

Workshop report

FAIR data maturity model Working Group

Online meeting #7 - 13th February 2020

Project	RDA FAIR data maturity model working group	Date & Time	13 February 2020 07:00 — 08:30 UTC 13 February 2020 15:00 — 16:30 UTC
Type	Online meeting	Location	Google Meet
Meeting Chairs	Keith Russel Edit Herczog	Issue date	20 February 2020

Objectives

The primary objective of this 7th workshop was to report back to the RDA FAIR data maturity model members about the testing phase. During January, several volunteers compared the FAIR data maturity model indicators against digital objects and methodologies. This exercise shed light upon several issues. The editorial team selected the most critical issues and proposed them for discussion / resolution. This meeting also tackled the potential scoring mechanisms that were put for discussion back in October 2019.

Agenda

1. Welcome, objectives of the meeting
2. Roundtable
3. State of play
4. Testing phase overview
5. Testing insights – feedback
6. Testing insights – general discussion
7. Potential scoring mechanisms
8. Action items and next steps

Useful links

- [RDA 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](#)

- [RDA FAIR data maturity model WG – Indicators prioritisation](#)
- [RDA FAIR data maturity model WG – Indicators prioritisation survey results](#)
- [RDA FAIR data maturity model WG – Guidelines](#)
- [RDA FAIR data maturity model WG – Mailing list](#)
- [RDA FAIR data maturity model WG – Workshop #6 material](#)

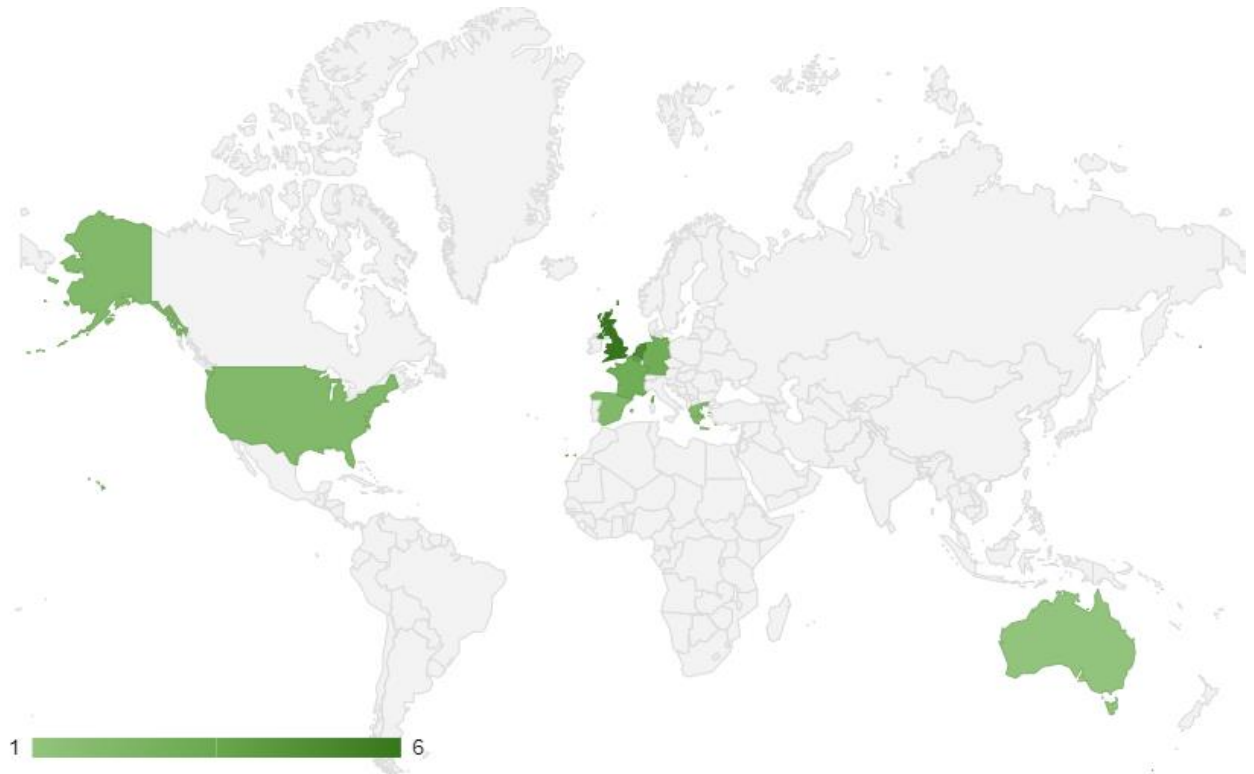
Participants

The workshop was well attended. Here below is a non-exhaustive list of the participants.

Name		Affiliation
Anne-Caroline Delétoille	FR	Institut Pasteur
Anusuriya Devaraju	DE	PANGAEA / University of Bremen
Carlos Casorrán Amilburu	BE	European Commission DG RTD
Carole Goble	GB	University of Manchester
Christophe Bahim	BE	PwC, Editor team
Dimitra Mavraki	GR	Hellenic Centre for Marine Research,
Ebtisam Aharbi	GB	PhD, University of Manchester
Edit Herczog	BE	Chair, Vision & values SPRL
Elli Papadopoulou	GR	ATHENA Research & Innovation Center
Erik Schultes	NL	GO FAIR
Françoise Genova	FR	Strasbourg Astronomical Data Centre
Ge Peng	US	North Carolina State University / NCEI
Ingrid Dillo	NL	DANS / H2020 FAIRsFAIR
Juan Bicarregui	GB	Science and Technology Facilities Council
Julianna Pakstis	US	Children's Hospital of Philadelphia
Keith Russell	AU	Chair, ARDC
Konstantinos Repanas	BE	European Commission DG RTD
Leyla Garcia	DE	ZBMED
Makx Dekkers	ES	Independent Consultant, Editor team
Mark Wilkinson	ES	GBGP, UPM – INIA
Marta Teperek	NL	TU Delft
Mustapha Mokrane	NL	DANS
Nick Juty	GB	University of Manchester, ELIXIR-UK

Oya Beyan	DE	EOSC FAIR WG & FAIRplus CMMI
Patricia Herterich	GB	Digital Curation Center
Pete McQuilton	GB	University of Oxford
Rob Hooft	NL	Dutch Techcentre for Life Sciences
Romain David	FR	INRA

Here below is a map representing the provenance of the different participants



Content¹


The workshop was designed to be as interactive as possible: interaction was encouraged during the presentation of the set of issues derived from the testing phase. The editorial team and the participants went over the issues one by one, discussing the different viewpoints. As a result, the meeting was fruitful and enabled lively discussions. The major issues discussed and the comments from the members of the Working Group can be found later in this document.

1. The Chairs opened the workshop, welcomed the participants and addressed the agenda. The approach to the Working Group was again presented:
 - Challenges rising from the different interpretations of FAIRness
 - Bringing together the relevant stakeholders to discuss and build on existing expertise and different approaches
 - Intended results: i) set of core assessment criteria for FAIRness ii) FAIR data maturity model & toolset iii) RDA recommendation and iv) FAIR data checklist.

 **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](#)

2020-02-13 www.rd-alliance.org - @resdata11 

¹ Please note that some of the slides are displayed for information purposes. The full presentation can be accessed via the RDA FAIR data maturity model WG web page.

Slide 3 | Welcome and objectives of the meeting

As usual, the Chairs insisted that despite all the challenges arising when designing indicators, the purpose of the WG was **NOT** to re-design the FAIR principles. As there are currently different interpretations of what the FAIR principles entail, the primary goal is to build a common understanding.

In addition, the chairs reminded the participants that all the presentations and reports are on the RDA FAIR data maturity model WG [web page](#) and that the members are encouraged to participate via the dedicated [GitHub repository](#).

2. The Chairs and the editor team introduced themselves, after which the participants were kindly invited to state their affiliation and write what their role is in their organisation via the chat window.
3. The editorial team reported on the current state of development: what steps have been taken and what steps remain to be taken.

RDA
RESEARCH DATA ALLIANCE

State of play

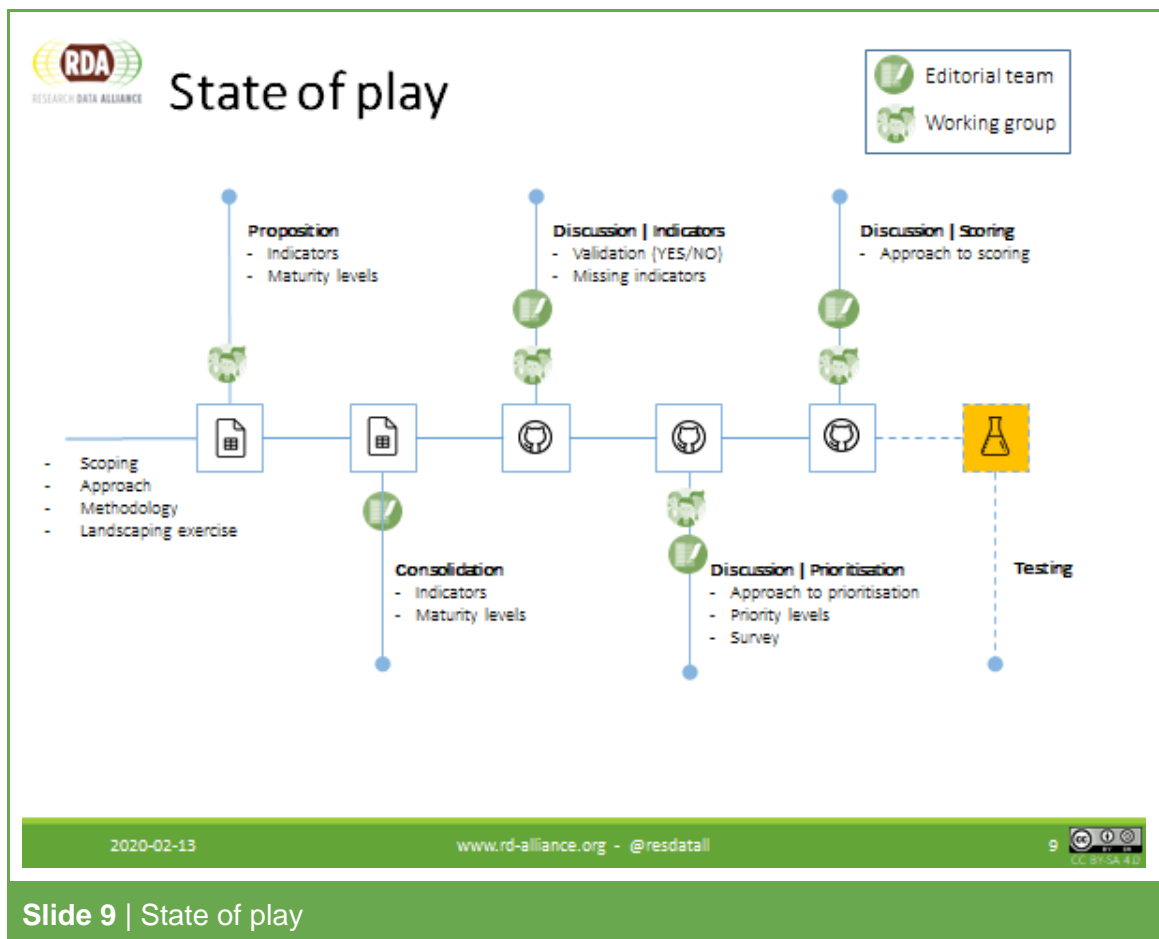
1. Definition	DONE
2. Development	DONE
i) First phase	DONE
ii) Second phase	DONE
3. Testing	ONGOING
4. Delivery	ON HOLD

* Any comments are still welcomed with regards to the output produced during the first phase | [GitHub](#)

2020-02-13 www.rd-alliance.org - @resdata11 8 CC BY-SA 4.0

Slide 8 | State of play

As illustrated on the slide above, the editorial team reminded participants that at the outset of the working group a methodology was designed. This methodology is composed of four main phases. In the beginning of 2020, the editorial team rolled out the testing phase. Then, the output of the testing phase – once aggregated and validated – will serve to update all the deliverables included in the fourth phase, the delivery phase.



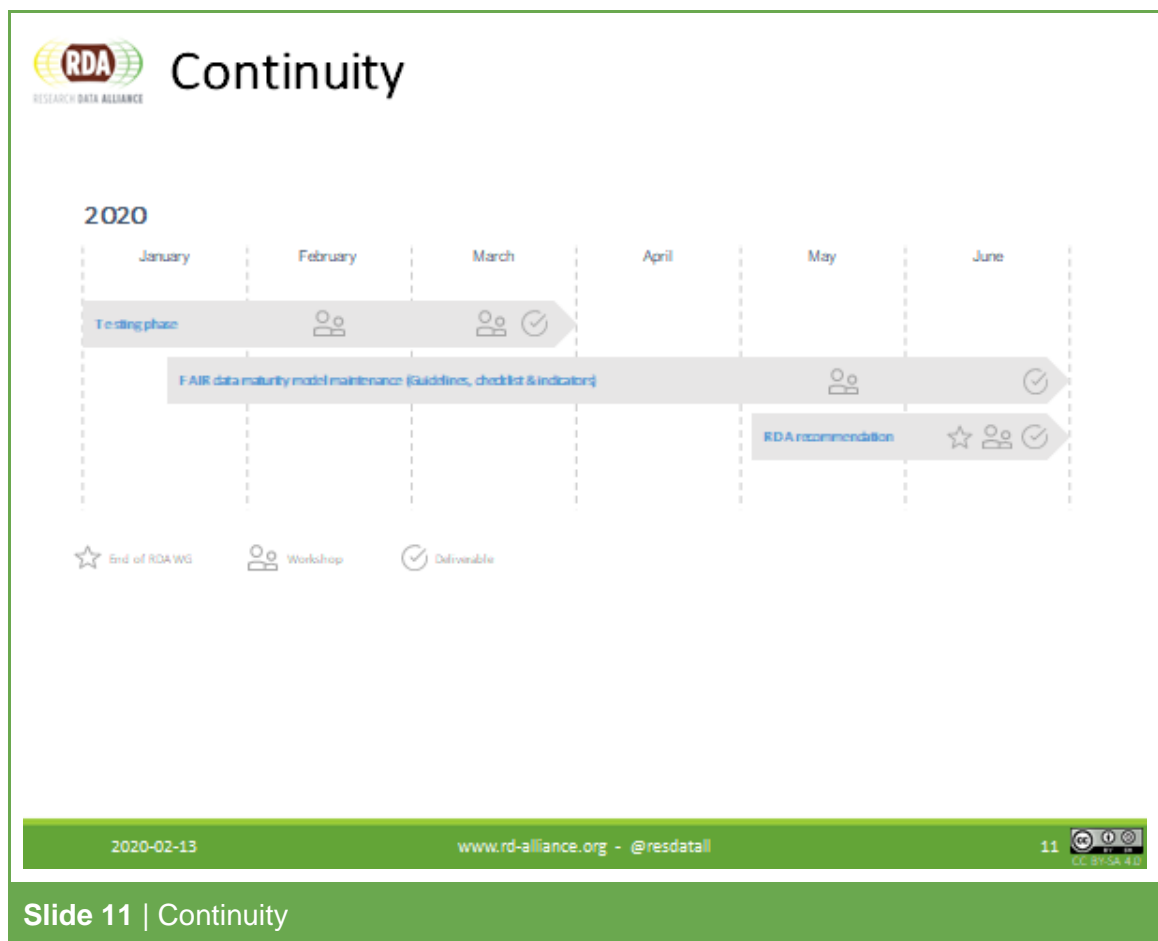
As illustrated by the slide above, the Working Group was first invited to propose potential indicators to measure the FAIRness of a digital resource. The editorial team then consolidated all the contributions, which resulted in a set of 51 indicators.

That consolidated set was shared for comments on the dedicated GitHub. Additionally, the editorial team made proposals for prioritisation and scoring. Discussions related to these three topics (i.e. indicators, prioritisation and scoring) were happening in parallel on the GitHub.

In order to facilitate the consensus process about prioritisation, the editorial team put together a survey. Based on the outcome of the survey, the priorities were frozen and further discussion was postponed to after the testing phase.

Last year, the editorial team initiated a pilot for testing the indicators. The feedback collected in the pilot testing phase allowed to better structure the full testing phase. This full testing phase has been set to run from January until mid-March.

As of February 2020, the editorial team is currently assisting volunteers who are testing the indicators and collecting / aggregating the feedback.



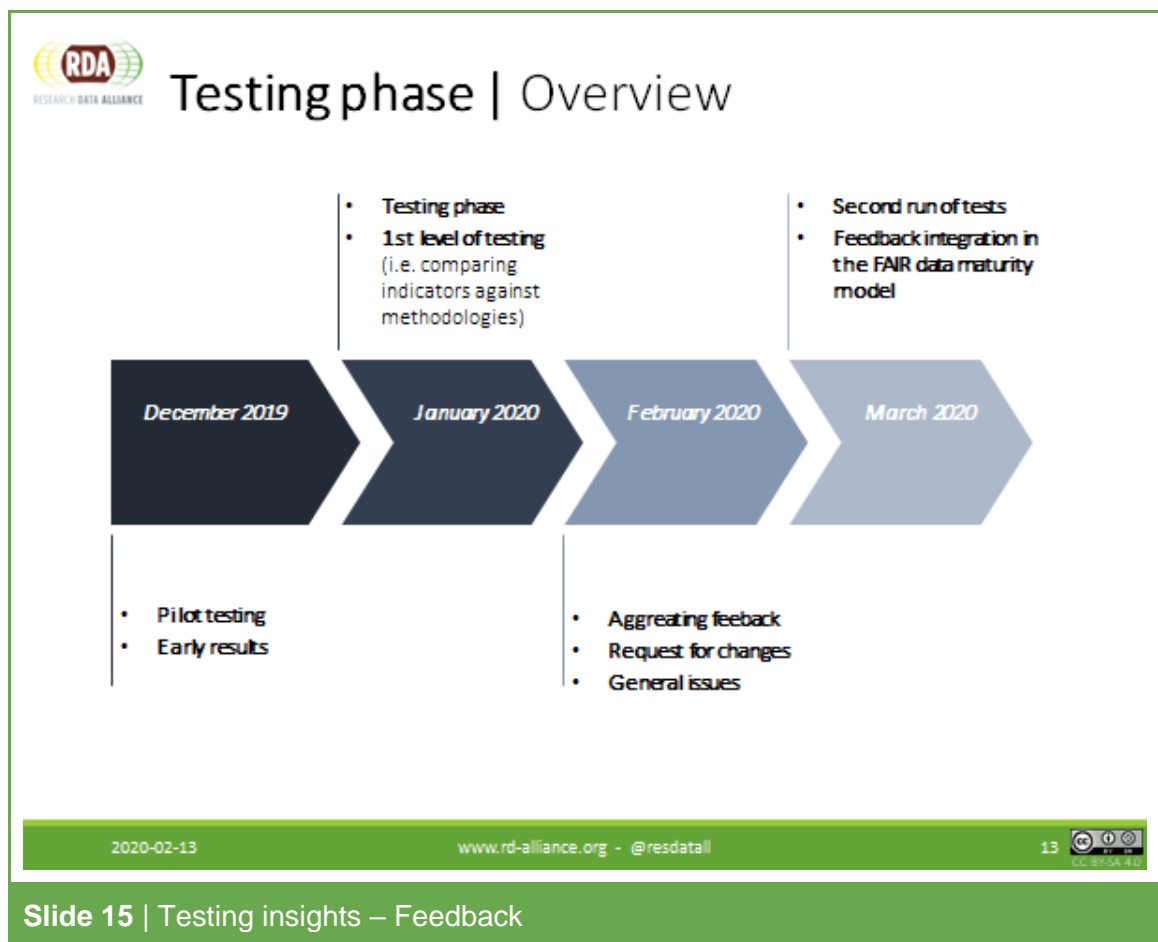
The editorial team touched upon the planning of 2020. Three work streams have been identified.

1. Testing phase – which will run until mid-March. The final results and conclusions will be presented at the next RDA plenary.
2. FAIR data maturity model maintenance – update of the guidelines and proposal for a checklist based on continued feedback during the testing phase.

A stable version of the guidelines will be presented in March at the next RDA plenary.

3. RDA Recommendation – the editorial team will work to submit the deliverables for publication as an RDA recommendation; this proposed Recommendation will be presented during a workshop in mid-June.

4. The editorial team walked the participants through the high-level testing phase timeline. It is worth mentioning that the two levels of testing previously identified -- (i) comparing the indicator against methodologies and (ii) comparing the indicators against datasets -- have been merged into a single testing phase.



Slide 15 | Testing insights – Feedback

5. As outlined in the introduction, the editorial team aggregated the feedback and categorised the issues into:

- Comments on indicators
- General issues
- Specific issues
- Information needs

The issues illustrated below were presented, explained and discussed during the workshop.

Comments on indicators General issues Specific issues Information needs

- There are (too) many indicators. However, others note that this level of granularity is useful, so you need to think about all the aspects
- Testing the indicators provided suggestions for improving existing evaluation approaches or existing standards
- Issue about distinguishing indicators for metadata separate from data, which does not work for resources with embedded metadata
- Overlap between indicators (e.g. across principles F1/A1 and F2/R1) which is the result of FAIR principles not being entirely independent
- Some indicators are conditional, e.g. the ones on authentication, authorisation, references and consent – if not applicable, they should not 'count'
- Several indicators require compliance with community standards, but the question is who defines them?
- If data is an ontology, different set of indicators or different priorities may be needed
- One tester proposes to do away with all priorities entirely

2020-02-13 www.rd-alliance.org - @resdata11 15

Here below is a non-exhaustive list of observations and comments made by the participants:

- Is it necessary to separate metadata from data? It is not written as such in the FAIR principles. F2 and F3 expect metadata to be separated from data, yet they are permanently linked.
- The “metadata should outlive the data” can be sufficiently satisfied if eternal life of a mixed metadata/data file is guaranteed.
- Some people advocate that the principles should be open for interpretation/evolution; others are of the opinion that they are not open to interpretation/evolution. It is the implementation that should be discussed.
- The application of the FAIR principles should take into account current community practices. Some communities may not have the possibility to move quickly towards

full FAIRness or even reach full FAIRness at all (i.e. people need to be aware that 100% FAIRness will not always be possible). It is more important to provide 'stepping stones' to more FAIRness.

- The word “community” is mentioned only once in the FAIR Principles (R1.3), but it has become clear that consensus within communities is a central concept in FAIR.
- GO-FAIR is developing for a matrix of community profiles to help communities to come together, building larger communities around common needs
- Metadata embedded in metadata is an example why metadata should have identifiers.
- F2 and R1 may both be about rich metadata but they have different purposes. F2 is about discovery, while R1 is about utility/reusability. Different kinds of metadata are required to comply with these facets. F1 and A1 don't overlap either. F1 is about identification, while A1 is about access.
- Priorities of indicators may depend on the way communities think about the objects relevant for their community, and on the existing community standards. Even where FAIR would aim for cross-community interoperability, at this point in time, community standards are more important than cross-community standards.
- The consistency of the wording of the indicators needs to be improved.

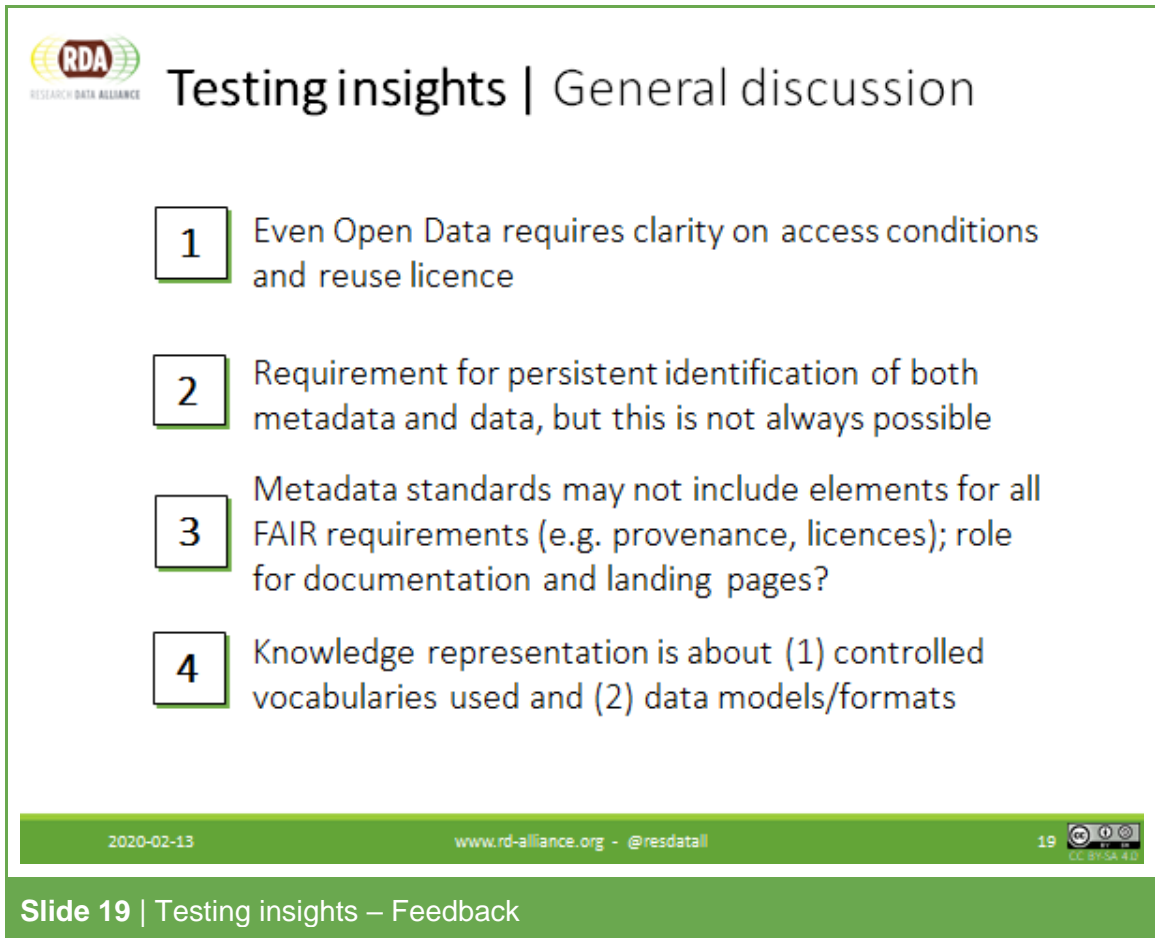


- FAIR principles are aspirational and ambitious, aiming at full machine-understandability, but current practices are not well aligned at this point in time
- Identification is a major issue: there are various comments, some favour identification of metadata over identification of data, others data over metadata, others see both as equally essential, but there is also a comment that having identifiers for both is not common practice
- There seems to be a role for landing pages and other human-readable documentation in providing information, in addition to structured metadata
- Requests for adding maturity levels > scoring
- Data comes in different granularities: whole dataset or part of dataset or individual data items (e.g. observations, concepts)
- Different perspectives on metadata and how it relates to data:
 - repository level / collection level / dataset / data item level metadata
 - separate metadata records or embedded metadata

Slide 16 | Testing insights – Feedback

- The FAIR principles should be seen as a goal to strive for. FAIR is a continuum towards good research data management practice. FAIR principles aim to set a new method and practice to handle data assets. It is very understandable that existing datasets cannot tick all boxes. The aim should be to have measurable pathways for better practice.
- FAIR is a view for the long term: the convergence among communities should be at the center of it, but this will take time.
- A focus on (i) minimal, (ii) important and (iii) appropriate elements could help the implementation of the FAIR principles. This is why the word 'plurality' was chosen in the FAIR principles – to not pre-define which facets are important for which communities.
- Complete FAIRness is the ideal in the context of good RDM, but it should not be used to judge. Fitness for use is something to keep in mind as well – very often this determines the choice for reuse.

6. Additionally, the editorial team proposed some issues for a general discussion. The purpose was to move towards a common understanding and agreement in order to reflect this in the FAIR data maturity model (i.e. indicators and guidelines)



1 Even Open Data requires clarity on access conditions and reuse licence

2 Requirement for persistent identification of both metadata and data, but this is not always possible

3 Metadata standards may not include elements for all FAIR requirements (e.g. provenance, licences); role for documentation and landing pages?

4 Knowledge representation is about (1) controlled vocabularies used and (2) data models/formats

2020-02-13 www.rd-alliance.org - @resdataall 19 CC BY-SA 4.0

Slide 19 | Testing insights – Feedback

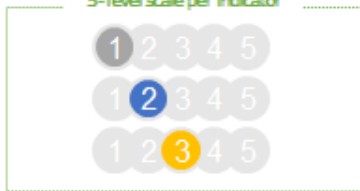
Here below are some key takeaways – views from the audience – from the discussion:

- Sometimes licences are applied at the level of collection. The challenge is to find that licence because it not in the metadata for an individual dataset.
- Application of licenses for all data whether public, restricted or for specific users should be encouraged. The legal position is that resources cannot be reused if there is no explicit licence. Adding a licence will make things easier in the long run.
- Open is not enough, a proper license – like CC-0 – is needed. The current DMPs require information about the licence.
- There is a need to define what Open Data is before starting a discussion about Open Data reuse licences. There are different [shades of Open Data](#).
- Provenance can require a full data model depending on the information provided

- Generic metadata serves for cross-domain interoperability, but within a community one needs a deeper level of detail. To move from the deeper level to the generic level, the starting point should be establishing cooperation with 'nearby' communities, where the benefits of reuse are clearer.
- If one can put elements in metadata one should do so. If current community practices and standards don't support some of the metadata elements necessary, those practices and standards might need to be changed.
- Good practice guidelines for the structure of landing pages might be useful for automated extraction of elements, such as the link to the data.
- In some standards – like DCAT – the landing page is explicitly included as a metadata component.
- Landing pages must be harvestable by all research engines which is not often the case and demonstrated by some case studies.
- Knowledge representation needs to be grounded in a shared understanding of each term (knowledge representation for machines always comes down to an agreement between humans about what a "tag" means, and how it should be treated by software).
- Knowledge representation can be seen as turning data into unambiguous knowledge.

Scoring mechanisms | Overview

5-level scale per indicator



- Five levels of compliance
- Per indicator – aggregated per FAIR area
- Non applicable or consideration/implementation as options
- Useful for giving credit for evolution and helping people to improve

FAIRNESS per area

	Essential	Important	Useful
Level 0	○		
Level 1	●		
Level 2	●	●	
Level 3	●	●	●
Level 4	●	●	●
Level 5	●	●	●

○ None of the indicators are satisfied
 ● Half of the indicators are satisfied
 ● All indicators are satisfied

- Measurement based on priorities
- Per indicator – aggregated per FAIR area
- Score determined based on the compliance to priorities
- Provides a 'measure of FAIRness'

Overall FAIRNESS



- Measurement based on priorities
- Per indicator – overall score
- Aggregated score
- Provides a quick view of how priorities are met -- but does not give detailed view



Slide 21 | Scoring mechanisms

7. Last year, the editorial team proposed a first method to score the indicators. Discussions (during [workshops](#) and [offline](#)) and comments from the early testing phase have shaped that first proposition. Two additional propositions have arisen. These 3 different approaches towards scoring the indicators were discussed with the WG to determine the way forward. As a result, the editorial team will refine the scoring mechanisms and propose a hybrid method comprising a 5-level scale and a score per FAIR area.

Here below are some observations made by the participants.

- Scoring is a dangerous phrase because it suggests that a score is meaningful, yet FAIR is not a competition. Maturity indicators should help providers find areas where they can increase FAIRness at a reasonable cost – cost/benefit analysis guided by a set of tests that highlight places where you could do better.
- FAIR principles can help to judge minimal FAIRness and efforts in RDM. It is an incentive to improve metadata standards.
- The first proposal (i.e. 5 scale per indicator) is worthwhile and aggregation is also useful (i.e. view on the journey and view on the results). An overall score for

FAIRness might not be useful, but some guidelines to improve the score per FAIR area (F, A, I and R) could be helpful; i.e. it would be better to say that resources are on a journey to be more FAIR and then provide suggestions for ways in which they could be more FAIR).

- It could be interesting to visually see where you stand per criterion, for example with radar charts.
- A standard score is quite difficult to provide – particularly as the indicators are so varied and may have different weights for different communities.
- Would it be possible to say that (meta)data are FAIR if they score above a certain threshold (e.g. 60%)? Or simply say “FAIR enough” and “Totally FAIR”?
- An average is misleading. Putting a number on something can be problematic. The score should be at the area level.
- “Maturity” stays a better word than “score” / “metric”.
- Radar plots are only good practice if there is a natural order in the measurements.
- The tests for the indicators are binary, but they should also include “not applicable” to cover cases where an indicator is not applicable.

The editorial team will further communicate about the scoring mechanism through [GitHub](#).

Follow-up action plan

Working Group members are invited to:

- Share feedback, comments & suggestions – on the [Guidelines](#)
- Discuss proposals for changes in priorities on [GitHub](#) (issues will be created)
- Contribute to GitHub discussion on [scoring](#)

We're also looking for volunteers for further testing; please contact us!

The next workshop will take place physically in Melbourne. A possibility to remotely participate in the meeting will be foreseen. The agenda will be shared soon through the usual channels.

WORKSHOP #8

15th RDA PLENARY IN MELBOURNE
19 March 2020

11.30 – 13h00 (GTM+11) | Breakout 4