Project Title	Fostering FAIR Data Practices in Europe
Project Acronym	FAIRsFAIR
Grant Agreement No	831558
Instrument	H2020-INFRAEOSC-2018-4
Topic	INFRAEOSC-05-2018-2019 Support to the EOSC Governance
Start Date of Project	1st March 2019
Duration of Project	36 months
Project Website	www.fairsfair.eu

D4.1 DRAFT RECOMMENDATIONS ON REQUIREMENTS FOR FAIR DATASETS IN CERTIFIED REPOSITORIES

Work Package	WP4, FAIR Certification
Lead Author (Org)	Anusuriya Devaraju (UniHB), Patricia Herterich (DCC)
Contributing Author(s) (Org)	Eliane Fankhauser (KNAW-DANS), Hervé L'Hours (UKDA), Joy Davidson (DCC), Ilona von Stein (KNAW-DANS), Jerry de Vries (KNAW-DANS), Michael Diepenbroek (UniHB), Kevin Ashley (DCC), Mustapha Mokrane (KNAW-DANS), Frans Huigen (KNAW-DANS)
Due Date	29.02.2020
Date	25.02.2020
Version	1.0
DOI	https://doi.org/10.5281/zenodo.3678715

Versioning and contribution history

Version	Date	Authors	Notes
0.1	31.07.2019	Anusuriya Devaraju	ToC and executive summary
0.2	24.09.2019	Anusuriya Devaraju, Eliane Fankhauser, Jerry de Vries, Ilona von Stein, Hervé L'Hours, Patricia Herterich	ToC updated, use cases and workflows, FAIR data assessment implementation requirements, FAIRsFAIR data assessment metrics (1st version).
0.3	03.12.2019	Anusuriya Devaraju, Eliane Fankhauser, Patricia Herterich, Michael Diepenbroek, Jerry de Vries, Frans Huigen	FAIRsFAIR data assessment metrics (2nd version), mapping between existing FAIR assessment frameworks, and requirements for implementing data FAIRness assessment were refined.
0.4	17.12.2019	Patricia Herterich, Anusuriya Devaraju, Ilona von Stein, Joy Davidson, Kevin Ashley	Changes made to the structure of the deliverable; methodology and steps to derive requirements and metrics were specified, assessment scenarios included.
0.5	27.12.2020	Anusuriya Devaraju, Patricia Herterich	References and appendices updated, the review of the first full draft.
0.6	27.01.2020	Anusuriya Devaraju, Patricia Herterich, Joy Davidson, Hervé L'Hours, Ilona von Stein, Jerry de Vries	Second review of the full draft, including Appendix 1 and 2. Formatted the deliverable for internal review.
1.0	25.02.2020	Simon Hodson, Helena Cousijn, Anusuriya Devaraju, Patricia Herterich, Mustapha Mokrane	Internal review of the report, improved the report based on reviewers' feedback and formatted it to be submitted to EC.

Disclaimer

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

2

Abbreviations and Acronyms

DMP	Data Management Plan
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
FAIR	Findable, Accessible, Interoperable, Reusable
TDRs	Trustworthy Digital Repositories

Executive Summary

The overall goal of FAIRsFAIR is to accelerate the realization of the goals of the European Open Science Cloud (EOSC) by compiling and disseminating all knowledge, expertise, guidelines, implementations, new trajectories, training and education on FAIR matters. FAIRsFAIR work package 4 (WP4) will support the provision of practical solutions for implementing the FAIR principles through the co-development and implementation of certification schemes for trusted data repositories enabling FAIR research data in the EOSC, and the provision of organizational support and outreach activities.

One of the objectives of WP4 is to develop requirements (e.g., metrics) and tools to pilot the FAIR assessment of digital objects, in particular research data objects in trustworthy digital repositories (TDRs). This report presents the first results of work carried out towards achieving the objective. We outline the context for our activities by summarizing related work both performed in other work packages within FAIRsFAIR and approaches from the wider community to address FAIR data assessment. We introduce a range of scenarios for assessing data objects for FAIRness before or after deposit in data repositories and outline two primary use cases that we want to focus on in the project:

- A trustworthy data repository will offer a manual self-assessment tool to educate and raise awareness of researchers on making their data FAIR before depositing the data into the repository, and
- A trustworthy data repository committed to FAIR data provision wants to programmatically assess datasets for their level of FAIRness over time. To facilitate this, FAIRsFAIR will develop an automated assessment for published datasets that will be piloted with some of the repositories selected for in-depth collaboration as part of the FAIRsFAIR open calls.

We present a set of preliminary metrics corresponding to FAIR principles that can be used to assess data objects through manual and automated testing. We discuss the development and key aspects of the metrics, including their initial alignments with the existing CoreTrustSeal requirements. The alignment forms a basis to develop the FAIR elaboration of CoreTrustSeal requirements, which is one of main ongoing activities of WP4. Additionally, we present draft requirements that any FAIR assessment implementation will need to consider and highlight how those requirements will impact the use cases for FAIR assessment that our upcoming work will address. We conclude by outlining the next steps in our work to iteratively improve the requirements through a number of pilots. Our priorities include the refinement of the suggested metrics based on the feedback elicited during pilot testing with several communities, in the context of the use cases developed.

Table of Contents

Table of Figures

1 Introduction

The FAIR principles[1] are high-level guidelines and leave the users to decide on their implementation (Wilkinson et al., 2019; 2016). This leads to a range of sometimes ambivalent or contradictory interpretations, thereby raising the need to define systematic measurements of data FAIRness (Bonaretti and Willighagen, 2019). To support translating FAIR into practice, the 'Turning FAIR into reality' report (European Commission Expert Group on FAIR Data, 2018) published an analysis and recommendations covering FAIR digital objects, the FAIR ecosystem necessary to support these objects, the implementation of FAIR practice, and steps to embed and sustain any culture change. One of the 15 priority recommendations is to develop metrics for FAIR digital objects (recommendation 12). In addition to metrics development, tools implementing the metrics need to be developed and piloted to facilitate the assessment of data FAIRness, both by humans and machines (recommendation 25).

Work Package 4 (WP4) of the FAIRsFAIR project focuses on the evaluation and certification of FAIR digital objects and FAIR-enabling repositories. The objective of Task 4.5 of the work package is to develop and offer practical solutions (i.e., requirements including metrics and toolset[2]) to support FAIRness assessment of individual digital objects from selected European members of the network of FAIR-enabling Trustworthy Digital Repositories (TDRs). In the context of this report, a TDR is a certified data repository (e.g., based on CoreTrustSeal) whose goal is to offer reliable and long-term data access and preservation to its community (Research Libraries Group, 2002). We are concerned with a subset of digital/data objects, which comprise quantitative and qualitative facts that are collected, measured, or created for purposes of scientific analysis. For example, measurements generated by instruments, field observations, and model simulation outputs. Overall, we aim to drive the adoption of FAIR data assessment by developing and piloting the practical solutions with TDRs.

1.1 Task Scope

Figure 1 illustrates the main components of Task 4.5, which are **use cases**, **requirements** (including **metrics**), the assessment **toolset** to be developed, and their relations. According to the (European Commission Expert Group on FAIR Data, 2018), it would be beneficial to define use cases to demonstrate the advantages and encourage communities to engage with a FAIR ecosystem. In agreement with the group, we follow a use-case based approach to scope FAIRness evaluation of data objects in TDRs. We explore several FAIR data assessment scenarios and identify primary use cases which we then categorize into **planned** and **potential** use cases.

- The planned use cases refer to FAIR data assessments applied pre or post-deposit into a TDR. Raising awareness of users on improving the fairness of their data is essential (European Commission Expert Group on FAIR Data, 2018). Therefore, the first planned use case focuses on **researchers** who might **self-assess** the FAIRness of their data manually before depositing them into a repository. This will encourage researchers to increase the FAIRness of their data

---

[1] https://www.force11.org/group/fairgroup/fairprinciples

[2] A collection of tools

and deposit this data in trusted repositories, and that outcome would lead to better production and availability of FAIR data which is an objective of FAIRsFAIR. The second planned use case focuses on the **automated assessment** of data objects deposited and published in selected **data repositories**.

- To promote future integration of FAIR data assessment into the wider repository community, we have also identified a potential use case, which is to apply the assessment of FAIR data objects to supplement the **CoreTrustSeal certification application**. This potential use case will not result in a technical implementation as part of FAIRsFAIR, but its scope and workflows will be documented (deliverable 4.5)[3] at the end of the project based on the outcomes of the planned use cases.

From the use cases, we identify (draft) **requirements** to be considered for implementing FAIR data evaluation. Here, an important contribution is that we develop a set of minimum viable **metrics** based on **existing work** on FAIR assessment. The metrics are intended to systematically measure to what extent data objects are FAIR. The requirements should not be regarded as final but rather as a first prioritized list of features for implementing FAIR assessment. Following the requirements, we will implement an assessment **toolset** (deliverable 4.5). We refer to the toolset as a collection of tools which will be applied by the actors (researchers and data repositories) of the planned use cases to measure the FAIRness of their data objects based on the FAIRsFAIR metrics. We will design a badging scheme to communicate the FAIR measurement results in a visually intuitive way. An essential aspect of our work is that in order to capture different disciplinary perspectives on FAIR assessment, we will test the metrics and the tools through several iterative pilots. The pilots will involve domain researchers and repositories that are currently being operated by project partners (DANS EASY and PANGAEA) and those selected through the FAIRsFAIR Open Calls[4].
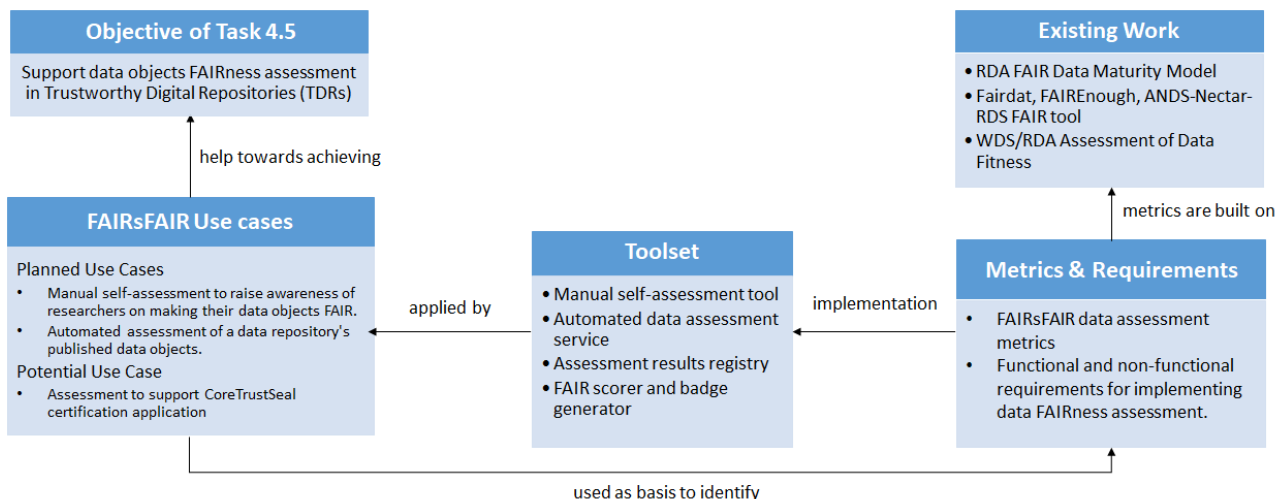

Figure 1. Main components of FAIRsFAIR Task 4.5.

1.2 Report Outline

This report is organized into 8 sections. Section 2 situates our work in relation to other FAIRsFAIR work packages and in the wider context of EOSC. Section 3 describes key concepts and related work on FAIR assessment. An overview of FAIR assessment scenarios is presented in Section 4. This leads to concrete use cases which we will implement during the project (section 5). Next, we cover FAIRsFAIR's objects assessment metrics (Section 6) and the requirements to be considered for implementing objects assessment (Section 7). We provide the mapping between FAIRsFAIR metrics and existing models in Appendix 1, and FAIRsFAIR metrics specification in Appendix 2. Section 8 outlines the next steps. Table 1 mapped the main components of Task 4.5 (as illustrated in Figure 1) with sections of this report.

Table 1. Report outline.

Components	Sections
Objective of Task 4.5	Section 1 (Introduction)
Existing Work	Section 3 (Definitions and Related Work)
FAIRsFAIR Use Cases	Section 5 (FAIRsFAIR Use Cases)
Metrics & Requirements	Section 6 (FAIRsFAIR Data Assessment Metrics) Section 7 (Draft Requirements for Implementing FAIR Data Assessment)
Toolset	To be addressed in deliverable 4.5.

2 Context and Dependencies

The requirements outlined in this report primarily focus on FAIRness evaluations of the individual data objects. However, in line with the Turning FAIR into Reality report (European Commission Expert Group on FAIR Data, 2018), FAIRsFAIR recognizes that in order for data and other research outputs to be FAIR, a broader ecosystem of shared concepts, technologies, services, skills and culture is required.

2.1 Repository Certification

An important role of TDRs in this ecosystem is the provision of long-term stewardship of FAIR data objects, including curation activities to ensure that the data objects remain FAIR (Mokrane and Recker, 2019). Therefore, FAIRsFAIR will augment existing certification mechanisms for digital data repositories. Specifically, WP4 will take an iterative approach to consider which elements of the CoreTrustSeal requirements best support the enabling of FAIR data objects and where additional clarification might be required. An initial set of proposals for repository evaluation including the assessment of maturity towards enabling long-term FAIR objects will be released in February 2020

(milestone 4.1). In February 2020, an evaluation of procedures and processes of certification mechanisms will be provided (milestone 4.5). A FAIR-oriented elaboration of core TDR requirements will provide the input for testing and revision of repository evaluation in the EOSC. An initial version of a repository certification mechanism, that is open to iterations afterwards, will be released in May 2020 (deliverable 4.2).

Since FAIRsFAIR will take an iterative approach to consider which elements of existing repository requirements, such as CoreTrustSeal Requirements (CoreTrustSeal, 2019), best support the enabling of FAIR data, it is expected that our work will be directly relevant to the CoreTrustSeal repository requirements. Accordingly, we have a dependency on the CoreTrustSeal Board to consider incorporating references to the FAIR Principles as well as to FAIR-enabling standards and technologies in the next scheduled review of the CoreTrustSeal requirements.

2.2 Link to Other FAIRsFAIR Activities

Within the FAIRsFAIR project, FAIR assessment is addressed at different levels by different work packages, e.g. by WP2 at infrastructure level for services and software. There is a clear overlap with FAIRsFAIR T2.4 which will propose FAIR recommendations for software and a framework for assessing the FAIRness of services. As found in a recent FAIRsFAIR analysis, funders' data policies increasingly require that data are made FAIR and often specify data deposit in repositories (Davidson et al., n.d.). As funder requirements become more specific and more demanding, self-assessment tools could become valuable for researchers wishing to test their compliance as their projects progress. One can envisage that funding bodies would also benefit from access to automated FAIR data assessment tools to assist with monitoring compliance. The findings of a recent FAIRsFAIR open consultation emphasized that there is a need for support on how to find TDRs and data (Herterich et al., n.d.). Automated assessment tools could become valuable for enabling researchers to determine the potential suitability of data for potential reuse. As the prototype tools and associated workflows are developed, they will be embedded into data stewardship training being delivered through the WP6 FAIR Competence Centre which will provide opportunities for seeking early feedback on the prototypes. In addition, access to prototype tools will be provided via the forthcoming FAIR Competence Centre to provide mechanisms for wider testing and feedback from the community.

We will make available the self-assessment tool to be developed for testing as part of the one-to-one support for repositories being implemented in 2020 via WP2 and WP4[5]. This support will be generalized through WP4's T4.3 Support for FAIR Certification and T3.4 transition support programme for repositories. The former will help repositories to prepare for FAIR-compliant certifiable status and the latter will promote elements of good practice more generally. Forthcoming recommendations on making domain-specific semantic resources FAIR will be valuable references as work on the self-assessment tools progresses. These will be based on the findings of a recent WP2 landscape analysis (Lehväslaiho et al., 2019).

---

[5] https://www.fairsfair.eu/application-results-open-call-data-repositories

2.3 Link to EOSC-relevant Initiatives

The EOSC-FAIR Working Group will develop recommendations[6] for implementing FAIR in practice based on the outcomes of selected projects, including FAIRsFAIR. FAIRsFAIR, OpenAIRE, EOSC-hub, FREYA and RDA Europe jointly organized workshops with different stakeholders to identify recommendations for services enabling FAIR data (Bangert et al., 2019). We will align our work with other criteria that might be set for involvement in the EOSC, the five ESFRI clusters, the thematic and regional projects ('INFRAEOSC 5b') as well as on the five established EOSC Executive Board Working Groups. For WP4 there is a direct connection to the work in the EOSC Working Group on 'FAIR', especially to its sub-teams 'FAIR metrics' and 'FAIR Service Certification'. Through the FAIRsFAIR Synchronization Force, a dialogue among the various projects and actors in the EOSC ecosystem has already been established and will be continued to maximize coordination.

3 Definitions and Related Work

This section briefly describes the FAIR principles and CoreTrustSeal Requirements. It presents related work on FAIR metrics and tools, which lays the foundation to develop FAIRsFAIR assessment metrics (section 6).

3.1 FAIR Principles

FAIR consists of 15 principles to enable digital resources to be Findable, Accessible, Interoperable, and Reusable. The principles are domain-agnostic, high-level guidelines that may be applied to any digital resource (Wilkinson et al., 2016). The implementation of the principles depends on the type of resources (e.g., dataset, software, and workflow) and their user communities. There are challenges in assessing and quantifying FAIRness levels of data objects in a trusted digital repository based on the principles. For example, some of the principles are ambiguous and include subjective terms such as 'rich', 'resource', 'necessary', 'broadly applicable', and 'detailed' which can be difficult to interpret and evaluate objectively (White et al., n.d.). The Interoperability principle focuses on using FAIR vocabularies and knowledge representation to describe digital resources, but the attributes needed of vocabularies and knowledge representation (i.e. formal, accessible, shared, and broadly applicable) require further definition. Numerous efforts have been made to provide implementation guidelines for the FAIR principles. One example is the FORCE11 extended guidelines[7]. Due to different communities' desire to evaluate data FAIRness, several groups have developed their own criteria for FAIRness and implemented them (Clarke et al., 2019), either in the form of manual self-assessment or programmatic assessment. A comprehensive analysis of existing FAIR assessment tools has been produced by the RDA FAIR Data Maturity Model Working Group (Bahim, Dekkers, and Wyns, 2019). To our knowledge, existing tools are currently used for a one-off assessment at a specific point in the research lifecycle. It is still unclear how different stakeholders (e.g., researchers, data stewards and repositories, publishers) representing different communities (domains/disciplines) can adopt and integrate these assessment approaches as part of their research data workflows. FAIRsFAIR

---

[6] https://www.eoscsecretariat.eu/working-groups/fair-working-group
[7] https://www.force11.org/fairprinciples

addresses this challenge through a use-case (researcher and repository) based approach to data FAIRness evaluation, which will be tested iteratively with real data repositories.

3.2 CoreTrustSeal Requirements

CoreTrustSeal is an international, community-driven organization that provides core level certification of data repositories. The certification assessment is based on 16 core requirements[8], which cover different aspects of a TDR including organizational, technical and data object management. CoreTrustSeal requires repositories to be reassessed every three years (CoreTrustSeal, 2019). Data stewardship is addressed by the FAIR data principles with a data-centric perspective whereas the CoreTrustSeal requirements have a repository-centric perspective. The principles can be used to assess if a data repository guarantees the provision of FAIR data objects, whereas the CoreTrustSeal requirements are important to ensure the long-term preservation of the objects. In light of FAIR-aligned repository certification, apart from the digital objects (data and metadata) to be assessed in terms of their FAIRness, business information management (e.g., policies, procedures and workflows) are essential as they provide supporting evidence for enabling FAIR[9].

A mapping between the FAIR data principles and CoreTrustSeal requirements shows that most of the FAIR data principles are either explicitly or implicitly covered by the CoreTrustSeal requirements. However, it cannot be systematically determined to what extent the data curated by a CoreTrustSeal-certified repository comply with the FAIR data principles (Mokrane and Recker, 2019). To better assess the areas of overlap, mappings between the principles and the requirements have been undertaken, for example (Austin et al., 2019; Mokrane and Recker, 2019; White et al., n.d.).

3.3 Related Work

There are many resources (e.g., assessment models and tools) to evaluate data FAIRness. Within FAIRsFAIR, a report on 'FAIRness of services' is currently being developed by WP2 (Task 2.4, milestone 2.7) which also analyses the FAIR evaluation landscape and how existing resources might contribute to a framework to assess the FAIRness of services. This section presents the relevant existing work, and highlights why they were selected and how they relate to our work.

3.3.1 FAIRdat and FAIREnough

In the summer of 2017, Data Archiving and Networked Services (DANS) developed a prototype of a FAIR data assessment tool (FAIRdat)[10]. The tool gives a rating of up to 5 stars for Findability, Accessibility, Interoperability and Reusability of a dataset, as well as its overall FAIRness score. After having received feedback from a broad range of users, it became clear that a step towards a full service version of the FAIRdat tool would require a reformulation of questions, provision of more

---

[8] For extended guidance of CoreTrustSeal, see https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf

[9] Milestone 4.2, Release of first draft of maturity model based on extensions and/or additions to CoreTrustSeal requirements for initial review, including the CoreTrustSeal Board [month 18].
[10] http://blog.ukdataservice.ac.uk/fair-data-assessment-tool/

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

12

guidance, and reconsidering the object level of FAIR assessment (repository, dataset, files). Presently, the FAIRdat tool, even though accessible online, remains a prototype.

In parallel, DANS also created a simple questionnaire ('FAIR enough? Checklist to evaluate FAIRness of data(sets)') for data creators (researchers) to 'roughly' evaluate the FAIRness of their datasets. The goal of the tool is to create awareness about FAIR principles and provide researchers with guidelines to support publishing their research resources following the principles. The questions are kept simple; additionally, short explanations of terms and concepts used in the subjects are provided. The questionnaire was first presented during the EOSC Stakeholder Forum in Vienna in November 2018 and is available online[11]. DANS tested the tools with actual users, and through several workshops[12] The development of both tools led to the formulation of one of the FAIRsFAIR planned use cases (section 5.1), which focuses on the development of a FAIR self-assessment tool predominantly for researchers. The first pilot of this use case will be tested with DANS EASY researchers.

3.3.2 WDS/RDA Assessment of Data Fitness for Use Working Group (WG)

The WDS/RDA Assessment of Data Fitness for Use WG developed a collection of criteria for evaluating research data fitness for use. The collection covers the FAIR principles (e.g., accessibility, findability, and interoperability) including aspects of data quality (e.g., metadata completeness, data completeness, and correctness) and data curation. The criteria are intended to serve as 'add-ons' to the CoreTrustSeal Repository Certification requirements, which may be applied by repository managers or certification reviewers to assess archived research data objects as part of the repository certification application. The main outcomes of the working group are (a) a comparison[13] of the criteria against the CoreTrustSeal requirements and the FAIR principles, and (b) an online questionnaire[14] (checklist) to enable manual data assessment based on the criteria. We referred to the comparative list to align the FAIRsFAIR metrics (Appendix 2) with the CoreTrustSeal requirements. This work also led to the potential use case (section 5.2), which focuses on the application of FAIR data objects assessment to support core certification of data repositories. Given the resources and time constraints of the project, it might not be feasible to implement the potential use case. Therefore, at the end of the project, we will document the possible procedures and workflows on how FAIR data assessment could be practically integrated into the CoreTrustSeal Data Repository certification.

3.3.3 RDA FAIR Maturity Model Working Group (WG)

The FAIR Data Maturity Model WG has developed a set of core assessment criteria[15] (known as indicators) for FAIRness. The first batch of the indicators derived from a landscape analysis of existing FAIR evaluation approaches (Bahim, Dekkers, and Wyns, 2019), which was further improved based

---

[11] https://docs.google.com/forms/d/e/1FAIpQLSf7t1Z9IOBoj5GgWqik8KnhtH3B819Ch6lD5KuAz7yn0I0Opw/viewform

[12] https://dans.knaw.nl/nl/over/diensten/training-consultancy/consultancy/presentaties

[13] https://www.rd-alliance.org/system/files/Fitness%20for%20use_%20compilation%20of%20criteria_0.xlsx

[14] https://www.rd-alliance.org/system/files/DataFitnessForUse_ChecklistForm_v2_20181218_RDADistribution.pdf

[15] https://docs.google.com/spreadsheets/d/1mkjElFrTBPBH0QViODexNur0xNGhJqau0zkL4w8RRAw/

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

13

on the members' feedback. Also, an online survey was conducted to identify the priority levels of the indicators. This work is ongoing and primarily focuses on 'what' to be evaluated, i.e., assessment indicators with their proposed priorities. It does not intend to define 'how' the core criteria could be evaluated. As stated by the working group, 'the exact way to evaluate data based on the core criteria is up to the owners of the evaluation approaches, taking into account the requirements of their community' (Herczog et al., n.d.). Therefore, there is a clear need to implement and test the FAIR criteria with several user communities. We address this need through use cases that target two FAIR stakeholders - researchers and repositories. The metrics we specified are built on the WG assessment indicators. We improved the clarity of the metrics and defined their practical tests (see Appendix 2).

During the 6th online workshop[16], two members of the WG provided their feedback on the indicators to improve the understandability of those indicators and accommodate disciplinary practices when developing practical tests against them. For example, the FAIRplus project (Burdett et al., n.d.), e.g., has worked on FAIRifying several datasets and assessed[17] their datasets against the FAIR Maturity Model Working Group indicators. Currently, the working group is consolidating feedback from various projects, including the feedback provided by FAIRsFAIR partners.

3.4 Technical Tool Candidates

Several tools have been developed to assess data FAIRness based on a set of criteria, either manually or programmatically. For more information about existing tools, see the FAIRassist list[18] and the survey of existing FAIR assessment tools and approaches[19] produced by the RDA FAIR Data Maturity WG. In examining the survey results, it is apparent that the majority of current work focuses on the manual FAIR assessment of data objects or repositories through case studies, checklists, and questionnaires. For example, (Dunning, Smaele, and Böhmer, 2017) investigated 37 data archives manually concerning their compliance with FAIR principles. Based on the study results, the authors concluded that different interpretations of FAIR principles that are compatible with the original principles should be allowed to accommodate disciplinary norms. We agree with the authors in recognizing the importance of taking into account domain practices when defining and implementing data FAIRness assessment. In order to better understand domain requirements, we will iteratively improve the FAIR assessment metrics through several pilots undertaken as part of Task 4.5; the pilots will involve different domain repositories selected through the FAIRsFAIR open calls.

ANDS-Nectar-RDS developed an online manual self-assessment tool[20] for data users, including data stewards, research support staff, and researchers, to evaluate the FAIRness of a data object. Currently, there are limited computational approaches to evaluating data FAIRness. Notable work in this direction has been done by the FAIRmetrics and the FAIRSharing groups, who developed a software framework (FAIR Evaluation Services[21]) to assess data objects programmatically based on

---

[16] https://www.rd-alliance.org/workshop-6
[17] An overview of the results is available at https://fairplus.github.io/fairification-results/
[18] https://fairassist.org
[19] https://docs.google.com/spreadsheets/d/14ojMSXVOITg3RoJn-PuDaPj8zuIGQz2Li-kl97HOBH4/edit#gid=0
[20] https://www.ands-nectar-rds.org.au/fair-tool
[21] https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

14

core maturity indicators (Wilkinson et al., 2019). We will consider reusing these tools to implement data FAIRness assessment in the context of the planned use cases (section 5.1).

In addition to the metrics to assess data FAIRness, one of the expected outcomes of Task 4.5 is the development of digital badges to showcase the FAIR aspects of a data object. Several developments have been carried out to quantify and visualize data FAIRness. FAIRshake[22] is a toolkit that facilitates the manual and automated assessment of different research assets based on existing individuals or a collection of metrics (aka. rubrics) (Clarke et al., 2019). An essential feature of the toolkit is that it allows users to submit a FAIR assessment of a biomedical digital object from the object's homepage to the FAIRshake site and to visualize the assessment through a browser extension or favelet. Bonaretti and Willighagen (2019) developed a semi-automated approach to evaluate datasets retrieved via a user search from two life sciences repositories. They used a plot with graphical elements (color and shape) to summarize and compare the datasets' FAIRness (Bonaretti and Willighagen, 2019). The Global Agricultural Research Data Innovation & Acceleration Network (Gardian) is an online harvester of agricultural data and publications from the Consortium of International Agricultural Research Centers (CGIAR). Gardian developed a guideline to score the FAIR aspects based on the DANS metrics for FAIR compliance (Global Agricultural Research Data Innovation & Acceleration Network, 2019). For each of the harvested records, the system displays its compliance score by means of a digital badge.

3.5 FAIR Stakeholders

As described in the Turning FAIR into Reality report, there are various FAIR stakeholders (Figure 2) in the ecosystem, such as research communities, data stewards, data service providers, publishers, funders and standards bodies (e.g., for repository certification). For a high-level vision of the FAIR ecosystem and its component interactions, see (L'Hours and von Stein, 2019) Not all of the stakeholders are interested in FAIR assessment of objects. Hence, in this deliverable, we will focus on:

- Researchers: we focus on individuals instead of the research communities considered in the Turning FAIR into Reality report. Researchers will be field practitioners that are data creators and users of FAIR data and are likely to be responsible for any practices and workflows to make their datasets FAIR.

- Data stewards: data stewards are support staff supporting researchers with guidance and help on various aspects of FAIR data management including the deposit into repositories and the FAIR assessment of datasets in preparation for deposit. While data stewards can also manage data repositories and have data curator roles, we separate these roles for clarity in this report.

- Data service providers: this stakeholder group covers a variety of services including domain data repositories and institutional repository services as well as providers of other tools and

---

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

15

services, e.g., to support data management planning. For this report, we include data curators here as they play a key role in running data repository workflows.

- Standards bodies: while the FAIR assessment of dataset will build on a variety of standards, this report will focus on standard bodies for repository certification as this is the context in which the work in this work package is carried out.

Figure 2. FAIR stakeholders (figure is derived from 8.3 Stakeholder Groups Assigned Actions (European Commission Expert Group on FAIR Data, 2018)). Dotted lines represent the stakeholders of the FAIRsFAIR use cases on FAIR data objects assessment.

3.6 FAIR Evaluation Mechanisms

Different approaches to FAIRness assessment can be employed, such as self-assessment, automated, task forces, and crowd-sourcing (Wilkinson et al., 2016). A manual self-assessment can be applied to engage and educate users on FAIR metrics, e.g., researchers and data stewards. In principle, assuming standardized inputs, an automated data FAIR assessment is more likely to be consistent and objective. Most importantly, an automated assessment will be scalable and able to evaluate the very large number of datasets from a given data repository.

In FAIRsFAIR, we will implement manual as well as automated FAIR data object assessment tools to support two use cases (researchers and data repositories), respectively. It should be noted that automated assessment is essential but might be difficult to be fully achieved during the active phase of the project due to heterogeneity of standards and requirements of the various science communities, and lack of machine-readable resources (e.g., registries and standards) to support the assessment. Besides, some evaluation metrics necessarily involve expert intervention. As part of the final report of Task 4.5 (deliverable 4.5), we will take into account and document these limitations and identify areas for improvements.

4 Assessment Scenarios

There are a variety of motivations, scenarios and approaches to assessing datasets for their FAIRness. A data object may be assessed by different actors before, during, or after it is curated or published (Figure 3), either manually, programmatically or a combination of these. In this section, we present an overview of various FAIR assessment scenarios that have come out of discussions and work carried out in FAIRsFAIR work packages 3 and 4 so far.

For each scenario, we identify the main stakeholders involved in the scenario using the stakeholders as detailed in section 3.5. Furthermore, we present considerations of the resources that would be needed to implement those assessments and highlight when in the research life cycle the assessment would occur. We also map the scenarios to existing tools that enable planning for FAIR datasets to be created (e.g. as part of data management plans and tools that support creating them) or FAIR assessment of datasets about to be or already published (using the FAIRassist[23] list). This will identify current gaps in FAIR assessment implementations and highlight areas for future work.

There might be additional scenarios that we are not aware off yet and this list will be extended if needed based on feedback from the community e.g., through our work on certification with repositories or calls for use cases (e.g., jointly with WP3).

---

[23]https://fairassist.org
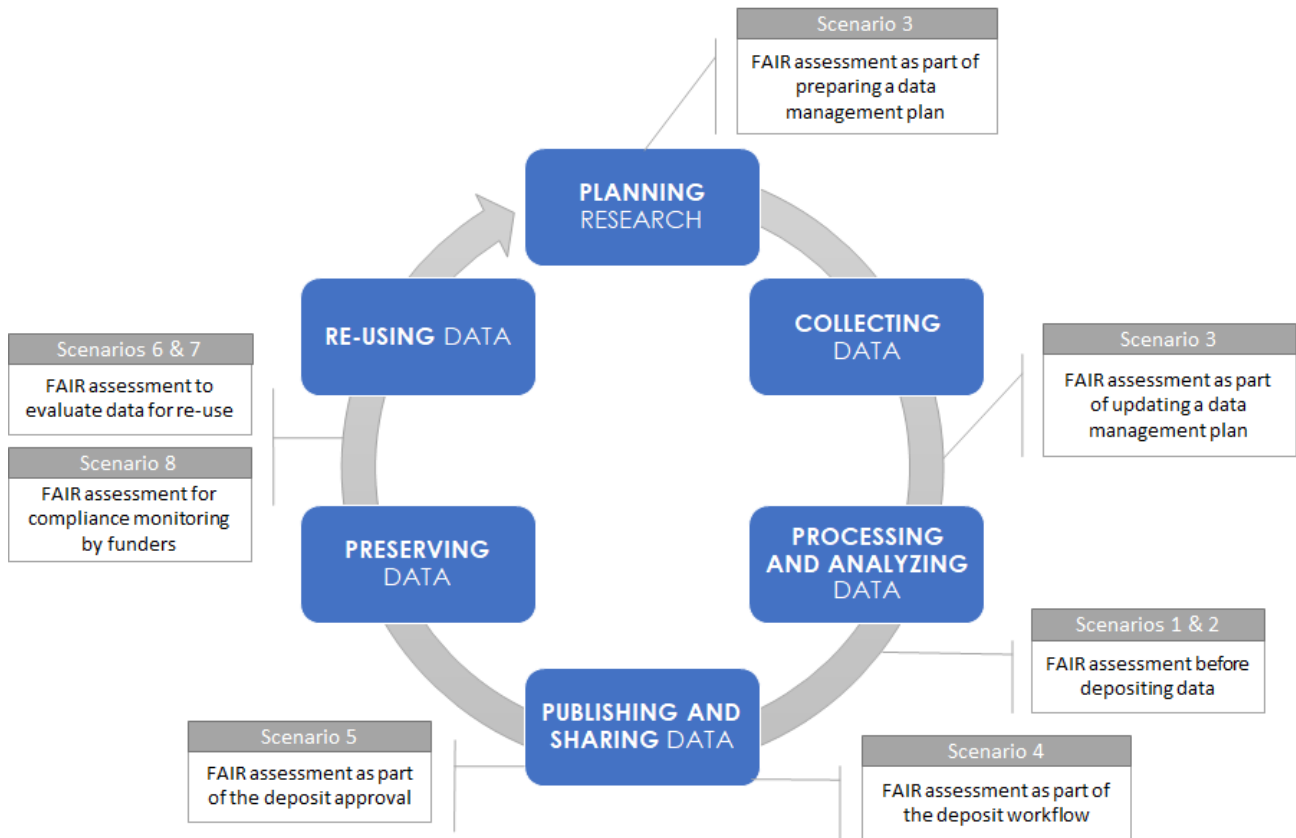
Figure 3. Research data lifecycle (figure adapted from (Mosconi et al., 2019) and scenarios of FAIR assessment of datasets therein.

Scenario 1	As a researcher, I want to check that my data is as FAIR as possible before depositing the data in a repository for wider sharing. To do so, I'm using a self-assessment tool to check certain aspects of the FAIRness of my dataset to gain an understanding where more documentation or metadata might be needed to make my dataset as useful as possible to others.
Stakeholders	Researchers, data stewards as evaluators (peer reviewer) Data repositories, research funders, publishers, or institutions to host the checklist
Implementation considerations	Providing a central checklist does not require much staff effort. However, it should provide sufficient clarity on how to interpret each checklist component, be kept up to date and align with new metrics and developments. The checklist questions should be simple and understandable to respondents. The self-assessment can be a read only document such as a PDF or markdown document for researchers to use or be implemented as an interactive web form.
Life cycle stage	Before deposit of the dataset. After deposit and publication of the dataset, the checklist can also be used to evaluate published datasets for re-use.

Existing implementations	FAIR enough[24]
	ANDS-Nectar-RDS FAIR Data Self-Assessment Tool[25]
	FAIRdat[26]to some degree
	FAIRness self-assessment grids[27]
	5 Star Data Rating tool[28]

Scenario 2	A trustworthy data repository has an interest in receiving FAIR data deposits that make it easier for it to provide access to FAIR data over time. To raise awareness around FAIRness of datasets, the TDR might offer a tool to its users to help them self-assess their datasets before deposit. This checklist is likely to be different from a generic assessment tool (scenario 1) to take local repository policies, requirements, implementations and workflows into account.
Stakeholders	Researchers, data stewards as evaluators
	Data repositories to host the self-assessment tool
Implementation considerations	Tailoring a checklist to specific repository workflows requires some staff resources and the material will need to be kept up to date. It can be provided in addition to the other guidelines and resources that the repositories offer to support its users and can be provided as a downloadable document or an interactive resource.
Life cycle stage	Upon deposit of the dataset
Existing implementations	FAIRdat tailored to EASY repository

Scenario 3	As a researcher, I want to check that my data is as FAIR as possible throughout the project lifecycle to understand if I have to change any of my research data management practices and adjust my data management plan. Ideally, this could be done via an updated data management plan and integrated in DMP tools. Alternatively, there could be checklists used that are more tailored for data that is still actively collected, analyzed and processed and not yet ready for publishing in a repository.
Stakeholders	Researchers, data steward as evaluators
	DMP tool providers

---

[24] https://docs.google.com/forms/d/e/1FAIpQLSf7t1Z9IOBoj5GgWqik8KnhtH3B819Ch6lD5KuAz7yn0I0Opw/viewform

[25] https://www.ands.org.au/working-with-data/fairdata/fair-data-self-assessment-tool

[26] https://www.surveymonkey.com/r/fairdat

[27] https://doi.org/10.5281/zenodo.1995645

[28] https://research.csiro.au/oznome/tools/oznome-5-star-data/

Implementation considerations	Enriching any of the existing DMP tools will require developer resources. Opportunities to do automated assessment could be investigated via machine-actionable DMPs and the RDA working group (Miksa, Walk, and Neish, n.d.). Adjusting checklists to support assessment throughout the research life cycle will likely take more staff resources than tailoring a self-assessment for publication as local, funder and/or disciplinary considerations will need to be taken into account.
Life cycle stage	Throughout the data life cycle as part of preparing or updating a data management plan
Existing implementations	Data Stewardship Wizard[29] as an automated feature DMPonline[30] via guidance integrated into the tool. For future more automated integrations, its API integration can be built on.

Scenario 4	As a researcher, I want to check that my data is as FAIR as possible when depositing the data in a repository. Thus, I want to get an automated FAIR assessment check when uploading data and entering metadata into a repository as well as receive pointers to how to make my data FAIRer at this stage.
Stakeholders	Researchers and data stewards supporting the deposit as evaluators. Data repositories to integrate the FAIR assessment tool into their deposit workflows.
Implementation considerations	Depending on the repository software used, this will take considerable effort to implement and keep up to date.
Life cycle stage	Upon deposit of the dataset
Existing implementations	None yet, but the JISC Open Research Hub (Fripp and Davey, 2019) seems to consider it.

Scenario 5	As a data curator for a data repository, I want to assess deposited data for FAIRness to understand if the dataset can be ingested into the repository as is or if any curation needs to be carried out by me in collaboration with the depositing researcher before ingesting. Results from the assessment can be used to update and improve deposit guidance and supporting researchers with planning for deposit and can help repositories with planning cost and effort of their data curators.
Stakeholders	Data curators as evaluators Data repositories to host the assessment tool

---

[29] https://ds-wizard.org/features.html
[30] https://dmponline.dcc.ac.uk/

Implementation considerations	Depending on the repository software used, this will take considerable effort to implement and keep up to date.
Life cycle stage	Upon approval of the deposited dataset into the repository
Existing implementations	None yet.

Scenario 6	A trustworthy data repository committed to FAIR data provision wants to regularly programmatically assess datasets for their level of FAIRness over time. This will benefit researchers wishing to access FAIR data from the data repository for re-use.
Stakeholders	Data repositories as evaluators and host of the assessment tool Researchers to use the assessment to decide on re-use of datasets
Implementation considerations	Depending on the implementation approach (e.g., web-service, plugin, bookmarklet), this will take considerable effort to implement and keep up to date.
Life cycle stage	After publication of the dataset
Existing implementations	None yet. A data repository may adapt the existing automated FAIR evaluation tools (e.g., FAIR Evaluation Service) to assess its datasets periodically.

Scenario 7	A standards body assessing data repositories for certification is interested to understand how repositories support their users in making their datasets FAIR. Thus, examples of FAIR data assessment results might become part of the certification application and review.
Stakeholders	Standards bodies for repository certification Data repositories aiming for certification
Implementation considerations	This is an enrichment of scenario 6 and should not take much additional development effort.
Life cycle stage	After publication of the dataset
Existing implementations	None yet.

Scenario 8	A funding body is interested to understand how automated assessment tools can support compliance monitoring with our requirement that researchers in receipt of funding make data FAIR.
Stakeholders	Funding bodies, researchers, institutions as evaluators Repositories to host/integrate with assessment
Implementation considerations	FAIR assessment can be run on selected datasets resulting from funding deposited in data repositories. Ideally, this is based on FAIR assessment carried out (programmatically) by repositories that will be tailored to funder policies with little effort. Making this a fully automated process would require additional effort to make funding body policies machine-readable to be fed into emerging automated FAIR assessment tools.
Life cycle stage	After publication of the dataset
Existing implementations	None yet.

5 FAIRsFAIR Use Cases

This section presents two use cases (corresponding to the Scenarios 2 and 6 laid out in section 4) which we will implement during the period of the project, and one potential use case which may be deployed in the future. The use cases are more granular than the scenarios and specify the actual implementation contexts and applications of the FAIR data assessment. Both planned use cases address scenarios where either no implementation exists yet (scenario 6) or the existing implementation is hosted by a FAIRsFAIR partner and will be improved upon and made available for the wider community to reuse.

5.1 Planned Use Cases

The first use case is centered around a **self-assessment tool for researchers** to educate and raise their awareness of FAIR data. The tool[31] is an online questionnaire with a set of closed-ended questions addressing the FAIR metrics. Its first pilot experiment will be tested with the DANS-EASY repository. EASY is an online system which archives and provides access to humanity, social science and other domain datasets. The tool was developed for researchers who would like to deposit their data in EASY and evaluate to what extent their datasets can be regarded as FAIR using the online questionnaire. This tool will be repurposed to cover the project's selected questions addressing the FAIR metrics. Importance will be given to formulating explicit metric questions and answer choices. For each of the questions, we will provide short supporting text to help the respondents to complete the questionnaire easily. While the tool is intended initially for depositors in DANS-EASY repository,

---

[31] The prototype of SATIFYD (Self-Assessment Tool to Improve the FAIRness of your Data) is available at https://satifyd.dans.knaw.nl/

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

22

depending on the available resources, we anticipate that it will be used by other partner repositories and as part of FAIRsFAIR engagement and training activities. To this end, the tool will be made customizable so that questions and supporting text can be adapted to the repository's context and training activities.

The second use case focuses on **trustworthy data repositories** who assess their published data objects **programmatically** based on the proposed FAIR metrics. We will develop an automated evaluation tool. We will first test the development with published objects in PANGAEA, a trustworthy repository for earth and environmental science datasets. We will reuse the development to test data objects from other project partners' repositories as well as the repositories that will participate through the FAIRsFAIR Open Call. It might be possible that datasets from some communities cannot be fully assessed programmatically based on the metrics due to limited resources available to run the assessment, e.g., lack of community-endorsed metadata standards. Therefore, as part of this use case implementation, we will document the limitations and required resources in intermediate (milestone 4.9) and final reports (deliverable 4.5). In addition to the automated data evaluation, we will develop a scoring and badging mechanism to communicate the overall FAIRness levels of data objects evaluated.

5.2 Potential Use Case

This use case addresses future integration of FAIR data assessment into the CoreTrustSeal repository certification landscape, which is to facilitate CoreTrustSeal Data Repository certification. To support this use case, as a first step, we aligned the proposed FAIR metrics to related CoreTrustSeal requirements (Appendix 2). Potentially, a data repository may use examples of FAIR data assessment results (generated through the automated evaluation tool as specified in section 5.1) to support its CoreTrustSeal certification application. The CoreTrustSeal reviewers may refer to the submitted results, or re-run the assessment over random datasets curated by the respective repository. This helps reviewers to gain some insights into the FAIR aspects of datasets when reviewing the repository's application. The implementation of this use case also depends on interfacing the assessment tool with external systems such as the CoreTrustSeal Application Management Tool and CoreTrustSeal procedures and governance.

6 FAIRsFAIR Data Assessment Metrics

This section provides the FAIRsFAIR draft recommendations of a minimum set of metrics to measure the FAIRness of data objects. During the project, we will improve the metrics based on the community feedback that will be solicited through the pilots testing.

6.1 Metrics Development

The metrics were built on 14 mandatory and recommended indicators (version 0.03)[32] proposed by the RDA FAIR Data Maturity Model Working Group, in addition to prior FAIR assessment models (sections 3.3.1, 3.3.2). The development involves three main steps. First, we built a mapping of the metrics used in the above models to observe similarities and differences in their representations (Appendix 1). Through the mapping, we identified common metrics. Second, we refined the common metrics, see the examples below (FAIRsFAIR metrics are prefixed with 'FsF' and their detailed descriptions are included in Appendix 2).

- Rephrasing the metrics and their maturity levels. For instance, the original indicator (F2-01M) contains a vague term ('sufficient metadata') and has 'dichotomous' maturity levels ('NO sufficient metadata to allow discovery is provided', 'Sufficient metadata to allow discovery is provided')[33]. We rephrased and extended the metric to take into account metadata elements required to support data citation and discovery (see 'FsF-F2-01M Descriptive Metadata'). Besides, we included a 'scale'-based maturity level such that users can measure their progress in supplying required metadata.

- Merging overlapping metrics. For instance, we combined the existing indicators 'I3-01D Data includes references to other data' and 'I3-02D Data includes sufficiently qualified references to other data' into one metric, see 'FsF-I3-01M Qualified References to Related Entities'.

- Excluding duplicate metrics. For example, we excluded the indicator 'A1-03D Data identifier resolves to a digital object' as the resolvability of a data identifier is tested as part of the metric 'FsF-F1-02D Persistent Identifier'[34].

- Selecting new metrics that apply to the use cases, but are not covered by RDA FAIR Data Maturity Model Working Group metrics (e.g., see 'FsF-R1-02M Data Content Description').
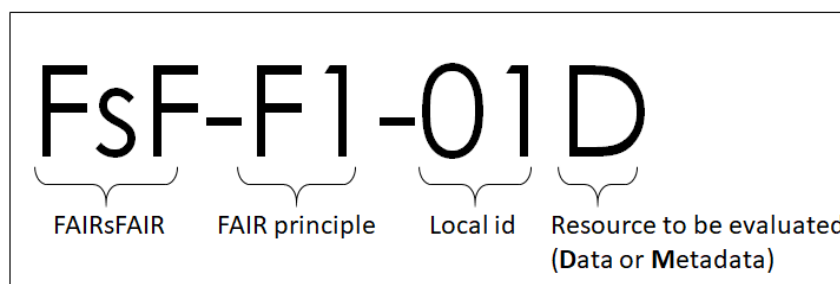

Figure 4. Anatomy of FAIRsFAIR metric identifier.

Each of the FAIRsFAIR metrics is identified following a naming convention. For example, in Figure 4, the identifier starts with the shortened form of the project's name, followed by the related FAIR principle identifier and local identifier. The last part of the identifier refers to the resource that will be evaluated based on the metric, e.g., data or metadata.

The following is a list of 13 FAIRsFAIR data assessment metrics. At present, the metrics cover the FAIR principles, except A1.1, A1.2 (communication protocol), I2 (FAIR vocabularies), and R1.2 (provenance) principles.

- FsF-F1-01D Universally Unique Identifier
- FsF-F1-02D Persistent Identifier
- FsF-F2-01M Descriptive Metadata
- FsF-F3-01M Inclusion of Data Identifier in Metadata
- FsF-F4-01M Searchable Metadata
- FsF-A1-01M Data Access Level
- FsF-A2-01M Metadata Preservation
- FsF-I1-01M Semantic Representation of Metadata
- FsF-I3-01M Qualified References to Related Entities
- FsF-R1-01M Community-Driven Metadata
- FsF-R1-02M Data Content Description
- FsF-R1.1-01M Data Usage Licence
- FsF-R1.3-01D Standard File Format

Finally, we represented the details of the FAIRsFAIR metrics, including their descriptions, evaluation methods, and limitations. We adapted the standard template recommended by the FAIR Metrics Authoring Group to represent the metrics (Table 1, (Wilkinson et al., 2018)). Figure 5 shows an example of a metric 'FsF-F1-02D' which evaluates if a data is assigned with a persistent identifier. The modified template includes Core Trustworthy Data Repositories requirements, assessment procedures (automated and manual), resources, and limitations of metrics implementation. For a detailed characterization of all metrics, see Appendix 2.

25

Field	Description
Metric Identifier	FsF-F1-02D **(The identifier of the metric)**
Metric Name	Persistent identifier **(The short name of the metric)**
Metric Description	The data is assigned with a persistent identifier to ensure the resolvability of the identifier in the long term. The identifier may be resolved to its digital object (e.g., a data file or a web service that returns the data), or to a data proxy (e.g., an online page that contains metadata, including the link to access the data). **(Metric definition and related examples)**

To which FAIR principle(s) does it apply?	F1	F2	F3	F4	A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
	X														

(The FAIR principle addressed by the metric)

To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R1
													X			

(The CoreTrustSeal requirements related to the metric)

For which digital resource is this relevant?	Data		Metadata
	X		

(The type of resource to be evaluated)

Manual Assessment (User Question) **(The metric question that will be addressed to users as part of the self-assessment tool)**	Question Type
Does the data have a persistent identifier assigned? • Yes • No	Single-choice question

Automated Assessment **(The details on the automated assessment of the metric, including inputs, methods and outputs)**

Input	Assessment Method	Output
• Data URI	The identifier is based on a persistent identifier scheme, and it resolves to a web address where the data can be accessed. Verify the identifier resolvability through the HTTP response status codes.	• Yes • No A successful request (Yes) returns the following additional results: • The persistent identifier scheme • The resolved URL A failed request (No) returns an error message.

Comments

Resources **(A list of related resources, constraints and limitations of the assessment)**
- A wiki entry on persistent identifier, https://en.wikipedia.org/wiki/Persistent_identifier

Known Limitations/Constraints
- The automated assessment verifies the resolvability of the specified identifier, but does not assert the type of the resolved object, e.g., whether the resolved object is landing page, a file or a web service query response.

Figure 5. An example of FAIRsFAIR metric representation.

6.2 Key Aspects of Metrics Developed

We regard the metrics as a systematic way to communicate to users how different facets of FAIRness of an object can be evaluated, and to allow users to compare data objects under the same assessment framework. We do not intend to classify a data object as either FAIR or not-FAIR by evaluating it based on the metrics, but rather to support the assessment of the degree/level of FAIRness. The following summarizes the key aspects of the metrics developed:

- **Metric scope and definition**: The metrics in their current form, cover most of the FAIR principles except A1.1, A1.2 (communication protocol), I2 (FAIR vocabularies), and R1.2 (provenance) principles. We assume a standardized and secured communication protocol is part of the requirements of the technical infrastructure of a data repository. Therefore, we defined a metric which evaluates the inclusion of data access level and conditions in the metadata. The criteria of a FAIR vocabulary are currently being developed by WP2 (deliverable 2.2), which we will consider when defining a metric that addresses I2 principle. Data provenance represents lineage information of a data object. For example, data operation, transformation, reduction, harmonization as well as for the scientific provenance information which is needed to understand the data (i.e. collection method and calibration). As data provenance depends on the disciplinary needs, we will involve domain researchers and data repositories when developing the metric. In Appendix 2, we supplied each metric with answer choices to support the objective FAIR assessment of a data object.

- **Digital resource:** Eight out of the thirteen metrics are centered on the evaluation of FAIRness based on the metadata of a data object. One metric ('FsF-R1-02M Data Content Description') involves the evaluation of both data and metadata of an object based on a quality aspect. This metric evaluates if the content descriptions (e.g., variables measured, MIME type, size) represented in the metadata and the representation conforms with the actual data.

- **Assessment mechanism:** We describe manual and automated evaluation procedures of the metrics. It should be noted that the automated assessment methods are preliminary and subject to further changes. Some of the metrics may not be 'completely' evaluated through automated testing, in comparison to manual testing. This is due to limited machine-readable resources available to run the evaluation, and some metrics require domain expertise (e.g., data access level, data usage license, and community metadata standards). We specified the constraints and limitations of the automated assessment in the 'comment' section of the metric specification (Appendix 2).

- **Repository evaluation and certification support:** FAIR assessment of data objects only address a part of the FAIR implementation. The objects' assessment should be further supported by the implementation of the TDR requirements, which we regard as essential to ensure long-term access, preservation, and reuse of the objects. A maturity model representing FAIR-aligned repository requirements is currently being developed by the project (WP4, Task 4.1).

7 Draft Requirements for Implementing FAIR Data Assessment

In this section, we recommend the (draft) requirements to be considered for implementing FAIR data objects assessment. The implementation here refers to building a toolset and applying it to evaluate FAIRness of data objects. The requirements are intended to be applicable to both evaluation mechanisms (automated and manual self-assessment) unless otherwise explicitly stated. The requirements were developed through the following activities. First, we (Task 4.5 team) examined the prior pilots (sections 3.3.1, 3.3.2) developed by the task partners and related documentation (e.g., reports and presentations) to identify limitations and areas of improvement. This led to the formulation of the FAIRsFAIR use cases. For each of the use cases, we developed coarse grained workflows, from which the WP4 members gathered the initial requirements. Then, we organized subsequent discussions with the members to improve the requirements. For example, we identified the requirements applicable to both use cases, decomposed complex requirements into 'atomic' requirements, and determined their level of importance. In addition, we took further steps to eliminate duplication and to refine ambiguous and contradictory statements.

We categorized the requirements into functional and non-functional. The functional requirements (7.1-7.6) specify the features of the implementation, whereas non-functional requirements (7.7-7.10) represent the overall characteristics of the implementation such as reliability, availability and ease of operation. We weighted the requirements based on RFC 2119[35] as follows:

- Mandatory: requirement MUST be satisfied (Essential)
- Recommended: requirement SHOULD be satisfied, if at all possible (Important)
- Optional: requirement MAY be satisfied, but not necessarily so (Useful)

7.1 Metrics Development

a. The metrics for assessing data objects must be congruent with FAIR principles. Further metrics that are essential for domain communities involved should be formulated and included as part of the assessment.

b. The metrics definition must be defined clearly and unambiguously.

c. To facilitate the adoption of the metrics, disciplinary and interdisciplinary practices must be taken into account when defining the metrics, their importance (priority), and compliance levels. This also includes the requirements of data with different access rights (e.g., open, embargoed, restricted), and at different granularities (i.e., data items and data collections).

d. The metrics should be specified systematically (e.g., using a standard template or user documentation) and made publicly available.

---

[35] https://tools.ietf.org/html/rfc2119

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

28

e. The specification of a metric should include its identity (i.e., identifier and name), description with examples, digital resource to be evaluated (e.g., data or metadata), and assessment procedures.

f. The descriptions of the metrics should be reviewed and improved periodically.

g. Each subsequent release of the metrics should be versioned such that it's transparent against which version of the metrics a data object was evaluated.

7.2 Assessment Object

a. The FAIR principles may be applied to evaluate any digital resource. In consideration of Task 4.5 scope, our assessment focuses on research data objects (data and their metadata). Various means of supplying metadata relating to a research data object should be taken into account when assessing the resources. In most of the domains, there is a clear distinction between metadata and data. However, in some domains (e.g., biodiversity), metadata is embedded alongside the data.

7.3 Stakeholders

a. There are various stakeholders who have an interest in the results of FAIR assessment. Some stakeholders may evaluate data FAIRness for deposit or reuse potential, such as research communities, data stewards, or data service providers; others may use the results of the evaluation, e.g., funders or standardization bodies for repository certification. The context of use (i.e., intended users, roles and goals) must be identified before implementing the FAIR data assessment.

b. High-level workflows on how stakeholders may participate in the assessment process should be developed and documented.

7.4 Assessment Method

a. Objective evaluation of FAIRness of data objects should be supported whenever possible. Here, different evaluation methods may be enabled, such as manual (human), automated (machine), or a combination of both. Importantly, the selected evaluation method must fit the purpose of conducting the assessment.

b. For an extensive collection of data objects, a bulk automated evaluation of the objects may be supported.

c. The automated data assessment should be modular and capable of iteration.

d. The assessment results should be communicated transparently with relevant stakeholders with suggestions for FAIRness improvement.

e. Data objects are not static, as they may be updated at any time. Therefore, the FAIRness of a data object should be assessed and updated as required when the changes occur.

7.5 FAIR Scoring and Badging

a. A 'sensible' scoring mechanism must be developed to quantify the overall FAIRness of a data object.

b. Scoring results should be made available in a machine-readable format to allow wide re-use by other services (e.g., visualization tools).

c. A simple and easy-to-understand visualization that summarizes the assessment results may be developed to engage and communicate the critical results to users.

7.6 External Dependency

a. An automated FAIR data assessment tool may be interfaced with appropriate external systems such as registries of metadata, data policies and standards, licenses, and vocabularies, which contain the resources to run the assessment. For example, FAIRsharing is an online catalog of community standards, repositories, and data policies. The registry re3data archives the metadata of more than 2,000 data repositories. SPDX License[36] lists standard licenses.

b. The access rights, availability, and development status of the external systems must be determined before interfacing the automated tool with external systems.

c. Depending on the operational implementation of the FAIR object testing (including the perceived authority of the testing body), the results of FAIR data assessments may be made available through various means as they may have value in the wider linked network of research data. For example, an object registry (e.g., Datacite) may harvest FAIR score/status of data objects as part of their metadata harvesting process.

7.7 Deployment and Operation

a. Instead of starting from scratch, relevant tools and designs should be reused when implementing a FAIR data assessment toolset to accelerate FAIR implementation.

b. The automated assessment tool should be able to support concurrent assessment requests without substantial system lag.

c. The toolset should be hosted at a server maintained by the task partner during the active phase of the project.

---

[36] https://spdx.org/licenses/

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

30

d. The source code of the toolset should be deposited into the FAIRsFAIR GitHub repository[37], under an open-source license to promote its reusability and collaborative development.

e. The collection and processing of personal information of users of the toolset must be in line with the General Data Protection Regulation (GDPR).

7.8 Usability

a. To ensure the toolset is easy to use, its guidelines or 'help' information must be made available to its end-users.

b. Novice FAIR users (e.g., researchers) should be able to use the self-assessment tool to evaluate their data objects after its introductory training. Intermediate and experienced FAIR users (e.g., data stewards) should be able to use the tool based on the guidelines and examples provided.

7.9 Organizational context

a. A repository must communicate the FAIR assessment that will be performed over their data objects to data depositors. For example, FAIR data assessment implementation may be specified in the repository documentation, e.g., terms and conditions, any service policies and articulated in associated workflows.

7.10 Training and Outreach

a. Implementers of the toolset must be guided on adopting the toolset to their systems and processes.

b. Stakeholders carrying out FAIR assessments must be trained on using the metrics and toolset developed by the project to assess data objects, interpreting the assessment results, and making changes to their data objects or repository systems and workflows as required.

---

8 Next Steps

One of the expected outputs of FAIRsFAIR is building pilots to support the assessment of FAIR data objects in trustworthy repositories. To contribute towards this objective, this report provides an overview of FAIR data assessment scenarios and the two use cases that will be developed in more detail going forward. We have provided draft recommendations for requirements, which we will refine and revise over the next few months through feedback from the community and our own work on the first technical implementations. The assessment scenarios will be made available for comment by repositories participating in FAIRsFAIR's support for FAIRification and/or certification at workshops and other events. We will also invite the wider community to comment on our assessment scenarios and contribute suggestions for additional scenarios that may be taken forward if resources allow. We will engage with cluster projects such as Photon and Neutron Open Science Cloud (PaNOSC) and ENVironmental Research Infrastructures building FAIR services Accessible for society (ENVRI-FAIR) to understand if there are discipline-specific needs that our requirements, metrics and implementation will need to address. We will focus efforts on implementing two use cases:

- a manual self-assessment tool that we will test with researchers, and pilot at upcoming FAIRsFAIR data stewardship schools run as part of Task 6.4.

- an automated assessment of published datasets in a repository that will be tested with the repositories selected for in-depth collaboration as part of the open call. Technical details will be elaborated in an upcoming milestone 4.9 report.

Feedback gathered as part of the implementation process will inform the transition support offered to repositories not selected in the open call as part of Tasks 4.3 *Support for FAIR Certification* and feed into *Task 3.4. Transitioning to FAIR data providers, stewardship and repositories*. Guidance for repositories on adopting the toolset will be developed as part of FAIRsFAIR WP3 transition support programme for repositories. The results and related guidelines will be of value to the European network of TDRs on enabling FAIR data that will be established as part of FAIRsFAIR. Training on using the metrics and toolset developed by the project to assess data objects and interpreting the assessment results will be embedded into future instances of the Data Stewardship Strand of the CODATA/RDA Research Data Science Schools being delivered as part of FAIRsFAIR WP6. In collaboration with the FAIRsFAIR Project Coordination Office and Synchronization Force, we will explore options for a sustainability plan for the maintenance and future development of Task 4.5 outputs (metrics and toolset) beyond the project duration. The plan will be documented as part of the first iteration of our FAIR implementation (milestone 4.9).

Currently, the WP4 partners are actively engaged with other FAIR initiatives, including RDA FAIR Data Maturity Model Working Group, ENVRI-FAIR, FAIRplus and EOSC-FAIR Working Group. The feedback gathered on the development and pilot testing will be fed back to these initiatives, to enable collective effort in translating FAIR data assessment into practice. Within WP4, the work on assessing FAIRness of objects runs in parallel to the work in T4.1 to map the CoreTrustSeal Requirements to the FAIR Principles and develop a supporting maturity model. Through the repository support process, we will seek to align these two strands of work into complementary processes through our iterative stages.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

32

Bibliography

Austin, Claire, Helena Cousijn, Michael Diepenbroek, Jonathan Petters, and Marina Soares E Silva. 2019. 'WDS/RDA Assessment of Data Fitness for Use WG Outputs and Recommendations'. http://dx.doi.org/10.15497/rda00034.

Bahim, Christophe, Makx Dekkers, and Brecht Wyns. 2019. 'Results of an Analysis of Existing FAIR Assessment Tools'. RDA Outputs & Recommendations. Research Data Alliance. https://doi.org/10.15497/RDA00035.

Bangert, Daniel, Emilie Hermans, René van Horik, Maaike de Jong, Hylke Koers, and Mustapha Mokrane. 2019. 'Recommendations for Services in a FAIR Data Ecosystem'. Zenodo. https://doi.org/10.5281/zenodo.3585742.

Bonaretti, Serena, and Egon Willighagen. 2019. 'Two Real Use Cases of FAIR Maturity Indicators in the Life Sciences'. biorxiv.org. http://dx.doi.org/10.1101/739334.

Burdett, Tony, Mélanie Courtot, Fuqi Xu, Wei Gu, Venkata Satagopam, Nick Juty, Oya Beyan, et al. n.d. 'FAIRplus: D3.01 First Phase Exemplar IMI Projects FAIRified'. Accessed 25 January 2020. https://doi.org/10.5281/zenodo.3587397.

Clarke, Daniel J. B., Lily Wang, Alex Jones, Megan L. Wojciechowicz, Denis Torre, Kathleen M. Jagodnik, Sherry L. Jenkins, et al. 2019. 'FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources'. *Cell Systems* 9 (5): 417–21. https://doi.org/10.1016/j.cels.2019.09.011.

CoreTrustSeal. 2019. 'CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022'. CoreTrustSeal. https://doi.org/10.5281/zenodo.3638211.

Davidson, Joy, Claudia Engelhardt, Vanessa Proudman, Lennart Stoy, and Angus Whyte. n.d. 'D3.1 FAIR Policy Landscape Analysis'. Accessed 15 January 2020. https://doi.org/10.5281/zenodo.3558173.

Dunning, Alastair, Madeleine de Smaele, and Jasmin Böhmer. 2017. 'Are the FAIR Data Principles Fair?' *International Journal of Digital Curation* 12 (2): 177–95. https://doi.org/10.2218/ijdc.v12i2.567.

European Commission Expert Group on FAIR Data. 2018. 'Turning FAIR into Reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data.' https://doi.org/10.2777/1524.

Fripp, Dom, and Tom Davey. 2019. 'A Proposal to Add a FAIR Metric to the Jisc Open Research Hub'. March 31. https://doi.org/10.5281/zenodo.2619357.

Global Agricultural Research Data Innovation & Acceleration Network. 2019. 'GARDIAN FAIR Metrics'. https://gardian.bigdata.cgiar.org/files/GARDIAN_FAIR_metrics_guide.pdf.

Herczog, Edit, Keith Russell, Vassilios Peristeras, and Makx Dekkers. n.d. 'FAIR Data Maturity Model'. presented at the 14th RDA Plenary session - Workshop #5. https://www.rd-alliance.org/system/files/documents/20191023_FAIR_WG_slides_v0.08.pdf.

Herterich, Patricia, Joy Davidson, Angus Whyte, Laura Molloy, Brian Matthews, and Gabin Kayumbi Kabeya. n.d. 'D6.1 Overview of Needs for Competence Centres'. Accessed 15 January 2020. https://doi.org/10.5281/zenodo.3549791.

Lehväslaiho, Heikki, Jessica Parland-von Essen, Leah Riungu-Kalliosaari, Claudia Behnke, Heidi Laine, Christine Staiger, Hylke Koers, and Yann LeFranc. 2019. 'FAIRsFAIR Data of Survey on

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

33

Semantics and Interoperability Solutions for D2.1 Report on FAIR Requirements for Persistence and Interoperability'. FAIRsFAIR. https://doi.org/10.5281/zenodo.3518922.

L'Hours, Hervé, and Ilona von Stein. 2019. 'FAIR Ecosystem Components: Vision'. Zenodo. https://doi.org/10.5281/zenodo.3565428.

Miksa, Tomasz, Paul Walk, and Peter Neish. n.d. 'RDA DMP Common Standard for Machine-Actionable Data Management Plans'. RDA. Accessed 23 December 2019. https://doi.org/10.15497/rda00039.

Mokrane, Mustapha, and Jonas Recker. 2019. 'CoreTrustSealCertified Repositories - Enabling Findable, Accessible, Interoperable, and Reusable (FAIR) Data'. In *Proceedings of the 16th International Conference on Digital Preservation. 2019*, edited by Marcel Ras, Barbara Sierman, and Angela Puggioni, 92–100. Amsterdam. https://ipres2019.org/static/pdf/iPres2019_paper_74.pdf.

Mosconi, Gaia, Qinyu Li, Dave Randall, Helena Karasti, Peter Tolmie, Jana Barutzky, Matthias Korn, and Volkmar Pipek. 2019. 'Three Gaps in Opening Science'. *Computer Supported Cooperative Work (CSCW)* 28 (3): 749–89. https://doi.org/10.1007/s10606-019-09354-z.

Research Libraries Group. 2002. 'Trusted Digital Repositories: Attributes and Responsibilities'. RLG-OCLC Report. RLG, Inc. https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf.

White, Marijane, Lily Winfree, Payal Mehndiratta, Kimberly Robasky, and Robin Champieux. n.d. 'Modeling the FAIR Rubrics Landscape'. Accessed 19 December 2019. https://osf.io/685sw/.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3 (1): 1–9. https://doi.org/10.1038/sdata.2016.18.

Wilkinson, Mark D., Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, et al. 2019. 'Evaluating FAIR Maturity through a Scalable, Automated, Community-Governed Framework'. *Scientific Data* 6 (1): 1–12. https://doi.org/10.1038/s41597-019-0184-5.

Wilkinson, Mark D., Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. 2018. 'A Design Framework and Exemplar Metrics for FAIRness'. *Scientific Data* 5 (June). https://doi.org/10.1038/sdata.2018.118.

Appendix 1: Mapping Between Existing FAIR Data Models

FAIR Principles	FAIRsFAIR Metric Identifier and Short Name	FAIRsFAIR Metric and Maturity Levels	RDA Data Maturity Model WG Indicator Identifier	RDA Data Maturity Model WG Indicator and Maturity Levels	FAIRDAT	FAIREnough	WDS/RDA Assessment of Data Fitness for Use WG
F1	FsF-F1-01D Universally Unique Identifier	Does the data have a universally unique identifier assigned? • Yes • No	F1-02D	Data is identified by a universally unique identifier • NO universally unique identifier • Universally unique and unambiguous identifier			
F1	FsF-F1-02D Persistent Identifier	Does the data have a persistent identifier assigned? • Yes • No	F1-01D	Data is identified by a persistent identifier • NO persistent identifier • Persistent identifier		Will your dataset have a Persistent Identifier after deposit?	Persistent identification of the dataset and related work (related literature and data, authors, projects, terms)
F2	FsF-F2-01M Descriptive Metadata	Are metadata elements to support data citation and discovery provided (e.g., creator, title, data identifier, publisher, title, creator, publication date/year, summary/keywords describing the data)? • Not provided • Partially provided • Completely provided	F2-01M	Sufficient metadata is provided to allow discovery, following domain/discipline-specific metadata standard • NO sufficient metadata to allow discovery is provided • Sufficient metadata to allow discovery is provided	Did you provide sufficient metadata (information) about your data for others to find, understand and reuse your data?	Did you provide enough information (metadata) about your data for others to understand and reuse your data?	Citation exists, including authorship, year, comprehensive title, persistent identifier (e.g., DOI)
F3	FsF-F3-01M Inclusion of data identifier in metadata	Does the metadata include the data identifier? • Yes • No	F3-01M	Metadata includes the identifier for the data • NO identifier of the data in the metadata • Identifier of the data present in the metadata			

F4	FsF-F4-01M Searchable metadata	Is the metadata offered in such a way that it can be harvested? • Metadata is not offered • Metadata is offered through a metadata registry, e.g., general-purpose, domain/discipline specific or institutional registries • Metadata is offered as structured data on the data page for use by a web search engine	F4-01M	Metadata is offered/ published/ exposed in such a way that it can be harvested and indexed • Metadata CANNOT be harvested and indexed • Metadata can be harvested and indexed	Is the metadata publicly accessible even if the data is no longer available?	Is the metadata publicly accessible?	
A1	FsF-A1-01M Data Access Level	Does the metadata include a statement on the level of access to the data? • Yes • No If yes, select the level of access to the data • Public access • Embargoed access • Restricted access • Metadata only access • None of the above (One of the following questions will be addressed to the user if the user selects 'Embargoed access' or 'Restricted access') If the data is an embargoed data, does the metadata include the date the data will be released publically? • Yes • No If the data is restricted, does the metadata include access conditions, e.g., point of contact and instructions to access the data? • Yes • No	A1-01M	Metadata includes information about access conditions • NO information about access conditions given • Information about access conditions given	Does your dataset contain personal data?		Terms of usage (licenses, other conditions of reuse, data protection, ethical issues)

A2	FsF-A2-01M Metadata Preservation	Is there a statement indicating that the metadata remain available once the dataset becomes unavailable? • Yes • No	A2-01M	Metadata is guaranteed to remain available after data is no longer available • No persistence policy/guarantee • Persistence policy/guarantee	Is the metadata publicly accessible even if the data is no longer available?	Is the metadata publicly accessible?	
I1	FsF-I1-01M Semantic Representation of Metadata	Does the metadata use any semantic vocabularies (e.g., ontologies, thesauri, taxonomies)? • No • Yes, semantic vocabularies embedded in the metadata • Yes, metadata is published as linked data using semantic vocabularies	I1-02M	Metadata uses machine-understandable knowledge representation • Knowledge representation is NOT machine-understandable • Knowledge representation is machine-understandable	Did you use standards such as controlled vocabularies, taxonomies (thesauri) or ontologies to describe your dataset?	Did you use standardized vocabulary?	Sufficient discovery metadata: metadata includes community accepted keywords and/or terms associated with relevant standards or terminologies
			I2-01M	Metadata uses standard vocabularies • NO standard vocabularies • Standard vocabularies			
I3	FsF-I3-01M Qualified References to Related Entities	Does the metadata include qualified references (links) between the data and its related entities? • Yes • No	I3-01D	Data includes references to other data • NO references to other data • References to other data	Did you provide rich and detailed additional documentation?	Did you provide rich additional documentation?	
			I3-02D	Data includes sufficiently qualified references to other data • NO qualification, e.g., simple URL link, without indication what it means • Specific, machine-understandable qualification, e.g., indicating author, publisher, etc.	Do you link to other (meta)data and is this (meta)data online resolvable?		
R1	FsF-R1-01M Community-Endorsed Metadata	Does the metadata meet a community-endorsed standard? • Yes • No • Standard unknown/unavailable	R1-01M	Sufficient metadata is provided to allow reuse, following domain/discipline-specific metadata standard • NO sufficient metadata to allow reuse is provided • Sufficient metadata to allow reuse is provided	Does your metadata meet domain standards?	Do you make use of relevant community standards?	Additional metadata adequate to respective research domain (if applicable)

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

37

R1	FsF-R1-02M Data Content Description	Do the content descriptions specified in the metadata conform with the data? • Content description not specified • Fully conformed • Somewhat conformed • Not conformed	Not specified	Not specified	Did you provide contextual information about your dataset?		Description of the dataset content adequate; this includes the following • An abstract • A listing of measurement & observation types, or parameterizations & simulation types including methods used • a description of size, structure, and data format/MIME type Structure, size and MIME type of the dataset agrees with description of the dataset content Content of the dataset agrees with description of the dataset content.
R1.1	FsF-R1.1-01M Data Usage License	Does the metadata include license information under which the data can be reused? • Yes • No • Not applicable	R1.1-01M	Metadata includes information about the licence under which the data can be reused • NO user license • Presence of a user license	Which of the usage licenses provided by EASY did you choose in order to comply with the access rights attached or the data? Which usage license are you planning to add to your dataset?	Does the dataset have a usage licence?	Terms of usage (licenses, other conditions of reuse, data protection, ethical issues)
R1.3	FsF-R1.3-01M Common File Format	Is the data available in standard file format? • Yes • No	R1.3-01D	Data complies with a community standard	Are the data in your dataset stored in preferred formats?	Are the data stored and archived in preferred archival formats?	Dataset is provided in a widely used or community accepted machine-readable format and using standard terminologies for nominal data and available standard protocols

Appendix 2: FAIRsFAIR Data Assessment Metric Specification

The following are metrics for assessing FAIRness of data objects. They are specified following the template below, which was modified from the template originally recommended by (Wilkinson et al., 2018).

Field	Description
Metric Identifier	The identifier of the metric.
Metric Name	The short name of the metric.
Metric Description	The definition of the metric, including its examples and supporting details.
To which FAIR principle(s) does it apply?	The FAIR principle addressed by the metric.
To which CoreTrustSeal requirement(s) does it apply?	The CoreTrustSeal requirements addressed by the metric. One metric may be related to one or more CoreTrustSeal requirements.
For which digital resource is this relevant?	The type of digital resource that will be assessed based on the metric, e.g., data or metadata.
Manual Assessment (User Question)	The metric question that will be addressed to users.
Automated Assessment	The details on the automated assessment of the metric, including inputs, methods and outputs.
Comments	A list of related resources, constraints and limitations of the proposed assessment.

1. Universally Unique Identifier

Field	Description														
Metric Identifier	FsF-F1-01D														
Metric Name	Universally unique identifier														
Metric Description	The data is assigned with a globally unique identifier such that it can be referenced unambiguously on the Web. In other words, the identifier should be associated with only one dataset at any time. Examples of unique identifiers of data are Uniform Resource Identifier (URI), Digital Object Identifier (DOI), the Handle System, identifiers.org, w3id.org and Archival Resource Key (ARK). We make a distinction between persistence (FsF-F1-02D) and uniqueness of an identifier. An HTTP URL is globally unique, but is not persistent, whereas a DOI is both globally unique and persistent.														
To which FAIR principle(s) does it apply?	F1	F2	F3	F4	A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
	X														
To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
													X			

For which digital resource is this relevant?	Data		Metadata	
	X			

Manual Assessment (User Question)	Question Type
Does the data have a universally unique identifier assigned? ● Yes ● No	Single-choice question

Automated Assessment

Input	Assessment Method	Output
● Data URI	Check if the identifier is based on a globally unique identifier scheme.	● Yes ● No A successful request (Yes) returns one of the following additional results: ● Web address (if a URL is specified) ● Unique identifier and the identifier scheme specified. A failed request (No) returns an error message.

Comments

Resources
The following are examples of sites that list identifier schemes:
- Identifier schema compiled by FAIRsharing from various sources, https://fairsharing.org/standards/?q=&selected_facets=type_exact:identifier%20schema
- Uniform Resource Identifier (URI) Schemes, https://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml#uri-schemes-1
- URI examples included in rfc3986, https://tools.ietf.org/html/rfc3986#section-1.1.2

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

40

2. Persistent Identifier

Field	Description
Metric Identifier	FsF-F1-02D
Metric Name	Persistent identifier
Metric Description	The data is assigned with a persistent identifier to ensure the resolvability of the identifier in the long term. The identifier may be resolved to its digital object (e.g., a data file or a web service that returns the data), or to a data proxy (e.g., an online page that contains metadata, including the link to access the data).

To which FAIR principle(s) does it apply?	F1	F2	F3	F4	A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
	X														

To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R1
													X			

For which digital resource is this relevant?	Data	Metadata
	X	

Manual Assessment (User Question)	Question Type
Does the data have a persistent identifier assigned? • Yes • No	Single-choice question

Automated Assessment

Input	Assessment Method	Output
• Data URI	The identifier is based on a persistent identifier scheme, and it resolves to a web address where the data can be accessed. Verify the identifier resolvability through the HTTP response status codes.	• Yes • No A successful request (Yes) returns the following additional results: • The persistent identifier scheme • The resolved URL A failed request (No) returns an error message.

Comments

Resources
- A wiki entry on persistent identifier, https://en.wikipedia.org/wiki/Persistent_identifier

Known Limitations/Constraints
- The automated assessment verifies the resolvability of the specified identifier, but does not assert the type of the resolved object, e.g., whether the resolved object is landing page, a file or a web service query response.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

41

3. Descriptive Metadata

Field	Description
Metric Identifier	FsF-F2-01M
Metric Name	Descriptive metadata
Metric Description	Metadata is descriptive information of data. Since the metadata required depends on users and their applications, this metric focuses on core metadata, which are the minimum metadata required for data citation and discoverability. We determine the required metadata based on the existing data citation guidelines, e.g., DataCite, ESIP, and IASSIST, and metadata recommendations for data discovery, e.g., DataCite Metadata Schema and RDA Metadata Interest Group. This metric focuses on domain-agnostic core metadata; we address domain or discipline-specific metadata through the metric FsF-R1-01M.

To which FAIR principle(s) does it apply?	F1	F2	F3	F4	A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
		X													

To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
													X			

For which digital resource is this relevant?	Data	Metadata
		X

Manual Assessment (User Question)	Question Type
Are metadata elements provided to support data discovery and citation (e.g., creator, title, data identifier, publisher, publication date/year, summary/keywords describing the data)? • Not provided • Partially provided • Completely provided	Single-choice question

Automated Assessment

Input	Assessment Method	Output
• Name/URL of the metadata standard • URL of the machine-readable metadata document	Use the metadata standard name/URL to identify elements (fields) representing core metadata in the standard. Verify presence/absence of the elements in the metadata document.	• No • Yes, partial metadata • Yes, all metadata A successful request (Yes) returns metadata url, and the key-value pairs of the analyzed metadata elements. A failed request (No) returns an error message.

Comments

Resources
The following sites list examples of metadata standards for data:
- https://fairsharing.org/standards/
- http://rd-alliance.github.io/metadata-directory/standards/
- http://www.dcc.ac.uk/resources/metadata-standards/list

Known Limitations/Constraints
- The URLs (metadata standard and metadata document) provided by the user might be broken or not accessible (e.g., proprietary standard).

4. Inclusion of Data Identifier in Metadata

Field	Description														
Metric Identifier	FsF-F3-01M														
Metric Name	Inclusion of data identifier in metadata														
Metric Description	The metadata includes the identifier of the data such that users can access the data through the metadata.														
To which FAIR principle(s) does it apply?	F1	F2	F3	F4	A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
			X												
To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
													X			

For which digital resource is this relevant?	Data	Metadata
		X

Manual Assessment (User Question)	Question Type
Does the metadata include the data identifier? • Yes • No	Single-choice question

Automated Assessment

Input	Assessment Method	Output
• Data URI • Name/URL of the metadata standard • URL of the machine-readable metadata document	Verify the presence/absence of the data identifier through metadata element(s). For example, the metadata element 'identifier' may be used to represent data identifier in the Datacite Metadata Schema v4. Check if it is active (or web-resolvable)	• Yes • No A successful request (Yes) returns the following additional results: • The metadata element and its value (data URI), and its status. A failed request (No) returns an error message.

Comments

Known Limitations/Constraints
- It might be possible that the user specifies the URL of a data landing page instead of the URI of the data in the metadata document.
- A metadata standard may include various elements (fields) through which a data URI may be specified.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

43

5. Searchable Metadata

Field	Description
Metric Identifier	FsF-F4-01M
Metric Name	Searchable metadata
Metric Description	This metric refers to various ways through which the metadata of data is exposed or offered in a machine-readable format. For example, metadata may be offered through a general or domain/discipline specific metadata registry. It may be embedded as structured data (e.g., schema.org implementation) on a data page for use by web search engines such as Google and Bing.

To which FAIR principle(s) does it apply?	F1	F2	F3		F4	A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
					X											

To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
													X			

For which digital resource is this relevant?	Data	Metadata
		X

Manual Assessment (User Question)	Question Type
Is the metadata offered in such a way that it can be harvested? • Metadata is not offered • Metadata is offered through a metadata registry, e.g., general-purpose, domain/discipline specific or institutional registries • Metadata is offered as structured data on the data page for use by a web search engine	Conditional multiple-choice question

Automated Assessment

Input	Assessment Method	Output
• Data URI • Query URL of the metadata registry for retrieving the machine-readable metadata document.	Resolve the data URI and test the structured data on the data landing page Test if the query URL returns the metadata of the data	• No • Yes, metadata registry • Yes, structured data A successful request (Yes) returns the metadata elements used in testing (e.g., title, author) and their values. A failed request (No) returns an error message.

Comments

Known Limitations/Constraints
• In order to verify that the metadata is findable through a web search engine, we should perform a search through a web search engine API based on the data identifier and its descriptive metadata. However, most of the web search engine APIs (e.g., Google Custom Search, Bing Web Search API) offer a limited number of free search queries.

6. Data Access Level

Field	Description
Metric Identifier	FsF-A1-01M
Metric Name	Data access level
Metric Description	This metric determines if the metadata includes the level of access to the data such as public, embargoed, restricted, or closed (metadata only) access. Public data are openly accessible, e.g., through FTP/HTTP-based direct download. Embargoed access refers to data that will become openly available after a specific date. For example, a data author may release their data after having published their findings from the data. Access to restricted data is limited, e.g. because of commercial, sensitive, or other confidentiality reasons or the data are only accessible via a subscription or a fee. Restricted data may be available to a particular group of users or after permission is granted. For embargoed access, the date the data will be made publicly available should be specified in the metadata. For restricted data, the metadata should include the conditions of access to the data (e.g., point of contact or instructions to access the data). This metric is related to the data usage license metric (FsF-R1.1-01M).

To which FAIR principle(s) does it apply?	F1	F2	F3	F4	A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
					X										

To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
															X	

For which digital resource is this relevant?	Data	Metadata
		X

Manual Assessment (User Question)	Question Type
Does the metadata include a statement on the level of access to the data? • Yes • No If yes, select the level of access to the data • Public access • Embargoed access • Restricted access • Metadata only access • None of the above (One of the following questions will be addressed to the user if the user selects 'Embargoed access' or 'Restricted access') If the data is embargoed, does the metadata include the date the data will be released publically? • Yes • No If access to the data is restricted, does the metadata include access conditions, e.g., point of contact or instructions to access the data? • Yes • No	Single-choice question

Automated Assessment

Input	Assessment Method	Output
• Name/URL of the metadata standard • URL of the machine-readable metadata document	Verify the presence/absence of data access level through metadata element(s). If it is embargoed data, check if the embargo end date is specified. If it is restricted data, check if the data access conditions are specified.	The response returns the access level of data, and related metadata (if the data is embargoed or restricted data). If no access level is specified in the metadata, it returns 'not found'.

Comments

Resources

- Open Digital Rights Language (ODRL) Information Model 2.2, https://www.w3.org/TR/odrl-model/
- Controlled Vocabulary for Access Rights, http://vocabularies.coar-repositories.org/documentation/access_rights/
- Archival Access Rights Vocabulary, http://sandbox.metadataregistry.org/concept/list/vocabulary_id/251.html
- Eprints AccessRights Vocabulary Encoding Scheme, http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_AccessRights_Vocabulary_Encoding_Scheme

Known Limitations/Constraints

- The metadata standard used may not include elements for representing access levels and conditions.
- The access level information may be expressed in an unstructured manner, e.g., as a 'comment' in the metadata document.
- This assessment should be complemented with the evaluation on the data access mechanism based on the specified access rights, e.g., data is not accessible, data is accessible semi-automated (mediated access to data via data custodian) or automated ways.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

46

7. Metadata Preservation

Field	Description
Metric Identifier	FsF-A2-01M
Metric Name	Metadata preservation
Metric Description	Metadata should be preserved even when the data they represent are no longer available or lost.

To which FAIR principle(s) does it apply?	F1	F2	F3	F4	A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
								X							

To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
										X						

For which digital resource is this relevant?	Data	Metadata
		X

Manual Assessment (User Question)	Question Type
Is there a statement indicating that the metadata remain available once the dataset becomes unavailable? • Yes • No	Single-choice question

Automated Assessment

Input	Assessment Method	Output
• The URL of the metadata preservation document	Check if the document URL returns a web resource through HTTP response status codes.	• Accessible over the web • Not accessible over the web

Comments

Known Limitations/Constraints
- Continued access to metadata depends on a data repository's preservation practice which is usually documented in the repository's service policies or statements. The automated assessment checks the accessibility of the document. Future assessment of the metric should also consider the machine readability of the document.

8. Semantic Representation of Metadata

Field	Description
Metric Identifier	FsF-I1-01M
Metric Name	Semantic representation of metadata
Metric Description	To make metadata more understandable to humans and machines, they are described with semantic vocabularies. Ontology, thesaurus, taxonomy are kinds of semantic vocabularies, and they come with different degrees of expressiveness, structure, and inferential power. Metadata may use semantic vocabularies in various ways. For example, semantic vocabularies may be embedded as controlled vocabularies in the metadata or the metadata web page (e.g., microformats and RDFa). Metadata may also be published as linked data using semantic vocabularies.

To which FAIR principle(s) does it apply?	F1	F2	F3	F4		A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
										X						

To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
															X	

For which digital resource is this relevant?	Data	Metadata
		X

Manual Assessment (User Question)	Question Type
Does the metadata use any semantic vocabularies (e.g., ontologies, thesauri, taxonomies)? • No • Yes, semantic vocabularies embedded in the metadata • Yes, metadata is published as linked data using semantic vocabularies	Conditional multiple-choice question

Automated Assessment

Input	Assessment Method	Output
• Metadata URI, if it resolves to a metadata page • URL to access the machine-readable metadata document	Resolve the metadata URI and extract structured data from the metadata page Check the content type of the return header of the metadata document, e.g., application/rdf+xml	• No • Yes, embedded vocabularies. • Yes, linked data. A successful request (Yes) returns the serialization format of the metadata. A failed request (No) returns an error message.

Comments

Resources
- A list of content types is available at https://www.iana.org/assignments/media-types/media-types.xhtml

Known Limitations/Constraints
- The automated assessment checks the presence of the structured data in the metadata document, not their contents, e.g., if the vocabularies used are in appropriate context and accessible over the web.
- RDF data can be expressed in a number of different ways, e.g., RDF/XML, Turtle, RDFa and RDF/JSON. Therefore, different serialization formats should be considered as part of the automated assessment.

9. Qualified References to Related Entities

Field	Description
Metric Identifier	FsF-I3-01M
Metric Name	Qualified references to related entities
Metric Description	Linking data to its related entities will increase its FAIRness, and the linking information should be captured as part of the metadata. A rich research graph (e.g., PID graph) can be formed by aggregating the entities connections from different data providers. A data object may be linked to its prior version, other datasets in the same data collection, related publications, source (instrument), data creators or collectors and organization (e.g., funder and hosting institution). Qualified references refer to the meaningful links between data and its related entities expressed through relation types. It is also essential to test if the URL of the related entities are active.

To which FAIR principle(s) does it apply?	F1	F2	F3	F4		A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
												X				

To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
											X					

For which digital resource is this relevant?	Data	Metadata
		X

Manual Assessment (User Question)	Question Type
Does the metadata include qualified references (links) between the data and its related entities? • Yes • No	Single-choice question

Automated Assessment

Input	Assessment Method	Output
• URL of the machine-readable metadata document	Check the metadata elements which indicate the relationship between data and related entities. Check if the URL of the related entity is active.	• Yes • No A successful request (Yes) returns the URL and type of the related entities and their status (active, not active). A failed request (No) returns an error message.

Comments

Resources
- The DataCite Metadata Schema specifies relation types between research entities, https://schema.datacite.org/meta/kernel-4.3/include/datacite-relationType-v4.xsd

Known Limitations/Constraints
- Different metadata schemas may use different properties to specify the relation between data and its related entities.
- The automated assessment regards any relation between a data and its related entities as success. It does not consider the number or types of relations.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

49

10. Community-Driven Metadata

Field	Description
Metric Identifier	FsF-R1-01M
Metric Name	Community-driven metadata
Metric Description	In addition to core metadata required to support data citation and discovery (FsF-F2-01M), metadata to support data reusability should be made available following community metadata standards. Community metadata standards may exhibit different levels of readiness. Some communities have well-established metadata standards, e.g., geospatial (19115), biodiversity (DarwinCore, ABCD, EML), social science (DDI). In contrast, others, including new domains, may have limited or standards that are under development, e.g., engineering and linguistics.

To which FAIR principle(s) does it apply?	F1	F2	F3	F4		A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
														X		

To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
														X		

For which digital resource is this relevant?	Data	Metadata
		X

Manual Assessment (User Question)	Question Type
Does the metadata meet a community-endorsed standard? • Yes • No • Standard unknown/unavailable	Single-choice question

Automated Assessment

Input	Assessment Method	Output
• Subject area • URL of the metadata standard • URL of the machine-readable metadata document that meets the specified standard	Based on the specified subject area, verify if the specified metadata standard is listed as one of the community-endorsed standards from a metadata registry. Use namespace elements and attributes to test if the metadata document is formatted based on the standard.	• Yes • No • Standard unknown A successful request (Yes) returns the URL of the community standard. A failed request (No) returns an error message.

Comments

Resources
- Examples of the community metadata standards and their status are available at FAIRsharing (https://fairsharing.org/standards/) and RDA Metadata Directory (http://rd-alliance.github.io/metadata-directory/subjects/).

Known Limitations/Constraints
- Metadata completeness is conditioned by the best practices and metadata standards within a given discipline or domain. Future evaluation of the metric should also consider metadata completeness, i.e., the degree to which the metadata are specified based on a standard of the data domain.

11. Data Content Description

Field	Description														
Metric Identifier	FsF-R1-02M														
Metric Name	Data content description														
Metric Description	This metric evaluates if the content descriptions specified in the metadata conform with the actual data. Examples of content descriptions are variables measured, methods, structure, MIME type, and size.														
To which FAIR principle(s) does it apply?	F1	F2	F3	F4		A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
													X			
To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
											X					
For which digital resource is this relevant?	Data								Metadata							
									X							

Manual Assessment (User Question)	Question Type
Do the content descriptions specified in the metadata conform with the data? • Content description not specified • Fully conformed • Somewhat conformed • Not conformed	Single-choice question

Automated Assessment

Input	Assessment Method	Output
• Name/URL of the metadata standard • URL of the machine-readable metadata document • Data URI	Verify the presence/absence of metadata elements representing content description of the data. If the related metadata are specified, retrieve the data and identify its properties such as size and media type (MIME). Compare the metadata values with the data properties values.	• No content descriptions • Yes • No A successful request (Yes) returns the key-value pairs of media type and size. A failed request (No) returns an error message.

Comments

Resources
• Model for Tabular Data and Metadata on the Web, https://www.w3.org/TR/tabular-data-model/

Known Limitations/Constraints
• The automated assessment is limited to the evaluation of the size and media type of the data. Other data content descriptions (e.g., method, variable measured) cannot be assessed programmatically due to the challenge of parsing different types of data, and unstructured content descriptions included in a data file.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

51

12. Data Usage License

Field	Description
Metric Identifier	FsF-R1.1-01M
Metric Name	Data Usage Licence
Metric Description	In general, all data should be licensed because otherwise, users cannot easily reuse them in a legally sound way. This includes standard (e.g., Creative Commons) or bespoke licenses, and rights statement which indicate the conditions under which data can be reused. It is highly recommended to use a standard, machine-readable license such that it can be interpreted by machines and humans. In order to inform users about what rights they have to use a dataset, the license information should be specified as part of the dataset's metadata.

To which FAIR principle(s) does it apply?	F1	F2	F3	F4		A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
														X		

To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
		X														

For which digital resource is this relevant?	Data	Metadata
		X

Manual Assessment (User Question)	Question Type
Does the metadata include licence information under which the data can be reused? • Yes • No • Not applicable	Single-choice question

Automated Assessment

Input	Assessment Method	Output
• URL of the metadata standard • URL of the machine-readable metadata document	Verify the presence/absence of licence information through metadata element(s). If a license URL is provided, test if it is active. Using the provided metadata, search a licence registry to determine if the license specified is free and open source.	• Yes • No A successful request (Yes) returns the status of the licence (active, not active), and additional information retrieved from the licence registry (e.g., free and open source), if applicable. A failed request (No) returns an error message.

Comments

Resources
- A list of commonly used licences is available at the SPDX License (https://spdx.org/licenses/).
- Rights statements of cultural heritage objects, https://rightsstatements.org/page/1.0/?language=en

Known Limitations/Constraints
- The automated assessment checks if the licence information provided as part of the metadata. It does not validate if the specified licence is the most appropriate licence for the data.
- In some cases, a data author may specify a brief contact information to convey permission to use his data. This method is insufficient, unless the information point to the location of the license or rights statement.

13. Standard File Format

Field	Description
Metric Identifier	FsF-R1.3-01D
Metric Name	Standard file format
Metric Description	File formats refer to methods for encoding digital information. For example, CSV for tabular data, NetCDF for multidimensional data and GeoTIFF for raster imagery. Data should be made available in a standard file format that is accepted by the research community to enable data sharing and reuse.

To which FAIR principle(s) does it apply?	F1	F2	F3	F4	A1	A1.1	A1.2	A2	I1	I2	I3	R1	R1.1	R1.2	R1.3
															X

To which CoreTrustSeal requirement(s) does it apply?	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
															X	

For which digital resource is this relevant?	Data	Metadata
	X	

Manual Assessment (User Question)	Question Type
Is the data available in a standard file format? • Yes • No	Single-choice question

Automated Assessment

Input	Assessment Method	Output
• Data URI	Retrieve media type of the resource from the header of HTTP response. Verify if the specified format is listed in a file format registry.	• Yes • No A successful request (Yes) returns the media type. A failed request (No) returns an error message.

Comments

Resources
- A list of common media types, https://www.iana.org/assignments/media-types/media-types.xhtml
- Examples of recommended file formats based on data types, https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats.aspx
- Public file format registry, https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=new
- List of open formats, https://en.wikipedia.org/wiki/List_of_open_formats

Known Limitations/Constraints
- The file formats recommended by the community of the data should be determined prior to testing the metric.
- A data URI may resolve to a compressed file (*.zip) which contains actual data files. The compressed file should be extracted to identify the actual file format.