A complex network diagram with nodes and connecting lines, transitioning from orange to blue, set against a gradient background. The nodes are represented by small circles, and the connections are thin lines. The overall structure is dense and interconnected, with some larger nodes and clusters.

# Accelerating data integration using true FAIR Implementations

## Disrupting Data Management

Martin Romacker  
Product Manager  
Roche Data Marketplace, Enterprise Data Product Line  
DIGI Foundational Domain

Roche Informatics, Basel

15th November 2023, Pistoia Alliance Conference, Boston  
MA

# Table of contents

1. Some words about FAIR
2. FAIR Identifiers – The Beauty
3. The Data Science Dilemma
4. Disrupting our Industry – True FAIR Data Management
5. FAIR Identifiers – The Beast (FAIR Silos)
6. FAIR Data Mastery & Mastering FAIR
7. Conclusions

# Some words about FAIR

# FAIR Principles

## Foundation for FAIR Maturity Models



**F**indable



**A**ccessible




**I**nteroperable



**R**eusable

### The FAIR Guiding Principles for scientific data management and stewardship

[Mark D. Wilkinson](#), [Michel Dumontier](#), [IJsbrand Jan Aalbersberg](#), [Gabrielle Appleton](#), [Myles Axton](#), [Arie Baak](#), [Niklas Blomberg](#), [Jan-Willem Boiten](#), [Luiz Bonino da Silva Santos](#), [Philip E. Bourne](#), [Jildau Bouwman](#), [Anthony J. Brookes](#), [Tim Clark](#), [Mercè Crosas](#), [Ingrid Dillo](#), [Olivier Dumon](#), [Scott Edmunds](#), [Chris T. Evelo](#), [Richard Finkers](#), [Alejandra Gonzalez-Beltran](#), [Alasdair J.G. Gray](#), [Paul Groth](#), [Carole Goble](#), [Jeffrey S. Grethe](#), ... [Barend Mons](#)  [+ Show authors](#)

<https://doi.org/10.1038/sdata.2016.18>

True FAIR implementations follow a methodology

- FAIR is not only about the **\*THAT\***
- FAIR is above all about the **\*HOW\***

# Data Management Value Chain - FAIR Data Mastery

## Vision

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



Applications have FAIR meta-models & data structures for FAIR digital assets (eg terminologies)



FAIR APIs for data exchange including semantics (community standards)



FAIR data described by rich metadata (common vocabularies)

# FAIR Identifiers – The Beauty

# FAIR Principles

## Foundation for FAIR Maturity Models

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



[FAIR Guiding Principles](#)

# Digital Objects & Identifiers

## Making Digital Objects FAIR



Drug Discovery Today  
Volume 24, Issue 4, April 2019, Pages 933-938



Feature  
Implementation and relevance of FAIR data principles in biopharmaceutical R&D

John Wise<sup>1</sup>, Alexandra Grebe de Barron<sup>2</sup>, Andrea Splendiani<sup>3</sup>, Beeta Balali-Mood<sup>1</sup>, Drashti<sup>4</sup>, Eric Little<sup>4</sup>, Gaspare Mellino<sup>5</sup>, Ian Harrow<sup>1</sup>, Ian Smith<sup>6</sup>, Jan Taubert<sup>7</sup>, Kees van Bochove<sup>8</sup>, Martijn<sup>9</sup>, Peter Walgemood<sup>9</sup>, Rafael C. Jimenez<sup>10</sup>, Rainer Winnenburg<sup>11</sup>, Tom Plasterer<sup>12</sup>, Vibhor Gupta<sup>13</sup>, Victoria Hedley<sup>14</sup>



`//DICOM/image/2355segrfdfsfdps2.dcm`

`134be220-9f42-11ed-a8fc-0242ac120002`

`https://doi.org/10.1016/j.drudis.2019.01.008`



**In a FAIR data ecosystem every digital object is represented by a resource (GUPRI)**



# 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.

# Digital Objects & Identifiers

## Data Linkage & Data Quality comes with Identity



Drug Discovery Today  
Volume 24, Issue 4, April 2019, Pages 933-938



Feature  
Implementation and relevance of FAIR data principles in biopharmaceutical R&D

John Wise<sup>1</sup>, Alexandra Grebe de Barron<sup>2</sup>, Andrea Splendiani<sup>3</sup>, Beeta Balali-Mood<sup>1</sup>, Drashti Vasant<sup>2</sup>, Eric Little<sup>4</sup>, Gaspare Mellino<sup>5</sup>, Ian Harrow<sup>1</sup>, Ian Smith<sup>6</sup>, Jan Taubert<sup>7</sup>, Kees van Bochove<sup>8</sup>, Martin Romacker<sup>5</sup>, Peter Walgemood<sup>9</sup>, Rafael C. Jimenez<sup>10</sup>, Rainer Winnenburg<sup>11</sup>, Tom Plasterer<sup>12</sup>, Vibhor Gupta<sup>13</sup>, Victoria Hedley<sup>14</sup>



<https://doi.org/10.1016/j.drudis.2019.01.008>

www.freeworldmaps.net

# Digital Objects/ Digital Assets are Resources

Resources represented by Global Unique Persistent Resolvable Identifiers (GUPRI)

A code list or a terminology is a resource

A code list element or a term in a terminology is a resource

A file on a file system is a resource

A database table or a table schema is a resource

A data model, a conceptual model or an ontology is a resource

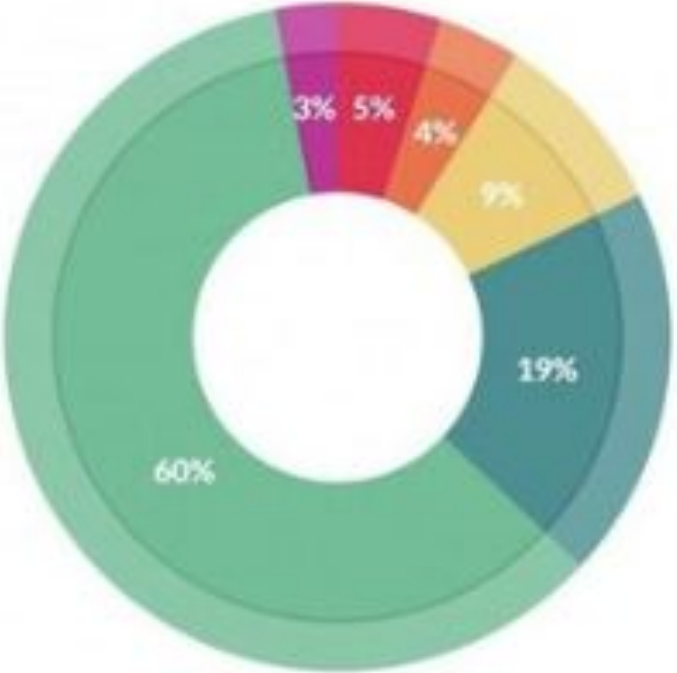
A concept or a business glossary term is a resource

A data element or a metadata element is a resource

# The Data Science Dilemma

# 80/20 Data Science Dilemma

## Data Science- and Analytics-Ready Data Assets



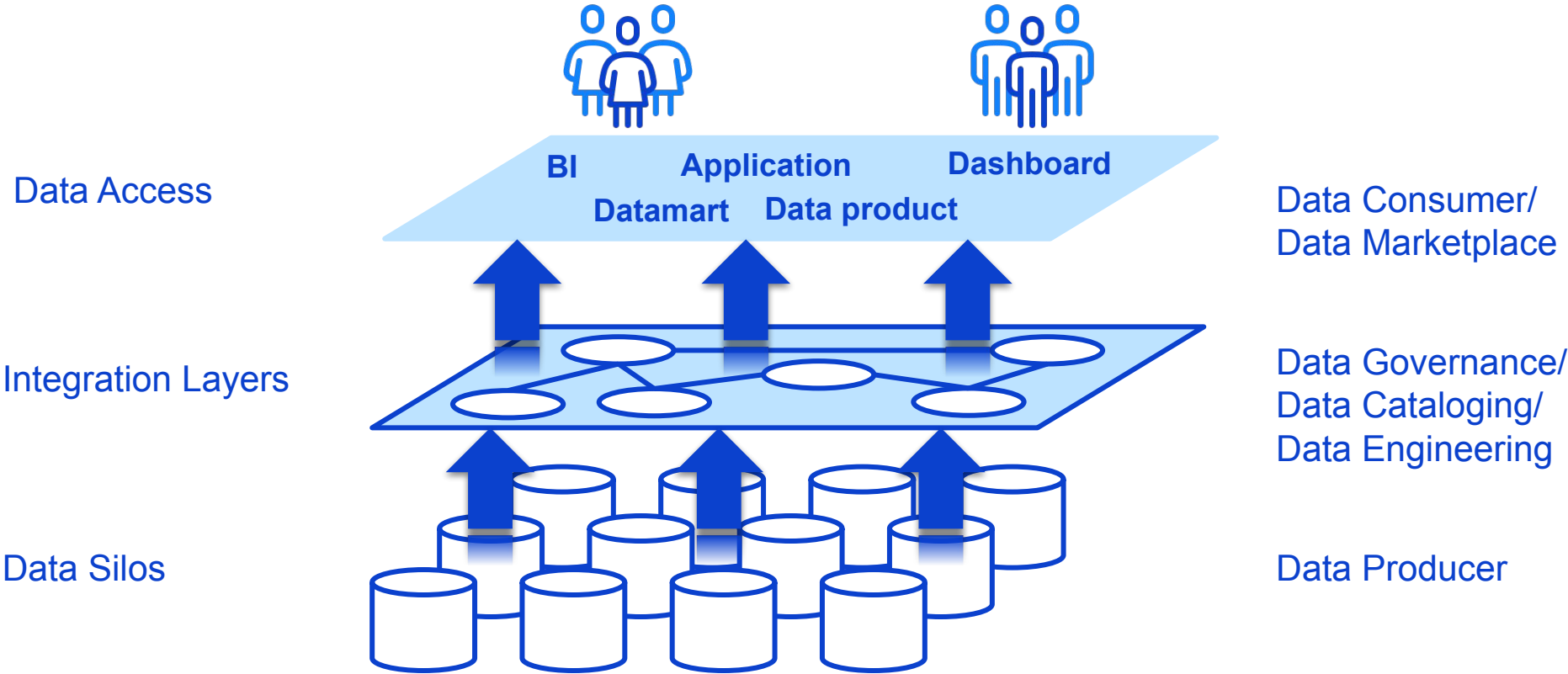
What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

# Disrupting our Industry – True FAIR Data Management

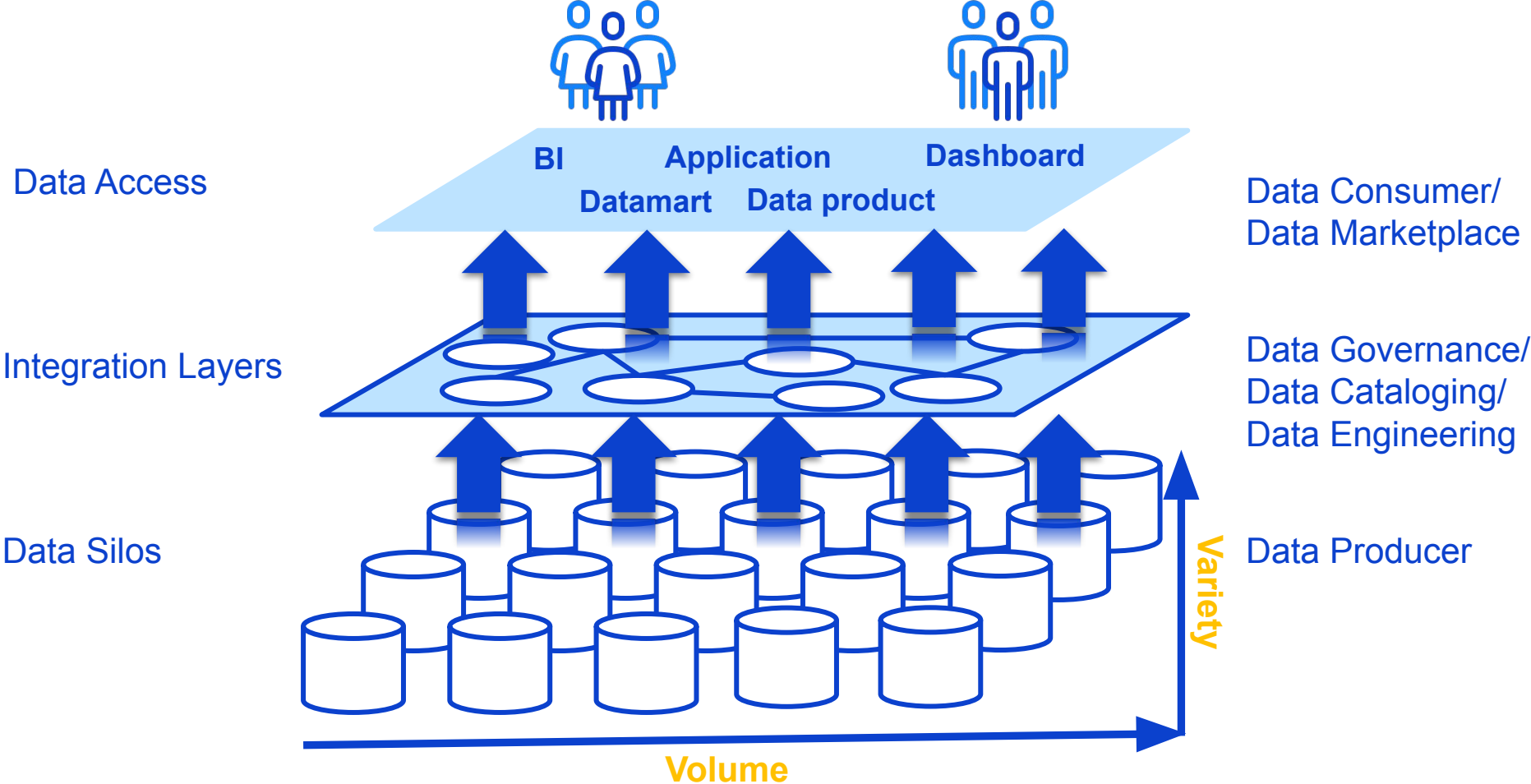
# Data Management Value Chain

Breaking Up Silos? Integration Layers on top of Silos!



# Data Management Value Chain

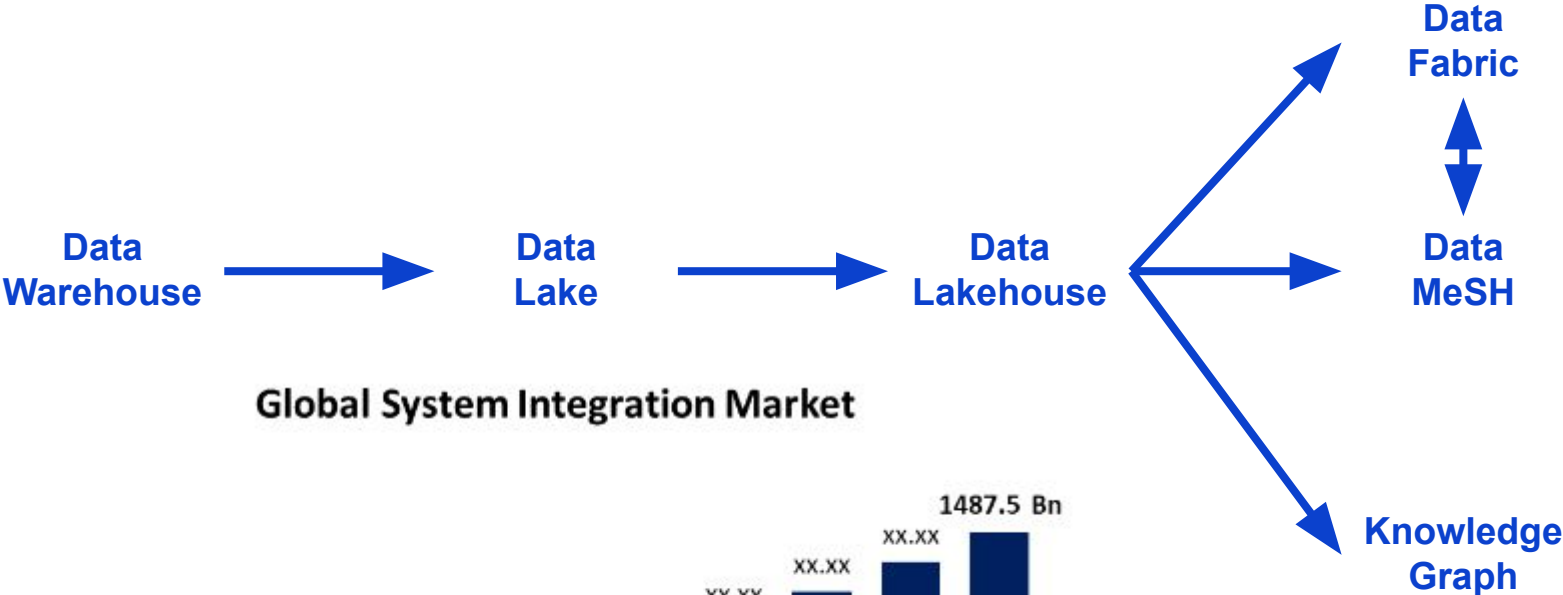
Growing Volume and Variety – Integration efforts ever growing



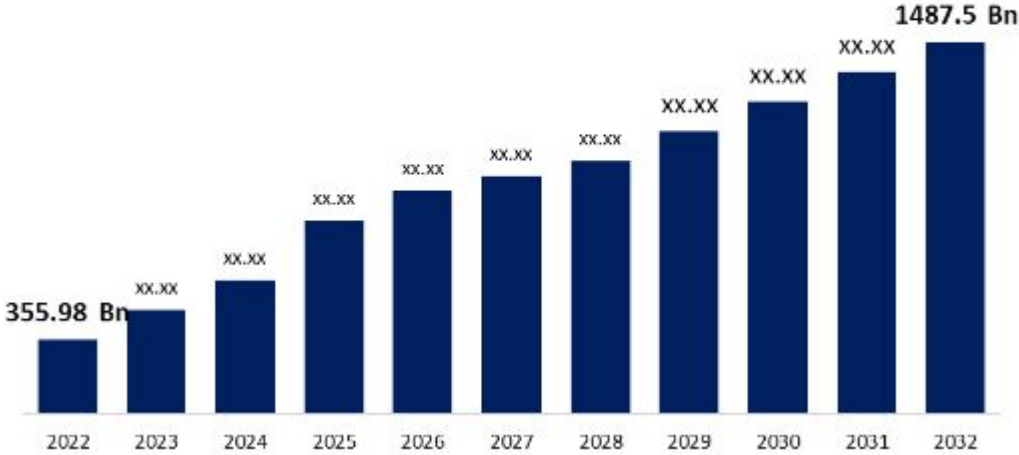


# Data Management Value Chain

## Technology Shift – Shifting Silos



**Global System Integration Market**

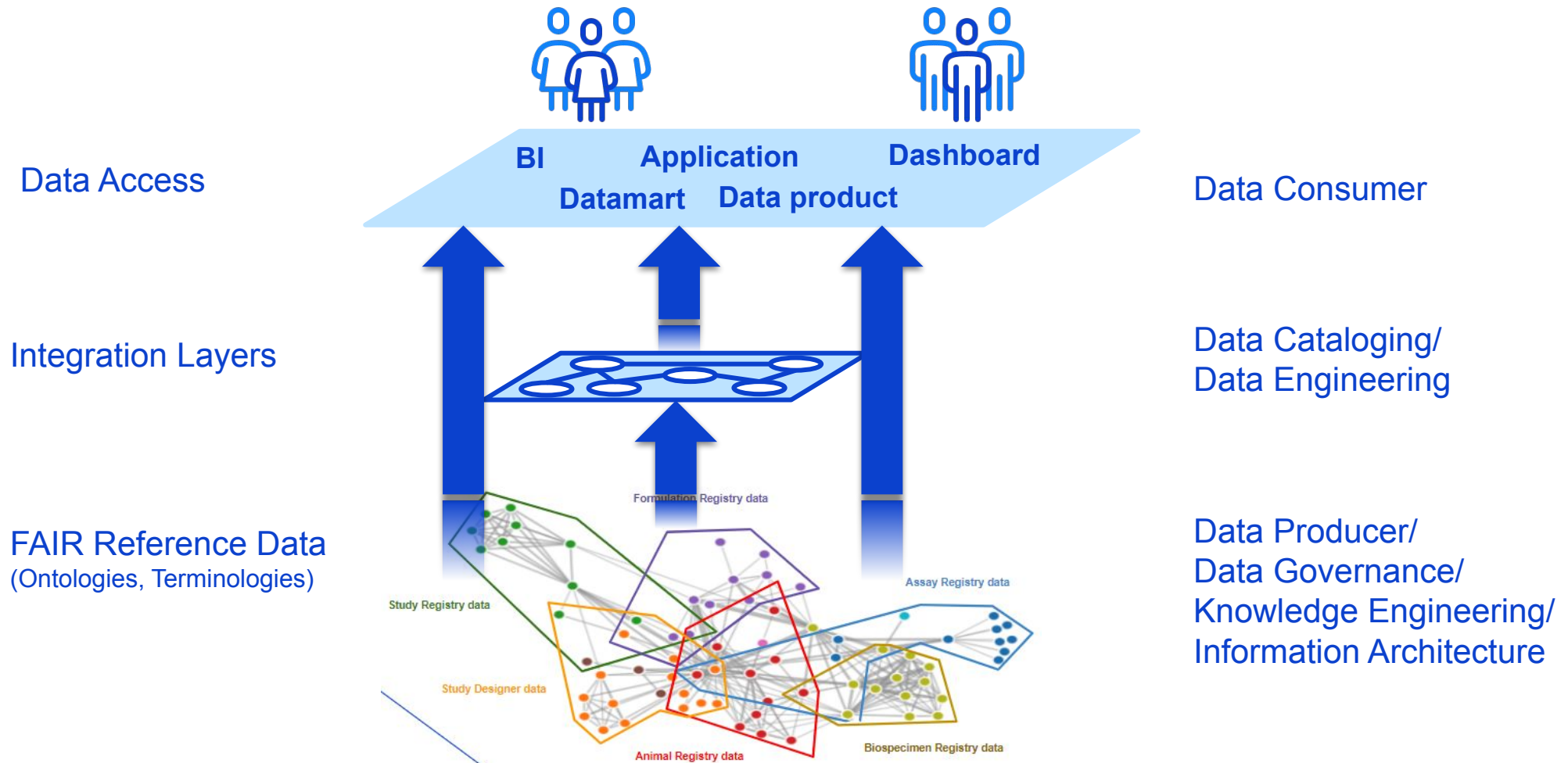


Source: Spherical Insights  
Thanks to Dave McComb

Technology does not solve data integration issues inherent to missing FAIRness

# Disrupting the Data Management Value Chain

Data Mastery – Eliminating Data Silos using Terminologies & Ontologies (FAIR by Design)



# FAIR Identifiers – The Beast (FAIR Silos)

# FAIRsharing Catalog of Biomedical Resources

## Proliferation and Fragmentation of Standards



FAIRsharing.org standards, databases, policies

Search all of FAIRsharing

Standards Databases Policies Collections Add/Claim Content Stats Log in or Register

Standards

Contribute by adding a standard Any problems? Please tell us!

The standards in FAIRsharing are manually curated from a variety of sources, including [BioPortal](#), [MIBBI](#) and the [Equator Network](#).

Manually done - no smart interfaces

30 %

Search Standards Search Search Reset Advanced

Showing records 1 - 50 of 1299.

View as Table View as Grid

Sort by Name

Recommended Records

Recommended

Associated Publication?

Registry	Name	Abbreviation	Type	Subject	Related Database	Related Standard	Related Policy	In Collection/Recommendation	Status
	ABA Adult Mouse Brain	ABA	Standard	None	None	None	None	None	R
	Access to Biological Collection Data	ABCD	Standard	Biodiversity Biology Life Sciences	GBIF Atlas of Living Australia IPT - GBIF Australia	ABCD EFG ABCDDNA	None	TDWG Biodiversity Information Standards	R

Bioportal has grown from 700 ontologies to 1081 ontologies in the last 6-8 years (15th Nov 2023)

# Managing FAIR Reference Data

## Proliferation and Fragmentation of Concept Identifiers



ChEBI:7798



SNOMED:412261005  
SNOMED:777006006



RxNORM:260101



chEMBL:1129



DB:00198 DRUGBANK

KEGGC:C08092  
KEGGD:D08306



MeSH:D053139

Anatomical Therapeutic  
Chemical Classification  
System

ATC:J05AH02

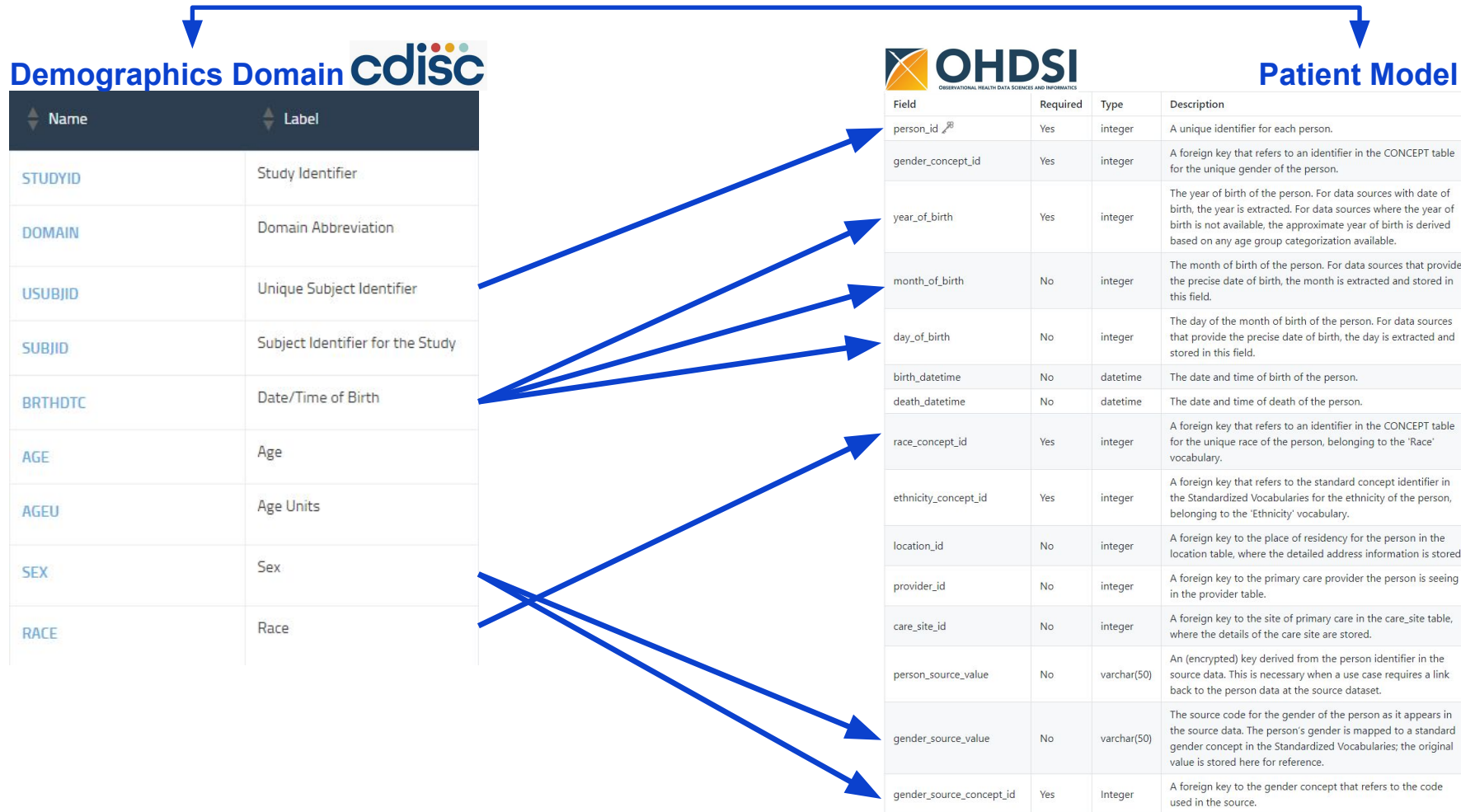


NCIt:C29305  
NCIt:C62061

Plethora of GUPRIs for the same semantic concept: welcome back to map & merge

# Data Standards & Interoperability

## CDISC vs OMOP OHDSI – Proliferation (Meta-)Data Elements/ Semantic Schemas



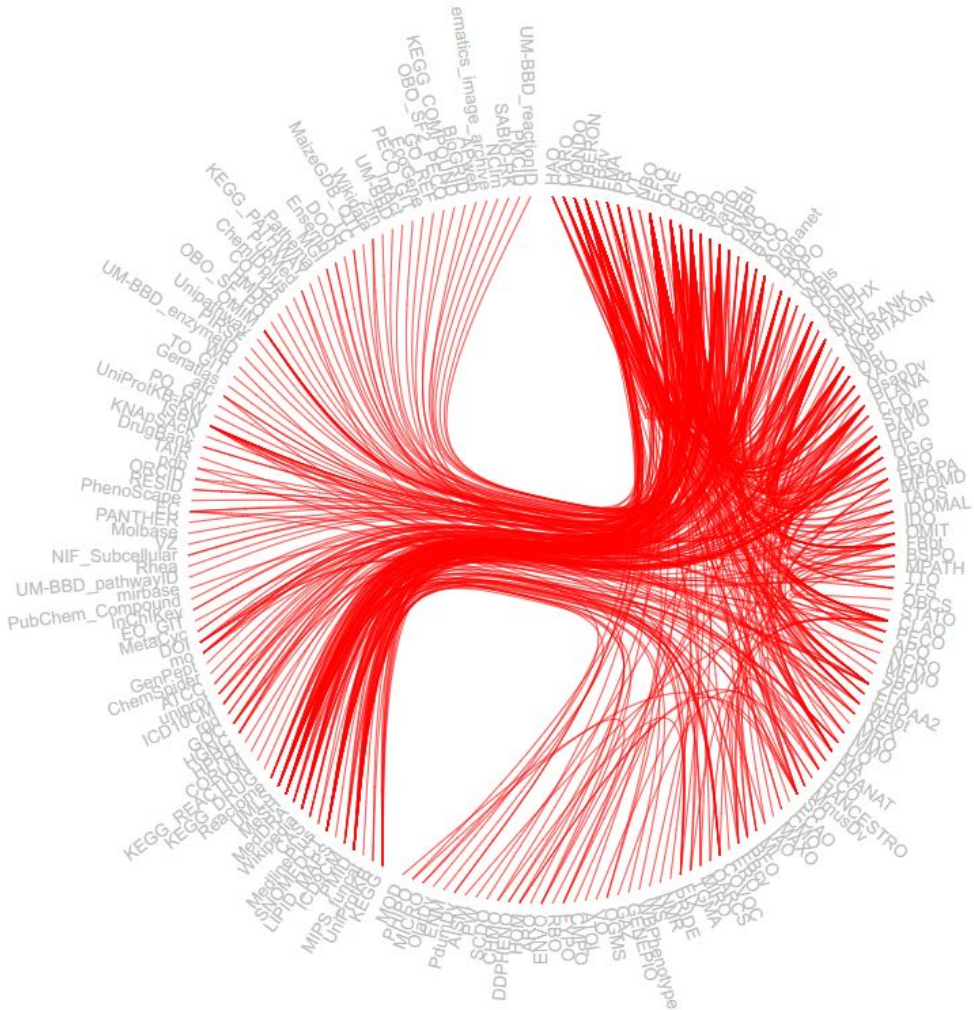
Creation of insights & analytics blocked: different schemas, variables and values

# EMBL-EBI Ontology OXO (Xref Service)

## Creating Referential Identity by Ontology Mapping

Welcome to the EMBL-EBI Ontology Xref Service (OxO).  
OxO is a service for finding mappings (or cross-references) between terms from ontologies, vocabularies and coding standards. OxO imports mappings from a variety of sources including the [Ontology Lookup Service](#) and a subset of mappings provided by the [UMLS](#). We're still developing the service so please [get in touch](#) if you have any feedback.

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



Source: EMBL-EBI OxO

# FAIR Data Master & Mastering FAIR



# FAIR Data Mastery & Mastering FAIR

Refining the Vision - Productivity Boost for Life Sciences R&D

*Standardizing FAIR to implement an open public-private semantic infrastructure of fully connected FAIR applications, services & data*

Govern & manage FAIR reference resources to make data machine-actionable

# E2E Mastery of Data Management Value Chain

## FAIR Digital Assets & Standardized FAIR



Significantly reducing global data integration efforts by truly eliminating data silos  
(cost avoidance)



Reducing time to make data assets consumable by transformation-less integration  
(time to market)



Making digital objects true data assets and machine-actionable  
(monetizing data assets)



Increasing productivity by better and more reliable insights  
(quantity & quality)

# Conclusions

# Conclusions

FAIR Data First for true Disruption !  
Digital Transformation requires FAIR Data Strategy at community level.

Transformation-less data integration using FAIR machine actionable digital assets.  
Intrinsically linked FAIR data ecosystem.  
More reliable insights in less time and at lower costs.

Pistoia Alliance (Data Driven Value) developing the community Master Plan.  
Umbrella & Driver for standardizing FAIR reference data (Ontologies, DataFAIRy).  
Connecting Content & Service Providers: FAIR applications, services & data.

It is all about Semantics !

Doing now what patients need next