FAIR Data Maturity Model

Workshop #6

4th December 2019
# Agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'</td>
<td>Welcome, objectives of the meeting</td>
</tr>
<tr>
<td>5'</td>
<td>Roundtable</td>
</tr>
<tr>
<td>5'</td>
<td>State of play</td>
</tr>
<tr>
<td>30'</td>
<td>Pilot testing</td>
</tr>
<tr>
<td>30'</td>
<td>Discussions about testing results</td>
</tr>
<tr>
<td>10'</td>
<td>Draft guidelines</td>
</tr>
<tr>
<td>5'</td>
<td>Action items &amp; next steps</td>
</tr>
</tbody>
</table>
Context

The principles are NOT strict

- Ambiguity
- Wide range of interpretations of FAIRness

Different FAIR Assessment Frameworks

- Different metrics
- No comparison of results
- No benchmark

SOLUTION is to bring together stakeholders to build on existing approaches and expertise

- Set of core assessment criteria for FAIRness
- FAIR data maturity model & toolset
- FAIR data checklist
- RDA recommendation

Join the RDA Working Group: RDA WG web page | GitHub
Objectives

What are to be evaluated to determine FAIRness?

- Identify the indicators that can serve as core criteria
- Propose guidelines and a checklist
- Test the core criteria
- Enable the development of automated tools for evaluation
- Update the core criteria based on feedback

FAIR data maturity model
Scope

**BUT** the Working Group does **NOT** have the purpose to ...

- **develop yet-another-evaluation-method:** the core criteria are intended to provide a common ‘language’ across evaluation approaches, not to be applied directly to datasets.

- **define how the core criteria need to be evaluated:** The exact way to evaluate data based on the core criteria is up to the owners of the evaluation approaches, taking into account the requirements of their community.

- **revise and re-design the FAIR principles**
Roundtable

In the chat window, please type...

▶ Your name
▶ Your affiliation
▶ Your role
  ▶ Researcher
  ▶ Librarian
  ▶ Service provider
  ▶ Policy maker
  ▶ Funder

▶ Introducing the editorial team
State of play
State of play

1. Definition  DONE
2. Development  CLOSING
   i) First phase  DONE
   ii) Second phase  CLOSING
3. Testing  ONGOING
4. Delivery  ON HOLD

* Any comments are still welcomed with regards to the output produced during the first phase | [GitHub](https://github.com)
State of play

Proposition
- Indicators
- Maturity levels

Discussion | Indicators
- Validation (YES/NO)
- Missing indicators

Discussion | Prioritisation
- Approach to prioritisation
- Priority levels
- Survey

Discussion | Scoring
- Approach to scoring

Consolidation
- Indicators
- Maturity levels

Scoping
- Approach
- Methodology
- Landscaping exercise

Testing

Editorial team
Working group
State of play

- Goal is to **finalise** indicators and priorities
- Indicators and priorities will be further **used in their current state**
- Indicators and priorities **will be re-evaluated** after the testing phase
Development
Overview | Indicators & levels

- F1 (Meta)data are assigned globally unique and persistent identifiers
- 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

- A1 (Meta)data are retrievable by their identifier using a standardised communication protocol
  - A1.1 The protocol is open, free and universally implementable
  - A1.2 The protocol allows for an authentication and authorisation where necessary
- A2 Metadata are accessible, even when the data are no longer available

- I1 (Meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation
  - I2 (Meta)data use vocabularies that follow the FAIR principles
  - I3 (Meta)data include qualified references to other (meta)data

- 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
Overview | Indicators & levels

Indicators for Findability

- [F1-01M] Metadata is identified by a persistent identifier
- [F1-01D] Data is identified by a persistent identifier
- [F1-02M] Metadata is identified by a universally unique identifier
- [F1-02D] Data is identified by a universally unique identifier
- [F2-01M] Sufficient metadata is provided to allow discovery, following domain/discipline-specific metadata standard
- [F2-02M] Metadata is provided for the discovery-related elements defined by the RDA Metadata IG, as much as possible and relevant, if no domain/discipline-specific metadata standard is available
- [F3-01M] Metadata includes the identifier for the data
- [F4-01M] Metadata is offered/published/exposed in such a way that it can be harvested and indexed

* The full list of indicators can be found on the following GSheet

www.rd-alliance.org - @resdatall
Weighting the indicators, developed as part of the WG, following the key words for use in RFC2119

- **Mandatory**: indicator **MUST** be satisfied for FAIRness (Essential)
- **Recommended**: indicator **SHOULD** be satisfied, if at all possible (Important)
- **Optional**: indicator **MAY** be satisfied, but not necessarily so (Useful)
Development | Weighting Stats

Distribution of the weight of the indicators

FAIR PRINCIPLES

- Mandatory
- Recommended
- Optional

FINDABLE
- 5
- 3
- 7

ACCESSIBLE
- 2
- 2
- 7

INTEROPERABLE
- 5
- 2
- 9

REUSABLE
- 1
- 3
- 9
Pilot testing Presentation
Evaluation of RDA FAIR indicators
Françoise Genova - Astronomic
Context

WHY
Finding, accessing, interoperating and reusing data is at the core of astronomical research. The community has long been working to develop its international, open data sharing framework fitting its requirements. The framework is implemented by data providers, including the ground and space-based telescope archives and the widely used added-value data services, and used by the community in its daily research work (it is often invisible from users, so people may not be aware that they use it when accessing data and using tools). It is essential to test how the disciplinary practices to find, access, interoperate and reuse data are fitting with the proposed FAIR Data Maturity criteria. Astronomy provides a good real-life operational example for that.

WHAT
The astronomical data sharing framework includes a common format, FITS, which integrates data and metadata, and the standards and tools of the astronomical Virtual Observatory (VO), which provide findability, discoverability and interoperability. The VO standards are developed and maintained by the International Virtual Observatory Alliance (IVOA, http://ivoa.net). The VO framework is used/customized by other disciplines (astroparticle physics, planetary sciences, solar physics, the Virtual Atomic and Molecular Data Centre), so the test of usefulness goes beyond astronomy.

HOW
We checked each proposed criterion wrt. the community requirements on finding, accessing, interoperating and reusing data and the astronomical data sharing framework. We identified which criteria are implemented ‘automatically’ when data is provided through the Virtual Observatory, which ones rely on the data provider for implementation, and which ones require a combination of the two. We paid particular attention to the criteria currently tagged as ‘Mandatory’, since they will play a key role in the acceptance and rejection of data as FAIR if evaluation is performed using yes/no procedures (which may well happen based on the WG recommendations even if it is not the WG aim).
VO is FAIR ... *wrt. our needs*
One view of the VO from an application:

- Built from VO Registry
- Data from many observatories and missions
- 1000s All-Sky data sets
- Largest catalogues: PanSTARRS, Gaia etc.
- Complex ADQL queries
- Multi-resolution techniques for Big Data
- Interoperability of data
- Interoperability between applications
Overall results (1)

What worked well:

- The Virtual Observatory enables a significant fraction of the criteria to be satisfied, which is a plus for data providers.
- Common data providers’ practices also align with some of the criteria.
- No proposal to change the criteria, except eventually to improve the understandability and usability of a few of them.

What didn’t work well:

- Overall comparison with community requirement
  - Data Reuse and Interoperability are the key user requirements in astronomy. Find and Access are developed to enable R and I, and not as an objective per se. This starting point is very different from the one of the disciplines for which reproducibility is the key requirement (e.g., biology), for which being able to Find and Access specific data sets is the key criteria.
  - Findability and accessibility are not defined the same way with the two different starting points.
  - These differences have critical consequences on which criteria should be defined as mandatory wrt. community practices.
Overall results (2)

What didn’t work well (cont’d):

- Comments on the individual principles
  - **Find:** Finding data is a dynamic process for astronomers, who want to find data of interest for their research. The VO allows users to make simple or complex queries on all the data services declared in the VO, using a wealth of metadata. The PID is an element of findability among many others and is not the be-all and end-all of FAIRness as they appear in the FAIR principles and criteria for F and A (F1-01D, F1-02D, F3.01M, A1-03D).

  - **Access:** Access to data manually or by a machine is at the core of our system. Astronomical data is mostly open by default, which means that A1.2-10M is not mandatory, but our metadata can include information relevant to access control when relevant.

  - **Interoperability:** Interoperability is the core objective of the VO. We note that this is the only principle which has no Mandatory criterion, which means that data can be evaluated as FAIR in the proposed system without fulfilling any of the I criteria.

  - **Reusability:** Reusability is a core requirement for astronomical data. Data is open by default and is massively reused. Usage rights rely mostly on disciplinary ethics: cite the origin of data when data is cited. In some cases, an explicit license for usage is provided, but not always. This does not impair the widespread acceptance of data sharing and reuse. Four R criteria, two of them Mandatory, deal with license information.
Overall results (3)

What didn’t work well (cont’d):

- Additional comments
  - In our case some metadata are attached to the data collection, others to the data item.
  - Cascading criteria: Twice two criteria exclusive from each other (F2-01M/F2-02M; R1-01M/R1-02M). This means that one of the two criteria is irrelevant if the other one is fulfilled.
  - Consent for reuse (R1.1-05M) is irrelevant in our case.
  - What is a ‘FAIR compliant vocabulary’ (I1-02M, I2-02D)? Is a vocabulary standard which has a DOI and is freely available and reusable a FAIR compliant vocabulary?
  - What is a ‘sufficiently qualified reference’ to something (I3-02D, I3-O3D, I3-04D)?
Discussion points

- We hear that disciplinary practices have to be taken into account when defining FAIRness, and also that to become FAIR is a process towards disciplinary and cross-disciplinary FAIRness. As shown by our analysis the weight of the different criteria is different with different disciplinary points of view. We do not want to change the criteria, but we strongly suggest to use compliance scales instead of yes/no compliance evaluation. This will provide an inclusive system and a way to set up goals and measure progress.

- If compliance scales are established, they should include a ‘non applicable’ level.

- Open by default should be considered as acceptable, in spite of the possible legal hurdles.

- We note that significant costs are induced when one has to modify well established characteristics of a legacy discipline-wide, world wide data system. The large research infrastructures are supported to serve their communities, which can be an issue for engaging resources to fulfil criteria not relevant to disciplinary requirements. However they can aim at making progress gradually with the help of the compliance scales described above.

- To check the set of criteria with different, diverse communities is critical to ensure usability, wide acceptance and take-up.

- We plan to write an IVOA Note describing the assessment reported here.
IMI FAIRplus Project: Evaluation of Biomedical Datasets by RDA FAIR Indicators

Comparing the Outcomes of Multiple Independent Evaluators for FAIRness Assessment
22 participants
12 academic, 7 EFPIA, 3 SME
**ELIXIR** - Project Coordinator
**Janssen** - Project Leader

€8.23M budget
€4M H2020 EC funding
€4.23M EFPIA in-kind

42 months
IMI Project Portfolio
> 100 translational research projects in Public-private partnerships

**Our challenge:**
How do we design, test and refine our tools and processes in a way that scales to our data volumes?

- FAIR maturity assessment
  - Standards, metrics
  - Capacity building
  - Support

- FAIR Cookbook
  [https://fairplus.github.io/the-fair-cookbook/intro](https://fairplus.github.io/the-fair-cookbook/intro)
  - Publish FAIR datasets from >20 projects for access and reuse
  - FAIRified internal EFPIA datasets
  - SME engagements, hackathons and fellowships
FAIRification of First 4 pilot Datasets

Increased levels of FAIR

1. 2. 3. 4. 5.
Motivation: Identify Ambiguous Indicators

Measure FAIRness before the FAIRification (RESOLUTE)

- We applied RDA FAIR Maturity Metrics to measure the initial FAIRness level of the RESOLUTE data set
- Two experts collectively discussed each metric and decided on a score
- They reported that they found some of the metrics difficult to assess, since it might depend on interpretations

Observe which metrics could potentially depend on evaluators interpretation

- Manual Assessment: a systematic assessment for the ETOX and ND4BB data sets
- Automated Assessment: with the FAIRevaluator tool
## Method

<table>
<thead>
<tr>
<th>ETOX Dataset</th>
<th>ND4BB Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>› Three independent evaluators</td>
<td>› Two independent evaluators</td>
</tr>
<tr>
<td>› Applied RDA FAIR Metrics v0.2</td>
<td>› Applied RDA FAIR Metrics v0.3</td>
</tr>
<tr>
<td>› Recorded outcomes separately (average time 90 min)</td>
<td>› Recorded outcomes separately (average time 45 min)</td>
</tr>
</tbody>
</table>

Dedicated squad sessions are conducted for each dataset to compare scores of the independent evaluators and record feedback regarding the encountered challenges during assessment.

- Out of 54 metrics 9 of them are evaluated differently
- Out of 50 metrics 6 of them are evaluated differently

(*) FAIRevaluator returned a fail report, since the assessment applied before any FAIRification and data sets did not provide a machine access and authentication
Data Set 1: ETOX

eTOX*Sys Sampler - https://etoxsys.eu/etoxsys.v3-demo-bk/dashboard/

Ask for login
Tick all filters
Select XLS format
1 file per compound
Data Set 2: ND4BB TRANSLOCATION

https://www.dsf.unica.it/translocation/abdb/

ANTIMICROBIAL COMPOUNDS DATABASE

Aminoglycosides

- Amikacin

Beta-lactamase inhibitors

- Clavulanic Acid
- Sulbactam
- Avibactam
- Tazobactam
- BAL29880

Efflux pumps inhibitors

- Amiritixipine (Chloramphenicol)
- NMP (1-(Naphthalen-1-ylmethyl)piperazine)

Oxazolidinones

- Linezolid
- Sutezolid

Tetracyclines

- Minocycline
- Tigecycline

Maintained by G. Malocci
Findability Indicators

- **F1-01M** Metadata is identified by a persistent identifier

- **F1-01D** Data is identified by a persistent identifier

- **F2-01M** Sufficient metadata is provided to allow discovery, following domain/discipline-specific metadata standard

Different levels (granularities) of metadata: data provenance, data protocol, properties of dataset..

Identifier persistently points to the data set vs the ability of identifier to persistently identify the same data over time

Sufficient for which purpose? to find.. to reuse ... for humans ... for machines..

How to refer to a domain-specific standard? could it be multiple domains?
Accessibility Indicators

- A1-02M Metadata identifier resolves to a metadata record
- A1-03M Metadata is accessed through standardised protocol
- A1.1-01M Metadata is accessible through a free access protocol
- A1.1-01D Data is accessible through a free access protocol
- A1.1-02M Metadata is accessible through an open-source access protocol

Protocols (what is the definition):
- standardized: http, ftp, csv, html?
- free access: no key?
- Open - source?
Accessibility Indicators

- A1-01M Metadata includes **information** about access conditions
- A1.1-03D Actions to be taken by a reuser to get access to the data are **well documented**
- A1-02D Data is available for **automatic download**

What type of information: who to contact? Licence? What if data is openly accessible, should it state explicitly that everyone can use?

What does automatic mean: cron job? API? What if it requires parsing or extraction?
Interoperability Indicators

- I1-01M Metadata uses **knowledge representation** expressed in **standardised format**
- I1-01D Data uses **knowledge representation** expressed in **standardised format**
- I2-01D Data uses **standard vocabularies**
- I1-03M Metadata uses **self-describing knowledge representation**
- I2-02M Metadata uses **FAIR-compliant vocabularies**

Required to assess vocabularies, can not be accomplished during evaluation of data sets

Exclusively ontologies?

Does controlled vocabularies included (defined within the project)?

Only RDF?
Reusability Indicators

- **R1.3-01D** Data complies with a community standard

- **R1.1-01M** Metadata includes information about the licence under which the data can be reused

- **R1.2-02M** Metadata includes provenance information according to a cross-domain language

- **R1.3-02D** Data is expressed in compliance with a machine-understandable community standard

---

**how to recognize a community standard?**

**what if data is publically available?** is that license on the original data, or on data from the hosting repository?

**Needs reference?**

**xml, json, rdf, html ... ?** is it the same as machine readable
Common confounding ‘concepts’

➢ ‘Sufficient’: (5) required thought but did not result in scoring discrepancy

➢ ‘Protocol’: (11) discussions on most; score discrepancy ~ 50%
  ➢ - discussion on free (cost?), level of detail of what ‘protocol’ means (technical term or access methodology)

➢ ‘Persistence’: (2) discussion on meaning of persistence (identifier, resolver, the link between both, resource host policy); score discrepancy ~ 50%

➢ ‘Metadata’: (34) discussion on where to separate data/metadata, and what ‘level’ of metadata (dataset vs records or headings); score discrepancy ~ 5%
Common confounding ‘concepts’

➢ ‘Self-describing’: (2) does this dictate RDF, machine processable vs human readable; no score discrepancy

➢ Data is available for ‘automatic download’ (1)

➢ ‘Standard vocabularies’ (4) are controlled vocabularies included? 25% discrepancy

➢ ‘FAIR compliant’ (2) creates a circular discussion/assessment burden

➢ ‘Metadata to allow reuse’: (1) hard to ascertain as the purpose (metadata) not known in advance

➢ Appropriate ‘community/domain standards’: (5) how to choose where multiple may be applicable
Suggestions

› ‘Definitions of concepts such ‘metadata’, ‘automated’, ‘standardized, free and open source protocol’, ‘persistency’ should be provided

› Community data and metadata standards should be referenceable via a community resource as FAIRsharing.org, which covers all disciplines

› Evaluators needs a guideline with some examples

› Different ways of publishing data (controlled access, openly available... / separate metadata, metadata embedded to data ) may lead to different interpretations. Examples should be provided

› FAIRification for a specified purpose has an impact on interpretation (e.g. what is sufficient, what is metadata)
Get in touch

Oya Beyan
Nick Juty
Vassilios Ioannidis
Jolanda Strubel
Ulrich Goldmann
Francesco Ronzano
Manfred Kohler
Andrea Zaliani
Fuqi Xu
Ebtisam Alharbi
Carole Goble
Melanie Courtot
Philippe Rocca-Serra
Susanna Assunta Sansone

Website: [www.fairplus-project.eu](http://www.fairplus-project.eu)
Twitter: [@FAIRplus_eu](http://twitter.com/FAIRplus_eu)
LinkedIn: [www.linkedin.com/company/fairplus/](http://www.linkedin.com/company/fairplus/)
Newsletter:
Sign-up: [http://eepurl.com/ghuHeT](http://eepurl.com/ghuHeT)
Email: FAIRplus-PM@elixir-europe.org
Discussion
Discussions about testing results

1. Should we define comiances scales instead of yes/no evaluation? E.g.:
   0= does not comply to indicator
   1= does not yet comply, under development
   2= fully complies to indicator

2. How does ‘Open by default’ fit with FAIR, especially for indicators related to access conditions and re-use licences?

3. How can we address improving terminology in some of the indicators and how can we get examples of good practices?

4. Should the evaluation of metadata concern the metadata attached to the data item and/or data collection?
Development

Next steps
# Draft guidelines

## INTRODUCTION
- Introduction
- Objectives
- Use of the document

## FRAMEWORK
- Indicators
- Maturity levels
- Prioritization
- Indicators description

## IMPLEMENTATION
- How to evaluate
Indicator description

- Description of each indicator and its respective assessment details
- Indicators ordered by their priorities

**EXAMPLE**

<table>
<thead>
<tr>
<th>F1-01M Metadata identified by a persistent identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principle – as defined by GO FAIR – to which the indicator relates</td>
</tr>
<tr>
<td>This indicator is linked to the following principle: <em>F1 (meta)data are assigned a globally unique and eternally persistent identifier</em>. More information about that principle can be found <a href="#">here</a>.</td>
</tr>
<tr>
<td>Description of the indicator F1-01M</td>
</tr>
<tr>
<td>This indicator evaluates whether or not the metadata is identified by a persistent identifier. A persistent identifier ensures that the metadata will remain findable over time, and reduces the risk of broken links.</td>
</tr>
<tr>
<td>Assessment details</td>
</tr>
<tr>
<td>The persistence of an identifier is determined by the commitment of the organisation that assigns and manages the identifier, so the evaluation of this indicator needs to take into account the persistence policy of that organisation. Such a commitment could be expressed by a university or research institute, by a research infrastructure or by an organisation that issues formal identifiers, such as the International DOI Foundation. A possible way to evaluate this indicator is to verify that the identifier used for the metadata is listed in a registry service like FAIRsharing.</td>
</tr>
</tbody>
</table>
Draft guidelines | Development

Working Group to share **remarks** and **suggestions** about the guidelines

**Testing phase** will bring out comments and suggestions for change and for additional guidance

**Stable version of the guidelines** to be published

https://docs.google.com/document/d/1pDGGL3-BbBJu18KlfZUI3AizKLHXGXdlI_mPtpEWmeg/
Next steps
Testing timeline

~ February 2020
1st Level of testing

March 2020
15th Research Data Alliance plenary meeting

~ June 2020
2nd Level of testing
Testing framework

In the coming week(s), the editorial team will share a *template document* allowing to individually report on the results.

- ID card of the **evaluator** (e.g. discipline, community, profile(s), etc.)
- ID card of the collection of **digital object/resource**
- **Methodology** followed
- **Observations per indicator** (e.g. ambiguity, misunderstanding, priority relevancy, etc.)
- **General recommendations**
Action item and next steps

➢ Share feedback – comments, remarks & suggestions – on the Guidelines
➢ Volunteers for testing

WORKSHOP #7
13 February 2020

09.00 - 10.30 CET | Morning session
17.00 - 18.30 CET | Afternoon session
Resources

- RDA FAIR data maturity model WG
  https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

- RDA FAIR data maturity model WG – Case Statement

- RDA FAIR data maturity model WG – GitHub

- RDA FAIR data maturity model WG – Collaborative document
  https://docs.google.com/spreadsheets/d/1gvMfbw46oV1idztsr586aG6-teSn2cPWe_RJZG0U4Hg/edit#gid=0

- RDA FAIR data maturity model WG – Indicators prioritisation
  https://docs.google.com/spreadsheets/d/1mkjElFrTBPBH0QViODexNur0xNGhJqau0zkL4w8RRAw/edit

- RDA FAIR data maturity model WG – Indicators prioritisation survey results
  https://drive.google.com/open?id=11hyAYCKz_NV0Ob9-vLPqjN9LCarOFmc3

- RDA FAIR data maturity model WG – Guidelines
  https://docs.google.com/document/d/1pDGGL3-BbBJu18KlfZUI3AizKLHXGXdli_mPtpEWmeg/

- RDA FAIR data maturity model WG – Mailing list
  fair_maturity@rda-groups.org
Thank you!