AstraZeneca

Enabling new medicines with FAIR Data-centricity:

Delivering Linked Data inside the enterprise

Ben Gardner

Data Standards, Interoperability and Governance, Data Office, Data Science & Artificial Intelligence, R&D, AstraZeneca, Cambridge, United Kingdom

September 2023

Contents

- **Enabling new science** – Information Discovery not Search
- **Standardise your infrastructure** – The data is everything
- **FAIR Data-centric Information Architecture** – Hiding the semantics from the enterprise

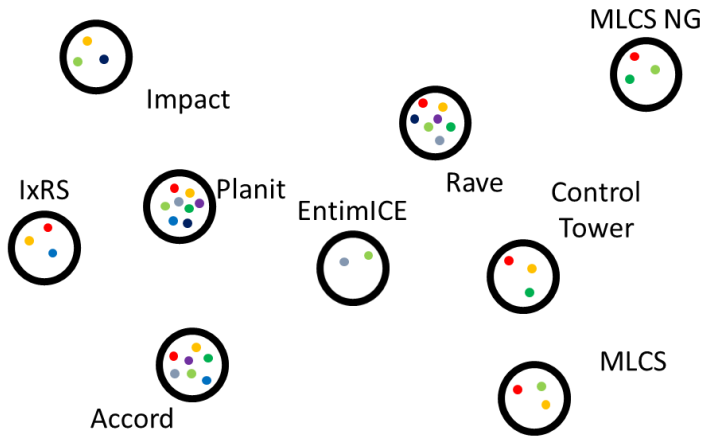
Enabling new science

Information Discovery not Search

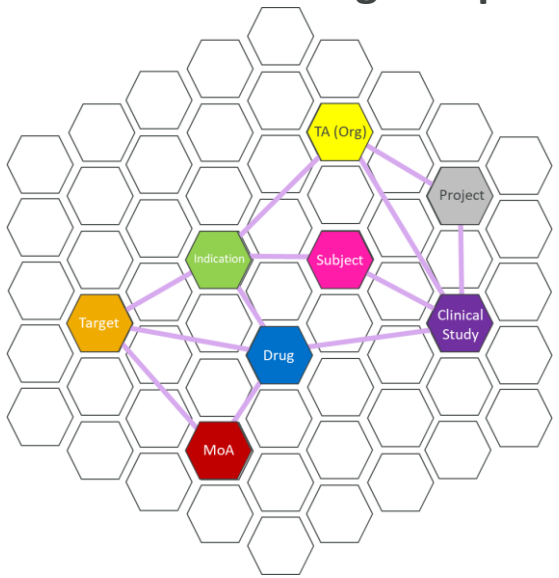
The evolving science is driving the need for data centricity

- As our understanding improves, we identify sub populations of diseases
  - Cancer → Lung Cancer → NSCLC → EGFR T790M mutation
- To drive better medicines for patients we need to connect the data at ever more granular levels
  - Clinical Study → Subject → Sample

Systems centric
Fragmented and silo'ed data



Data centric
FAIR & Knowledge Maps



Scientific Intelligence

Use case gathering

*As a data scientist, to be able to access **eGFR measurements** and **patient level clinical data** across **clinical trials** to build Risk Models to Predict **Chronic Kidney Disease** and Its Progression*

As an Oncology Bioinformatician, I want to develop a model that explains how lung cancer tumours can resist to xxxxxx or xxxxxxxxxx, so that biomarkers can be created to predict drug resistance

As a translational scientist, I want to access patient derived cancer models multimodality data (imaging omic and SOC dosing/response) and patient data of corresponding indication so I can build predictive models for preclinical drug evaluation

Identify Data sources

clinical data

Omics data

Imaging data

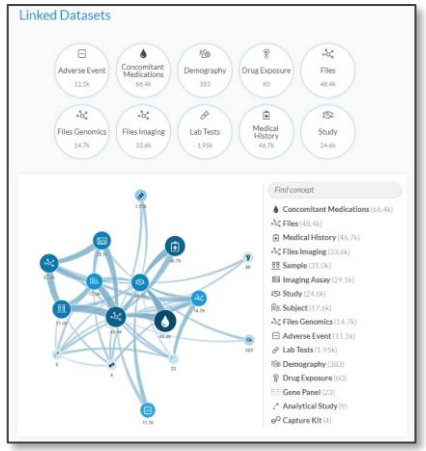
Sample data

FAIR data sources within platforms

R&D Knowledge Map

Integrated using a knowledge map

Exploration UI

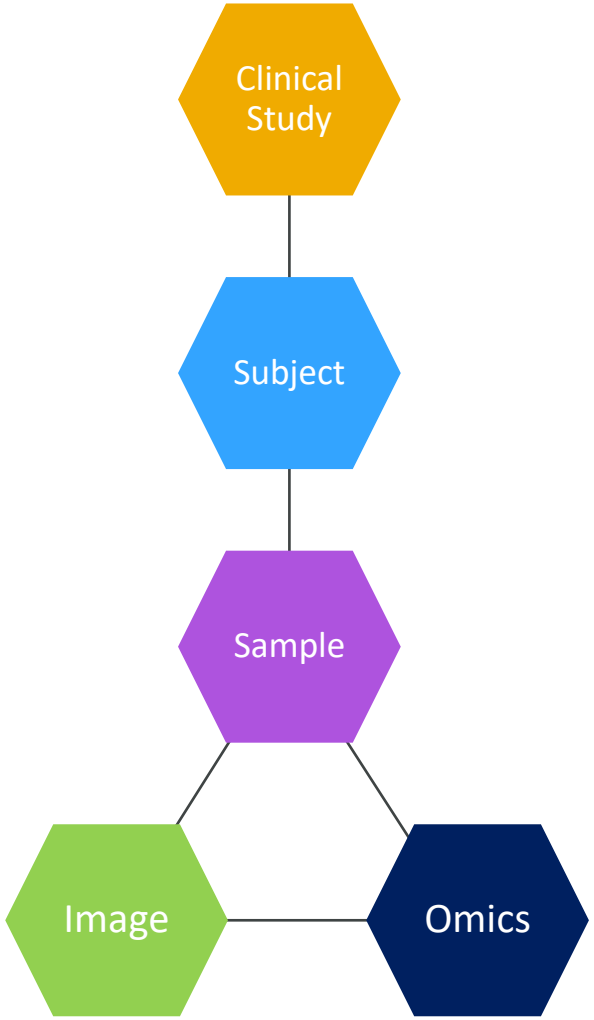


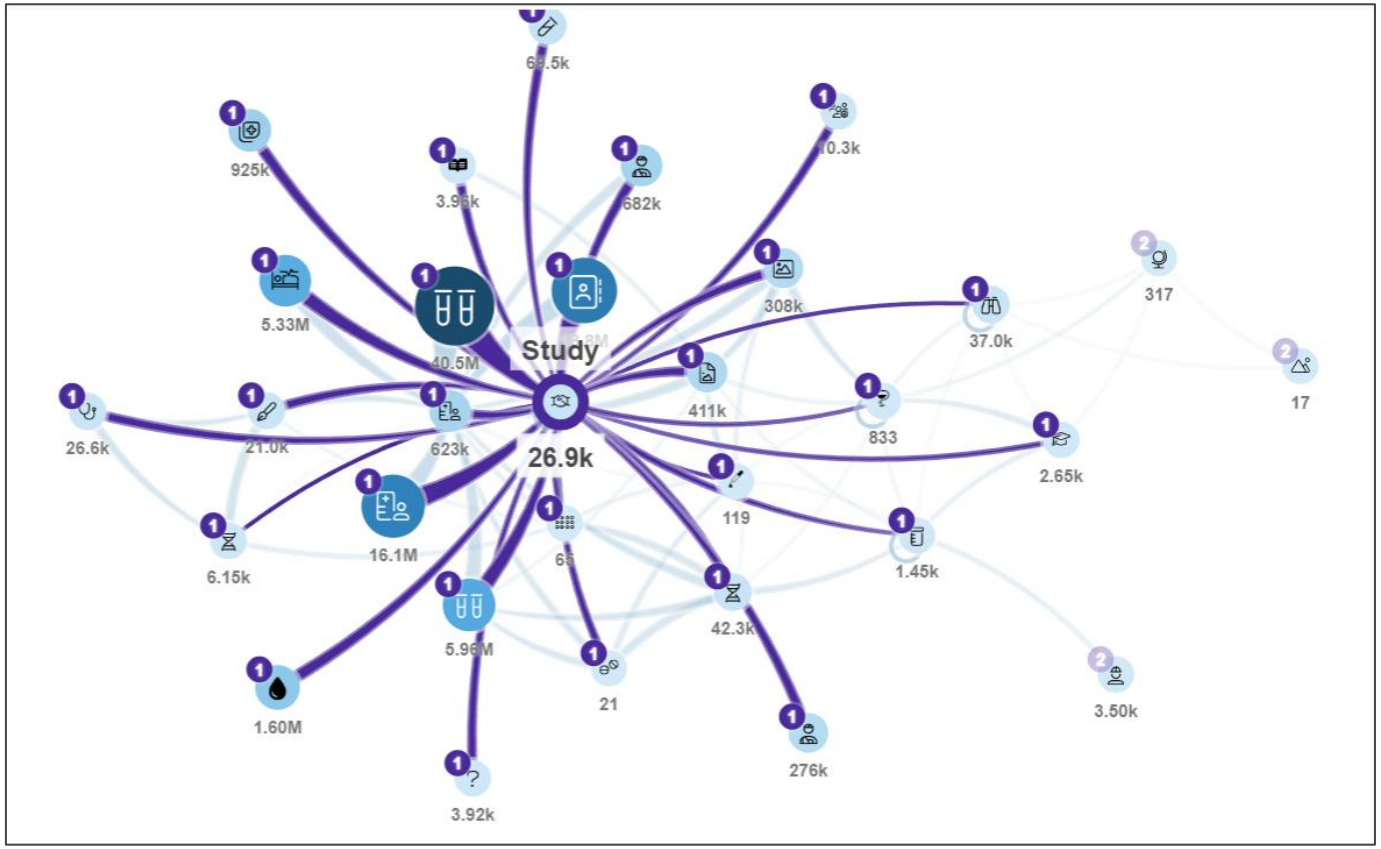
Accessed via *Scientific Intelligence*

Data access according to compliance i.e. iDAP

Scientific Intelligence

Study

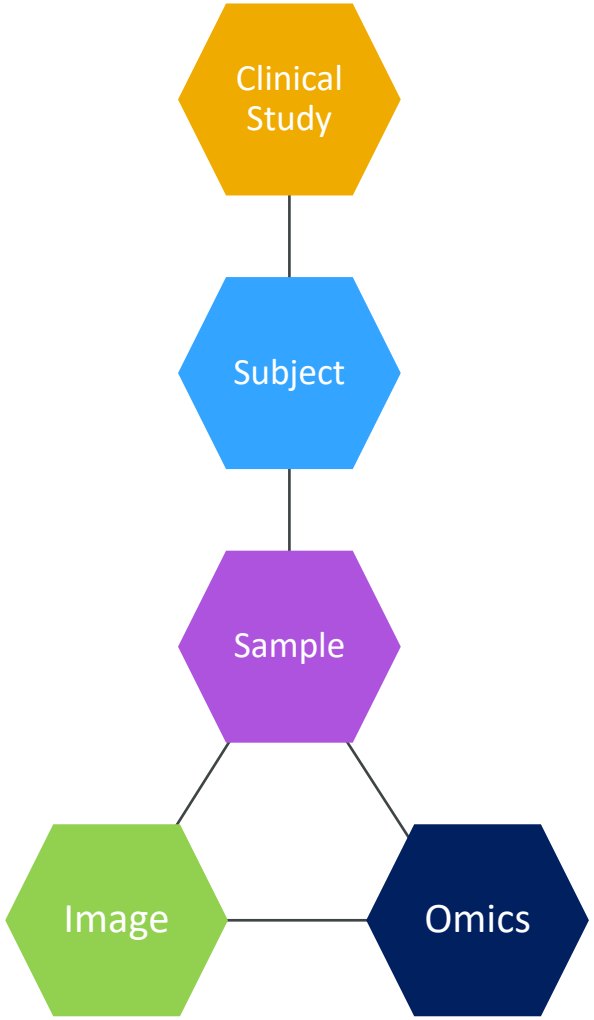


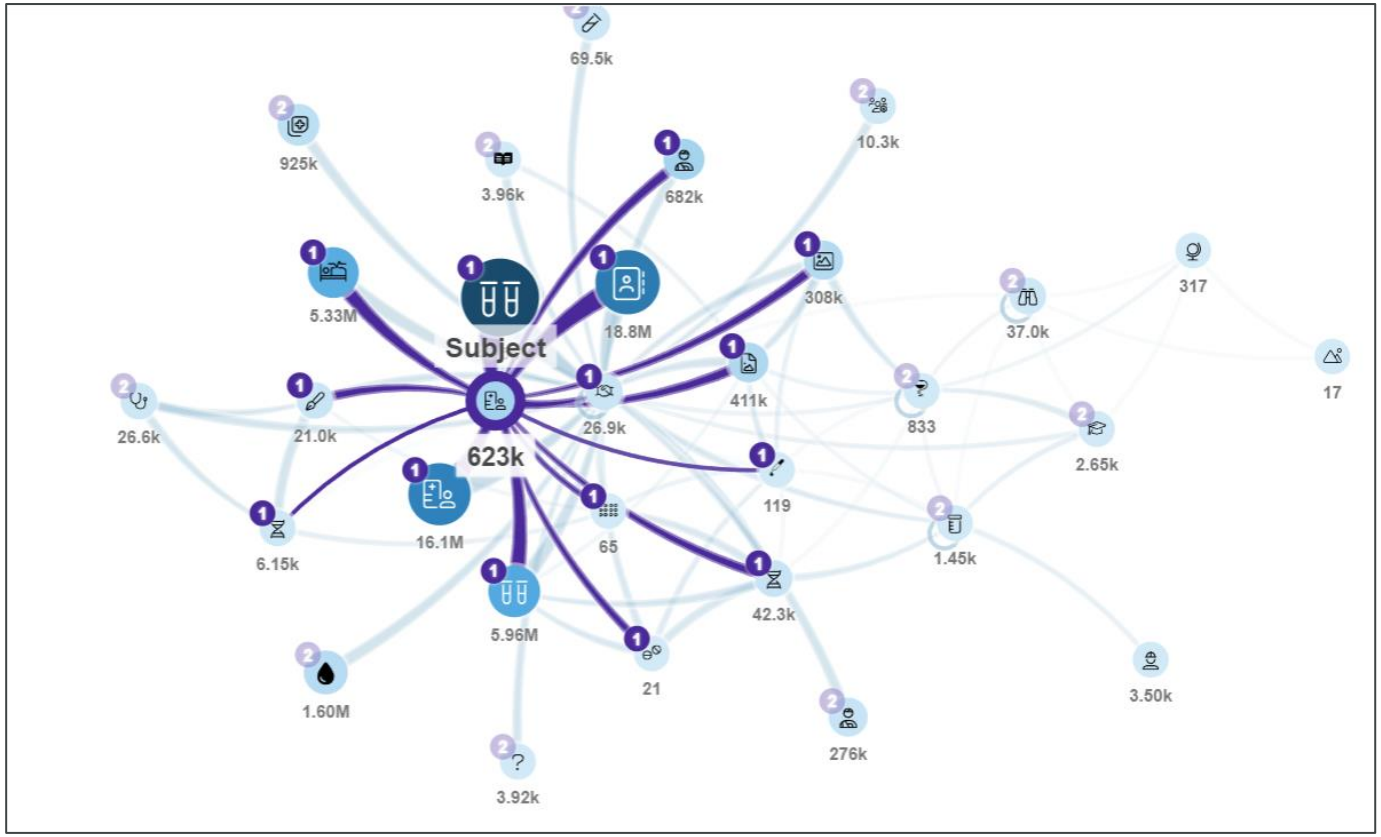


Summary statistics	Instance data
Counts of studies by Indication, Drug, etc	Title, Drug, Indication, Status, No. of patients recruited, Milestones, CSP, CSR, etc

Scientific Intelligence

Subject

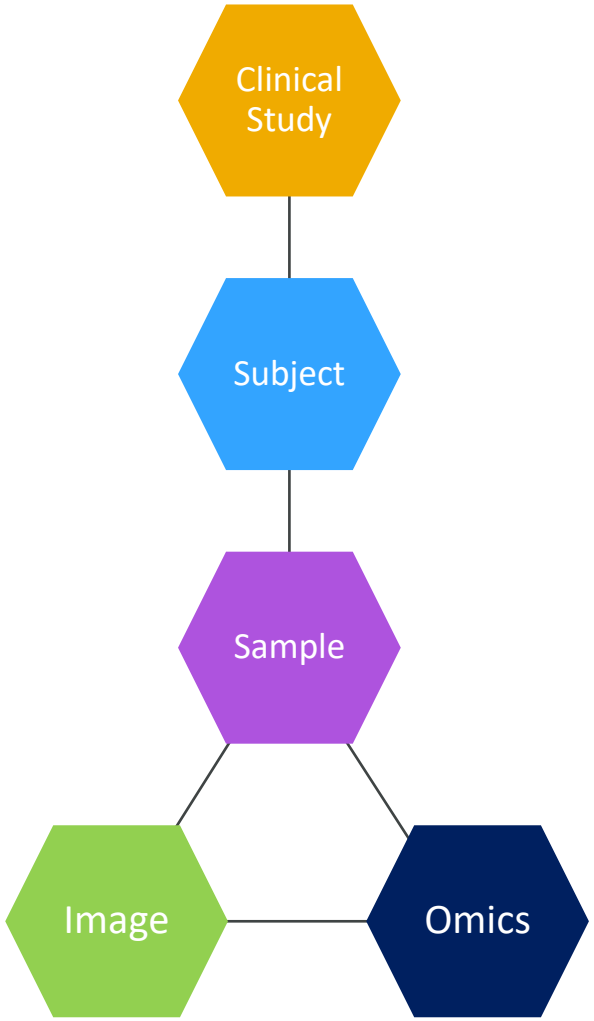


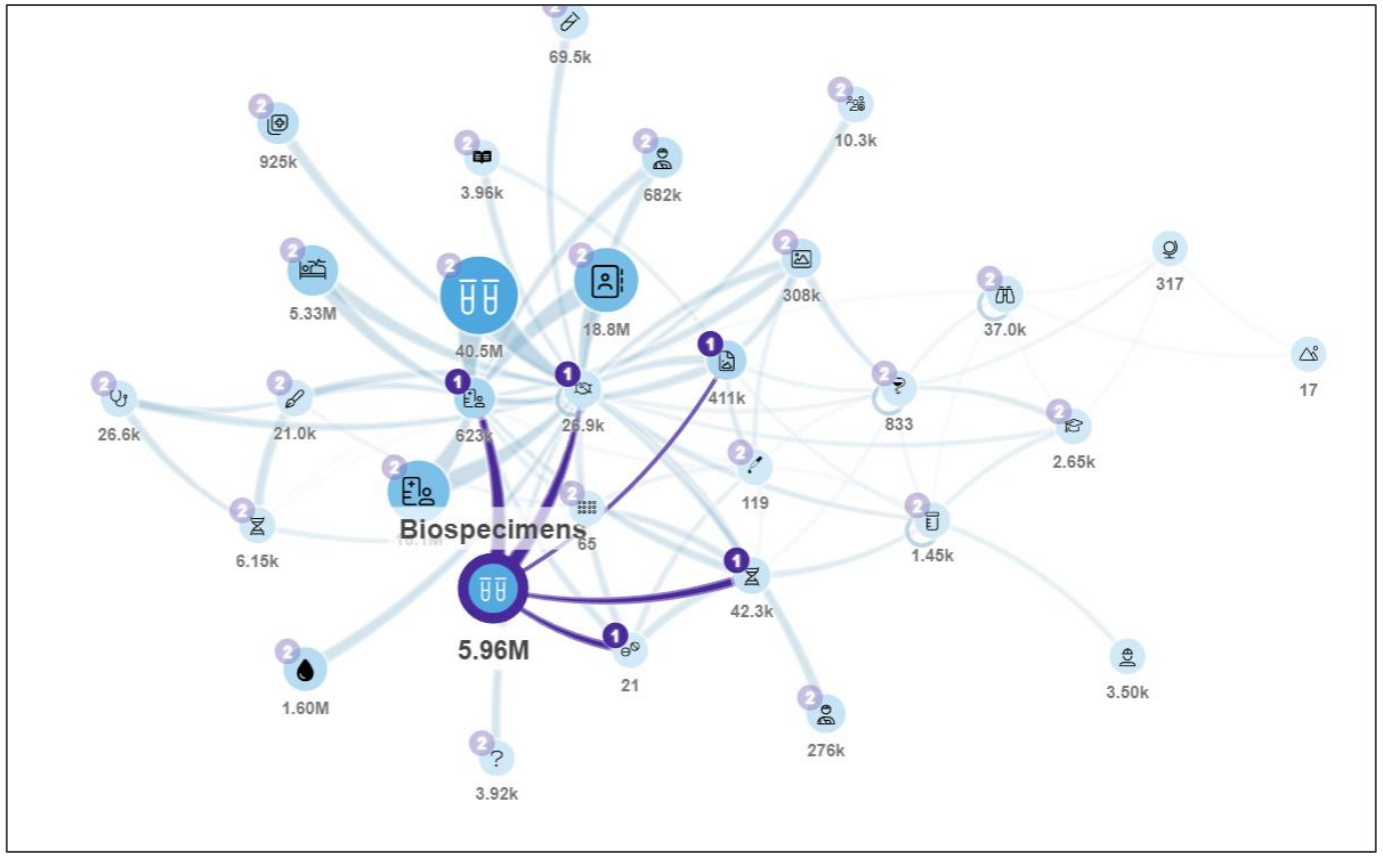


Summary statistics	Instance data
For each subject display counts of Serious AEs, total Lab Tests, etc	Individual subject data for common non-sensitive modules i.e. Demographics, Adverse Events, Lab Tests, etc

Scientific Intelligence

Sample

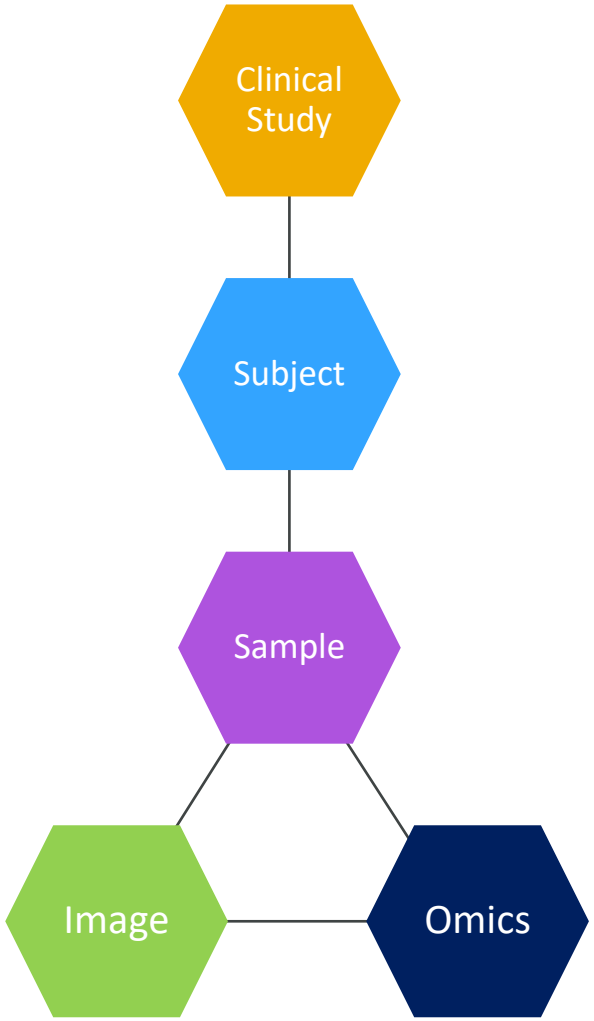


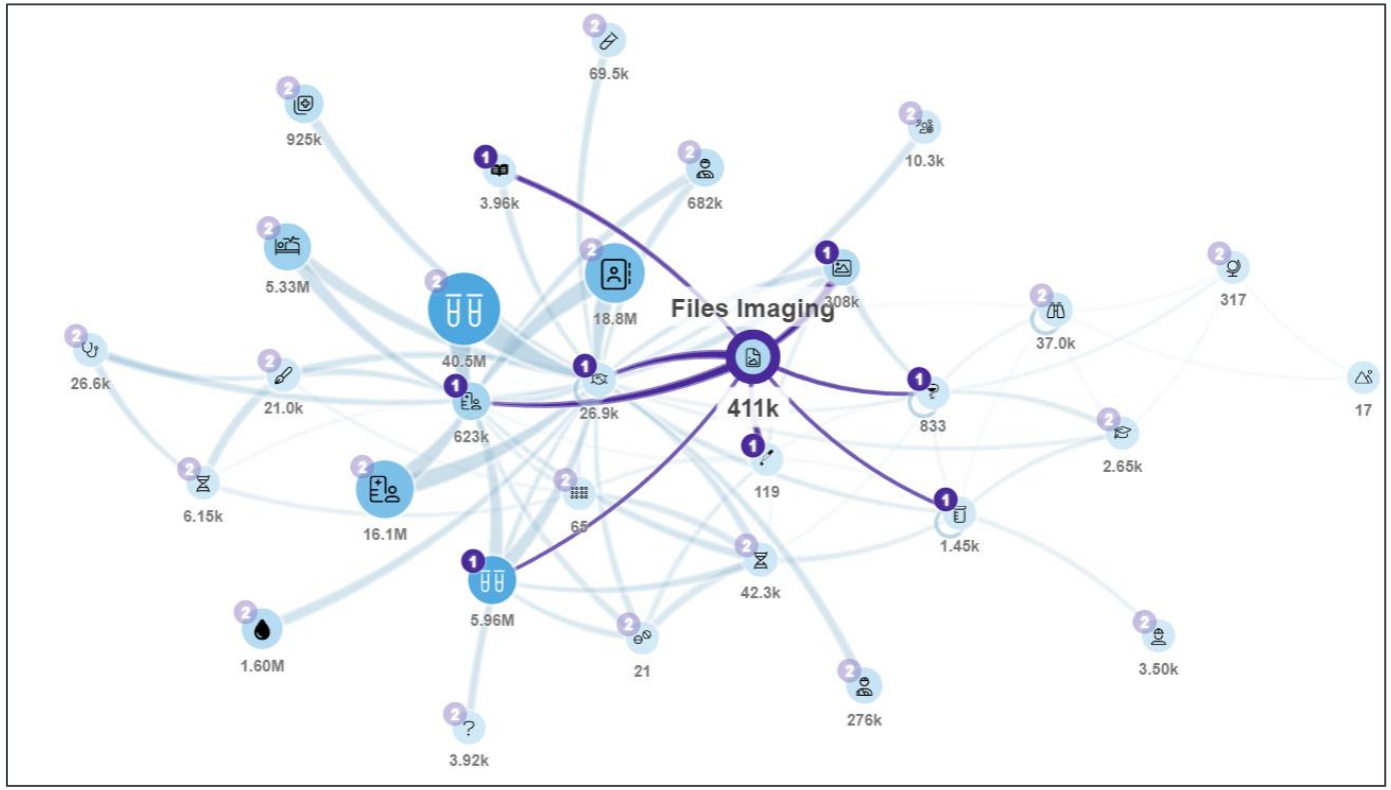


Summary statistics	Instance data
For each Subject display a counts of samples, etc	For each Sample display inventory, sample type, etc

Scientific Intelligence

Observations





Summary statistics	Instance data
For each subject display a counts of images of a type, sequencing files, etc	Variant observed, tumour typed, stain performed, image readout, etc

Scientific Intelligence use cases

Questions we could not answer easily before

Landscape views

Show the distribution of subjects by biomarker, and disease stage supplemented with insights into the different observations (Omics, Imaging and Clinical data) and BioSample availability.

Efficiency saving – months to days (build), months to minutes (update)

Study design optimisation

Determine the variability of various lab tests within a specific cohort of subjects to help determine the number of subjects to include in a clinical study.

Efficiency saving – weeks to hours

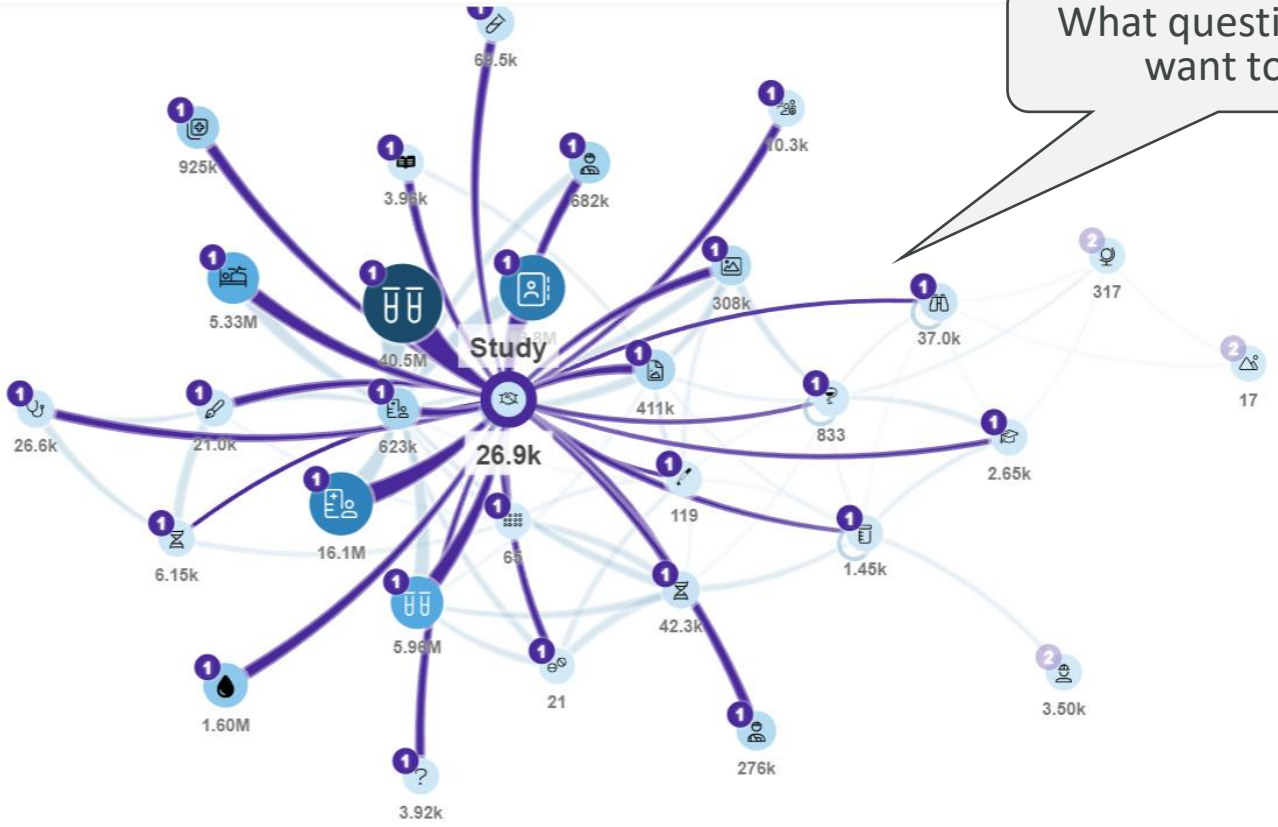
Cohort selection

Find subjects with indication X who were treated with Y and who devolved adverse event Z where we have CT scans and sequencing data.

Efficiency saving – weeks to minutes

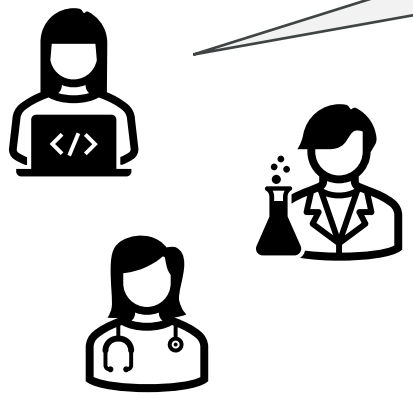
Scientific Intelligence

Biggest challenge, understanding the art of the possible

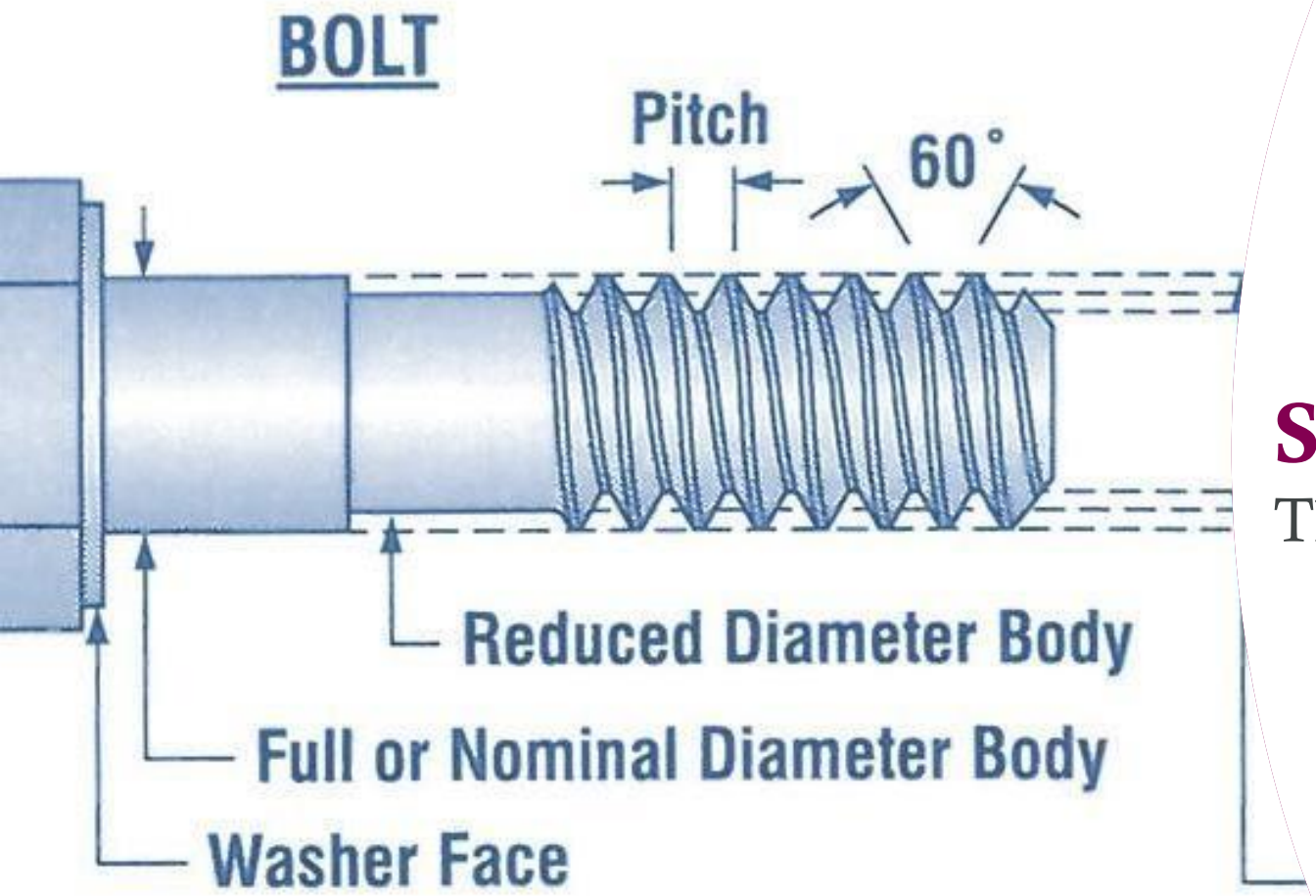


What question do you want to ask?

We don’t know. We haven’t been able to ask this sort of question so haven’t even thought about it



Hugely powerful but not everyday questions



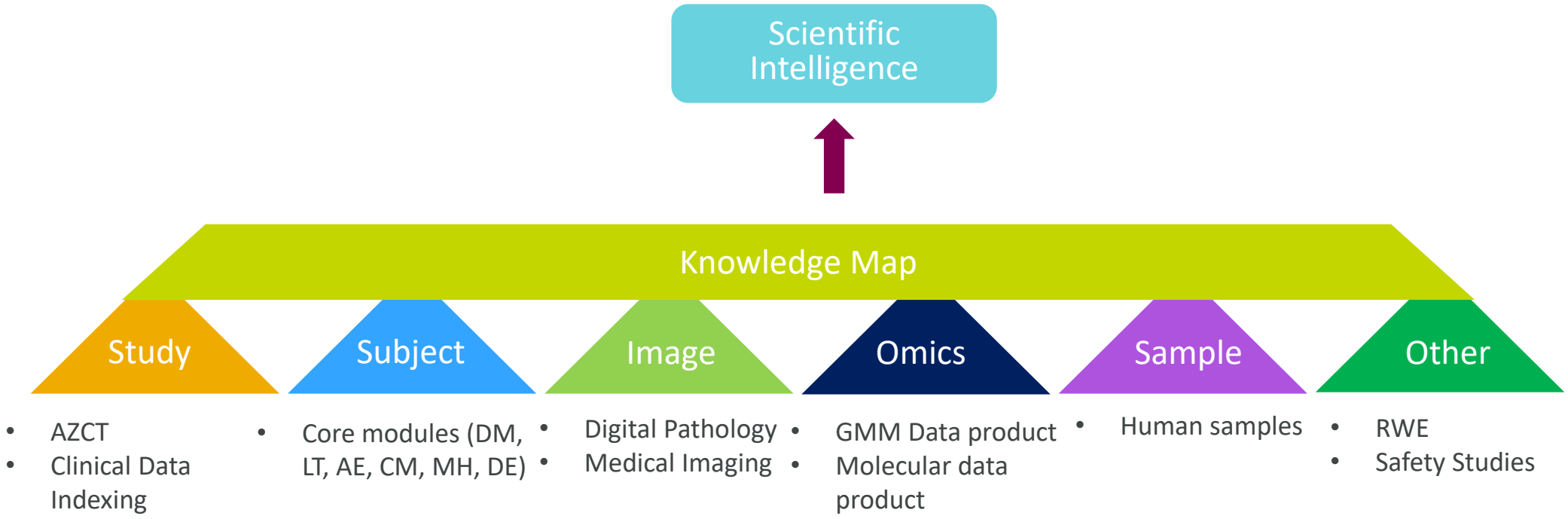
Standardise your infrastructure

The data is everything

Image from https://mechanical.engineeringtalks.com/gallery/image/2781-bolt-and-nut-screw-thread-terminologyjpg/

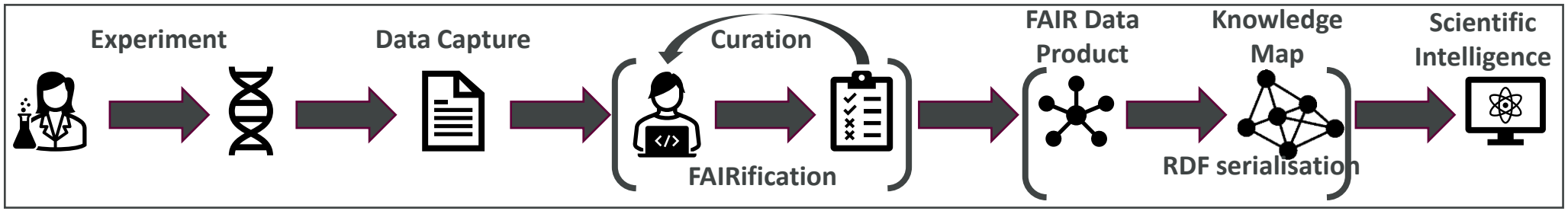
Success breeds headaches

Bottom up is fast but incurs technical debt



What we learnt and why we ended up where we are

We did everything and killed ourselves



How can we simplify how we do FAIRification?

- Could Data Mesh & Data Products help?
- What can we do to standardise?
- Can we maximise our data engineering skills?

How do we better manage the graph and ecosystem?

- We spend too much time managing infrastructure?
- How to evolve from virtualisation to physical graph?
- Orchestration how?

How do we democratise the graph?

- Can we simplify data pipeline in Sci-Int?
- How do we play nicely with the rest of AZ?
- How do we extract more value?

Standardise, standardise, standardise

Focus our skills on the data not the infrastructure

FAIRification | **Graph ecosystem** | **Democratisation**

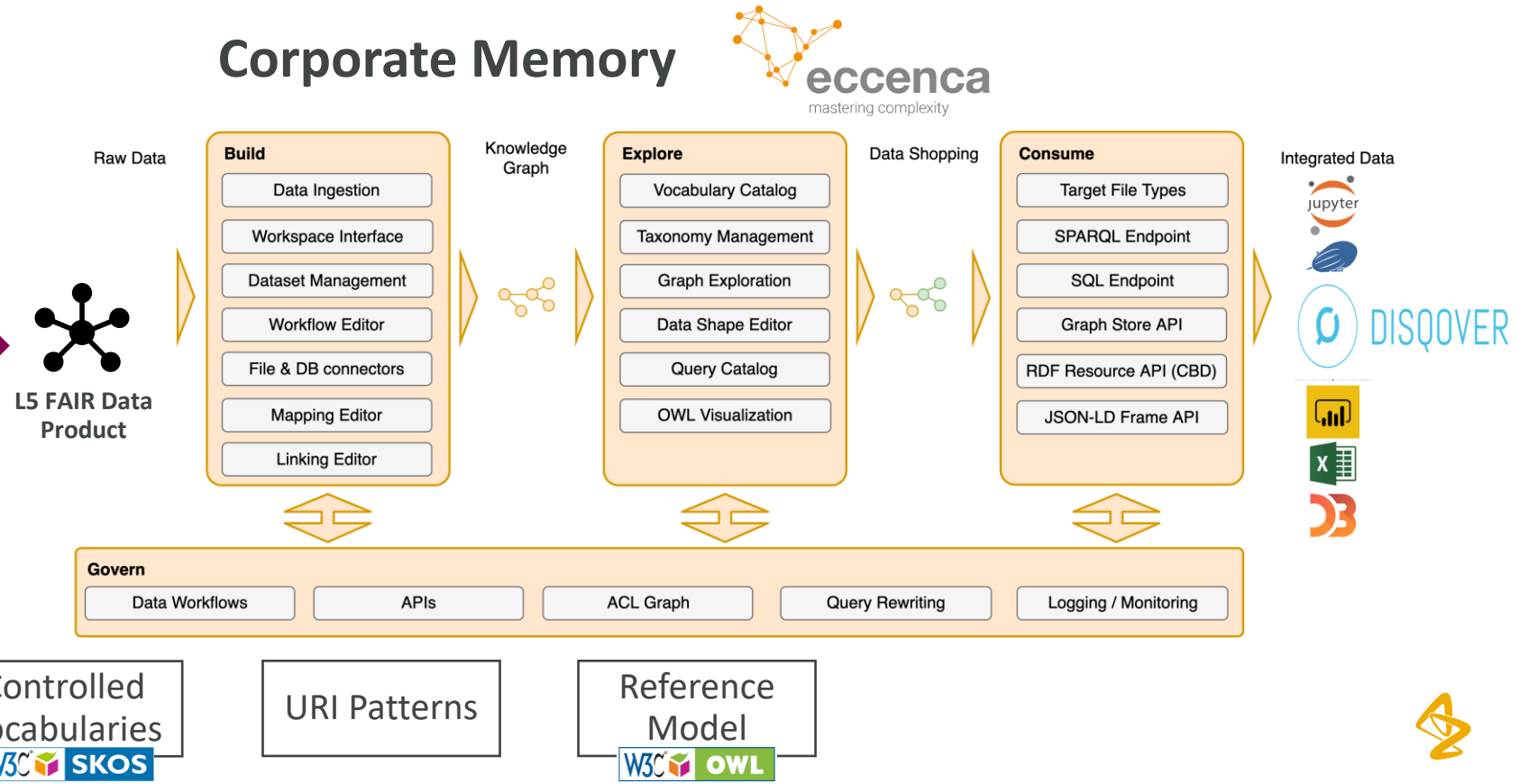
Consume Data Products from Platforms not individual sources

Build our FAIRification pipelines using enterprise standards

Apache Airflow · Quantified FAIR · dbt · snowflake

Data cleaning, applying CVs, enriching, flattening & minting PIDs

Even better how can we get other people to do FAIR@Source or FAIRification?



FAIR Data-centric Information Architecture

Hiding the semantics from the enterprise

What is Data Centricity?

Data-Centricity puts **data at the centre of the enterprise**.

Applications are optional visitors to the data. (Data-centric manifesto)

Data-centricity involves structuring our **data around the science** that we do **rather than the systems** that we use. It promotes data reusability over system-centric design.

Defining our data-centric transformation using FAIR metrics

Evolution of System Centric Thinking

Evolution of Data Centric Thinking to Knowledge Centric

L1

- **Data is catalogued in situ**
- **Raw data** in an unconformed format, external to a data lake
- **Requires expert** technical and subject matter **knowledge** to access and use

L2

- **Raw data catalogued and accessible** with governance in a data lake, unconformed.
- **Requires average** technical and **expert subject matter knowledge** to use

L3

- **Conformed data catalogued** and available in a data lake in accordance with AZ patterns
- **Data is conformed on a system by system basis** and may not map to a domain data model
- **Controls on access at the system level**
- **Requires average level technical and subject matter knowledge** to use: Data Scientist.

Sweet spot for many AZ domains – where deep science and/or broad data integration is not required

L4 – Domain level FAIR

- **Data conformed, integrated, processed and audited** to support specific analytics patterns/enquiries
- Data is conformed using a **domain level data model**
- Data embeds **local master and reference data**
- **Controls on access at the data level**
- **Creation of analytics ready 'Marts'** enabling self service analytics: Citizen Data Scientist

Some data domains will require additional investment to support scientific use cases. This maximises use of Enterprise data and fully supports AI

L5 – Enterprise Level FAIR

Extend L4 further by:

- **Data is described in a cross domain data or industry model** and is integrated in a Data Mesh
- Data embeds **enterprise master and reference data**
- **Enterprise Metadata** standard applied
- **URI's and PURLs** are implemented
- **Cross domain Analytics enabled:** Citizen Data Scientist using all relevant AZ data

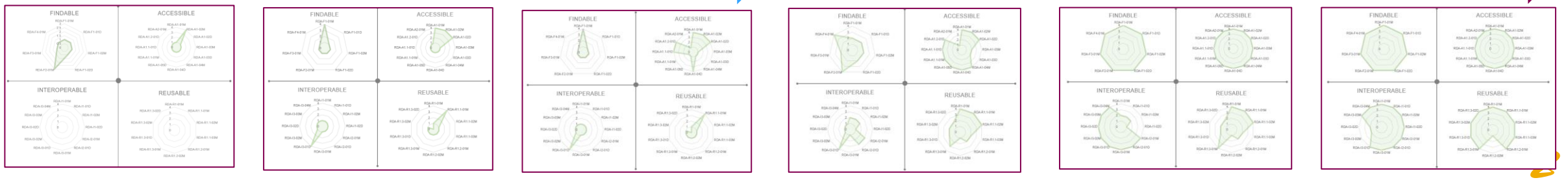
L6 – Knowledge Level FAIR

Extend L5 further by:

- **Data is fully described using a knowledge language or ontology**
- **Data is aggregated by business concepts** and users can navigate from concept to concept
- **Automated AI enabled**: AI and semantic solutions can act directly on data sets without need for interpretation
- **Knowledge enabled citizens**

System-centric Catalogs and API's only, giving minimal FAIR capabilities

Requires strong and increasing linkage of FAIR with TRUST



Credit to Information Architecture and Ben Gardner

The value of data centricity

Think about your shopping experience….

Junk Yard

- **Randomly distributed stuff**
- Lots of effort to find stuff
- **Easy to miss what you are looking for**
- Only the 'owner' knows where stuff might be at best

Car Boot Sale

- **Random stuff by car**
- Ordered by seller
- Sellers have assigned pitches
- **No insights about what other sellers have**
- **Random level of quality** i.e. first edition vs latest edition book

Moroccan Street Market

- **Grouping by product category** i.e. spices vs carpets vs etc
- **Improved level of quality**
- Seller can explain the **provenance of the product**
- Good intentions but **not really scalable**

Local Market

- **Well organised**
- Lots of categories
- **Well labelled items including provenance** (made where)
- **Limited produce**

Hyper Market

- **Scale and organisation** at the next level
- **Everything in once place**
- Store guide/map
- Optimised for society/community
- **Specialist by product category**
- **Data driven display/groupings** - I.e. Christmas, summer vs winter; mining shopping patterns, context sensitivity
- **Click and collect service**
- **Price comparison with other retailers**

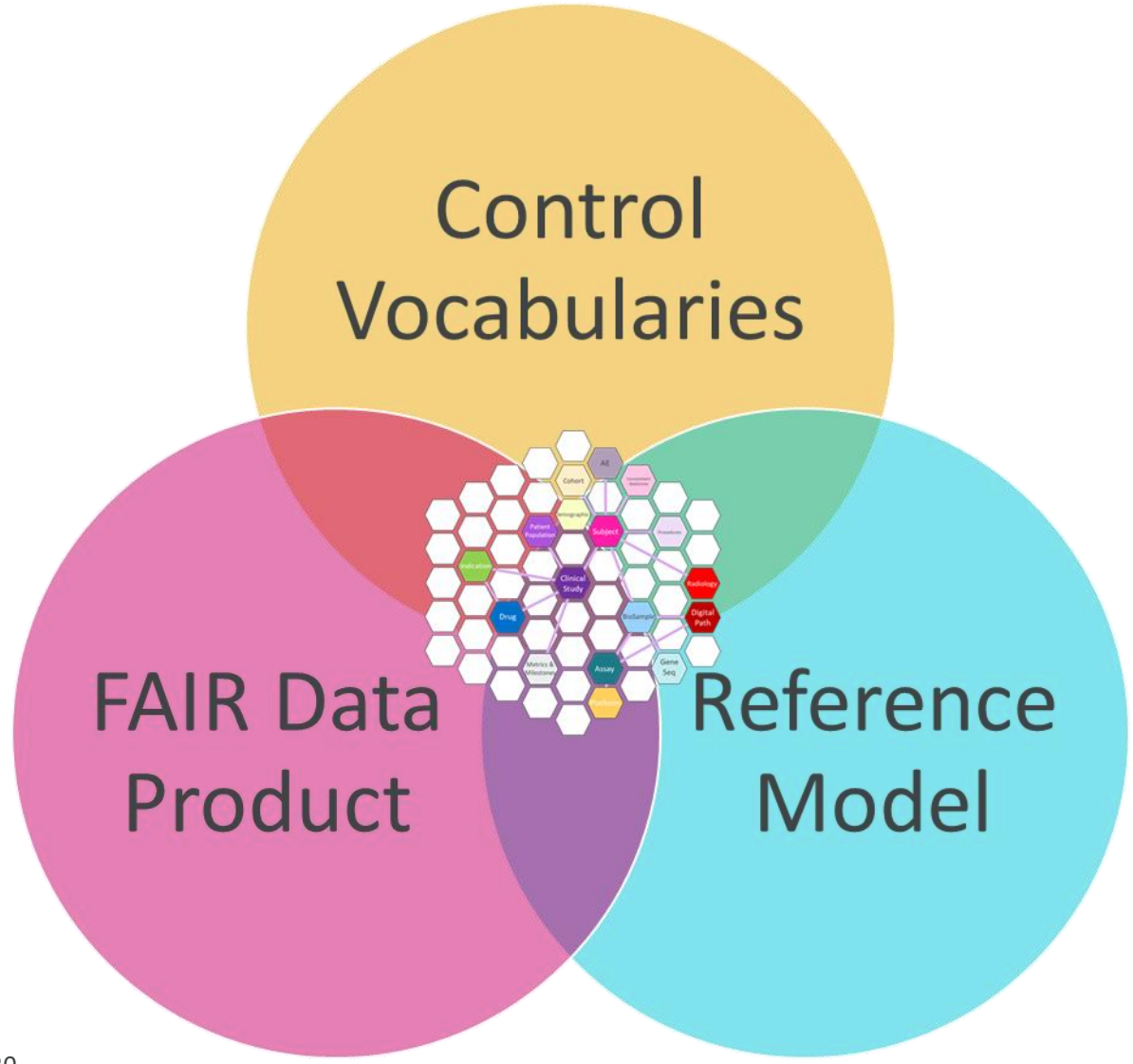
Amazon Prime

- **Digital, no longer physically constrained**
- **Recommendation engines** - other categories/products related to this, **push** information to you vs **pull** provides information on request
- **Scale** of products and variety of manufactures
- **Fuzzy search** - helps me find what I might be looking for - I'll know it when I find it searches
- **Amazon subscription services** - schedule delivery, Dash buttons
- Alexa - AI guided
- Services - Music and video
- **Market place for other vendors**

Credit to Gareth Henry, Sandra McGarry, Kerstin Forsberg and Ben Gardner

Building Knowledge Graphs

A three legged stool

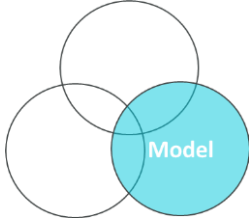


"Start with meaning"

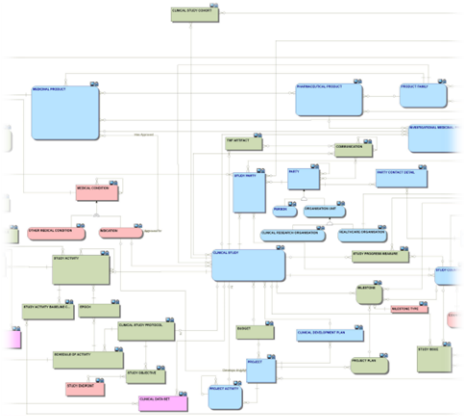
Dave McComb, Semantic Arts

Stool built by Derek Scuffell

Ontology Architecture overview



Conceptual Model

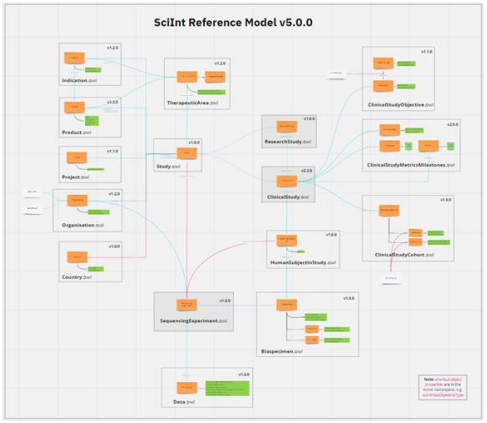


A conceptual model that provides the minimum linking of shared entities used across domain and application models.

Entity Ontology

Entity Ontologies describe individual concepts and act as building blocks for importing into application ontologies.

Application Ontology



Application Ontologies are built to support a specific application.

Honeycomb

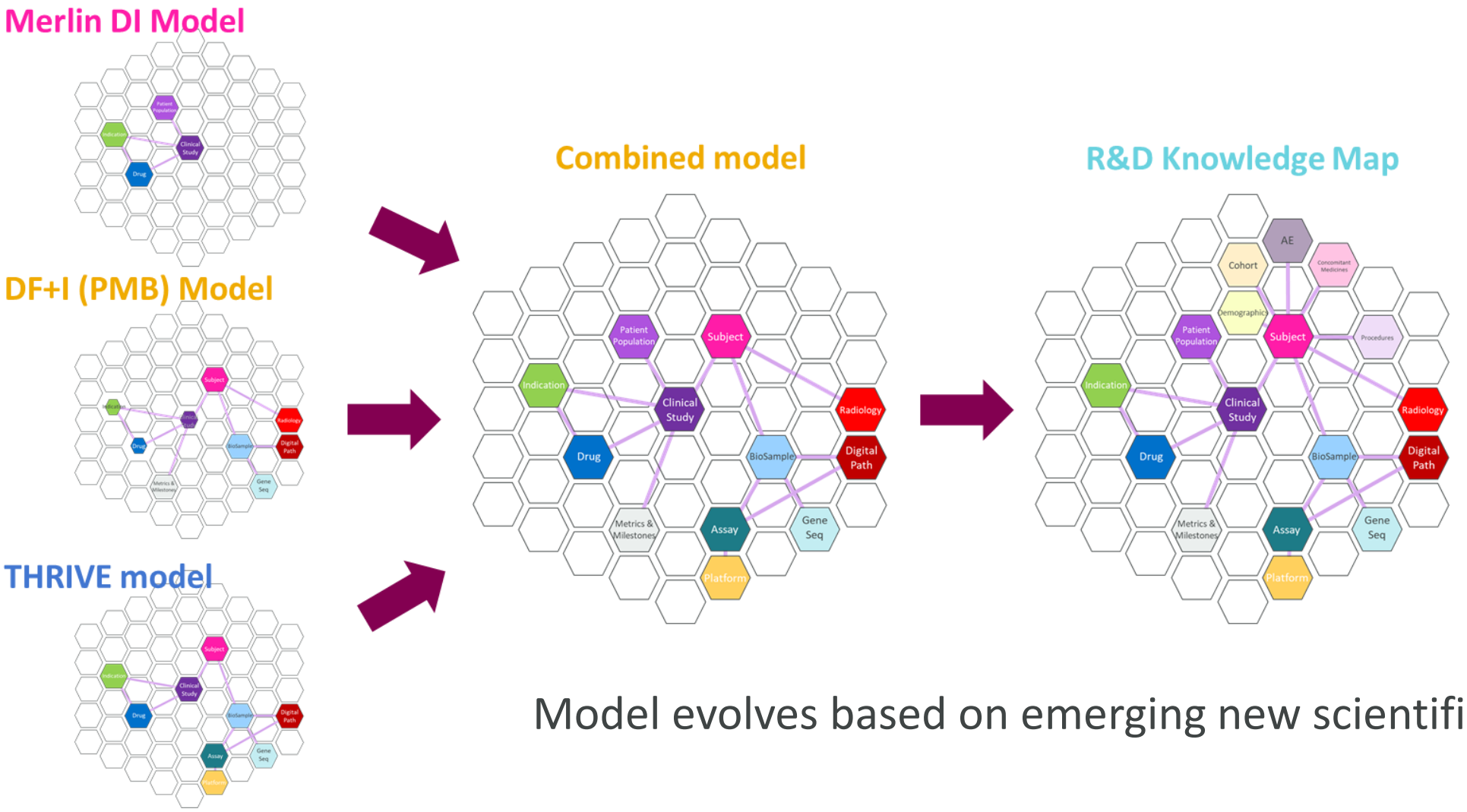
Honeycombs used to communicate concept of knowledge map, illustrate use case coverage and organic evolution.

Credit to Jon Ison

Honeycomb motif inspired by Dan Gschwend - Amgen's Enterprise Data Fabric, DCAF 2021

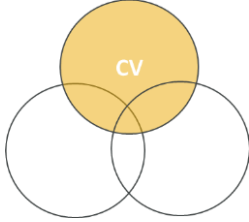
Strategy to organically grow a knowledge map

Increasing the breadth, depth and complexity of questions enabled



Model evolves based on emerging new scientific use cases

Foundational CV's deliver *Quality all the way up*



SKOS-XL

`https://pid.astrazeneca.net/ref/cvname/{ID}`

Silo'ed picklists

Uncoordinated list of strings used by an application

Local CV

Well designed atomic controlled vocabulary built to a common standard

Narrow coverage for a domain/application

Decided by use case specific governance, enabled by editorial capability & appropriate tools

Foundational CV

Well designed atomic controlled vocabulary built to a common standard

Broad coverage for multiple domains/applications

Decided by SMEs, enabled by specialist curators

Collections

A collection of terms derived from existing atomic controlled vocabularies that meet specific application needs/use cases

Credit to Varsha Khodiyar

Controlled Vocabularies

SKOS-XL supports a multitude of labels and signifiers

Preferred Labels

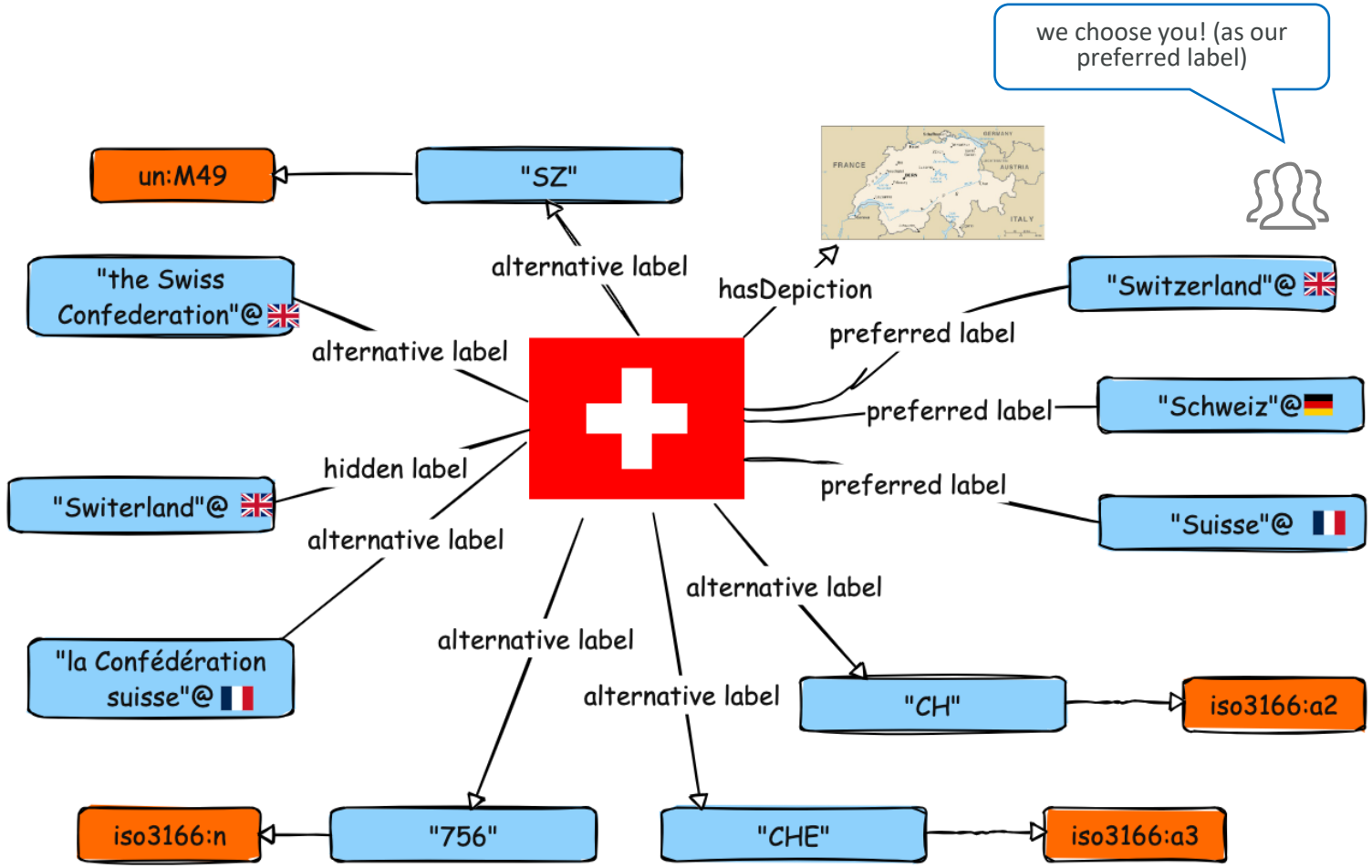
The preferred label for AstraZeneca, that resonates with the business vernacular

Alternative Labels

Well defined non-case variant, alternative labels that are used for this concept. – some may call these “synonyms”

Hidden Labels

Common mis-representations (spelling mistakes, etc) of the concept that exist and we don’t want used by humans. Often used to support NLP and AI activity

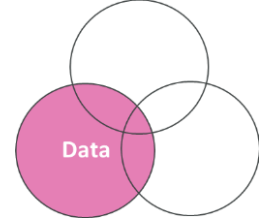


A pool of lexical labels exist for each concept. They are common use OR attributed to systems and vocabularies.
AZ curators decide which one will be preferred (for AZ) and whether other labels will be alternative or hidden.
Each label should be further characterized by a signifier.

Credit to Nathalie Conte, Derek Scuffell & Jon Ison – OWG Summer project team

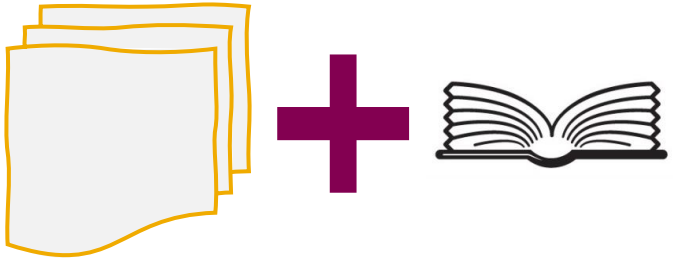
Data

Domain level FAIR (L4) a great first step



Data

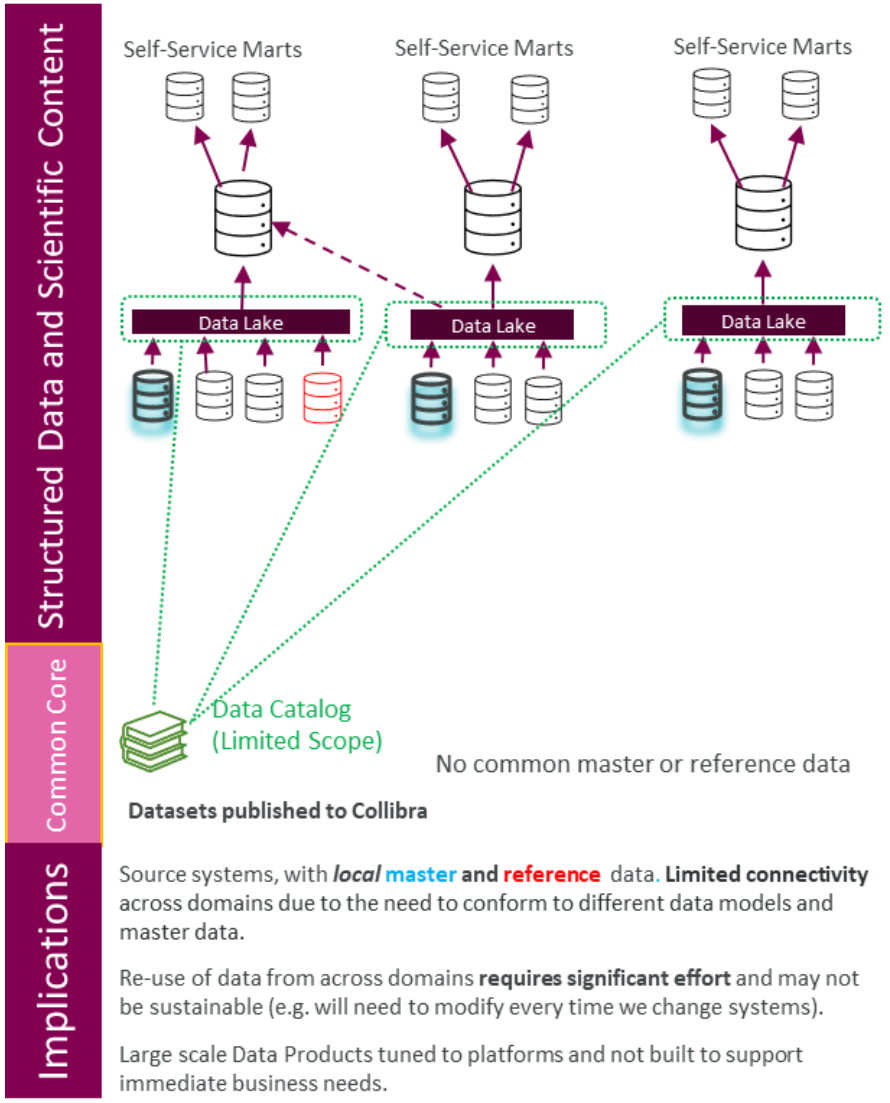
Catalogue



Find - Data registered in Collibra and tagged with CMM

Access – Access controlled via Collibra request service

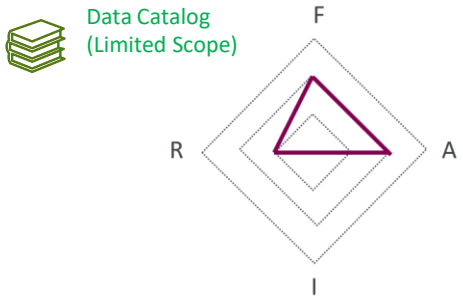
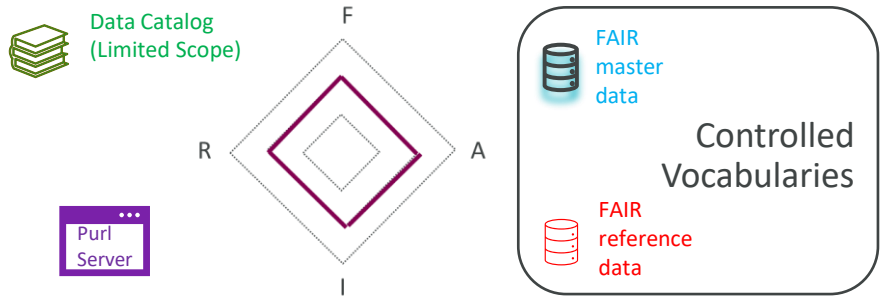
Reuse – Documentation captured against data



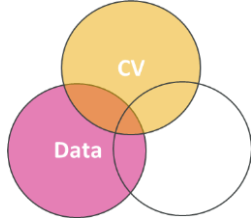
Credit to – Data Office Collibra team

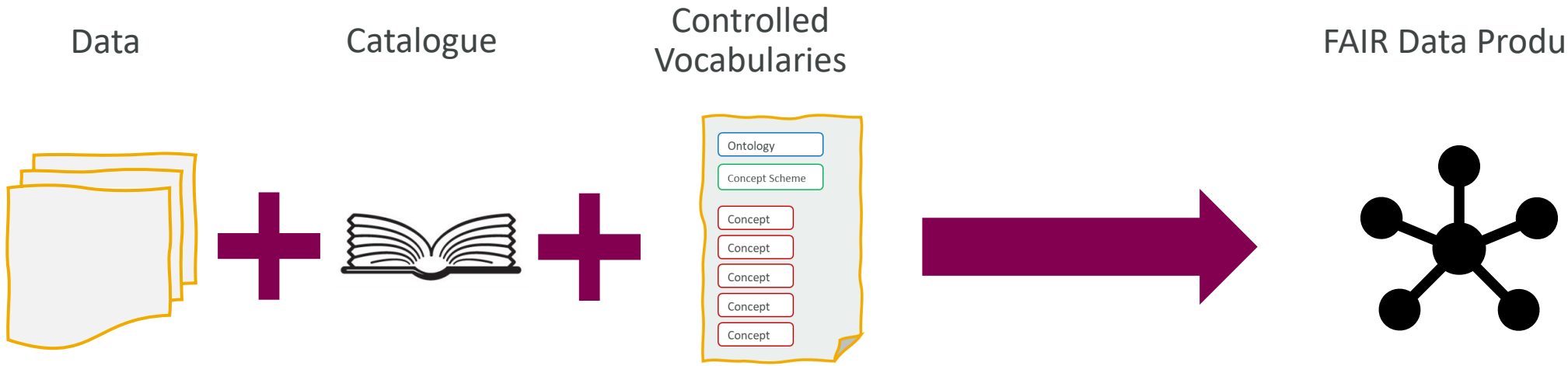
But we are FAR from FAIR

We MUST go further

FAIR metrics (Level 4)	FAIR metrics (Level 5)
	
Find - Data registered in Collibra and tagged with CMM **A**ccess – Access controlled via Collibra request service **R**euse – Documentation captured against data	**F**ind - Data registered in Collibra and tagged with CMM **A**ccess – Access controlled via Collibra request service **I**nteroperable – Data enhanced with shared CV and PIDs **R**euse – Documentation captured against data, includes data dictionary, etc
Data record discoverable in Collibra	Data record discoverable in Collibra Data enriched to create interoperability Data is Machine Readable

Enterprise Level FAIR (L5)
A FAIRe(nough) data product



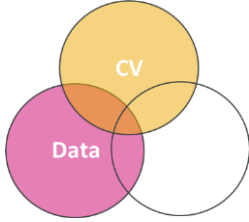


The minimum viable FAIR Data standard should deliver

Findable	Accessible	Interoperable	Reusable
Registered and discoverable in a Data Catalogue	Mechanism for requesting and receiving data	The data has been aligned to AZ standards where they exist	Documentation describing the constraints associated with using the data
		The PID for each instance in the standard is included to make the data machine readable	Documentation describing the data i.e. data dictionary, schema, etc

Data and Controlled Vocabularies

Putting Interoperability into FAIR



Dirty data

Study	Indication	Drug
D1234C00001	Non small cell lung cancer	Tagrisso
ADORA	NSCLC	Osimertinib
CP11278-CMA33G	Diabetes type 2	Forxiga

- Inconsistent identifiers & terms
- Column values can be concatenated
- etc

Data Transform

Term

prefLabel & PID

Controlled Vocabularies

Interoperable data

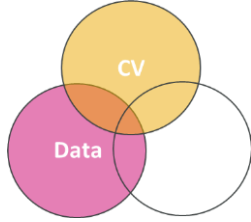
Study_ID	Study_ID__URI	Indication	Indication__URI	Drug	Drug__URI
D1234C00001	https://pid.astrazeneca.com/1/12345	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456	Tagrisso	https://pid.astrazeneca.com/Product/965723
D1234C00012	https://pid.astrazeneca.com/1/48373	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456	Tagrisso	https://pid.astrazeneca.com/Product/965723
D4568L00007	https://pid.astrazeneca.com/1/97538	Diabetes type 2	https://pid.astrazeneca.com/Indication/9857	Forxiga	https://pid.astrazeneca.com/Product/853584

prefLabel PID

- Shared Controlled Vocabularies
  - Enrich with preferred label and PIDs
- Uses common files format CSV, JSON, etc
- Is machine readable, graph enabling and relationally world friendly
- Well documented - Data Dictionary/Data Schema/etc

L5 FAIR Data Products benefit all

Inclusion of PIDs simplifies data integration irrespective of target data model



L5 FAIR Data Product

Study_ID	Study_ID__URI	Drug	Drug__URI
D1234C00001	https://pid.astrazeneca.com/1/12345	Tagrisso	https://pid.astrazeneca.com/Product/965723
D1234C00012	https://pid.astrazeneca.com/1/48373	Tagrisso	https://pid.astrazeneca.com/Product/965723
D4568L00007	https://pid.astrazeneca.com/1/97538	Forxiga	https://pid.astrazeneca.com/Product/853584

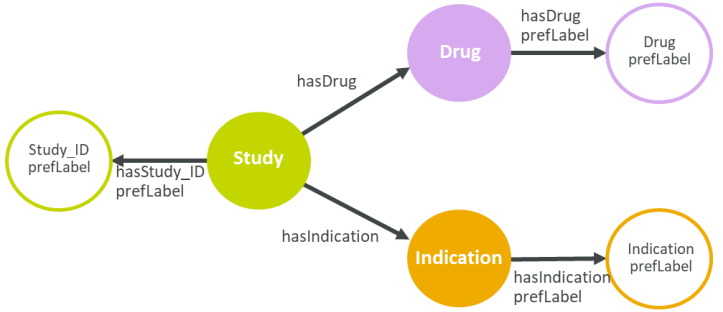
+

Study_ID	Study_ID__URI	Indication	Indication__URI
D1234C00001	https://pid.astrazeneca.com/1/12345	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456
D1234C00012	https://pid.astrazeneca.com/1/48373	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456
D4568L00007	https://pid.astrazeneca.com/1/97538	Diabetes type 2	https://pid.astrazeneca.com/Indication/9857

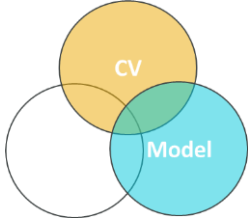
Relational

Study_ID	Study_ID__URI	Indication	Indication__URI	Drug	Drug__URI
D1234C00001	https://pid.astrazeneca.com/1/12345	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456	Tagrisso	https://pid.astrazeneca.com/Product/965723
D1234C00012	https://pid.astrazeneca.com/1/48373	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456	Tagrisso	https://pid.astrazeneca.com/Product/965723
D4568L00007	https://pid.astrazeneca.com/1/97538	Diabetes type 2	https://pid.astrazeneca.com/Indication/9857	Forxiga	https://pid.astrazeneca.com/Product/853584

Graph



Aligning entities with controlled vocabularies is key



Ontologies

(things and relationships)

Vocabularies

(terms that categorise things)

Build from

Application Ontology

Entity Ontology

Foundational Vocabularies

Build from

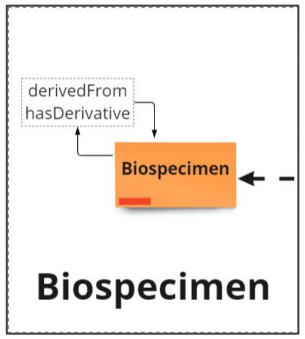
Collections

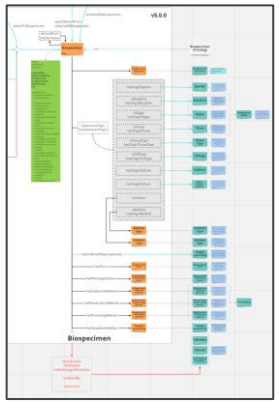
A model built from one or more entity ontologies and extended as required based on use case

Discrete models of things within a domain that provide the foundational building blocks

Atomic sets of terms that are used to describe the things identified within entity ontologies

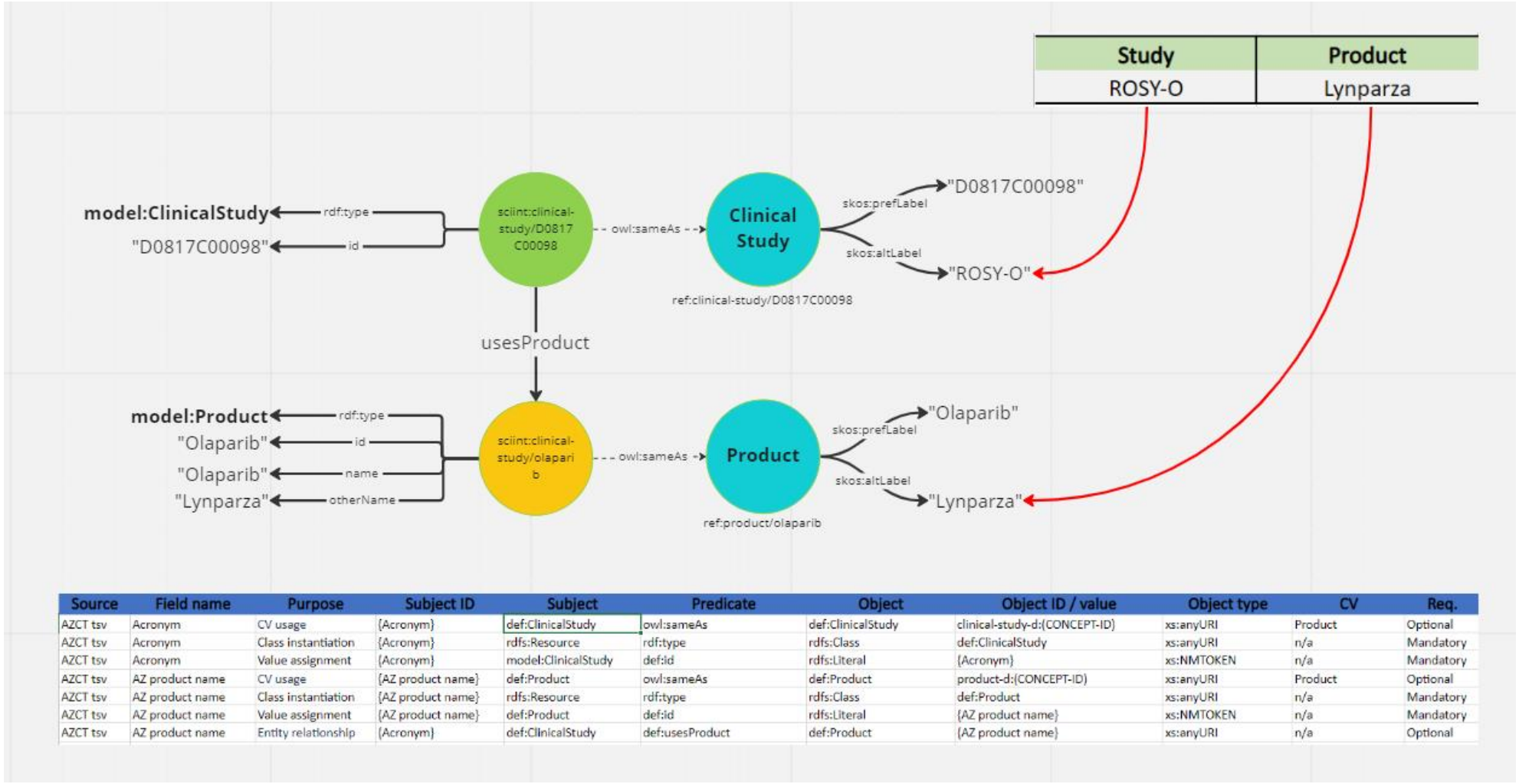
The collection of term sets and terms from one or more controlled vocabularies that are required to support a use case





Model and Data

Strings to things a mapping standard

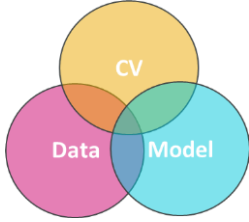


Source	Field name	Purpose	Subject ID	Subject	Predicate	Object	Object ID / value	Object type	CV	Req.
AZCT tsv	Acronym	CV usage	{Acronym}	def:ClinicalStudy	owl:sameAs	def:ClinicalStudy	clinical-study-d:{CONCEPT-ID}	xs:anyURI	Product	Optional
AZCT tsv	Acronym	Class instantiation	{Acronym}	rdfs:Resource	rdf:type	rdfs:Class	def:ClinicalStudy	xs:anyURI	n/a	Mandatory
AZCT tsv	Acronym	Value assignment	{Acronym}	model:ClinicalStudy	def:id	rdfs:Literal	{Acronym}	xs:NMTOKEN	n/a	Mandatory
AZCT tsv	AZ product name	CV usage	{AZ product name}	def:Product	owl:sameAs	def:Product	product-d:{CONCEPT-ID}	xs:anyURI	Product	Optional
AZCT tsv	AZ product name	Class instantiation	{AZ product name}	rdfs:Resource	rdf:type	rdfs:Class	def:Product	xs:anyURI	n/a	Mandatory
AZCT tsv	AZ product name	Value assignment	{AZ product name}	def:Product	def:id	rdfs:Literal	{AZ product name}	xs:NMTOKEN	n/a	Mandatory
AZCT tsv	AZ product name	Entity relationship	{Acronym}	def:ClinicalStudy	def:usesProduct	def:Product	{AZ product name}	xs:anyURI	n/a	Optional

Credit to – Jon Ison, Nicola Ellingham and Arinjay Jadeja

Knowledge Level FAIR (L6)

A FAIR data product minimises the gap between relational and graph worlds



L4 FAIR Data

L4 – Domain level FAIR

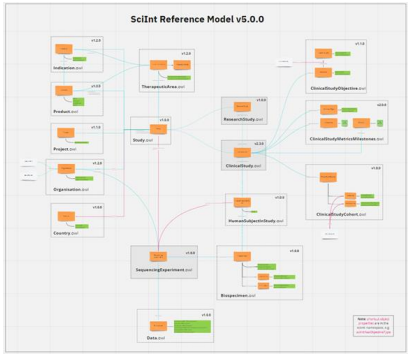
- **Data conformed, integrated, processed and audited** to support specific analytics patterns/enquiries
- Data is conformed using a **domain level data model**
- Data embeds **local master and reference data**
- **Controls on access at the data level**
- **Creation of analytics ready 'Marts'** enabling self service analytics: Citizen Data Scientist

Controlled Vocabularies

Ontology
Concept Scheme
Concept
Concept
Concept
Concept
Concept

vocab RDF file

Reference Model



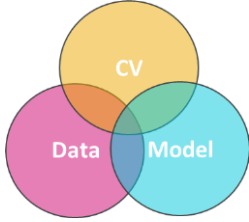
RDF Serialisation

Knowledge Map

L5 FAIR Data Product

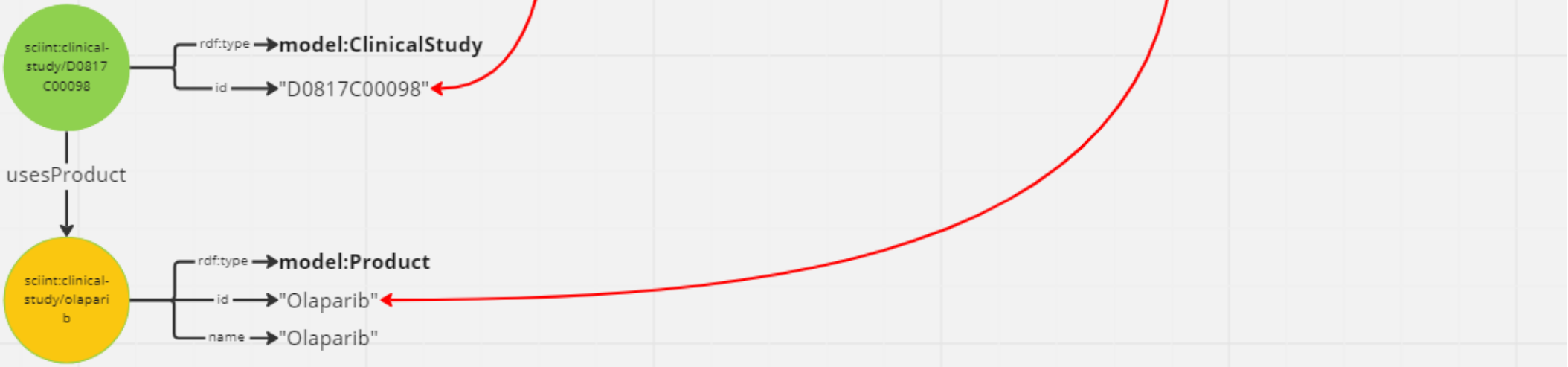
- An enterprise standard
- Adds value for everyone
- Two thirds of the way to graph without impacting non-graph consumers

Model, Controlled Vocabularies and Data



In a data-centric world the PIDs just snap the data and model together

Study	Study ID	Product	Product ID
D0817C00098	ref:clinical-study/D0817C00098	Olaparib	ref:product/olaparib

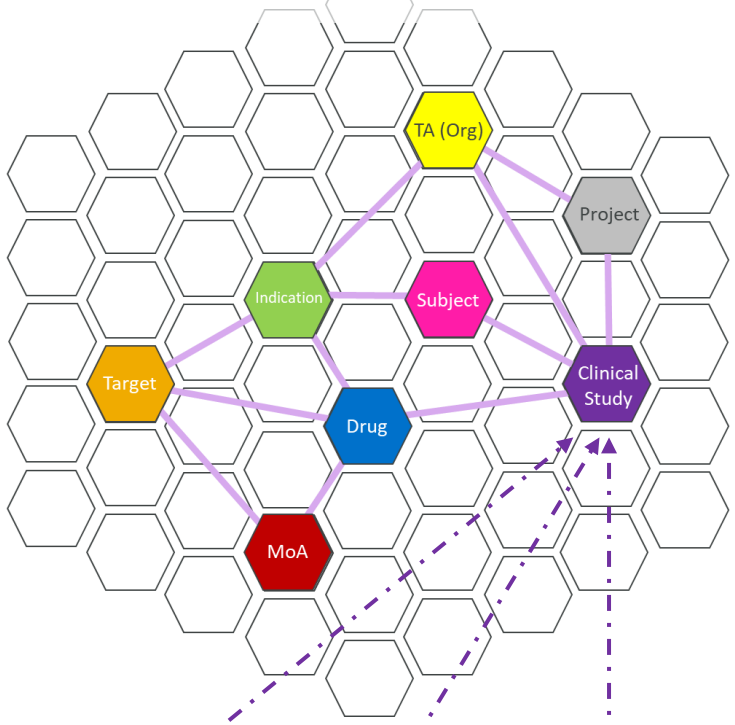


From System-centric to Data-centric

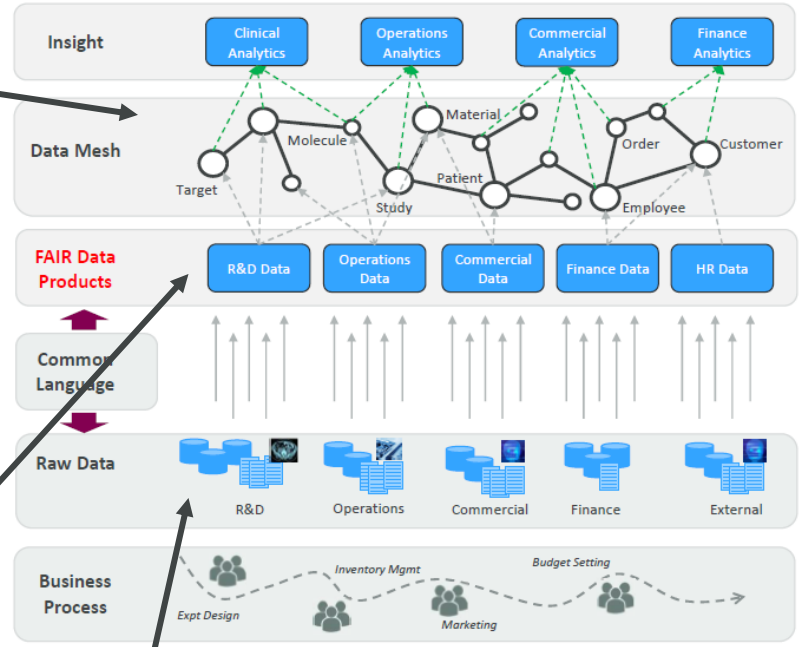
Knowledge Map

(Data-centric)

Data is FAIR and aggregated by business concept (semantic)



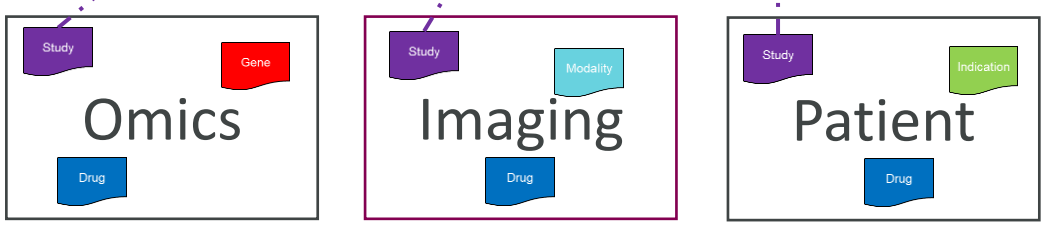
Reference Architecture



Platform Data Products

(Platform-centric)

Data is FAIR but fragmented across domains



Application Data

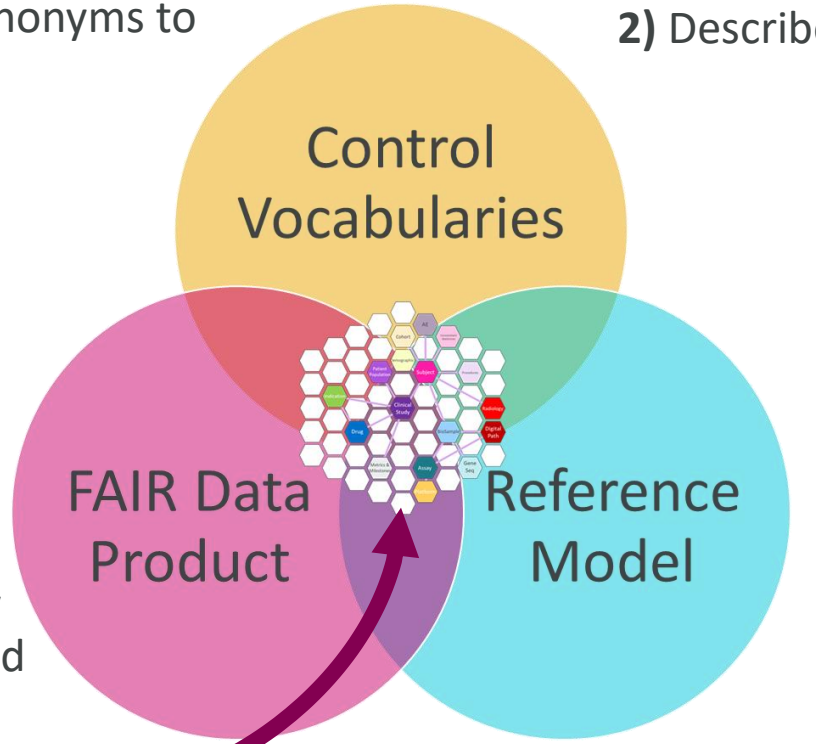
(System-centric)

Data is aligned around processes

Bringing it back together

3) Map terms and synonyms to unique identifier

2) Describe the things identified



1) Start with meaning by identifying the things that matter

4) Create interoperability by enriching data with preferred term and unique identifier

5) Populate the knowledge graph

Reuse data

Take home messages

Information Discovery

- Evolving science requires greater granularity
- New demands on our data
- Librarians, librarians, librarians

Common patterns

- Open source is great, but it hurts
- Blend your skill sets, relational and graph
- Design for extensibility

Data Interoperability

- Invest in Controlled Vocabularies
- Editorial governance is critical
- Hide the semantics if you want adoption

It takes village ……

DF+I

Daniel Roythorne
Jon Ison
Nathalie Conte
Nicola Ellingham
Arun Balaji
Induja Mohan
Arinjay Jadeja
Ben Gardner
Hans Ienasescu
Bhavna Khilnani
Michael Neylon
Mathew Woodwark
Rob Hernandez

Collaborators

Derek Scuffell
Varsha Khodiyar
Pablo Porras Millan
John Berrisford
Bijay Jassal
Rafa Jimenez
Philippe Rocca-Serra
Victor Kim
Alex Wood
Linda Zander-Balderud
Antonio Fabregat
Mark Reuter
Tom Plasterer
James Holman
Martina Devoti
Stacy Mather
Di Elvers
Colin Wood
Sandra Mc Garry
Gareth Henry
Kerstin Freberg
Calle Nordmark