FAIR in Action

From application-centric to data-centric using FAIR architectures and resources

Martin Romacker
Data and Information Architect
Data and Analytics
Pharma Research and Early Development
Roche Innovation Center Basel

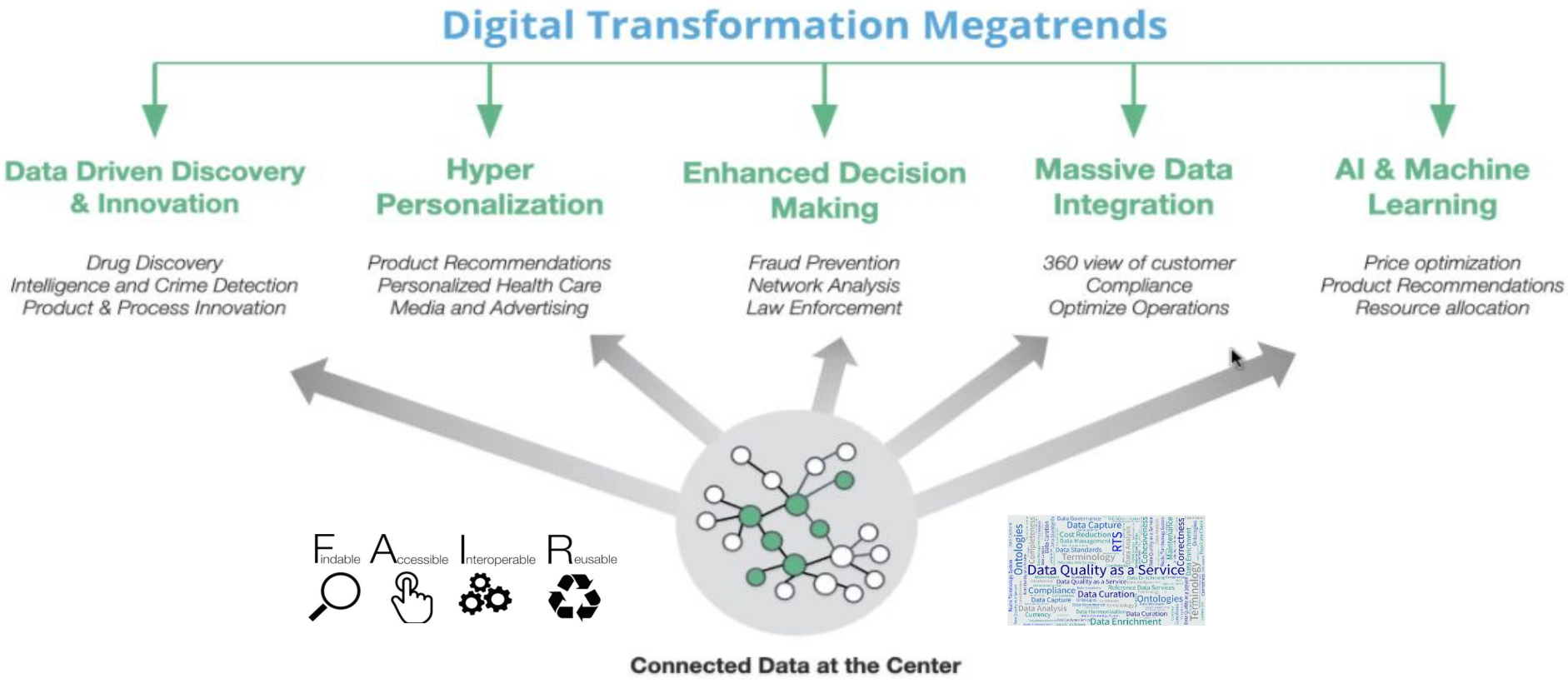
15 November 2022, EOSC Symposium 2022, Prague

EOSC SYMPOSIUM
14-17 November 2022

Table of contents

Business Case for FAIR and Data Quality

Harnessing Connections Drives Business Value



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/ Visible Costs

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

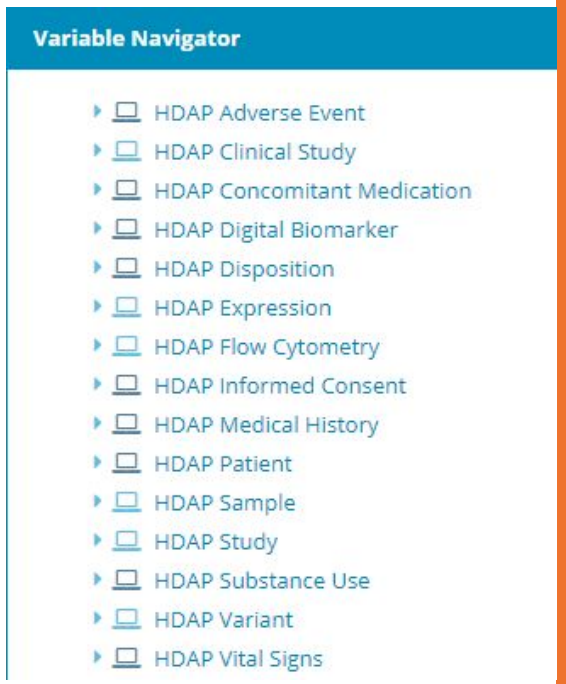
Unplanned/ Invisible Costs

- 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

FAIR and Roche Data Commons

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

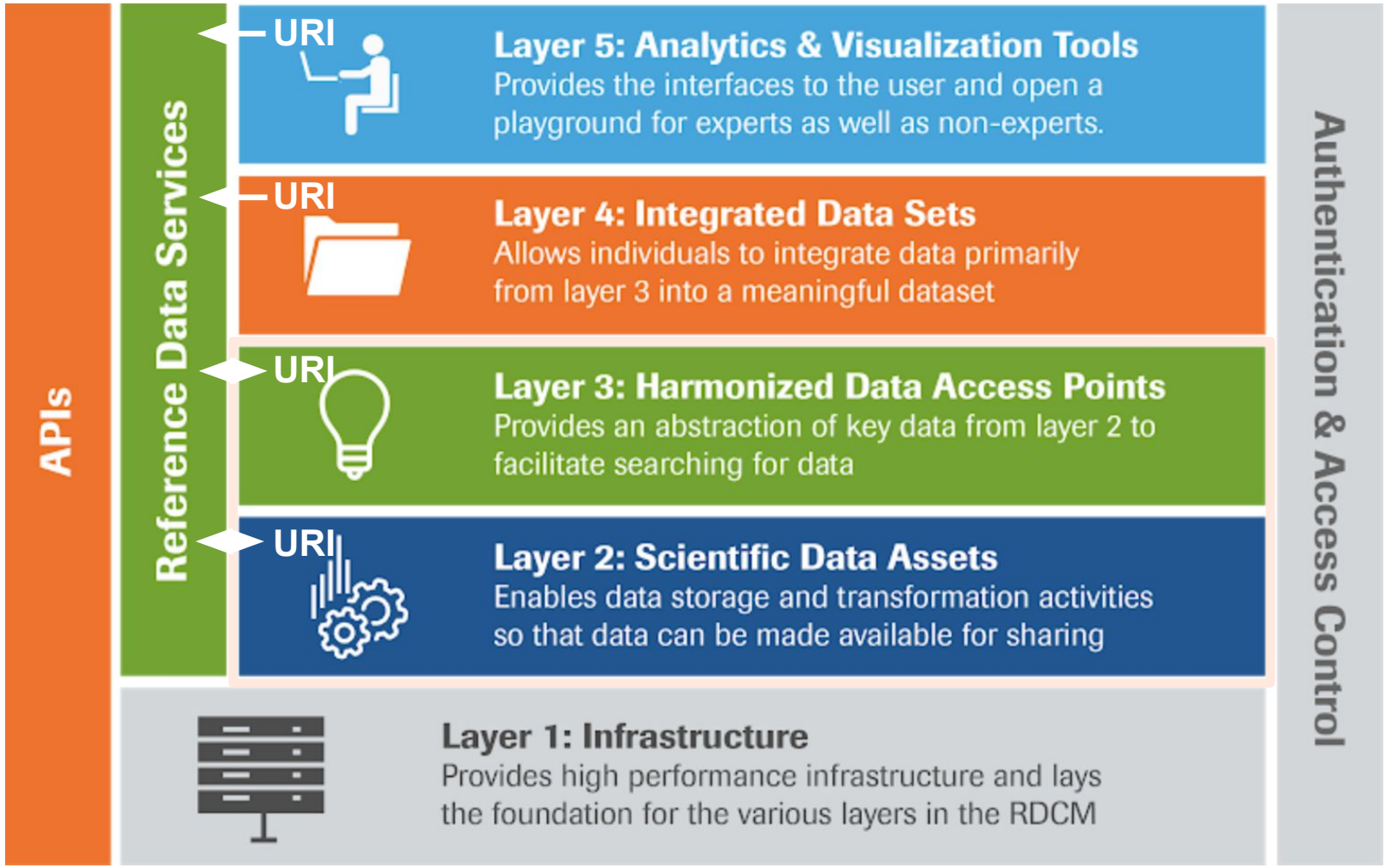
Layer 2: Scientific Data Assets
Enables data storage and transformation activities so that data can be made available for sharing

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

Authentication & Access Control

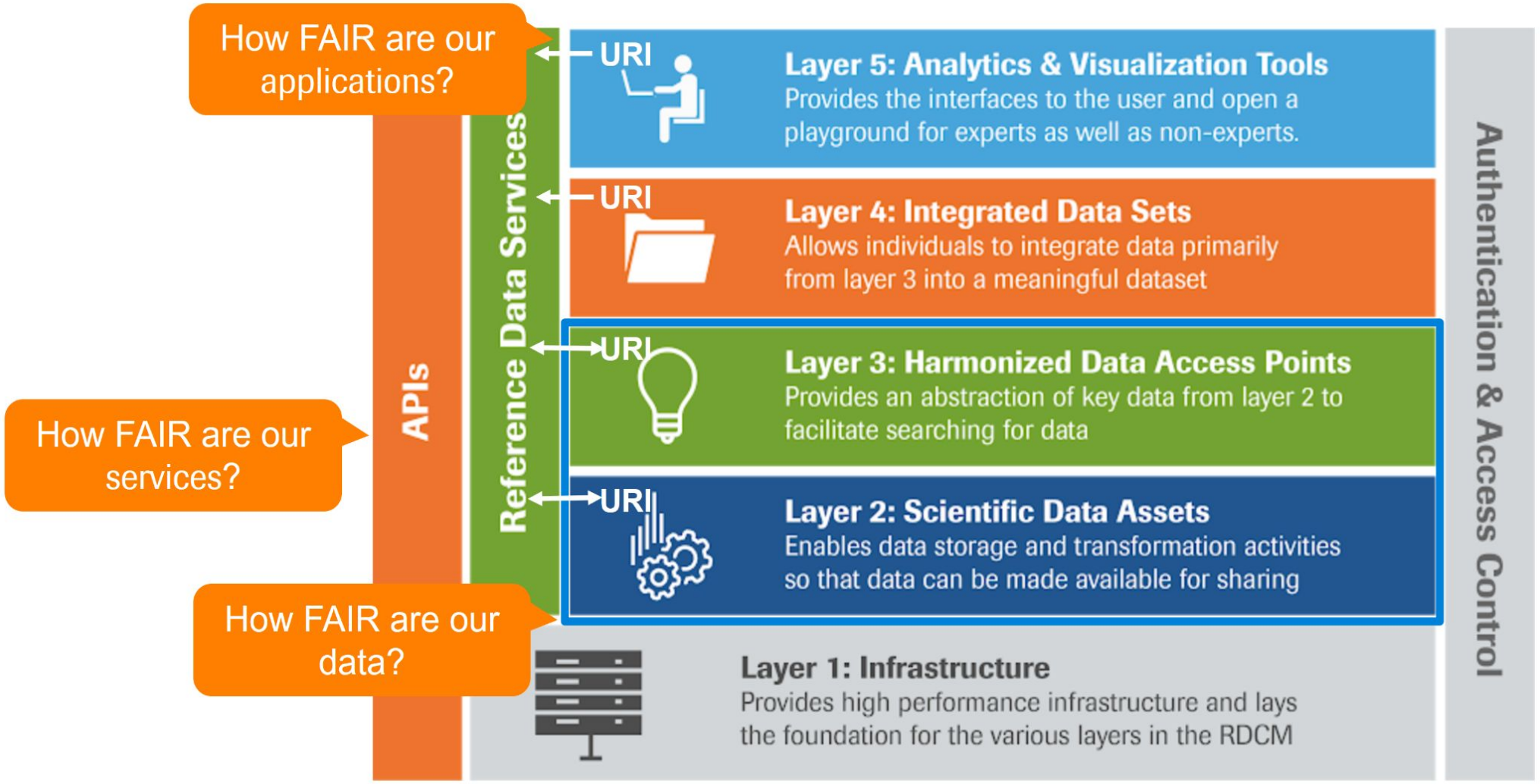
HDAPs organize data in
Information Types Interoperability (URIs):
semantic data dictionary
semantic models
Data FAIRification only in layer 2 & 3

No more transformation between layer 3 & 4,5



Roche Data Commons (RDC)

Semantic Infrastructure of FAIR Data, Services and Applications



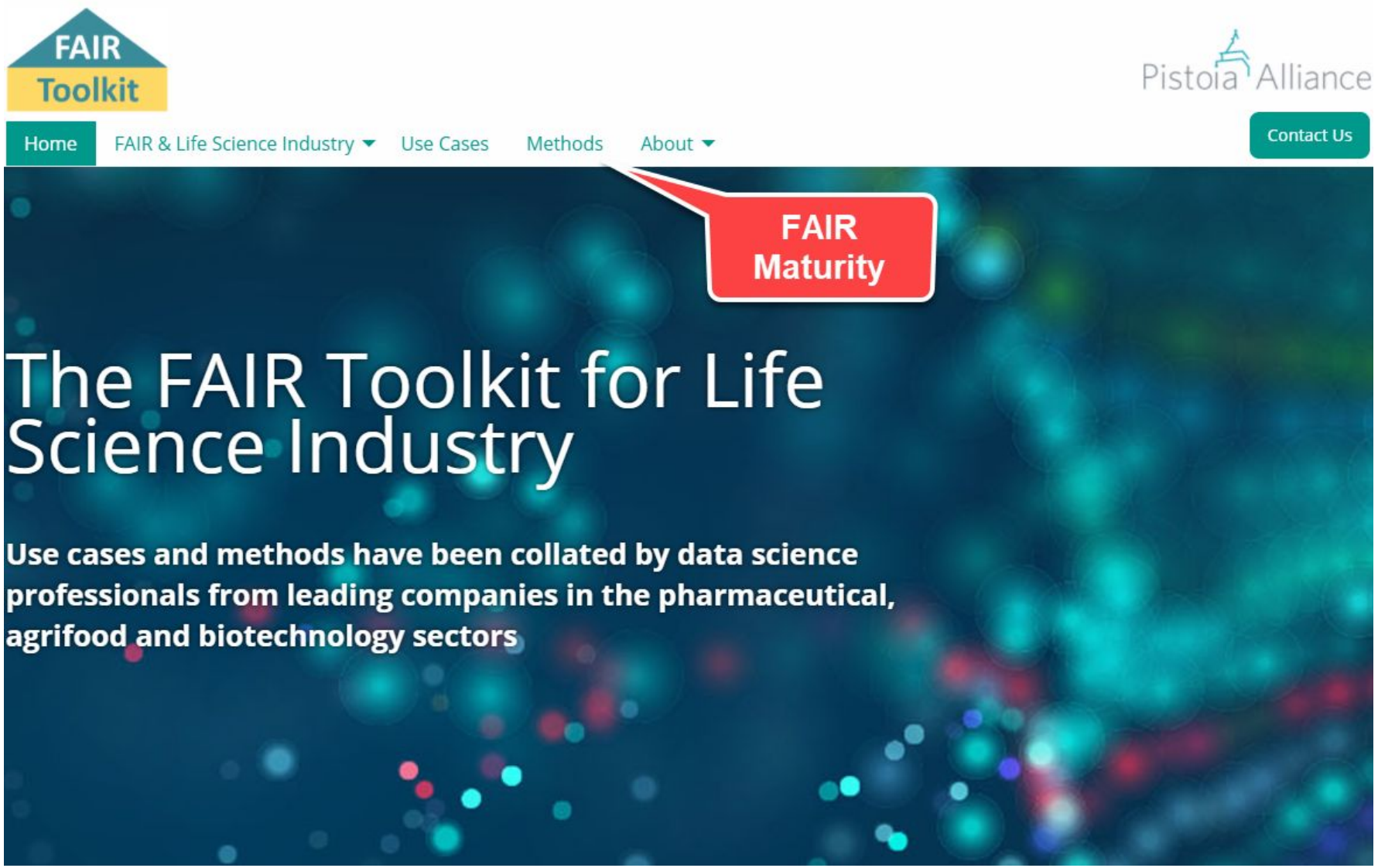
FAIR scientific data management

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)



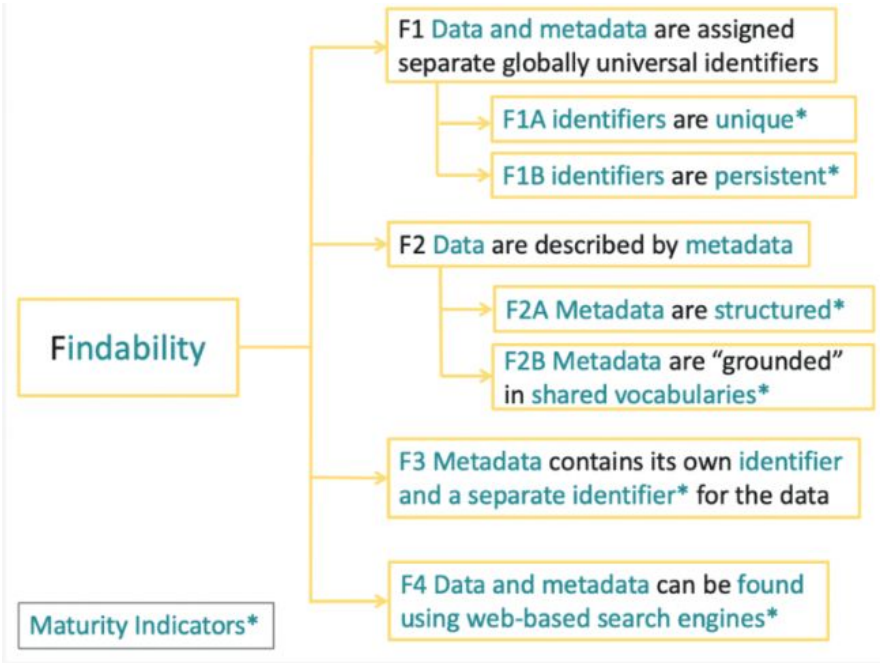
FAIR Toolkit

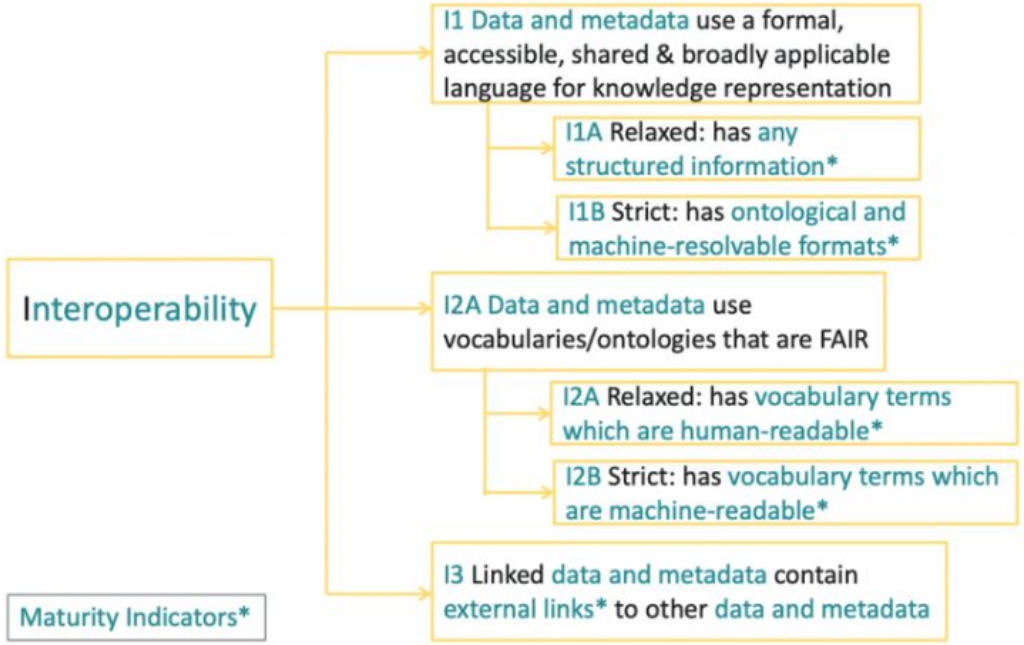
FAIR Toolkit

Pistoia Alliance

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

https://fairtoolkit.pistoiaalliance.org/





FAIR is about data *and* metadata

FAIR in action: Semantic Product Line

Scientific 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.

Semantic Interoperability Hub - Capability Stack

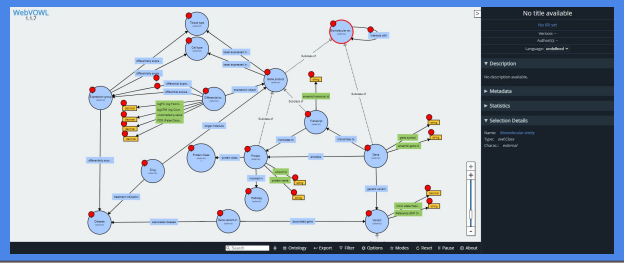
Data Management Value Chain - From Terminologies to a Unified Domain Model

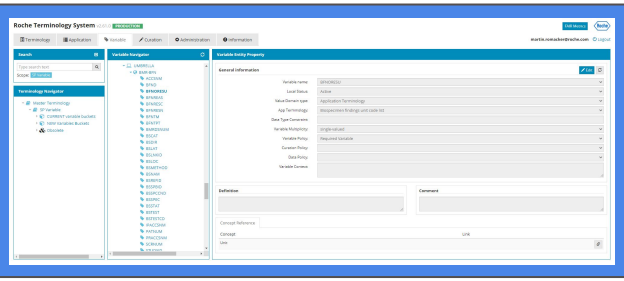
Terminology Managment (RTS)

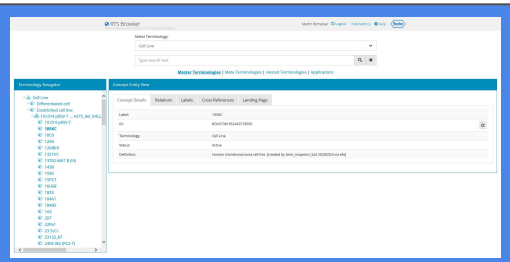
Dataset Model (MDR)

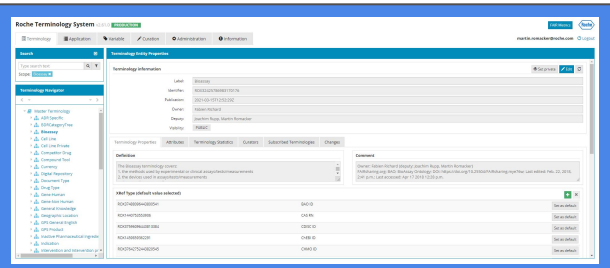
Conceptual Model

FAIR Unified Domain Model (KG)



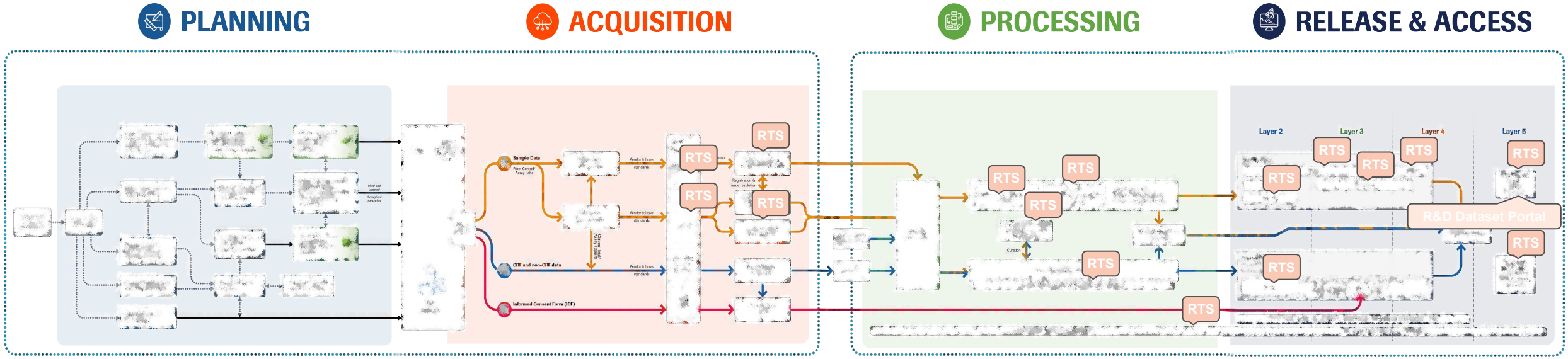






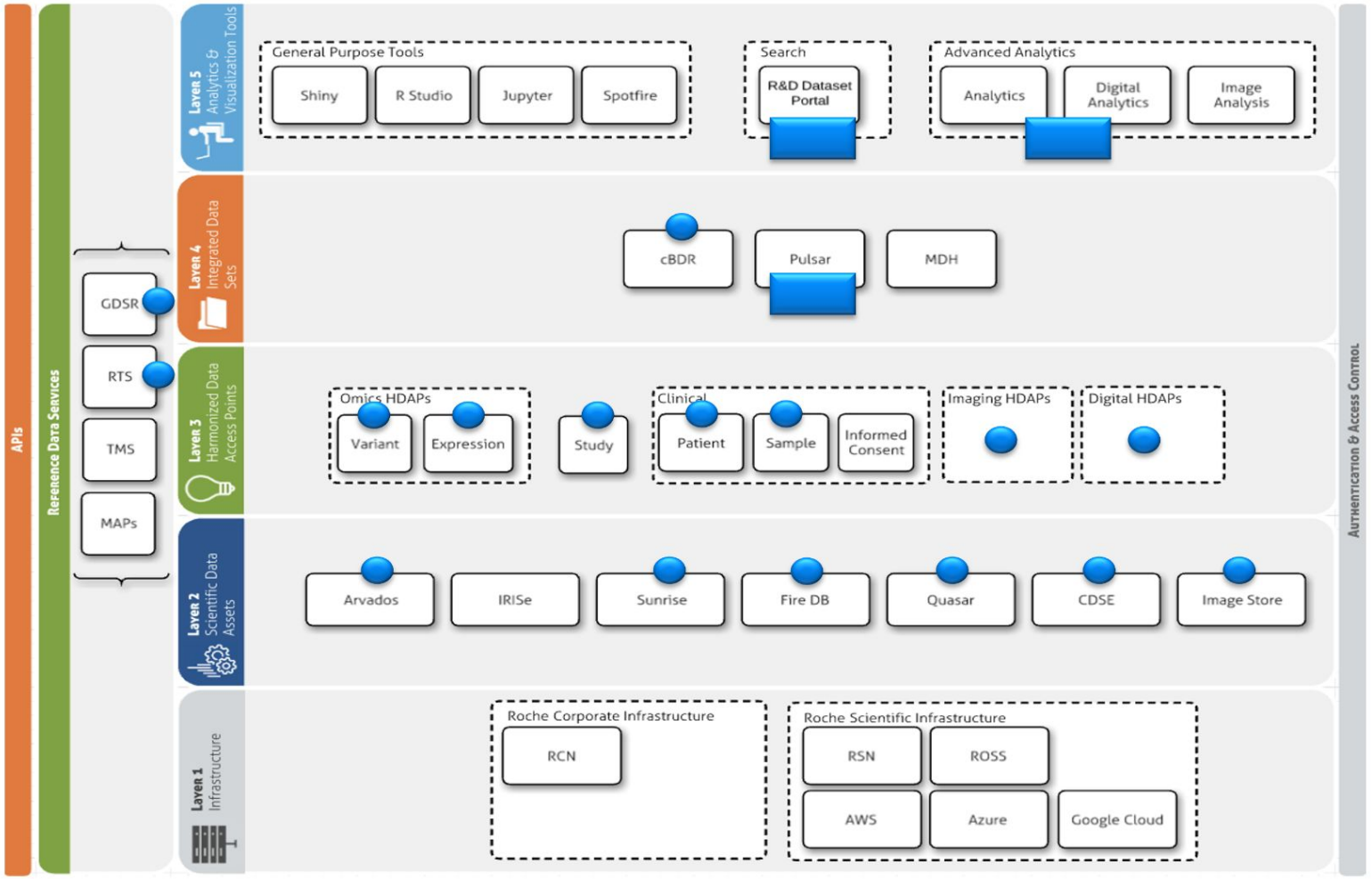
EDIS E2E Engine

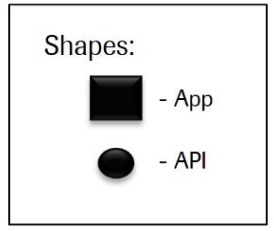
RTS Integration (born FAIR)



Roche Data Commons

Fully Integrated Transformationless FAIR Architecture (FAIR by Design)

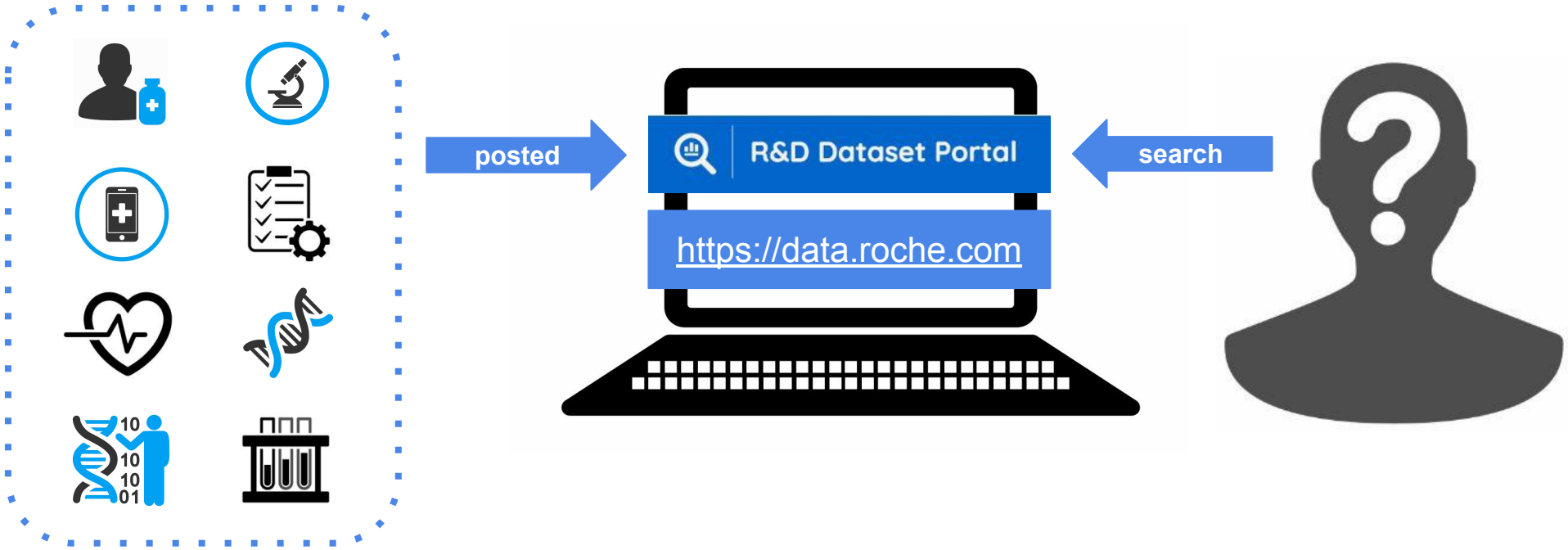




Transformationless Data Integration: Roche R&D Dataset portal

R&D Dataset Portal

Data Catalog of Data Catalogs



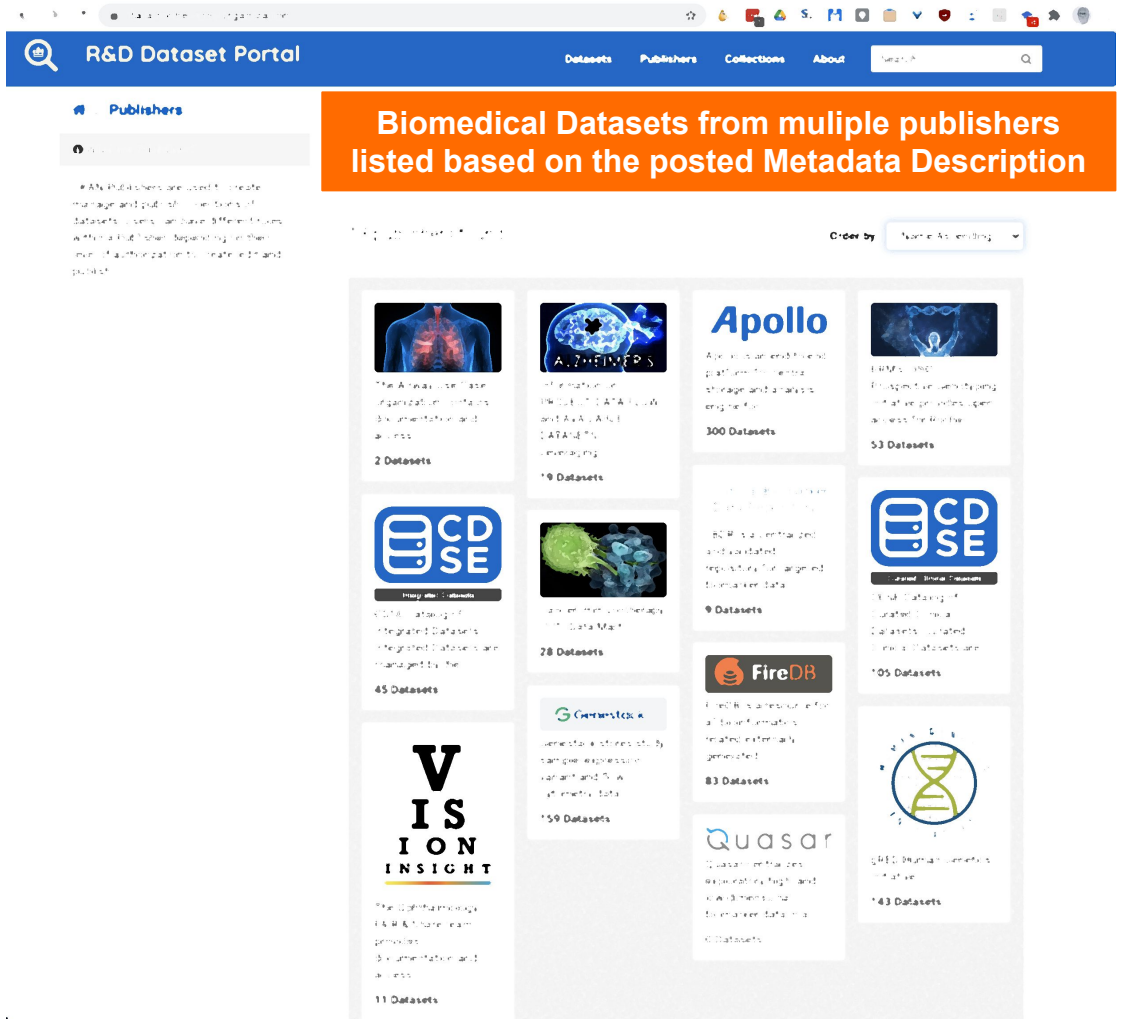
Biomedical datasets from **Roche R&D data catalogs** *e.g. biomarker, clinical, digital, imaging, omics or real world datasets*

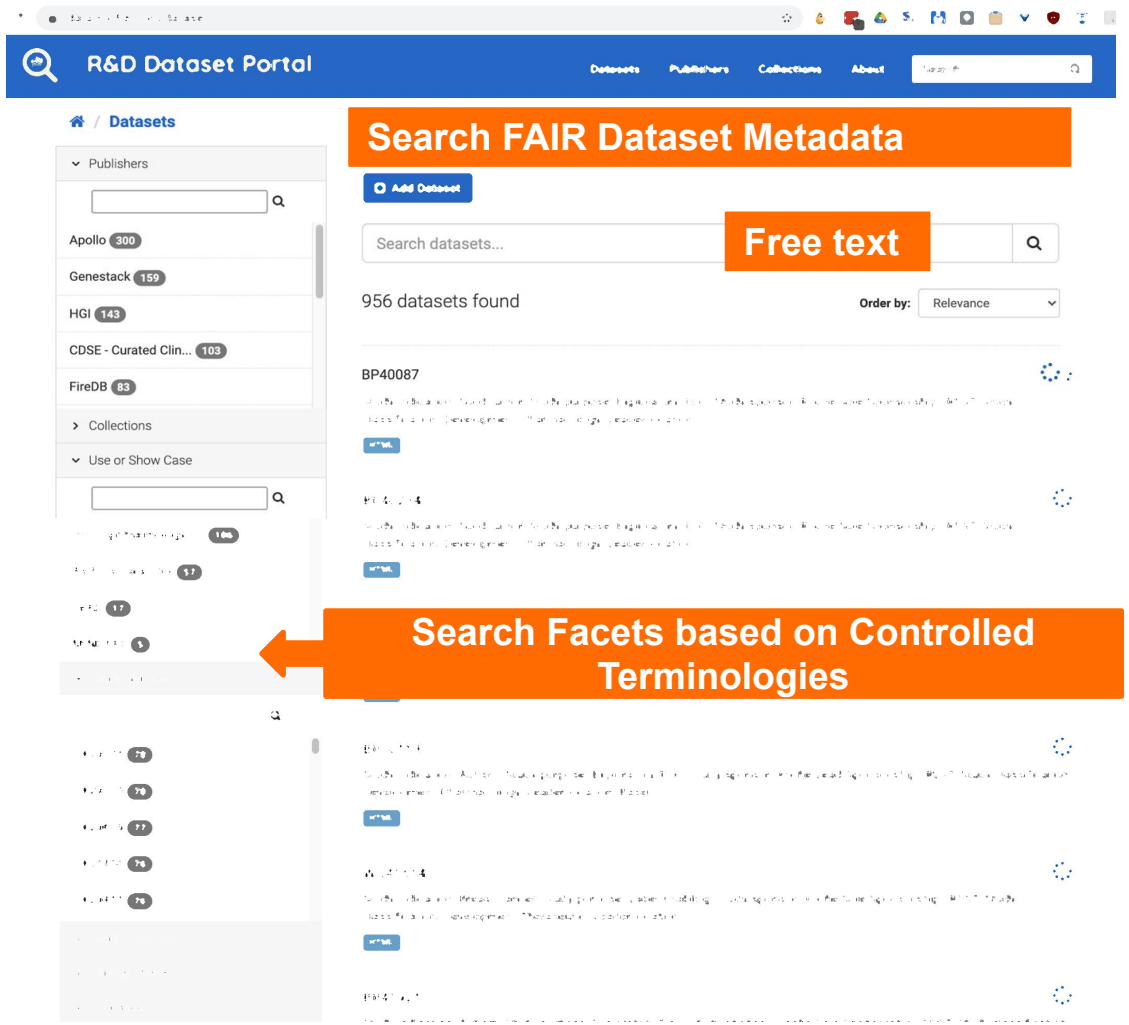
Cataloged and stored in source systems **published to the R&D Dataset Portal** as a central place to search & access corporate data assets

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

Roche Dataset Portal

Find Biomedical Datasets Across R&D





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)

Concept Entity View

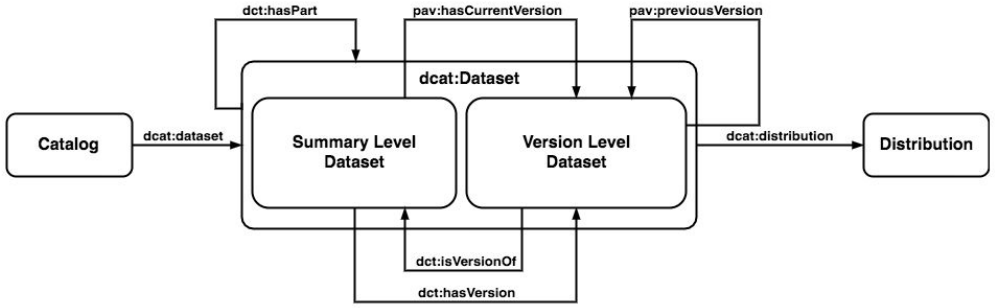
Concept Details | Relations | Labels

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

identifier.roche.com

language en


```
},
- 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,
        }
    ],
},
```

Terminology

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)

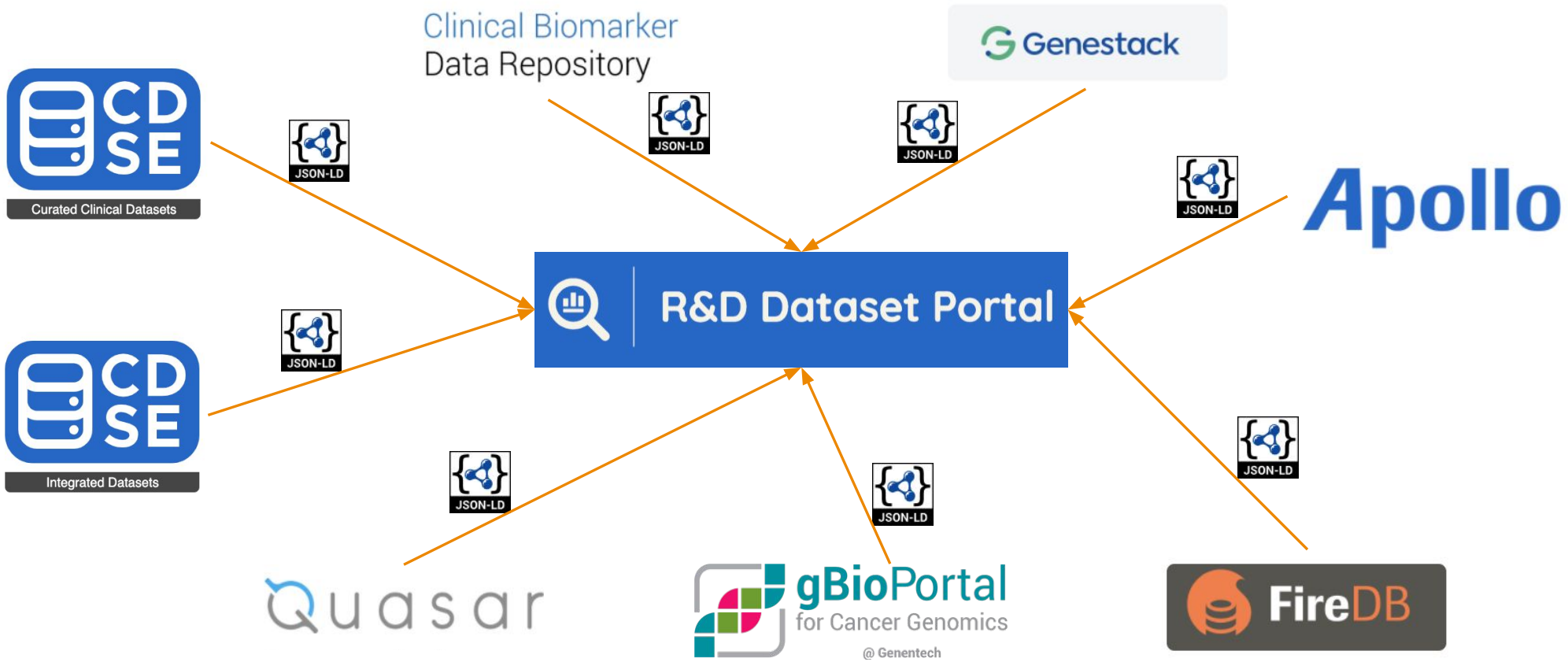
Published: 2021-03-10 00:43:15

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)

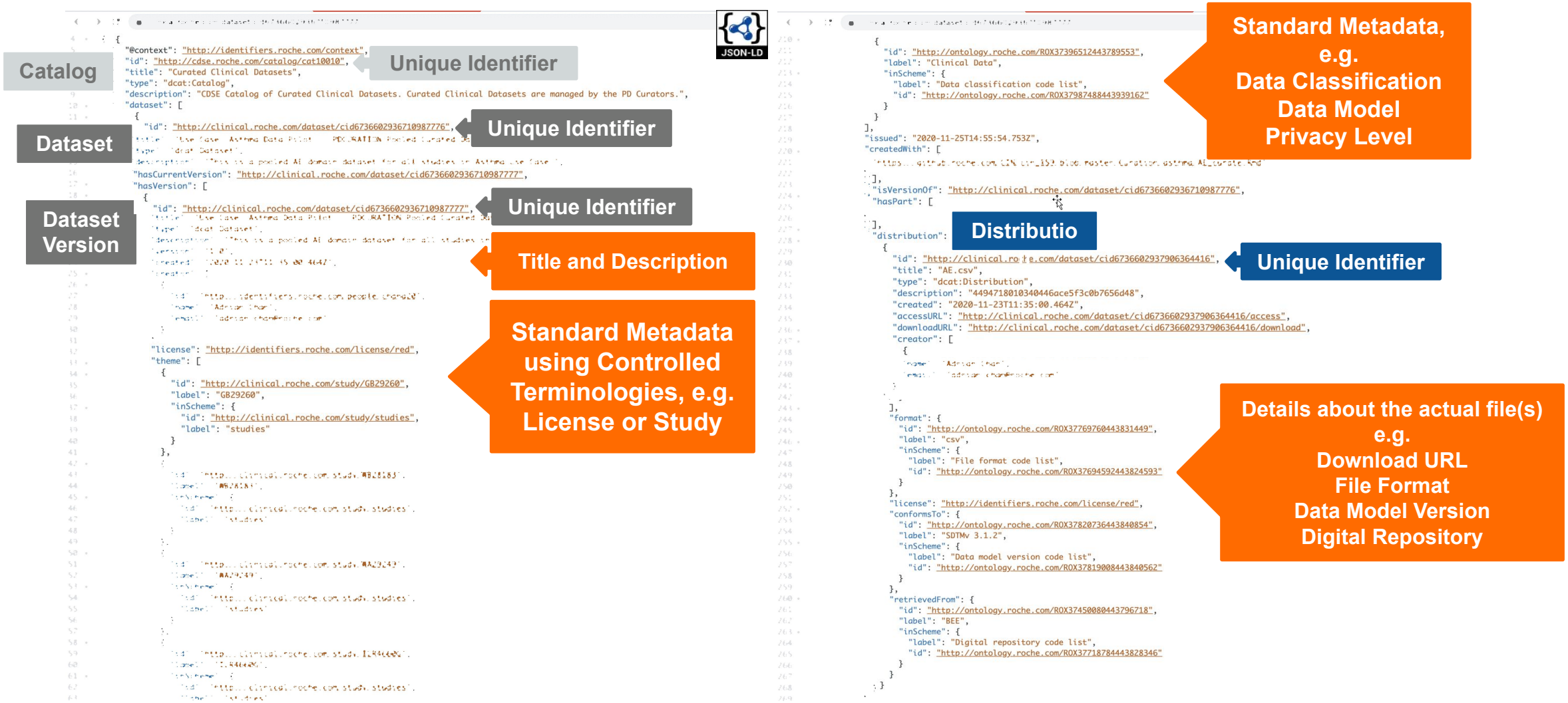
JSON-LD format specified in R&D Dataset Metadata Standards (Data contracts)



Transformationless data integration based on fully FAIRified machine-actionable data and dataset models (no data connctors)

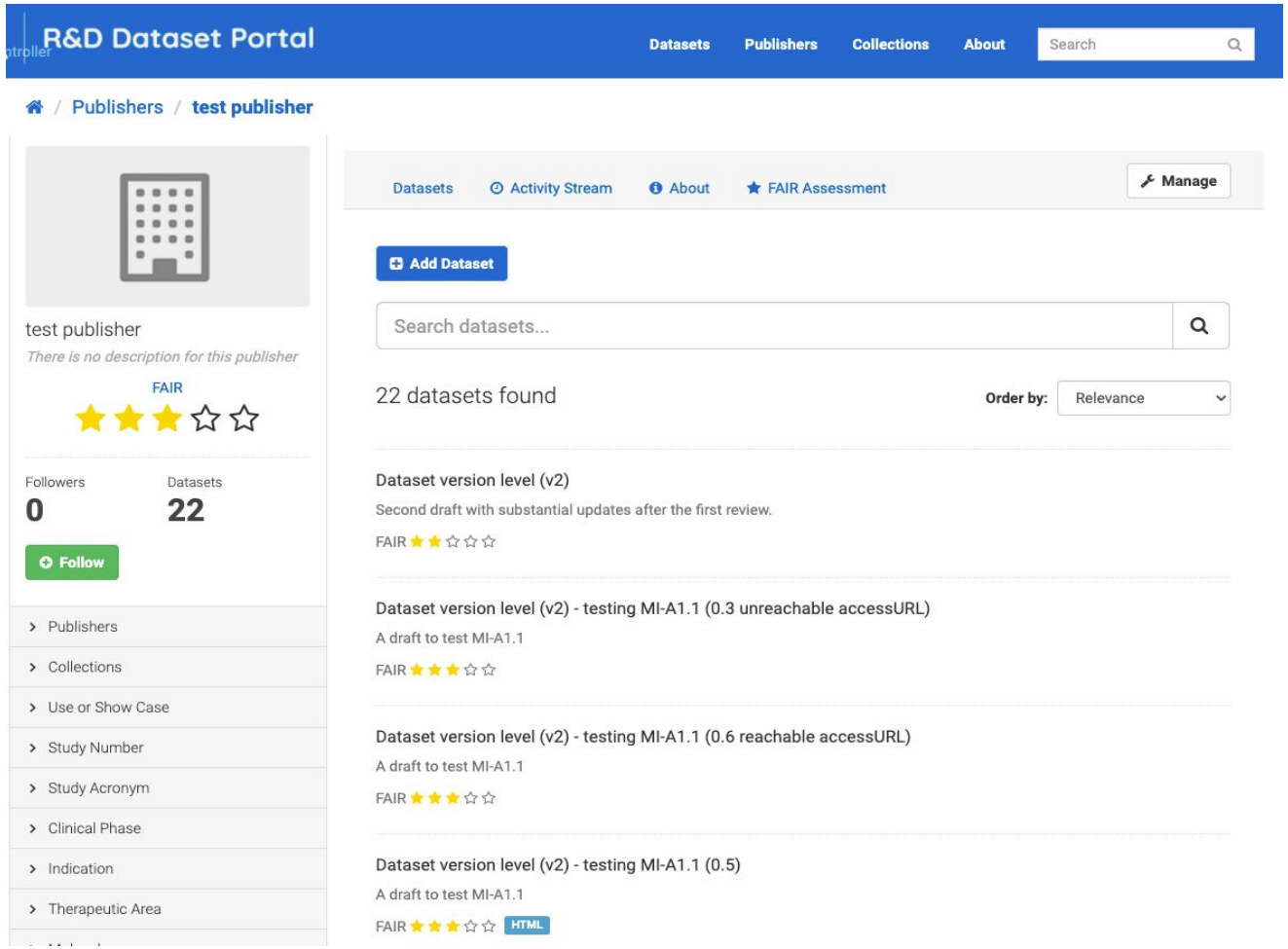
R&D Dataset Metadata

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

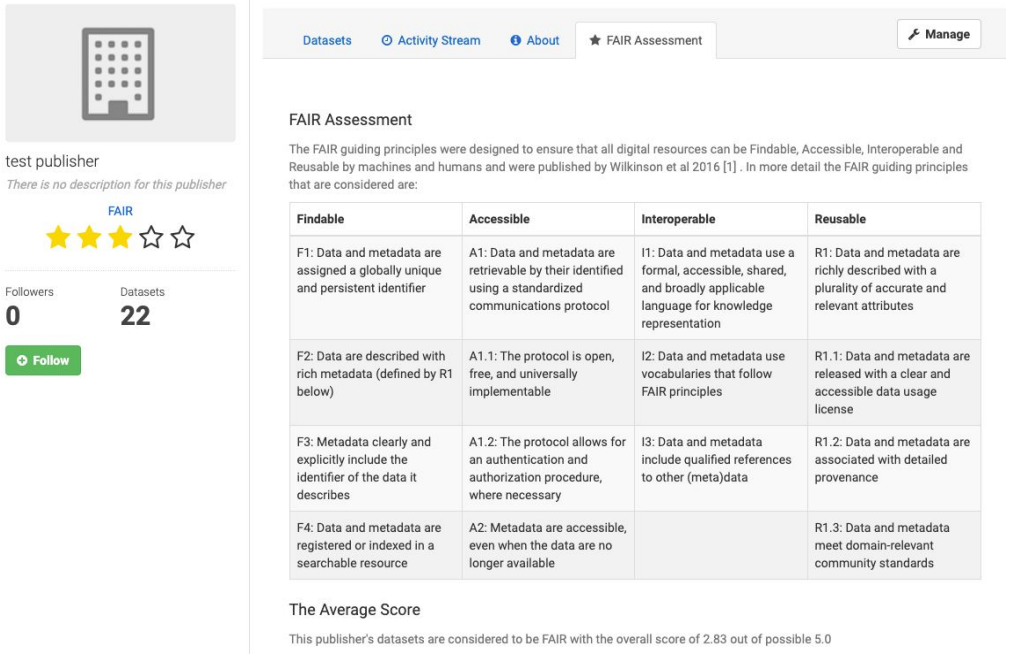


Roche Dataset Portal

Machine-Actionable Data - Automatic FAIR Assessment



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



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/ Data born FAIR) 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

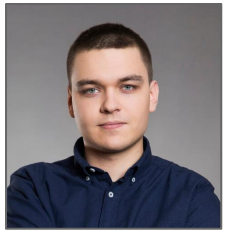
Michal Paradowski

Adam Zawada

Konrad Borowka

Robert Trypuz

Agnieszka Bananszynska-Krolikiewicz



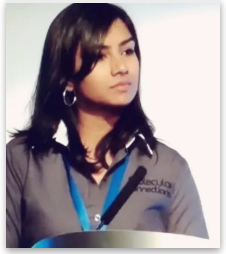
Majewski Krzysztof

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