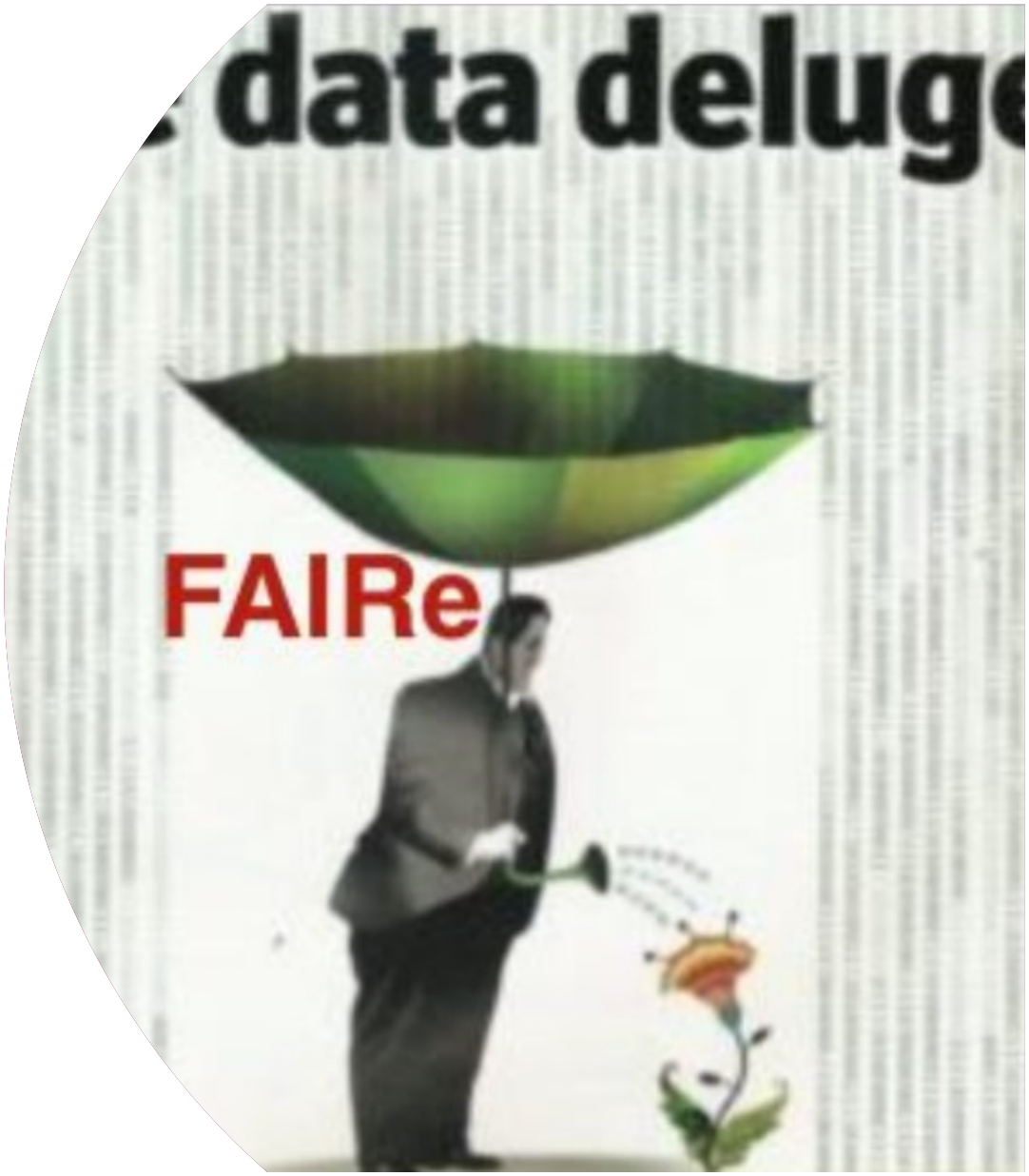
AstraZeneca

Measuring FAIR in AZ: Lessons learned from implementation of FAIR benchmarks

Pablo Porras Millán, PhD

Data Curation Director, AZ

18/04/2023



Findable Accessible Interoperable Reusable

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...]Barend Mons

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

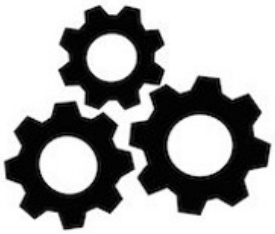
334k Accesses | **2882** Citations | **1902** Altmetric | Metrics

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE



Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

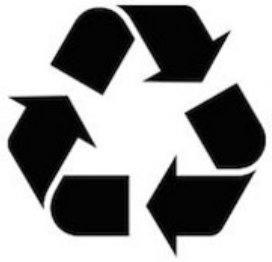
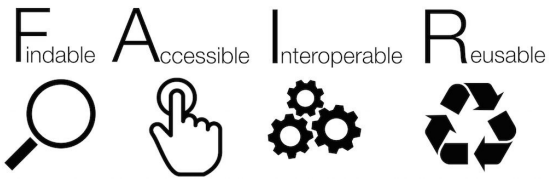
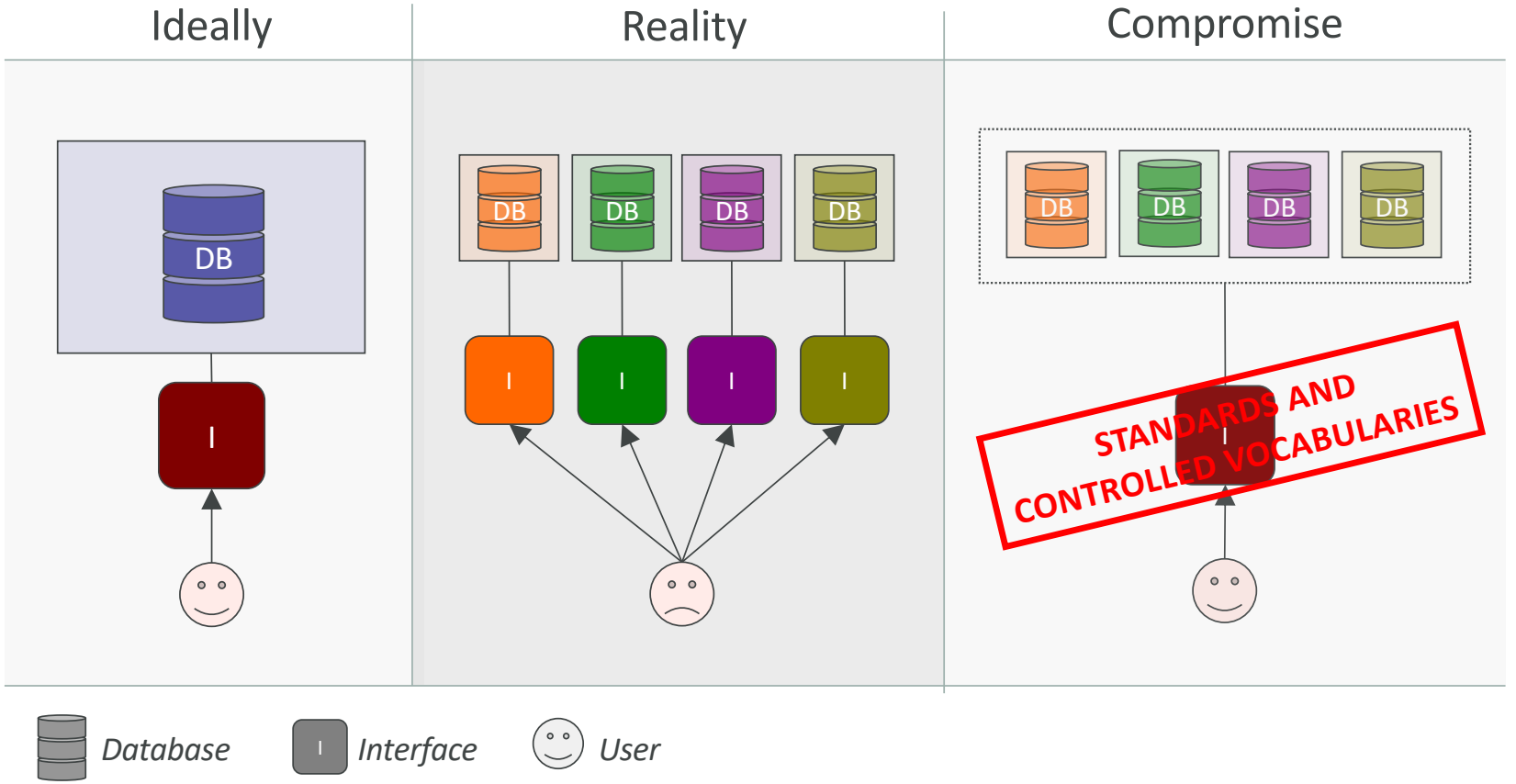


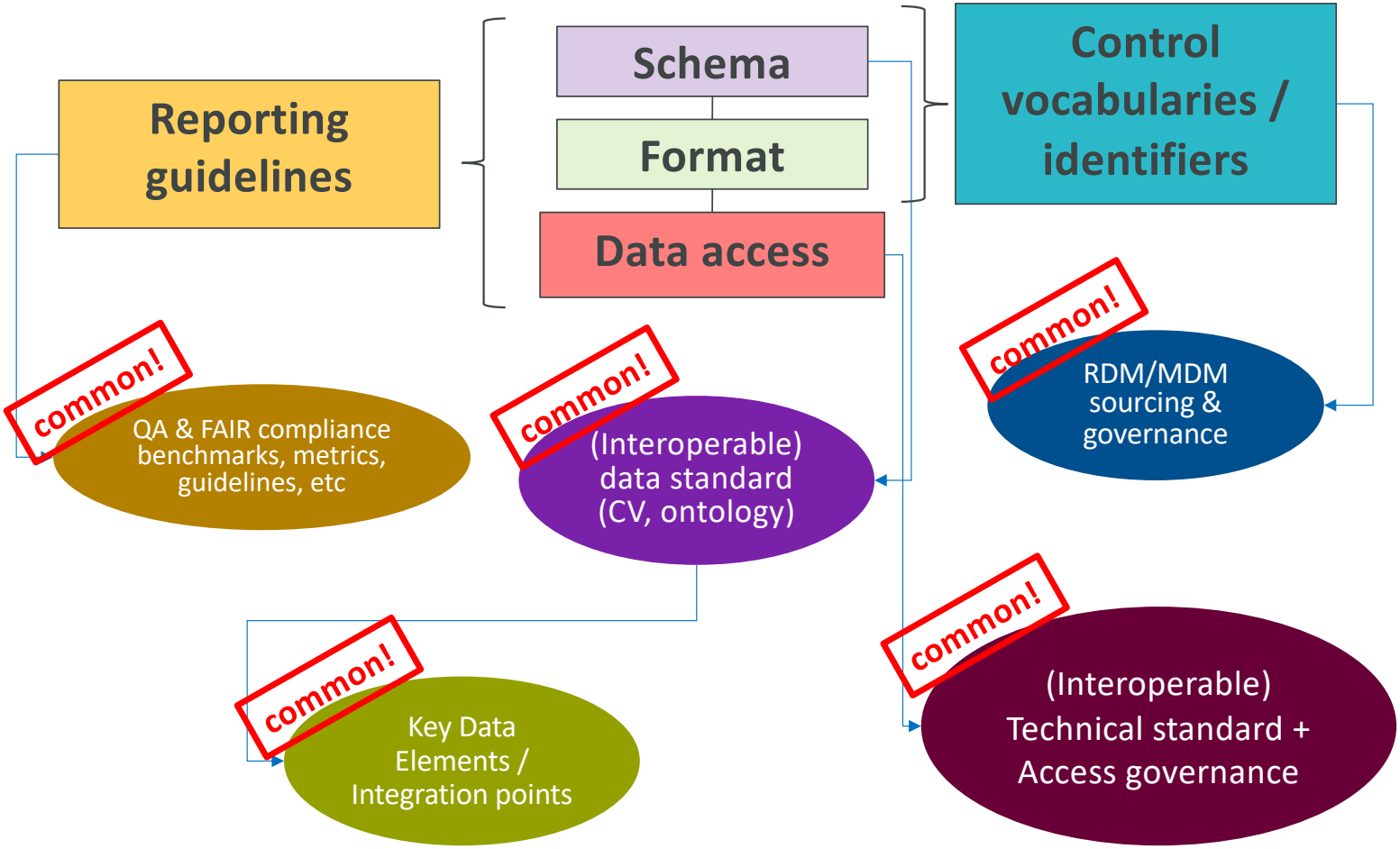
Image from https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf

Application centricity:
A problem of data integration

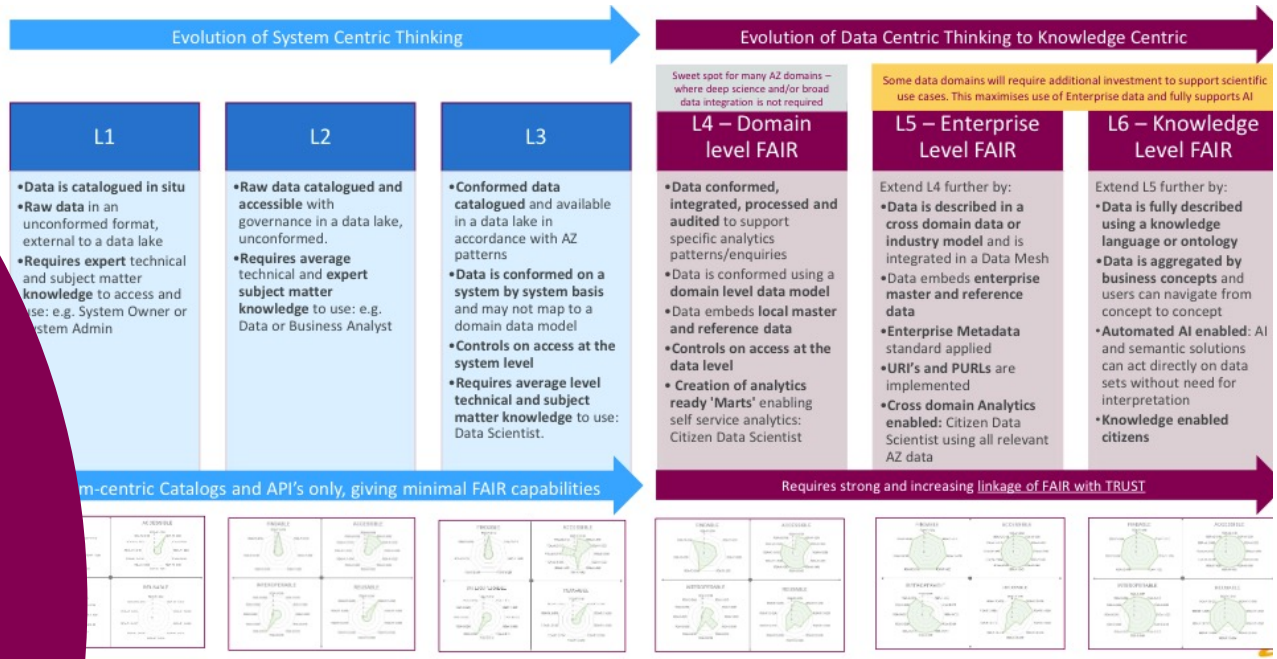




What do we need to achieve FAIR?:
The need for structured data and its consequences



AZ Enterprise FAIR levels



FAIR indicators, defined externally, can be used to benchmark and measure each level

ID		PRINCIPLE	INDICATOR_ID	INDICATORS	PRIORITY	METRIC	VIZ	
1	F	F1	RDA-F1-01M	Metadata is identified by a persistent identifier	Essential	1 – not being considered this yet	1	0
2		F1	RDA-F1-01D	Data is identified by a persistent identifier	Essential	2 – under consideration or in planning phase	2	0
3		F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	Essential	1 – not being considered this yet	1	0
4		F1	RDA-F1-02D	Data is identified by a globally unique identifier	Essential	2 – under consideration or in planning phase	2	0
5		F2	RDA-F2-01M	Rich metadata is provided to allow discovery	Essential	1 – not being considered this yet	1	0
6		F3	RDA-F3-01M	Metadata includes the identifier for the data	Essential	1 – not being considered this yet	1	0
7		F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	Essential	1 – not being considered this yet	1	0
8	A	A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	Important	1 – not being considered this yet	1	0
9		A1	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	Essential	3 – in implementation phase	3	0
10		A1	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	Essential	3 – in implementation phase	3	0
11		A1	RDA-A1-03M	Metadata identifier resolves to a metadata record	Essential	1 – not being considered this yet	1	0
12		A1	RDA-A1-03D	Data identifier resolves to a digital object	Essential	3 – in implementation phase	3	0
13		A1	RDA-A1-04M	Metadata is accessed through standardised protocol	Essential	1 – not being considered this yet	1	0
14		A1	RDA-A1-04D	Data is accessible through standardised protocol	Essential	3 – in implementation phase	3	0
15		A1	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	Important	3 – in implementation phase	3	0
16		A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	Essential	1 – not being considered this yet	1	0
17		A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	Important	3 – in implementation phase	3	0
18		A1.2	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	Useful	3 – in implementation phase	3	0
19		A2	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	Essential	1 – not being considered this yet	1	0
20	I	I1	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	Important	1 – not being considered this yet	1	0
21		I1	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	Important	1 – not being considered this yet	1	0
22		I1	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	Important	1 – not being considered this yet	1	0
23		I1	RDA-I1-02D	Data uses machine-understandable knowledge representation	Important	1 – not being considered this yet	1	0
24		I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	Important	1 – not being considered this yet	1	0
25		I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies	Useful	2 – under consideration or in planning phase	2	0
26		I3	RDA-I3-01M	Metadata includes references to other metadata	Important	1 – not being considered this yet	1	0
27		I3	RDA-I3-01D	Data includes references to other data	Useful	3 – in implementation phase	3	0
28		I3	RDA-I3-02M	Metadata includes references to other data	Useful	1 – not being considered this yet	1	0
29		I3	RDA-I3-02D	Data includes qualified references to other data	Useful	3 – in implementation phase	3	0
30		I3	RDA-I3-03M	Metadata includes qualified references to other metadata	Important	1 – not being considered this yet	1	0
31		I3	RDA-I3-04M	Metadata include qualified references to other data	Useful	1 – not being considered this yet	1	0
32	R	R1	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	Essential	3 – in implementation phase	3	0
33		R1.1	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	Essential	1 – not being considered this yet	1	0
34		R1.1	RDA-R1.1-02M	Metadata refers to a standard reuse licence	Important	1 – not being considered this yet	1	0
35		R1.1	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	Important	1 – not being considered this yet	1	0
36		R1.2	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	Important	1 – not being considered this yet	1	0
37		R1.2	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	Useful	1 – not being considered this yet	1	0
38		R1.3	RDA-R1.3-01M	Metadata complies with a community standard	Essential	1 – not being considered this yet	1	0
39		R1.3	RDA-R1.3-01D	Data complies with a community standard	Essential	3 – in implementation phase	3	0
40		R1.3	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	Essential	1 – not being considered this yet	1	0
41		R1.3	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	Important	1 – not being considered this yet	1	0

Enterprise FAIR standard: Defining AZ FAIR levels

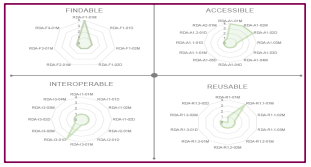
Evolution of System Centric Thinking → **Evolution of Data Centric Thinking to Knowledge Centric**

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

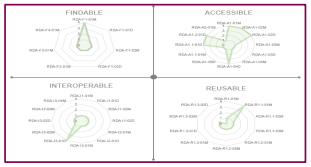
L2

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



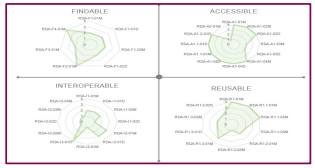
L3

- 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.



L4 – Domain level FAIR

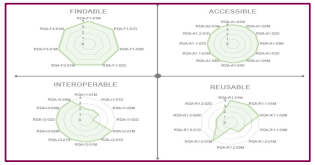
- Data conformed, integrated, processed and audited to support specific analytics patterns/enquiries
- **Data is conformed using a domain level data model**
- **Data embeds local master and reference data**
- Controls on access at the data level
- Creation of analytics ready 'Marts' enabling self service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



L5 – Enterprise Level FAIR

Extend L4 further by:

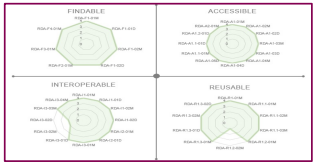
- **Data is described in a cross domain** data or industry model and is integrated in a Data Mesh
- **Data embeds enterprise master and reference data**
- Enterprise Metadata standard applied
- **URI's and PURLs are implemented**
- Cross domain Analytics enabled: Citizen Data Scientist using all relevant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



Adapted from Colin Wood

FAIR indicators, defined externally, can be used to benchmark and measure each level

How is the assessment carried out?

Focus on Level 4

Criteria for achieving Level 4 - Domain Level FAIR

Programmatic Data Provisioning

4. Is there programmatic provisioning, for example a data level API to access data? (ACCESSIBLE)

- Is there programmatic provisioning to access the data?
- Can data be accessed at a grain that allows consumption by grouping e.g. by study (or equivalent grouping) as well as bulk level basis?
- Are outputs provided in a commonly used format?
- Does the programmatic access mechanism provide metadata at entity and attribute level of the data it exposes?
- If API is available, is it registered in Anypoint Exchange so access control can be coordinated via Mulesoft DevOps?

Access Control

5. Are data product centric access controls in place to manage access to the data? (ACCESSIBLE)

- Are the access control determined by metadata that defines access roles?
- Is there a system or process in place that allows the data owner to revoke access if necessary?
- Is there a data access policy from Data Office in place for the data / data set in question?
- Is the access control orchestrated by a workflow tool?
- Are the access control at an appropriate access grain and tied into metadata about critical data as defined b

Note: Permission to access this data is independent of the system the data may have originated from.

Reference and Master Data (RMD) Identification

6. Has the data model (minimum metadata or conceptual/logical) undergone an IA and Data Office review (I and reference data sources? (INTEROPERABLE)

- Has the data model for the data product undergone an IA review through IARB?
- Has Master and Reference Data been identified??*
- Has Master and Reference Data sources been confirmed with Data Office (e.g. through IARB review)?

Note: Depending on the solution, conceptual or logical level might need to be reviewed. Level to be determine
*In future will need to have defined minimum/critical Master & Reference Data per domain[1]

RDM Service

7. Is the Platform consuming and/or publishing relevant reference data terms from a reference data solution

- Are all domain related entities included?
- Is the reference data in scope aligned to a standard vocabulary? /To what degree is the reference data in
- Is the vocabulary available in CENtree or will it be requested to be added to CENtree?

Note:
- Interim tactical solutions do not count, unless it is approved as a standards by the Data Standards Councils
- Accountability for Completeness is with the Data Office, as part of TRUST framework. The Data Office prio may specify expected measures for completeness for current and legacy data.

8

Criteria for achieving Level 4 - Domain Level FAIR

MDM Service

8. Is the Platform consuming and publishing identifiers from domain specific Master Data Solutions? (INTEROPERABLE)

- Is the platform/data product consuming identifiers from domain specific Master Data Solution?
- Or is the platform/data product sourcing the master data in scope from the appropriate system of record?

Note: Interim tactical solutions do not count, unless it is approved as a standards by the Data Standards Councils.

Meta Data

9a. Is there a description of the minimum metadata terms available within Collibra? ?* (REUSABLE)
9b. OR Have business terms, defining tables and entities, been published to Collibra using the logical data model as a source? (REUSABLE)
9c. OR Has an industry data model, with associated metadata, been loaded into Collibra (REUSABLE))

Datasets/Data Products must be mapped to the appropriate metadata loaded into Collibra.

Note to 9a: The question here is whether the metadata used to describe the dataset/data product, itself is described in Collibra. The requirement is:
- Dataset/data product is described through metadata in Collibra.
- The metadata used to describe the dataset/data product is also defined in Collibra.

Example:
- Scenario – our data catalogue specifies 5 attributes as mandatory – e.g. GRAD code, security classification, study identifier, indication, material entity id
- To address this scenario, metadata giving a human readable definition of what all 5 of those attributes mean should also appear in Collibra
- E.g. GRAD code may have a definition "A code defining the AZ retention policy for any records associated with this term"
- Best practice is to link this to a Collibra glossary, so that all terms are consistently defined.
- Level 5 (and 6) will increase the criteria to make these machine readable.

*In future will need to have Enterprise vocabulary of meta data (inclusive of domain specific vocabulary)[1]

Data Audit

10. Are audit logs for the platforms published e.g. by sending logs to a suitable logging tool such as Splunk? (REUSABLE)

This is not specifically aligned to FAIR practices, but is a requirement for GxP and should align to MHRA's GXP data integrity guide published - MHRA Inspectorate (blog.gov.uk)
The audit trail is to provide secure recording of life-cycle details - creation, additions, deletion or alteration - of information in a record.
The purpose of this is to embed TRUST in the data, but primarily to provide evidence of the journey of the data to allow it to be used with confidence. This should include:
- Where the data originated
- Loading information (Status and Time)
- Data consumption: who or what utilised/accessed the data
- Any alteration of the data

Data Lineage

11. Stretch: Has upstream and downstream Data Lineage metadata been published to Collibra, identifying sources and consumers of data? (REUSABLE)

- To what degree/coverage is data lineage available in Collibra for the dataset/data product in question?

9

ID		PRINCIPLE	INDICATOR ID	INDICATORS	PRIORITY	METRIC	VIZ	%
1	F	F1	RDA-F1-01M	Metadata is identified by a persistent identifier	Essential	1 – not being considered this yet	1	0
2		F1	RDA-F1-01D	Data is identified by a persistent identifier	Essential	2 – under consideration or in planning phase	2	0
3		F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	Essential	1 – not being considered this yet	1	0
4		F1	RDA-F1-02D	Data is identified by a globally unique identifier	Essential	2 – under consideration or in planning phase	2	0
5		F2	RDA-F2-01M	Rich metadata is provided to allow discovery	Essential	1 – not being considered this yet	1	0
6		F3	RDA-F3-01M	Metadata includes the identifier for the data	Essential	1 – not being considered this yet	1	0
7		F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	Essential	1 – not being considered this yet	1	0
8	A	A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	Important	1 – not being considered this yet	1	0
9		A1	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	Essential	3 – in implementation phase	3	0
10		A1	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	Essential	3 – in implementation phase	3	0
11		A1	RDA-A1-03M	Metadata identifier resolves to a metadata record	Essential	1 – not being considered this yet	1	0
12		A1	RDA-A1-03D	Data identifier resolves to a digital object	Essential	3 – in implementation phase	3	0
13		A1	RDA-A1-04M	Metadata is accessed through standardised protocol	Essential	1 – not being considered this yet	1	0
14		A1	RDA-A1-04D	Data is accessible through standardised protocol	Essential	3 – in implementation phase	3	0
15		A1	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	Important	3 – in implementation phase	3	0
16		A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	Essential	1 – not being considered this yet	1	0
17		A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	Important	3 – in implementation phase	3	0
18		A1.2	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	Useful	3 – in implementation phase	3	0
19		A2	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	Essential	1 – not being considered this yet	1	0
20	I	I1	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	Important	1 – not being considered this yet	1	0
21		I1	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	Important	1 – not being considered this yet	1	0
22		I1	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	Important	1 – not being considered this yet	1	0
23		I1	RDA-I1-02D	Data uses machine-understandable knowledge representation	Important	1 – not being considered this yet	1	0
24		I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	Important	1 – not being considered this yet	1	0
25		I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies	Useful	2 – under consideration or in planning phase	2	0
26		I3	RDA-I3-01M	Metadata includes references to other metadata	Important	1 – not being considered this yet	1	0
27		I3	RDA-I3-01D	Data includes references to other data	Useful	3 – in implementation phase	3	0
28		I3	RDA-I3-02M	Metadata includes references to other data	Useful	1 – not being considered this yet	1	0
29		I3	RDA-I3-02D	Data includes qualified references to other data	Useful	3 – in implementation phase	3	0
30		I3	RDA-I3-03M	Metadata includes qualified references to other metadata	Important	1 – not being considered this yet	1	0
31		I3	RDA-I3-04M	Metadata include qualified references to other data	Useful	1 – not being considered this yet	1	0
32	R	R1	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	Essential	3 – in implementation phase	3	0
33		R1.1	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	Essential	1 – not being considered this yet	1	0
34		R1.1	RDA-R1.1-02M	Metadata refers to a standard reuse licence	Important	1 – not being considered this yet	1	0
35		R1.1	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	Important	1 – not being considered this yet	1	0
36		R1.2	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	Important	1 – not being considered this yet	1	0
37		R1.2	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	Useful	1 – not being considered this yet	1	0
38		R1.3	RDA-R1.3-01M	Metadata complies with a community standard	Essential	1 – not being considered this yet	1	0
39		R1.3	RDA-R1.3-01D	Data complies with a community standard	Essential	3 – in implementation phase	3	0
40		R1.3	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	Essential	1 – not being considered this yet	1	0
41		R1.3	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	Important	1 – not being considered this yet	1	0

Solution	HBS Data Hub
Service Now Link	TBD
Blueprint Link	HBS Data Hub Solution Blueprint
ARB Review Date	28-Sep-22
FAIR Assessor	Colin Wood

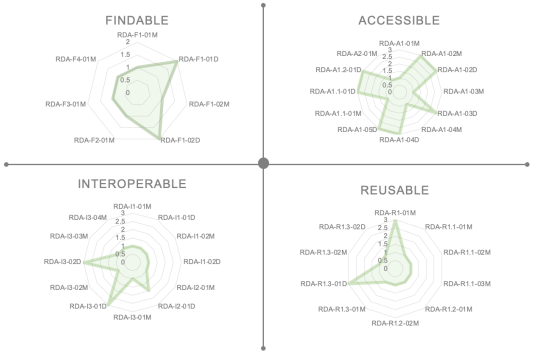
FAIR Criteria	Score	Comments	Actions required to make FAIR
1. Lake	1	Solution is fully integrated with Data Lake	
2. Catalogued	0	Solution does not use data catalogues or metadata repositories	Catalogue in Collibra.
3. Critical Data	0	Solution and metadata not catalogued, so it is not possible to identify this.	Annotate critical data in Collibra.
4. Confirmation	1	Conceptual and Logical Data Models reviewed and approved at IARB	
5. Programmatic Data Provisioning	1	API, delivered via Snowflake	
6. Access Control	1	Full access control in scope for the blueprint	We recommend a separate access control review later in the project lif
7. Reference and Master Data Identification	1	Full list of master and reference data in the blueprint and in this workbook	
8. RDM Service	0.38	No use of RDM solutions - using data directly from operational sources	Build in support for ontologies sourced from CENTREE
9. MDM Service	0.70	Using best available master data sources where available.	More detailed plan for master data integration required
10. Metadata Vocabulary	0	Solution does not use data catalogues or metadata repositories	Catalogue in Collibra and publish the logical data model.
11. Data Audit	1	Auditing is built into the solution	
12. Data Lineage	0	No use of metadata repositories to capture lineage	Publish lineage, specifying sources of data, into Collibra
LEVEL 4 FAIR Score	64.37%		

Note that target score is 100% - but 120% is possible by embedding PURL identifiers

FAIRNESS PROGRESS PER INDICATOR

Maturity level per indicator (per FAIR area)

- 0 – not applicable
- 1 – not being considered this yet
- 2 – under consideration or in planning phase
- 3 – in implementation phase
- 4 – fully implemented



FAIRe(nough) Framework

How FAIR is my data: the FAIRe(nough) benchmark at AstraZeneca

Ensure we target resources where the value is using Use case/business value as a driver

Prioritised
Use case example

As a ML/statistics scientist,
I want to be able to build
a **Drug Sensitivity Predictive Model**
for **Disease X** patient
treated with **drug Y**
using **gene expression profiles**

Use case example (Data Science)

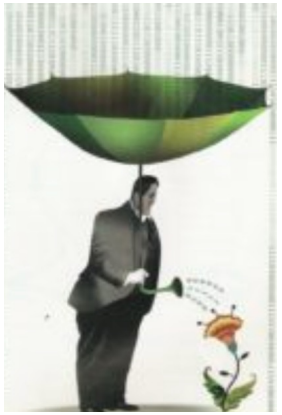
01 **Concepts and Data sources discovery**

Find Patient level data sources where **indication X**, **response** following **a drug treatment X** and **OMIC data** are available

02 **Source and key elements gathering**

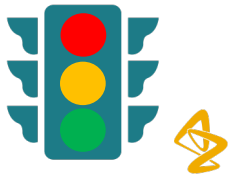
Data Sources - High level metrics

FAIRe Data – low level metrics



03 **FAIRe implementation**

- FAIRe Data assessment (High/low level)
- Compute result and translate to impact
- FAIRe driven action:
  - Send back
  - Curate/ingest/buy

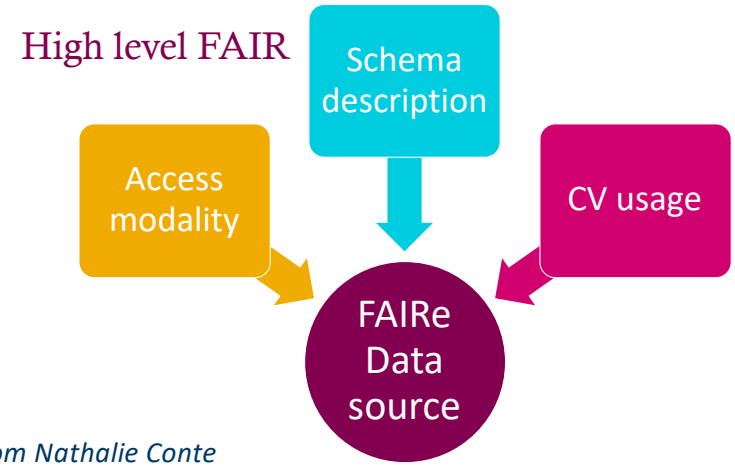


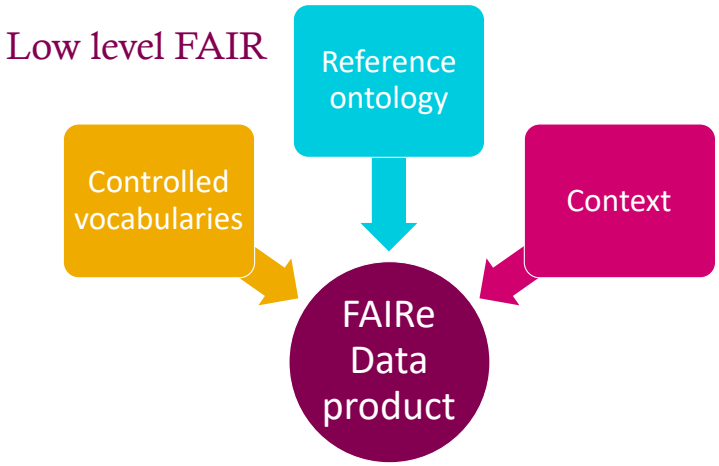
Adapted from Nathalie Conte

FAIRe: Producing FAIRe(nough) data products

Two dimensions of FAIRness …

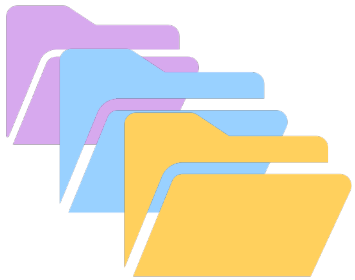
High-level FAIRness	Low-level FAIRness: FAIRe for a purpose
FAIRness of the data resource itself.	FAIRness of the data.
Does not require a data sample to calculate.	Requires a data sample to calculate.
Based on aspects of resource documentation, provisioning, access methods etc. (metadata attributes).	Based on metrics calculated from the data itself.
Fairly easy / cheap to define and calculate.	Less easy to define / calculate.
Calculated per resource.	Calculated via data assessment tools for a defined set of critical data elements in the context of a use case within/across data resources Completeness/adherence to Std/CV



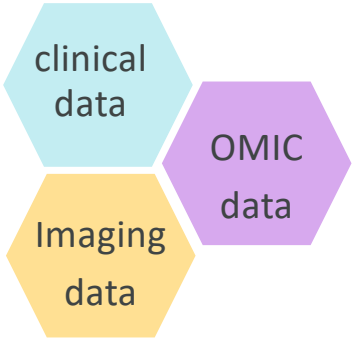


Adapted from Nathalie Conte

High-level FAIR assessment



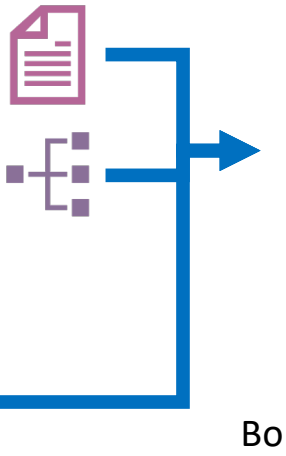
Sources



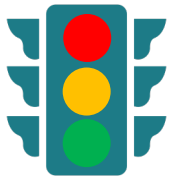
Data access docs

Schema docs

Compliance docs



FAIRe assessment

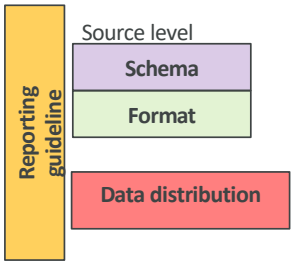


FAIRe output

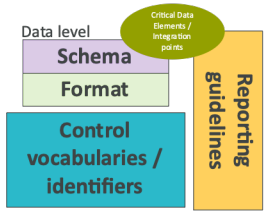
Valid modelling specification

Boiled down to 4 aspects:

- Data access documentation
- Data field descriptions
- Entity definitions & references
- CV usage



Low-level FAIR assessment –
Data assessment / curation loop

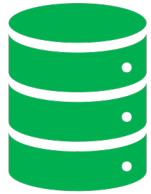


Raw data product

Assess

Review

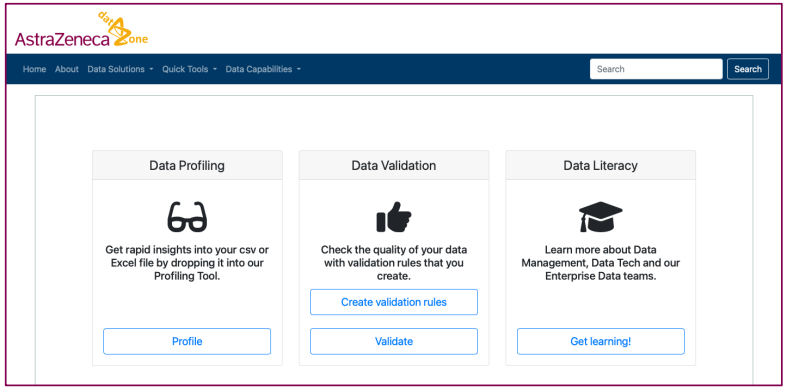
Curate



FAIR (enough) data product

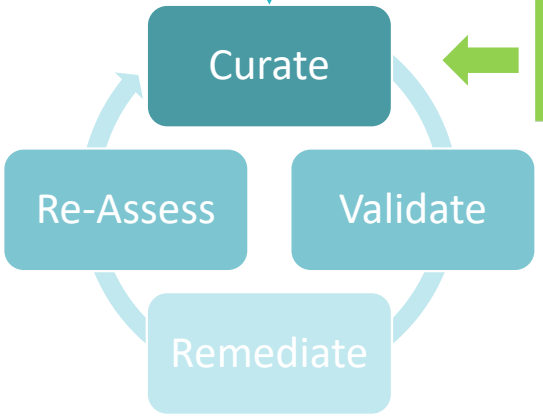
Modelling specification

AZ Data Zone tool

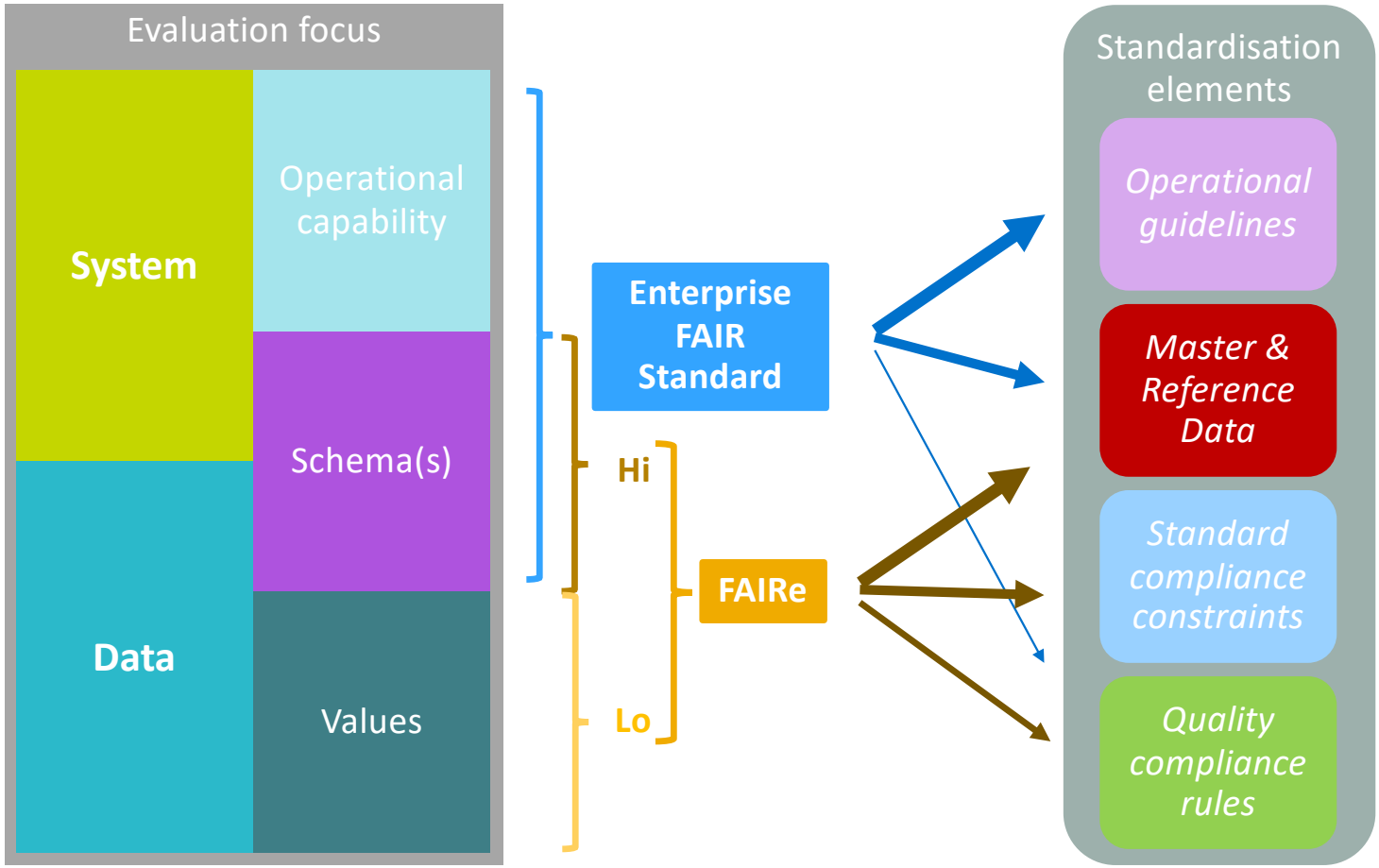


Targeted Curation

Standards



Benchmarks scope



Benchmarks assessment

Enterprise FAIR Standard

- **Test cases**: Two independent system-level evaluations
- **Lessons learned**:
  - Defined as a **high-level** assessment benchmark
  - **Easy to measure**
  - **Quantitative output** for qualitative yes-no questions
  - **In scope**: system operational capability, data schema
  - **Out of scope**: data value level, qualitative/validation dimension of the data

FAIRe(nough)

- **Test cases**: Data asset records in a single application
- **Lessons learned**
  - Current form is **integration specs guideline** rather than assessment benchmark
  - Implementation as a benchmark **challenging** at both high and low levels
  - **High level**:
    - Most elements can be met with Enterprise FAIR Standard
  - **Low level**:
    - Formalised most FAIRe questions as **validation rule**s in DataZone Validator
    - Turning guidelines into quantitative scores is a challenge

In practice

High-level FAIR

- Work in progress to get a single score per level
- **North star** for system-level FAIR

Low-level FAIR

- Generation of validation rule set
- Assessment of data against **validation rules**
- Governance of individual **data elements** is key

Next steps

Global Strategic partnership

- Mature development of **low-level assessment** into **defined set of rules**
  - Consolidation of **Core Data Elements** reference list
  - Representation and implementation of **business rules** around Core Data Elements into validation toolkit
  - Mature into **common validation toolkit**
- Mature **high-level assessment**
  - **Consolidation of alignment** between FAIRe high-level information standard and Enterprise FAIR levels
- FAIR **measures to be computable, cohesive and transparent** to all Data users
  - FAIR **toolkit** development
  - **Governance & policies** alignment
  - Focus on **sustainability**
- **Improve/publish the framework** through iteration & strategic alignment

- **Open for Collaboration** with relevant initiatives! Please contact us ☺

Thanks for listening!