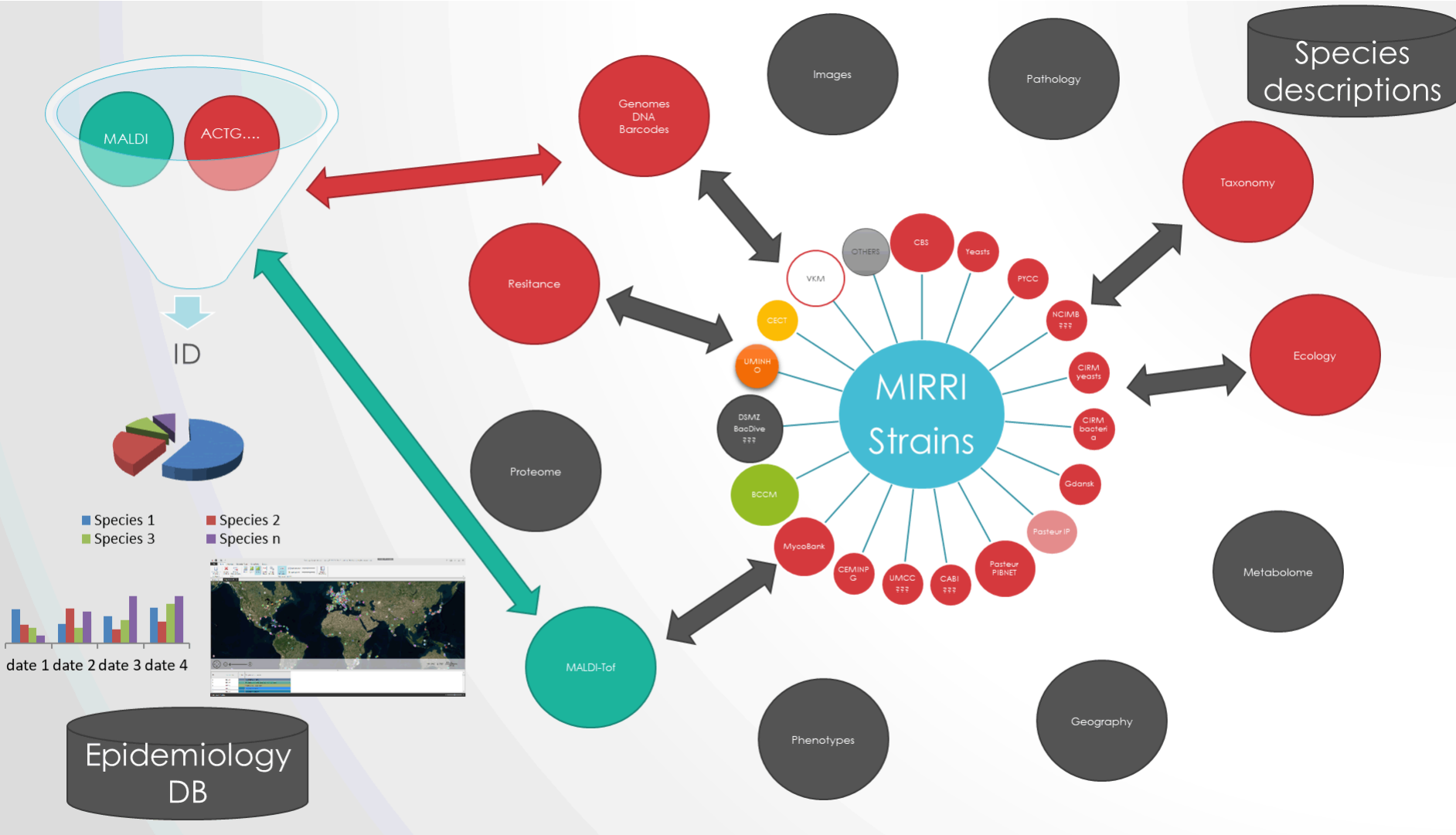# FAIR OPTIONS
IN MBRC SPECIFICS

Alexander Vasilenko, Vincent Robert, José Miguel López Coronado, Oleg Stupar, Galina Kochkina, Svetlana Ozerskaya, Serge Casaregola

MIRRI data processing

Key principle: FAIR (FORCE11)

Findable

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

A1. (Meta)data 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

Interoperable

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

FAIR recommendations hierarchy

1. **FORCE11***

2. **The EC EG Action Plan, RDA GEDE, GO FAIR,** BD2K/NIH, FAIR Data Commons (USA), RDA, etc.

3. **EOSC-Life (WP6, WP1, WP2, WP4)**
   FAIRplus(Elixir)

4. **MIRRI**

* https://www.force11.org/fairprinciples

MIRRI domain specific recommendations

Findable

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata: MTA, MAA, MDA, PIC, IRCC, taxonomy used, mCC, curator, storage method, supply method, price, provenance, licensing

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

A1. (Meta)data 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

Interoperable

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance: MTA, MAA, MDA, PIC, IRCC

R1.3. (Meta)data meet domain-relevant community standards: OECD MDS+RDS+FDS, WDCM MDS, WDCM+VKM RDS, AWI 21710

Implementation process participants

- **Research communities**: practitioners from all research fields, clustered around disciplinary interests, data types or cross-cutting grand challenges.
- **Data service providers**: domain repositories, research infrastructures (e.g. ESFRIs) and e-infrastructures, institutional, community and commercial tools and services.
- **Data stewards**: support staff from research communities and research libraries, and those managing data repositories.
- **Standards bodies**: formal organisations and consortia coordinating data standards and governing procedures relevant to FAIR, e.g. repository certification, curriculum accreditation (e.g. W3C, NIST).
- **Coordination fora**: global and national bodies such as the Research Data Alliance, CODATA, WDS Communities of Excellence, GO FAIR, German Data Forum (RatSWD), Dutch Coordination Point (LCRDM) and similar initiatives.
- **Policymakers**: governments, international entities like OECD, research funders, institutions, publishers and others defining data policy.
- **Research funders**: the European Commission, national research funders, charitable organisations and foundations, and other funders of research activity.
- **Institutions**: universities and research performing organisations.
- **Publishers**: not-for-profit and commercial, Open Access and paywall publishers of research papers and data.

Action Plan recommendation example

Rec. 20: Deposit in Trusted Digital Repositories

Research data should be made available by means of Trusted Digital Repositories, and where possible in those with a mission and expertise to support a specific discipline or interdisciplinary research community.

- **Action 20.1**: Policy should require data deposit in certified repositories and specify support mechanisms (e.g. incentives, structural funding and/or funding for deposit fees, and training) to enable compliance.

Stakeholders**:** Policymakers; Funders; Publishers.

- **Action 20.2**: Mechanisms need to be established to support research communities to determine the optimal data repositories and services for a given discipline or data type.

Stakeholders**:** Data service providers; Institutions; Data stewards; Coordination fora.

- **Action 20.3**: Concrete steps need to be taken to ensure the development of domain repositories and data services for interdisciplinary research communities so the needs of all researchers are covered.

Stakeholders**:** Data service providers; Funders; Institutions; Research communities.

- **Action 20.4**: Outreach is required via scholarly societies, scientific unions and domain conferences so researchers in each field are aware of the relevant disciplinary repositories.

Stakeholders**:** Data service providers; Research communities.

Related recommendations: Rec. 13: Develop metrics to certify FAIR services; Rec. 17: Align and harmonise FAIR and Open data policy.

FAIR Recs presented/not presented in MIRRI

KEY SYSTEMS TO MAKE

Cloud of registries
- Policies
- DMPs
- Identifiers
- Standards
- Repositories
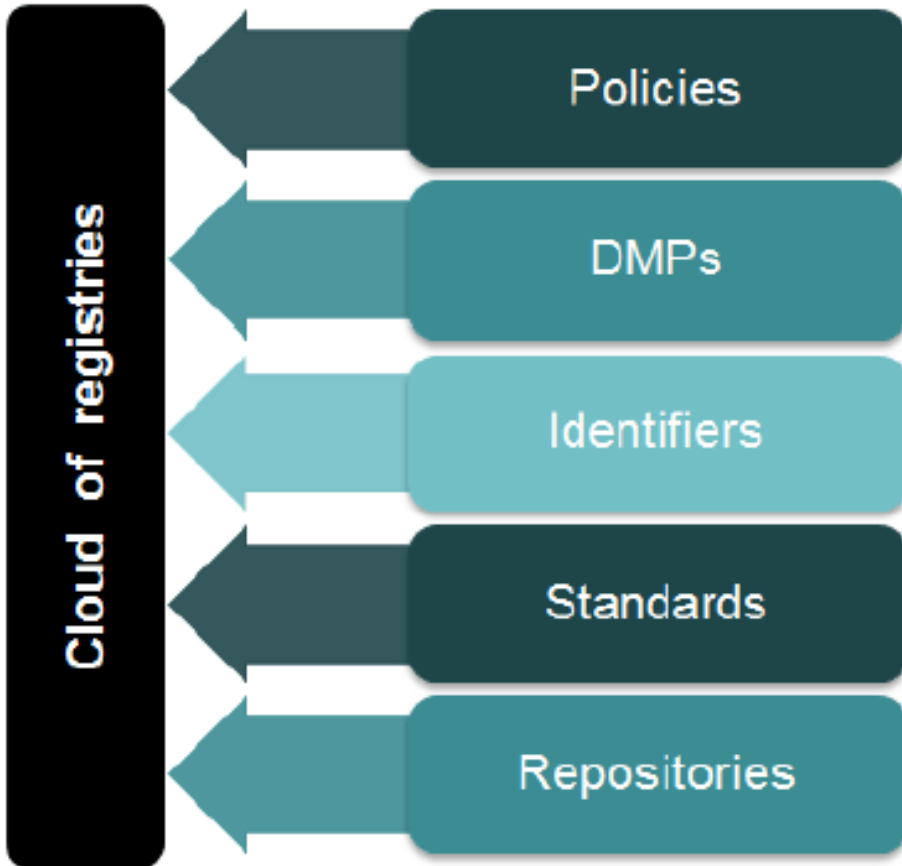
FAIR ACTION PLAN RECS PARTIALLY PRESENTED IN MIRRI PLANS

1.2, 3.2, 4.5, 5.1, 7.1, 8.1, 22.3
(7 total)

FAIR ACTION PLAN RECS NOT PRESENTED IN MIRRI AT ALL

1.1, 1.3, 2, 3.1, 3.3, 3.4, 4.1, 4.2, 4.3, 4.4, 5, 6, 7.2, 7.3, 8.2, 8.3, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22.1, 22.2, 22.4, 22.5, 23, 24, 25, 26, 27
(124 total)

KEY SYSTEMS OPTIONS

- ## Policies:
  rules for exceptional conditions in data processing: in CABRI (http://www.cabri.org/guidelines.html), Guidelines for Collection Quality Management Standards http://www.cabri.org/guidelines/gl-framed.html, OECD BPG - http://www.oecd.org/sti/emerging-tech/38777417.pdf (and must be available for computer programs), WFCC Guidelines http://www.wfcc.info/guidelines/, Guide to the Deposit of Microorganisms under the Budapest Treaty.

- ## Data Management Plans (DMPs):
  DMPs are SOPs in information processing, many culture collections have inner SOPs, MIRRI did not collect them, empty position yet.

- ## Identifiers: Digital Object Identifier (DOI) format looks popular as a persistent identifier,
  possibly for each strain, and possibly for each mCC, we discovered the price 1$ for each DOI. Example: doi:10.1000/182 refer to URL https://doi.org/10.1000/182

- ## Standards: OECD MDS+RDS+FDS, WDCM MDS, WDCM+VKM RDS, AWI 21710

- ## Registries: Popular global registry: https://www.re3data.org/, local in EOSC-Life
  community: https://fairsharing.org/

- ## Repositories: MIRRI-IS

Repository: Core Trust Seal Rules (9)

R0. Provide context for your repository (repository type, designated community, level of curation, outsource partners)

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).

R7. The repository guarantees the integrity and authenticity of the data.

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

R9. The repository applies documented processes and procedures in managing archival storage of the data.

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

R12. Archiving takes place according to defined workflows from ingest to dissemination.

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

R7. ... integrity and authenticity of the data (*)

Integrity ensures that changes to data and metadata are documented and can be traced to the rationale and originator of the change.

Authenticity covers the degree of reliability of the original deposited data and its provenance, including the relationship between the original data and that disseminated, and whether or not existing relationships between datasets and/or metadata are maintained.

For this Requirement, responses on data integrity should include evidence related to the following:

● Description of checks to verify that a digital object has not been altered or corrupted (i.e., fixity checks).
● Documentation of the completeness of the data and metadata.
● Details of how all changes to the data and metadata are logged.
● Description of version control strategy.
● Usage of appropriate international standards and conventions (which should be specified).

Evidence of authenticity management should relate to the follow questions:

● Does the repository have a strategy for data changes? Are data producers made aware of this strategy?

● Does the repository maintain provenance data and related audit trails?

● Does the repository maintain links to metadata and to other datasets? If so, how?

● Does the repository compare the essential properties of different versions of the same file? How?

● Does the repository check the identities of depositors?

This Requirement covers the entire data lifecycle within the repository, and thus has relationships with workflow steps included in other requirements - for example, R8 (Appraisal) for ingest, R9 (Documented storage procedures) and R10 (Preservation plan) for archival storage, and R12–R14 (Workflows, Data discovery and identification, and Data reuse) for dissemination. However, maintaining data integrity and authenticity can also be considered a mindset, and the responsibility of everyone within the repository.

Repository: Additional Requirements

FAIR supporting requirements for repositories. Some can be directly derived from the FAIR principles:

F: Repositories need to ensure that its digital objects are assigned a PID and are described by "rich" metadata which also include the PID, and that metadata can be harvested.

A: Repositories need to ensure that the PID can be used to retrieve the DOs bit sequence using standard protocols which are open, free and universal, that authentication and authorisation is being checked and that metadata exists even if the bit sequence is not accessible anymore.

I: Repositories need to ensure that well-known languages are used to represent (structure and) semantics, that the vocabularies used in the DOs are FAIR and that relevant relationships are included in an explicit way.

R: Repositories need to ensure that DOs are being described by accurate attributes that include clear usage licenses and provenance descriptions, and that domain-relevant community standards are being used.

Digital Object (DO) definitions (1)

D1: The discussions about Digital Objects since the start of RDA in 2013 resulted basically in two major definitions from RDA DFT and DOIP and several others that are based on different discussion roots such as from ITU, SAA or the one suggested by Hermon Sorin.

RDA DFT 2 : A digital object (DO) is represented by a bitstream, is referenced and identified by a persistent identifier and has properties that are described by metadata.
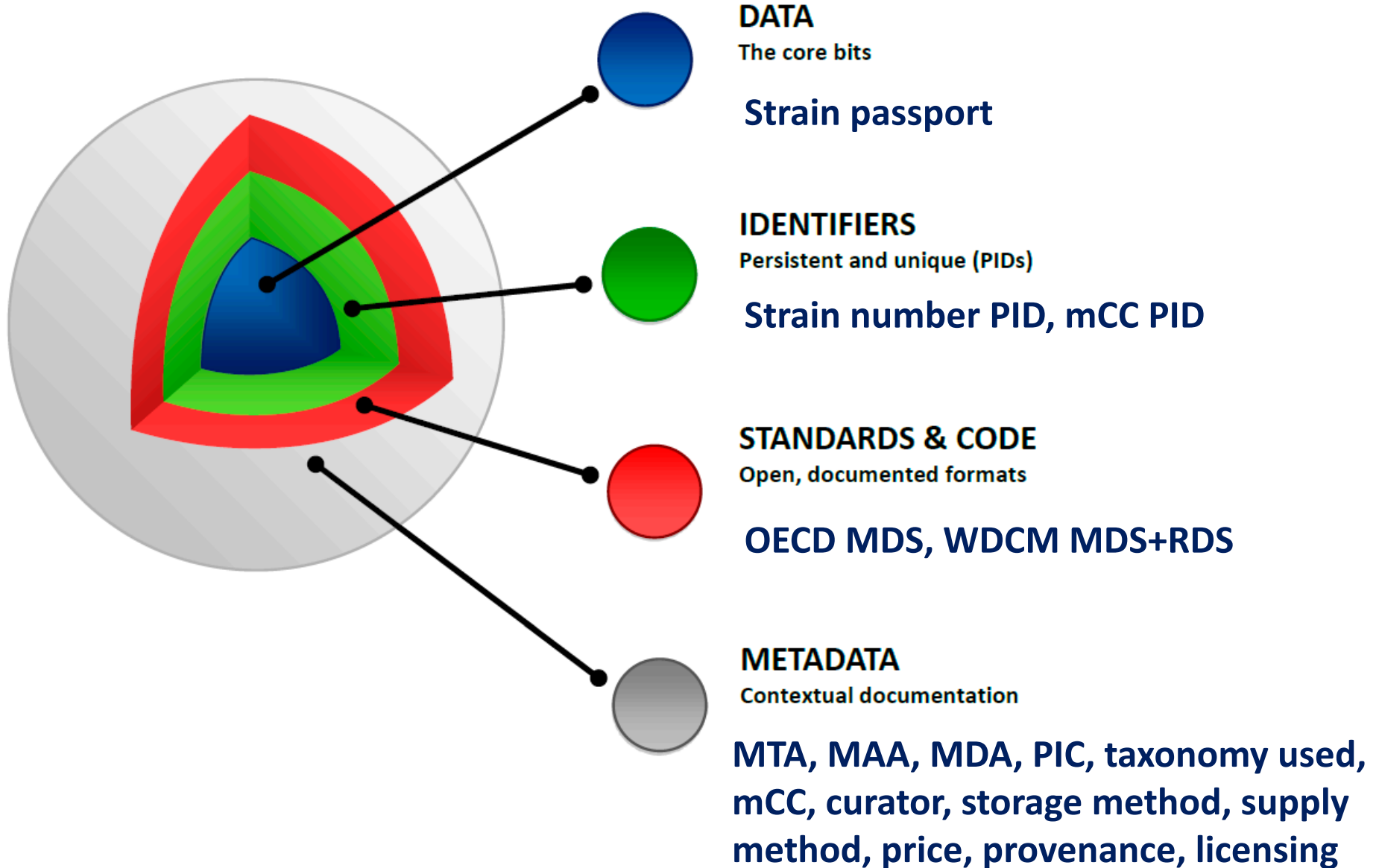
DOIP 3 : A digital object (DO) is a sequence of bits, or a set of sequences of bits, incorporating a work or portion of a work or other information in which a party has rights or interests, or in which there is value, each of the sequences being structured in a way that is interpretable by one or more of the computational facilities, and having as an essential element an associated unique persistent identifier.

ITU 4 : A digital entity is an entity represented as, or converted to, a machine-independent data structure consisting of one or more elements in digital form that can be parsed by different information systems; the structure facilitates interoperability among diverse information systems in the Internet.

Society of American Archivists (SAA) 5 : A digital object (DO) is a unit of information that includes properties (attributes or characteristics of the object) and may also include methods (means of performing operations on the object).

Hermon Sorin: DO is an abstract element that gathers information about a physical or a virtual entity, sufficient for the discovery and inquiry on this entity.

mCC strains digital object model

DATA
The core bits

Strain passport

IDENTIFIERS
Persistent and unique (PIDs)

Strain number PID, mCC PID

STANDARDS & CODE
Open, documented formats

OECD MDS, WDCM MDS+RDS

METADATA
Contextual documentation

MTA, MAA, MDA, PIC, taxonomy used, mCC, curator, storage method, supply method, price, provenance, licensing

A good FAIR metric should be:*

- Clear: anyone can understand the purpose of the metric
- Realistic: it should not be unduly complicated for a resource to comply with the metric
- Discriminating: the metric should measure something important for FAIRness; distinguish the degree to which that resource meets that objective; and be able to provide instruction as to what would maximize that value
- Measurable: the assessment can be made in an objective, quantitative, machine-interpretable, scalable and reproducible manner, ensuring transparency of what is being measured, and how.
- Universal: The metric should be applicable to all digital resources.

(4), http://fairmetrics.org

FAIR Metrics tools in FAIRshering.org

14 Core FAIR Metrics

Findable:

FM-F1A **FM-F1B**
F1 (meta)data are assigned a globally unique and persistent identifier;

FM-F2
F2 data are described with rich metadata;

FM-F3
F3 metadata clearly and explicitly include the identifier of the data it describes;

FM-F4
F4 (meta)data are registered or indexed in a searchable resource;

Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally **FM-A1.1** implementable;

FM-A1.2
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; **FM-A2**

Interoperable:

I1 (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge **FM-I1** representation.

FM-I2
I2 (meta)data use vocabularies that follow FAIR principles;

I3 (meta)data include qualified references to other (meta)data; **FM-I3**

Reusable:

R1 meta(data) are richly described with a plurality of accurate and relevant attributes;

R1.1 (meta)data are released with a clear and accessible data usage license; **FM-R1.1**

R1.2 (meta)data are associated with detailed **FM-R1.2** provenance;

R1.3 (meta)data meet domain-relevant community standards; **FM-R1.3**

The Evaluator Workflow
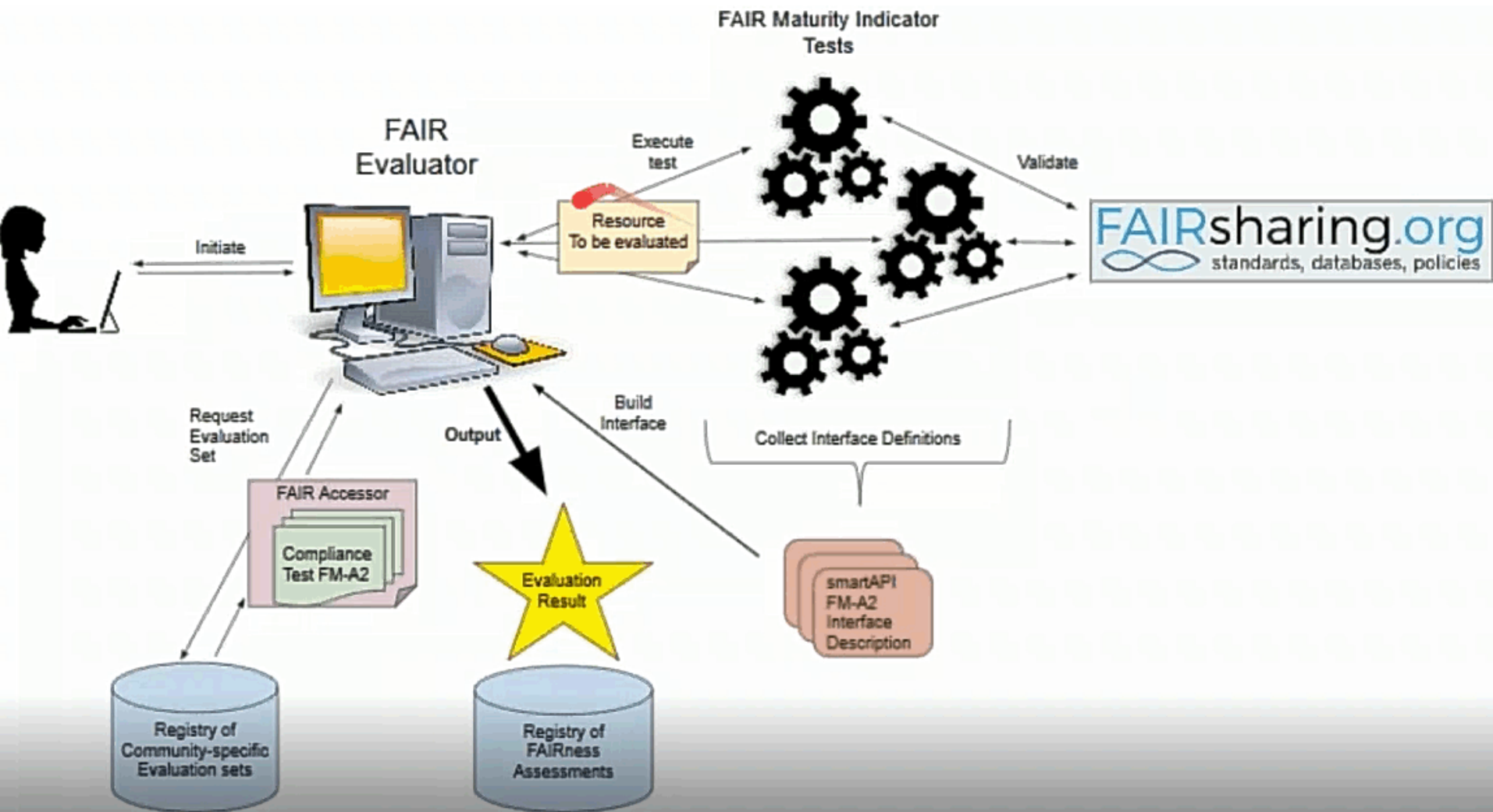
How to join GO FAIR

- Rules of Engagement in Implementation Network
- Envisioned Internet of FAIR Data & Services (IFDS)
- Manifesto preparation
- Communication with GO FAIR International Support & Coordination Office (GFISCO) for approval
- Active Implementation Network period on approval

Publications

1. GEDE DO Topic Group: Digital Object Assertions. Version 2.0 - February 2019 (in DO share) https://drive.google.com/drive/folders/1E9OB3xUhdKfyAFNHEeZvb4v1kte7CjWN

2. GEDE DO Topic Group: Digital Object Roadmap Document. Version 2.0 - February 2019

3. GEDE Webinar on Maturity Indicators for FAIRness and Certification of Repositories (done by GEDE RDA), 22 March 2019. https://rd-alliance.org/sites/default/files/GEDE%20fair-cts-webinar-1.pdf

4. Wilkinson M.D., Sansone S-A., Schultes E., Doorn P., Bonino L.O., Dumontier M. A design framework and exemplar metrics for FAIRness. Scientific Data volume 5, Article number: 180118 (2018). https://doi.org/10.1038/sdata.2018.118

5. Boulanger D., Haggstrom I. (EISCAT), Hellstrom M. (ICOS), Satagopam V.(ELIXIR), Christophe, Wittenburg P. (RDA GEDE). Role of repositories in research infrastructure building. Result of the RDA GEDE Repository Topic Group (in share, 2019)

6. Wilkinson M., Schultes E., Bonino L.O., Sansone S-A., Doorn P., Dumontier M. FAIRMetrics/Metrics: FAIR metrics, evaluation results, and initial release of automated evaluator code [Internet]. 2018. doi:10.5281/zenodo.1305060

7. Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data volume 3, Article number: 160018 (2016). https://doi.org/10.1038/sdata.2016.18

8. https://www.force11.org/fairprinciples

9. https://www.coretrustseal.org/

Thank you