FAIRSFAIR Data Object Assessment Metrics

Authors	Anusuriya Devaraju, Robert Huber, Mustapha Mokrane, Linas Cepinskas, Joy Davidson, Patricia Herterich, Herve L'Hours, Jerry de Vries, Angus Whyte.
DOI	10.5281/zenodo.3934401
Date Published	10.07.2020
Version	0.3
License	https://creativecommons.org/licenses/by/4.0/

Versioning History

Version	Date	Notes
0.3	10 July 2020	This specification contains 15 metrics. The metric (FsF-I1-01M) from v0.2 was divided into two metrics (FsF-I1-01M and FsF-I1-02M). Metrics were improved/updated based on the focus group-based evaluation and the final version of the RDA FAIR Data Maturity Model (https://doi.org/10.15497/RDA00050).
0.2	30 April 2020	Metrics were refined based on the feedback provided by FAIRsFAIR partners. New metric (FsF-R1.2-01M Data Provenance) is added to the specification, sums up to a total of 14 metrics. Link: https://doi.org/10.5281/zenodo.3775793
0.1	25 February 2020	Includes 13 metrics to assess the FAIRness of data objects, which were developed based on existing work (FAIRdat/FAIREnough, WDS/RDA Assessment of Data Fit Checklist and RDA FAIR Data Maturity Model v0.03). Link (Appendix II): https://doi.org/10.5281/zenodo.3678715

Disclaimer

FAIRsFAIR has received funding from the European Commission's Horizon 2020 research and innovation programme under the Grant Agreement no. 831558 The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

Table of Contents

1. Introduction

The overall goal of FAIRsFAIR[1] is to develop practical solutions to facilitate the application of the FAIR principles[2] throughout the research data life cycle. One of the expected outcomes of FAIRsFAIR is building pilots to support the assessment of FAIR digital objects from selected members of the European network of FAIR-enabling Trustworthy Digital Repositories (TDRs). While FAIR principles may apply to any digital objects, we are concerned with the subset of digital objects: research data[3] that are collected, measured, or created for purposes of scientific analysis.

1.1 Purpose

This specification presents 15 minimum viable metrics to systematically measure to what extent research data objects are FAIR. A research data object[4] may comprise data, metadata, and documentation (e.g., policies and procedures). These components influence the implementation of the FAIR assessment. For instance, they can either be resources to be evaluated or evidence of enabling FAIR. The metrics are developed in stages, and are based on indicators proposed by the RDA FAIR Data Maturity Model Working Group[5], in addition to prior work conducted by the project partners such as FAIRdat[6] and FAIREnough[7], and WDS/RDA Assessment of Data Fitness for Use checklist[8]. We will apply the metrics by implementing tools to support FAIR assessment in selected use cases. Nonetheless, we welcome the possibilities of adapting the metrics to support different FAIR assessment scenarios[9] in the research data lifecycle.

1.2 Scope

In its current form, the specification applies metrics that may correspond to all or part of one or more FAIR principles. To be inclusive of current data practices, we will refine and revise the metrics through several iterations based on feedback from stakeholders interested in FAIR, and on the implementation of our use cases to demonstrate FAIR assessment. A new metric will be incorporated into the specification if required by a majority of participating TDRs. Ultimately, we strive to define metrics to address most FAIR principles and as explicitly as possible, both at data and metadata level. We recognize that data quality elements (e.g., completeness, precision/accuracy, validity, ease of data use), and data archival, preservation, and retention aspects are essential, but they are not within the scope of this specification.
In addition to defining metrics against FAIR principles, the assessment of the metrics proposed in this specification depends on several factors.

[1] https://www.fairsfair.eu/
[2] https://www.force11.org/group/fairgroup/fairprinciples
[3] http://www.rdm.kit.edu/english/research.php
[4] In this specification, we use the terms 'data object' and 'dataset' synonymously.
[5] RDA FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model: specification and guidelines. Research Data Alliance. DOI: 10.15497/RDA00050
[6] Research Data Journal - FAIR Data Review, https://docs.google.com/forms/d/e/1FAIpQLSd8_pd2r2SnjCVfCC3CHhEUHZzv2MTRC3RTh0S2YTvbVJj87Q/viewform
[7] https://docs.google.com/forms/d/e/1FAIpQLSf7t1Z9IOBoj5GgWqik8KnhtH3B819Ch6ID5KuAz7yn0I0Opw/viewform
[8] Austin, C., Cousijn, H., Diepenbroek, M., Petters, J., Soares E Silva, M. (2019). WDS/RDA Assessment of Data Fitness for Use WG Outputs and Recommendations. DOI: 10.15497/rda00034
[9] An overview of FAIR data assessment scenarios is available at Devaraju, A, and Herterich, P. (2020). D4.1 Draft Recommendations on Requirements for Fair Datasets in Certified Repositories (Version v1.0_draft). Zenodo. DOI: 10.5281/zenodo.3678715

- In the FAIR ecosystem[10], FAIR assessment must go beyond the object itself. FAIR enabling services and repositories are vital to ensure that research data objects remain FAIR over time. Importantly, machine-readable services (e.g., registries) and documents (e.g., policies) are required to enable automated tests.
- In addition to repository and services requirements, automated testing depends on clear, machine assessable criteria. Some aspects (rich, plurality, accurate, relevant) specified in FAIR principles still require human mediation and interpretation.
- The tests must focus on generally applicable data/metadata characteristics until domain/community-driven criteria have been agreed (e.g., appropriate schemas and required elements for usage/access control). For example, for some of the metrics (i.e., on I and R principles), the automated tests we proposed only inspect the 'surface' of criteria to be evaluated. Therefore, tests are designed in consideration of generic cross-domain metadata standards such as dublin core, dcat, datacite, schema.org, etc.

For each of the metrics, we include further details on the limitations and constraints of its assessment.

1.3 Metric Outline

The metrics are specified following the template (Table 1), modified from Wilkinson et al. (2018)[11]. In each metric table, we provide the descriptions and assessment details of the metric, and its alignment with the relevant FAIR principle and CoreTrustSeal requirement(s).

Table 1. Modified Metric Template

Field	Description
Metric Identifier	The local (FAIRsFAIR) identifier of the metric (for more details, see Figure 1).
Metric Name	Metric name in a human readable form.
Description	The definition of the metric, including examples.
FAIR Principle	The FAIR principle most related to the metric.
CoreTrustSeal Alignment	The CoreTrustSeal requirement(s) most related to the metric.
Assessment	Requirements and methods to perform the assessment against the metric.
Comments	A list of related resources which may be used as a reference basis to implement the assessment, constraints and limitations of the proposed assessment.

Each of the FAIRsFAIR metrics is identified following a naming convention. For example, in Figure 1, 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 distinguishes the resource that will be evaluated based on the metric, e.g., data or metadata.

[10] L'Hours, H. and von Stein, I. (2020). FAIR Ecosystem Components: Vision (Version 02.00). Zenodo. DOI: 10.5281/zenodo.3734273
[11] Wilkinson, MD., Sansone, SA., Schultes, E., Doorn, P., Bonino da Silva Santos, LO., Dumontier, M. (2018). A design framework and exemplar metrics for FAIRness. Sci Data. 2018;5:180118. DOI:10.1038/sdata.2018.118

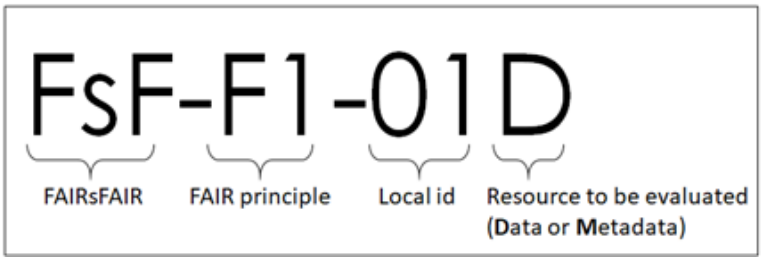

Figure 1. Anatomy of FAIRsFAIR metric identifier.

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

Table 2. List of Metrics.

Identifier	Name
FsF-F1-01D	**Data** is assigned a globally unique identifier.
FsF-F1-02D	**Data** is assigned a persistent identifier.
FsF-F2-01M	**Metadata** includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability.
FsF-F3-01M	**Metadata** includes the identifier of the data it describes.
FsF-F4-01M	**Metadata** is offered in such a way that it can be retrieved by machines.
FsF-A1-01M	**Metadata** contains access level and access conditions of the data.
FsF-A2-01M	**Metadata** remains available, even if the data is no longer available.
FsF-I1-01M	**Metadata** is represented using a formal knowledge representation language.
FsF-I1-02M	**Metadata** uses semantic resources.
FsF-I3-01M	**Metadata** includes links between the data and its related entities.
FsF-R1-01MD	**Metadata** specifies the content of the data.
FsF-R1.1-01M	**Metadata** includes license information under which data can be reused.
FsF-R1.2-01M	**Metadata** includes provenance information about data creation or generation.
FsF-R1.3-01M	**Metadata** follows a standard recommended by the target research community of the data.
FsF-R1.3-02D	**Data** is available in a file format recommended by the target research community.

2. Metric Specification

2.1 Globally Unique Identifier

FIELD	DESCRIPTION
Metric Identifier	FsF-F1-01D
Metric Name	**Data** is assigned a globally unique identifier.
Description	A data object may be assigned with a globally unique identifier such that it can be referenced unambiguously by humans or machines. Globally unique means an identifier should be associated with only one resource at any time. Examples of unique identifiers of data are Internationalized Resource Identifier (IRI)[12], Uniform Resource Identifier (URI) such as URL and URN, Digital Object Identifier (DOI), the Handle System, identifiers.org, w3id.org and Archival Resource Key (ARK). A data repository may assign a globally unique identifier to your data or metadata when you publish and make it available through its curation service.
FAIR Principle	F1. (Meta) data are assigned globally unique and persistent identifiers
CoreTrustSeal Alignment	R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation
ASSESSMENT	
Requirement(s)	● Data identifier (IRI, URL) ● List of globally unique identifier schemes
Method	Check if the identifier is specified based on a globally unique identifier scheme.
COMMENTS	
Related Resources: ● Identifiers compiled by FAIRsharing, https://fairsharing.org/standards/?q=&selected_facets=type_exact:identifier%20schema ● A list of Uniform Resource Identifier (URI) schemes, available in different formats, https://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml#uri-schemes-1 ● Uniform Resource Identifier (URI) Generic Syntax (RFC 3986), https://tools.ietf.org/html/rfc3986	

2.2 Persistent Identifier

FIELD	DESCRIPTION
Metric Identifier	FsF-F1-02D
Metric Name	**Data** is assigned a persistent identifier.
Description	In this specification, we make a distinction between the uniqueness and persistence of an identifier. An HTTP URL (the address of a given unique resource on the web) is globally unique, but may not be persistent as the URL of data may

[12] IRI is a generalization of URI that permits Universal Character Set.

	be not accessible (link rot problem) or the data available under the original URL may be changed (content drift problem). Identifiers based on, e.g., the Handle System, DOI, ARK are both globally unique and persistent. They are maintained and governed such that they remain stable and resolvable for the long term. The persistent identifier (PID) of a data object may be resolved (point) to a landing page with metadata containing further information on how to access the data content, in some cases a downloadable artefact, or none if the data or repository is no longer maintained. Therefore, ensuring persistence is a shared responsibility between a PID service provider (e.g., datacite) and its clients (e.g., data repositories). For example, the DOI system guarantees the persistence of its identifiers through its social (e.g., policy) and technical infrastructures, whereas a data provider ensures the availability of the resource (e.g., landing page, downloadable artefact) associated with the identifier.
FAIR Principle	F1. (Meta) data are assigned globally unique and persistent identifiers
CoreTrustSeal alignment	R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation
ASSESSMENT	
Requirement(s)	● Data identifier (IRI, URL) ● Landing page of the identifier ● List of commonly accepted persistent identifiers for data
Method	Check if the data identifier specified is based on a commonly accepted persistent identifier scheme, and it resolves to a landing page with metadata containing further information on how to access the data object. Note that this assessment method follows the current best practice to have a PID resolve to a landing page instead of its actual content.
COMMENTS	

Related Resources

- A wiki entry on persistent identifier, https://en.wikipedia.org/wiki/Persistent_identifier
- Generic PID definitions, Initial Persistent Identifier Policy for the EOSC, https://doi.org/10.5281/zenodo.3574202
- FREYA Deliverable 3.1 (Survey of Current PID Services Landscape), https://doi.org/10.5281/zenodo.1324295
- FREYA Deliverable 2.1 PID Resolution Services Best Practices, https://doi.org/10.5281/zenodo.1324299

Known Limitations/Constraints

- The assessment verifies the resolvability of the specified identifier to a landing page, but a PID may resolve to a data file or a web service response.
- A registry of persistent identifiers should provide the list of identifiers as well as associated policy documents for ensuring persistence that may be periodically reviewed and updated. If a policy document is issued with a validity period, this should be captured by the registry.
- A PID service provider may periodically check if an identifier within its registry is resolvable (e.g., https://support.datacite.org/docs/link-checker). While the PID itself may be persistent, it may not resolve to a downloadable artefact if the data or repository is no longer maintained.

2.3 Descriptive Core Metadata

FIELD	DESCRIPTION
Metric Identifier	FsF-F2-01M
Metric Name	**Metadata** includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability.
Description	Metadata is descriptive information about a data object. Since the metadata required differs depending on the users and their applications, this metric focuses on core metadata. The core metadata is the minimum descriptive information required to enable data finding, including citation which makes it easier to find data. We determine the required metadata based on common data citation guidelines (e.g., DataCite, ESIP, and IASSIST), and metadata recommendations for data discovery (e.g., EOSC Datasets Minimum Information (EDMI), DataCite Metadata Schema, W3C Recommendation Data on the Web Best Practices and Data Catalog Vocabulary). This metric focuses on domain-agnostic core metadata. Domain or discipline-specific metadata specifications are covered under metric FsF-R1.3-01M. A repository should adopt a schema that includes properties of core metadata, whereas data authors should take the responsibility of providing core metadata.
FAIR Principle	F2. Data are described with rich metadata
CoreTrustSeal Alignment	R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation
ASSESSMENT	
Requirement(s)	● Data identifier (IRI, URL) ● Machine-accessible and readable metadata
Method	Use the data identifier to access its metadata document. Parse or retrieve core metadata, e.g., through one or more options below, combine the results and then verify presence/absence of the core elements in the metadata. ● Structured data embedded in the landing page of the identifier (e.g., Schema.org, Dublin Core and OpenGraph meta tags) ● Typed Links in the HTTP Link header; for more information, see https://signposting.org/conventions/ ● If the identifier specified is a persistent identifier, use it to retrieve the metadata of the data from its PID provider, e.g., see DataCite Content Resolver at https://datacite.org/content.html
COMMENTS	
Related Resources ● Examples of metadata recommendations: ○ EOSC EDMI metadata properties, https://eosc-edmi.github.io/properties ○ W3C Recommendation Data on the Web Best Practices, https://www.w3.org/TR/dwbp/#metadata ○ W3C Data Catalog Vocabulary, https://www.w3.org/TR/vocab-dcat-2/ ● Sites that provide a list of metadata standards: ○ FAIRsharing standards, https://fairsharing.org/standards/	

- DCC List of Metadata Standards, http://www.dcc.ac.uk/resources/metadata-standards/list
- RDA Metadata Directory (based on the DCC list), http://rd-alliance.github.io/metadata-directory/

- Examples of domain agnostic metadata standards for describing research data:
  - Dublin Core Metadata Initiative (DCMI) Metadata Terms, https://www.w3.org/TR/dwbp/#bib-DCTERMS
  - DataCite Metadata Schema, https://doi.org/10.14454/7xq3-zf69
  - Schema.org, https://schema.org/Dataset
  - Data Catalog Vocabulary (DCAT), https://www.w3.org/TR/vocab-dcat-2/

Known Limitations/Constraints

- The assessment assumes that the identifier resolves to a landing page (e.g., html) that contains the metadata of the data. Landing page may not necessarily be an html page.
- Data providers may use different standards to expose the metadata of their data.
- The metadata records maintained by a data provider might not be accessible, due to, e.g., broken link of the landing page, proprietary metadata standard used, and restricted metadata.

2.4 Inclusion of Data Identifier in Metadata

FIELD	DESCRIPTION
Metric Identifier	FsF-F3-01M
Metric Name	**Metadata** includes the identifier of the data it describes.
Description	The metadata should explicitly specify the identifier of the data such that users can discover and access the data through the metadata. If the identifier specified is persistent and points to a landing page, the data identifier and links to download the data content should be taken into account in the assessment.
FAIR Principle	F3: Metadata clearly and explicitly include the identifier of the data they describe
CoreTrustSeal Alignment	R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation
ASSESSMENT	
Requirement(s)	• Data identifier (IRI, URL) • Machine-accessible and readable metadata
Method	Use the data identifier to access its metadata document. Verify if the data identifier provided is the same as the identifier specified in the metadata. Check if the identifier (link) to access data content is included in the metadata (e.g., use the metadata elements 'schema:Distribution', 'foaf:isPrimaryTopicOf' or Typed Links), and test if the content identifier is active.
COMMENTS	

Related Resources

- Signposting the Scholarly Web, https://signposting.org/conventions/

Known Limitations/Constraints

- A metadata standard may not support any element or include multiple elements through which a data identifier may be specified.
- Different practices of associating data with its metadata should be handled as part of the assessment:
- Data is assigned with an identifier that resolves to a page that contains metadata of the data. The metadata may contain the identifier and a URL to access the data (contents). In this case, the access URL should be tested.
- Data and metadata are assigned with separate identifiers. Therefore, the data identifier should be tested.

2.5 Searchable Metadata

FIELD	DESCRIPTION
Metric Identifier	FsF-F4-01M
Metric Name	**Metadata** is offered in such a way that it can be retrieved by machines.
Description	This metric refers to ways through which the metadata of data is exposed or provided in a standard and machine-readable format. Assessing this metric will require an understanding of the capabilities offered by the data repository used to host the data. Metadata may be available through multiple endpoints. For example, if data is hosted by a repository, the repository may disseminate its metadata through a metadata harvesting protocol (e.g., via OAI-PMH) and/or a web service. Metadata may also be embedded as structured data on a data page for use by web search engines such as Google and Bing or be available as linked (open) data.
FAIR Principle	F4. (Meta)data are registered or indexed in a searchable resource
CoreTrustSeal Alignment	R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation
ASSESSMENT	
Requirement(s)	● Data identifier (IRI, URL) ● Metadata provision endpoint (if it is not included in the metadata or landing page of the identifier)
Assessment	The following methods may be applied to determine if metadata of the data is accessible programmatically: ● Check if the metadata provision endpoint returns metadata records based on a request using the data identifier (see comment* below) ● Check if search engine friendly structured data is embedded in the data landing page with a proper resource type, e.g., schema.org representation of type ‘Dataset’ or ‘Collection’.
COMMENTS	
Related Resources	

- Google reference documentation on representing structured data of Dataset, https://developers.google.com/search/docs/data-types/dataset

Known Limitations/Constraints

- *Data providers may expose their metadata through different ways, e.g., OAI-PMH, REST API using JSONAPI specification, and Catalog Service for the Web (CSW). Their endpoints (URLs) should be machine discoverable and accessible. The metadata access endpoints of a repository can be found through FAIRsharing and re3data. However, at present, it is not possible to programmatically discover the metadata endpoints of a repository based on a data identifier, unless they are explicitly specified in the metadata or the landing page of the data. Mapping the client ids from DataCite's PID service to re3data identifiers is in progress and might provide a starting point for the assessment.
- Structured data may be provided in different formats, JSON-LD, RDFa or Microdata. The variety of formats should be handled as part of the assessment.
- The assessment only verifies if structured data is present on the data landing page with a proper type (e.g., Dataset or Collection). Embedding structured data does not guarantee that the data will be present on search results. To verify that the data is findable through a web search engine, we should perform a search through the search engine API based on the data identifier and its descriptive metadata (e.g., title, author). 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.

2.6 Data Access Information

FIELD	DESCRIPTION
Metric Identifier	FsF-A1-01M
Metric Name	**Metadata** contains access level and access conditions of the data.
Description	This metric determines if the metadata includes the level of access to the data such as public, embargoed, restricted, or metadata-only access and its access conditions. Both access level and conditions are necessary information to potentially gain access to the data. It is recommended that data should be as open as possible and as closed as necessary. ● There are no access conditions for public data. Datasets should be released into the public domain (e.g., with an appropriate public-domain-equivalent license such as Creative Commons CC0 license) and openly accessible without restrictions when possible. ● Embargoed access refers to data that will be made publicly accessible at a specific date. For example, a data author may release their data after having published their findings from the data. Therefore, access conditions such as the date the data will be released publically is essential and should be specified in the metadata. ● Restricted access refers to data that can be accessed under certain conditions (e.g. because of commercial, sensitive, or other confidentiality reasons or the data is 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 restricted data, the metadata should include the conditions of access to the data such as point of contact or instructions to access the data.

	● Metadata-only access refers to data that is not made publicly available and for which only metadata is publicly available.
FAIR Principle	A1: (Meta)data are retrievable by their identifier using a standardized communication protocol **Note**: This metric is about ensuring provision of metadata related to data access. This metadata is important to retrieve data using a standardized communication protocol, thus we mapped it to the principle A1.
CoreTrustSeal Alignment	R2. The repository maintains all applicable licenses covering data access and use and monitors compliance R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community
ASSESSMENT	
Requirement(s)	● Data identifier (IRI, URL) ● Machine-accessible and readable metadata
Assessment	Use the data identifier to access its metadata document. Check 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.
COMMENTS	

Related Resources

- Public domain licenses, https://creativecommons.org/share-your-work/public-domain
- EU Vocabulary on access rights, https://op.europa.eu/en/web/eu-vocabularies/at-dataset/-/resource/dataset/access-right
- 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 (test vocabulary, not yet available through the production metadata registry), http://sandbox.metadataregistry.org/concept/list/vocabulary_id/251.html
- Eprints Access Rights Vocabulary Encoding Scheme, http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_AccessRights_Vocabulary_Encoding_Scheme

Known Limitations/Constraints

- The metadata standard considered as part of the assessment may not include all of the elements for representing data access levels and related access information. The access information may be expressed in an unstructured manner, e.g., as a 'comment' in the metadata document.
- The assessment of this metric only checks the metadata of access restrictions, but it does not validate if the access conditions specified are correct.
- The assessment should be complemented with the evaluation of the data access mechanism based on the specified access levels, e.g., data is not accessible, accessible in a semi-automated (mediated access to data via data custodian), or automated fashion.
- A data object may consist of several files with different access levels; some are with open access while others are with restricted access. So mixed access levels may apply to the object.

2.7 Metadata Preservation

FIELD	DESCRIPTION
Metric Identifier	FsF-A2-01M
Metric Name	**Metadata** remains available, even if the data is no longer available.
Description	This metric determines if the metadata will be preserved even when the data they represent are no longer available, replaced or lost. Similar to metric FsF-F4-01M, answering this metric will require an understanding of the capabilities offered, data preservation plan and policies implemented by the data repository and data services (e.g., Datacite PID service). Continued access to metadata depends on a data repository's preservation practice which is usually documented in the repository's service policies or statements. A trustworthy data repository offering DOIs and implementing a PID Policy should guarantee that metadata will remain accessible even when data is no longer available for any reason (e.g., by providing a tombstone page)
FAIR Principle	A2. Metadata should be accessible even when the data is no longer available
CoreTrustSeal Alignment	R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way
ASSESSMENT	
Requirement(s)	--
Assessment	Programmatic assessment of the preservation of metadata of a data object can only be tested if the object is deleted or replaced. So this test is only applicable for deleted, replaced or obsolete objects. Importantly, continued access to metadata depends on a data repository's preservation practice. Therefore, we regard that the assessment of metric applies to at the level of a repository, not at the level of individual objects. For this reason, we excluded its assessment details from this specification. Depending on the supported persistent identifier type, some metadata may be by default preserved in a registry maintained by a PID provider (e.g. datacite). In addition to a repository's preservation policy or statement, exchange protocol may indicate the status of records in an archive. For instance, OAI-PMH harvesting protocol which offers a field to declare one of three levels (no, persistent, and transient) of support for deleted records.
COMMENTS	
Related Resources ● DMPonline, https://dmponline.dcc.ac.uk/public_plans ● DMP Common Standards WG, https://www.rd-alliance.org/groups/dmp-common-standards-wg ● ezDMP, https://ezdmp.org/index ● Best Practices for offering tombstone pages, https://support.datacite.org/docs/tombstone-pages **Known Limitations/Constraints** ● Data preservation statements are usually found in a repository's data policy or other governance documents. Machine-actionable representation of preservation policies in repository catalogues and registries such as re3data is important to enable an automated assessment of the statements. Further work in this areas is needed, for example to enable data producers to receive repository	

recommendations, based on preservation requirements expressed in machine-actionable DMPs, e.g. Oblasser et al 2020 http://dx.doi.org/10.2218/ijdc.v15i1.704

- Currently, PID providers (e.g., DataCite) do not offer any tombstone pages automatically for unavailable objects. Data providers may maintain the pages instead, for example https://doi.pangaea.de/10.1594/PANGAEA.715333

2.8 Formal Representation of Metadata

FIELD	DESCRIPTION
Metric Identifier	FsF-I1-01M
Metric Name	**Metadata** is represented using a formal knowledge representation language.
Description	Knowledge representation is vital for machine-processing of the knowledge of a domain. Expressing the metadata of a data object using a formal knowledge representation will enable machines to process it in a meaningful way and enable more data exchange possibilities. Examples of knowledge representation languages are RDF, RDFS, and OWL. These languages may be serialized (written) in different formats. For instance, RDF/XML, RDFa, Notation3, Turtle, N-Triples and N-Quads, and JSON-LD are RDF serialization formats.
FAIR Principle	I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation **Note**: The I1 principle loosely defines the use of knowledge representation. Therefore, we define two metrics corresponding to the principle concerning metadata. The metric FsF-I1-01M focuses on making the metadata available for machine-mediated interpretation, whereas the metric FsF-I1-02M focuses on the use of semantic resources to enrich metadata.
CoreTrustSeal Alignment	R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community
ASSESSMENT	
Requirement(s)	• Data identifier (IRI, URL) • Metadata provision endpoint (e.g., SPARQL endpoint)
Assessment	Machine-actionable representation (e.g., RDF) of the metadata may be retrieved as follows: • If content negotiation is supported, use the identifier to perform a request, e.g., an RDF-based document. • Use the 'typed links' given in the HTML header section of the landing page to access the RDF-based metadata of the data, e.g., https://data.gov.lv/dati/lv/dataset/covid-19 • Query the SPARQL endpoint using the identifier (or optionally title) of the data, for example by using metadata elements from dcterms and dcat standards. Perform a full text-search within the SPARQL query if it is supported.

COMMENTS
Related Resources ● RDF MIME types by serializer, http://librdf.org/raptor/api/raptor-formats-types-by-serializer.html ● SPARQL Protocol for RDF, https://www.w3.org/TR/rdf-sparql-protocol/ ● Best Practice Recipes for Publishing RDF Vocabularies, https://www.w3.org/TR/swbp-vocab-pub/ ● Community-defined models and formats via FAIRsharing, https://fairsharing.org/standards/?q=&selected_facets=type_exact:model/format **Known Limitations/Constraints** ● Based on a data identifier, it is not possible to programmatically discover the SPARQL endpoint provided by a data repository, unless the endpoint information is specified in the repository metadata, e.g., https://www.re3data.org/repository/r3d100012203 ● The RDF-based metadata may not be supported by the data repository which curates the data, but it may be available through external linked data repositories, e.g., bio2rdf. ● RDF data may be serialized in a number of different ways. Therefore, the variety of serialization formats (and their respective MIME types) should be considered when performing the SPARQL query.

2.9 Metadata with Semantic Resources

FIELD	DESCRIPTION
Metric Identifier	FsF-I1-02M
Metric Name	**Metadata** uses semantic resources.
Description	A metadata document or selected parts of the document may incorporate additional terms from semantic resources (also referred as semantic artefacts) that unambiguously describe the contents so they can be processed automatically by machines. This metadata enrichment may facilitate enhanced data search and interoperability of data from different sources. Ontology, thesaurus, and taxonomy are kinds of semantic resources, and they come with varying degrees of expressiveness and computational complexity. Knowledge organization schemes such as thesaurus and taxonomy are semantically less formal than ontologies.
FAIR Principle	I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
CoreTrustSeal Alignment	R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community
ASSESSMENT	
Requirement(s)	● Data identifier (IRI, URL) ● Metadata provision endpoint (SPARQL endpoint) ● Machine-accessible and readable metadata ● Registry of semantic resources

Assessment	This assessment is the continuation of the assessment FsF-I1-01M, but focuses on the metadata contents. ● Extract namespaces declared from the machine-actionable metadata document. Filter out common namespaces (e.g., rdf, rdfs, xsd, owl). ● Compare the remaining namespaces with entries from existing (known) ontology registries (see examples listed in Related Resources).
COMMENTS	

Related Resources

- Publishing and consuming Linked Data embedded in HTML, https://www.w3.org/2001/sw/interest/ldh/
- Examples of repositories or look-up services for semantic resources (the list is not exhaustive):
  - Linked Open Vocabularies (LOV), https://lov.linkeddata.es/dataset/lov
  - OBO Foundry, http://www.obofoundry.org/
  - BioPortal, https://bioportal.bioontology.org/
  - Basel Register of Thesauri, Ontologies & Classifications (BARTOC), https://bartoc.org/
  - NERC Vocabulary Server, https://vocab.nerc.ac.uk/
  - Research Vocabulary Australia, https://vocabs.ands.org.au/
  - MMI Ontology Registry and Repository (ORR), https://mmisw.org/
  - Industrial Ontologies Foundry (IOF), https://www.industrialontologies.org/
  - CESSDA Vocabulary Service, https://vocabularies.cessda.eu/

Known Limitations/Constraints

- The assessment checks the inclusion of semantic markup in the metadata page, not their contents and quality, e.g., if the terms used are in appropriate context and accessible over the web.
- There is no up-to-date, maintained, cross domain ontology catalogue, registry or ontology library available.
- It is hard to verify if the metadata uses FAIR vocabularies as the criteria defining a FAIR vocabulary have not fully developed and recommended yet.

2.10 Links to Related Entities

FIELD	DESCRIPTION
Metric Identifier	FsF-I3-01M
Metric Name	**Metadata** includes links between the data and its related entities.
Description	Linking data to its related entities will increase its potential for reuse. The linking information should be captured as part of the metadata. A dataset may be linked to its prior version, related datasets or resources (e.g. publication, physical sample, funder, repository, platform, site, or observing network registries). Links between data and its related entities should be expressed through relation types (e.g., DataCite Metadata Schema specifies relation types between research objects through the fields 'RelatedIdentifier' and 'RelationType'), and preferably use persistent Identifiers for related entities (e.g., ORCID for contributors, DOI for publications, and ROR for institutions).

FAIR Principle	I3. (Meta)data include qualified references to other (meta)data
CoreTrustSeal Alignment	R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations
ASSESSMENT	
Requirement(s)	• Data identifier (IRI, URL) • Machine-accessible and readable metadata
Assessment	Use the data identifier to access its metadata record. Check the metadata elements which indicate the relationship between data and related entities. Test if the URLs of the related entities are active (not broken links).
COMMENTS	

Related Resources

- DataCite Metadata Working Group. (2019). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.3, https://doi.org/10.14454/7xq3-zf69
- Link Relation Types, https://www.iana.org/assignments/link-relations/link-relations.xhtml

Known Limitations/Constraints

- Different metadata schemas may use different properties to specify the relation between data and its related entities.
- The assessment regards any relation between a data and its related entities as success. It does not consider the quantity or types of relations.
- Links to related resources are not necessarily expressed as actionable links but may also be strings such as citations.

2.11 Metadata of Data Content

FIELD	DESCRIPTION
Metric Identifier	FsF-R1-01MD
Metric Name	**Metadata** specifies the content of the data.
Description	This metric evaluates if the content of the dataset is specified in the metadata, and it should be an accurate reflection of the actual data deposited. Examples of the properties specifying data content are : resource type (e.g., data or a collection of data), variable(s) measured or observed, method, data format and size. Ideally, ontological vocabularies should be used to describe data content (e.g., variable) to support interdisciplinary reuse.
FAIR Principle	R1: (Meta)data are richly described with a plurality of accurate and relevant attributes **Note**: Data quality aspect is not explicitly addressed by FAIR principles. However, an accurate description of the data content is important for assessing the quality of the data. We regard the properties of data content as part of rich metadata, therefore we map this metric to its closest principle R1.

CoreTrustSeal Alignment	R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations
ASSESSMENT	
Requirement(s)	● Data Identifier ● Machine-accessible and readable metadata ● Data file(s)
Assessment	Use the data identifier to access its metadata document. Verify the presence/absence of elements representing data content descriptions in the metadata document. Use the data access URL specified in the metadata to retrieve the actual data. Check if ontology terms are used to describe data content. Compare the content descriptions found with actual data properties (see comment* below).
COMMENTS	

Related Resources

- Frictionless Data, https://frictionlessdata.io/
- CSV on the Web: A Primer, https://www.w3.org/TR/tabular-data-primer/
- Apache Tika (an example of content analysis toolkit), https://tika.apache.org/

Known Limitations/Constraints

- *The proposed assessment has some general limitations and some cases where future expansion is dependent on contexts:
    - Descriptors (mandatory and optional properties of a schema) may influence metadata completeness.
    - Validation of descriptor content is beyond the scope of this test as it would depend on human judgement.
    - A detailed assessment of data files properties would depend on some agreed mechanism for defining and agreeing domain requirements.
- General-purpose metadata standards such as Datacite Metadata Schema and Schema.org provide elements to represent content descriptions. Thus, it is possible to check programmatically if the descriptions required are present in the metadata. However, the conformance/matching test may become a challenge due to a variety of data types and data size. Standardized tabular data and self-describing data formats (e.g., HDF, NetCDF, Parquet) are promising, but not the solution to every research domain. Another challenge is that unstructured content descriptions might be included in a data file; fuzzy text-matching algorithms can be useful here.

2.12 Data Usage License

FIELD	DESCRIPTION
Metric Identifier	FsF-R1.1-01M
Metric Name	**Metadata** includes license information under which data can be reused.
Description	This metric evaluates if data is associated with a license because otherwise users cannot reuse it in a clear legal context. We encourage the application of licenses for all kinds of data whether public, restricted or for specific users. Without an explicit license, users do not have a clear idea of what can be done with your data. Licenses can be of standard type (Creative Commons, Open Data Commons Open Database License) or bespoke licenses, and rights statements 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.
FAIR Principle	R1.1. (Meta)data are released with a clear and accessible data usage license
CoreTrustSeal Alignment	R2. The repository maintains all applicable licenses covering data access and use and monitors compliance
ASSESSMENT	
Requirement(s)	● Data identifier (IRI, URL) ● Machine-accessible and readable metadata
Assessment	Use the data identifier to access its metadata document. Verify the presence/absence of metadata element(s) corresponding to license information of the data. The license information (e.g., name or URI) may be used to request additional information (e.g., OSI approved) from an external license registry (e.g., SPDX).
COMMENTS	

Related Resources

- SPDX license registry, https://spdx.org/licenses/
- Rights statements of cultural heritage objects, https://rightsstatements.org/page/1.0/?language=en
- ARDC Data Rights Management Guide, https://ardc.edu.au/guides/research-data-rights-management
- The Landscape of Rights and Licensing Initiatives for Data Sharing, https://doi.org/10.5334/dsj-2019-029
- Open Digital Rights Language (ODRL), https://www.w3.org/TR/odrl-model/
- Creative Commons, https://creativecommons.org/
- Creative Commons Rights Expression Language, https://creativecommons.org/ns

Known Limitations/Constraints

- The assessment checks if the license information is provided as part of the metadata. It does not validate if the specified license is the most appropriate license for the data. There may be quite specific circumstances related to the data that cannot be explicitly expressed in the metadata as to why a license was chosen.

2.13 Data Provenance

FIELD	DESCRIPTION
Metric Identifier	FsF-R1.2-01M
Metric Name	**Metadata** includes provenance information about data creation or generation.
Description	Data provenance (also known as lineage) represents a dataset's history, including the people, entities, and processes involved in its creation, management and longer-term curation. It is essential that data producers provide provenance information about the data to enable informed use and reuse. The levels of provenance information needed can vary depending on the data type (e.g., measurement, observation, derived data, or data product) and research domains. For that reason, it is difficult to define a set of finite provenance properties that will be adequate for all domains. Based on existing work, we suggest that the following provenance properties of data generation or collection are included in the metadata record as a minimum. ● Sources of data, e.g., datasets the data is derived from and instruments ● Data creation or collection date ● Contributors involved in data creation and their roles ● Data publication, modification and versioning information There are various ways through which provenance information may be included in a metadata record. Some of the provenance properties (e.g., instrument, contributor) may be best represented using PIDs (such as DOIs for data, ORCIDs for researchers). This way, humans and systems can retrieve more information about each of the properties by resolving the PIDs. Alternatively, the provenance information can be given in a linked provenance record expressed explicitly in, e.g., PROV-O or PAV or Vocabulary of Interlinked Datasets (VoID).
FAIR Principle	R1.2. (Meta)data are associated with detailed provenance
CoreTrustSeal Alignment	R7. The repository guarantees the integrity and authenticity of the data
ASSESSMENT	
Requirement(s)	● Data identifier (IRI, URL) ● Machine-accessible and readable metadata
Assessment	Use the data identifier to access its metadata record. Verify the presence/absence of metadata element(s) corresponding to the minimum data provenance properties. ● Presence of basic 'proxy' metadata elements related to data creation (creator, contributors, date, and version, modification date, etc.) ● Presence of process indicator, e.g. dc:source or relation type (isVersionOf, isBasedOn, isFormatOf) addressed in FsF-I3-01M. ● Presence of PROV-O or PAV information in RDFa microformats (landing page) or in RDF metadata.
COMMENTS	
Related Resources	

- PROV Model Primer, https://www.w3.org/TR/prov-primer/
- Dublin Core to PROV Mapping, https://www.w3.org/TR/prov-dc/
- Checklist for Evaluation of Dataset Fitness for Use produced by the WDS/RDA Assessment of Data Fitness for Use WG, https://www.rd-alliance.org/system/files/DataFitnessForUse_ChecklistForm_v2_20181218_RDADistribution.pdf
- W3C Recommendation Data on the Web Best Practices (8.4 Data Provenance), https://www.w3.org/TR/dwbp/#metadata
- PROV-O as RDFa, https://www.w3.org/2011/prov/wiki/PROV-O_as_RDF
- OPMV, the Open Provenance Model Vocabulary, http://purl.org/net/opmv/ns
- Business Process Model and Notation, https://www.omg.org/spec/BPMN/
- PAV- Provenance, Authoring and Versioning ontology: https://pav-ontology.github.io/pav/

Known Limitations/Constraints

- The proposed minimum provenance properties are not final; new properties may be incorporated into the assessment if the requirement emerges. Properties such as processes/methods (incl. model, instrument, etc.) used in the data creation depend on domain standards.
- We regard references to related works (scholarly articles, data papers, preceding or associated data) as useful provenance information. This property of provenance is considered as part of FsF-I3-01M, therefore we excluded it from the assessment.
- Metadata may include a specific element (e.g., dcmi:provenance) and/or 'proxy' elements (e.g., datacite:Contributor, schema.org:measurementTechnique) to convey data provenance.
- Data may be published at different analysis stages (raw, processed, derivative, product). The completeness of the provenance information may depend on the stage at which the data is published.

2.14 Community Metadata Standard

FIELD	DESCRIPTION
Metric Identifier	FsF-R1.3-01M
Metric Name	**Metadata** follows a standard recommended by the target research community of the data.
Description	In addition to core metadata required to support data discovery (covered under metric FsF-F2-01M), metadata to support data reusability should be made available following community-endorsed metadata standards. Some communities have well-established metadata standards (e.g., geospatial: ISO19115; biodiversity: DarwinCore, ABCD, EML; social science: DDI; astronomy: International Virtual Observatory Alliance Technical Specifications) while others have limited standards or standards that are under development (e.g., engineering and linguistics). The use of community-endorsed metadata standards is usually encouraged and supported by domain and discipline-specific repositories.
FAIR Principle	R1.3. (Meta)data meet domain-relevant community standards
CoreTrustSeal Alignment	R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data

ASSESSMENT	
Requirement(s)	• Data identifier (IRI, URL) • Metadata provision endpoints including SPARQL endpoint
Assessment	Gather all metadata standards used by a data repository; this list can be requested, e.g., from the metadata endpoint (e.g., OAI-PMH). Filter out domain-agnostic standards (e.g., Datacite Metadata Schema, Dublin Core, Schema.org) from the list. Cross check the remaining standards with an external metadata registry, e.g., RDA Metadata Standards Catalog. Request metadata of the data identifier specified based on one of the remaining standards as a test case (see comment* below).

COMMENTS

Related Resources

Examples of the metadata standards with subject areas:

- RDA Metadata Standards Catalog, https://rdamsc.bath.ac.uk/
- FAIRsharing, https://fairsharing.org/standards/
- OAI-PMH Data Provider Validation and Registration, https://www.openarchives.org/Register/ValidateSite
- OAI-PMH Tools, https://www.openarchives.org/pmh/tools/
- Metadata standards supported by a repository may be available through re3data, https://www.re3data.org/

Known Limitations/Constraints

- *The data identifier provided (e.g., PID) may not be the same as the identifier used in the metadata record harvested. For example, in OAI-PMH, the nature of a record identifier is outside the scope of the harvesting protocol; for more information, see http://www.openarchives.org/OAI/openarchivesprotocol.html#UniqueIdentifier
- The assessment focuses on a specific metadata harvesting protocol. It might not be supported by all data repositories.
- Future evaluation of the metric should also consider the extent to which the metadata of a dataset reflects the community-endorsed metadata standard.
- Some of these discipline-specific standards might not be properly formalized so an automatic validation of the metadata based on the standards can be problematic. External tools might be necessary to check compliance with metadata standards.

2.15 Data File Format

FIELD	DESCRIPTION
Metric Identifier	FsF-R1.3-02D
Metric Name	**Data** is available in a file format recommended by the target research community.
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 file format that is backed by the research community to enable data sharing and reuse. Consider for example, file formats

	that are widely used and supported by the most commonly used software and tools. These formats also should be suitable for long-term storage and archiving, which are usually recommended by a data repository. The formats not only give a higher certainty that your data can be read in the future, but they will also help to increase the reusability and interoperability. Using community-endorsed formats enables data to be loaded directly into the software and tools used for data analysis. It makes it possible to easily integrate your data with other data using the same preferred format. The use of community-endorsed formats will also help to transform the format to a newer one, in case an older format gets outdated.
FAIR Principle	R1.3. (Meta)data meet domain-relevant community standards
CoreTrustSeal Alignment	R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community
ASSESSMENT	
Requirement(s)	● Data identifier (IRI, URL) ● Machine-accessible and readable metadata
Assessment	Extract file format information (mime-type) from the metadata based on elements, e.g., datacite:Format, schema.org: fileFormat, dc:format. Check if the format is an open and long-term file format (see comment* below).
COMMENTS	

Related Resources

- A list of commonly used as well as domain specific scientific file formats
  - http://justsolve.archiveteam.org/index.php/Scientific_Data_formats
  - https://en.wikipedia.org/wiki/List_of_file_formats#Scientific_data_(data_exchange)
- Examples of recommended file formats based on data types, https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats.aspx
- PRONOM file format registry, https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=new
- A recommended format statement by the US Library of Congress, https://www.loc.gov/preservation/resources/rfs
- Long-term file formats: ISO/TR 22299. Document management - Digital file format recommendations for long-term storage, https://www.iso.org/standard/73117.html
- File type support lists provided by open source and commercial statistics (e.g. https://de.mathworks.com/help/matlab/import_export/supported-file-formats.html) or spreadsheet processing software vendors (e.g. https://support.microsoft.com/en-us/office/file-formats-that-are-supported-in-excel-0943ff2c-6014-4e8d-aaea-b83d51d46247?ui=en-us&rs=en-us&ad=us).

Known Limitations/Constraints

- *At present, there is a lack of reference resources (registries) against which a file format test can be checked programmatically. Common file formats endorsed by communities are not available through a registry but on static web pages (see resources above). This is an issue for the scientific community as a whole. Further work is needed to develop a standard approach to defining which formats are

open and suitable for long-term preservation and use and managing those community-specific lists over time.

- Not all data can be made available in an open, non-proprietary, widely supported format, such as most 3D data, CAD data, dynamic spreadsheets or databases with specific significant characteristics which cannot be exported.
- Standard formats in earth system modeling (atmosphere, ocean) are netCDF and GRIB. GRIB is used for internal storage rather than for publication.
- Commonly used community file formats are not necessarily very domain specific. Some very generic file formats for e.g. spreadsheets are widely used by the scientific community.
- Data files may be made available using an archive file format (e.g., *.zip). In addition to the archive format, the actual file formats should be specified in the metadata such that machines can extract/unzip the downloaded file and read the actual files programmatically.
- Many scientific formats do not have an associated mime-type (e.g. BUFR), thus are hard to detect.