Data First!

FAIR Principles Implementation at Roche
A Pharma Perspective

Martin Romacker & Nick Perry
Product Managers
Enterprise Data
Digital Integrations Generating Insights (DIGI)
Roche Informatics, Basel

06 April 2023, OntoCommons Workshop, Berlin
Table of contents

1. Data Economics
2. FAIR & Roche Data Commons
3. FAIR is simple and beautiful
4. FAIR is complex and ugly
5. Pharma Interoperability Hub
6. Conclusions
7. Acknowledgements
Data Economics
Megatrends in Digitilization
FAIR plus Q Data

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

Connected Data at the Center

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

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

Source: Rik van Bruggen, Neo4J
Planned and unplanned Costs in Data Management

Business Case for Prospective FAIRification and High Data Quality

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

Unplanned/ Invisible Costs
- Business Analysis
- ETL processes/ Data Cleansing
- Searching & accessing
- Data Curation/ Semantic Data Integration
- IT Infrastructure supporting unplanned activities

Backcharge the costs for processing to the data producers
Standards in Pharma Industry
An open public-private semantic infrastructure of fully standardized FAIR applications, services & data
Roche Data Commons
From Application-Centric to Information-Centric

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

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

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

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

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

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

HDAPs organize data in Information Types

Interoperability (URIs): semantic data dictionary semantic models

Data FAIRification only in layer 2 & 3

No more transformation between layer 3 & 4,5
Roche Data Commons
Semantic Infrastructure of FAIR Applications, Services & Data

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

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

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

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

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

How FAIR are our applications?

How FAIR are our services?

How FAIR are our data?
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 is about data *and* metadata

https://fairtoolkit.pistoiaalliance.org/
Standards in Pharma Industry
FAIR scientific data management

FAIR guiding principles

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

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

by Olivier Roche (pREDi)
### FAIR data architecture

**Terminology & Concepts**

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Sex</th>
<th>Age Value</th>
<th>Age Unit</th>
<th>Substance Name</th>
<th>Dose Value</th>
<th>Dose Unit</th>
<th>Dosage Form</th>
<th>Route</th>
<th>Dosage Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>S32345</td>
<td>male</td>
<td>230</td>
<td>day</td>
<td>bevacizumab</td>
<td>50</td>
<td>mg/ml</td>
<td>tablet</td>
<td>oral</td>
<td>daily</td>
</tr>
<tr>
<td>S93784</td>
<td>female</td>
<td>year</td>
<td>oseltamivir</td>
<td>100</td>
<td>mg/l</td>
<td>solution</td>
<td>intravenous</td>
<td>weekly</td>
<td></td>
</tr>
<tr>
<td>S11324</td>
<td>male</td>
<td>88</td>
<td>year</td>
<td>trastuzumab</td>
<td>35</td>
<td>g/l</td>
<td>solution</td>
<td>subcutaneous</td>
<td>hourly</td>
</tr>
<tr>
<td>S92833</td>
<td>unknown</td>
<td>1</td>
<td>year</td>
<td>MabThera</td>
<td>50</td>
<td>umol/L</td>
<td>solution</td>
<td>intraocular</td>
<td>hourly</td>
</tr>
<tr>
<td>S33021</td>
<td>female</td>
<td>11</td>
<td>month</td>
<td>Polivy</td>
<td>12.5</td>
<td>mg/ml</td>
<td>inhalant</td>
<td>intrabronchial</td>
<td>weekly</td>
</tr>
</tbody>
</table>

**Concept**

- ROX1380015111414
## FAIR data architecture

### Terminology & Concepts

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Sex</th>
<th>Age Value</th>
<th>Age Unit</th>
<th>Substance Name</th>
<th>Dose Value</th>
<th>Dose Unit</th>
<th>Dosage Form</th>
<th>Route</th>
<th>Dosage Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>S32345</td>
<td>male</td>
<td>230</td>
<td>day</td>
<td>bevacizumab</td>
<td>50</td>
<td>mg/ml</td>
<td>tablet</td>
<td>oral</td>
<td>daily</td>
</tr>
<tr>
<td>S93784</td>
<td>female</td>
<td>11</td>
<td>year</td>
<td>oseltamivir</td>
<td>100</td>
<td>mg/l</td>
<td>solution</td>
<td>intravenous</td>
<td>weekly</td>
</tr>
<tr>
<td>S11324</td>
<td>male</td>
<td>88</td>
<td>year</td>
<td>trastuzumab</td>
<td>35</td>
<td>g/l</td>
<td>solution</td>
<td>subcutaneous</td>
<td>hourly</td>
</tr>
<tr>
<td>S92833</td>
<td>unknown</td>
<td>1</td>
<td>year</td>
<td>MabThera</td>
<td>50</td>
<td>umol/L</td>
<td>solution</td>
<td>intraocular</td>
<td>hourly</td>
</tr>
<tr>
<td>S33021</td>
<td>female</td>
<td>11</td>
<td>month</td>
<td>Polivy</td>
<td>12.5</td>
<td>mg/ml</td>
<td>inhalant</td>
<td>intrabronchial</td>
<td>weekly</td>
</tr>
</tbody>
</table>
## FAIR data architecture

### Variable & Schema

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Sex</th>
<th>Age Value</th>
<th>Age Unit</th>
<th>Substance Name</th>
<th>Dose Value</th>
<th>Dose Unit</th>
<th>Dosage Form</th>
<th>Route</th>
<th>Dosage Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>S32345</td>
<td>male</td>
<td>230</td>
<td>day</td>
<td>bevacizumab</td>
<td>50</td>
<td>mg/ml</td>
<td>tablet</td>
<td>oral</td>
<td>daily</td>
</tr>
<tr>
<td>S93784</td>
<td>female</td>
<td>11</td>
<td>year</td>
<td>oseltamivir</td>
<td>100</td>
<td>mg/l</td>
<td>solution</td>
<td>intravenous</td>
<td>weekly</td>
</tr>
<tr>
<td>S11324</td>
<td>male</td>
<td>88</td>
<td>year</td>
<td>trastuzumab</td>
<td>35</td>
<td>g/l</td>
<td>solution</td>
<td>subcutaneous</td>
<td>hourly</td>
</tr>
<tr>
<td>S92833</td>
<td>unknown</td>
<td>1</td>
<td>year</td>
<td>MabThera</td>
<td>50</td>
<td>umol/L</td>
<td>solution</td>
<td>intraocular</td>
<td>hourly</td>
</tr>
<tr>
<td>S33021</td>
<td>female</td>
<td>11 month</td>
<td>Polivy</td>
<td>12.5</td>
<td>mg/ml</td>
<td>inhalant</td>
<td>intrabronchial</td>
<td>weekly</td>
<td></td>
</tr>
</tbody>
</table>
## FAIR data architecture

### Variable & Schema

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Sex</th>
<th>Age Value</th>
<th>Age Unit</th>
<th>Substance Name</th>
<th>Dose Value</th>
<th>Dose Unit</th>
<th>Dosage Form</th>
<th>Route</th>
<th>Dosage Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>S32345</td>
<td>male</td>
<td>230</td>
<td>day</td>
<td>bevacizumab</td>
<td>50</td>
<td>mg/ml</td>
<td>tablet</td>
<td>oral</td>
<td>daily</td>
</tr>
<tr>
<td>S93784</td>
<td>female</td>
<td>11</td>
<td>year</td>
<td>oseltamivir</td>
<td>100</td>
<td>mg/l</td>
<td>solution</td>
<td>intravenous</td>
<td>weekly</td>
</tr>
<tr>
<td>S11324</td>
<td>male</td>
<td>88</td>
<td>year</td>
<td>trastuzumab</td>
<td>35</td>
<td>g/l</td>
<td>solution</td>
<td>subcutaneous</td>
<td>hourly</td>
</tr>
<tr>
<td>S92833</td>
<td>unknown</td>
<td>1</td>
<td>year</td>
<td>MabThera</td>
<td>50</td>
<td>umol/L</td>
<td>solution</td>
<td>intraocular</td>
<td>hourly</td>
</tr>
<tr>
<td>S33021</td>
<td>female</td>
<td>11</td>
<td>month</td>
<td>Polivy</td>
<td>12.5</td>
<td>mg/ml</td>
<td>inhalant</td>
<td>intrabronchial</td>
<td>weekly</td>
</tr>
</tbody>
</table>
### FAIR data architecture

**Simple Graph Generation**

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

- S32345 - hasSex - male
- S32345 - ageValue - 230
- S32345 - ageUnit - day
FAIR data architecture

Semantic Graph Generation
### FAIR data architecture

#### Semantic Graph Generation

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Sex</th>
<th>Age Value</th>
<th>Age Unit</th>
<th>Substance Name</th>
<th>Dose Value</th>
<th>Dose Unit</th>
<th>Dosage Form</th>
<th>Route</th>
<th>Dosage Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>S32345</td>
<td>male</td>
<td>230</td>
<td>day</td>
<td>bevacizumab</td>
<td>50</td>
<td>mg/ml</td>
<td>tablet</td>
<td>oral</td>
<td>daily</td>
</tr>
</tbody>
</table>
FAIR is ugly and complex
FAIRsharing Catalog of Biomedical Resources
Proliferation and Fragmentation of Standards

Manually done—no smart interfaces

1299 entries for Standards
Interoperability - Standards
Clinical Data vs FHIR

```
"name": {
  "extension": [
    {
      "url": "http://hl7.org/fhir/StructureDefinition/humanname-assembly-order",
      "valueCode": "NLL"
    }
  ],
  "use": "official",
  "family": "001",
  "extension": [
    {
      "url": "http://hl7.org/fhir/StructureDefinition/humanname-own-name",
      "valueString": "001"
    }
  ]
},
"gender": "male",
"birthdate": "1975-06-12",
"deceasedBoolean": false
```

IF deceasedBoolean = “false” THEN SurvivalStatus = “alive”
Data Standards & Interoperability Challenges
CDISC vs OMOP/ OHDSI

Creation of insights & analytics blocked: different model, variables and values
Data Standards & Interoperability Challenges
CDISC vs OMOP/ OHDSI vs DICOM

DICOM Standard Browser by Innovitics

<table>
<thead>
<tr>
<th>Domain</th>
<th>Name</th>
<th>STUOID</th>
<th>STUOID</th>
<th>DOMAIND</th>
<th>SUBIID</th>
<th>SUBIID</th>
<th>SUBIID</th>
<th>SUBIID</th>
<th>CIOD</th>
<th>CIOD</th>
<th>CIOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Patient</td>
<td>(0008,1120)</td>
<td>Referenced Patient Sequence</td>
<td>3</td>
<td>Sequence</td>
<td>with date of birth of the event in seconds</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0010)</td>
<td>Patient’s Name</td>
<td>2</td>
<td>Person Name</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0020)</td>
<td>Patient ID</td>
<td>2</td>
<td>Long String</td>
<td>so that provide and stored in</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0021)</td>
<td>Issuer of Patient ID</td>
<td>3</td>
<td>Long String</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0022)</td>
<td>Type of Patient ID</td>
<td>3</td>
<td>Code String</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0024)</td>
<td>Issuer of Patient ID Qualifiers Sequence</td>
<td>3</td>
<td>Sequence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0026)</td>
<td>Source Patient Group Identification Sequence</td>
<td>3</td>
<td>Sequence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0027)</td>
<td>Group of Patients Identification Sequence</td>
<td>3</td>
<td>Sequence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0030)</td>
<td>Patient’s Birth Date</td>
<td>2</td>
<td>Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0032)</td>
<td>Patient’s Birth Time</td>
<td>3</td>
<td>Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0033)</td>
<td>Patient’s Birth Date In Alternative Calendar</td>
<td>3</td>
<td>Long String</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0034)</td>
<td>Patient’s Death Date In Alternative Calendar</td>
<td>3</td>
<td>Long String</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0035)</td>
<td>Patient’s Alternative Calendar</td>
<td>1C</td>
<td>Code String</td>
<td>It appears in a standard</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0040)</td>
<td>Patient’s Sex</td>
<td>2</td>
<td>Code String</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0200)</td>
<td>Quality Control Subject</td>
<td>3</td>
<td>Code String</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0010,0212)</td>
<td>Strain Description</td>
<td>3</td>
<td>Unlimited Characters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Creation of insights & analytics blocked: different model, variables and values
EMBL-EBI Ontology Xref Service
Creating referential identity by ontology mapping

1. Allocating significant resources to inflate a problem
2. Allocating significant resources to reduce a problem (loss of information & interoperability)

Source: EMBL-EBI
OxO
Interoperability for Ontology Mappings

RDF standard for a FAIR representation of OM

A Simple Standard for Sharing Ontology Mappings (SSSOM)

A Simple Standard for Sharing Ontology Mappings (SSSOM)

SSSOM is a simple metadata standard for describing semantic mappings:

1. Introducing a machine-readable and extensible vocabulary to describe metadata of mappings.
2. Defining an easy to use table-based format that can be integrated into existing data science pipelines without the need to parse or query ontologies, and that integrates seamlessly with Linked Data standards.
3. Implementing open and community-driven collaborative workflows designed to evolve the standard continuously to address changing requirements and mapping practices.
4. Providing reference tools and software libraries for working with the standard.

A SSSOM mapping comprises three major components:

1. The mapping itself, that is, a triple <subject, predicate, object> that reflects a correspondence of a subject entity, for example a class in an ontology, to an object entity, for example an identifier in some database, via a semantic mapping predicate, such as skos:exactMatch.
2. A mapping justification, which is the process or activity that led us to consider the mapping to be correct or reasonable (typical examples: labels match exactly; two classes are logically equivalent; a domain expert determined that two terms reflect the same real world concept).
3. Provenance metadata, including information about author and mapping tool.

Reference: SSSOM

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

Linked Open Vocabularies
Linked Open Data Cloud
The Linked Data Illusion

Data in the Linked Data Cloud is not linked
Linkage comes with referential identity
Referential identity comes with interoperability
LODD is not FAIR
Pharma Interoperability Hub & Knowledge Graphs
Pharma Interoperability Hub & Data Harmonization Service
FAIR by Design to support FAIRification at Scale

Product Line “Scientific Interoperability Hub” offering three products: terminology management, semantic dataset definition & conceptual modeling (purpose-driven ontologies)

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

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

Supporting more than 100 productive applications across all Roche functions and sites. The Data Harmonization Services guarantees currency and ongoing support.
Reference Data Services for Data Management
Terminology Management - Contextualize Concepts (FAIR)

Roche Terminology System 92.52.0 (PRD)

More than 110 productive applications
Reference Data Services for Data Management
Metadata Registry/ Dataset Models – Metadata Harmonization (FAIR)
Reference Data Services for Data Management

Conceptual Model - Purpose-build FAIR Ontologies
Semantic Interoperability Hub - Capability Stack
Data Management Value Chain - From Terminologies to a Unified Domain Model
Infectious Disease Ontology
Instantiation of a Knowledge Graph
FAIR Data Integration

Federation of Knowledge Graphs (Zero Integration)

I2O Knowledge Graph

Competitor Information

Knowledge Graph
Conclusions
Conclusions

- Successful and value-generating Digitilization requires true machine-actionable data, machine-readability is not sufficient. Application of FAIR principles is mandatory.

- FAIR data principles intrinsically tie Data Management to Semantic Technologies. (usage of terminologies, dataset definitions & ontologies)

- Transformationless data integration based on fully harmonized and standarized machine-actionable data assets (FAIR by design) results in fully linked data ecosystem to produce more reliable insights in less time at lower costs.

- Data Management Value Chain: new architectural approaches around data and information. Semantic Interoperability of terminologies, dataset definitions and ontologies is key to make our data assets machine-actionable.

- It's all about Semantics.
Acknowledgements
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 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

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