FAIR Data Implementation Progress Report

Tom Plasterer[1] & Ben Gardner[2]

[1]Strategy, Business Development & Alliances, Oncology R&D, AstraZeneca, Boston, MA, United States

[2]Data Standards & Interoperability, Data Office, Data Science & Artificial Intelligence, R&D, AstraZeneca, Cambridge, United Kingdom

November 2023

What is Data Centricity?

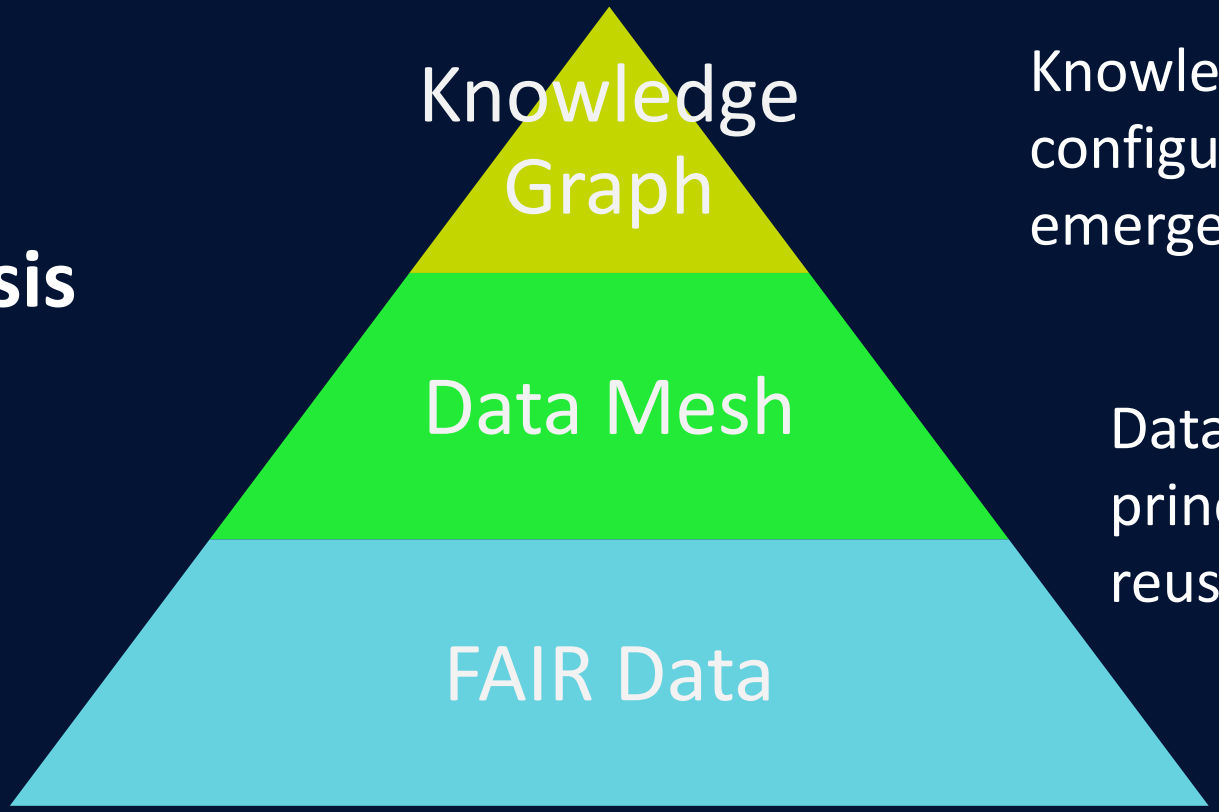
Data-Centricity puts **data at the centre of the enterprise**.

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.

Data Centricity Enables Knowledge Applications

FAIR Data is the Foundation for all inquiry and analysis



Knowledge Graph(s) and configurable apps/analytics emerge

Data Mesh utilizes FAIR principles for interoperable, reusable Data Products

Data-centricity can be realized with a commitment to a **FAIR Data foundation** and a loosely-coupled **Data Mesh** manifested in a **Knowledge Graph**.

The value of data centricity

- **Data conformed, integrated, processed and audited** to support specific consumption patterns
- 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

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

L4

L6

- Digital, no longer physically constrained
- Recommendation engines - other categories/products related to this, push information to you vs pull / provide information on
- Scale of products and variety of manufacturers
- Fuzzy search - helps me find what I'm 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**

Extended L5 functionality:
- **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**

L5

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

- Extended L4 functionality
- **Data is described in a cross domain data or industry model**
- and is integrated/mapped 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 serving all [illegible]

- **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

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

L2

L3

- **Conformed data catalogued** and available in a data lake
- **Data is conformed on a system by system basis** [illegible] data model
- Controls on access at the system level
- Grouping by product category i.e. spices vs carpets vs etc
- **Improved level of quality**
- Seller can explain the provenance of the product
- **Requires average level technical and subject matter knowledge** to use: Data Scientist.

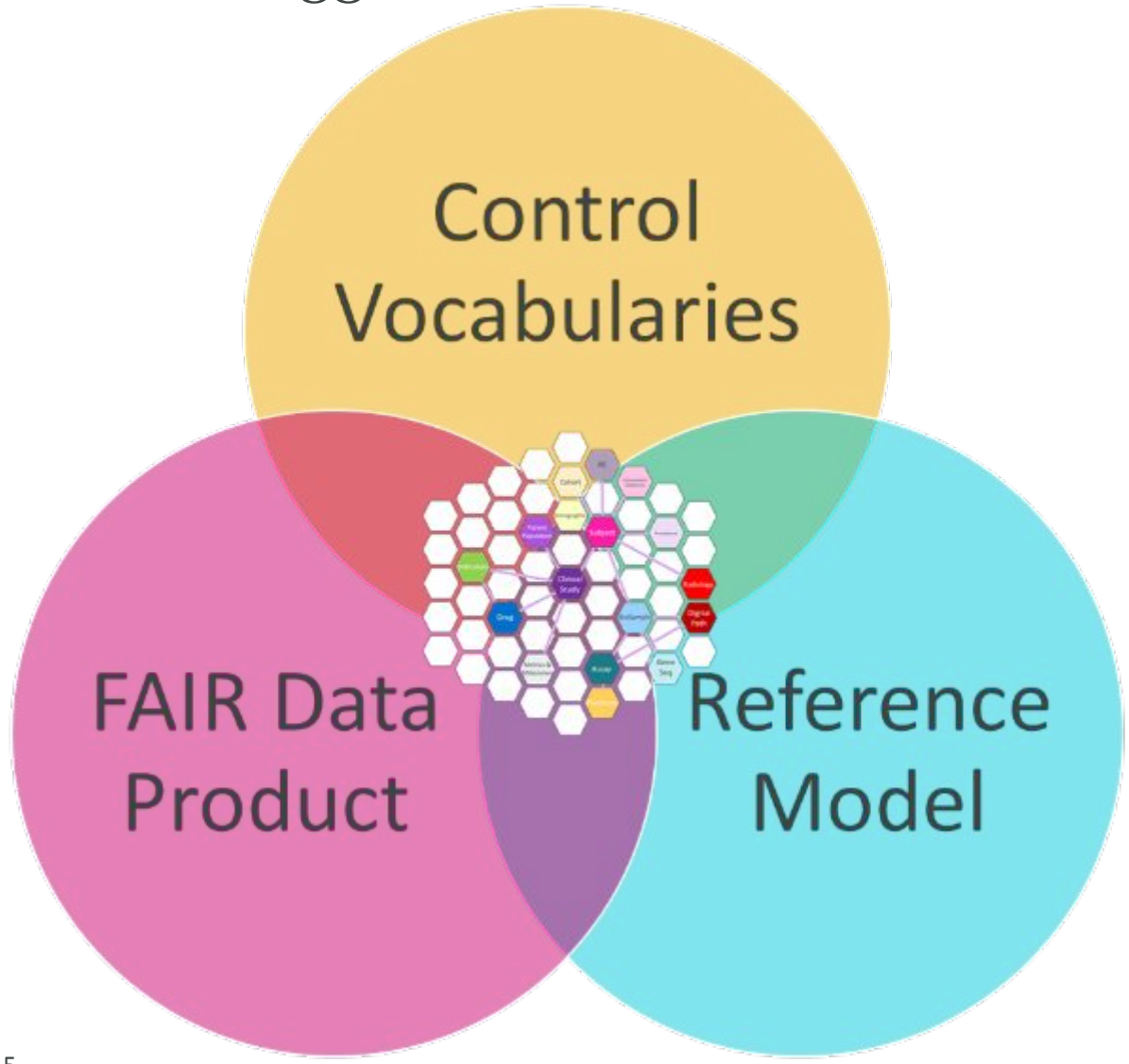
L1

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

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

Building Knowledge Graphs

A three-legged stool

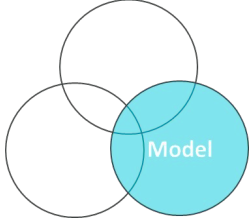


"Start with meaning"

Dave McComb, Semantic Arts

Stool built by Derek Scuffell

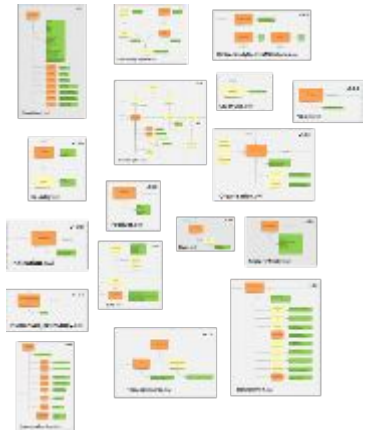
Ontology Architecture overview



Conceptual Model

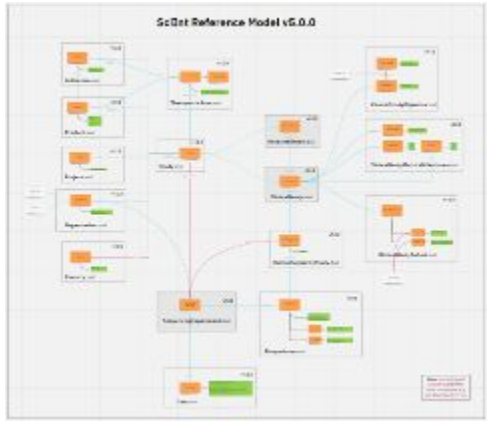
A conceptual model that provides the minimum linking of shared entities used across domain and application models.

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

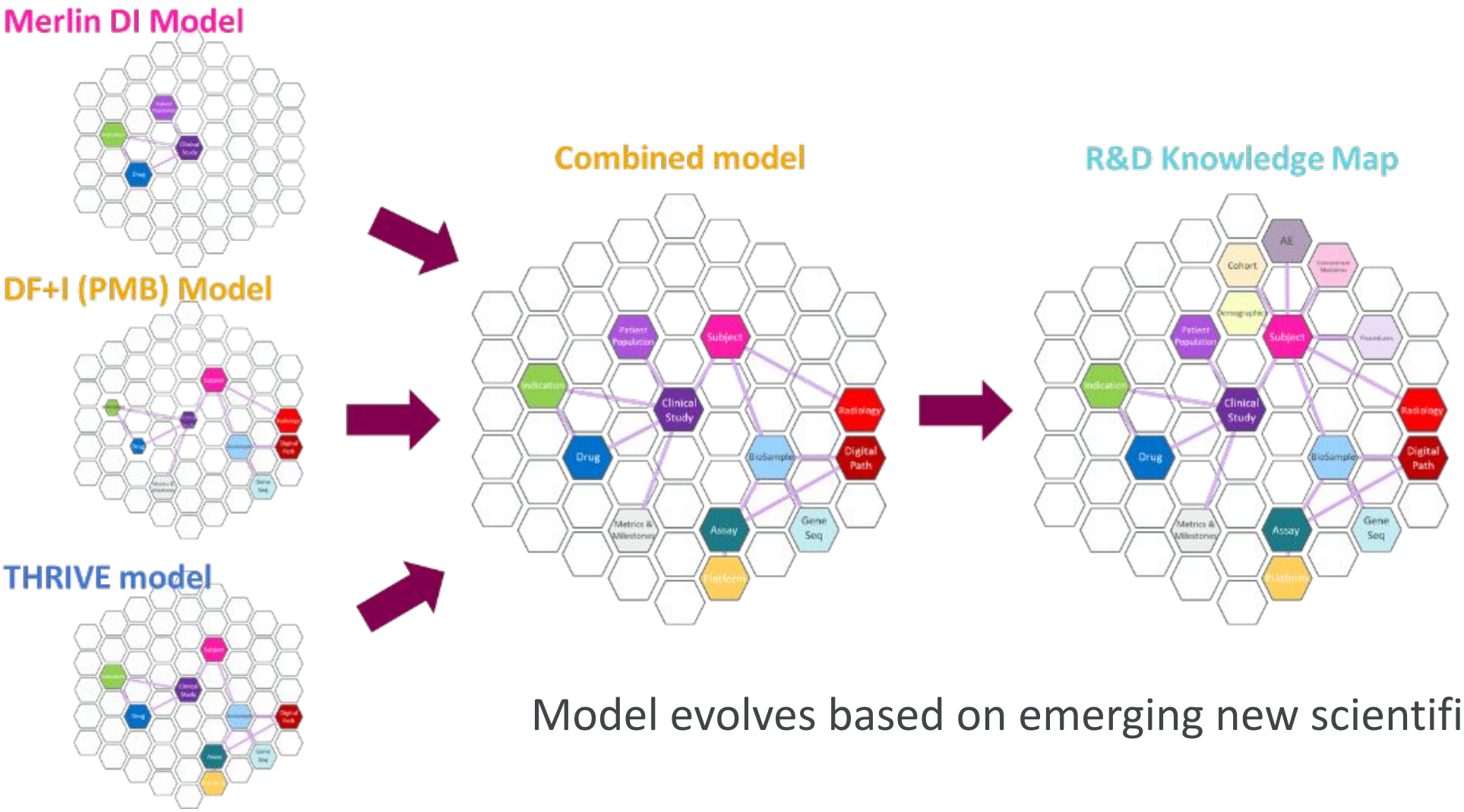
Honeycombs used to communicate concept of knowledge map, illustrate use case coverage and organic evolution.

Credit to Jon Ison

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

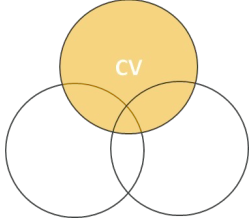
Strategy to organically grow a knowledge map

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

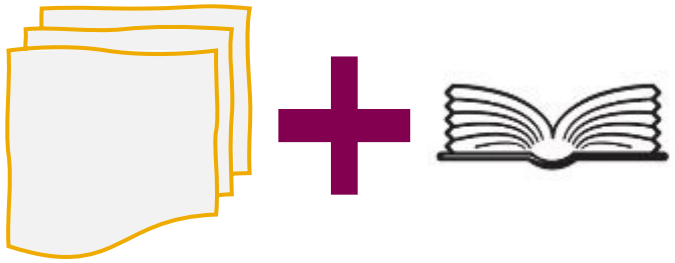
Credit to Varsha Khodiyar

Data

Domain level FAIR (L4) a great first step

Data

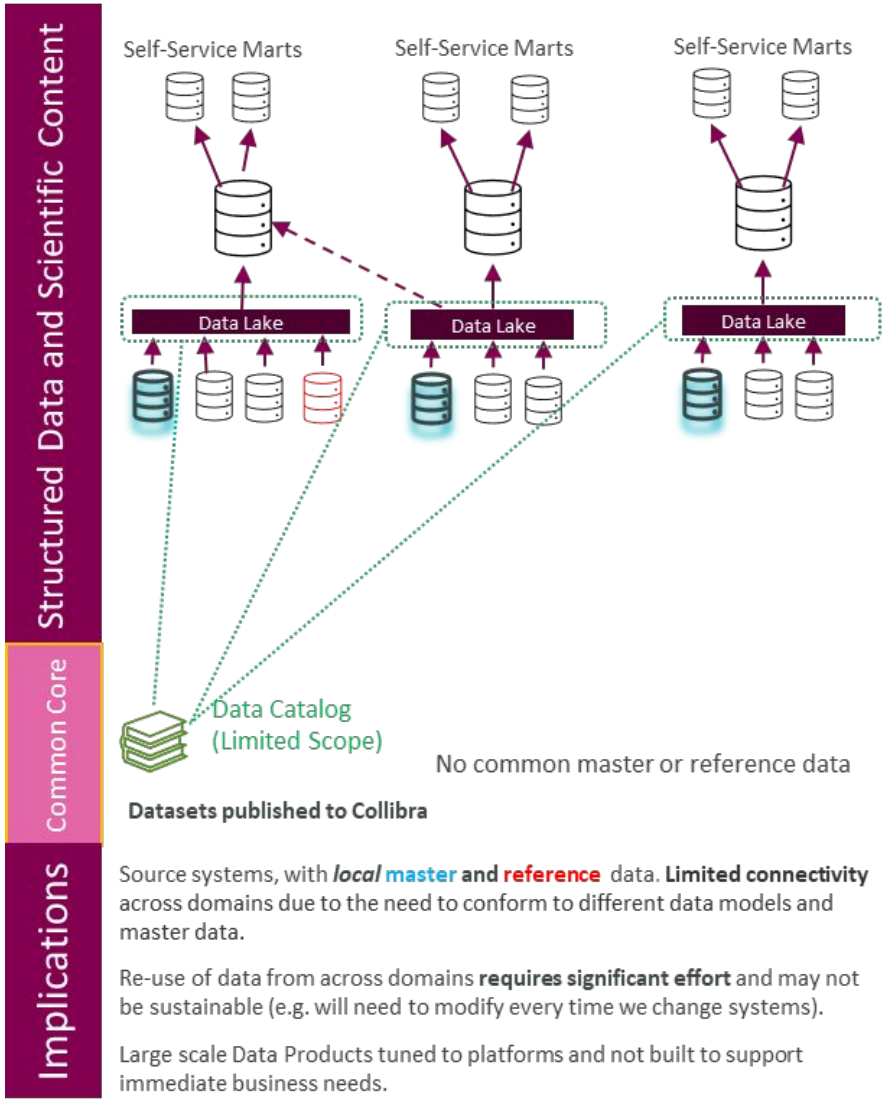
Catalogue



Find - Data registered in Collibra and tagged with CMM

Access – Access controlled via Collibra request service

Reuse – Documentation captured against data

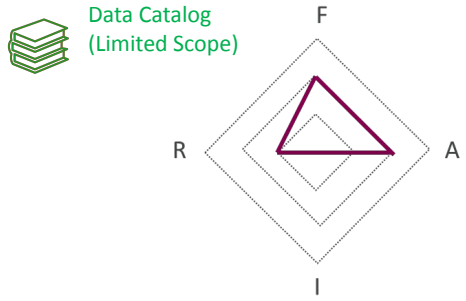


Credit to – Data Office Collibra team

But we are FAR from FAIR

We MUST go further

FAIR metrics (Level 4)



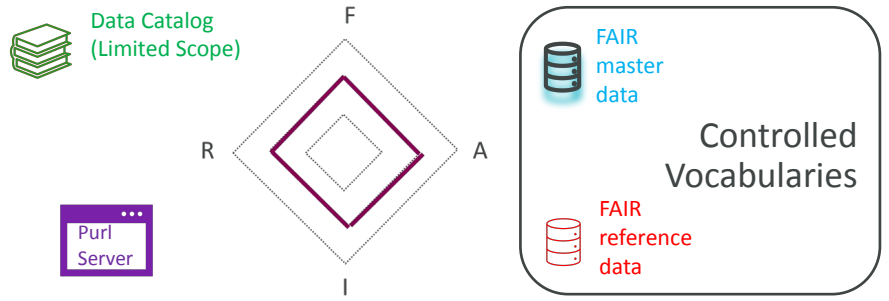
Find - Data registered in Collibra and tagged with CMM

Access – Access controlled via Collibra request service

Reuse – Documentation captured against data

Data record discoverable in Collibra

FAIR metrics (Level 5)



Find - Data registered in Collibra and tagged with CMM

Access – Access controlled via Collibra request service

Interoperable – Data enhanced with shared CV and PIDs

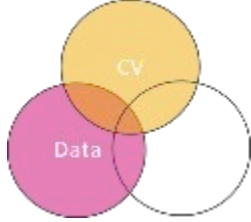
Reuse – Documentation captured against data, includes data dictionary, etc

Data record discoverable in Collibra

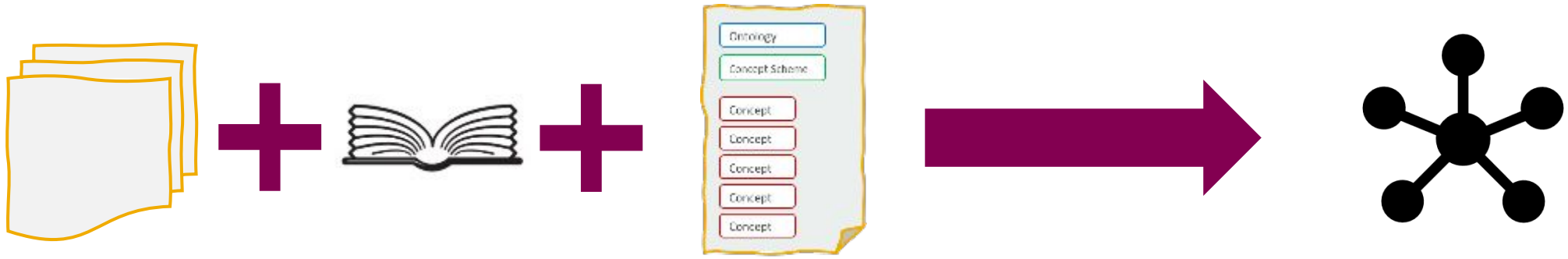
Data enriched to create interoperability

Data is Machine Readable

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



Data + Catalogue + Controlled Vocabularies → FAIR Data Product

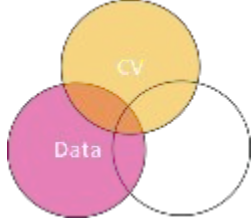


The minimum viable FAIR Data standard should deliver

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

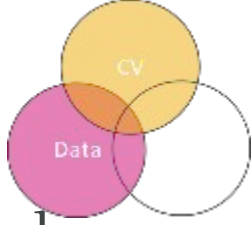
Study_ID	Study_ID__URI	Indication	Indication__URI	Drug	Drug__URI
D1234C00001	https://pid.astrazeneca.com/1/12345	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456	Tagrisso	https://pid.astrazeneca.com/Product/965723
D1234C00012	https://pid.astrazeneca.com/1/48373	Non small cell lung 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

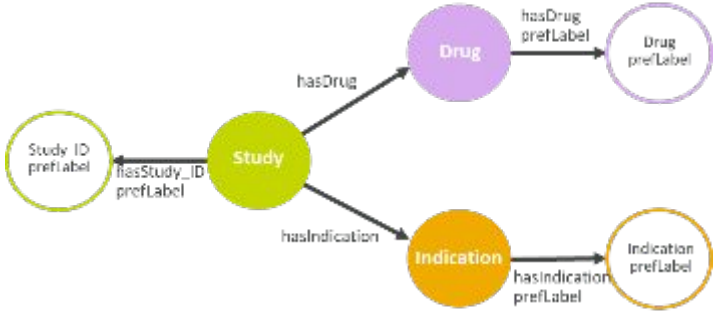
+

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

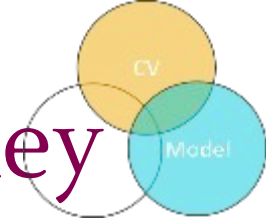
Relational

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

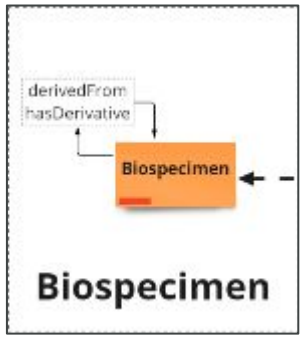
Collections

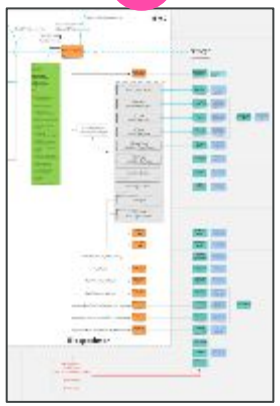
A model built from one or more entity ontologies and extended as required based on use case

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

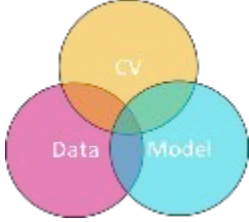
The collection of term sets and terms from one or more controlled vocabularies that are required to support a use case





Knowledge Level FAIR (L6)

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

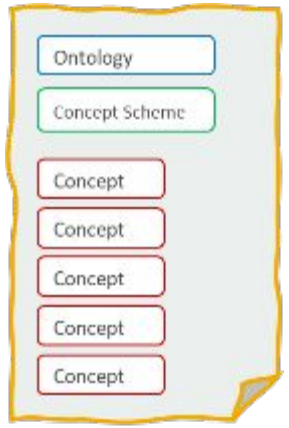


L4 FAIR Data

L4 – Domain level FAIR

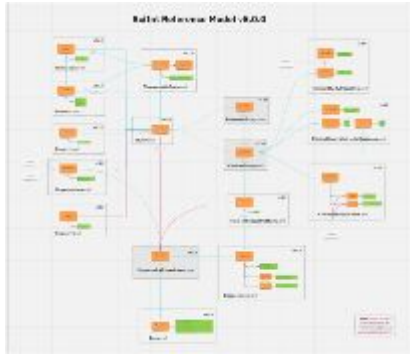
- **Data conformed, integrated, processed and audited** to support specific analytics patterns/enquiries
- Data is conformed using a **domain level data model**
- Data embeds **local master and reference data**
- **Controls on access at the data level**
- **Creation of analytics ready 'Marts'** enabling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

Controlled Vocabularies



vocab RDF file

Reference Model



RDF Serialisation

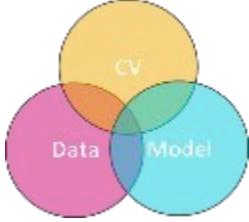
Knowledge Map

L5 FAIR Data Product

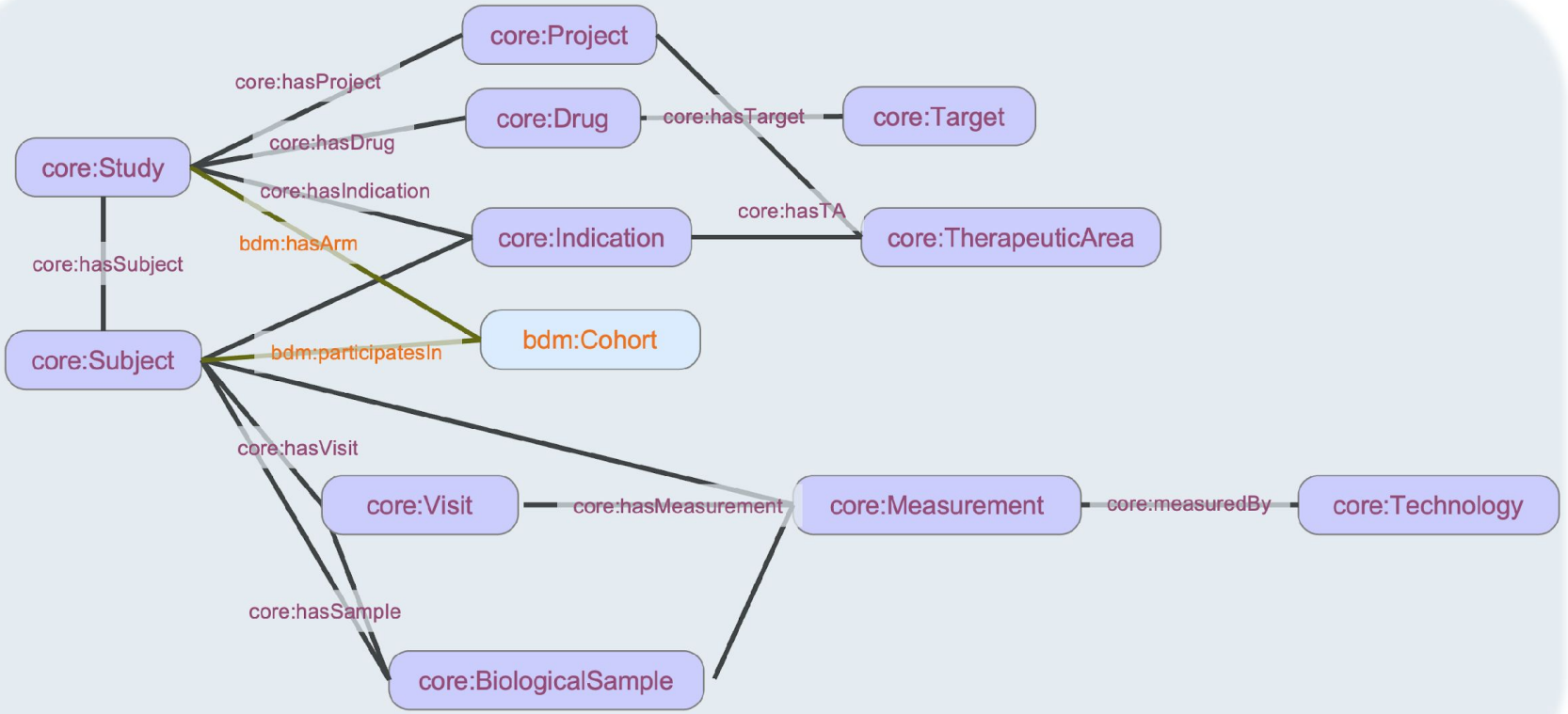
- An enterprise standard
- Adds value for everyone
- Two thirds of the way to graph without impacting non-graph consumers

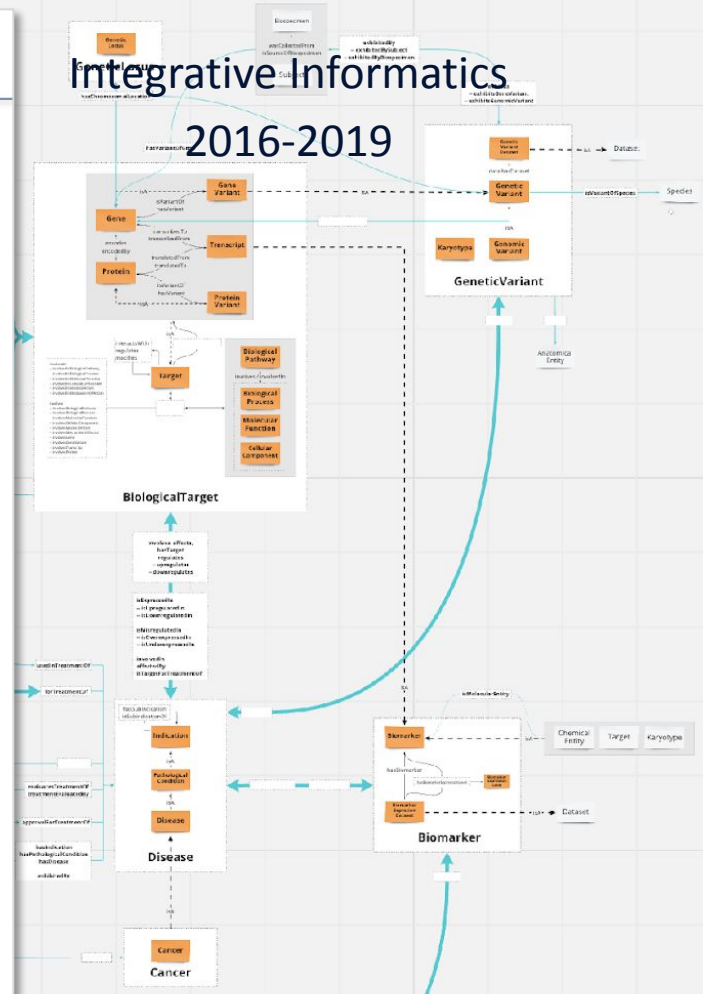
Accepting Change

You may need to give up your tactical solution

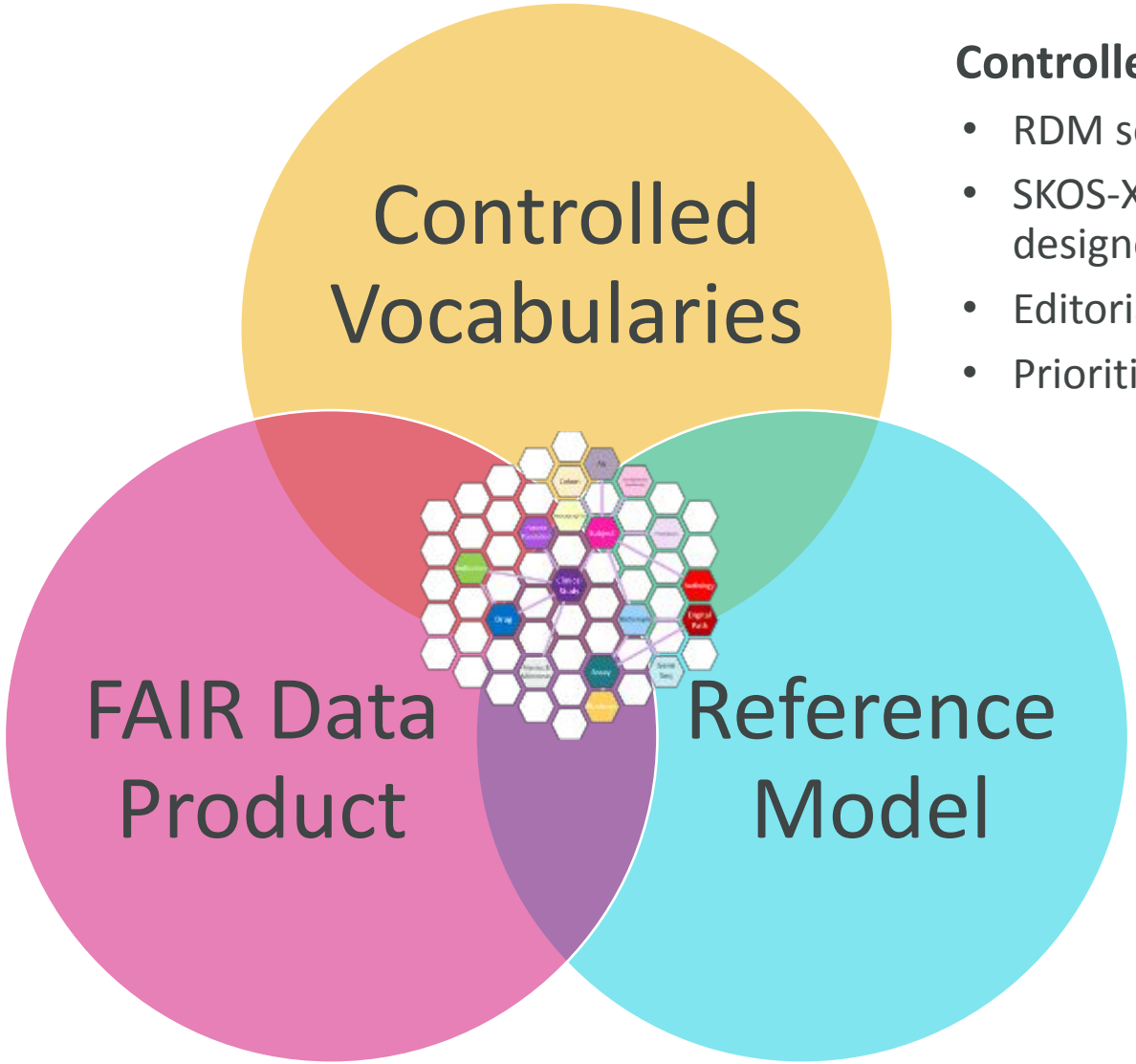


Starting Point: Modeling Business Questions





Bringing it all back together



Controlled Vocabularies

- RDM service implemented
- SKOS-XL based CV framework & workflow designed
- Editorial controls in place
- Prioritisation & creation of CV underway

FAIR Data

- FAIR metrics goal 80% Data FAIR by 2025
- Standards, resources, tooling and governance
- FAIR assessment in place
- PID server and standard patterns agreed

Ontology Architecture

- Scalable and sustainable
- Reusable library of entity ontologies
- Organic evolution of reference ontology

Across AstraZeneca Acknowledgements

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

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
Justin Johnson

Mark Reuter
Tom Plasterer
James Holman
Martina Devoti
Stacy Mather
Di Elvers
Colin Wood
Sandra Mc Garry
Gareth Henry
Kerstin Forsberg
Calle Nordmark

Questions?