IMI2 Project 802750 - FAIRplus

FAIRification of IMI and EFPIA data

WP2 – Standards Definition and Process Development

D2.5 FAIRplus FAIR Data Maturity Framework

Lead contributor	Oya Beyan (02 – Fraunhofer)
	beyan@fit.fraunhofer.de
Other contributors	Ibrahim Emam (11 – Imperial College) i.emam@imperial.ac.uk
	Philippe Rocca-Serra, Susanna-Assunta Sansone (03 – University of Oxford)
	Nick Juty (04 – UNIMAN)
	Ebtisam Alharbi (04 – University of Manchester)
	Colin Wood (17 – AstraZeneca)
	David Henderson (21 – Bayer)
	Tony Burdett (01 – EMBL-EBI)
	Melissa Konopko (01 – ELIXIR Hub)

Due date	28 Feb 2021
Delivery date	29 Jun 2021

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 802750. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation and EFPIA companies.

Deliverable type	R
Dissemination level	PU

Description of Work	Version	Date
	V1.0	29 Jun 2021

Document History

Version	Date	Description
V0.1	25.01.2021	First Draft
V0.2	15.02.2021	Revised Draft
V0.3	22.02.2021	Comments
V0.4	28.04.2021	Revised Draft
V0.5	30.04.2021	Comments
V0.6	07.06.2021	Revised Draft
V0.7	18.06.2021	Comments
V1.0	29.06.2021	Final version

Table of Contents

1. Executive Summary

The goal of the FAIRplus Maturity Framework is to help the IMI office, IMI projects, and EFPIA partners analyse and plan their advancement in creation and maintenance of FAIR data. To this end, the work package aims to deliver two maturity models. One is a Data Management Maturity Model which focuses on ensuring that processes in place support the creation of new FAIR datasets, the second is the Dataset Maturity Model which is regarding the FAIRness of existing datasets. These two maturity models fit into the updated FAIRplus FAIRification process; the Dataset Maturity Model supports the data requirements task while the Data Management Maturity Model supports the task to identify FAIRification capabilities and resources.

These models will be aligned with and based on the RDA FAIR indicators. However, before including them, an analysis was done on their suitability for use with human data as well as their overall clarity. This led to the creation of FAIR Assessment Indicators on which the Maturity Models will be built.

2. Background

The FAIR data guiding principles have succeeded in setting expectations for what characteristics FAIR data should exhibit. However, the *path* as well as the *destination* to deliver FAIR data in a research environment are still subject to individual interpretation and project-specific needs.

The FAIRplus project consortium established the FAIRplus Capability Maturity Integration (CMMI) team, now the FAIR Data Management Maturity (FAIR-DMM) team as a cross-work-package task force. This team assimilates the learnings from the emerging WP2 (metadata and standards) and WP3 (FAIR implementation) iterative FAIRification processes to help define a systematic and graduated approach towards achieving higher levels of FAIRness.

The deliverables of the FAIR-DMM team are planned in two phases. Phase one: the 'learning and examination' phase, and Phase two: 'development and evaluation' phase. In this report we present the outcomes and learnings of phase one and based on these we also present the roadmap for phase two, which is to deliver the FAIRplus Data Maturity Framework.

The 'FAIRplus Data Maturity Framework' is a FAIRplus cross-work-package deliverable that aims to offer a guide and a reference model for building FAIR data management processes (FAIRification), as well as a model for assessing FAIR datasets maturity. This framework will act as a guide to advance the cause of FAIR data within the IMI office, IMI projects, and EFPIA partners by showing the benefits of stepwise FAIR advancement and how to reach these milestones.

3. Results

3.1 FAIRplus Data Management and Dataset Maturity models

The FAIR-DMM team was originally tasked with developing a maturity model following the well established general model from the CMMI institute[1]. The aims of the FAIRplus maturity model were twofold: first to help evaluate and improve the processes through which FAIR data is produced; secondly to help evaluate and improve the FAIRness level of a given dataset. The first is a model to evaluate and measure *process* improvement, while the second is a model to assess and measure *product* improvement.

The FAIRplus DMM team adapted the CMMI model to suit the context of FAIR data management in the life sciences domain and to fulfil the expected aims. One of the first outcomes of the FAIRplus DMM team was to define and distinguish between two separate yet related deliverables.

1. **The FAIRplus Data Management Maturity (FAIRplus-DMM) model:** a process maturity model that guides an organization's or project's data management activities for the delivery of FAIR data by-design. The levels within this maturity model contain the underlying data management capabilities required to enable the creation of FAIR datasets at different levels of maturity and to guide an organization's or project's work to develop and improve their own FAIR data management capabilities.

2. **The FAIRplus Dataset Maturity (FAIRplus-DSM) model**: A life-sciences domain-specific, indicator-based *dataset maturity model* to serve as both an assessment and a maturation guide towards FAIR maturity of a dataset. The levels within this Maturity Model contain the metadata required to achieve a certain level of Dataset FAIRness along with the FAIR benefits achieved at each level.

3.2 Revisiting the FAIRplus FAIRification Process

The development, implementation and execution of the *FAIRification process* in FAIRplus is carried out by a cross-work package team of WP2 and WP3 members called the Squad team. A major deliverable of FAIRplus, this process is used to define the workflow that FAIRplus adopts in FAIR*ifying* research datasets from IMI projects and EFPIA partners.

[1] https://cmmiinstitute.com/cmmi

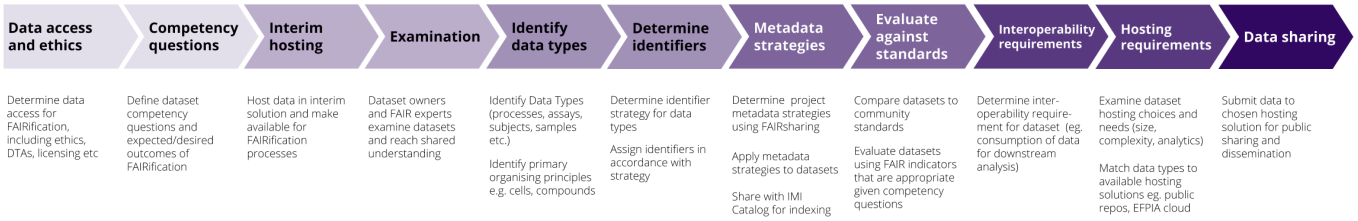

Figure 1 Linear version of FAIRplus FAIRification process

The FAIRplus FAIRification process is a major deliverable of FAIRplus. In this report we only focus on the modifications applied to the FAIRplus FAIRification process in relation to the development of the FAIRplus-DMM and FAIRplus-DSM models.

Figure 1 illustrates an early version of the FAIRplus FAIRification process. This version presented an end-to-end workflow focusing on the tasks that are involved in the FAIRification of a dataset. As the concepts of maturing datasets and maturing processes were still developing, this version lacked an iterative component that reflects the cyclic nature of improving a dataset's FAIRness levels as well as improving levels of FAIRification capabilities. Based on this observation, a revised version of the FAIRification process was developed (Figure 2). In this version, a cyclic phase was introduced to the process allowing maturity assessment to take place before and after the execution of each cycle of a FAIRification task. Furthermore, tasks in the linear version were categorized into two categories: 1) Fulfilling data-related requirements to align with the FAIRplus-DSM model and 2) Establishing FAIR management-related capabilities to align with the FAIRplus-DMM model. We discuss these two categories below.

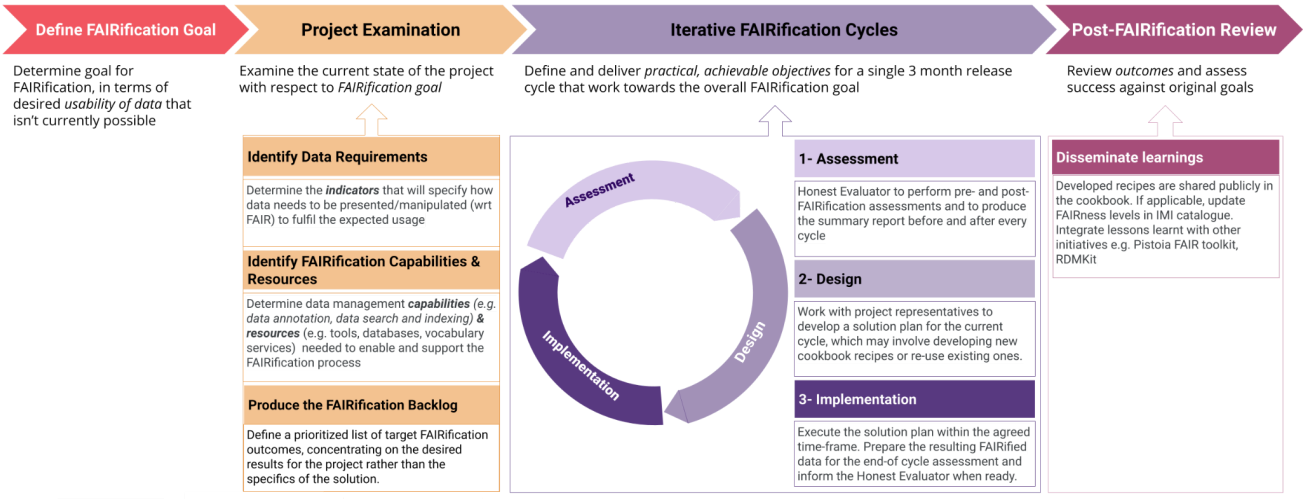

Figure 2 Revised cyclic version of FAIRplus FAIRification process

Requirements related to the characterization of data such as data types, identifiers, interoperability, metadata and data standards were categorized as 'Data Requirements' tasks. These tasks are expected to have varying levels of complexity depending on the FAIRness level targeted for the dataset. In the revised version, identifying the characteristics that a FAIR dataset should exhibit has been explicitly added as part of the 'Project Examination' phase of the FAIRification process. This phase, therefore, becomes the target phase to employ the FAIRplus Dataset Maturity (FAIRplus-DSM) model to assess the level of FAIRness exhibited by the data and identify, accordingly, the next incremental level and the necessary steps and requirements needed to achieve it.

The second category of tasks are related to the capabilities that a FAIR data management environment (i.e. the FAIR process) should exhibit to enable and support the realization of a FAIR dataset. In the early version of the FAIRplus FAIRification process, these were: Data access, data hosting, ontology services and data sharing amongst others. In the revised version these tasks are categorized as 'FAIR capabilities and resources', and is another part of the `Project Examination' phase. These capabilities are also expected to vary depending on the level of maturity achieved (as well as targeted) for the FAIRification process being developed. Identifying these capabilities is delegated to the FAIRplus Data Management Maturity (FAIRplus-DMM) model, which would classify and order them according to a graduated maturation path.

The revised version of the FAIRification process established an alignment between the development efforts of the FAIR-DMM team and the hands-on practices of the squad team, running and validating the FAIRification process against real-world data. Establishing this alignment also helped the FAIRplus-DMM team to develop a clear roadmap for delivering the components of the FAIRplus Data Maturity Framework[2].

3.3 Developing FAIR Assessment indicators

In this section we highlight early results of the FAIR-DMM team efforts in developing the FAIRplus Dataset Maturity (FAIRplus-DSM) model. The first task was to evaluate already existing FAIR indicators, which are an essential component of the maturity model. The FAIRplus Squad team first evaluated the use of RDA indicators as a basis for measuring the FAIR maturation of datasets. This evaluation identified some gaps, which the team tried to address by developing a set of new indicators presented in section 3.3.2.

3.3.1 Evaluation of RDA FAIR Data Maturity Assessment

The RDA FAIR Data Maturity Model specifications and guidelines[3] define a set of

[2] https://docs.google.com/document/d/1exCJbCm0rgtint9QnoIm7wjeVVD_prg1WJVYMZZKn0A/edit?usp=sharing

[3] https://www.rd-alliance.org/system/files/FAIR%20Data%20Maturity%20Model_%20specification%20and

Indicators and their priorities for evaluating the level of FAIRness of a dataset. The FAIRplus DMM team evaluated the usability and fitness of the RDA Indicators in the life sciences domain with available (IMI) FAIRplus data sets. Compliance with these Indicators is used to generate a score which reflects FAIR 'maturity' for the assessed dataset.

The RDA Indicators were applied against the RESOLUTE, eTOX, and ND4BB datasets and reported the results on GitHub[4]. The RESOLUTE dataset was evaluated in the first phase where two experts collectively discussed each metric and decided on a score. Experts reported that they found some of the metrics difficult to assess, noting assessment outcomes might depend on subjective interpretation of the metric itself. In the second phase, the eTOX and ND4BB datasets were assessed by two or three evaluators working independently. Due to the independent nature of the assessments, it was possible to quantify subjective interpretation: nine out of 54 Indicators were scored differently in the eTOX evaluation, while 6 out of 50 were scored differently in the ND4BB dataset. Afterwards, dedicated Squad sessions were conducted for each dataset to compare scores of the independent evaluators and record feedback regarding the challenges encountered during assessment. Appendix 1 lists the challenges identified.

Besides the aforementioned, indicator-specific challenges, the evaluators noted the following feedback about the wording used in some of the indicators, which led to the ambiguity they witnessed during their assessments:

- Definitions of concepts such as 'metadata', 'automated', 'standardized, 'free and open source protocol', 'persistence' should be provided.
- Evaluators suggested a guideline with some examples.
- Different ways of publishing data (controlled access, openly available, separate metadata) may lead to different interpretations. Examples should be provided.
- FAIRification for a specified purpose has an impact on interpretation (e.g. what is sufficient?)

All of the above guidance was shared back with the RDA 'FAIR Data Maturity Model' group for future improvements, and was also used within FAIRplus to develop a set of new indicators that target the life-science domain, to try to alleviate the ambiguity that is commonly associated with generic (RDA) indicators.

3.3.2 Developing FAIRplus Indicators 1.0

To mitigate the ambiguities in the RDA Indicators, we developed a set of new FAIRplus Indicators[5] to address metadata requirements specific to the life sciences domain.

FAIRplus indicators are derived from, and aligned to, the set of RDA data maturity indicators, generated by community agreement. The FAIRplus indicators were also

%20guidelines_v1.00.pdf

[4] https://fairplus.github.io/fairification-results/

[5] https://fairplus.github.io/fairification-results/2020-10-11-FAIRplus-Indicators-v0.1/

aligned to the ISA framework[6], which is a domain-aware community standard developed for the life-sciences community. The first version of the FAIRplus indicators includes 18 indicators. 12 indicators relate to the 'Study' concept of the ISA model (F+S01 - F+S08d) and 6 indicators relate to the 'Assay' concept of the ISA model, which focuses on the actual measured data. Table 4 lists these indicators with a short description for each. More details including examples and best practices are available from the FAIRplus GitHub repository[7].

Table 4: FAIRplus Indicators version 0.1

ID	Indicator Short Description
F+S01	Study level documentation is available in a human readable format.
F+S02	Data is reported by following community specific minimum information guidelines
F+S03	Metadata provides references about all biological data types reported in the data.
F+S04	Relationships between different datasets in a study are well defined.
F+S05	A versioning policy is applied to uniquely differentiate between a particular form of a dataset and an earlier form or other forms of itself.
F+S06	Data generated in early phases of research data workflow such as primary data is available for sharing.
F+S07	Negative results are shared.
F+S08	The study is described with metadata including context, biological samples and data acquisition, methods for analyzing and processing data, quality control, and restrictions for reuse.
F+S08a	Metadata includes information about the study design, protocols and data collection methods.
F+S08b	Metadata includes explicit references to research resources such as samples, cell lines.
F+S08c	Metadata contains information about data processing methods, data analysis and quality assurance metrics.

[6] https://www.isacommons.org
[7] https://github.com/FAIRplus/CMM/blob/master/docs/FAIR+Indicators.md

F+S08d	Metadata includes information about data ownership, license and reuse constraints for sensitive data.
F+A01	Data is organized and documented in a human understandable way.
F+A02	Data is encoded in a community specific exchange standard.
F+A03	A machine and human readable formal description of the structure of data is available including types and properties.
F+A04	Data is structured following a life sciences domain model, core classes and their semantic relations refer to a common data model.
F+A05	Data is described with terminology standards.
F+A06	Core data classes (important data elements) follow a common master and reference data entity.

Another innovation introduced by the FAIRplus indicators is their alignment with another concept: the "Data Usage Area" (DUA). A DUA is a representation of a FAIRification goal that motivates the FAIRification process and directs the FAIRification efforts towards a data usage scenario; each organization or research community might have different business goals for using the data. These different goals require different FAIR data management investments for improving the FAIRness of data being used. Table 5 lists the four proposed DUA's each with their corresponding goals and relevant indicators. The DUA concept is also used to contextualize the meaning of the FAIRness score, which the RDA's arbitrary scoring mechanism lacks.

Table 5: FAIRplus Data Usage Areas

Data Usage Area & Associated Indicators	**Goal**	**Expected Benefit**
Data Interpretability	Improve reuse of data by another person. - Provide metadata about how data is organized and structured. - Document the content of data files and their relations. - Identify and validate data types and formats	- Reduces time spent for each person examining and understanding existing research data.
F+A02, F+A03, F+A04, F+A05, F+A06		

Data Integration	Improve data consolidation & harmonization. - Annotate data with common vocabularies, ontologies. - Use of common master data and reference data (if available). - Use common data profiles, models, schemas for semantic modelling (if available). - Improve interoperability by mapping terminologies (e.g. via identifier linksets).	- Reduces time for data cleaning and integration. - Increases the likelihood of linking datasets with other sets automatically.
F+A02, F+A03, F+A04, F+A05, F+A06		
Data Repurposing	Improve reuse of data in another context, such as with different research hypotheses. - Document research hypothesis and data inclusion and exclusion criteria - Document reference materials, such as cell lines and microorganisms. - Document different steps of the research lifecycle and their data outputs. - Provide raw data or primary data, not only derived and analyzed data sets. - Provide a variety of research outcomes, such as negative results.	- Reduces the resources spent for generating data for research hypotheses. - Improves repurposing of data as part of a new study.
F+S01, F+S02, F+S04, F+S05, F+S06, F+S07, F+S08a, F+S08b, F+S08c, F+S08d		

Data Reproducibility	Improve repeatability, replicability, reproducibility of research outcomes. - Provide documentation and guidelines for describing research protocols. - Provide provenance of experiment such as measuring tools, locations, conditions, hypothesis, time periods, study design (power analysis, sample sizes) - Identify the key resources such as antibodies, model organisms and software. - Share materials, software, and other tools used for data analysis	- Ensures transparency, gives confidence in understanding study. - Increases the likelihood of attaining results by a different or same research team, using the same or different experiment setups.
F+S02, F+S05, F+S08a, F+S08b, F+S08c		

4. Discussion

4.1 FAIR process vs FAIR dataset maturity

The CMMI institute model, which is the reference maturity model originally adopted by FAIRplus, is a five-level **process improvement** framework. It describes how organizations across multiple disciplines improve their processes in order to consistently deliver high-quality products and services[5]. The FAIR-DMM faced two challenges when trying to adopt it for its own purposes.

First, the underlying concept of the CMMI model assumes a '*business organization*' context, which is not directly portable to FAIR data management in a research setting. Secondly, the model focuses on improving *processes and capabilities* of an organization from 'unmanaged' to 'optimized' and it does not offer an insight into the path to *product-maturity* and the necessary dimensions that drive that maturation path.

Based on these learnings the FAIR-DMM team decided to refine the original FAIRplus maturity model into two separate maturity models as described in the results section. This separation was necessary to distinguish between two user stories that became more obvious after the first few FAIRplus squad team iterations working with different stakeholder groups from IMI project representatives and EFPIA partners. The FAIRplus-DSM model is driven by a user story more commonly brought by IMI Projects: "As a researcher and a data consumer, I want to know what level of maturity my data is at? How do I enhance the value of my dataset to the next level?". On the

other hand, the FAIRplus-DMM model is driven by a user story that is more commonly brought by industry stakeholders such as EFPIA partners: “How FAIR is my organization’s data management process? What capabilities do I need to invest in to generate FAIR data by design?”

This separation enabled the FAIRplus-DMM team members to focus on developing each model on its own making it easier to communicate with each stakeholder group more clearly.

4.2 Using the RDA and FAIRplus Assessment Indicators

Indicators in the FAIR assessment process often have two roles. First, the ***text*** of the indicator itself explains to the evaluator ‘what’ they need to assess in order to decide whether or not the requirement specified by the indicator is satisfied. Secondly, the cumulative result of evaluating the individual indicators is used to provide an ‘overall’ ***score*** of the dataset’s FAIRness level. We refer to these two roles separately in our discussion below.

Results of the FAIR assessments using the RDA indicators first highlighted a problem in interpreting the text specified by some of the indicators. The use of ambiguous terms such as, ‘rich metadata’ (F2-01M), ‘plurality of attributes’ (R1-01M), ‘other data’ (I3-02D) and ‘other metadata’ (I3-01M), led to a subjective interpretation that differed from one evaluator to another. The FAIRplus indicators tried to be less prone to this subjectivity by aligning its language to more concretely defined concepts defined by the ISA framework. However, results of assessments done using these revised indicators highlighted another challenge related to the *scoring approach.*

The FAIRplus indicators followed the same ‘scoring approach’ adopted by the RDA FAIR indicators. This approach uses an arbitrary criteria to calculate a score for each FAIR principle separately. For each principle, a group of indicators are classified into three subgroups: essential, important, and useful. Based on an arbitrary percentage achieved for each group of indicators, a data object is judged to be at a certain maturity level for that particular FAIR principle. For example, level 2 of Findability is achieved when 100% of the ‘essential’ indicators, 50% of the ‘important’ indicators and none of the ‘useful’ indicators are met. The problem with this scoring approach is twofold.

First within each subgroup all indicators have the same weight so similar scores can be achieved with a great variability of indicators satisfied. Accordingly, this scoring approach does not necessarily reflect actual FAIRness improvement of the dataset. Even if improvement of the score is achieved, comparing FAIRness across different datasets cannot be made since satisfying different combinations of indicators representing very different states of FAIRness can still produce the same score. This results in giving a false indication of similarity or equivalence of the level of FAIRness achieved. Solving this problem will motivate future revisions of the FAIRplus indicators and the associated scoring mechanism.

Secondly, the RDA maturity 'level' achieved for each principle does not provide an overall meaning to the result of the assessment. What does it mean for a data object to be at level 1 for Findability, level 0 for Accessibility, Level 2 for Interoperability and Level 1 for Reusability? The FAIRplus indicators offer an improvement in this aspect by aligning the levels of maturation to Data Usage Areas instead of the FAIR principles as shown in Figure 3. Refining these data usage areas and their relevance to different stakeholders will also be considered in future versions of the indicators and the associated FAIRplus-DSM model.

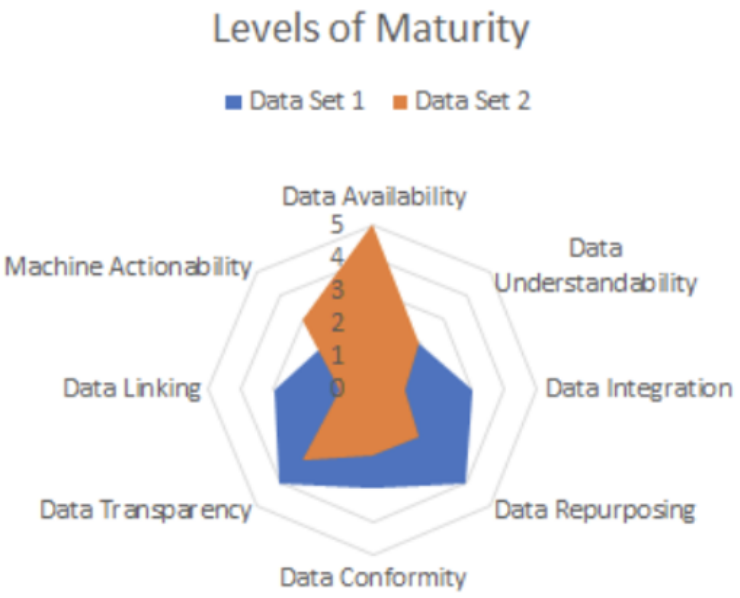

Figure 3 Maturity Levels defined with respect to a Data Usage

More details about the use and evaluation of the RDA indicators and the FAIRplus indicators for Dataset FAIR Assessment are reported in D3.2 IMI FAIR Metrics Publication[8]. In that report the authors present the results of using each set of indicators to assess different project datasets using manual and automated approaches. They also discuss a number of challenges and gaps that remain unresolved and hence invite more improvements to be made to future versions of the FAIRplus-DSM model.

5. CMM Repository

FAIRplus CMM development is documented and made publicly available in the following git repository: https://github.com/FAIRplus/CMM.

[8] https://zenodo.org/record/4428633#.YIhAdn1Kgwu

6. Appendix 1: RDA FAIR Indicator Challenges

RDA Indicator ID	Assessment Feedback
Findability Indicators	
All indicators	There are different levels of metadata. Metadata might refer to data provenance, data protocol, or properties of a dataset. Lack of reference to metadata type creates an ambiguity.
RDA-F1-01M RDA-F1-01D	The concept of persistence is not clear: identifier persistence points to the data set vs the ability of the identifier to persistently identify the same data over time.
RDA-F2-01M	"rich metadata" is not quantifiable: rich for what purpose?
RDA-F2-01M	Metadata should follow a domain specific standard. However in some cases there might be multiple domains.
Accessibility Indicators	
RDA-A1-03M	A separate metadata record is not a common practice in the biomedical domain.
RDA-A1-04M RDA-A1-04D RDA-A1.1-01M RDA-A1.1-01D	Standardized, free, open source protocol definitions were not clear for all experts.
RDA-A1.1-01D RDA-A1.1-01M	Metadata includes information about access conditions, however the definition of the access conditions are vague. What type of information is required is not defined, e.g. who to contact? licence? what if data is openly accessible, should it state explicitly that everyone can use it?
RDA-A1-05D	What automatic means is not clear. If it requires parsing or extraction can it be classified as an automated download?
Interoperability Indicators	
RDA-I1-01M RDA-I1-01D	The scope of knowledge representation expressed in standardized format is not clear, e.g. is it limited with ontologies.
RDA-I2-01M RDA-I2-01D	The scope of standard vocabularies is not clear, e.g. does it include controlled vocabularies defined within the project.
RDA-I1-02M RDA-I1-02D	The scope of self-describing knowledge representation is not clear. Does it only refer to RDF, or are other forms possible?
RDA-I2-01M RDA-I2-01D	FAIR-compliance of applied vocabularies cannot be evaluated during data set assessment.

Reusability Indicators	
RDA-R1.3-01M	How a community standard can be recognized is not well defined.
RDA-R1.2-01M RDA-R1.2-02M	Providing provenance metadata according to a cross-domain language requires a reference.
RDA-R1.3-02M RDA-R1.3-02D	Machine understandable community standard requires examples about what is included, what is excluded, e.g. xml, json, rdf, html
RDA-R1.1-01M	How to interpret the publishing licence requirement for publicly available data is not clear.