



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

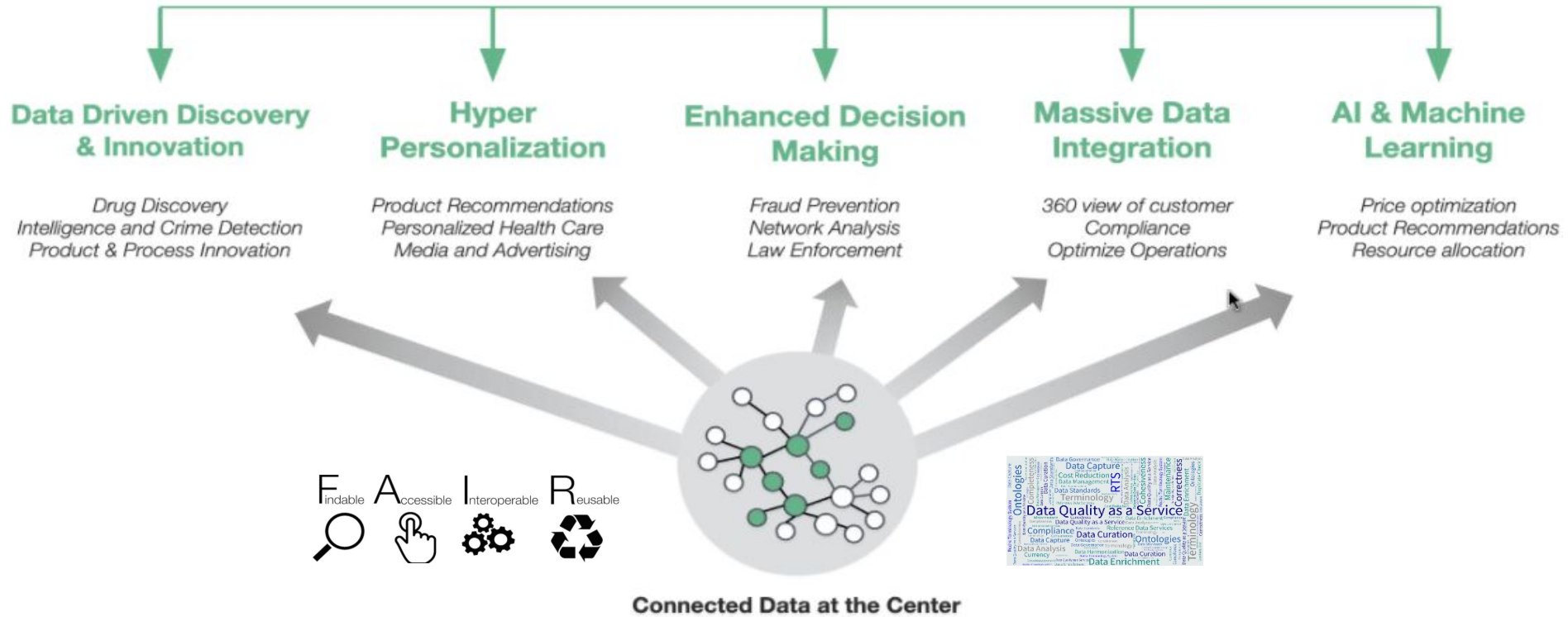
Table of contents

1. Business Case for FAIR & Data Quality
2. FAIR and Roche Data Commons
3. FAIR in action: Semantic Product Line
4. Transformationless Data Integration: Roche R&D Dataset portal
5. Conclusions
6. Acknowledgements

Business Case for FAIR and Data Quality

Harnessing Connections Drives Business Value

Digital Transformation Megatrends



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*

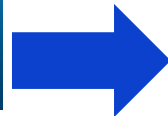


Planned/ Visible Costs

- 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

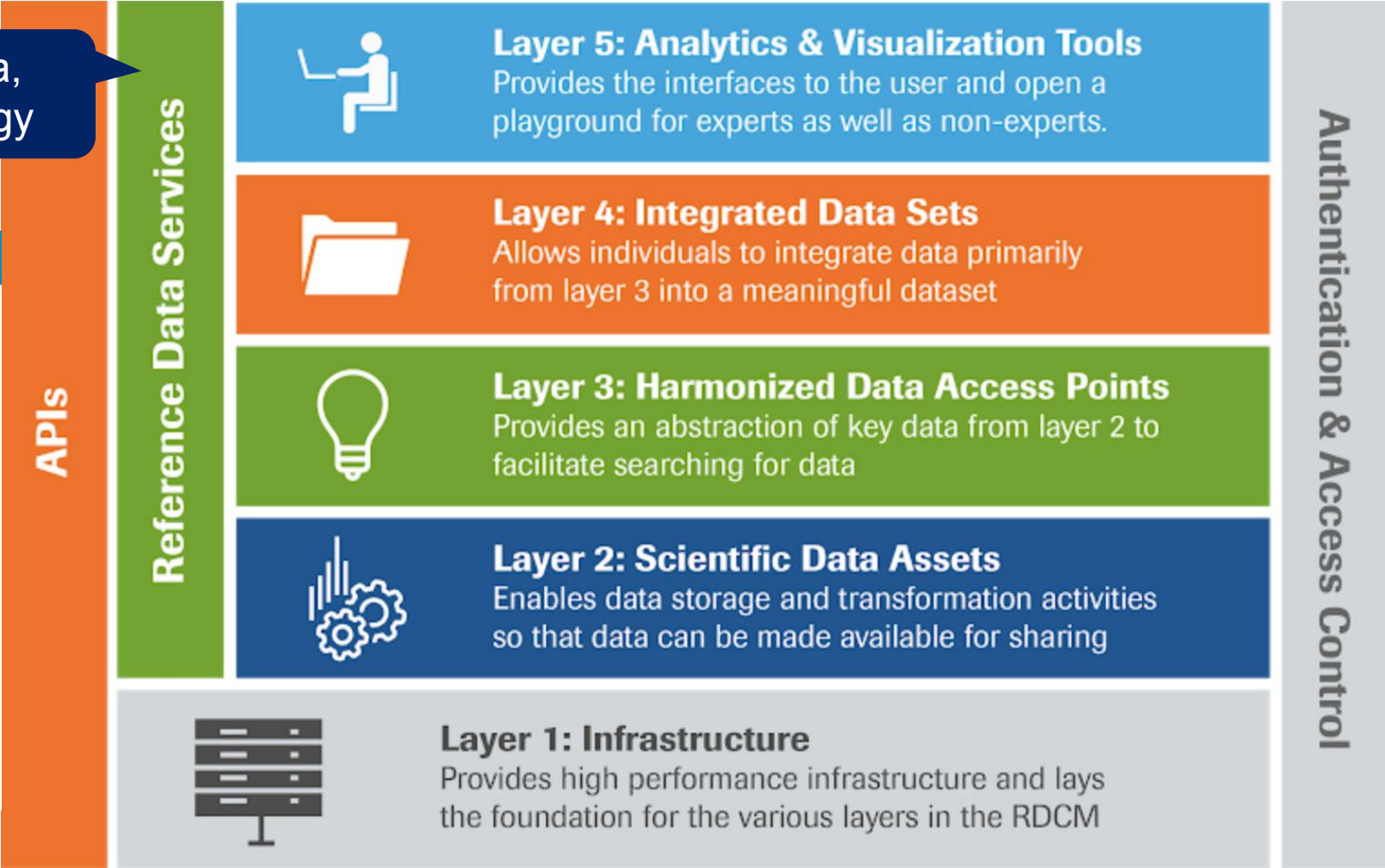


Backcharge the costs for processing to the data producers

FAIR and Roche Data Commons

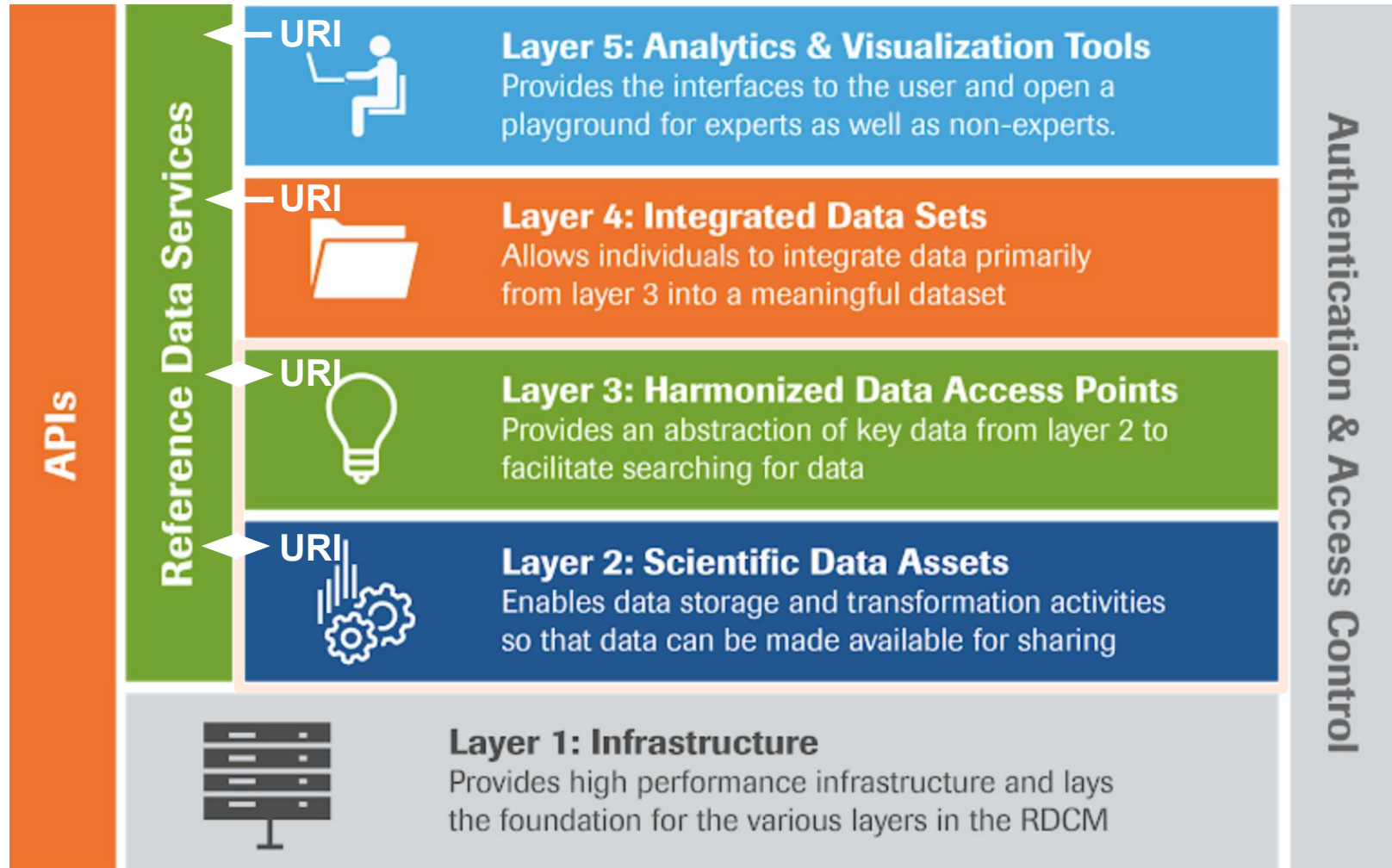
Terminology, Metadata, Dataset Model, Ontology

- 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



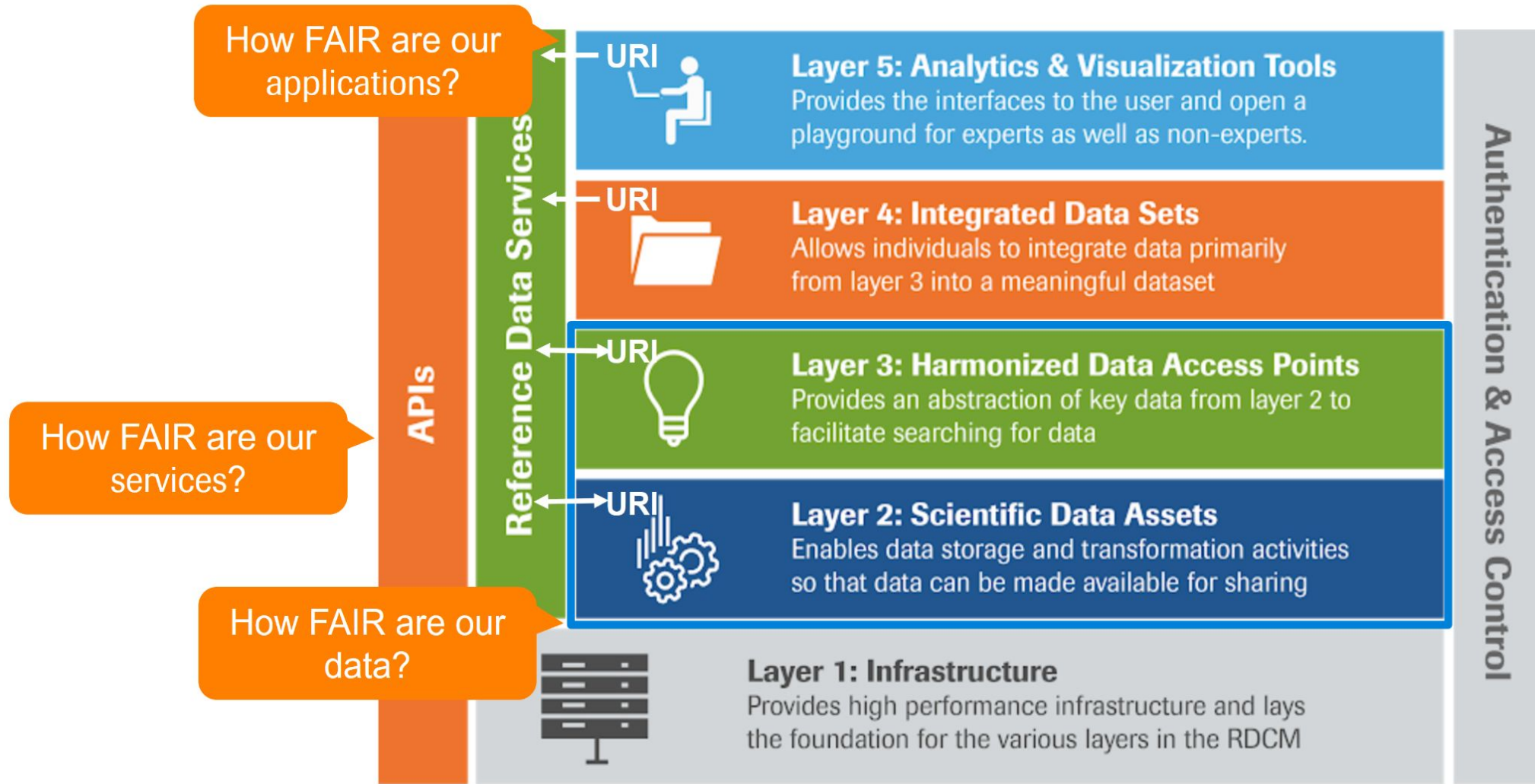
HDAPs organize data in Information Types Interoperability (URIs): semantic data dictionary
 semantic models
 Data FAIRification only in layer 2 & 3

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)

A red callout box with a white border and a pointer pointing towards the top navigation menu. It contains the text "FAIR Maturity" in white.

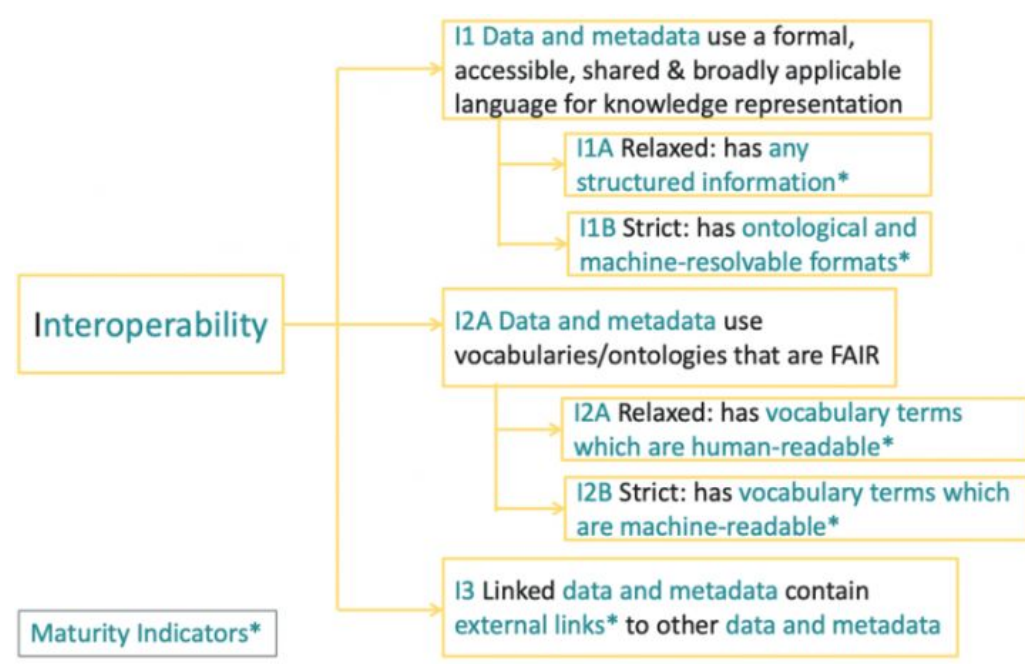
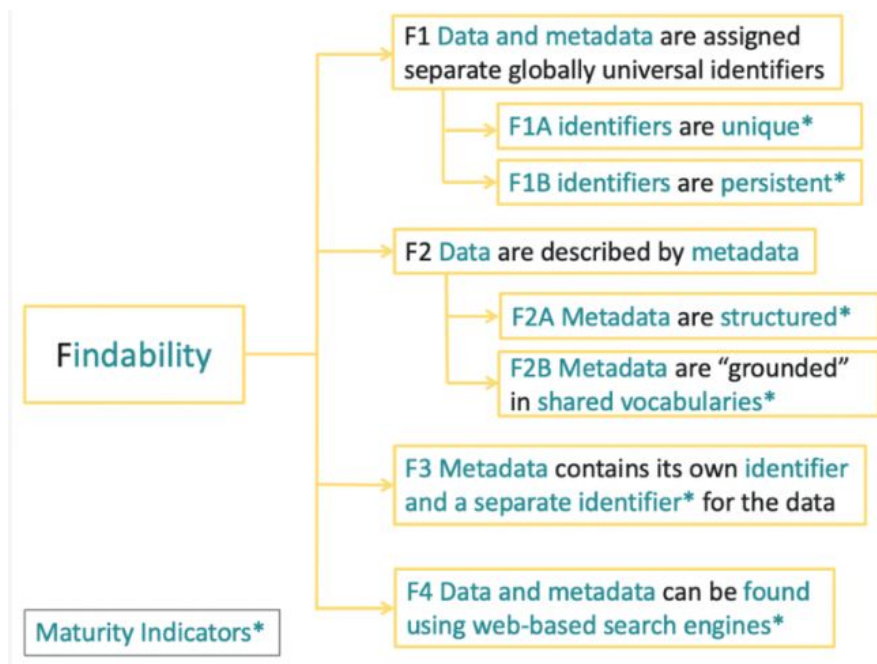
FAIR
Maturity

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



<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.



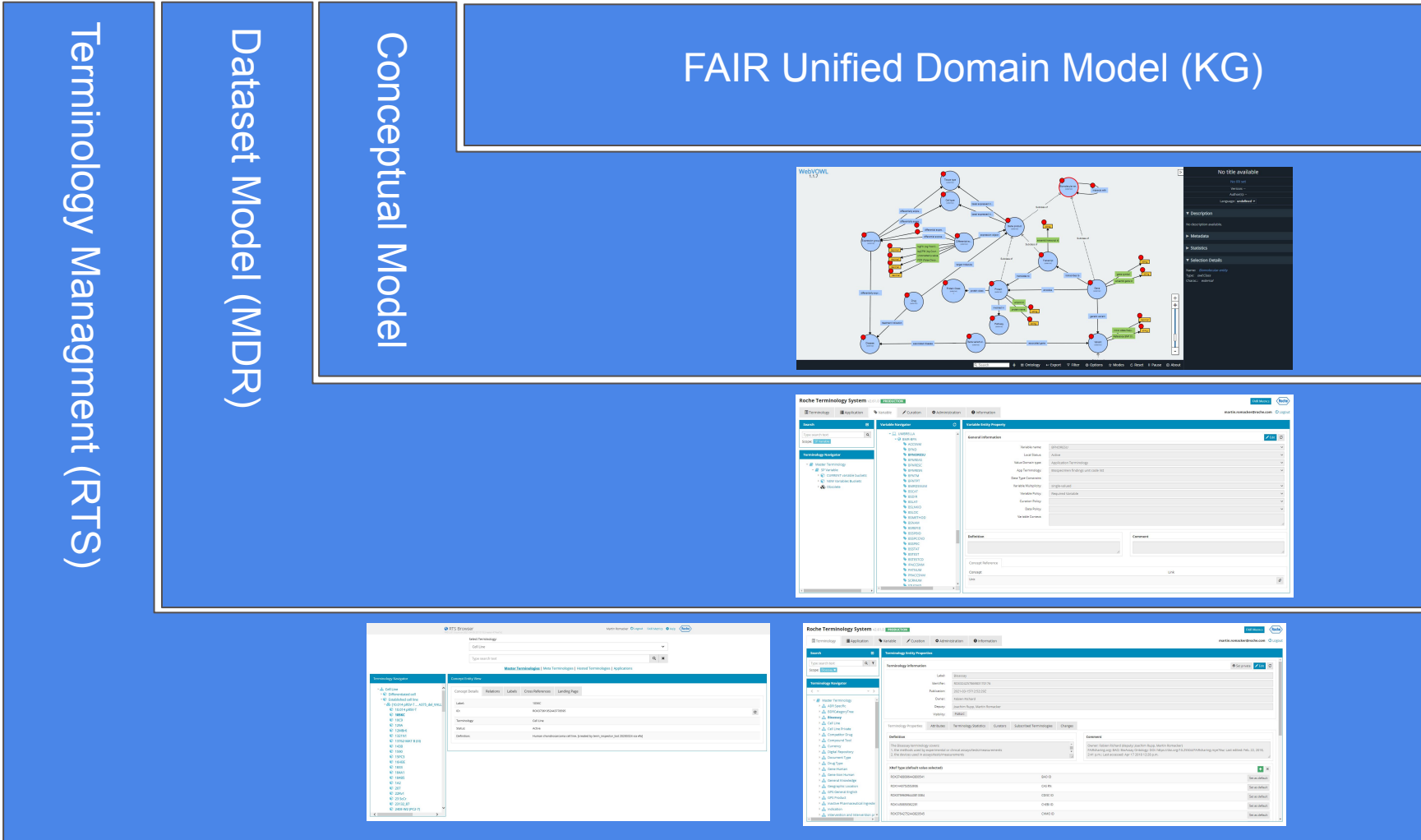
Products serve as reference data for standardized terminologies, metadata & conceptual models semantically linking internal and external data assets for data acquisition and data integration.



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



EDIS E2E Engine

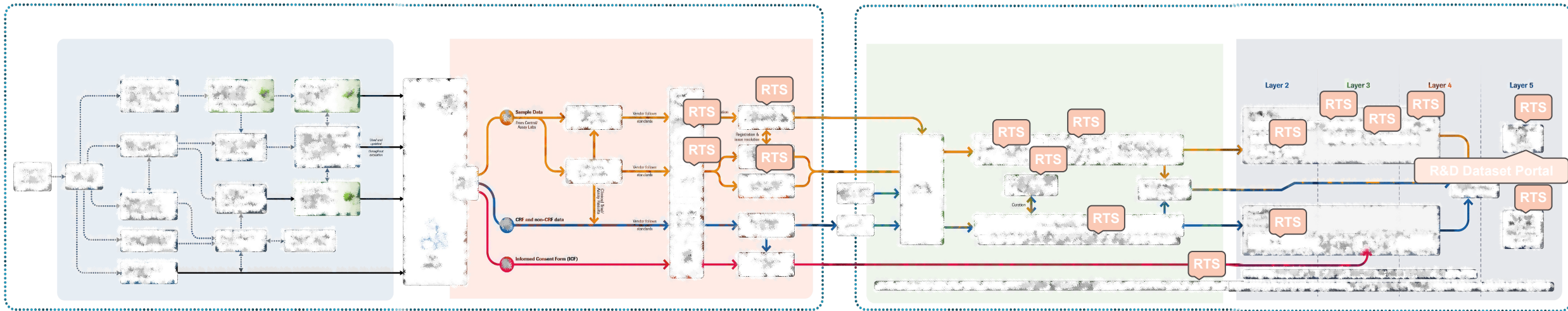
RTS Integration (born FAIR)

PLANNING

ACQUISITION

PROCESSING

RELEASE & ACCESS

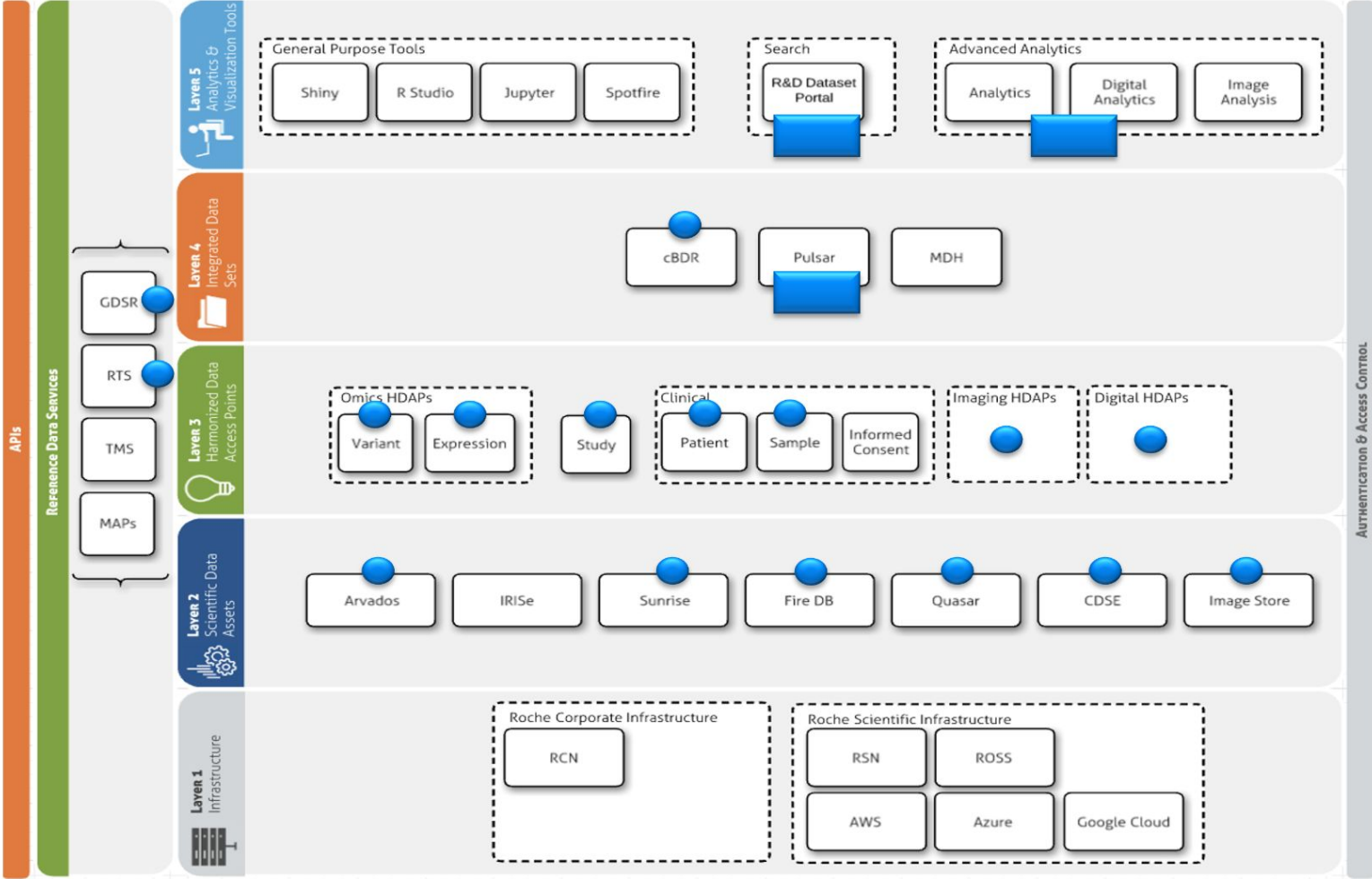


FOR ALL STUDIES

PRIMARY EXPLORATORY & SECONDARY REUSE

Roche Data Commons

Fully Integrated Transformationless FAIR Architecture (FAIR by Design)



Shapes:

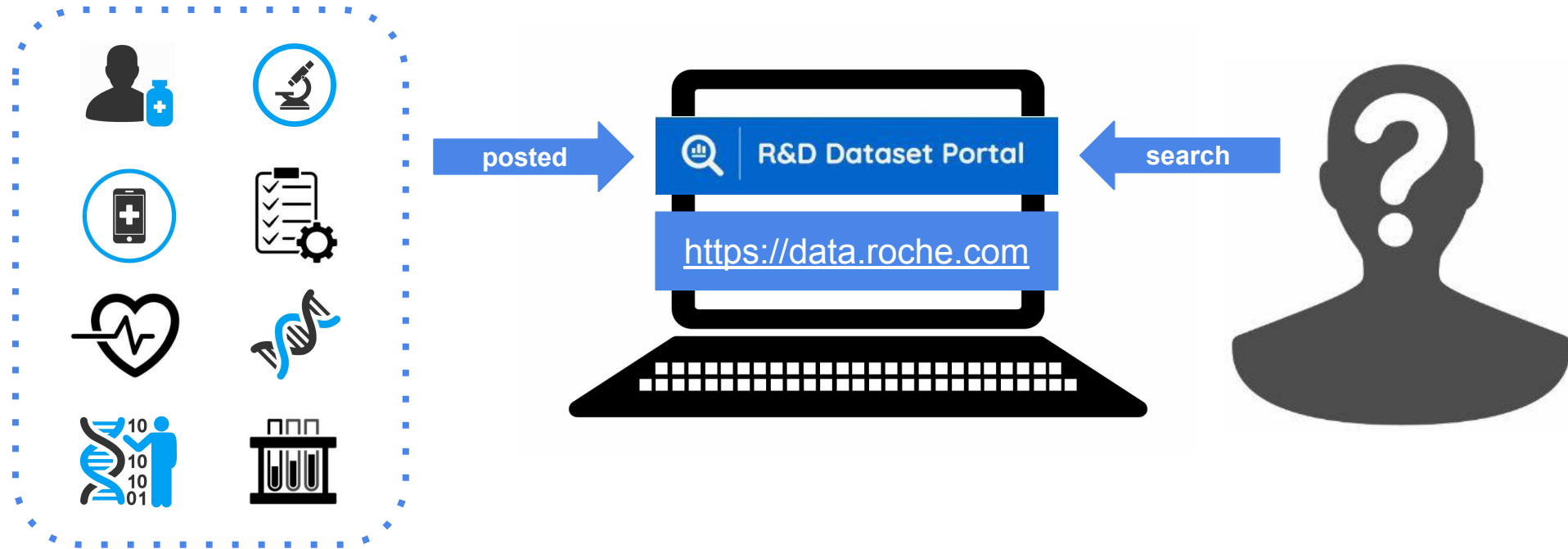
- App
- API



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

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

Order by: Name Ascending

- Apollo** (300 Datasets): Apollo is an extensible platform for meta-storage and analysis engine for...
 - 17k Analyses: Large organization contains 3k experimental and assays
 - 19 Datasets
- CDSE - Curated Clin...** (103 Datasets): CDSE is a centralized and validated repository for aligned biomarker data...
 - 28 Datasets
- FireDB** (83 Datasets): FireDB is a database for all biomarkers related externally generated...
 - 9 Datasets
- Quasar** (143 Datasets): Quasar provides experimental, high and low dimensional biomarker data in...
 - 11 Datasets

R&D Dataset Portal

Search FAIR Dataset Metadata

Free text search: Search datasets... [Free text] [Search]

956 datasets found

Order by: Relevance

Search Facets based on Controlled Terminologies

- Publishers**
 - Apollo: 300
 - Genestack: 159
 - HGI: 143
 - CDSE - Curated Clin...: 103
 - FireDB: 83
- Collections**
- Use or Show Case**

BP40087: 956 datasets found based on the search criteria... [Free text] [Search]

BP40087: 956 datasets found based on the search criteria... [Free text] [Search]

BP40087: 956 datasets found based on the search criteria... [Free text] [Search]

BP40087: 956 datasets found based on the search criteria... [Free text] [Search]

R&D Dataset Portal

FAIR Representation of Metadata & Data

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, \(gRED PS BiOmics\)](#)

Concept Entity View	
Concept Details	Relations
Label:	Pharma Informatics
ID:	ROX38029824443945995
Terminology:	Roche Organization
Status:	Active
Definition:	Pharma Informatics organization led by Steve Guise.

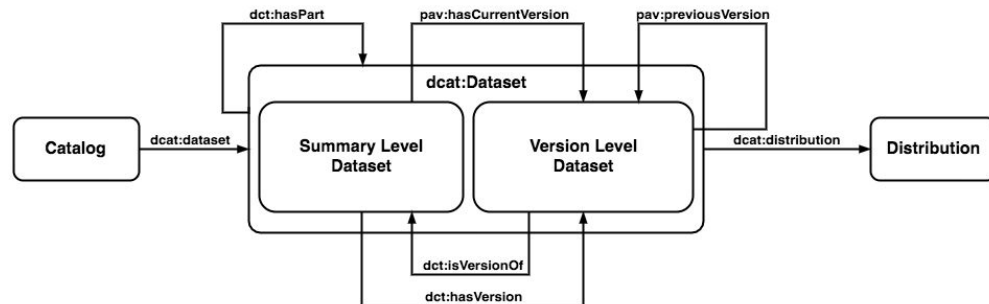
identifier.roche.com

```

},
  "response": {
    "countFound": 1,
    "start": 0,
    "docs": [
      {
        "user_defined_job_title": "Senior Principal Scientist",
        "preferred_last_first": "Romacker Martin",
        "unix_id": "romacker",
        "unix_id_gram": "romacke",
        "email": "martin.romacker@roche.com",
        "preferred_full_name": "Martin Romacker",
        "cost_center_number": "1005312300",
        "hire_date": "2013-01-01",
        "building": "992",
        "company_code": "1201",
        "id": "p780032",
        "type": "p",
        "office_phone": "+41 61 687 40 14",
        "manager_dn": "gned=exmjpknn,ou=people,dc=gene,dc=com",
        "manager_full_name": "Rupp, Joachim",
        "site": "R&D",
        "guid": "729032",
        "job_title": "Senior Principal Scientist",
        "manager_guid": "663886",
        "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": "gned=mgfssaga,ou=people,dc=gene,dc=com",
        "preferred_first_name": "Martin",
        "room": "06.NBH03",
        "_version_": 1700869243876147200,
      }
    ]
  }
}

```

language [en](#)



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

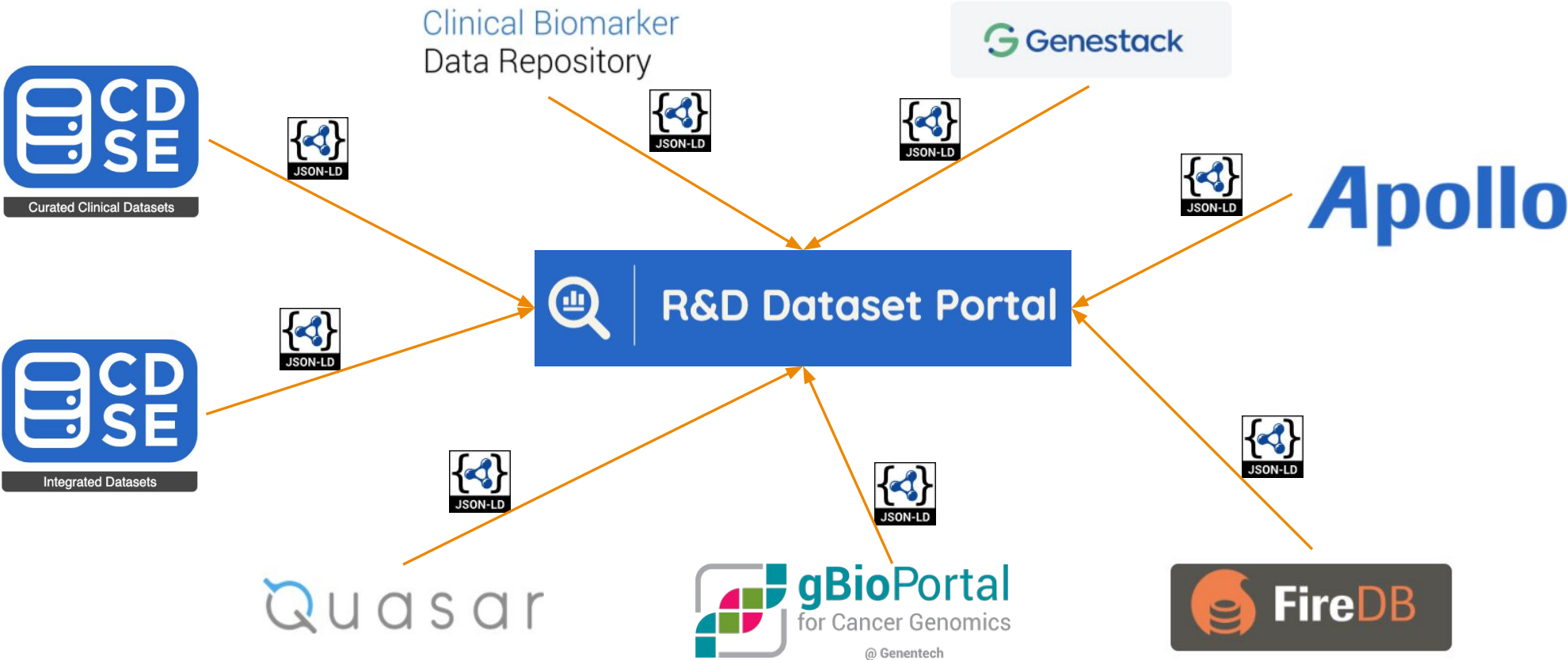
An ADaM dataset is a particular type of analysis dataset that either:

- (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 connectors)

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",
      "title": "Use Case - Asthma Data Point - POCURATION: Pooled Curated CDSE Clinical Dataset",
      "description": "This is a pooled AI domain dataset for all studies on Asthma use case.",
      "hasCurrentVersion": "http://clinical.roche.com/dataset/cid6736602936710987777",
      "hasVersion": [
        {
          "id": "http://clinical.roche.com/dataset/cid6736602936710987777",
          "title": "Use Case - Asthma Data Point - POCURATION: Pooled Curated CDSE Clinical Dataset",
          "description": "This is a pooled AI domain dataset for all studies on Asthma use case.",
          "created": "2020-11-23T11:35:00.464Z",
          "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"
              }
            },
            {
              "id": "http://clinical.roche.com/study/GB28183",
              "label": "GB28183",
              "inScheme": {
                "id": "http://clinical.roche.com/study/studies",
                "label": "studies"
              }
            },
            {
              "id": "http://clinical.roche.com/study/AA29249",
              "label": "AA29249",
              "inScheme": {
                "id": "http://clinical.roche.com/study/studies",
                "label": "studies"
              }
            },
            {
              "id": "http://clinical.roche.com/study/studies",
              "label": "studies"
            }
          ]
        }
      ]
    }
  ]
}

```

```

{
  "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": [
    {
      "id": "https://github.com/CDSE/ai_153/blob/master/Curation/asthma/AE/curated.Rd"
    }
  ],
  "isVersionOf": "http://clinical.roche.com/dataset/cid6736602936710987776",
  "hasPart": [
    {
      "distribution": {
        "id": "http://clinical.roche.com/dataset/cid6736602937906364416",
        "title": "AE.csv",
        "type": "dcat:Distribution",
        "description": "4494718018340446ace5f3c0b7656d48",
        "created": "2020-11-23T11:35:00.464Z",
        "accessURL": "http://clinical.roche.com/dataset/cid6736602937906364416/access",
        "downloadURL": "http://clinical.roche.com/dataset/cid6736602937906364416/download",
        "creator": [
          {
            "name": "Adrian Chohan",
            "email": "adrian.chohan@roche.com"
          }
        ],
        "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": "SDTM 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

Machine-Actionable Data - Automatic FAIR Assessment

R&D Dataset Portal | Datasets | Publishers | Collections | About | Search

Home / Publishers / test publisher

test publisher
There is no description for this publisher

FAIR ★★☆☆☆

Followers: 0 | Datasets: 22

[Follow](#)

Activity Stream | About | FAIR Assessment | Manage

Add Dataset

Search datasets... [Q]

22 datasets found | Order by: Relevance

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 ★★☆☆☆ [HTML](#)

- 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

[Follow](#)

Activity Stream | About | FAIR Assessment | Manage

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



Conclusions

Conclusions

- Successful and value-generating Digitilization requires true machine-actionable data, machine-readability is not sufficient. Application of FAIR principles is mandatory.
- FAIR data principles intrinsically tie Data Management to Semantic Technologies. (usage of terminologies, dataset definitions & ontologies)
- Transformationless data integration based on fully harmonized and standardized machine-actionable data assets (FAIR by design/ 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

Dataset portal team:

- Hugo de Schepper
- Oliver Steiner
- Roy Weiler



Felix Schwagereit

Scientific Technical Manager,
Basel



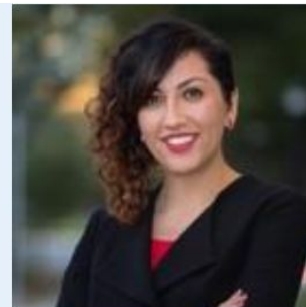
Pratishtha Duhan

Business Manager, SSF



Rama Balakrishnan

Biomedical Ontology Specialist,
SSF



Shima Dastgheib

Semantic Integrator, SSF

Roche Terminology System

Dev and Ops Team, Curation Team

RTS Dev and OPS Team



Michal Bielak



Michal Paradowski



Adam Zawada



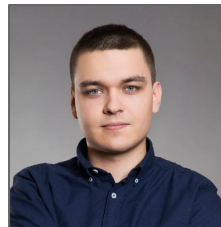
Konrad Borowka



Robert Trypuz



**Agnieszka
Banaszynska-
Krolikiewicz**



**Majewski
Krzysztof**

Additional Members:

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

Molecular Connections Team



Arathi Raghunath
Technical & Project
Lead for Roche



Krishna K Chinnaiah
Business & Account
Manager for Roche

Curation supported by:

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

Rancho Biosciences Team



Erfan Younesi
Sr. Data Curator



Svetlana Koltsova
Sr. Data Curator



Maxim Papin
Sr. Data Curator

Doing now what patients need next