FAIR Data Maturity Model

Online meeting #1
21 & 22 February 2019
Agenda

› Welcome, objectives of the meeting
› Round table
› Introduction to the Working Group
› Survey results
› Presentations from existing approaches
   › DANS FAIR data assessment tool, FAIR checklist
   › FAIR Metrics
   › Data Stewardship Wizard
   › RDA SHARC IG
   › Dataset Fitness for Use
   › ARDC FAIR self-assessment tool
› Results of preliminary analysis
› How to contribute
› Logistics
› Conclusion
Roundtable

▶ Short introduction of the chair and editor team
▶ All other participants, please type your name and affiliation in the chat window
Problem:
- Ambiguity and wide range of interpretations of FAIRness
- Lack of a common set of core assessment criteria and a minimum set of shared guidelines

Approach:
- Bring together stakeholders
- Build on existing approaches and expertise

Intended results:
- RDA Recommendation of core assessment criteria
- Generic and expandable self-assessment model
- Self-assessment toolset
- FAIR data checklist
Introduction to the Working Group - 2

Target audiences:
- Researchers, data stewards, other data professionals
- Data service owners, e.g. infrastructure, repositories
- Organisations that manage research data
- Policymakers

Connections:
- RDA Disciplinary Framework Interest Group
- RDA Domain Repositories Interest Group
- Other RDA groups

Scope of the assessment:
- Datasets
- Data-related aspects (e.g. algorithms, tools, workflows)
Any questions about the *approach* outlined?

① Do you agree with the proposed approach and intended results?

② Do you have other suggestions concerning the scope of the work?

③ ...
Work methodology

Method step 1
Articulate objectives

Method step 2
Define stakeholders and users

Method step 3
Establish liaisons with other RDA groups

Method step 4
Identify and analyse existing approaches

Method step 5
Identify issues and additional areas of interest

Method step 6
Agree work structure and time plan

Method step 7
Consider each of the FAIR principles and their facets

Method step 8
Compare and consolidate metrics per facet

Method step 9
Identify levels per metric

Method step 10
Propose pathway for improvement per metric

Method step 11
Identify dependencies, overlaps and gaps

Method step 12
Harmonise metrics across FAIR areas

Method step 13
Identify overall maturity levels and pathways

Method step 14
Draft core assessment criteria

Method step 15
Map existing approaches to draft assessment criteria

Method step 16
Apply draft assessment criteria to selected collections

Method step 17
Compare results and improve criteria

Method step 18
Finalise core assessment criteria

Method step 19
Describe overall pathways/guidelines

Method step 20
Publish results

Definition
Development
Testing
Delivery
Proposed approach to development

Consider the assessment of the four FAIR principles in four ‘strands’

Possibly create a fifth strand for issues related to the environment around the FAIR principles, e.g.

- Characteristics of projects, workflows and tools
- Open vs. closed/embargoed data
- Curation, maintenance and governance
- Certification (what and who/how)
Introduction to the Working Group - 5

Tentative timeline 2019

**Workshop #1** [February]
- Methodology and scope
- Infrastructure issues
- Documents sharing

**Workshop #2 (face-to-face at RDA13)** [April]
- Discussion on first set of consolidated questions per principle

**Workshop #3** [June]
- Presentation of results
- Discussion

**Issues and comments tracking**

Q1 Q2 Q3 Q4 Q5 Q6

M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17 M18
Any questions about the *methodology*

① Do you agree with the proposed methodology?

② ..
Survey results

Respondents
- Big Data Readiness
- FAIR Metrics
- FAIR evaluator
- Data Stewardship Wizard
- FAIR data assessment tool
- FAIR enough? Checklist to evaluate FAIRness for researchers
- Checklist for evaluation of Dataset Fitness for Use
- Support your Data
- Fairness assessment tools for crediting/rewarding research data sharing activities

Some discussion items derived from the survey
- Scope of the assessment
  - What does the tool assess? [e.g. DMP, dataset, way of conducting research, anything]
  - Cross-domain or domain-specific?
- Audience [e.g. researcher, repository manager, data librarian, data steward]
- Automation of the assessment [i.e. what proportion to automate and how]
- Certification [e.g. quality label, scoring system]
- Maintenance and governance [e.g. GitHub]
- Guidance [e.g. checklist]
Relevant initiatives

» Presentation of existing approaches
  » DANS FAIR data assessment tool, FAIR checklist
    » Eliane Fankhauser, DANS
  » FAIR Metrics
    » Luiz Olavo Bonino, GO-FAIR
  » Data Stewardship Wizard
    » Rob Hooft, DTL
  » RDA SHARC IG
    » Laurence Mabile & Romain David, University of Toulouse
  » Dataset Fitness for Use
    » Jonathan Petters, Virginia Tech
  » ARDC FAIR self-assessment tool
    » Keith Russell, ARDC

» Summary of lessons learnt and open issues
  » Makx Dekkers, editor team
Contributing to FAIR policy and practice in the EOSC: The FAIRsFAIR Project

Eliane Fankhauser

RDA FAIR Data Maturity Model WG,
First virtual meeting, 21/22 February 2019
Overall aim

- Development and concrete realisation of an overall knowledge infrastructure based on the FAIR data principles on academic quality
  - data management
  - procedures
  - standards
  - metrics ...

- Delivering FAIR aspects of essential Rules of Participation (RoP) and regulatory compliance for participation in the EOSC

- Contribute to a FAIR infrastructure of the EOSC

- Implementation of recommendations from the EOSC HLEG and the Expert Group on FAIR Data.

FAIRsFAIR in a nutshell

- Budget: 10 million euro
- Time plan: 36 months
- Start: March 1 2019
- 22 partners from 8 MS
- 6 core partners
FAIRsFAIR work and the FAIR Data Maturity Model WG

- Technical implementation of FAIR principles: review of commonalities and gaps regarding semantic interoperability, use of metadata and PIDs [WP2, T2.1]
- Mapping emerging data assessment mechanisms with the FAIR principles to develop pragmatic concepts for FAIRness evaluations at dataset level [WP4, T4.5]
- Badging scheme for assessing the compliance of data resources with the FAIR principles [WP4, T4.5]
- Further development of FAIR data assessment tools including the FAIRdat tool [WP4, T4.5]

WP2 (CSC)
- FAIR Practices: Semantics, Interoperability and Services

WP3 (DCC)
- FAIR Data Policy and Practice

WP 4 (DANS)
- FAIR Certification

WP6 (STFC)
- Competence Centre

WP7 (EUA)
- FAIR Data Science and Professionalisation
Main challenges

• Being coherent within the project (collaboartion accross WPs)
• Serving an EOSC Governance structure under development
• Creating synergies with all FAIR related projects, initiatives and activities in Europe and beyond
FAIRNESS ASSESSMENT CHALLENGES

RDA FAIR Maturity Model - February 20-21 2019
FAIR PRINCIPLES
FAIR PRINCIPLES

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 it describes;
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;

https://www.nature.com/articles/sdata201618
FAIR DATA PRINCIPLES - METADATA

Findable:
F1. metadata 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 it describes;
F4. metadata are registered or indexed in a searchable resource;

Interoperable:
I1. metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. metadata use vocabularies that follow FAIR principles;
I3. metadata include qualified references to other metadata;

Accessible:
A1. metadata 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;

Reusable:
R1. metadata are richly described with a plurality of accurate and relevant attributes;
    R1.1. metadata are released with a clear and accessible data usage license;
    R1.2. metadata are associated with detailed provenance;
    R1.3. metadata meet domain-relevant community standards;

https://www.nature.com/articles/sdata201618
FAIR DATA PRINCIPLES – DATA/DIGITAL RESOURCES

Findable:
F1. 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 it describes;
F4. metadata are registered or indexed in a searchable resource;

Accessible:
A1. metadata 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. metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. metadata use vocabularies that follow FAIR principles;
I3. metadata include qualified references to other (meta)data;

Reusable:
R1. metadata are richly described with a plurality of accurate and relevant attributes;
   R1.1. metadata are released with a clear and accessible data usage license;
   R1.2. metadata are associated with detailed provenance;
   R1.3. metadata meet domain-relevant community standards;

https://www.nature.com/articles/sdata201618
FAIR DATA PRINCIPLES – SUPPORTING ELEMENTS

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 it describes;
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;

https://www.nature.com/articles/sdata201618
FAIRNESS ASSESSMENT CHALLENGES
WHY TO ASSESS?

- Because everybody is talking about FAIR and my resources should be seen as FAIR, whatever this means?

- To satisfy funders requirements?

- To serve as a guideline for achieving higher levels of interoperability and reuse with clarity on the concrete benefits (help improve)?
WHAT TO ASSESS?

- Metadata and data?

- Only metadata?

- Only data?
  - What do you mean by data?
  - In the FAIR principles, data refers to a variety of different resources, e.g., “traditional” data, services, software, APIs, vocabularies, ontologies, articles, etc.
HOW TO ASSESS?

- **Manual**
  - Takes advantage of human understandable artifacts, which are currently prevalent
  - May lead to subjective assessments and, therefore, harder to compare resources
  - Harder to scale
  - Harder to evaluate FAIR for machines, which is the main goal of the FAIR principles

- **Automatic**
  - Requires more rigor on the assessed resources
  - More likely to produce objective assessments
  - Easier to scale
  - Able to check if machines can, in fact, “work” with the (meta)data
HOW TO “READ” THE ASSESSMENTS?

- Need for a scoring system
  - One score for as 4 aspects of FAIR? Does not seem useful.
  - One score per aspect (F, A, I and R)?
  - One score per principle? What about the sub-principles?
  - Is there a hierarchy among the principles? Is there an order of precedence? Or different weights?
  - Is there an acceptable minimal FAIR level? Should it be across domains and applications or domain/community-dependent?
  - Do we use a pass/fail approach or introduce intermediary compliance levels in each/some evaluation?

- Need for a visual representation of the scores
  - To facilitate quick perception of the FAIRness level, a visual representation of the FAIR scores is required, e.g., stars, bars, etc.
GENERAL CHALLENGES

- Clarify that nobody has been asked to be 100% FAIR. Many times a lower FAIRness level is perfectly adequate.
- How to deal with the conflicting forces that, from one side want to push the communities towards a better (and FAIRer) data landscape and, from the other side, want to preserve the status quo (existing “kingdoms”) but labeling themselves FAIR?
- Who will define the assessment criteria?
- Who will execute the assessments based on the defined criteria?
- Should we have a unique set of assessment criteria? Or a core set for general comparison and domain-specific sets on top of the core for the specific needs of a given domain/application?
Moving from metrics to maturity indicators

- The Maturity Indicator tests are also going to be "incremental". e.g. for the new indicators there are "weak" and "strong" forms... with loose interpretation of "knowledge representation language” (e.g., CSV) vs strong interpretation (i.e. RDF)

- Full set of fully automatic evaluators almost complete

- Clear separation between the evaluation of metadata and data

- Used (together with the Data Stewardship Wizard) in the “FAIR Funders Pilot”, involving Dutch ZonMW and Irish Health Research Board
Luiz Bonino
International Technology Coordinator – GO FAIR
Associate Professor BioSemantics – LUMC

E-mail: luiz.bonino@go-fair.org
Skype: luizolavobonino
Web: www.go-fair.org
DATA STEWARDSHIP WIZARD
ROB HOOFT / ROBERT PERGL

21 and 22 February 2019
MOTIVATION FOR DATA STEWARDSHIP WIZARD

- Software tool for Smart Data Management Plans for FAIR Open Science
- Help researcher with Data Management
  - Smart questionnaire system
  - Expert system
  - Not: fulfil requirements
- Target audience:
  - Researcher (awareness of options, pointers)
  - Data Steward (checklist)
  - Data Expert (being found)
  - Funder (evaluate DMP)

https://ds-wizard.org/
APPROACH

- Hierarchical
- Based on mind-map
- Relevant questions in context
- No attempt to limit it
- Links to supporting materials and experts
- Localizable: add your local experts and intranet resources
**METRICS FOR FAIR**

- No dedicated questions to probe the FAIR metrics
- *Instead*: Every question helps to measure
- Fully based on answers in the questionnaire

**Data design and planning**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Findability</td>
<td>0.00</td>
</tr>
<tr>
<td>Accessibility</td>
<td>1.00</td>
</tr>
<tr>
<td>Interoperability</td>
<td>1.00</td>
</tr>
<tr>
<td>Reusability</td>
<td>0.88</td>
</tr>
<tr>
<td>Good DMP Practice</td>
<td>1.00</td>
</tr>
<tr>
<td>Openness</td>
<td>0.00</td>
</tr>
</tbody>
</table>
DMP

- No dedicated questions to fill a DMP template
- Instead: template engine (under development) uses the answers in the wizard to write a required DMP
- Fully based on answers in the questionnaire
EXPERIENCE / LESSONS LEARNED / OPEN ISSUES

- Worked on components (mind map) since 2013
- Very successful NL/CZ collaboration with clear task split
- Currently advertised as “source of inspiration for making a DMP” for researcher writing a proposal
- Many ideas on how the wizard can be further improved
- Interest in approach from many academic organisations
- Installation/collaboration at companies (e.g. DSM)
- Broad interest but adoption of new approach takes time
- Concrete plans with ZonMw funder (+HRB, Science Europe)
- Acquisition of funding for further development
What for? to foster data sharing by improving recognition of the work required
How? by providing a set of recommendations to guide researchers and other relevant stakeholders (research institutions administrators, funders, policy makers and publishers/editors) in moving through the necessary steps towards crediting and rewarding in the data/resources-sharing process (in progress); and to encourage the adoption of data sharing activities-related criteria in the research evaluation process at the institutional, national and European/international levels.

As part of it, 3 human readable assessment tools are under development that will assess semi-quantitatively the fairness knowledge & practices of scientists:

1.1 extensive FAIRness external assessment grid
   52 criteria so far
1.2 simplified FAIRness external assessment grid (can be used as a quick self-assessment grid)
   18 essential criteria
   https://zenodo.org/record/2551500#.XGK4llxKg2w
2.2 extensive checklist for fairness self-assessment (adapted from the 2 previous grids)
Objectives: credit & reward for FAIRness in researchers sharing behaviors

-> necessity to improve FAIRness *(understandable and step by step processes)*

Main properties:

- As simple as possible (understandable by non IT people)
- Easy to complete (due to FAIR skills availability in evaluation processes)
- Based on informations given by researchers in careers doc / activity reports
- CC author license (can be reused by anyone at the end of the implementation)

Assessment process: leading to recommendations to improve fairness

- Designed as a decision tree in each FAIR Principle
- 3 Level of criterion importance: essential / recommended / desirable
- 4 possible answers/criteria:
  - Never/NA  If Mandatory  Sometimes  Always
- Evaluation based on scoring each answer for each F.A.I.R. principle
  *ex: Findable 2/8 Never/NA; 3/8 If Mandatory; 1/8 Sometimes; 2/8 Always*
- Recommendations based on this scoring
Lessons learnt from the first tests:

- Essential criteria not always understandable without training
- Implementation of some criteria can be time consuming / need technical advisor / operator

Possible open issues:

- Develop iterative assessment of the researcher FAIRness Literacy
- Help identify needs to build FAIRness guidelines for a better researcher sharing capacity (based on rewards and credits / How to do and step by step tools)

Next steps:

- Upcoming SHARC-survey launch to evaluate the external assessment extensive grid usability: please participate!
- RDA P13 Sharc’s session: please attend!
- Tools experimentation in specific networks (IMI FAIRplus; BiodiFAIRse; Citizen science networks...)

Fairness assessment grids
WDS/RDA Assessment of Data Fitness for Use WG

Goals:
• Specify criteria of dataset reusability expanding on FAIR principles
• Develop process by which a repository/data provider could assess their holdings for reusability

Outputs:
• Criteria for fitness for use, compared against CoreTrustSeal requirements and FAIR principles (spreadsheet)
• Checklist for evaluation of dataset for fitness for use (form) (pdf)
  • designed as a CoreTrustSeal certification add-on
Lessons learned/open issues

• CoreTrustSeal certification goes a ways towards providing for data reuse (covers F and A, less so I and R
• Our practical assessment approach has caveats (also see notes)
  • Manual approach; hard to automate checks for metadata completeness and data correctness that require domain expertise to evaluate
  • Domain expertise of evaluator matters in assessment
  • Neglects data user perspective
• Many domains have not established metadata standards towards reusability
FAIR Self Assessment Tool
Choices

- For research support staff
- Kept close to the principles
- Assessing a data set
- Hard to provide a score across disciplines
- Just a bar rather than a score
- With guidance included
Uptake/Feedback

Used in workshops in Australia by institutions

Used in paper form

Used for developers in Agriculture and BioSciences projects

International interest

‘We want a score’
Links

The Self Assessment tool

Survey on ‘How well does your repository enable FAIR?’
https://www.slideshare.net/kgrussell/how-well-does-your-repository-support-fair-poll-results

Training resources categorised by FAIR

Training for tool designers to enable FAIR

10 FAIR data things
https://librarycarpentry.org/Top-10-FAIR/
Summary of open issues

- Scope of the assessment
  - Data versus metadata, DMP, data sharing activities
  - General versus domain-specific

- Standards maturity

- Responsibilities
  - Criteria definition
  - Measurement execution

- FAIRness literacy

- Manual vs automated

- Scoring / Levels

- Certification
Any questions about the *lessons learnt and open issues* presented?

① Which open issues could be considered in this exercise?

② ...
Results of preliminary analysis - 1

- Landscaping exercise as a *starting point*
- Analysis of existing approaches
  - Publicly available documentation and the [survey](www.rd-alliance.org)
  - Clustering questions and options
    - FAIR facets [e.g. F1, A2] per principle
    - Beyond the FAIR principles [e.g. data storage]
  - Identification of potential overlaps
- WG to compare questions and derive common aspects
So far, **11** approaches are on the radar

**Approaches considered**
- ANDS-NECTAR-RDS-FAIR data assessment tool
- DANS-Fairdat
- DANS-FAIR enough?
- The CSIRO 5-star Data Rating Tool
- FAIR Metrics questionnaire
- Checklist for Evaluation of Dataset Fitness for Use
- RDA-SHARC Evaluation
- FAIR evaluator

**Approach partially considered***
- Data Stewardship Wizard

**Approaches not considered***
- Big Data Readiness
- Support Your data: A Research Data Management Guide for Researchers

*Methodologies analysed but partially/not included in the results because of questions that could not be classified
Early observations

- On average, six questions per facet
  - Overlaps and different terminologies used
  - Some facets are underused [e.g. A1, A1.1, A1.2, A2]
  - Some facets are overused [e.g. F1, F2]

- Different options
  - YES/NO
  - TRUE/FALSE
  - URL
  - Multiple choice
  - Free text

- Different scoring mechanisms
  - Stars
  - Grade
  - Loading bar
  - None

123 questions
5 types of option
4 scoring approaches
Results of preliminary analysis - 4

Five slide decks classifying questions

- **FAIR – Findable** [Link]
- **FAIR – Accessible** [Link]
- **FAIR – Interoperable** [Link]
- **FAIR – Reusable** [Link]
- Beyond the FAIR principles (X) [Link]

Questions, options and potential overlaps

- **A2** metadata is accessible, even when the data are no longer available
  - 1. Will the metadata record be available even if the data is no longer available?
    - No
    - Unsure
    - Yes
  - 2. Are the metadata accessible?
    - No
    - Yes
  - 5. Please provide the URL to a metadata longevity plan
  - 7. The existence of metadata even in the absence/removal of data
Results of preliminary analysis - 5

Beyond the FAIR principles

- Characteristics of projects, workflows and tools
- Open vs. closed/embargoed data
- Curation, maintenance and governance
- Certification (what and who/how)
- Others?

Should the WG consider these additional aspects as one or more separate strands?
Any opinions about the *additional aspects* to be considered?

<table>
<thead>
<tr>
<th>①</th>
<th>Which other aspects should the WG consider?</th>
</tr>
</thead>
<tbody>
<tr>
<td>②</td>
<td>...</td>
</tr>
</tbody>
</table>
Contribution is sought and welcomed for

- **METHODOLOGY**
  - E.G.
    - Missing items
    - Alternative approach
    - ...

- **ANALYSIS**
  - E.G.
    - Scope
    - Irrelevant items
    - Missing items
    - Additional aspects
    - ...

- **AOB**
  - ...
How to contribute - 2

Issue tracking on GitHub ([Join GitHub](https://github.com))

Create an issue:
- Provide a clear title and a detailed description
- Label and categorize the issue [e.g. Methodology, Principle_F]
Logistics

➤ 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 – Mailing list
fair_maturity@rda-groups.org
Conclusion

Action items

- Feedback via GitHub
  - Work methodology
  - Work process
  - Tentative timeline
  - Results of preliminary analysis

Next steps

- *Issues and comments* review period
- RDA 13th Plenary Session [Philadelphia, USA]
- Online workshop #3