Workshop report

FAIR data maturity model Working Group

Online meeting #12 — 3rd December 2020

Project	RDA FAIR data maturity model working group	Date & Time	3rd December 2020 12:00 — 14:00 CET
Type	Online meeting	Location	Zoom meeting
Meeting Chairs	• Keith Russell • Shelley Stall • Edit Herczog	Issue date	2020-12-11

Agenda

1. Presentation of the agenda and objectives of the meeting [5 minutes]
2. History and achievements of the work in the FDMM WG [5 minutes]
3. Results of the survey on FAIR assessments [5 minutes]
4. Moderated discussion [90 minutes]
5. Conclusions and agreement on further work [15 minutes]

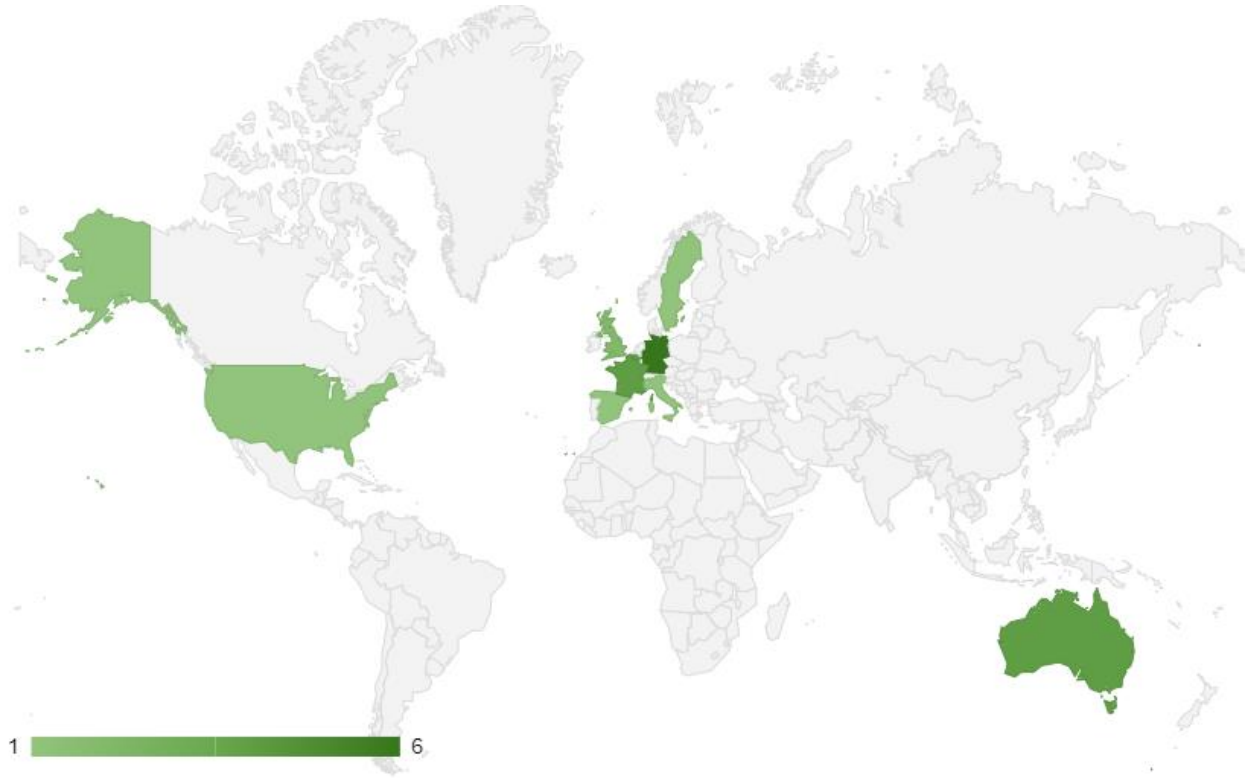
Attendance

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

Name		***Affiliation***
Amandine Kaiser	DE	German Climate Computing Center (DKRZ)
Andrey Vukolov	IT	Elettra Sincrotrone Trieste
Anette Ganske	DE	Technische Informationsbibliothek (TIB)
Anusuriya Devaraju	AU	TERN Australia
Carlos Casorrán Amilburu	BE	European Commission DG RTD
Carmen Reverté	ES	IRTA
Christophe Bahim	BE	PwC, Editor team
Edit Herczog	BE	Chair, Vision & values SPRL
Felix Rau	DE	University of Cologne
Françoise Genova	FR	Strasbourg Astronomical Data Centre
Irina Bastrakova	AU	Spatial Data Architecture
Keith Russell	AU	Chair, ARDC
Konstantinos Repanas	BE	European Commission DG RTD
Maarten Vermeyen	AU	University of Antwerp
Maggie Hellström	SE	ICOS ERIC's Carbon Portal data center
Makx Dekkers	ES	Independent Consultant, Editor team
Margie Smith	AU	Geoscience Australia
Olivier Rouchon	FR	CINES

Parnian Kiani	DE	German Center for neurodegenerative diseases - DZNE
Parnian Kiani	DE	German Center for neurodegenerative diseases - DZNE
Robin Burgess	GB	University of New South Wales
Robin Rice	GB	University of Edinburgh
Romain David	FR	INRA
Roman Gerlach	DE	University of Jena
Shelley Stall	US	Chair, American Geophysical Union
Sophie Aubin	FR	INRAE

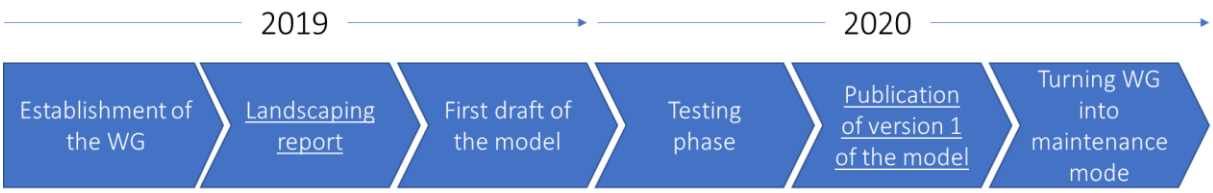
Here below is a map representing the provenance of the different participants



Meeting notes

The chair welcomed the participants and introduced the agenda of the meeting. On the agenda of the meeting were the history of the working group, the report on the survey on bridging the gap between funders and communities, discussion items and the outline of the maintenance phase.

The meeting started with a tour de table, where participants were asked to share their name and affiliation. Then the chairs introduced the context and what the FAIR data maturity model was trying to address and the output of this working group. Further, the chairs clarified what this working group does not aim to do (e.g. develop yet another evaluation method nor define how the core criteria need to be evaluated) and listed the milestones for the past one year and half.



Besides, the chairs unveiled the tentative schedule until the end of the 2020, which starts with a report on the survey among funders and communities, establish a work plan for 2021 and wrap up the year by participation in the CODATA International FAIR convergence symposium.

On top of that, the chair reported three major publications of the working group.

- Publication of the FAIR Data Maturity Model: Specification and Guidelines as an RDA recommendation, 25 June 2020
- EOSC-SYNERGY Intermediate report on technical framework for FAIR principles implementation, 2 Sept. 2020
- Publication of The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments as a paper in the Data Science Journal, 27 Oct. 2020

The editors introduced the survey on FAIR assessments which tries to improve the understanding of benefits and challenges of the FAIR assessments from the perspective of funders and research communities. This survey had a twofold objective: (i) formulate conclusions and recommendations on the level of policy, and (ii) finding out how the research community and the funders' community want to use the model and what changes they want to see.

Results were analysed and the editors reported some trends already. First of all, the survey results were divided into two parts. On the one hand, results were about policy & adoption matters. For instance, it turned out that both communities and funders see the RDA as a neutral platform to bring stakeholders together and create cross-community understanding. On the other hand, results were about the future work (for the FDMM). Both communities and funders agreed that any scoring mechanism needs to be considered in the context of community targets and practices, and that it is necessary to make the assessment approach more practical and that there is a need to provide actionable guidelines.

The full report will be published on the RDA FAIR data maturity model WG web page.

Makx Dekkers introduced the three discussion topics:

- Evolution of metadata practices to improve FAIRness within and across communities
- Challenges of different data granularities for FAIRness (collections, datasets, data items)
- Aspects to be considered for evaluation of FAIR assessment tools and services

Topics	**Questions**	**Observations from the discussions**
#1 Evolution of metadata practices to improve FAIRness within and across communities	If there are different approaches to **'rich' metadata**, is there a need for mapping services to enable wider interoperability? If so, who would be	○ There is no way around using metadata standards. Well-established ones, e.g. DCMI, should be preferred. ○ Focus needs to be put on the purpose for reuse, i.e. there are different levels of reuse, each one requiring different amounts and/or types of metadata ○ The following paper 'A multi-level metadata approach for a Public Sector Information data infrastructure' can help around the use cases for different levels of metadata

best placed to develop such services?	○ The work of the RDA Metadata IG, including the discussion of 'minimum' metadata fields, how to prioritise them and especially how to provide useful content for these fields can help too. ○ There is currently ongoing work in different venues to discuss a minimal common set of metadata. ○ A disciplinary 'rich' metadata profile will not mean minimal ○ *'Rich'* shouldn't be interpreted as just *'quantitative'.* Metadata *quality* must also be considered (just as much as for the data). ○ *'Rich enough'* is entirely related to the use case. To cite the data, one just need basic bibliographic metadata. But, to reuse the dataset in an experiment, one need more '*richness'*
What could be the **role of repositories** in improving FAIRness of domain-specific metadata? How can repositories pave the way towards cross-domain metadata exchange?	○ It's unrealistic to have all information in one place or stored according to one mechanism. In other words, different parts of the metadata set associated with a given digital object may be stored at different places. ○ FAIR needs to be implemented in repositories, i.e. *'if you control the storage you control the metadata […]'* ○ To share metadata between institutions in a distributed way, metadata needs to be stored in a distributed way. ○ Mirroring of metadata is crucially important for reuse of data – be organised like the Web, i.e. in a decentralised way. ○ Metadata are the link between users, i.e. machine and human. However, minimum metadata and cross-domain schema for metadata is a (FAIR) challenge ○ There are different levels of metadata necessary for FAIR compliance; cross-domain metadata and community-specific metadata ○ There is an absolute need for a minimal cross-domain base for metadata. ○ Various metadata fields could be created and administrated by different organisations – as long as the digital object's PID is used as index. However, this 'fragmentation of responsibilities' will make the management more complex. ○ Mapping between cross-domain profiles can be one-sided. For instance, there is only a small set of core elements between Geospatial (ISO19115) and Dublin Core. ○ Content-based addressing and packaging of the metadata inside repositories is needed ○ Metadata profiles need to be modular, extensible (to allow for use by different communities), public, formally defined, documented, updated and promoted. ○ Technical standards may never solve the issue. Engagement of user/scientific communities is required. ○ The W3C issued a best practice for descriptive metadata .
What could be **the role of Semantic Web/Linked**	○ There is a lot of work on "mapping" in groups dealing with semantics, ontologies, etc. Tools already exist – more people need to be engaged with testing these in practice.

	Data/Knowledge Graph approaches in understanding knowledge representation?	
	General observations	○ FAIR targets, i.e. '*FAIR enough'* can be high for some community and be very different in another. ○ The input of different stakeholders is needed to define *'FAIR enough'* in a given situation. E.g. *'FAIR enough'* could be defined based on targets set by researchers, communities. They could define what are the minimal requirements, i.e. the minimal set of metadata for a dataset. ○ Minimal set of interoperable and understandable metadata across communities should be further discussed ○ The idea beyond requirements is to encourage FAIRness and reuse. ○ Funders struggle with the tools at their disposal to say what is *'FAIR enough'* ○ The geographical data initiative (OGC) which makes interoperability and reusability of data possible can be an inspiration.
#2 Challenges of different data granularities for FAIRness (collections, datasets, data items)	Is the granularity something that individual data creators decide, or is it standardised? Should there be domain-specific or cross-domain agreements?	○ The type of granularity was questioned (e.g. temporal, geographical, semantic…) ○ Different levels of granularity are sometimes required, e.g. export of a wide image to a sub-image in astronomy. Granularity depends on the purpose. ○ Granularity will depend on the ability to respond to research questions, which is the ultimate goal. ○ Metadata varies also across different granularities, e.g. experiment, collection, data item… ○ The collection will have its own metadata ○ When talking about *'collections',* this really moves into the structural metadata domain, which is probably not encompassed by the FAIR data maturity model. ○ The *'right'* granularity of data items (or FAIR digital objects) is in the eye of the beholder. ○ If there is a *'too high'* granularity, there is a risk that many fields [of the metadata set] remain empty ○ PIDs need also to be granular and connected (e.g. when dividing datasets)
	How do decisions on granularity affect interoperability and wider reuse? Should the FAIR Data Maturity Model include examples of best practices for this?	○ It all depends on the community's metadata standards and which data are made Open and/or FAIR. ○ The context in which the object is evaluated should be communicated in addition to the assessments results ○ DCAT, which is a widely accepted standard for data catalogs and repositories, distinguishes between catalog and datasets, but doesn't specify whether a dataset should be a collection, spreadsheet image, etc. The FAIR data maturity model should be similarly agnostic ○ Apply data maturity standards to processes rather than individual datasets, i.e. having FAIR data maturity

		indicators applied in organisations with the explicit goal of improved data interoperability. It is more interesting to assess whether the process produces FAIR data by default. ○ In the PID community, what should be considered as a *'data entity'* should be decided at the time and point of use, when it needs to be referenced (what is '*granule'* for one person is a *'dataset'* for another)
	General observations	○ It might be too late at the point of reuse to consider whether something is FAIR. If it isn't, there is not much you can do. ○ ExPaNDS project will publish a deliverable on what metadata do one need to capture for FAIR as moving through the experimental stages ○ FAIR enabling processes and FAIR processes are very different. FAIRsFAIR is working on FAIR enabling services. ○ FAIRsFAIR latest report on the assessment of services ○ EC's report on 'Turning FAIR into reality' has also a process view on FAIR ecosystem.
#3 Aspects to be considered for evaluation of FAIR assessment tools and services	Is it important to look at the distribution of effort across the four FAIR areas? E.g. not just F and A, but also sufficient attention to I and R.	○ Different stakeholders have different motivation for assessing FAIRness ○ It may be that researchers' expectations are higher, notably concerning the 'R' aspect while funding agencies are mostly interested in F & A. ○ The weight of the criteria depends on the use case and community needs ○ Reproducibility is very important yet a bit underdeveloped. It depends more on intended use which is difficult to state in generic terms. ○ Funders need to acknowledge the need for resources to support the development of technological solutions that can support interoperability. ○ *'Fitness for use'* or *'suitability'* may change over time and the community changes in composition. Thus, reuse should be re-evaluated regularly.
	How can tools tailored to a specific resource type or domain be compared? Does that stop at assessing whether they correctly implement a specific approach, or should the evaluation also consider whether the tools and services contribute to cross-domain interoperability?	○ The evaluation issue is very difficult. One must be careful to be as inclusive as possible when considering the list of tools to evaluate – as well as mapping them with the FAIR aspects being addressed, e.g. 'F', 'A'. ○ Each tool has biases. There is a need to assess the tools and identify where the biases are. ○ The EC FAIR report has as priority that the evaluation tools are compared, and their biases are identified. ○ Detailed description of each tool, in a given situation, is needed so the usefulness is well understood. Also, the context of the assessment needs to be communicated when providing the assessment results. ○ Tools must be adapted to what is assessed, e.g. data access, dataset, etc. ○ In bibliometrics, each of the citations databases give slightly different scores. If they are different, the subject area looked at is not covered well enough by the citation database and there is need to find another way to evaluate.

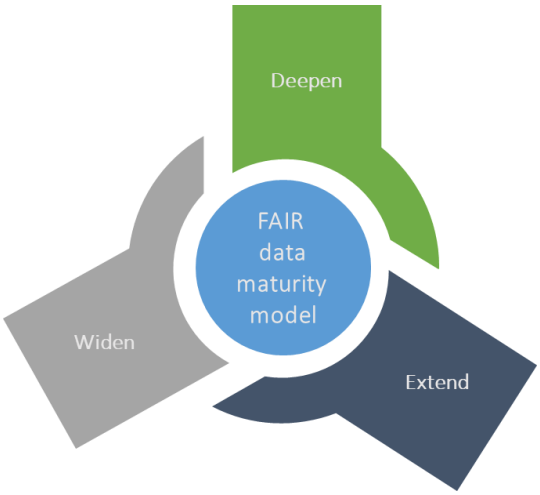
	○ The limitation of tools for automated assessment, is that you can only test what is available for machines to read. So, they naturally will be very skewed towards linked data / semantic web aspects of datasets. When it comes to tools recording the results of human assessments, it's down to the question asked.
How can an initiative like the GO-FAIR FAIR Implementation Profiles help in understanding the results of assessments done by different tools?	○ What if the EC would prefer a single FAIR evaluation tool? ○ Extensive tests in different community contexts are essential ○ Implementation profiles, i.e. how one wants to implement FAIR, are needed. ○ FIPs could help to understand which FAIR enabling tools/practices are used
General observations	○ OpenAIRE is working on linking the OpenAIRE Provide guidelines to the FDMM indicators ○ FAIRsFAIR metrics were built on the FDMM indicators ○ There is work in progress on Agroportal towards specifying FAIR maturity evaluation for semantic resources and implementing the FDMM indicators in the portal as a service ○ It is essential to assess how the FDMM is *fit-for-purpose* in order to suggest useful updates for the next version ○ It was proposed to organise a workshop where different tool developers would present their approach and where communities could contribute with questions and concerns (co-located with RDA 17). ○ The workshops should serve to identify biases of (FAIR) assessments tools.

Lastly, the chair listed the next steps for the FAIR data maturity model (WG), which are:

- Create further connections with other groups – having an interest in contributing to FAIR assessments – for instance:
  - FAIR for Research Software (FAIR4RS) WG
  - FAIRsharing Registry WG: connecting (meta)data standards, repositories and policies
  - CURE-FAIR WG (FAIR curation)
  - WDS/RDA Assessment of Data Fitness for Use WG
  - Etc.
- Identify which indicators of the FDMM are the ones to begin the FAIR journey with
- Map the FDMM indicators to the different stages of the lifecycle of (FAIR) data

- Look into establishing a 'community of practice' (CoP) for the FDMM

Next to that, the chair presented the three-pronged approach towards updating the model:



- Deepen – Work on different elements to understand them better (e.g. metadata, granularity etc.)
- Widen – Follow up on the FDMM indicators and the FAIR assessment methods in different domains, sectors (e.g. private and public research) and regions.
- Extend – Follow the FDMM indicators through the data lifecycle (e.g. from the creation to archiving)

Finally, Edit Herczog thanked the participants and the editorial team for their contribution to the work of the Working Group and welcomed continued contributions for the members of the Working Group in the next year.