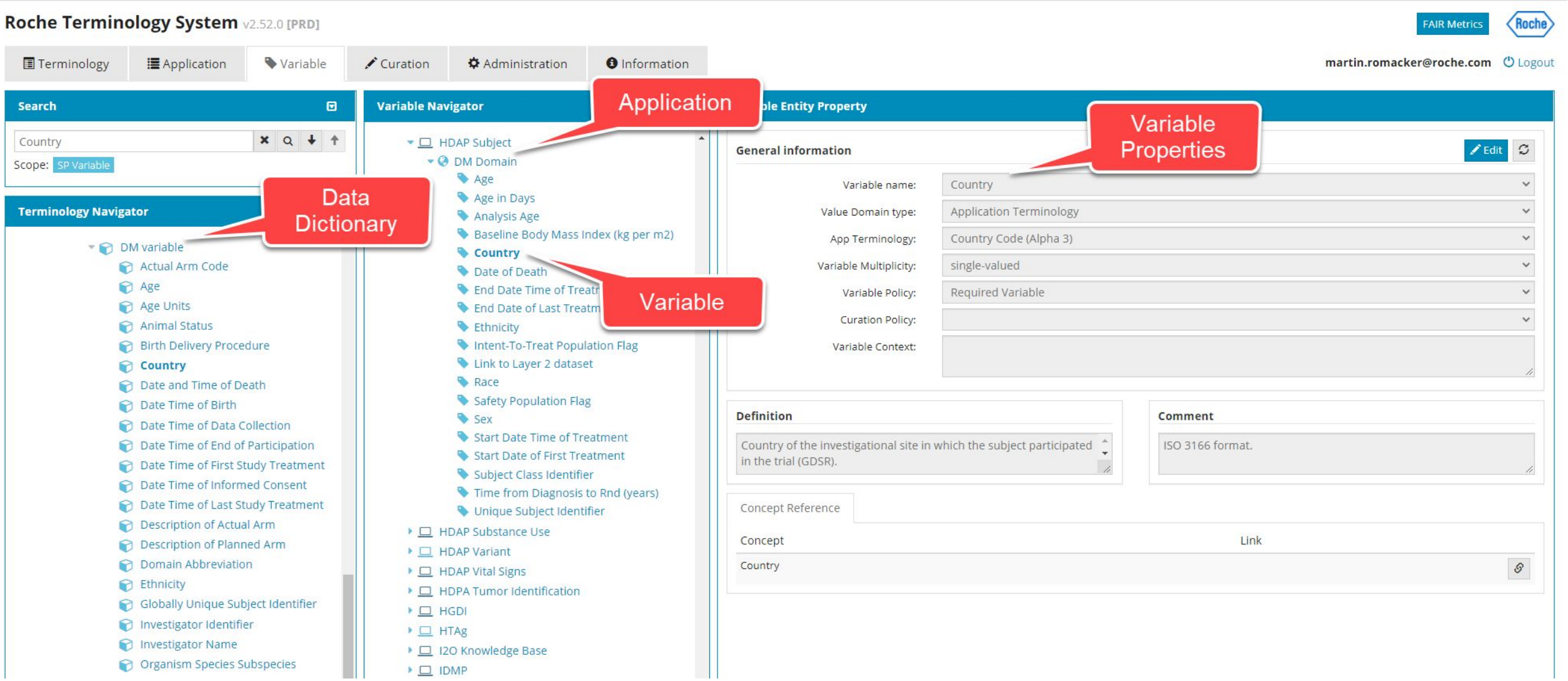
Reference Data Services for Data Management

Terminology Management - Contextualize Concepts (FAIR)



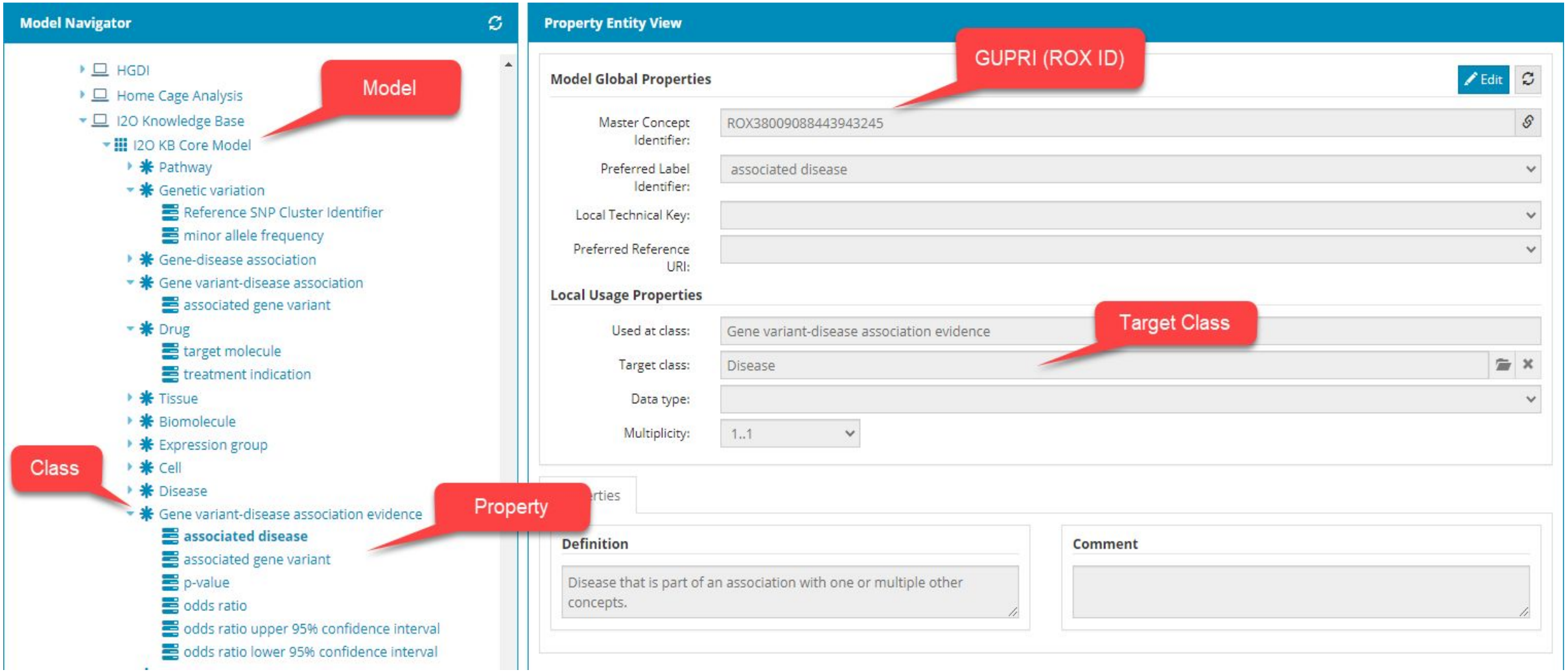
Reference Data Services for Data Management

Metadata Registry/ Dataset Models – Metadata Harmonization (FAIR)



Reference Data Services for Data Management

Concptual Model - Purpose-build FAIR Ontologies



Semantic Interoperability Hub - Capability Stack

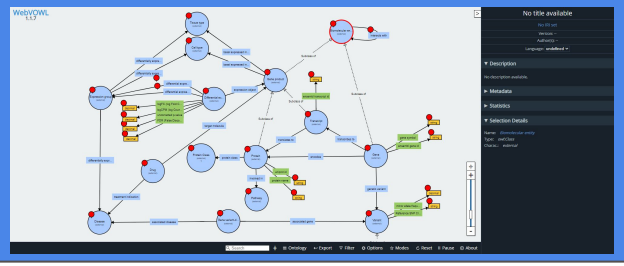
Data Management Value Chain - From Terminologies to a Unified Domain Model

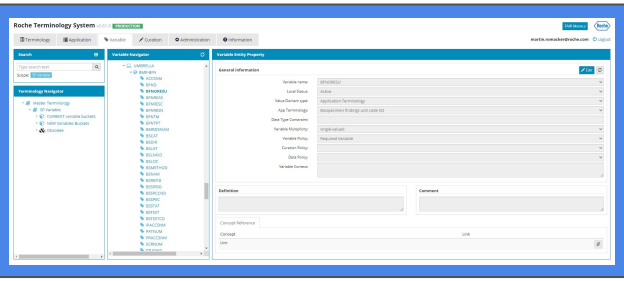
Terminology Managment (RTS)

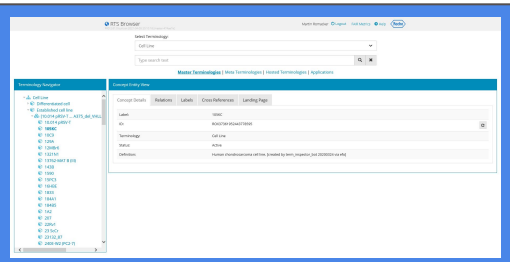
Dataset Model (MDR)

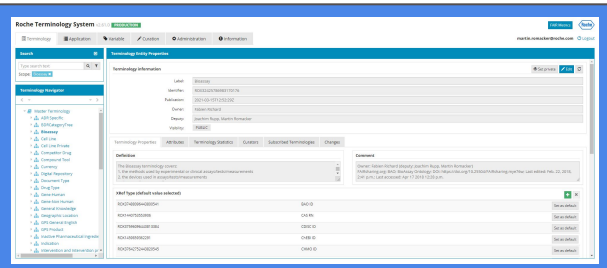
Conceptual Model

FAIR Unified Domain Model (KG)



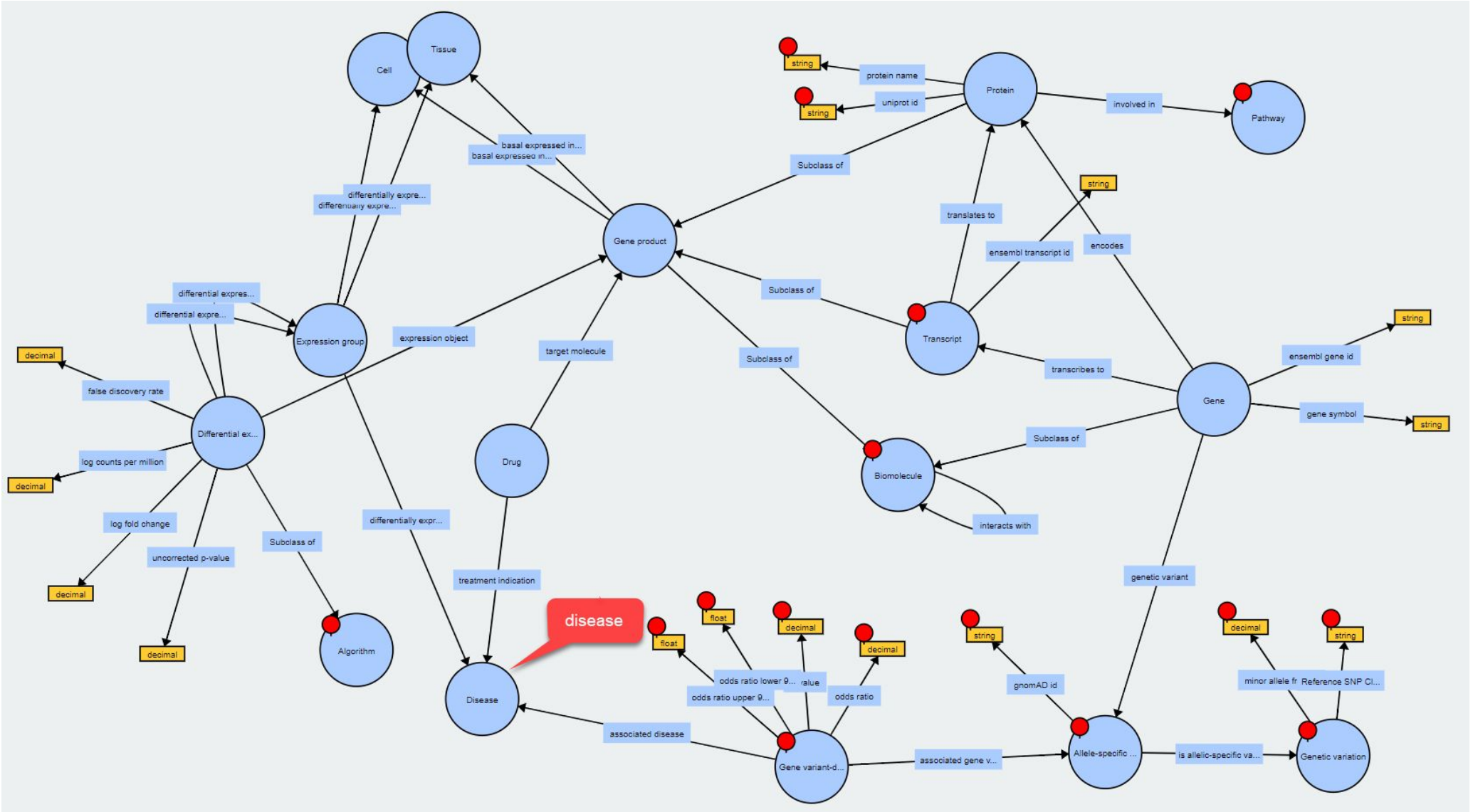






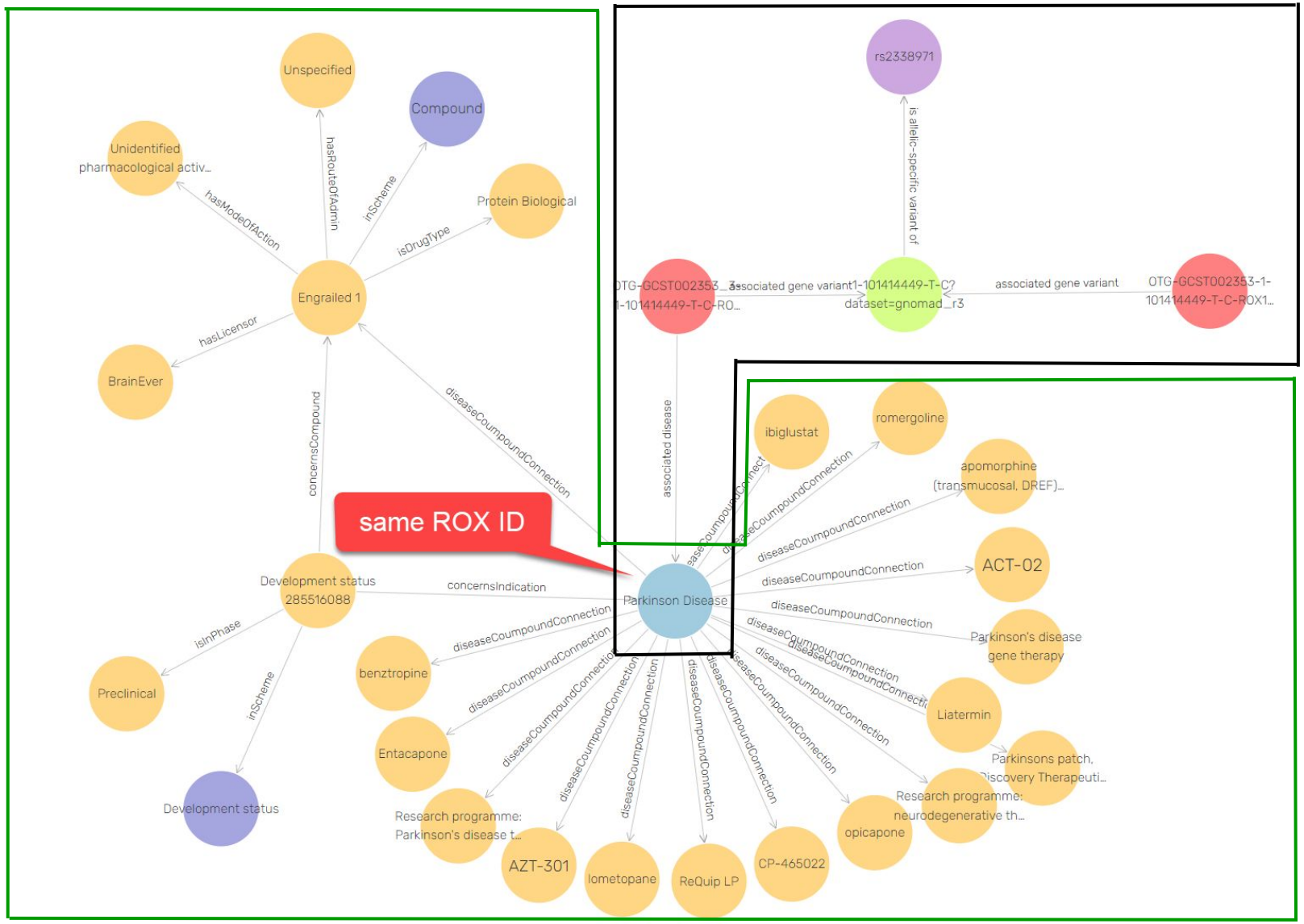
Infectious Disease Ontology

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 standarized 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

Felix Schwagereit

Scientific Technical Manager, Basel

Pratishtha Duhan

Business Manager, SSF

Rama Balakrishnan

Biomedical Ontology Specialist, SSF

Shima Dastgheib

Semantic Integrator, SSF

Dataset portal team:

- Hugo de Schepper
- Oliver Steiner
- Roy Weiler

Roche Terminology System

Dev and Ops Team, Curation Team

RTS Dev and OPS Team

Michal Bielak

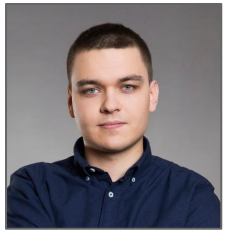
Michal Paradowski

Adam Zawada

Konrad Borowka

Robert Trypuz

Agnieszka Bananszynska-Krolikiewicz



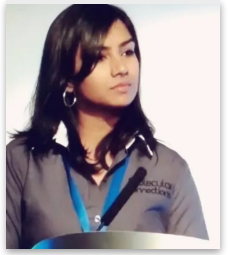
Majewski Krzysztof

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

Rancho Biosciences Team

Erfan Younesi
Sr. Data Curator

Svetlana Koltsova
Sr. Data Curator

Maxim Papin
Sr. Data Curator

Curation supported by:

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

Doing now what patients need next