Beauty & the Beast – FAIR Linked Data and the Reality (a pharma perspective)

Martin Romacker, Data and Information Architect
Scientific Solution Engineering and Architecture (S2EA)
Data & Analytics
Roche Innovation Center Basel
PLDN 10 Jaar Conference, 27th September 2022, Hilversum, Netherlands

Platform Linked Data Nederland

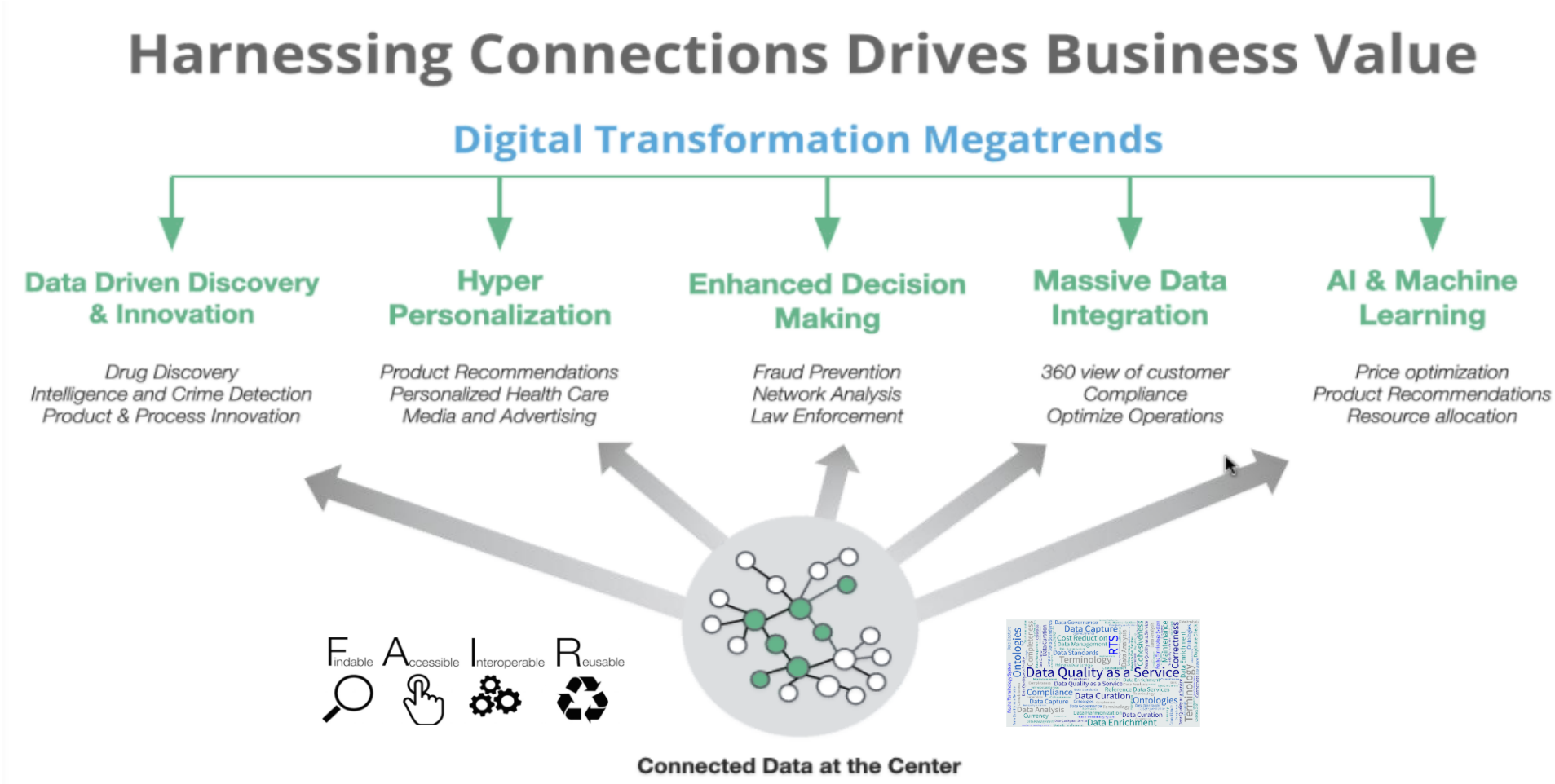
Roche pRED

Digital Transformation & Management of Data Assets

FAIR plus Q

Digital Transformation

Megatrends & Data Management Strategy



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

Data as an Asset

True Costs of Data Management

Planned/ Visible Costs

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

Unplanned/ Invisible Costs

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

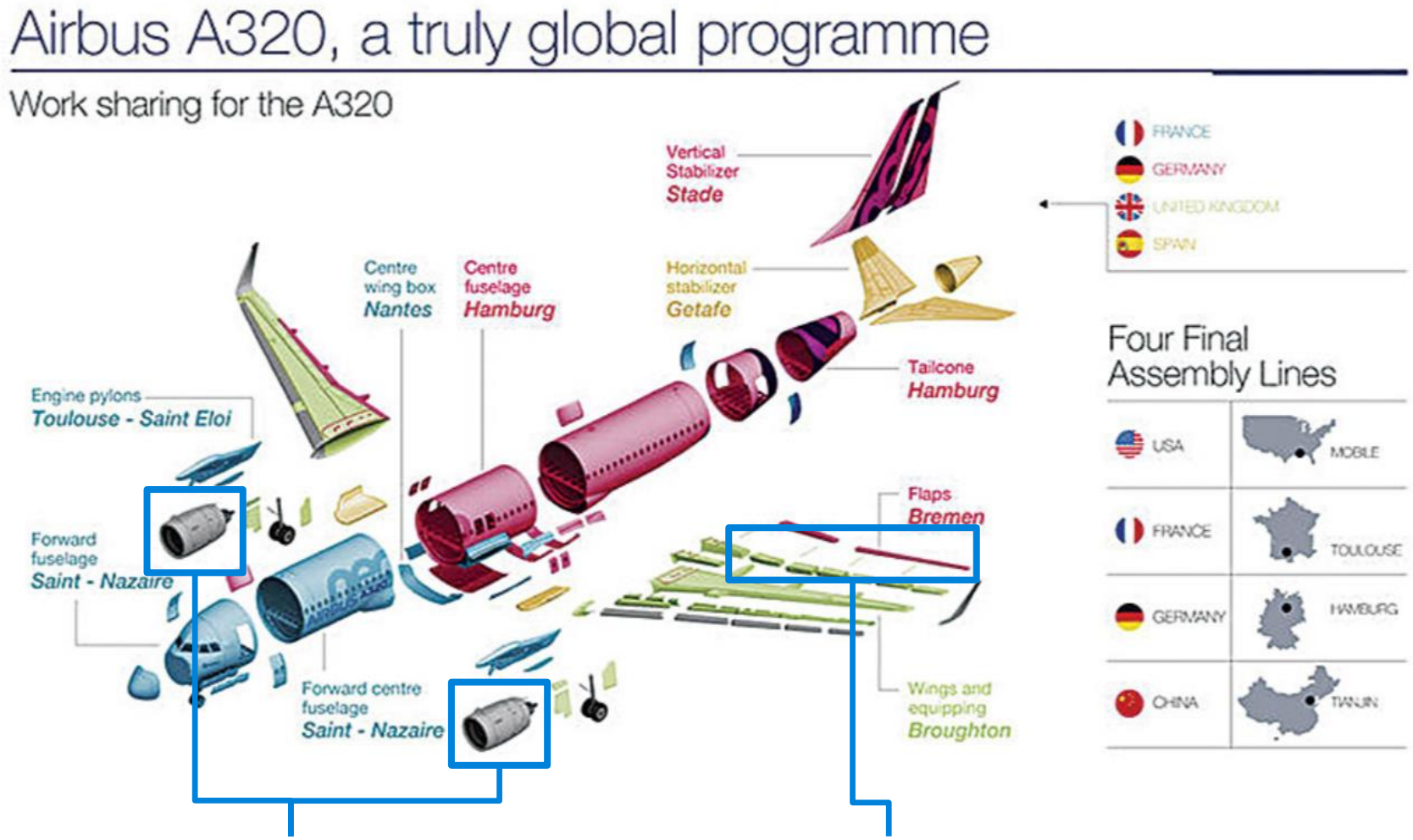
Backcharge the costs for processing to the data producers

Information Procurement

Transformationless Integration of Data Assets

Manufacturing

Components & C-Parts Procurement



- highly complex product
- highly distributed process
- vertical integration (CMO)
- assembly in different global geographic regions

Are data-driven data management/ information assembly processes different?

Contract Manufacturing Organization

Technical Connections/ Interfaces

Information Procurement

Data-Driven and Knowledge-Based Pharma R&D (Ontologies)

- Information Architecture:
  Information-centric data organization - semantically sound and meaningful (ontologies)

- Information Procurement:
  the effective and efficient process of creating, acquiring and integrating standardized information types into information-driven R&D activities

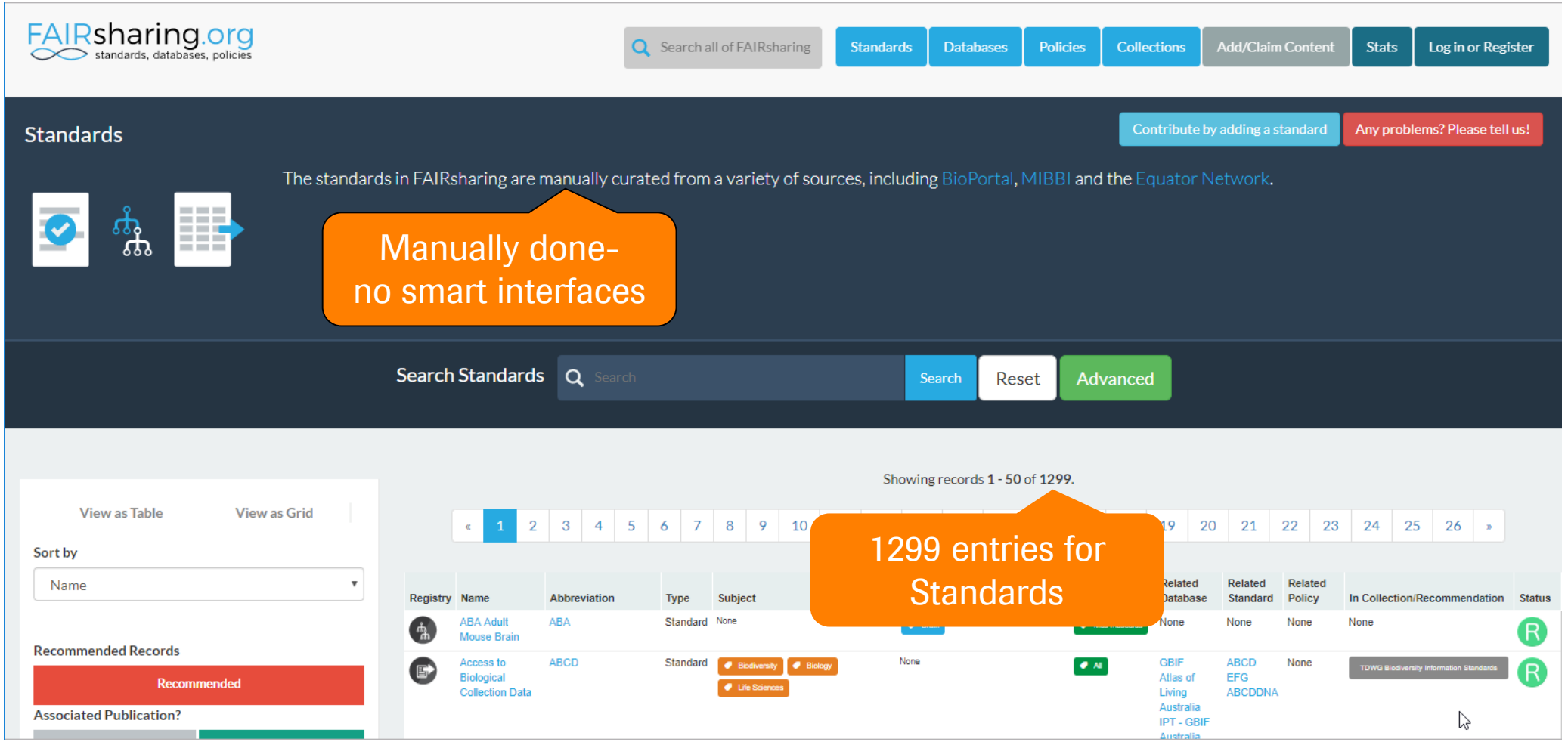
  - creation: an in-house activity which will result in a new data asset
  - acquisition: an internalization of a new data asset created by an external organization
  - integration: assembly of internal and external data assets into larger meaningful assets
  - information type: primary building blocks for representation of data assets based on interoperable Minimal Models using community standards (FAIR data principles)

Biomedical Ontologies & Terminologies

Missing Community Strategy – Intractable Knowledge Space

FAIRsharing Catalog of Biomedical Resources

Proliferation and Fragmentation of Standards

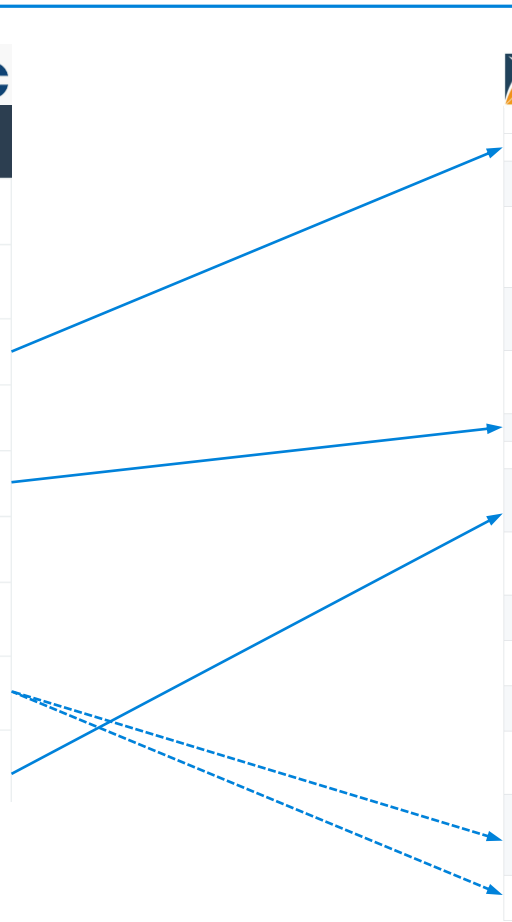


Data Standards & Interoperability Challenges

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



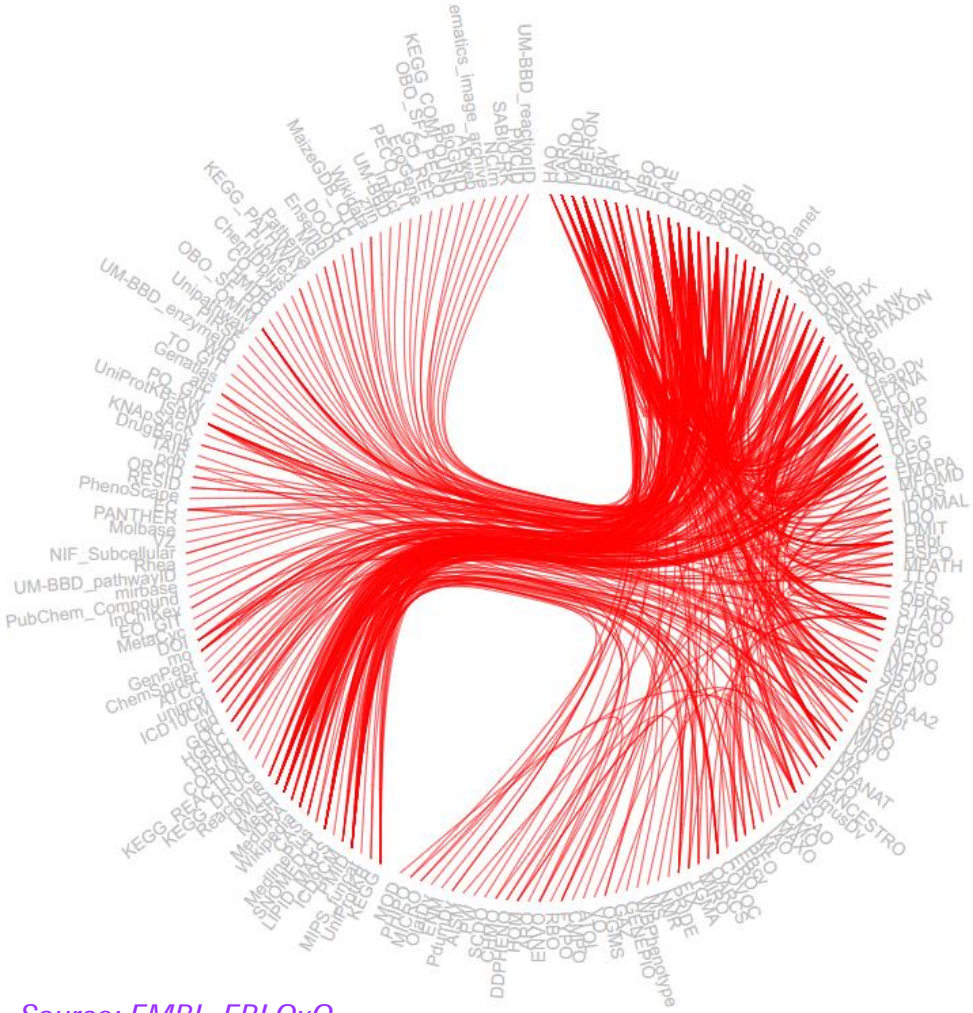
Patient Model (OHDSI)

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

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)

Source: EMBL-EBI OxO

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

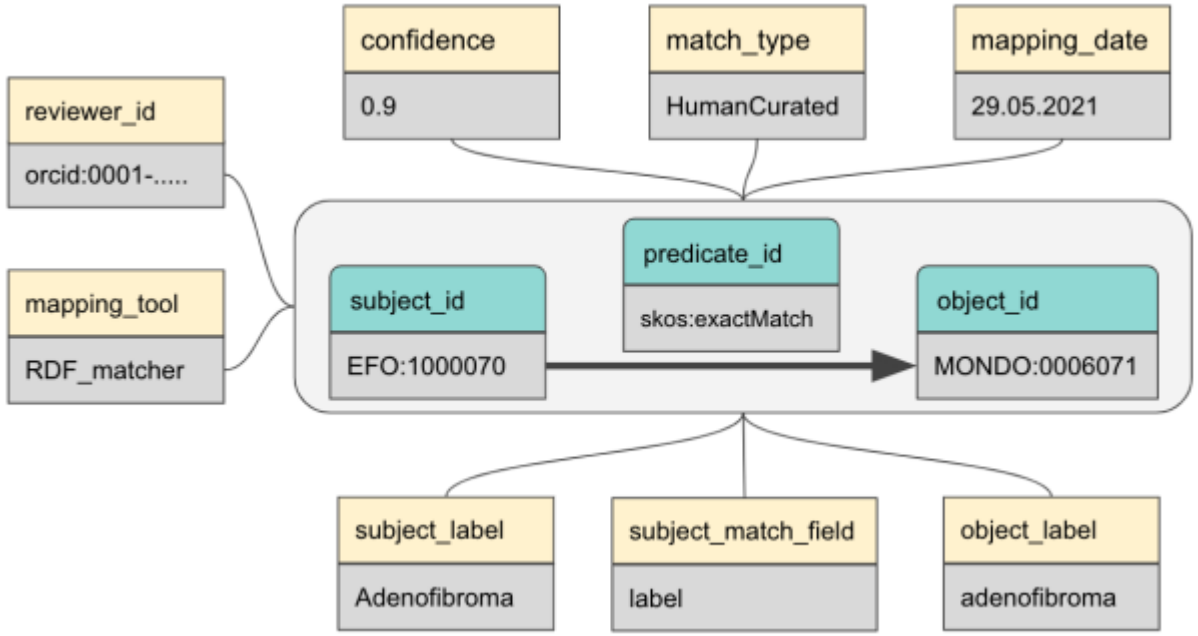
SSSOM is a simple metadata standard for describing semantic mappings:

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

Digital Transformation & FAIRification at Scale

Industry Approach –Vision & Reality

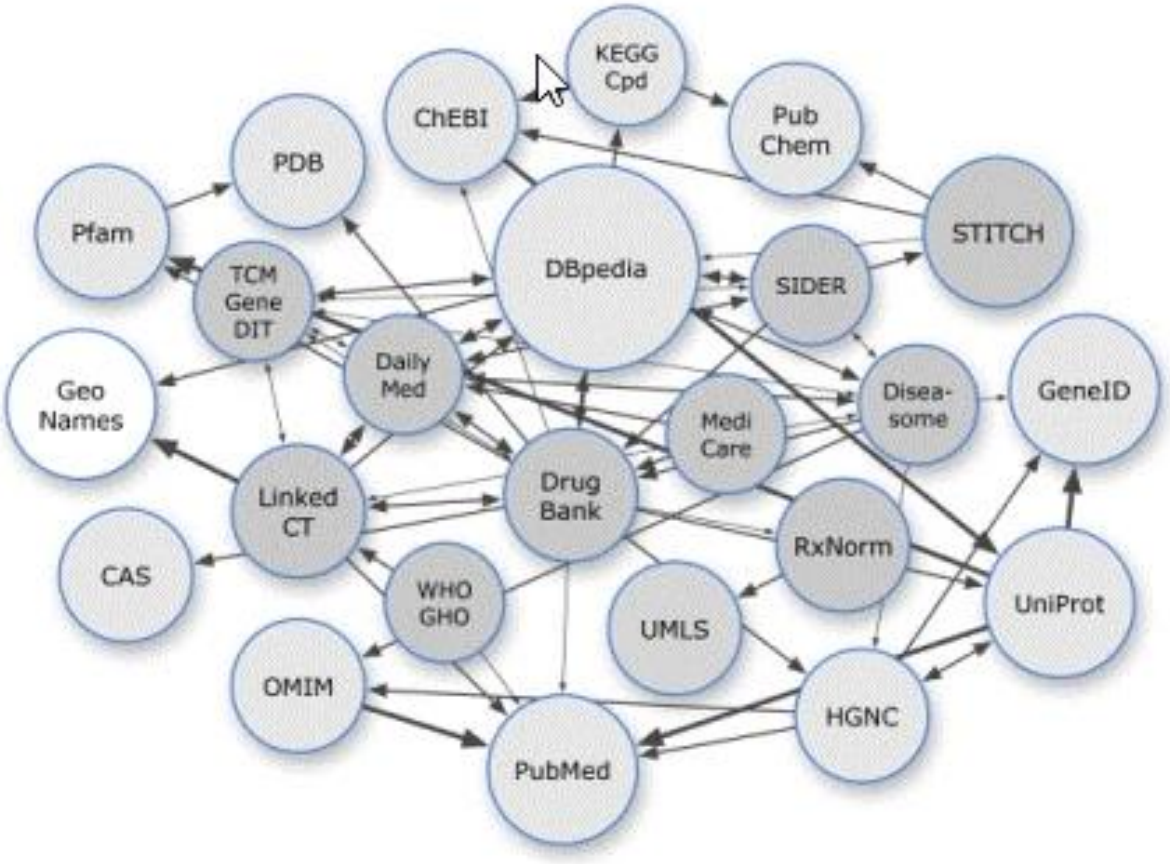
FAIRification at Scale using Biomedical Ontologies

Vision

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

Linked Open Data Cloud

Linked Open Drug Data Cloud (LODD)



Data in the Linked Data Cloud is not linked

Linkage comes with referential identity

Referential identity comes with interoperability

LODD is not FAIR

Digital Data Assets & Data Management

The Reality - No Standardization, unlinked, unFAIR

A	B	C	D	E	F	G	H	I
External_ID	**MOLECULE**	**ORIGIN**	**Target_class_orig**	**Target_orig**	**MoA_orig**	**MoA_class_orig**	**Action_o**	**Pathway_orig**
	43	Data Provider 1			Less active enantiomer of naturally-occurring (-)-nicotine	Cholinergic	Agonist	
Sata 3	ENVxEa232412	Data Provider 2		PARP	is found to be a inhibitor of poly (ADP-ribose) synthetase of rat pancreatic islet cells.			DNA Damage
reposENVxEa3792	ENVxEa37847	Data Provider 3			Hypertension - a beta blocker. A metabolite of , ENVxEa34787-xEa0			
ZyQ1572291	ENVxEa901222	Data Provider 4	Ion Channel	Ca2+ channel				
Z203042168	ENV0145777	Data Provider 5		MetabotENVpic glutamate receptor				
cENVwdENV07114	ENV07166121	Data Provider 6		mGlu1				
ENV050307	ENV065122	Data Provider 7			Antibiotic betalactam carbapenem			
ENVpr 8949	ENV0232156	Data Provider 8			VEGFR inhibitor. Also inhibits KIT, RET, MET and FLT3			

8 different data providers

different metadata

Free Text

different value domains

Coordination

- Data Curation: creation of evitable additional costs
- Data Curation: creation of evitable project delay
- Data Consolidation/ Data Integration: misallocation of expensive resources (data scientists)
- Information loss (not all information can be recovered)
- Insight creation (data monetization) serverly impacted

Why FAIR and Linked Data might fail

The seven deadly sins

Seven Deadly Sins of the FAIR Community

Deadly Sin #1

"The FAIR community fails to make clear what FAIR really is. In particular, the implications of FAIRifation on how we work in Knowledge Management and IT projects are not clear."

Seven Deadly Sins of the FAIR Community

Deadly Sin #2

"There are foundational misunderstandings about the scope of Data FAIRification in particular in the context of Data Quality Frameworks. FAIR Data and High Quality Data are not the same."

Seven Deadly Sins of the FAIR Community

Deadly Sin #3

"We cannot really measure FAIR maturity as the FAIR maturity model is almost incomprehensible. This heavily impacts the correct adoption of the FAIR principles"

The discovery of digital object should be possible from its metadata. For this to happen, the metadata must explicitly contain the identifier for the digital resource it describes, and this should be present in the form of a qualified reference, indicating some manner of "about" relationship, to distinguish this identifier from the numerous others that will be present in the metadata.

In addition, since many digital objects cannot be arbitrarily extended to include references to their metadata, in many cases the only means to discover the metadata related to a digital object will be to search based on the GUID of the digital object itself.

Seven Deadly Sins of the FAIR Community

Deadly Sin #4

"The biomedical community does not really converge on standards. We rather increase the chaos instead of harmonizing both in terms of semantic resources and harmonization projects resulting in an intractable knowledge space."

Seven Deadly Sins of the FAIR Community

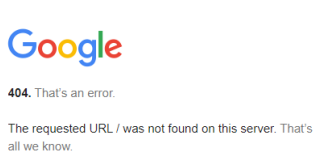
Deadly Sin #5

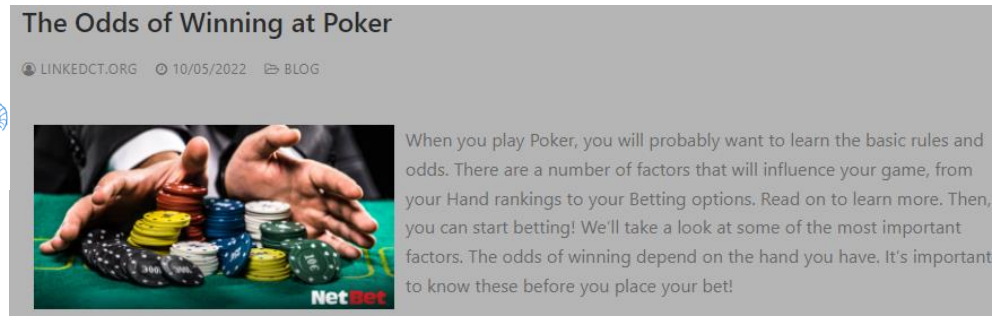
"The FAIR community stays in a bubble. Insiders connect with insiders. The outreach and integration with the business is poor. Besides a vague understanding there is little support for a true break through from management."

Seven Deadly Sins of the FAIR Community

Deadly Sin #6

"Scaling up knowledge management based on FAIR resources and standards requires an operational backbone of FAIR data and services – but no community strategy about the basic resources and their maintenance"





http://www.ebi.ac.uk/rdf/ *http://lov.okfn.org* *http://linkedct.org*

Seven Deadly Sins of the FAIR Community

Deadly Sin #7

"Implementation of a landscape of FAIR data, services and application requires a high engagement with key communities such as IT Architects, Master Data Management, Data Managers. Until now FAIR does not speak to them."

snowflake MuleSoft Collibra databricks Veeva PEGA RELTIO

Disclaimer: This is an arbitrary listing of tools and vendors without implying any recommendation or preference

Seven Deadly Sins of the FAIR Community

Deadly Sin #X

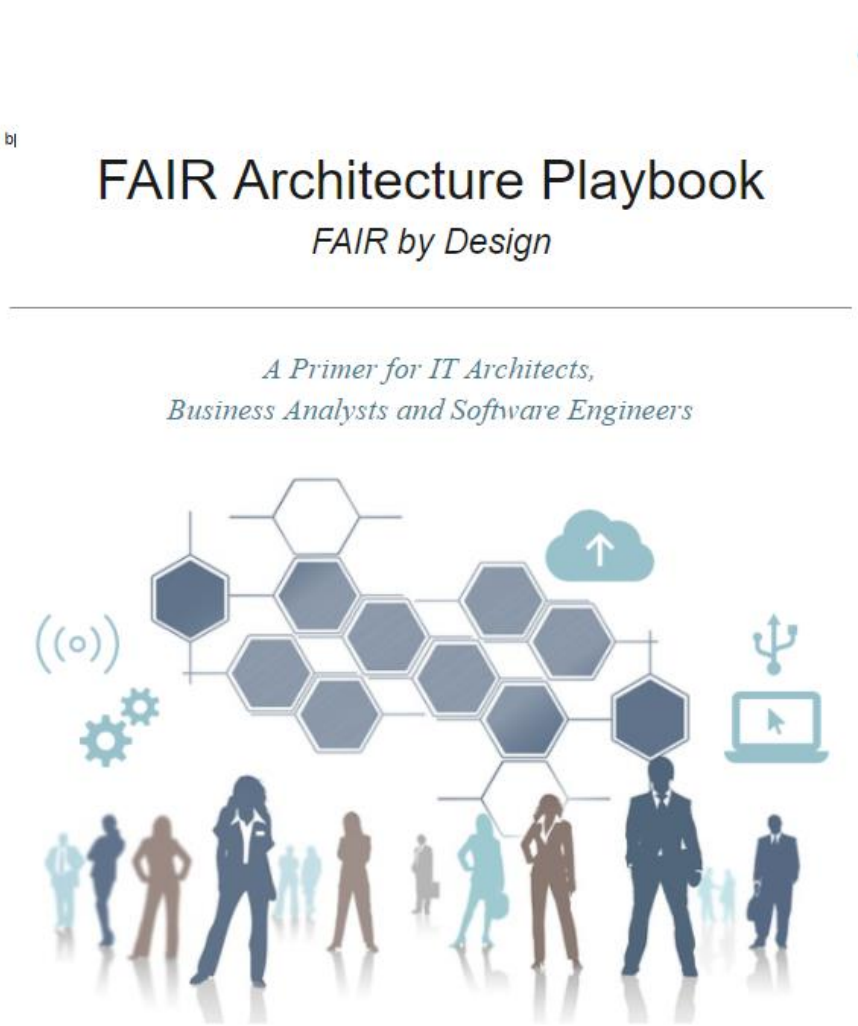
"Using RDF and OWL to build Ontologies as well as creating Knowledge Graphs does not at all prevent you from establishing new data silos or producing unFAIR data."

Digital Transformation & FAIRification at Scale

FAIR by Design

FAIR Playbook for IT Professionals

Targeting the Key Enablers



IT & its professionals as key enabler.

Data Managers should not care about FAIR.

FAIR Data & Identifiers

Global Unique Persistent Resolvable Identifiers (GUPRI)

Globally Unique: *Uniqueness* means that any identifier refers to exactly one Digital Object. *Global validity* means that every Digital Object should have exactly one identifier for reference where *global* is not limited to our organization but ideally would also include the external universe of discourse.

Persistent: An identifier never ever changes. An identifier never gets deleted even if the related Digital Object ceases to exist. The metadata of the identifier should also be maintained.

Resolvable: Identifiers are resolved by a service that returns the latest version of the object, including its metadata.

Transposing these principles to our organization and establishing FAIR identifier management, we need to define and enforce company-wide or even global policies:

- Namespace **registration**: Provision of a repository and a service supporting the definition and governance of namespaces used for the creation of identifiers.
- **GUPRI policies**: Definition of the format and structure for namespaces and identifiers.
- **Generation/minting of GUPRIs**: Unambiguous creation of unique identifiers by a service.
- **GUPRI resolution service**: Service enabling the resolution of GUPRIs for finding and accessing resources.

Conclusion:

FAIR applications, services, and data require governance, policies, and infrastructure to manage the identifiers space at the global scale.

***Opaque* GUPRI:** no semantics is encoded in the structure of the GUPRI, and it consists solely of the namespace and an identifier. For example, RTS follows this principle by combining the namespace "http://ontology.roche.com/" with a random but unique identifier "ROX1302017050223" to "http://ontology.roche.com/ROX1302017050223". The GUPRI does not reveal any semantically relevant information about the entity it refers to.

***Speaking* GUPRI:** There are additional elements in the GUPRI giving the consumer hints about the context of this resource. Table REF offers an example. The namespace "http://clinical.roche.com/study/" exposes the semantic type of the resource "Study" in the name. This supports the human readability of GUPRIs. Systems for defining speaking GUPRIs can be very sophisticated[10].

Digital Transformation & FAIRification at Scale

Standardization & Capability Stack

FAIRification at Scale: Capability Stack

From Terminologies to Domain Models

Terminology Management: The concepts used in our scientific and technical domains are properly defined, typed and organized in a *Terminology Management System*. Each *concept* is given an unambiguous, complete, *preferred label* and a *textual definition*. The concept is complemented by a rich *synonym set* and *cross-references* linking semantically equivalent concepts in other internal or external repositories.
Every concept is represented by a global, unique, persistent, and resolvable identifier serving as a reference.

Dataset Model Management: In essence, a *dataset model* describes a fully harmonized representation of a *table-like data structure*. The column headers refer to *metadata elements* (variables, field names, properties, attributes – many different names are used). All the metadata elements are defined in a *Metadata Registry* and share the same rich descriptions as concepts in a terminology management system. The set of all metadata elements forms a *(meta)data dictionary* or a *(meta)data catalog*. When a metadata element is selected as a column header to define a dataset, additional properties are set to determine its *value domain*. Value domains are either *data types* (string, date, boolean, etc.) or terminologies. Value domains establish the constraints for the values occurring in the column of the metadata element.
Every metadata element is represented by a global, unique, persistent, and resolvable identifier serving as a reference.
Every dataset model is represented by a global, unique, persistent, and resolvable identifier serving as a reference.

Conceptual & Logical Model (Domain Ontology) Management: Following modern data and information architecture approaches, conceptual models support a reasonably grained division of the knowledge space in *data domains* and *subdomains*. In contrast to the table-like dataset models, conceptual models are purpose-driven *Ontologies* representing the *classes* and *properties* of a domain using a directed acyclic graph as a data structure. Domain ontologies can be used as a blueprint for knowledge graphs.
Every class or property is represented by a global, unique, persistent, and resolvable identifier serving as a reference.
Every conceptual or logical model is represented by a global, unique, persistent, and resolvable identifier serving as a reference.

Semantic Interoperability Hub – Capability Stack

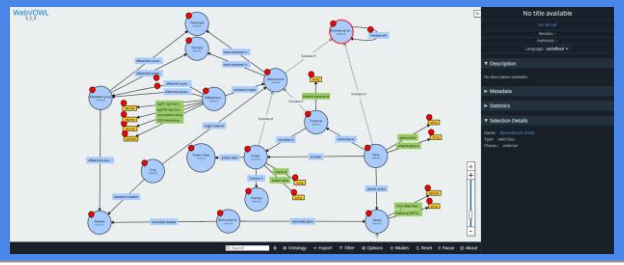
Data Management Value Chain - From Terminologies to a Unified Domain Model

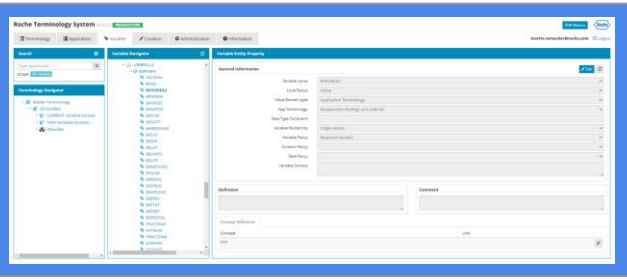
Terminology Managment (RTS)

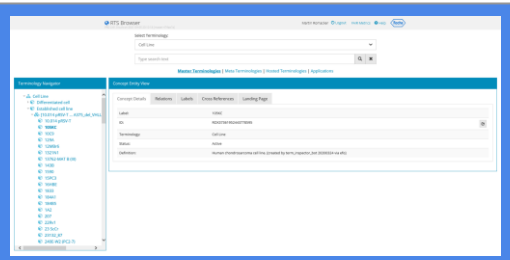
Dataset Model (MDR)

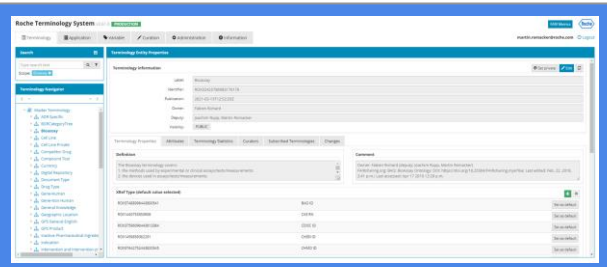
Conceptual Model

FAIR Unified Domain Model (KG)









Data and Information Architecture

Mapping Semanti capabilities - Fully FAIR Representation

Conceptual Model
(concepts and relationships)

Represented as

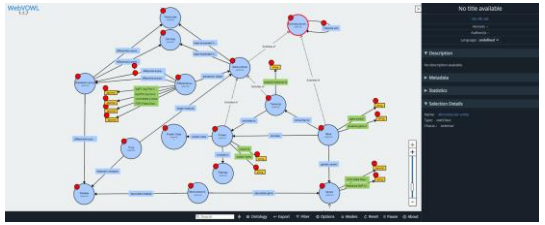
Logical Model
(data elements and relations)

Implemented as

Physical
(technical specification)

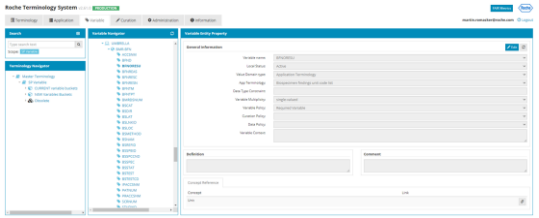
What data means

Defined concepts and relationships that are used in the real world / universe of discourse
Example: "Patient Identifier: unique value that identifies a single patient or subject of care"



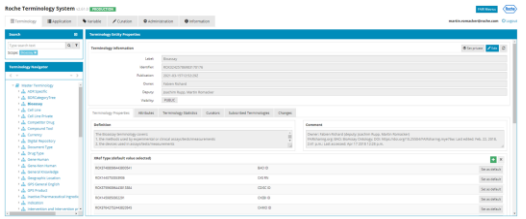
How data is modelled

Structures for how data is modelled, with data elements, groups, relations, cardinality, data types, etc.
Example: Patient.PatientID: 0..1: string



How data is implemented

An actual implementation in a physical system, e.g. a database or a field in a file
Examples: "Patient_ID: VARCHAR(25)"



Conceptual Modelling

Dataset Modelling

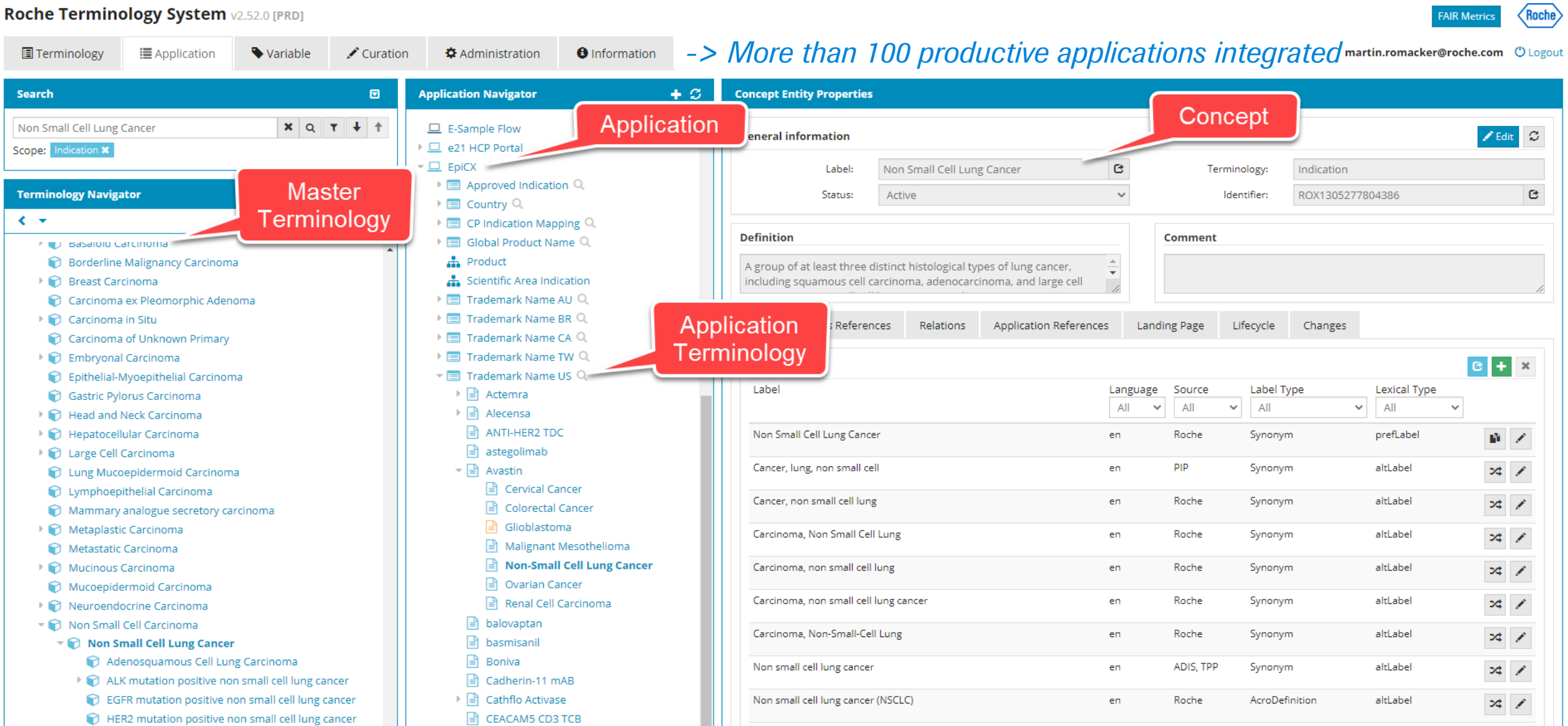
Terminology

Semantic Interoperability Hub

Terminology Management, Metadata, Dataset Model & Ontology

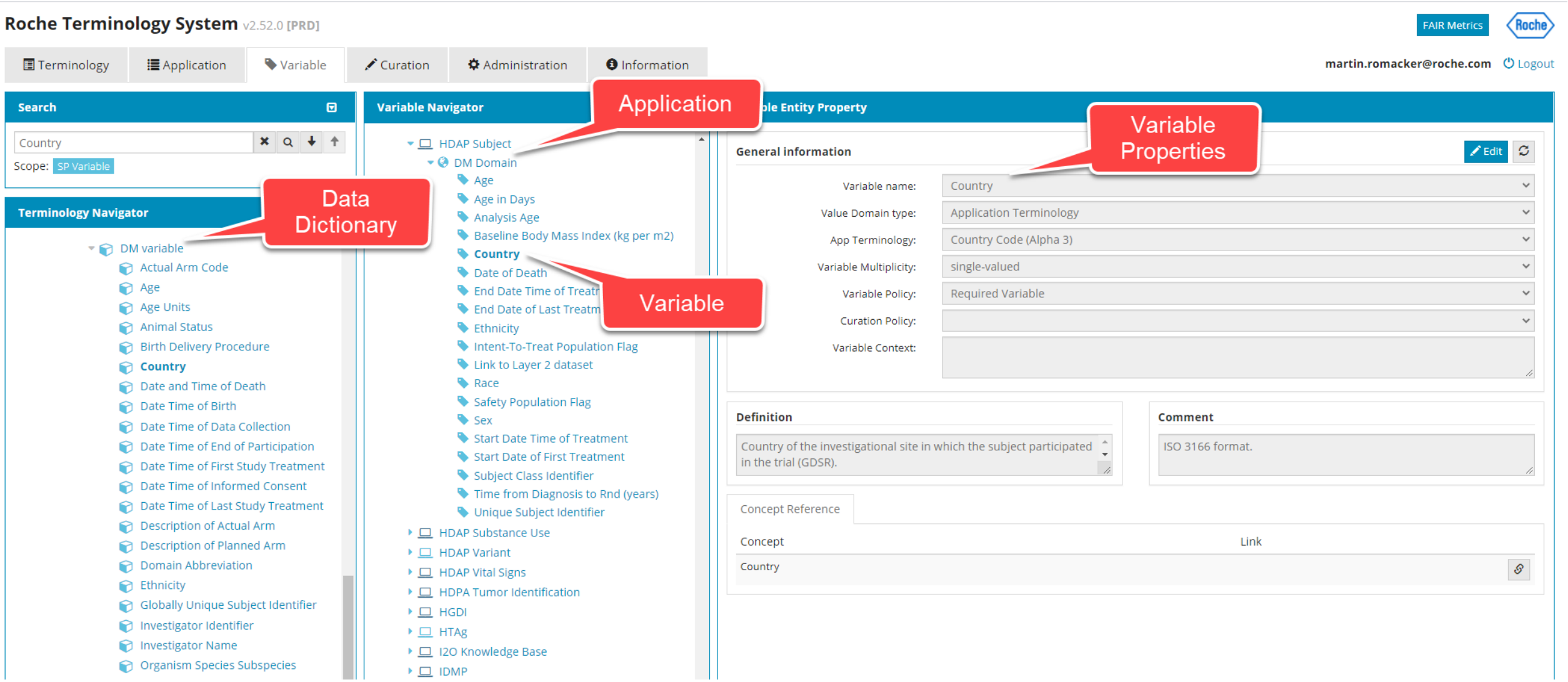
Reference Data Services for Data Management

Terminology Management - Contextualize Concepts (FAIR)



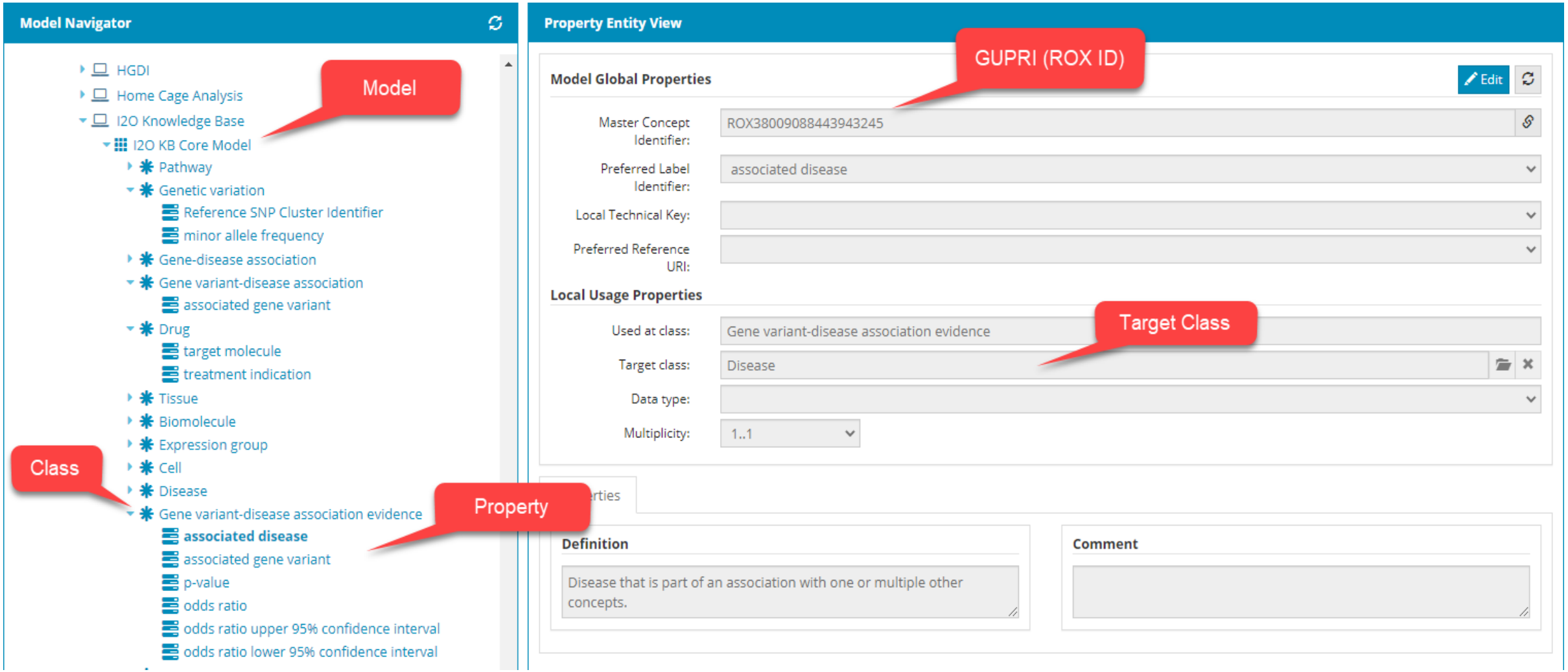
Reference Data Services for Data Management

Metadata Registry/ Dataset Models – Metadata Harmonization (FAIR)



Reference Data Services for Data Management

Conecptual Model - Purpose-build FAIR Ontologies



FAIR Linked Data

Conclusion

Conclusions

- High-Quality, standardized and linked data: foundation for digitilization & insight generation.
- FAIR data principles intrinsically tie Data Management to Semantic Technologies.
  (FAIR data is Linked Data by Design)
- Information Procurement based on FAIR supporting transformationless data integration.
- FAIR is primarily about the *HOW* and not only about the *THAT* (FAIR maturity indicators).
- Data Management Value Chain: new architectural approaches around data and information.
  Interoperability of terminologies, metadata, dataset models and ontologies is key.
- Data Management Strategy - urgency to build semantic capabilities at community level:
  open public-private semantic infrastructure of FAIR applications, services and data.
- It’s all about Semantics.

FAIR Linked Data

Acknowledgements

Acknowledgements

RTS Data Harmonization Service Team

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

Doing now what patients need next