Roche

Data FAIRification at scale- Semantic Core Capabilities & Implementation

Martin Romacker, Data and Information Architect
Scientific Solution Engineering and Architecture (S2EA)
Data & Analytics
Roche Innovation Center Basel
ENDORSE, 31st May 2022, Virtual

THE EUROPEAN DATA CONFERENCE ON REFERENCE DATA AND SEMANTICS
ENDORSE FOLLOW-UP EVENT
#ENDORSE2021

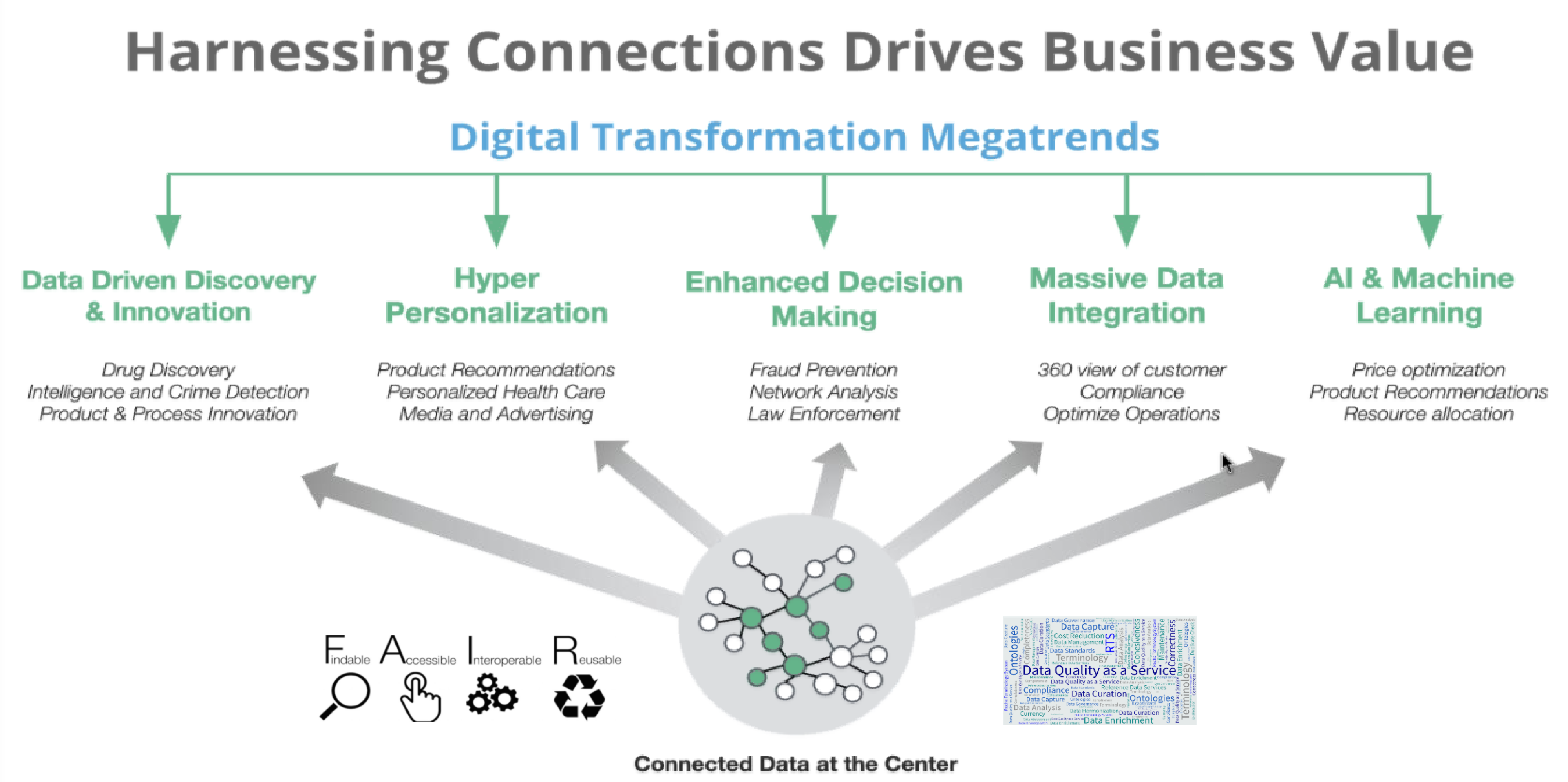
Roche *pRED*

Digital Transformation & Management of Data Assets

FAIR plus Q

Digital Transformation

Megatrends & Data Management Strategy



Harnessing Connections Drives Business Value

Digital Transformation Megatrends

Data Driven Discovery & Innovation

Drug Discovery
Intelligence and Crime Detection
Product & Process Innovation

Hyper Personalization

Product Recommendations
Personalized Health Care
Media and Advertising

Enhanced Decision Making

Fraud Prevention
Network Analysis
Law Enforcement

Massive Data Integration

360 view of customer
Compliance
Optimize Operations

AI & Machine Learning

Price optimization
Product Recommendations
Resource allocation

Findable Accessible Interoperable Reusable

Connected Data at the Center

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

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

Data as an Asset

Fundamental Change in Data/ Information Management needed

Data is the new **oil** of the digital economy

World Bank: 147 billion m^3 natural gas were flared in 2015.
Price in Europe about 0.5 € per m^3 (75 billion € value)
http://www.worldbank.org/en/news/press-release/2016/12/12/new-data-reveals-uptick-in-global-gas-flaring

? How many data assets vanish every year due to poor data management ?

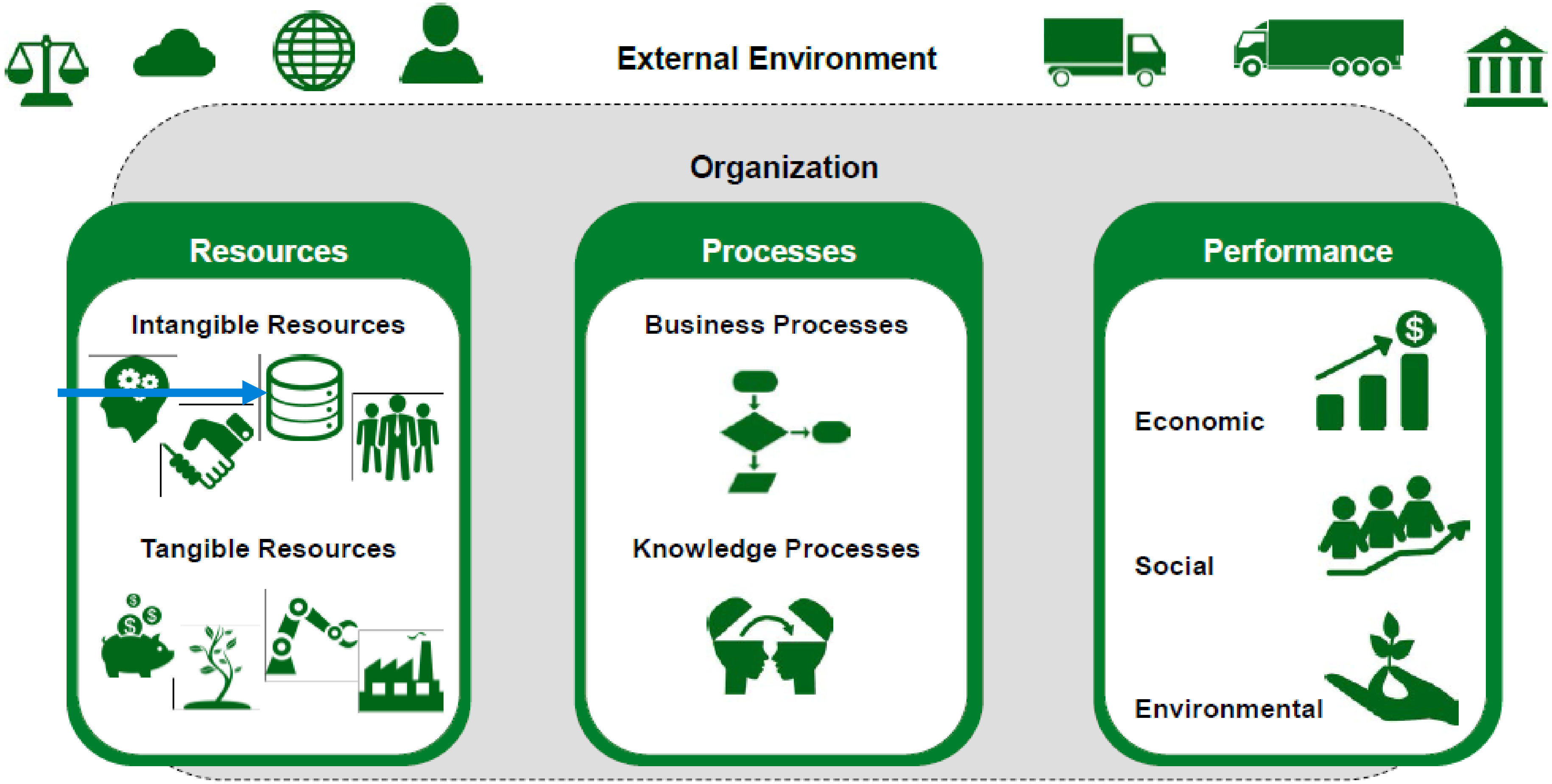
Pharma Industry: we consider data as an asset but we *do not* treat it as an asset

Management of Corporate Data Assets

Economic Perspective: Data should be in the Balance Sheet

DATA

- Cost
- Value
- Amortization



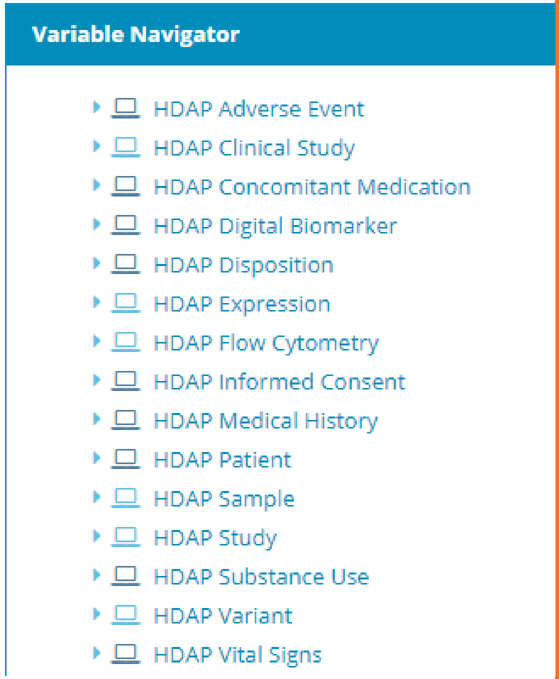
FAIR and Roche Data Commons

From application-centric to information-centric

Roche Data Commons (RDC) – Flipping the Coin in Data Mgt

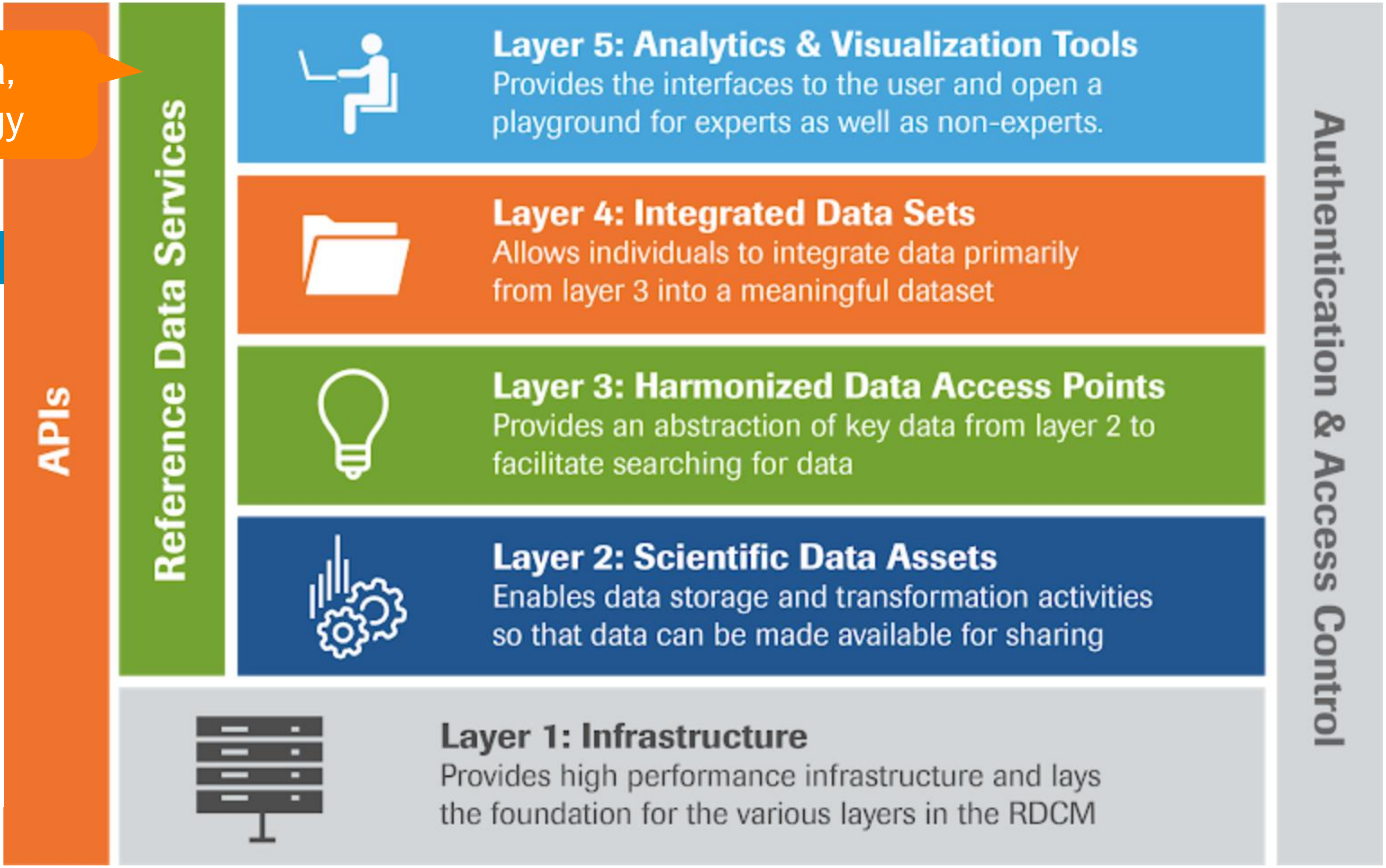
Moving from an application-centric to an information-centric organization

Terminology, Metadata, Dataset Model, Ontology



Variable Navigator

- HDAP Adverse Event
- HDAP Clinical Study
- HDAP Concomitant Medication
- HDAP Digital Biomarker
- HDAP Disposition
- HDAP Expression
- HDAP Flow Cytometry
- HDAP Informed Consent
- HDAP Medical History
- HDAP Patient
- HDAP Sample
- HDAP Study
- HDAP Substance Use
- HDAP Variant
- HDAP Vital Signs



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

Roche Data Commons (RDC)

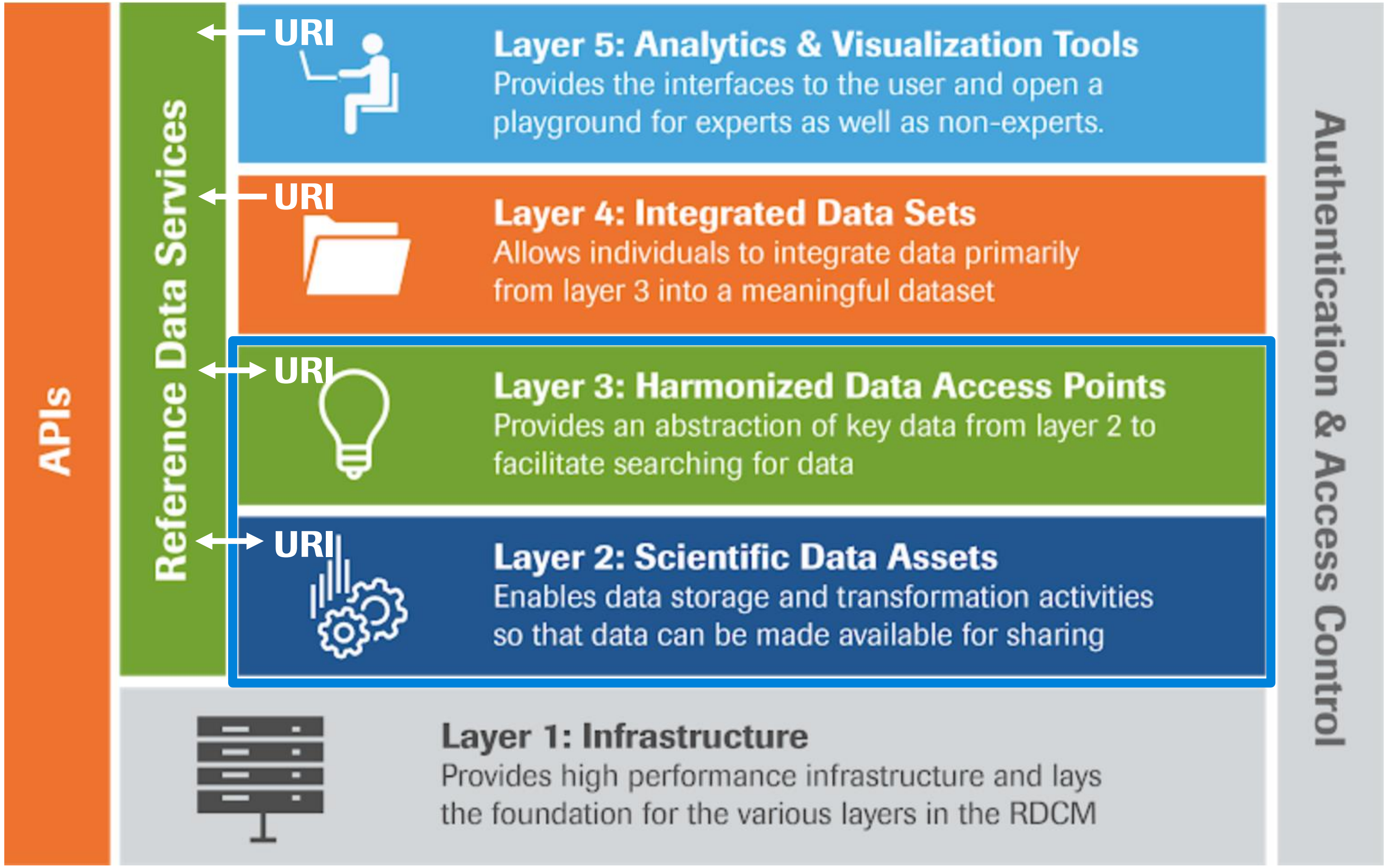
Data FAIRification – Everything is a Resource (URI)!

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



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

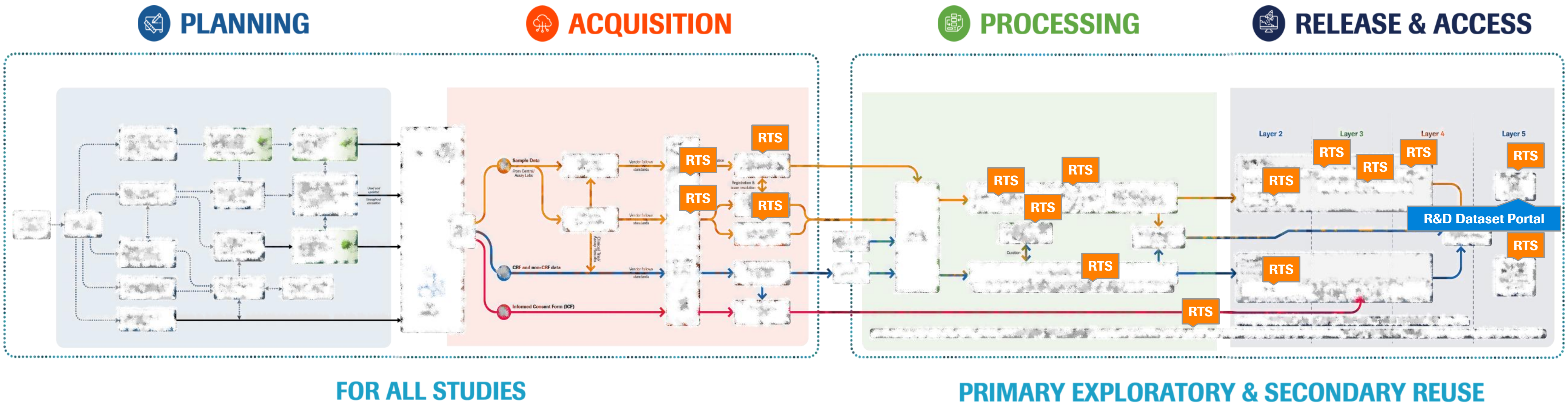
Layer 3: Harmonized Data Access Points
Provides an abstraction of key data from layer 2 to facilitate searching for data

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

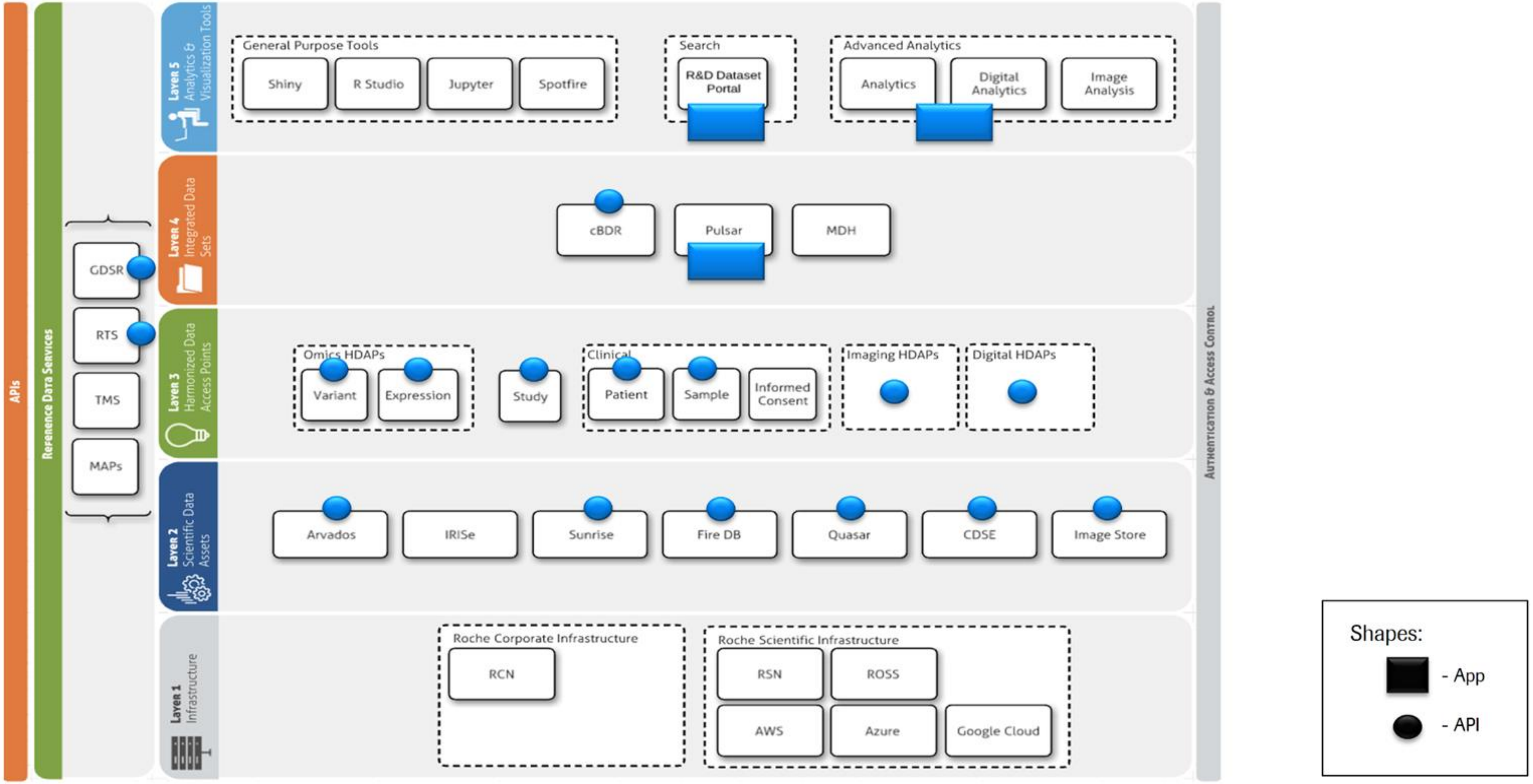
EDIS E2E Engine

RTS Integration (born FAIR)



Roche Data Commons

Fully Integrated Transformationless FAIR Architecture (FAIR by Design)

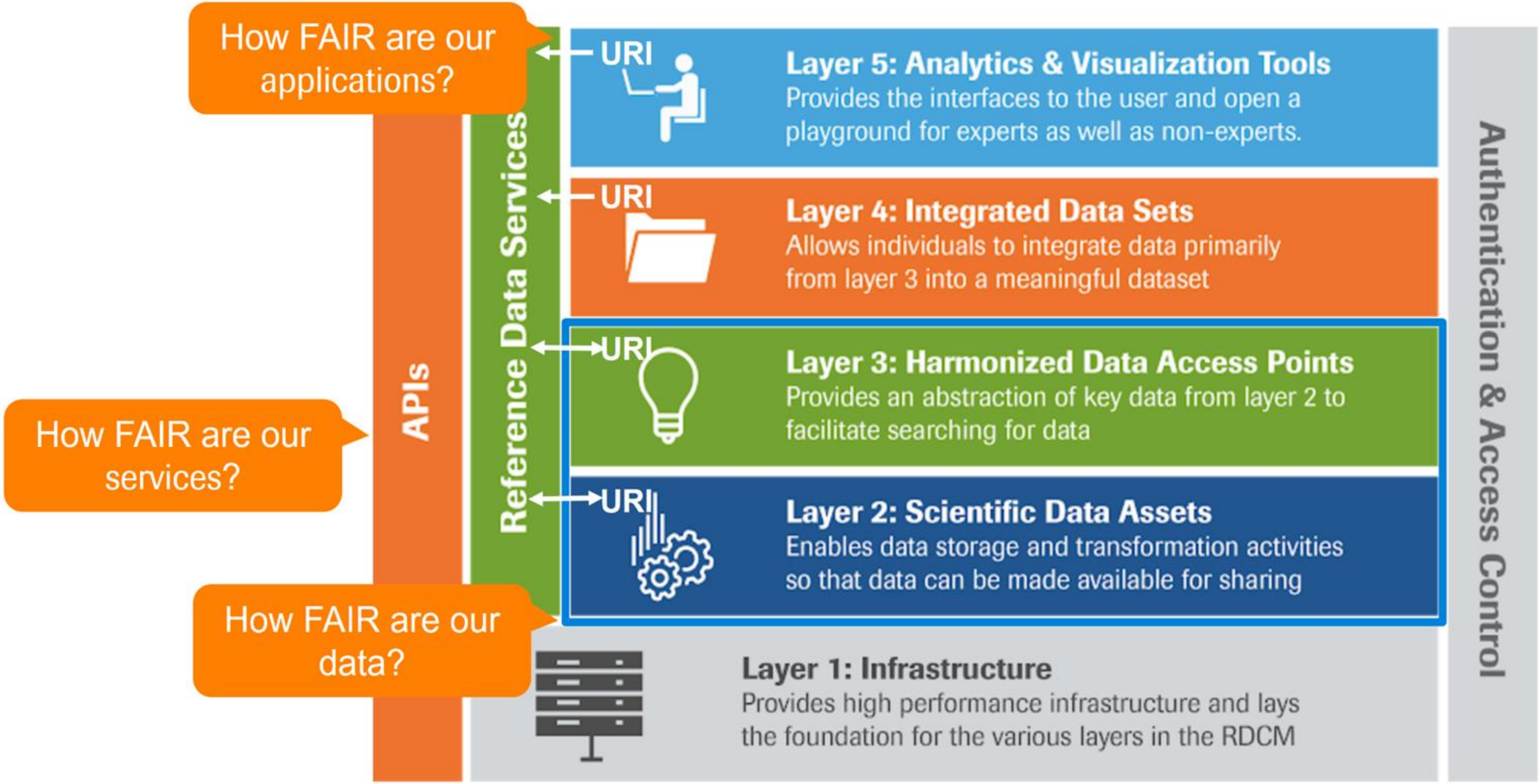


FAIRification at Scale

Scientific Interoperability Hub – Key Capabilities

Roche Data Commons (RDC)

Semantic Infrastructure of FAIR Data, Services and Applications



How FAIR are our applications?

How FAIR are our services?

How FAIR are our data?

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

FAIR scientific data management

FAIR guiding principles

F A I R

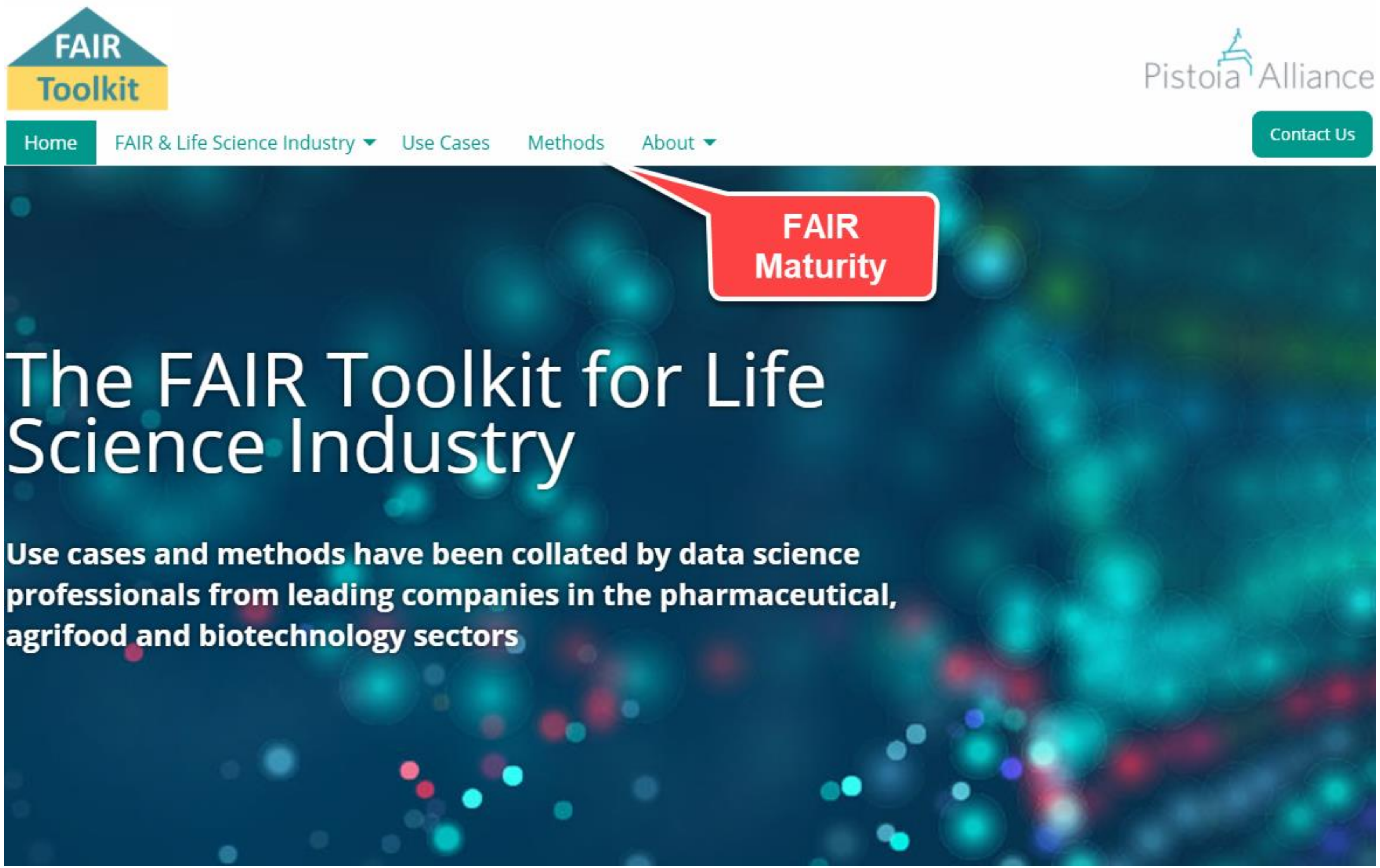
Ability for scientist/data consumer to find, access and understand the data
(without the presence of the data owner)

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

by Olivier Roche (pREDi)

FAIR Assessment

Pistoia Alliance



The FAIR Toolkit for Life Science Industry

Use cases and methods have been collated by data science professionals from leading companies in the pharmaceutical, agrifood and biotechnology sectors

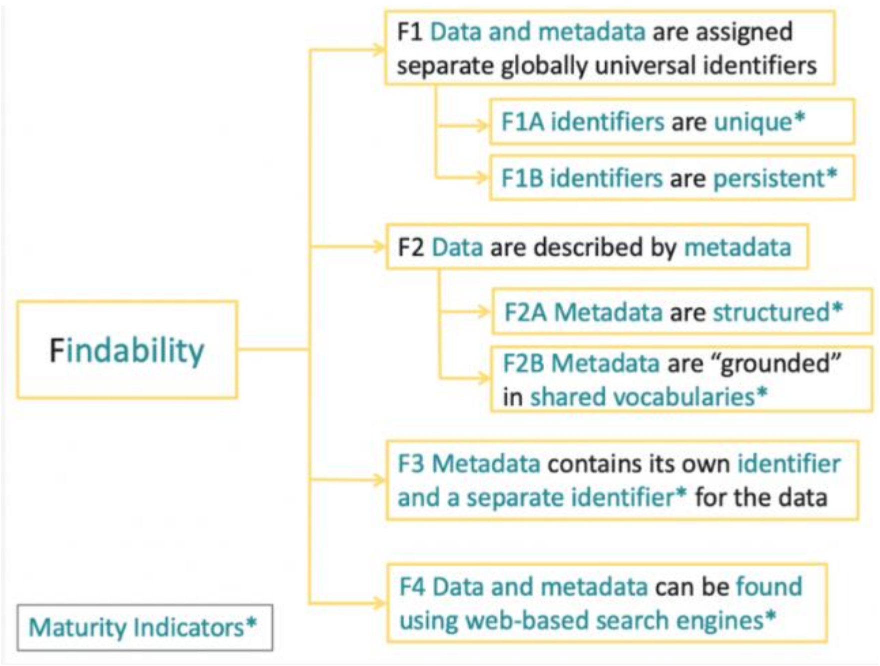
FAIR Maturity

FAIR Toolkit

Implementation for FAIR Data Principles in Life Science R&D

Maturity Indicators: FAIR Metrics

https://fairtoolkit.pistoiaalliance.org/



Findability

F1 Data and metadata are assigned separate globally universal identifiers

F1A identifiers are unique*

F1B identifiers are persistent*

F2 Data are described by metadata

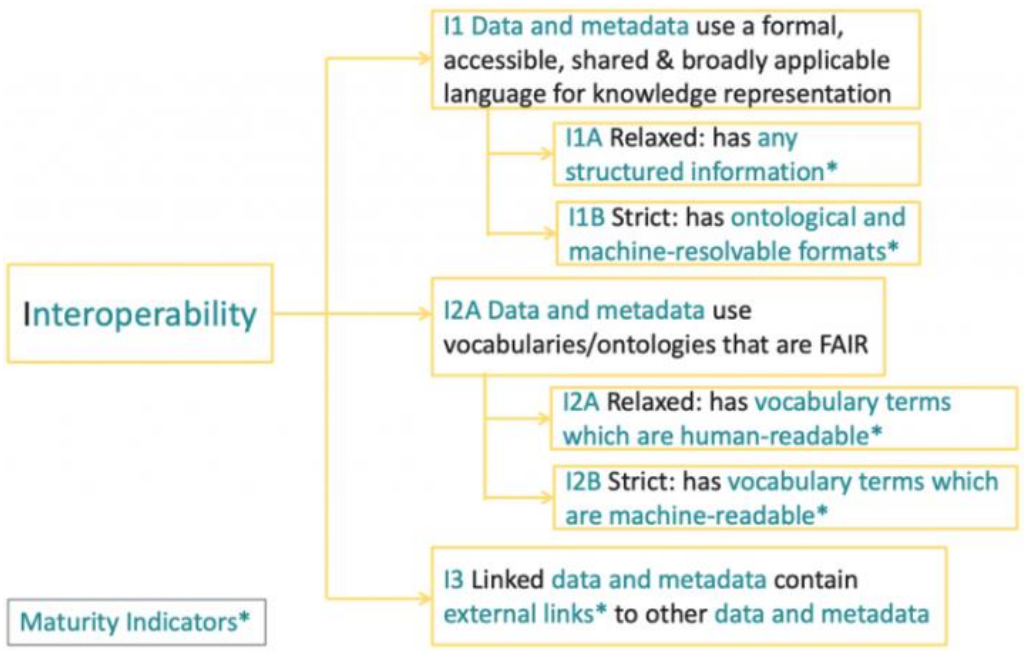
F2A Metadata are structured*

F2B Metadata are "grounded" in shared vocabularies*

F3 Metadata contains its own identifier and a separate identifier* for the data

F4 Data and metadata can be found using web-based search engines*

Maturity Indicators*



Interoperability

I1 Data and metadata use a formal, accessible, shared & broadly applicable language for knowledge representation

I1A Relaxed: has any structured information*

I1B Strict: has ontological and machine-resolvable formats*

I2A Data and metadata use vocabularies/ontologies that are FAIR

I2A Relaxed: has vocabulary terms which are human-readable*

I2B Strict: has vocabulary terms which are machine-readable*

I3 Linked data and metadata contain external links* to other data and metadata

Maturity Indicators*

FAIR is about data *and* metadata

FAIR Playbook for IT Professionals

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

b|

FAIR Architecture Playbook

FAIR by Design

A Primer for IT Architects,
Business Analysts and Software Engineers

IT as key enabler.

Data Managers should not care about FAIR.

Digitalization building on a FAIR architecture

Digital Objects and Data FAIRification

Definition 1: A *Digital Object* is any kind of data that exists in a digital modality.
A digital representation of a physical object or a process is also a *Digital Object*.

Definition 2a: A *First-class Digital Object* is a Digital Object born in a digital modality (born-digital).

Definition 2b: A *Digital Twin* is a Digital Object that represents a Non-Digital Object.

Born-digital does not necessarily mean born FAIR.

(FAIR Maturity Indicators)

FAIR is above all about the *how* not only about the *that*.

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:

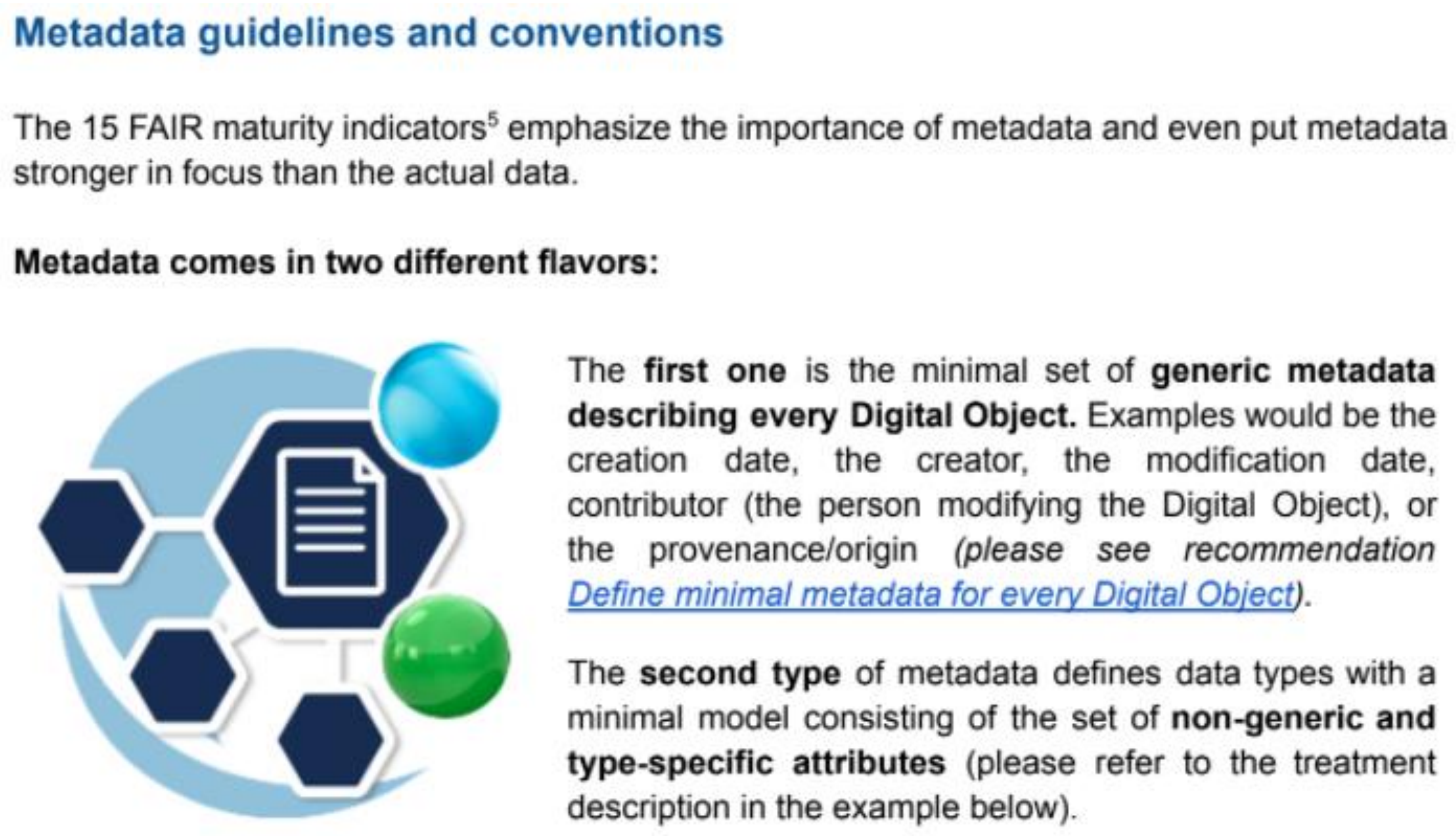
FAIR applications, services, and data require governance, policies, and infrastructure to manage the identifiers space at the global scale.

***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].

Metadata Management

Generic and type specific metadata



Metadata guidelines and conventions

The 15 FAIR maturity indicators[5] emphasize the importance of metadata and even put metadata stronger in focus than the actual data.

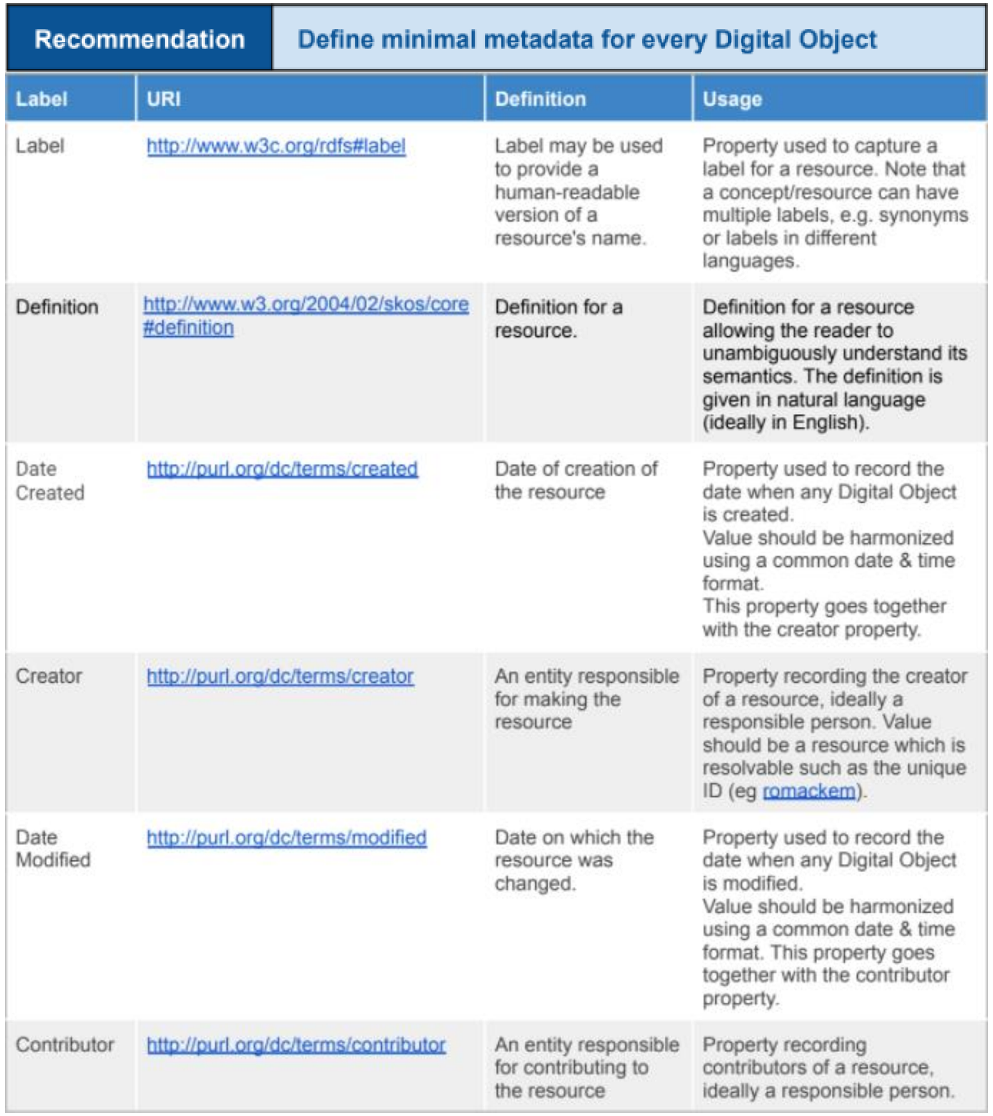
Metadata comes in two different flavors:

The **first one** is the minimal set of **generic metadata describing every Digital Object.** Examples would be the creation date, the creator, the modification date, contributor (the person modifying the Digital Object), or the provenance/origin *(please see recommendation [Define minimal metadata for every Digital Object](#))*.

The **second type** of metadata defines data types with a minimal model consisting of the set of **non-generic and type-specific attributes** (please refer to the treatment description in the example below).

Metadata – Recommendation

Minimal Metadata & Conceptual Models



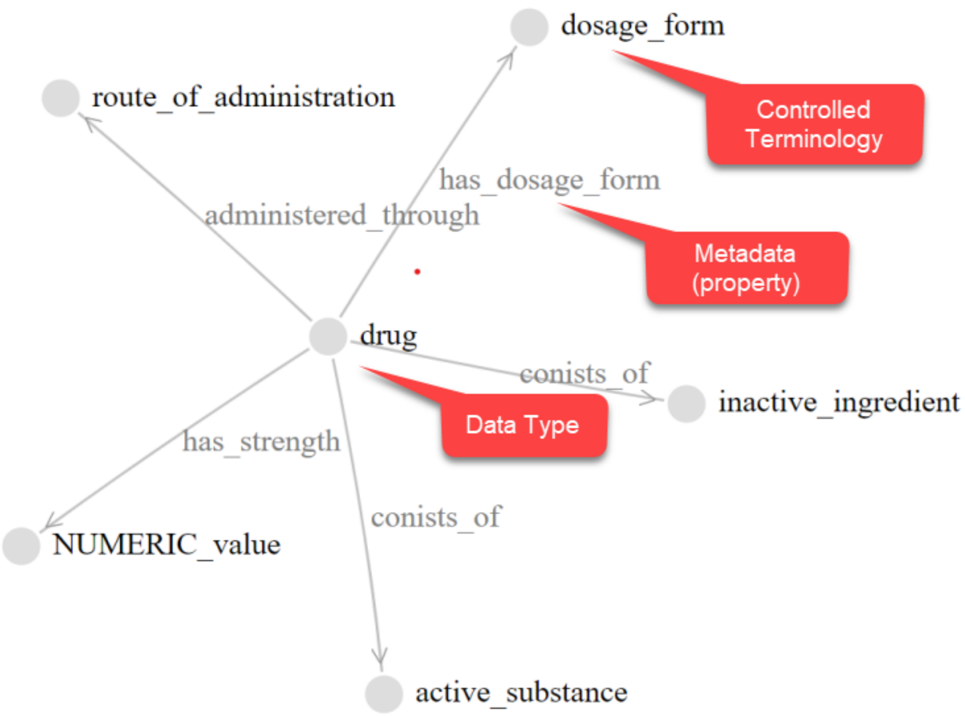
Recommendation	Define minimal metadata for every Digital Object

Label	URI	Definition	Usage
Label	http://www.w3c.org/rdfs#label	Label may be used to provide a human-readable version of a resource's name.	Property used to capture a label for a resource. Note that a concept/resource can have multiple labels, e.g. synonyms or labels in different languages.
Definition	http://www.w3.org/2004/02/skos/core#definition	Definition for a resource.	Definition for a resource allowing the reader to unambiguously understand its semantics. The definition is given in natural language (ideally in English).
Date Created	http://purl.org/dc/terms/created	Date of creation of the resource	Property used to record the date when any Digital Object is created. Value should be harmonized using a common date & time format. This property goes together with the creator property.
Creator	http://purl.org/dc/terms/creator	An entity responsible for making the resource	Property recording the creator of a resource, ideally a responsible person. Value should be a resource which is resolvable such as the unique ID (eg romackem).
Date Modified	http://purl.org/dc/terms/modified	Date on which the resource was changed.	Property used to record the date when any Digital Object is modified. Value should be harmonized using a common date & time format. This property goes together with the contributor property.
Contributor	http://purl.org/dc/terms/contributor	An entity responsible for contributing to the resource	Property recording contributors of a resource, ideally a responsible person.

Recommendation	Define complex types with properties and classes

Example (taken from RxNorm): *oseltamivir 6 mg/ml oral suspension*

- oseltamivir - *active substance*
- 6mg/ml - *strength* (6 numeric value and mg/ml *unit*)
- oral - *route of administration*
- suspension - *dosage form*



Controlled Terminology

Metadata (property)

Data Type

FAIRification at Scale: Capability Stack

From Terminologies to Domain Models

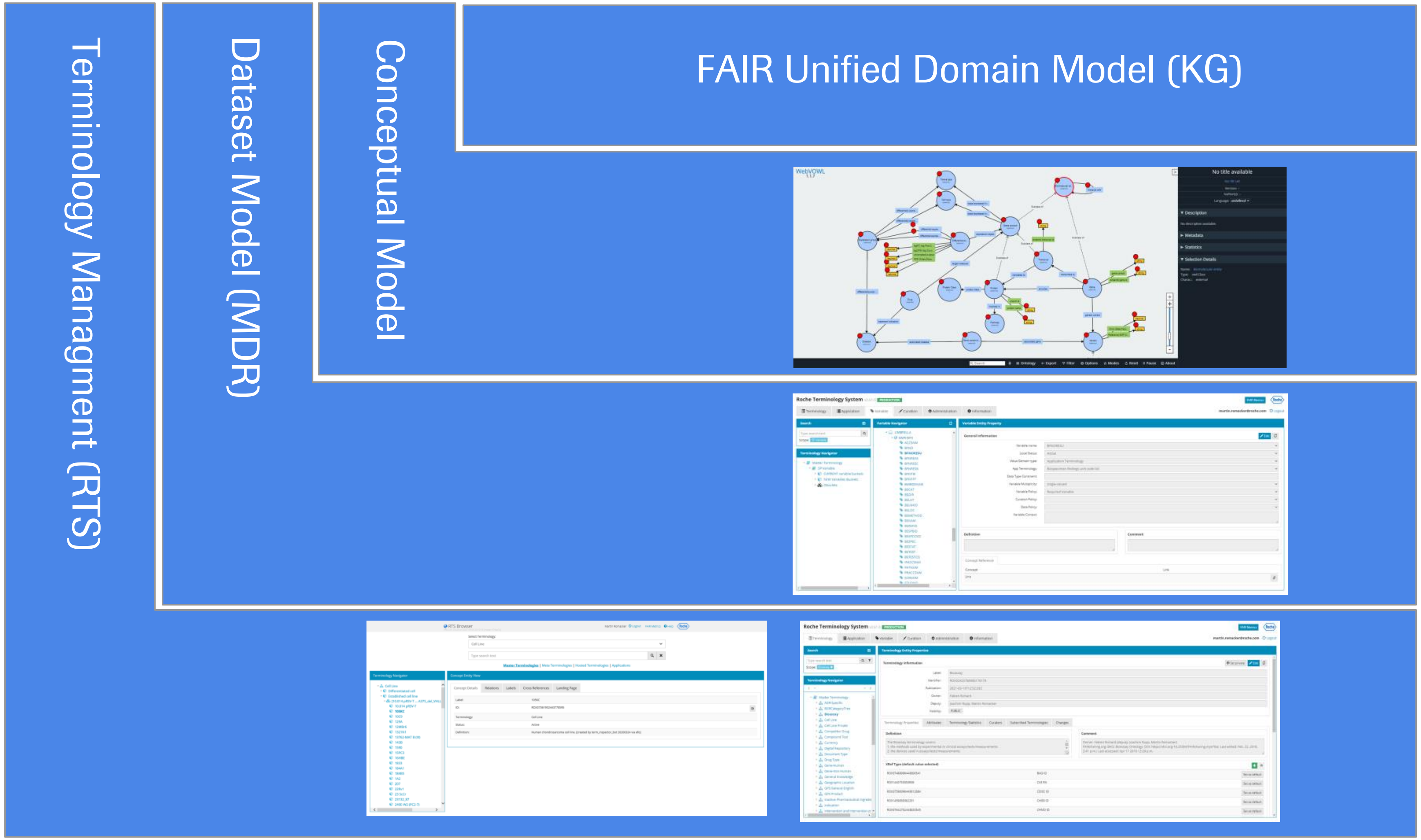
Terminology Management: The concepts used in our scientific and technical domains are properly defined, typed and organized in a *Terminology Management System*. Each *concept* is given an unambiguous, complete, *preferred label* and a *textual definition*. The concept is complemented by a rich *synonym 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.

Scientific Interoperability Hub – Capability Stack

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



Terminology Managment (RTS)

Dataset Model (MDR)

Conceptual Model

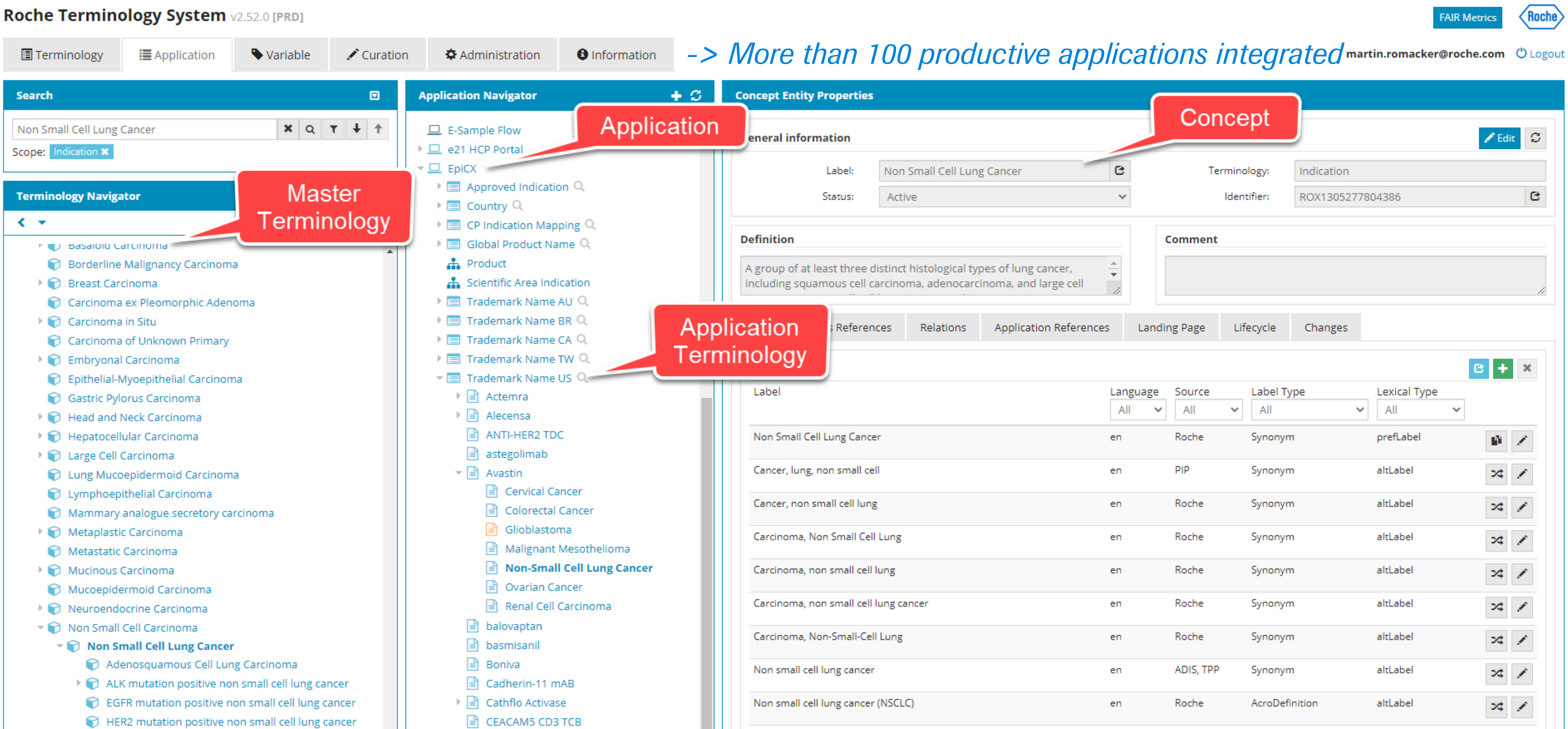
FAIR Unified Domain Model (KG)

Scientific Interoperability Hub

Terminology Management, Metadata, Dataset Model & Ontology

Reference Data Services for Data Management

Terminology Management - Contextualize Concepts (FAIR)



-> More than 100 productive applications integrated

Application

Concept

Master Terminology

Application Terminology

Roche Terminology System v2.52.0 [PRD]

FAIR Metrics

Terminology | Application | Variable | Curation | Administration | Information

martin.romacker@roche.com Logout

Search

Non Small Cell Lung Cancer

Scope: Indication ✖

Terminology Navigator

Basaloid Carcinoma
Borderline Malignancy Carcinoma
Breast Carcinoma
Carcinoma ex Pleomorphic Adenoma
Carcinoma in Situ
Carcinoma of Unknown Primary
Embryonal Carcinoma
Epithelial-Myoepithelial Carcinoma
Gastric Pylorus Carcinoma
Head and Neck Carcinoma
Hepatocellular Carcinoma
Large Cell Carcinoma
Lung Mucoepidermoid Carcinoma
Lymphoepithelial Carcinoma
Mammary analogue secretory carcinoma
Metaplastic Carcinoma
Metastatic Carcinoma
Mucinous Carcinoma
Mucoepidermoid Carcinoma
Neuroendocrine Carcinoma
Non Small Cell Carcinoma
  Non Small Cell Lung Cancer
    Adenosquamous Cell Lung Carcinoma
    ALK mutation positive non small cell lung cancer
    EGFR mutation positive non small cell lung cancer
    HER2 mutation positive non small cell lung cancer

Application Navigator

E-Sample Flow
e21 HCP Portal
EpiCX
  Approved Indication
  Country
  CP Indication Mapping
  Global Product Name
  Product
  Scientific Area Indication
  Trademark Name AU
  Trademark Name BR
  Trademark Name CA
  Trademark Name TW
  Trademark Name US
    Actemra
    Alecensa
    ANTI-HER2 TDC
    astegolimab
    Avastin
      Cervical Cancer
      Colorectal Cancer
      Glioblastoma
      Malignant Mesothelioma
      Non-Small Cell Lung Cancer
      Ovarian Cancer
      Renal Cell Carcinoma
    balovaptan
    basmisanil
    Boniva
    Cadherin-11 mAB
    Cathflo Activase
    CEACAM5 CD3 TCB

Concept Entity Properties

eneral information Edit

Label: Non Small Cell Lung Cancer
Status: Active
Terminology: Indication
Identifier: ROX1305277804386

Definition

A group of at least three distinct histological types of lung cancer, including squamous cell carcinoma, adenocarcinoma, and large cell

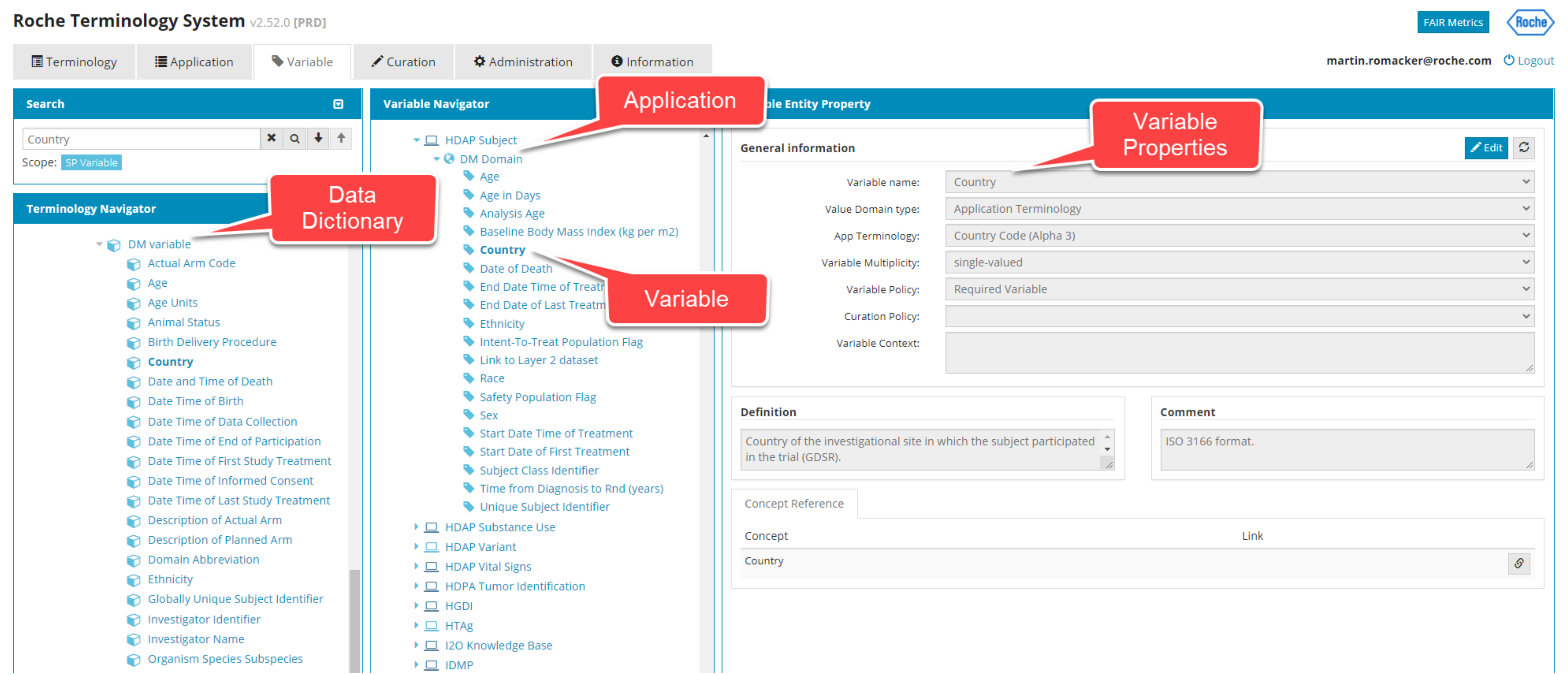
Comment

References | Relations | Application References | Landing Page | Lifecycle | Changes

Label	Language	Source	Label Type	Lexical Type
	All	All	All	All
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)



Application

Data Dictionary

Variable

Variable Properties

Roche Terminology System v2.52.0 [PRD]

FAIR Metrics

Terminology | Application | Variable | Curation | Administration | Information

martin.romacker@roche.com Logout

Search

Country

Scope: SP Variable

Terminology Navigator

- DM variable
  - Actual Arm Code
  - Age
  - Age Units
  - Animal Status
  - Birth Delivery Procedure
  - **Country**
  - Date and Time of Death
  - Date Time of Birth
  - Date Time of Data Collection
  - Date Time of End of Participation
  - Date Time of First Study Treatment
  - Date Time of Informed Consent
  - Date Time of Last Study Treatment
  - Description of Actual Arm
  - Description of Planned Arm
  - Domain Abbreviation
  - Ethnicity
  - Globally Unique Subject Identifier
  - Investigator Identifier
  - Investigator Name
  - Organism Species Subspecies

Variable Navigator

- HDAP Subject
  - DM Domain
    - Age
    - Age in Days
    - Analysis Age
    - Baseline Body Mass Index (kg per m2)
    - **Country**
    - Date of Death
    - End Date Time of Treatr
    - End Date of Last Treatm
    - Ethnicity
    - Intent-To-Treat Population Flag
    - Link to Layer 2 dataset
    - Race
    - Safety Population Flag
    - Sex
    - Start Date Time of Treatment
    - Start Date of First Treatment
    - Subject Class Identifier
    - Time from Diagnosis to Rnd (years)
    - Unique Subject Identifier
- HDAP Substance Use
- HDAP Variant
- HDAP Vital Signs
- HDPA Tumor Identification
- HGDI
- HTAg
- I2O Knowledge Base
- IDMP

ble Entity Property

General information Edit

Variable name:	Country
Value Domain type:	Application Terminology
App Terminology:	Country Code (Alpha 3)
Variable Multiplicity:	single-valued
Variable Policy:	Required Variable
Curation Policy:	
Variable Context:	

Definition

Country of the investigational site in which the subject participated in the trial (GDSR).

Comment

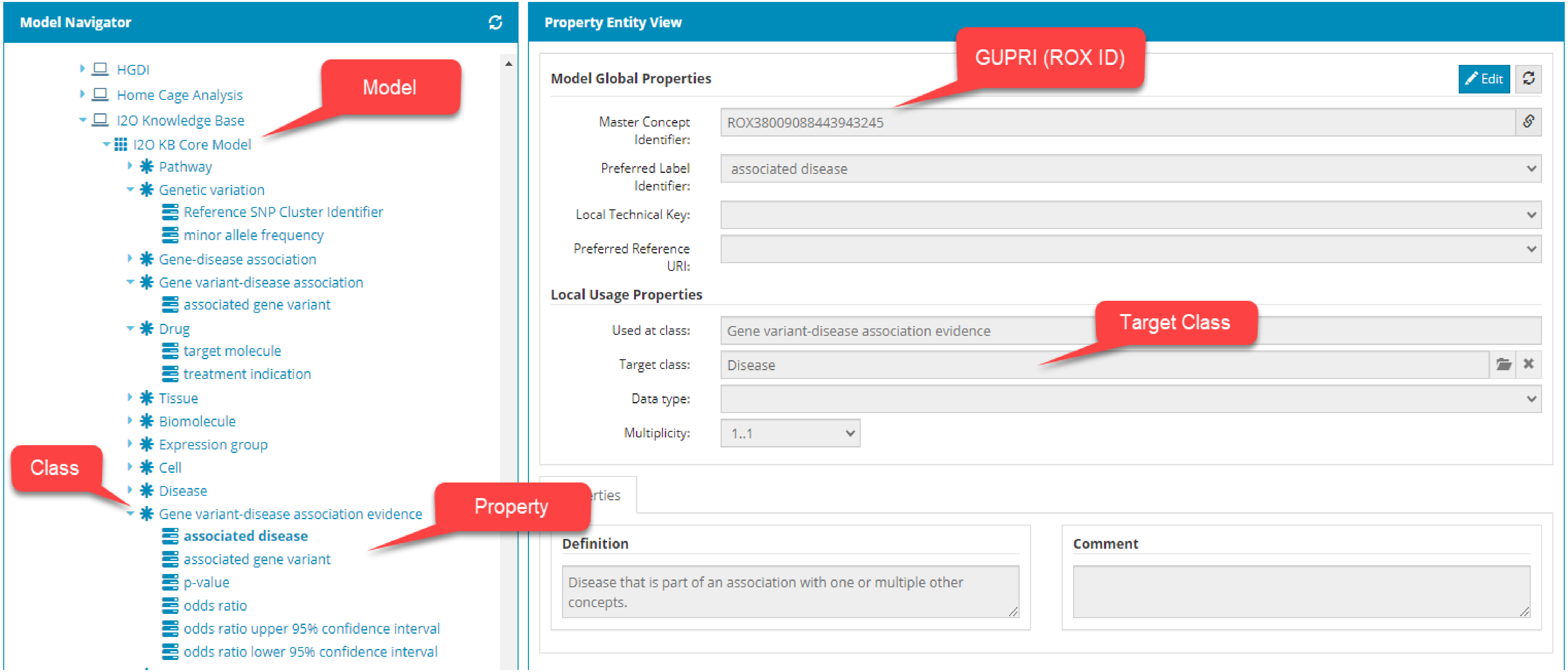
ISO 3166 format.

Concept Reference

Concept	Link
Country	

Reference Data Services for Data Management

Conecptual Model - Purpose-build FAIR Ontologies



Model Global Properties

Master Concept Identifier: ROX38009088443943245

Preferred Label Identifier: associated disease

Local Technical Key:

Preferred Reference URI:

Local Usage Properties

Used at class: Gene variant-disease association evidence

Target class: Disease

Data type:

Multiplicity: 1..1

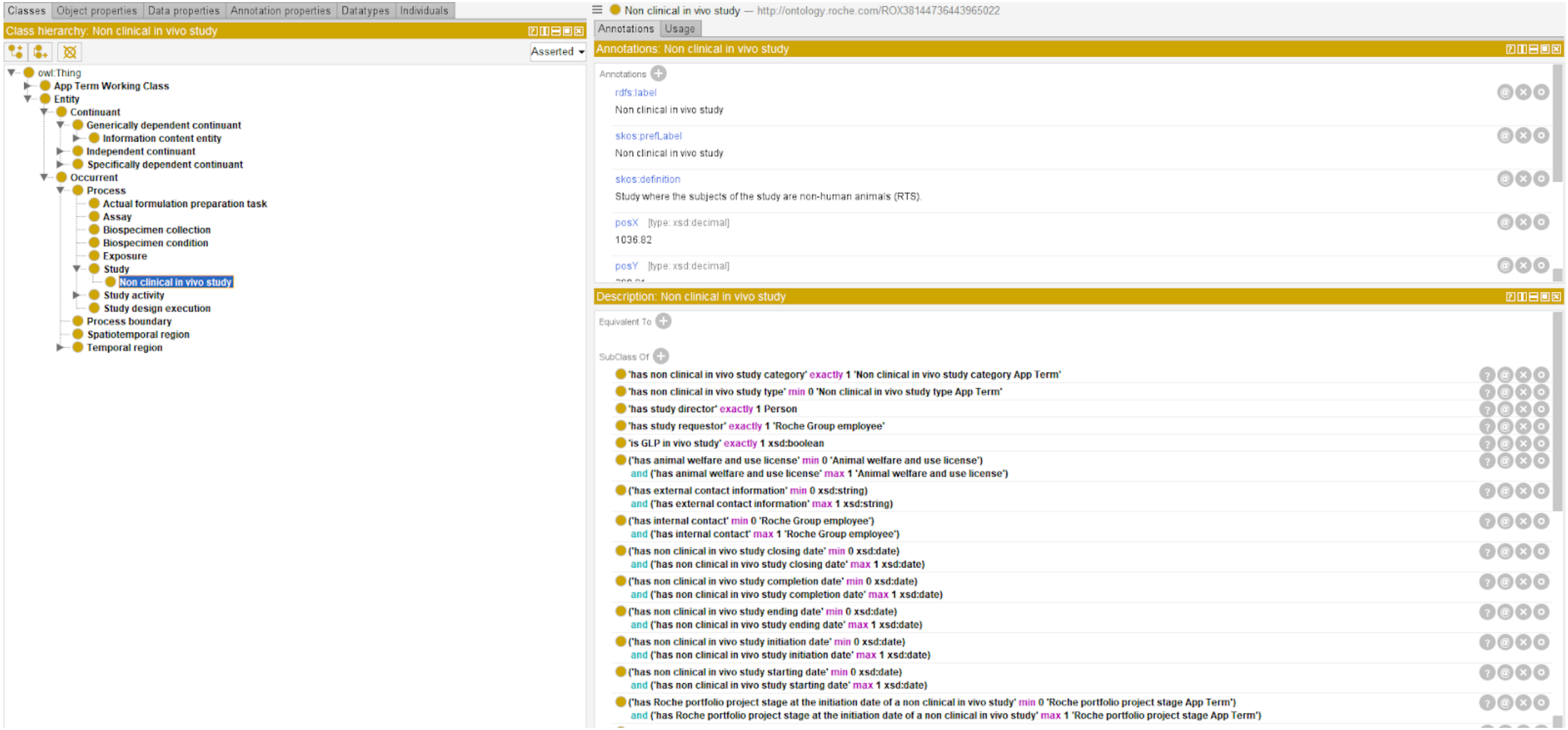
Definition

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

Comment

Reference Data Services for Data Management

Native support of RDF/ OWL



Class hierarchy: Non clinical in vivo study

- owl:Thing
  - App Term Working Class
  - Entity
    - Continuant
      - Generically dependent continuant
        - Information content entity
      - Independent continuant
      - Specifically dependent continuant
    - Occurrent
      - Process
        - Actual formulation preparation task
        - Assay
        - Biospecimen collection
        - Biospecimen condition
        - Exposure
        - Study
          - Non clinical in vivo study
        - Study activity
        - Study design execution
      - Process boundary
      - Spatiotemporal region
      - Temporal region

Non clinical in vivo study — http://ontology.roche.com/ROX38144736443965022

Annotations: Non clinical in vivo study

Annotations

rdfs:label
Non clinical in vivo study

skos:prefLabel
Non clinical in vivo study

skos:definition
Study where the subjects of the study are non-human animals (RTS).

posX [type: xsd:decimal]
1036.82

posY [type: xsd:decimal]

Description: Non clinical in vivo study

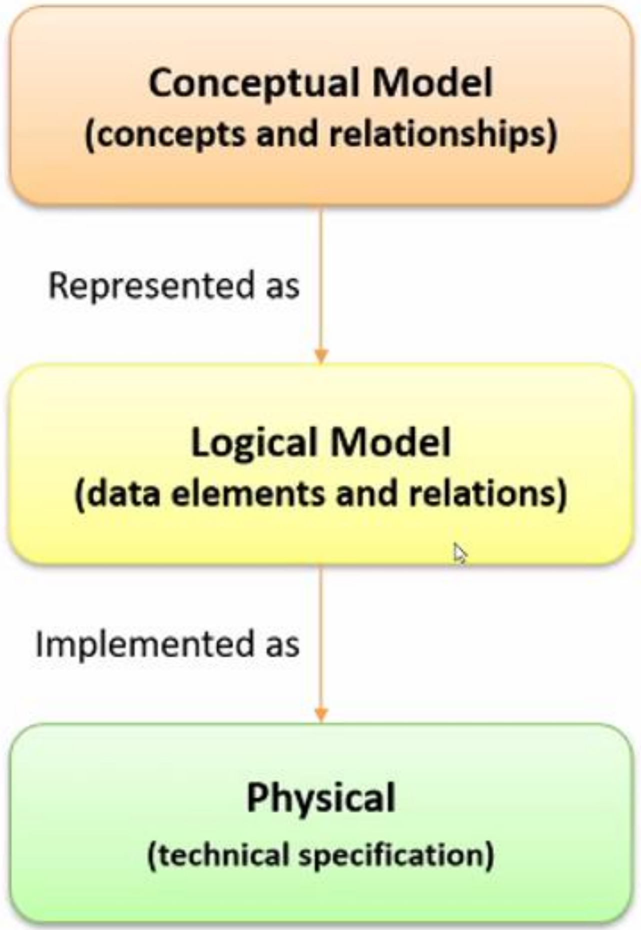
Equivalent To

SubClass Of

- 'has non clinical in vivo study category' exactly 1 'Non clinical in vivo study category App Term'
- 'has non clinical in vivo study type' min 0 'Non clinical in vivo study type App Term'
- 'has study director' exactly 1 Person
- 'has study requestor' exactly 1 'Roche Group employee'
- 'is GLP in vivo study' exactly 1 xsd:boolean
- ('has animal welfare and use license' min 0 'Animal welfare and use license')
  and ('has animal welfare and use license' max 1 'Animal welfare and use license')
- ('has external contact information' min 0 xsd:string)
  and ('has external contact information' max 1 xsd:string)
- ('has internal contact' min 0 'Roche Group employee')
  and ('has internal contact' max 1 'Roche Group employee')
- ('has non clinical in vivo study closing date' min 0 xsd:date)
  and ('has non clinical in vivo study closing date' max 1 xsd:date)
- ('has non clinical in vivo study completion date' min 0 xsd:date)
  and ('has non clinical in vivo study completion date' max 1 xsd:date)
- ('has non clinical in vivo study ending date' min 0 xsd:date)
  and ('has non clinical in vivo study ending date' max 1 xsd:date)
- ('has non clinical in vivo study initiation date' min 0 xsd:date)
  and ('has non clinical in vivo study initiation date' max 1 xsd:date)
- ('has non clinical in vivo study starting date' min 0 xsd:date)
  and ('has non clinical in vivo study starting date' max 1 xsd:date)
- ('has Roche portfolio project stage at the initiation date of a non clinical in vivo study' min 0 'Roche portfolio project stage App Term')
  and ('has Roche portfolio project stage at the initiation date of a non clinical in vivo study' max 1 'Roche portfolio project stage App Term')

Data and Information Architecture

Mapping RTS 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

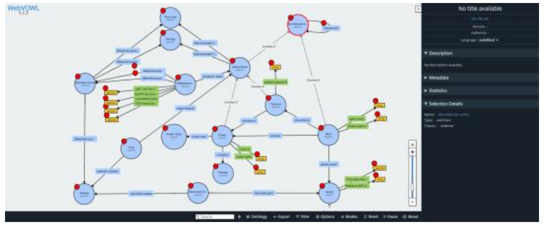
Defined concepts and relationships that are used in the real world / universe of discourse
Example: “Patient Identifier: unique value that identifies a single patient or subject of care”

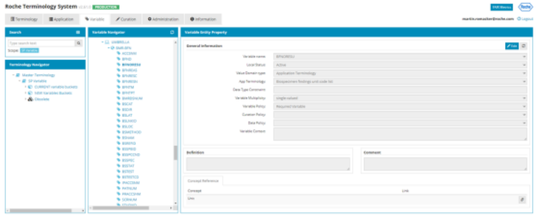
How data is modelled

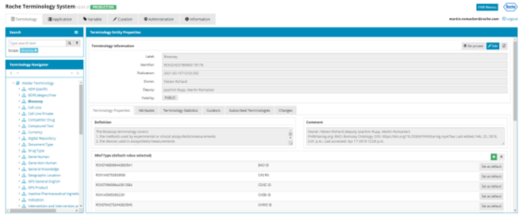
Structures for how data is modelled, with data elements, groups, relations, cardinality, data types, etc.
Example: Patient.PatientID: 0..1: string

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)”







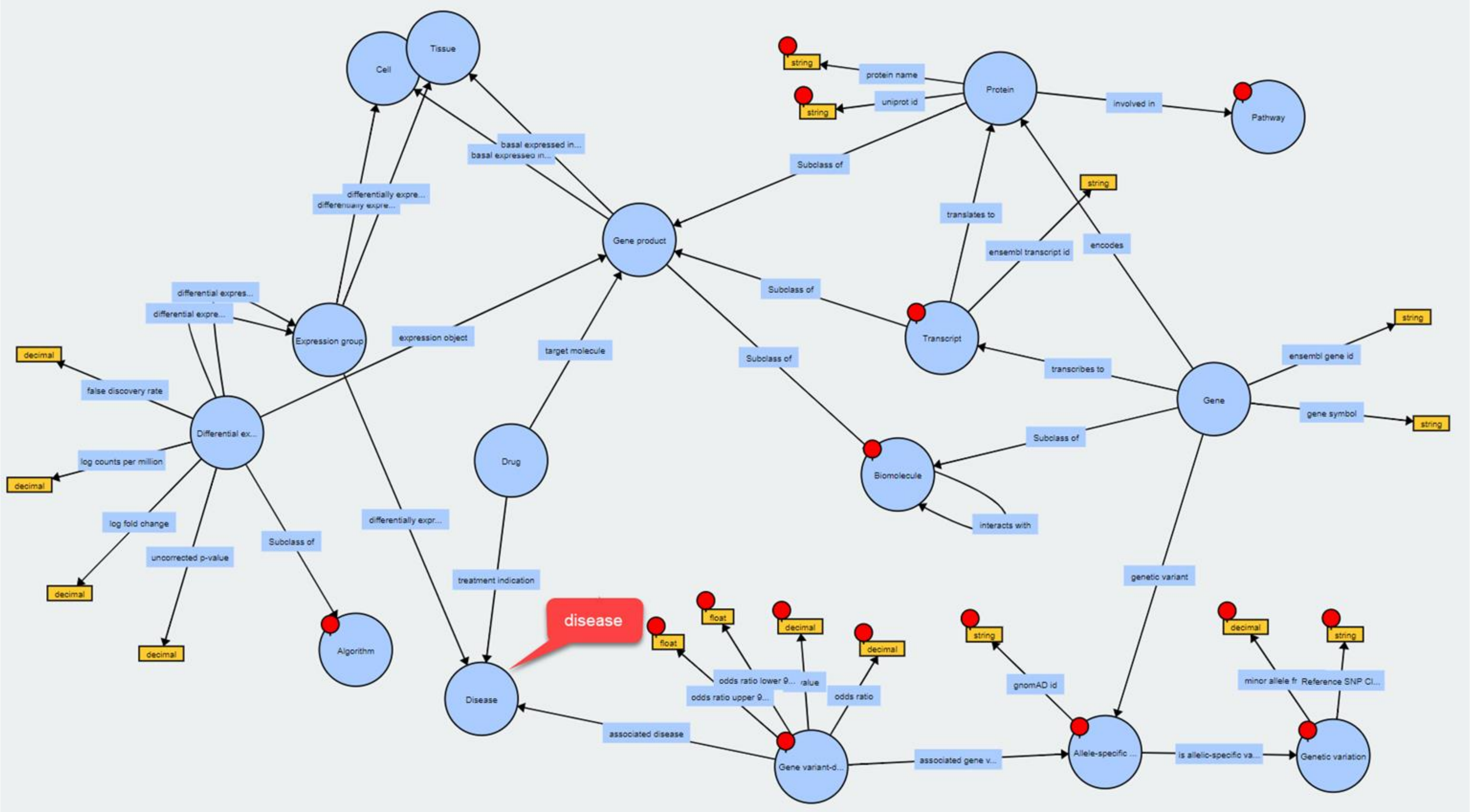
Conceptual Modelling

Dataset Modelling

Terminology

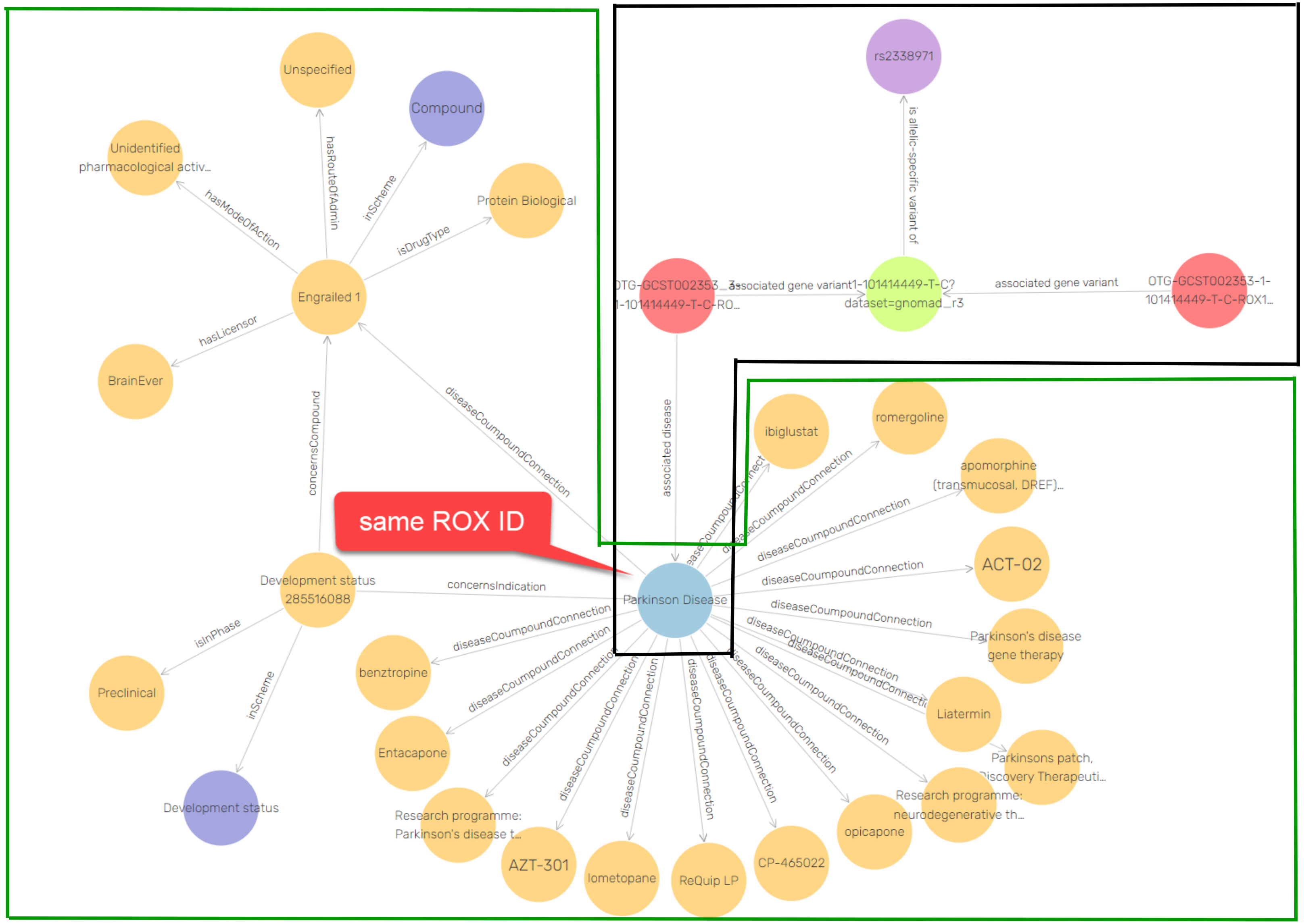
I2O Ontology

Instantiation of a Knowledge Graph



FAIR Data Integration

Federation of Knowledge Graphs (Zero Integration)



Unspecified
Compound
Unidentified pharmacological activ...
Protein Biological
hasRouteOfAdmin
inScheme
hasModeOfAction
isDrugType
Engrailed 1
hasLicensor
BrainEver
concernsCompound
diseaseCoumpoundConnection
same ROX ID
Development status 285516088
concernsIndication
isInPhase
Preclinical
inScheme
Development status
benztropine
Entacapone
Research programme: Parkinson's disease t...
AZT-301
lometopane
ReQuip LP
CP-465022
opicapone
Research programme: neurodegenerative th...
Parkinsons patch, Discovery Therapeuti...
Liatermin
Parkinson's disease gene therapy
ACT-02
apomorphine (transmucosal, DREF)...
romergoline
ibiglustat
Parkinson Disease
associated disease
rs2338971
is allelic-specific variant of
OTG-GCST002353_1-101414449-T-C-RO...
associated gene variant
1-101414449-T-C?dataset=gnomad_r3
associated gene variant
OTG-GCST002353-1-101414449-T-C-ROX1...

I2O Knowledge Graph

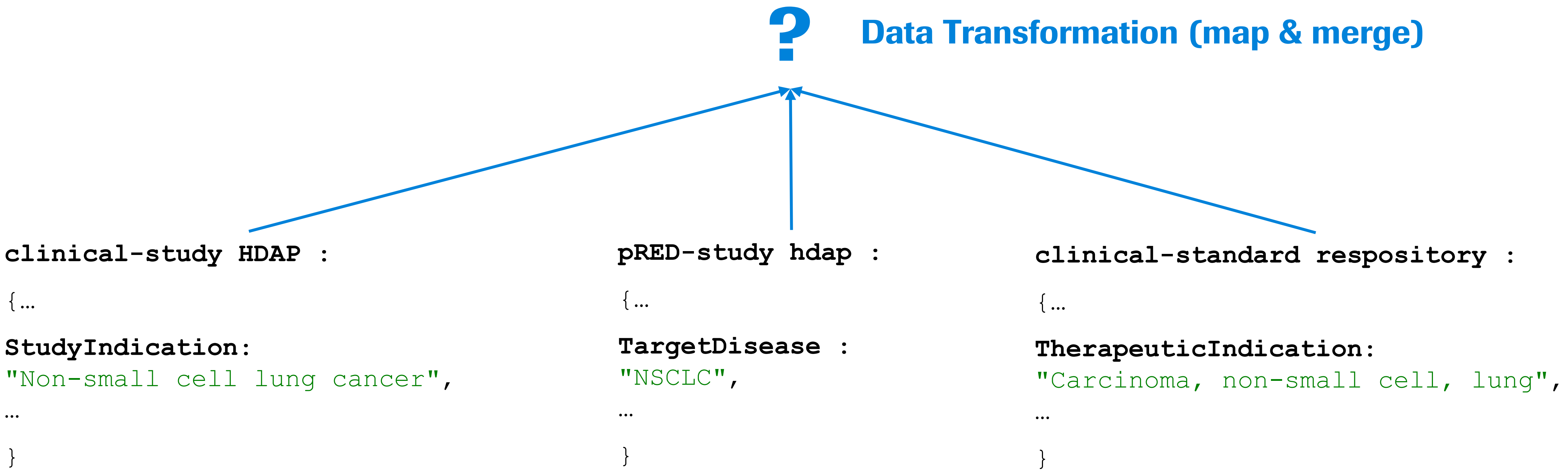
Competitor Information Knowledge Graph

FAIRification at Scale -

FAIR APIs with JSON Linked Data (LD)

Digital Data Assets & Data Management

The Hamster Wheel (Why FAIR Data is not sufficient)



? Data Transformation (map & merge)

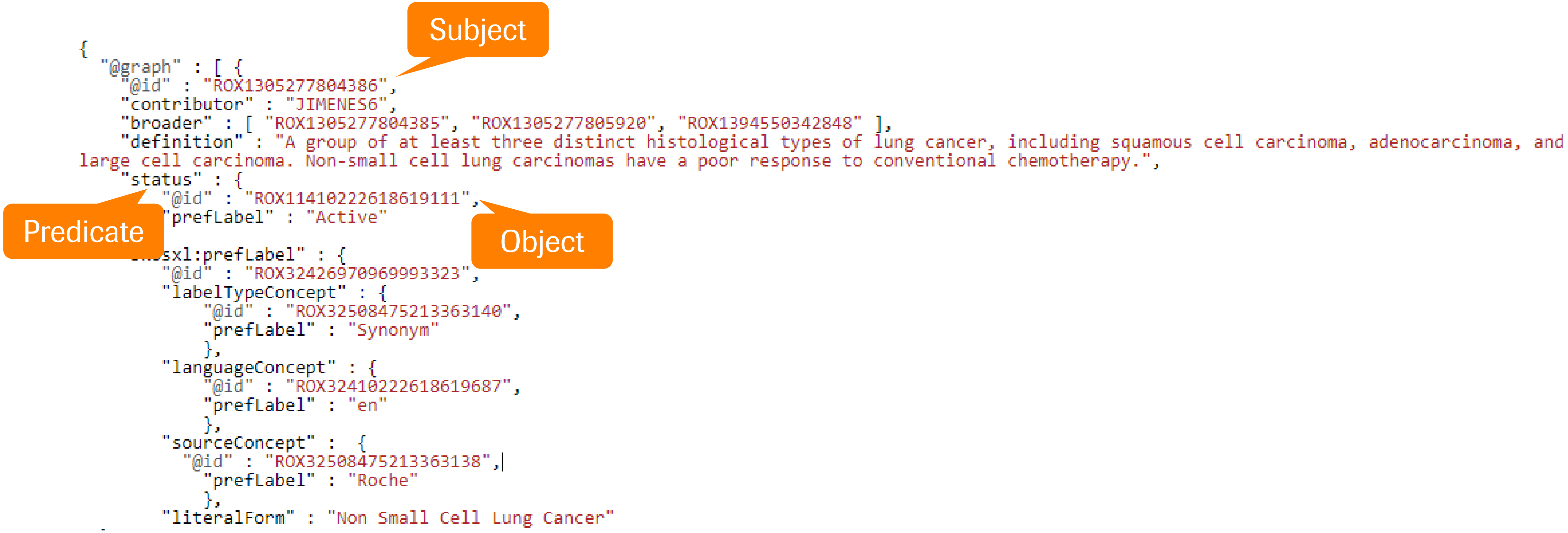
```
clinical-study HDAP :

{…

StudyIndication:
"Non-small cell lung cancer",
…

}
```

```
pRED-study hdap :

{…

TargetDisease :
"NSCLC",
…

}
```

```
clinical-standard respository :

{…

TherapeuticIndication:
"Carcinoma, non-small cell, lung",
…

}
```

JSON-Linked Data (JSON-LD)

Leveraging on a Semantic Infrastructure

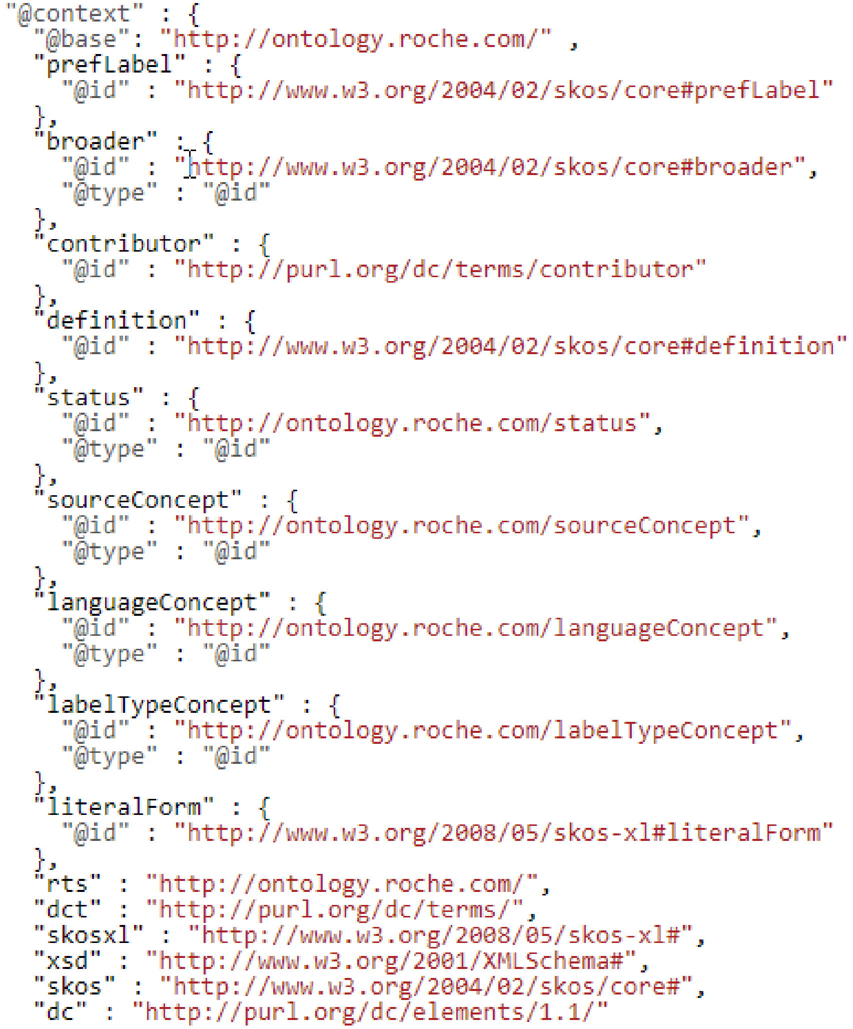


```
{
  "@graph" : [ {
    "@id" : "ROX1305277804386",
    "contributor" : "JIMENES6",
    "broader" : [ "ROX1305277804385", "ROX1305277805920", "ROX1394550342848" ],
    "definition" : "A group of at least three distinct histological types of lung cancer, including squamous cell carcinoma, adenocarcinoma, and
large cell carcinoma. Non-small cell lung carcinomas have a poor response to conventional chemotherapy.",
    "status" : {
        "@id" : "ROX11410222618619111",
        "prefLabel" : "Active"

      sxl:prefLabel" : {
        "@id" : "ROX32426970969993323",
        "labelTypeConcept" : {
            "@id" : "ROX32508475213363140",
            "prefLabel" : "Synonym"
            },
        "languageConcept" : {
            "@id" : "ROX32410222618619687",
            "prefLabel" : "en"
            },
        "sourceConcept" :  {
          "@id" : "ROX32508475213363138",|
            "prefLabel" : "Roche"
            },
        "literalForm" : "Non Small Cell Lung Cancer"
```

➡ In a universe of FAIR applications, data and services *everything* should be considered as a resource

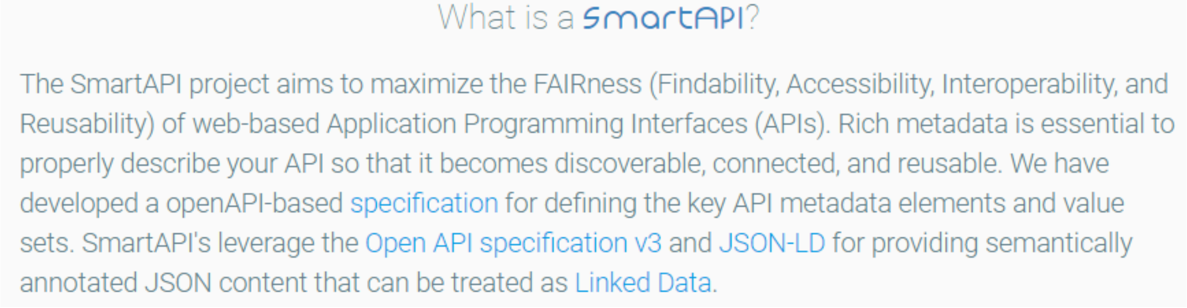
JSON-Linked Data (JSON-LD)

Context provides Model for unambiguous interpretation



```
"@context" : {
  "@base": "http://ontology.roche.com/" ,
  "prefLabel" : {
    "@id" : "http://www.w3.org/2004/02/skos/core#prefLabel"
  },
  "broader" : {
    "@id" : "http://www.w3.org/2004/02/skos/core#broader",
    "@type" : "@id"
  },
  "contributor" : {
    "@id" : "http://purl.org/dc/terms/contributor"
  },
  "definition" : {
    "@id" : "http://www.w3.org/2004/02/skos/core#definition"
  },
  "status" : {
    "@id" : "http://ontology.roche.com/status",
    "@type" : "@id"
  },
  "sourceConcept" : {
    "@id" : "http://ontology.roche.com/sourceConcept",
    "@type" : "@id"
  },
  "languageConcept" : {
    "@id" : "http://ontology.roche.com/languageConcept",
    "@type" : "@id"
  },
  "labelTypeConcept" : {
    "@id" : "http://ontology.roche.com/labelTypeConcept",
    "@type" : "@id"
  },
  "literalForm" : {
    "@id" : "http://www.w3.org/2008/05/skos-xl#literalForm"
  },
  "rts" : "http://ontology.roche.com/",
  "dct" : "http://purl.org/dc/terms/",
  "skosxl" : "http://www.w3.org/2008/05/skos-xl#",
  "xsd" : "http://www.w3.org/2001/XMLSchema#",
  "skos" : "http://www.w3.org/2004/02/skos/core#",
  "dc" : "http://purl.org/dc/elements/1.1/"
```

Model

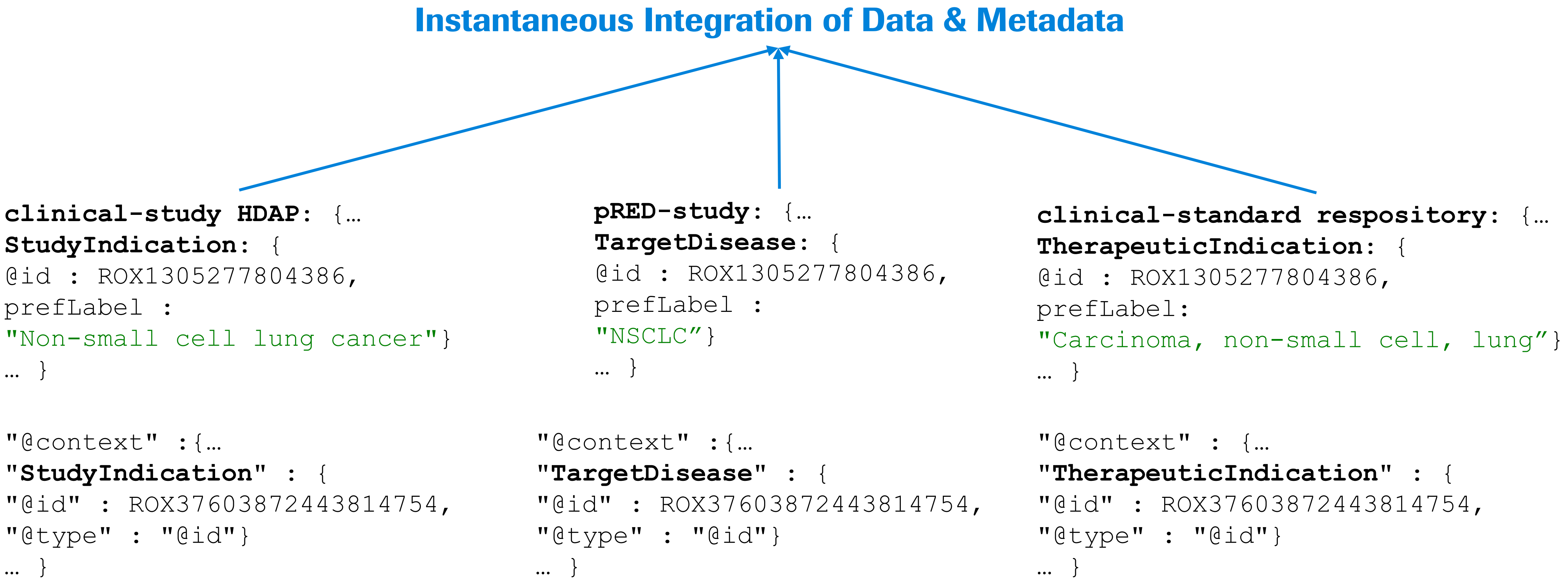


What is a SmartAPI?

The SmartAPI project aims to maximize the FAIRness (Findability, Accessibility, Interoperability, and Reusability) of web-based Application Programming Interfaces (APIs). Rich metadata is essential to properly describe your API so that it becomes discoverable, connected, and reusable. We have developed a openAPI-based specification for defining the key API metadata elements and value sets. SmartAPI's leverage the Open API specification v3 and JSON-LD for providing semantically annotated JSON content that can be treated as Linked Data.

Digital Data Assets & Data Management

Breaking up the Vicious Circle



Instantaneous Integration of Data & Metadata

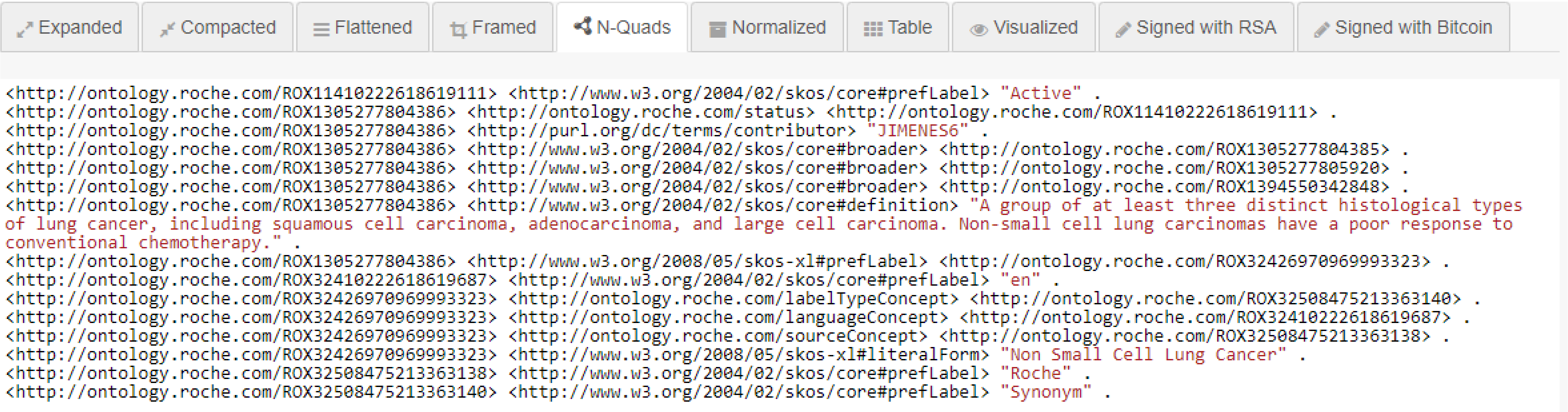
```
clinical-study HDAP: {…
StudyIndication: {
@id : ROX1305277804386,
prefLabel :
"Non-small cell lung cancer"}
… }

"@context" :{…
"StudyIndication" : {
"@id" : ROX37603872443814754,
"@type" : "@id"}
… }
```

```
pRED-study: {…
TargetDisease: {
@id : ROX1305277804386,
prefLabel :
"NSCLC”}
… }

"@context" :{…
"TargetDisease" : {
"@id" : ROX37603872443814754,
"@type" : "@id"}
… }
```

```
clinical-standard respository: {…
TherapeuticIndication: {
@id : ROX1305277804386,
prefLabel:
"Carcinoma, non-small cell, lung”}
… }

"@context" : {…
"TherapeuticIndication" : {
"@id" : ROX37603872443814754,
"@type" : "@id"}
… }
```

JSON-Linked Data (JSON-LD)

RDF Serialization – immediate usage



```
<http://ontology.roche.com/ROX1141022261861911l> <http://www.w3.org/2004/02/skos/core#prefLabel> "Active" .
<http://ontology.roche.com/ROX1305277804386> <http://ontology.roche.com/status> <http://ontology.roche.com/ROX1141022261861911l> .
<http://ontology.roche.com/ROX1305277804386> <http://purl.org/dc/terms/contributor> "JIMENES6" .
<http://ontology.roche.com/ROX1305277804386> <http://www.w3.org/2004/02/skos/core#broader> <http://ontology.roche.com/ROX1305277804385> .
<http://ontology.roche.com/ROX1305277804386> <http://www.w3.org/2004/02/skos/core#broader> <http://ontology.roche.com/ROX1305277805920> .
<http://ontology.roche.com/ROX1305277804386> <http://www.w3.org/2004/02/skos/core#broader> <http://ontology.roche.com/ROX1394550342848> .
<http://ontology.roche.com/ROX1305277804386> <http://www.w3.org/2004/02/skos/core#definition> "A group of at least three distinct histological types
of lung cancer, including squamous cell carcinoma, adenocarcinoma, and large cell carcinoma. Non-small cell lung carcinomas have a poor response to
conventional chemotherapy." .
<http://ontology.roche.com/ROX1305277804386> <http://www.w3.org/2008/05/skos-xl#prefLabel> <http://ontology.roche.com/ROX3242697096999323> .
<http://ontology.roche.com/ROX32410222618619687> <http://www.w3.org/2004/02/skos/core#prefLabel> "en" .
<http://ontology.roche.com/ROX3242697096999323> <http://ontology.roche.com/labelTypeConcept> <http://ontology.roche.com/ROX32508475213363140> .
<http://ontology.roche.com/ROX3242697096999323> <http://ontology.roche.com/languageConcept> <http://ontology.roche.com/ROX32410222618619687> .
<http://ontology.roche.com/ROX3242697096999323> <http://ontology.roche.com/sourceConcept> <http://ontology.roche.com/ROX32508475213363138> .
<http://ontology.roche.com/ROX3242697096999323> <http://www.w3.org/2008/05/skos-xl#literalForm> "Non Small Cell Lung Cancer" .
<http://ontology.roche.com/ROX32508475213363138> <http://www.w3.org/2004/02/skos/core#prefLabel> "Roche" .
<http://ontology.roche.com/ROX32508475213363140> <http://www.w3.org/2004/02/skos/core#prefLabel> "Synonym" .
```

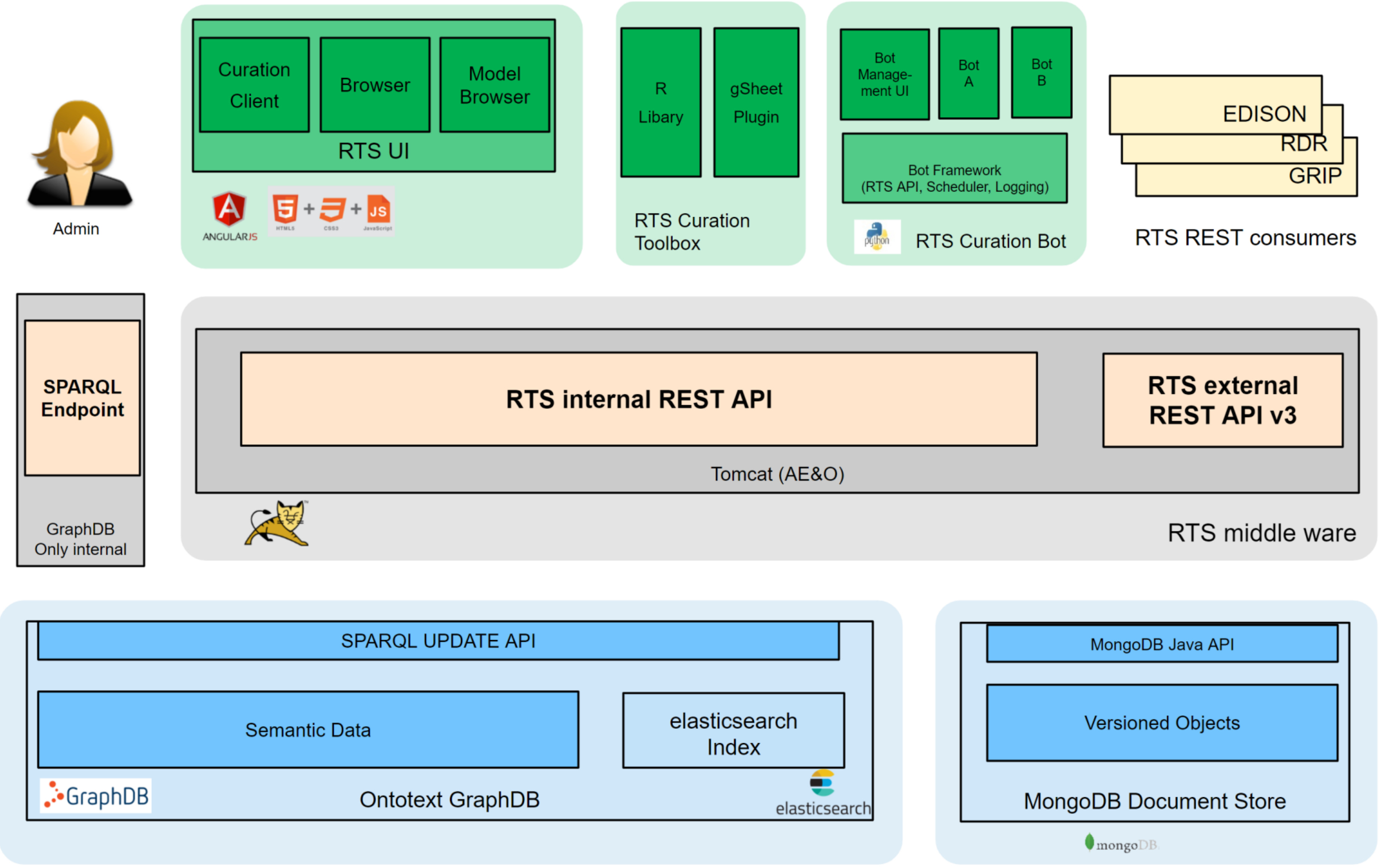
FAIRification at Scale - Implementation

FAIR plus Data Quality as a Service (DQaaS)

Semantic Interoperability Hub

Architecture/ FAIR by Design

Roche



Admin

Curation Client

Browser

Model Browser

RTS UI

ANGULARJS

HTML5 + CSS3 + JS

R Libary

gSheet Plugin

RTS Curation Toolbox

Bot Manage-ment UI

Bot A

Bot B

Bot Framework (RTS API, Scheduler, Logging)

python

RTS Curation Bot

EDISON

RDR

GRIP

RTS REST consumers

SPARQL Endpoint

GraphDB Only internal

RTS internal REST API

RTS external REST API v3

Tomcat (AE&O)

RTS middle ware

SPARQL UPDATE API

Semantic Data

elasticsearch Index

GraphDB

Ontotext GraphDB

elasticsearch

MongoDB Java API

Versioned Objects

MongoDB Document Store

mongoDB

Data Harmonization Service – Offerings

Data assets born FAIR

Data Quality based on Standardized Terminologies

Custom Tailored Terminologies, Dataset model & Ontologies (FAIRification at Scale)

Technical Support

Data Links for Seamless Contextual Navigation

Maintenance & Enhancements of Standards

Data Standards and Data Harmonization (born FAIR)

Knowledge Integration Hub for Identifier Mapping

Shared Semantics (Conceptual models)

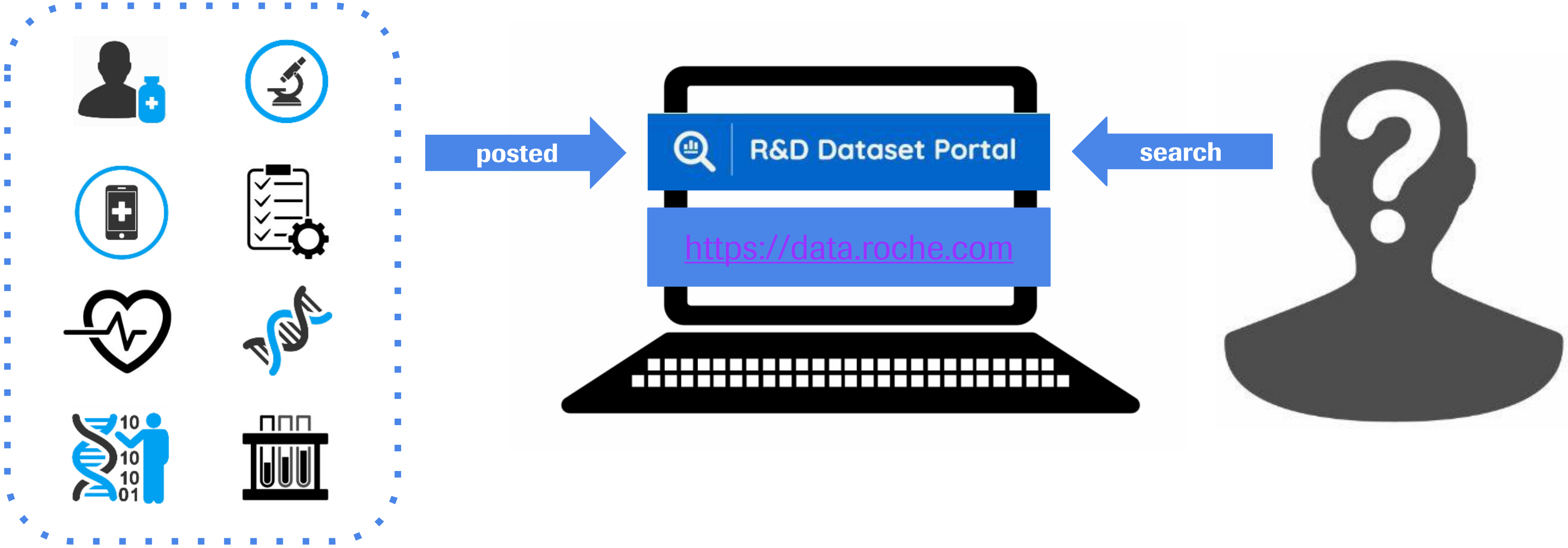
Toolbox (Service Layer Windows)

FAIRification at Scale Use Case

A true FAIRy Tale – the Roche Dataset Portal

R&D Dataset Portal

Data Catalog of Data Catalogs



posted

R&D Dataset Portal

https://data.roche.com

search

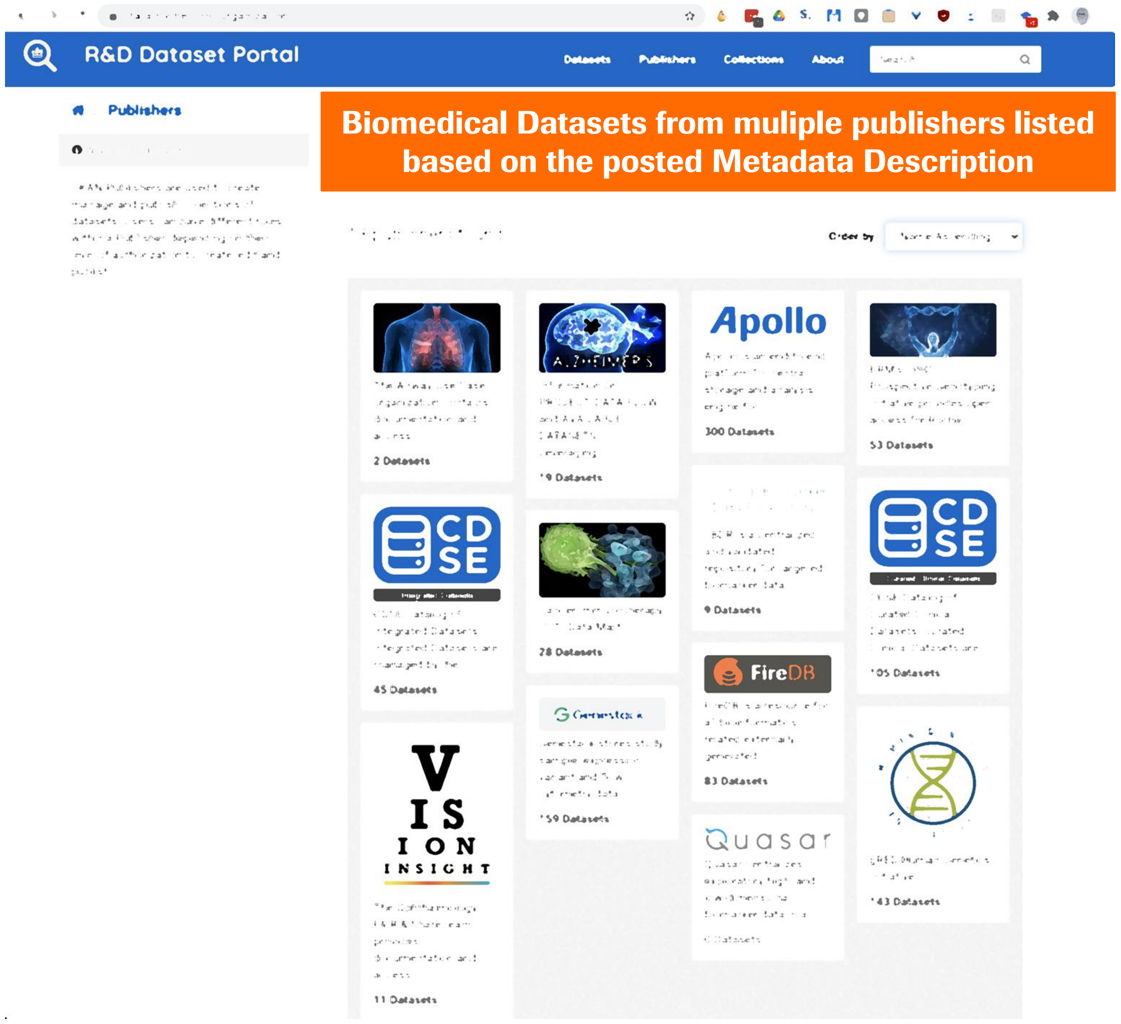
Biomedical datasets from **Roche R&D data catalogs** *e.g. biomarker, clinical, digital, imaging, omics or real world datasets*

Cataloged and stored in source systems **published to the R&D Dataset Portal** as a central place to search & access corporate data assets

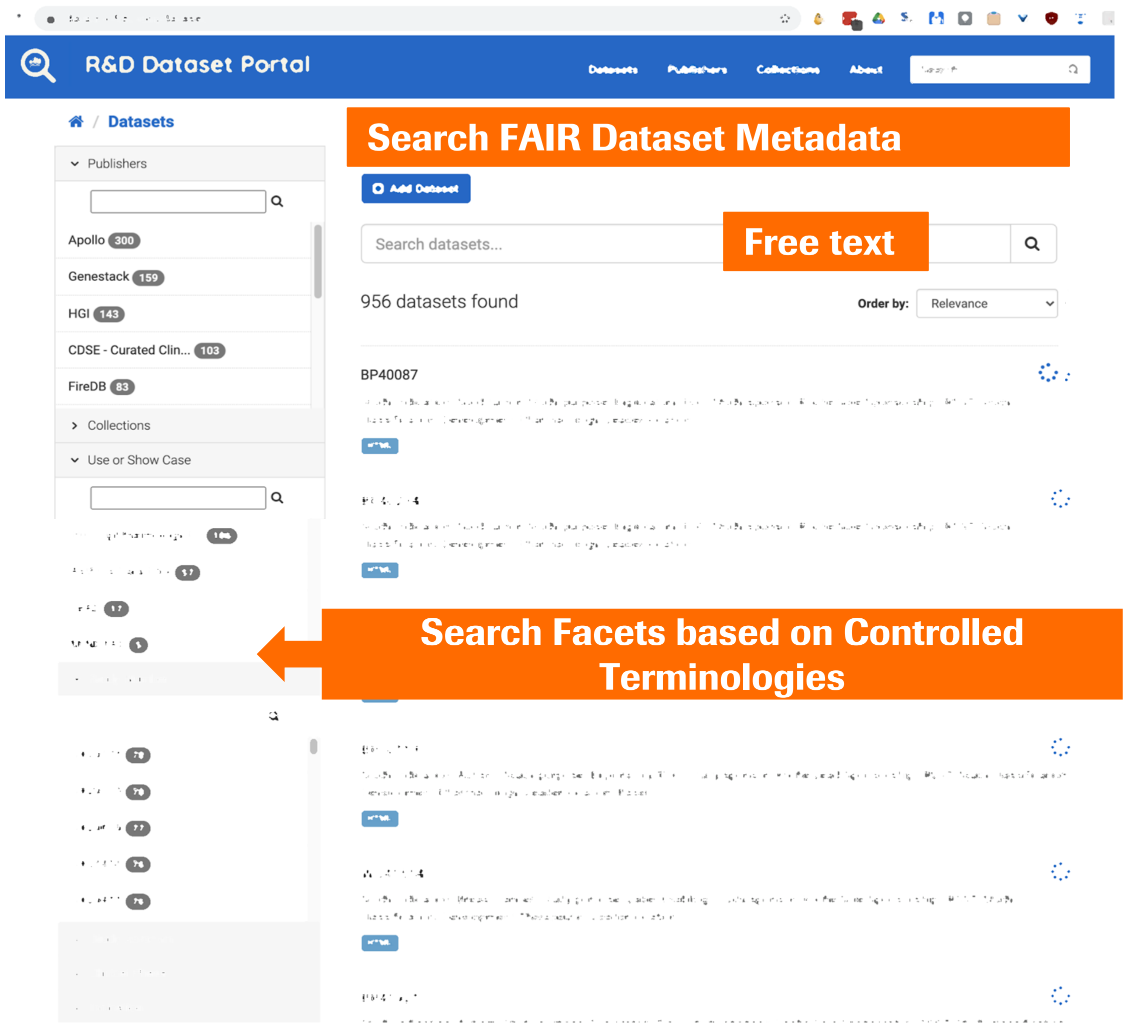
Scientists in Roche can **search** for Biomedical datasets from PD, pRED, gRED, DIA, etc.

Roche Dataset Portal

Find Biomedical Datasets Across R&D



Biomedical Datasets from muliple publishers listed based on the posted Metadata Description



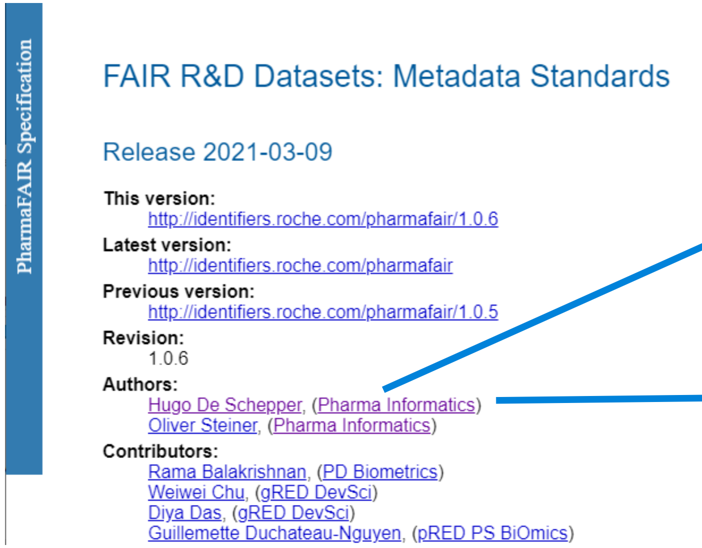
Search FAIR Dataset Metadata

Free text

Search Facets based on Controlled Terminologies

R&D Dataset Portal

FAIR Representation of Metadata & Data



PharmaFAIR Specification

FAIR R&D Datasets: Metadata Standards

Release 2021-03-09

This version:
http://identifiers.roche.com/pharmafair/1.0.6

Latest version:
http://identifiers.roche.com/pharmafair

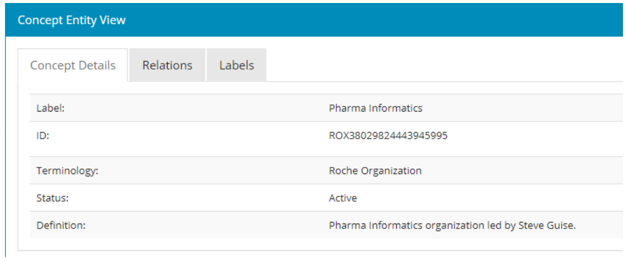
Previous version:
http://identifiers.roche.com/pharmafair/1.0.5

Revision:
1.0.6

Authors:
Hugo De Schepper, (Pharma Informatics)
Oliver Steiner, (Pharma Informatics)

Contributors:
Rama Balakrishnan, (PD Biometrics)
Weiwei Chu, (gRED DevSci)
Diya Das, (gRED DevSci)
Guillemette Duchateau-Nguyen, (pRED PS BiOmics)

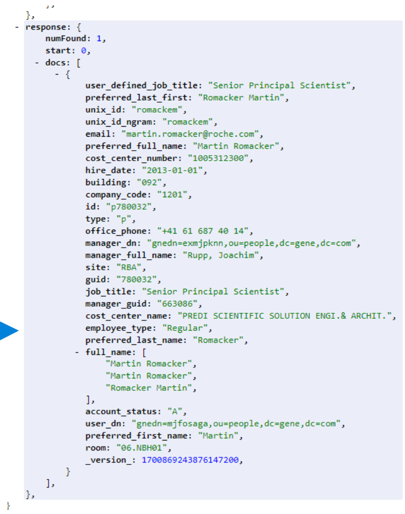
language en



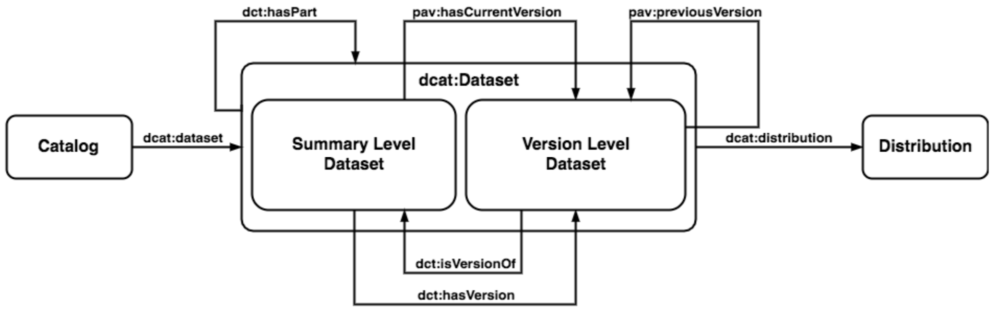
Concept Entity View

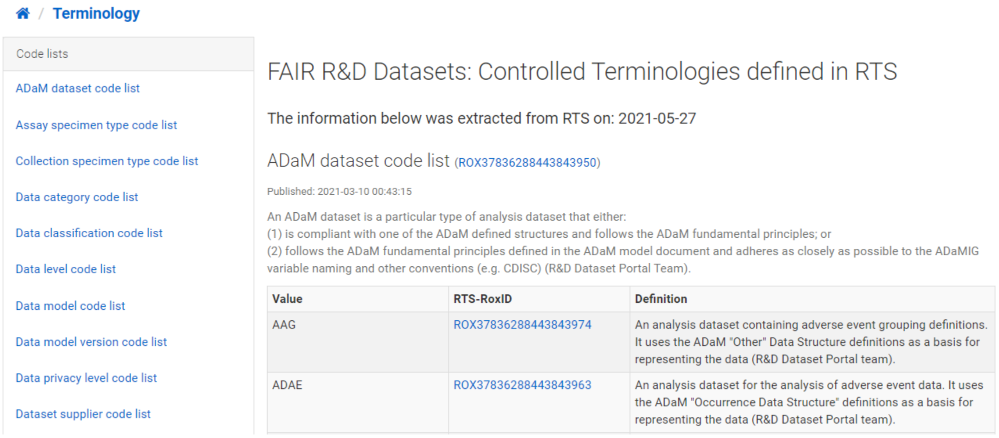
Label:	Pharma Informatics
ID:	ROX38029824443945995
Terminology:	Roche Organization
Status:	Active
Definition:	Pharma Informatics organization led by Steve Guise.

identifier.roche.com



```
    },
  },
- response: {
      numFound: 1,
      start: 0,
    - docs: [
        - {
              user_defined_job_title: "Senior Principal Scientist",
              preferred_last_first: "Romacker Martin",
              unix_id: "romackem",
              unix_id_ngram: "romackem",
              email: "martin.romacker@roche.com",
              preferred_full_name: "Martin Romacker",
              cost_center_number: "1005312300",
              hire_date: "2013-01-01",
              building: "092",
              company_code: "1201",
              id: "p780032",
              type: "p",
              office_phone: "+41 61 687 40 14",
              manager_dn: "gnedn=exmjpknn,ou=people,dc=gene,dc=com",
              manager_full_name: "Rupp, Joachim",
              site: "RBA",
              guid: "780032",
              job_title: "Senior Principal Scientist",
              manager_guid: "663086",
              cost_center_name: "PREDI SCIENTIFIC SOLUTION ENGI.& ARCHIT.",
              employee_type: "Regular",
              preferred_last_name: "Romacker",
            - full_name: [
                  "Martin Romacker",
                  "Martin Romacker",
                  "Romacker Martin",
              ],
              account_status: "A",
              user_dn: "gnedn=mjfosaga,ou=people,dc=gene,dc=com",
              preferred_first_name: "Martin",
              room: "06.NBH01",
              _version_: 1700869243876147200,
          }
      ],
  },
}
```





Code lists

- ADaM dataset code list
- Assay specimen type code list
- Collection specimen type code list
- Data category code list
- Data classification code list
- Data level code list
- Data model code list
- Data model version code list
- Data privacy level code list
- Dataset supplier code list

FAIR R&D Datasets: Controlled Terminologies defined in RTS

The information below was extracted from RTS on: 2021-05-27

ADaM dataset code list (ROX37836288443843950)

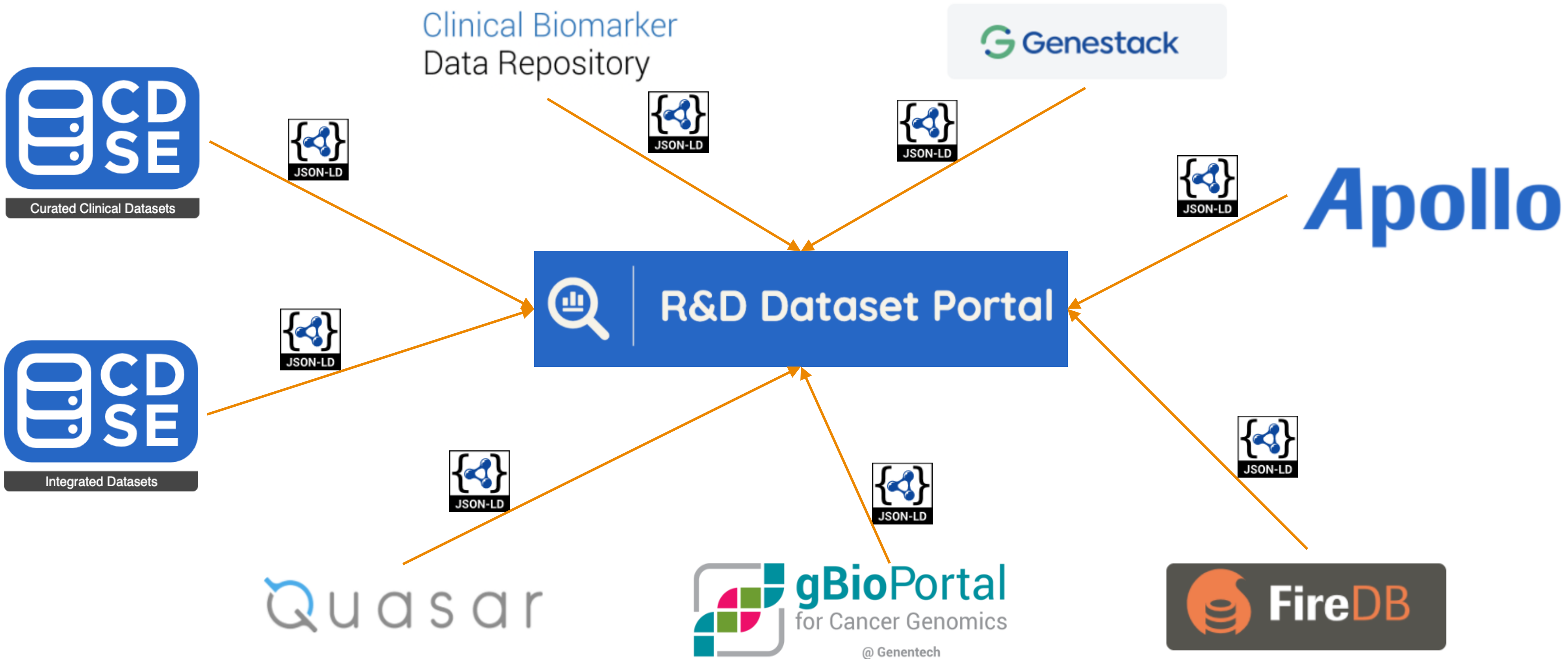
Published: 2021-03-10 00:43:15

An ADaM dataset is a particular type of analysis dataset that either:
(1) is compliant with one of the ADaM defined structures and follows the ADaM fundamental principles; or
(2) follows the ADaM fundamental principles defined in the ADaM model document and adheres as closely as possible to the ADaMIG variable naming and other conventions (e.g. CDISC) (R&D Dataset Portal Team).

Value	RTS-RoxID	Definition
AAG	ROX37836288443843974	An analysis dataset containing adverse event grouping definitions. It uses the ADaM "Other" Data Structure definitions as a basis for representing the data (R&D Dataset Portal team).
ADAE	ROX37836288443843963	An analysis dataset for the analysis of adverse event data. It uses the ADaM "Occurrence Data Structure" definitions as a basis for representing the data (R&D Dataset Portal team).

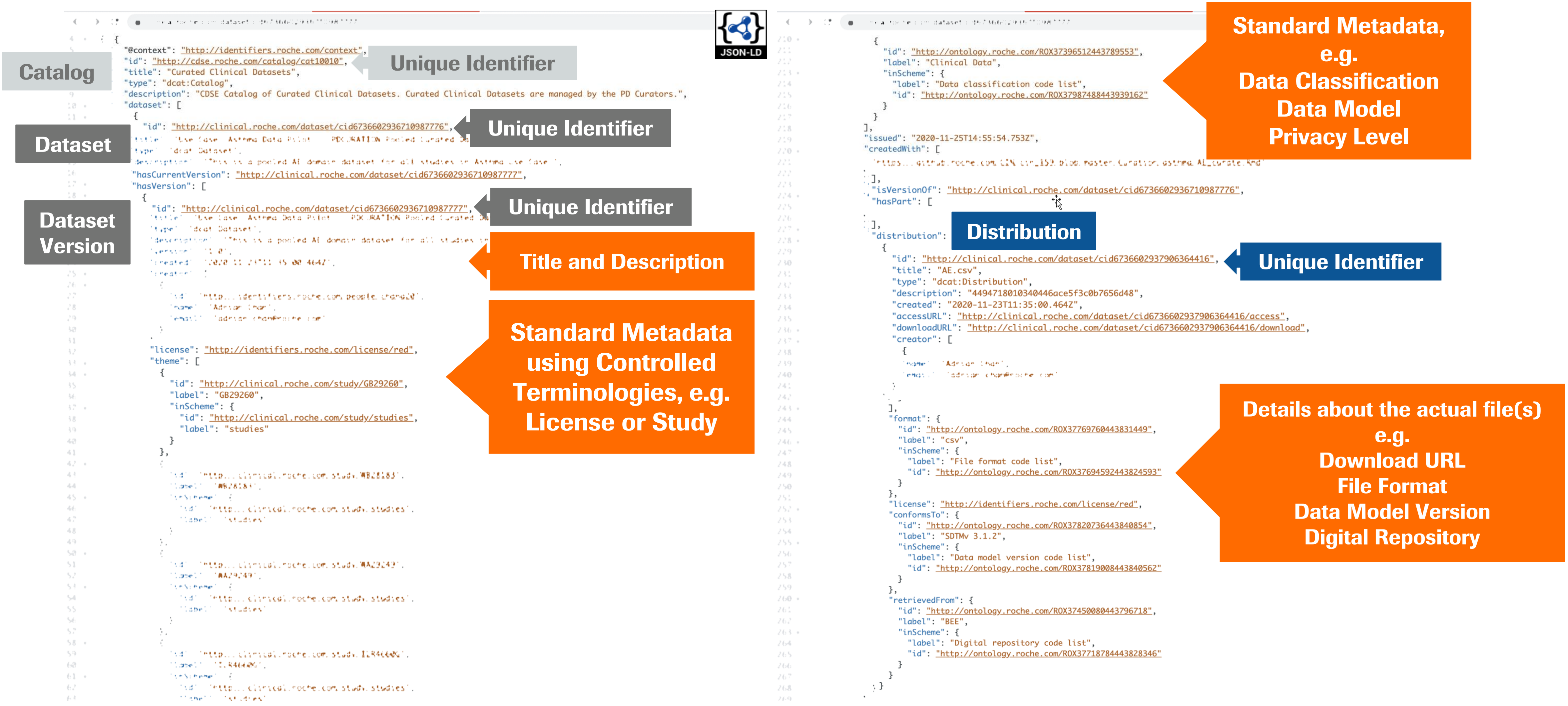
Standardized Dataset Metadata & Data (Terminology)

JSON-LD format specified in R&D Dataset Metadata Standards (Data contracts)



R&D Dataset Metadata

JSON-LD API (served by all data catalogs based on prospective FAIRification)



Catalog

Unique Identifier

Dataset

Unique Identifier

Dataset Version

Unique Identifier

Title and Description

Standard Metadata using Controlled Terminologies, e.g. License or Study

Standard Metadata, e.g. Data Classification Data Model Privacy Level

Distribution

Unique Identifier

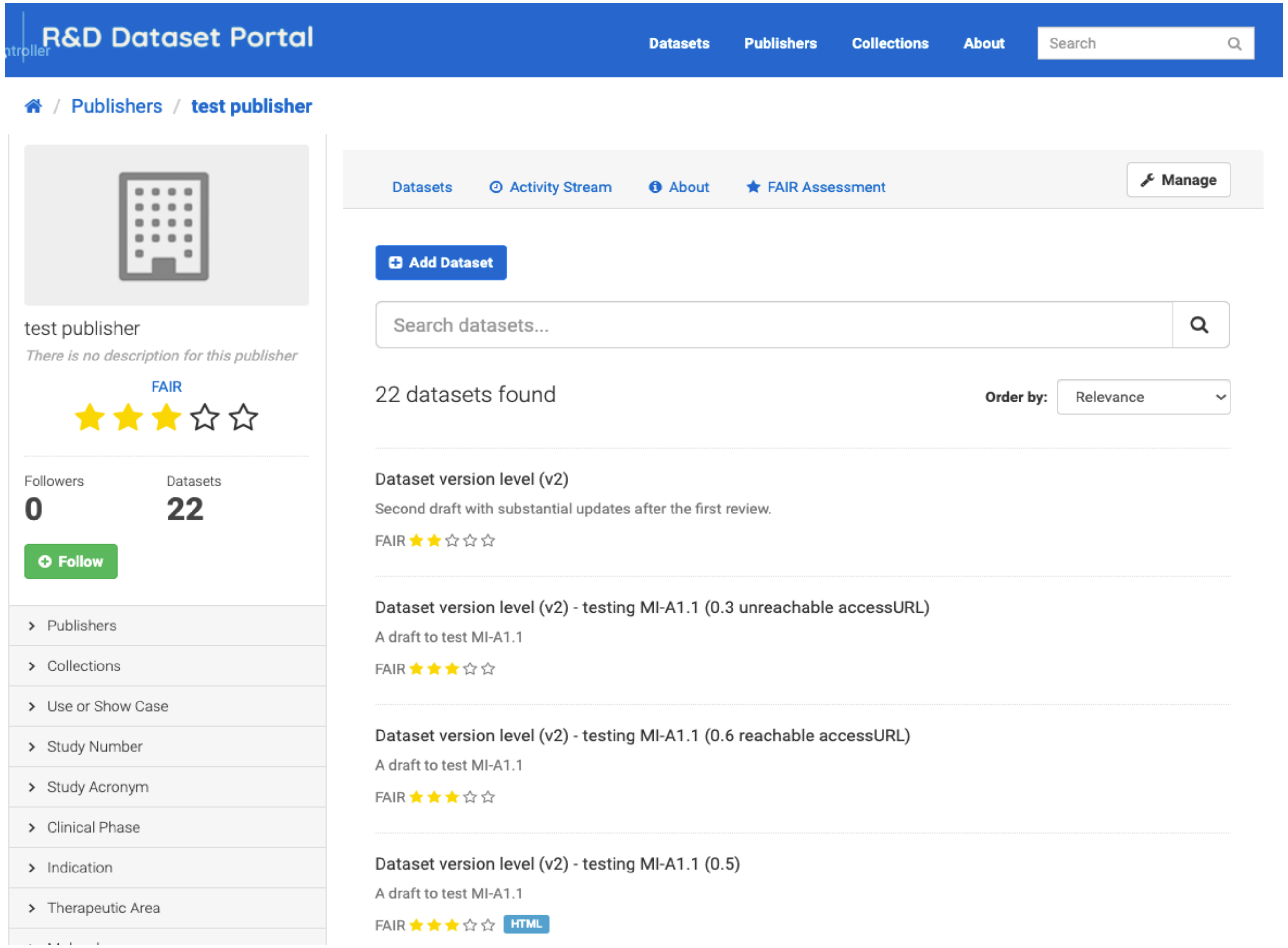
Details about the actual file(s) e.g. Download URL File Format Data Model Version Digital Repository

```
{
  "@context": "http://identifiers.roche.com/context",
  "id": "http://cdse.roche.com/catalog/cat10010",
  "title": "Curated Clinical Datasets",
  "type": "dcat:Catalog",
  "description": "CDSE Catalog of Curated Clinical Datasets. Curated Clinical Datasets are managed by the PD Curators.",
  "dataset": [
    {
      "id": "http://clinical.roche.com/dataset/cid6736602936710987776",
      [illegible]
      "hasCurrentVersion": "http://clinical.roche.com/dataset/cid6736602936710987777",
      "hasVersion": [
        {
          "id": "http://clinical.roche.com/dataset/cid6736602936710987777",
          [illegible]
          "license": "http://identifiers.roche.com/license/red",
          "theme": [
            {
              "id": "http://clinical.roche.com/study/GB29260",
              "label": "GB29260",
              "inScheme": {
                "id": "http://clinical.roche.com/study/studies",
                "label": "studies"
              }
            },
            [illegible]
```

```
            {
              "id": "http://ontology.roche.com/ROX37396512443789553",
              "label": "Clinical Data",
              "inScheme": {
                "label": "Data classification code list",
                "id": "http://ontology.roche.com/ROX37987488443939162"
              }
            }
          ],
          "issued": "2020-11-25T14:55:54.753Z",
          "createdWith": [
            [illegible]
          ],
          "isVersionOf": "http://clinical.roche.com/dataset/cid6736602936710987776",
          "hasPart": [
          ],
          "distribution":
            {
              "id": "http://clinical.roche.com/dataset/cid6736602937906364416",
              "title": "AE.csv",
              "type": "dcat:Distribution",
              "description": "4494718010340446ace5f3c0b7656d48",
              "created": "2020-11-23T11:35:00.464Z",
              "accessURL": "http://clinical.roche.com/dataset/cid6736602937906364416/access",
              "downloadURL": "http://clinical.roche.com/dataset/cid6736602937906364416/download",
              "creator": [
                {
                  [illegible]
                }
              ],
              "format": {
                "id": "http://ontology.roche.com/ROX37769760443831449",
                "label": "csv",
                "inScheme": {
                  "label": "File format code list",
                  "id": "http://ontology.roche.com/ROX37694592443824593"
                }
              },
              "license": "http://identifiers.roche.com/license/red",
              "conformsTo": {
                "id": "http://ontology.roche.com/ROX37820736443840854",
                "label": "SDTMv 3.1.2",
                "inScheme": {
                  "label": "Data model version code list",
                  "id": "http://ontology.roche.com/ROX37819008443840562"
                }
              },
              "retrievedFrom": {
                "id": "http://ontology.roche.com/ROX37450080443796718",
                "label": "BEE",
                "inScheme": {
                  "label": "Digital repository code list",
                  "id": "http://ontology.roche.com/ROX37718784443828346"
                }
              }
            }
}
```

Roche Dataset Portal

Automatic FAIR Assessment



test publisher

There is no description for this publisher

FAIR ★★★☆☆

Followers 0

Datasets 22

22 datasets found

Dataset version level (v2)

Second draft with substantial updates after the first review.

FAIR ★★☆☆☆

Dataset version level (v2) - testing MI-A1.1 (0.3 unreachable accessURL)

A draft to test MI-A1.1

FAIR ★★★☆☆

Dataset version level (v2) - testing MI-A1.1 (0.6 reachable accessURL)

A draft to test MI-A1.1

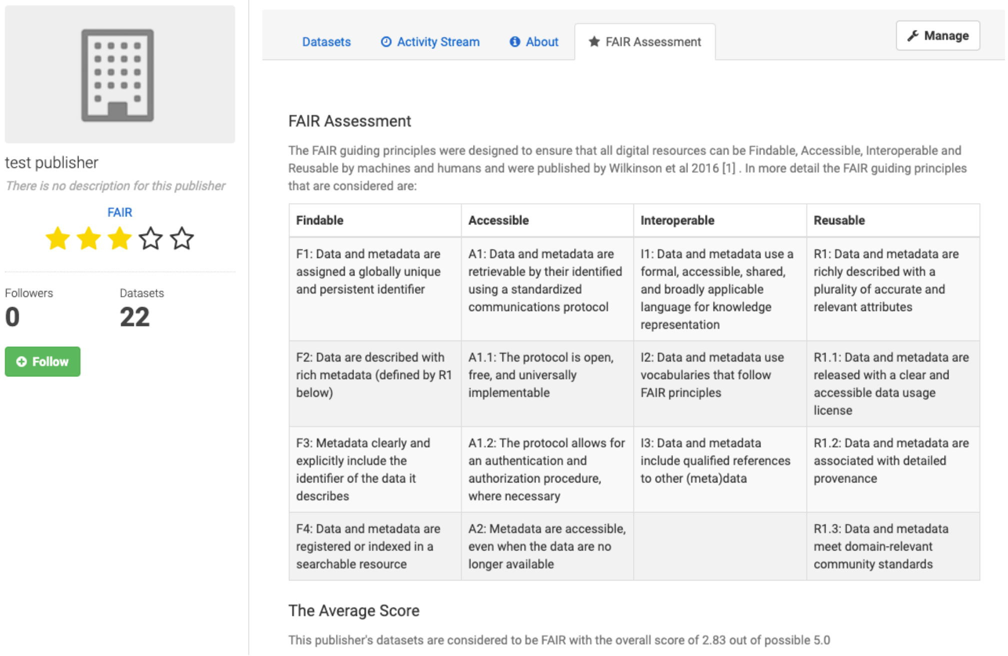
FAIR ★★★☆☆

Dataset version level (v2) - testing MI-A1.1 (0.5)

A draft to test MI-A1.1

FAIR ★★★☆☆

- FAIR representation of Model, Metadata and Data
- Entirely machine-readable FAIR Data Standards
- Automated FAIR Assessment



test publisher

There is no description for this publisher

FAIR ★★★☆☆

Followers 0

Datasets 22

FAIR Assessment

The FAIR guiding principles were designed to ensure that all digital resources can be Findable, Accessible, Interoperable and Reusable by machines and humans and were published by Wilkinson et al 2016 [1] . In more detail the FAIR guiding principles that are considered are:

Findable	Accessible	Interoperable	Reusable
F1: Data and metadata are assigned a globally unique and persistent identifier	A1: Data and metadata are retrievable by their identified using a standardized communications protocol	I1: Data and metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation	R1: Data and metadata are richly described with a plurality of accurate and relevant attributes
F2: Data are described with rich metadata (defined by R1 below)	A1.1: The protocol is open, free, and universally implementable	I2: Data and metadata use vocabularies that follow FAIR principles	R1.1: Data and metadata are released with a clear and accessible data usage license
F3: Metadata clearly and explicitly include the identifier of the data it describes	A1.2: The protocol allows for an authentication and authorization procedure, where necessary	I3: Data and metadata include qualified references to other (meta)data	R1.2: Data and metadata are associated with detailed provenance
F4: Data and metadata are registered or indexed in a searchable resource	A2: Metadata are accessible, even when the data are no longer available		R1.3: Data and metadata meet domain-relevant community standards

The Average Score

This publisher's datasets are considered to be FAIR with the overall score of 2.83 out of possible 5.0

FAIRification at Scale– Capabilities and Implementation

Conclusions

Conclusions

- High-Quality, standardized and linked data: foundation for digitilization & insight generation.
- FAIR data principles intrinsically tie Data Management to Semantic Technologies.
- Data Integration based on interoperable Domain Knowledge Graphs – Unified Domain Model.
  FAIR Ecosystems at scale based on Knowledge Graphs becomes reality.
- 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 and integrate semantic capabilities:
  open public-private semantic infrastructure of FAIR applications, services and data.
- It’s all about Semantics.

Doing now what patients need next