Project Title	Expanding FAIR solutions across EOSC
Project Acronym	FAIR-IMPACT
Grant Agreement No.	101057344
Start Date of Project	2022-06-01
Duration of Project	36 months
Project Website	fair-impact.eu

[Version for community input]

M5.2 - Guidelines for repositories and registries on exposing repository trustworthiness status and FAIR data assessments outcomes

Work Package	**WP5 - Metrics, certification, and guidelines**
Lead Author (Org)	**Maaike Verburg (KNAW-DANS)**
Contributing Author(s) (Org)	**Robert Ulrich (KIT), Hervé L'Hours (UKDS), Robert Huber (UBremen), Mike Priddy (KNAW-DANS), Joy Davidson (UEDIN-DCC), Alejandra Gonzalez-Beltran (UKRI-STFC), Gabriela Meijas (DataCite), Charlotte Neidiger (KIT)**
Date	**2023-10-31**
Version	**V1.0**

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Funded by the European Union

Versioning and contribution history

Version 1.0 of this Deliverable can be found on Zenodo at:

> Verburg, M., Ulrich, R., L'Hours, H., Huber, R., Priddy, M., Davidson, J., Gonzalez-Beltran, A., Meijas, G., & Neidiger, C. (2023). M5.2 - Guidelines for repositories and registries on exposing repository trustworthiness status and FAIR data assessments outcomes (1.0). Zenodo. https://doi.org/10.5281/zenodo.10058634

This version of the document is intended to invite community input on this version of the guidelines for transparency. Please provide your input through suggestions or comments. The authors of the document will review the input to advance their work further. Your input is much appreciated!

Disclaimer

FAIR-IMPACT has received funding from the European Commission's Horizon Europe funding programme for research and innovation programme under the Grant Agreement no. 101057344. 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.

Executive Summary

This report presents a first iteration of guidelines that will help to expose relevant information at the organisational and object level to facilitate discovery, provide context, and support interoperability between repositories, registries, and other related stakeholders. By improving the flow of information between stakeholders more trusted perceptions and relationships about the services provided and objects held or referred to by them can be fostered. The guidelines will focus on exposing information about characteristics (e.g. repository name, contact information), information that can inform a sense or status of trust (e.g. a certificate, preservation policy), and information relating to FAIR assessment (e.g. assessment results, tool(s) used).

Supporting this first list of guidelines, the rationale, methodology, scope, and added value scenarios are all discussed to place the creation of the guidelines in their full context. The report also covers the plans around the further development of the guidelines and the prototype that will implement them, to offer a full picture of the envisioned line of work. The main aim of this report is to continue to invite community-input on the guidelines presented, as it is of great importance that the guidelines are relevant to the community that would implement them. Continuous opportunity for providing feedback on the document (presented on the FAIR-IMPACT website), as well as targeted outreach activities from the project will result in future iterations of the guidelines, and reports on the prototype explorations will depict a practical implementation of the model.

The main stakeholders of interest for these guidelines are repositories, registries, and all information providers (e.g. depositors) and information consumers (e.g. harvesters, funders, policy makers) that engage with them. Throughout the document, the term ‘repositories’ is used to reference digital repositories and repositories exposing any information about their holdings digitally. Similarly, the ‘objects’ discussed also include both (born) digital objects and all objects that have information about them exposed digitally. The term ‘service’ in the document refers to the provision of a facility to meet the needs of (human or machine) users. Service providers can indicate a variety of organisations, including registries and repositories, which will be the main focus of the guidelines. The guidelines focus on how information can be exposed and shared between these different stakeholders in a way that is understandable and useful to both humans and machines. Information will be exposed on both the organisational level as well as the object level, to support a meaningful and relevant presentation of information.

Funded by the European Union

Table of Contents

Funded by the European Union

Terminology

Terminology/Acronym	Description
CESSDA	Consortium of European Social Science Data Archives
DCAT	Data Catalog Vocabulary
DDI	Data Documentation Initiative
DMP	Data Management Plan
DQV	Data Quality Vocabulary
DRAWG	Data Repositories Attributes Working Group. A working group from the Research Data Alliance (RDA).
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
FIP	FAIR Implementation Profile
HTML	HyperText Markup Language
ISO	International Organization for Standardization
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OGC	Open Geospatial Consortium
ORCiD	Open Researcher and Contributor IDentifier
PID	Persistent IDentifier
RDF	Resource Description Framework
TDR	Trustworthy Digital Repository
W3C	World Wide Web Consortium

Funded by the European Union

1 Introduction

The FAIR-IMPACT project supports the implementation of FAIR-enabling practices, tools, and services. To this end, guidelines and a prototype will be developed to improve the transparency of, and trust in, repositories. These guidelines help to expose relevant information as metadata at the organisational and object level to facilitate discovery, provide context, and support interoperability. The guidelines will also recommend accompanying evidence in a uniform and transparent way, to create a sense of trust in the services and their providers. This Milestone report presents an initial set of these guidelines, informed by community input on the previously presented methodology and recommendations[1]. To ensure continued relevance to the community both the guidelines, and the prototype that will implement them, will undergo multiple iterations.

Specifically, these guidelines will focus on exposing information relating to:

- the trustworthiness of a repository and the transparent exposure of information that helps to inform such a status,
- the FAIRness of the digital objects held by the repository, in the form of assessment results and required information for FAIR assessment.

This Milestone report will present an update on the methodology, the first set of guidelines, and the next steps. This document will reiterate the most important and relevant parts of its predecessor with minor or major changes based on subsequent developments and community feedback, while also referring the reader back to the document on occasion for more extensive considerations or descriptions.

1.1 Role of the milestone

This Milestone report presents the first set of the guidelines as introduced above. These guidelines will increase the transparency and availability of information about repositories, and the objects they hold. The creation of these guidelines will facilitate the creation of the prototype, implementing and exposing FAIR assessments for data and code. The guidelines and prototype will be tested and then further refined and iterated. The specific guidelines and the way they are presented may change over the course of the project. At the end of the project, the most up to date version of the guidelines will be presented as a final reference point.

1.1.1 Means of verification

The Milestone report is currently publicly available through its publication on Zenodo and displayed on the FAIR-IMPACT website. The guidelines will be open for continuous community input via the FAIR-IMPACT website and specific outreach activities will also focus on collecting input from specific stakeholders and potential users.

[1] Maaike Verburg, Hervé L'Hours, Robert Huber, Robert Ulrich, Mike Priddy, Joy Davidson, & Alejandra Gonzalez-Beltran. (2023). Introduction to the guidelines for repositories and registries on exposing repository trustworthiness status and FAIR data assessments outcomes (0.5). Zenodo. https://doi.org/10.5281/zenodo.8224359

Funded by the European Union

2 Background and rationale

FAIR-IMPACT supports the implementation of FAIR-enabling practices, tools, and services to support a FAIR European Open Science Cloud (EOSC). The actualisation of the EOSC depends on FAIR data and a FAIR-enabling landscape[2], and requires the translation of FAIR into guidelines and frameworks to make them applicable and relevant to the wider community and facilitate interconnectedness and communication. The work presented here improves connections in the FAIR-enabling landscape by exposing relevant information in a transparent and uniform manner to improve and simplify how it can be expressed, interpreted, and shared.

This section of the report focuses on the context of the current landscape and the different considerations in terms of scoping and development that have resulted in the first set of guidelines, which are presented in Section 3.

2.1 The current landscape

The current research data environment is populated with digital objects, each with a range of characteristics, including those that, if exposed transparently and understandably, can inform perceptions and decisions related to FAIRness and trustworthiness. Complex collaborative coordination is required by all (human) actors and (machine) agents involved to carry out the processes and activities required throughout the object lifecycle to enable FAIRness. Providing and maintaining FAIR data for the long term depends on repositories that are prepared to take action when user needs or technologies change over time[3].

The ability to cooperate, interoperate, and deliver services to researchers depends on mutual trust between these different stakeholders. This trust can be fostered through the transparent sharing and exchange of relevant information about objects and the organisations that care for them. Transparent and linked information between the **objects**, the **repositories** that hold or interact with them, and the **registries** that use this information, can be indexed, harvested, reused, and potentially validated against agreed criteria or by designated '**validation authorities**'. These are authoritative entities that can logically be called on to provide the validation of the assertion(s) made in the exposed metadata (e.g. when a repository asserts that they are CoreTrustSeal[4] certified, the CoreTrustSeal Board would be the validation authority). Much of the information that needs to be made transparently available to inform trust is currently entered into multiple different services, registries, and identity providers, making it complex to navigate and collect. Figure 1 provides a specific perspective on organisations and objects. It aims to outline the desired connections between the entities addressed by these guidelines. Each stakeholder has varying perspectives on the flow of information and is interested in a different subset of

[2] https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en
[3] Hervé L'Hours, Recker, J., & Kleemola, M. (2023). CoreTrustSeal+FAIRenabling: Alignment between the FAIR Principles and CoreTrustSeal 2023-25 (01.00). Zenodo. https://doi.org/10.5281/zenodo.7564703
[4] https://www.coretrustseal.org/

Funded by the European Union

connections and metadata elements depending on their goals and activities. The guidelines are relevant to a wide range of (meta)data services, including registries, but the immediate organisational focus of the project is on repositories. The digital objects could be any item including data, metadata, software, and semantic artefacts, but the immediate focus of this report is on research data and will be extended to software in further iterations (see Section 2.3 on Scoping).

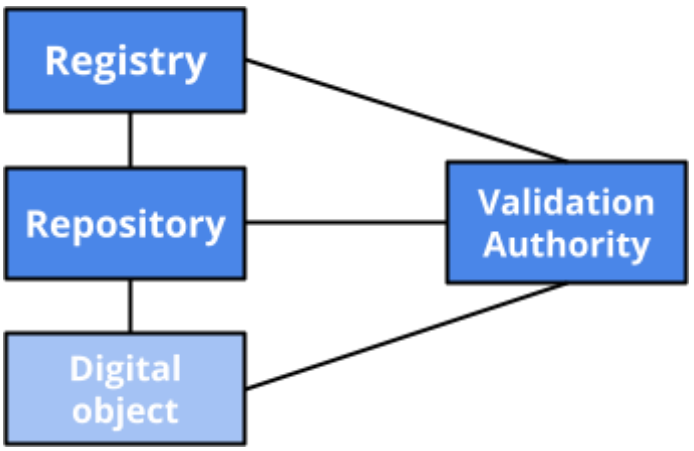

Figure 1 - The main interactions of relevance to the guidelines and prototype.

The guidelines will focus on the exposure of information pertinent to the trustworthiness and certification status of a repository[5], as well as the FAIR related qualities of digital objects, including the presentation of assessment results. These are areas of importance to a FAIR-enabling landscape built on mutual trust, that can help all stakeholders to better understand each other and to specify measures that increase interoperability. Analysis of the current landscape identifies the starting point from which these developments can be designed and implemented. Our preceding publication[1] goes into more depth on the current landscape and complexities in these different areas. This section presents a condensed overview of the current landscape and the recommendations derived from this for the guidelines, enriched by feedback from the community[6].

Exposing characteristics of repositories and objects

Finding and assessing digital objects and repositories depends on their unambiguous identification and discovery. Our analyses of the current landscape show some issues and inconsistencies in the way that digital objects, datasets, and the web pages of repository data catalogues identify and categorise themselves. This complicates machine-based evaluation, including those relating to trustworthiness and quality. For example, if a

[5] Andreu, T., Anglada, L., Antos, D., Bähr, T., Brzeźniak, M., Burgi, P.-Y., Cavet, C., Celjak, D., Crépé-Renaudin, S., De Loof, C., Dillo, I., Dubois, O., Fernandes, R., Forsström, P.-L., Ganis, G., Gibney, E., Holl, A., L'Hours, H., Lamers, D., … Wyns, R. (2023). EOSC Preservation: Overview Discussion Paper. Zenodo. https://doi.org/10.5281/zenodo.7516259

[6] The initial publication was open to community feedback July-September 2023, accessible via the FAIR-IMPACT website: https://fair-impact.eu/fair-impact-materials-community-review Specific community members were also contacted directly to share input.

Funded by the European Union

repository holds different types of resources (datasets, software, publications), the appropriate evaluation metrics must be used to assess their FAIRness. Using the 23 CESSDA[7] repository websites, a preliminary analysis demonstrates how archives identify their homepages or catalogue pages via embedded metadata. Standard categories for identifying these archive's offerings such as DCAT[8,9], schema.org[10] or Dublin Core[11] terms representing data catalogues are not used at all and instead mostly only generic terms such as 'website', 'article' or 'blog' are used. An, as yet unpublished, dataset consisting of a high number (> 11M) of digital objects' metadata (via OAI-PMH interfaces listed by re3data[12]) demonstrated how data repositories apply the Dublin Core Type element 'type'. A very large number of different types (>100k) were detected, many of which are free-text entries rather than being selected from controlled vocabularies as can be observed in Figure 2.

Figure 2 - A wordle illustrating the distribution of different terms used in the Dublin Core 'type' property in metadata delivered via re3data listed OAI-PMH interfaces.

This is challenging for machine agents like F-UJI[13] to detect the intended content and this can complicate how repositories and their contents are indexed and showcased in registries and discovery portals and consequently will obstruct large scale comparisons and analyses. These observations of the current landscape led to a specific focus on the characteristics of

[7] The Consortium of European Social Science Data Archives (CESSDA) consists (currently) of 22 member countries and one observer that are service providers. They are named 'data archives', but fit the description of repositories used in this document. https://www.cessda.eu/About/Consortium
[8] https://www.w3.org/TR/vocab-dcat-2/
[9] https://www.w3.org/TR/vocab-dcat-3/
[10] https://schema.org/
[11] https://www.dublincore.org/
[12] https://www.re3data.org/
[13] Anusuriya Devaraju, & Robert Huber. (2020). F-UJI - An Automated FAIR Data Assessment Tool (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.4063720

Funded by the European Union

organisational entities and their holdings in the guidelines. Baseline transparency about core identity is a dependency for considering the topics of trustworthiness and FAIR assessment.

Trust through transparency

Transparency between parties including repositories, object depositors, object users, and other (meta)data services is a critical precursor to trusted relationships. For repositories that provide active long term preservation of the digital objects they hold, a range of certifications exist to award 'Trustworthy Digital Repository' (TDR) status (e.g. CoreTrustSeal, ISO16363[14], Nestor seal[15]). For these types of repository, exposing metadata about their certification status is a way of demonstrating trustworthiness, but not all service providers have such defined criteria, standard processes, and external authorities to assess and validate their trust characteristics. Therefore, it is important that the transparent disclosure of self-declared assertions about trust-related characteristics is also possible. This information informs perceptions and decisions around interactions between relevant stakeholders, even in the absence of a certificate.

Information can be presented at organisational level based on the different activities and functions[16] the repository undertakes, while granular characteristics of repositories can also be included in repository registries (e.g. organisational mission, or governance structure).

A repository may also list the levels of curation and preservation[17] it provides to the digital objects it cares for. This is valid organisational information for a registry, but is insufficient if the repository offers different levels of care to different digital objects. In this case the best approach to transparency would be to specify the level of care applied at the object level as well as the organisational level.

Absolute trustworthiness can never be guaranteed. However the public assertion of information about an organisation or an object demonstrates transparency and supports mutual trust between human actors (e.g. researcher or funder) and, increasingly, interoperability between machine agents.

Transparency of FAIR assessment outcomes

Many different FAIR assessment tools are currently available for use, with a wide diversity of aims, audiences, purposes, target objects, execution types, and interpretations of FAIR. This leads to a diversity of results and their presentation, which illustrates the complexity of exposing FAIR assessment in the current landscape. Without a clear presentation of relevant information, it is more difficult to develop a sense of trust in the assessment results presented.

[14] https://www.iso.org/standard/56510.html
[15] Nestor seal for Trustworthy Digital Archives: https://www.langzeitarchivierung.de/Webs/nestor/EN/Zertifizierung/nestor_Siegel/siegel.html
[16] Using the broad activities and functions headings from: L'Hours, H., & Bell, D. (2023). Repository & (Meta)Data Services Activities & Functions Overview (v01.00). Zenodo. https://doi.org/10.5281/zenodo.7689090
[17] CoreTrustSeal Standards and Certification Board. (2023). Curation & Preservation Levels: CoreTrustSeal Discussion Paper (v02.00). Zenodo. https://doi.org/10.5281/zenodo.8083359

Funded by the European Union

FAIR assessment of digital objects can be undertaken by anyone, including researchers, research organisations, repositories, and other service providers. However, a clear understanding of the results of an assessment and its implications depends on an understanding of the methodology of the evaluation process executed by a specific tool. This is particularly important when assessment metrics are selected or adapted for a disciplinary context. It is therefore important that the guidelines include the exposure of certain assessment characteristics, to allow both humans and machines to understand the context of the assessment results and foster a sense of trust in them.

These summarised insights into the current landscape inform the scope and focus of the guidelines. The guidelines are further informed by the concept designs and other scoping considerations presented in the next sections.

2.2 Design concepts

The development of the guidelines must be anchored by a set of design concepts to ensure the original purpose and goals are maintained during the iterative developmental process that will span the project's duration. To this end, the following concepts have been defined:

- **The guidelines and accompanying prototype should focus on how to expose information and are neutral on which information should be exposed.**

This line of work is focused on facilitating the connections between repositories, registries, validation authorities, and other stakeholders. The guidelines and prototype therefore focus on addressing the 'how' of information exposure. The guidelines and prototype will not not include mandates or requirements of minimal information to be exposed. Instead, they will focus on providing recommendations and examples of information that can be exposed to help stakeholders to draw their own conclusions on the trustworthiness of the service. .. Although the guidelines focus on the exposure of information related to trustworthiness and FAIR assessment, they do not specify how the judgement of a repository or object as 'trustworthy' or 'FAIR(-enabling)' should be reached. Specific information consumers can design their own requirements or standards about the specific elements to expose or the content of the information exposed depending on their purposes. Such requirements on content can be communicated to their desired audience or counterparts directly. For example, F-UJI as an information consumer can detail which metadata elements the tool needs to see exposed to run all its tests, and what content in those elements the tool looks for to satisfy the metrics. Moreover, if this mechanism of information exchange will be used in the future to align a Network of TDRs[18], it will be the decision of the Network governance or decision making body to set any expectations around the sharing or content of specific metadata fields. Our focus is primarily on the exposure and linking of information.

- **The model should be flexible enough to fit all relevant assertions of information.**

[18] Philipp Conzett, Ingrid Dillo, Francoise Genova, Natalie Harrower, Vasso Kalaitzi, Mari Kleemola, Amela Kurta, Pedro Principe, Olivier Rouchon, Hannes Thiemann, & Maaike Verburg. (2022). Towards a European network of FAIR-enabling Trustworthy Digital Repositories (TDRs) - A Working Paper (v2.0). Zenodo. https://doi.org/10.5281/zenodo.7034315

Funded by the European Union

The purpose of the guidelines and the prototype is to connect the stakeholders in the landscape on characteristics, trustworthiness, and FAIR assessment. Acknowledging that the landscape is vast and varied, it is important that the guidelines are broadly implementable and not too specific to support a wide range of assertions to be made. For example, the guidelines on exposing FAIR assessment should facilitate different types of results (e.g. labels, numerical, including different ranges of scores) to be exposed. However, the model should also remain meaningful and not risk being weakened by its broadness to the extent that it is no longer adding anything of relevance. Striking this balance and ensuring the model is applicable to different scenarios will be an important part of the prototype testing and will require community testing and input. Throughout the process, focus may be given to more specific implementations at times to fit with the timeline and scoping.

- **The guidelines should be relevant to a broad range of repository and object assessment processes**

The guidelines address repositories, assessment tools, certifications, and registries and will be applicable by different instances of these. This enables the adoption of the guidelines in a variety of scenarios, e.g. discipline specific assessment tools and registries or other existing and emerging data services. The testing of the prototype will focus on specific approaches (e.g. F-UJI for FAIR assessment, CoreTrustSeal for certification of repositories), but the guidelines will remain agnostic. The guidelines are not meant to offer a replacement for any of the current instances of services, or the standards and requirements they offer. The use of the guidelines will be in parallel with the existing instances, and the aim of the prototype testing is to validate that the guidelines are suitable to be used in parallel, or need to be updated to better reflect the relevant community standards and established services.

- **The final guidelines and prototype should be meaningful and usable by both human actors and machine agents.**

The exposure of information related to identification, trustworthiness, and FAIR assessment is relevant to both humans and machines. It is therefore important that the final presentation of the guidelines and prototype should be interpretable and usable by both types of stakeholders. Throughout the development of the guidelines and prototype, focus may shift more to one of the two to reach intermediary goals. For example, the first presentation of the guidelines in this report will be geared towards human understandability, as our initial goal is to gather human feedback on these. As prototype development and testing moves forwards, the focus will shift towards the machine-relevant expressions. Ultimately the guidelines should be able to communicate to a range of stakeholders. To this end, the decision has already been made to use basic RDF[19], which allows for the presentation of information in ways that are understandable to both human actors and machine agents. The levels of readability and actionability (e.g. conversion to HTML text for humans or native RDF for machines) in the prototype model will depend on the consuming applications and the decisions made by the stakeholders themselves in exposing information according to certain standards or requirements.

2.2.1 Addressing functions and characteristics of organisations

[19] https://www.w3.org/RDF/

Funded by the European Union

A visualisation has been created that details the flow of information envisioned in the model, from the perspective of the repository. The terms used to indicate parts of this process, and their relationships. This overview gives insight into the overall methodology that the guidelines are based on and the way this can effectively facilitate communication.

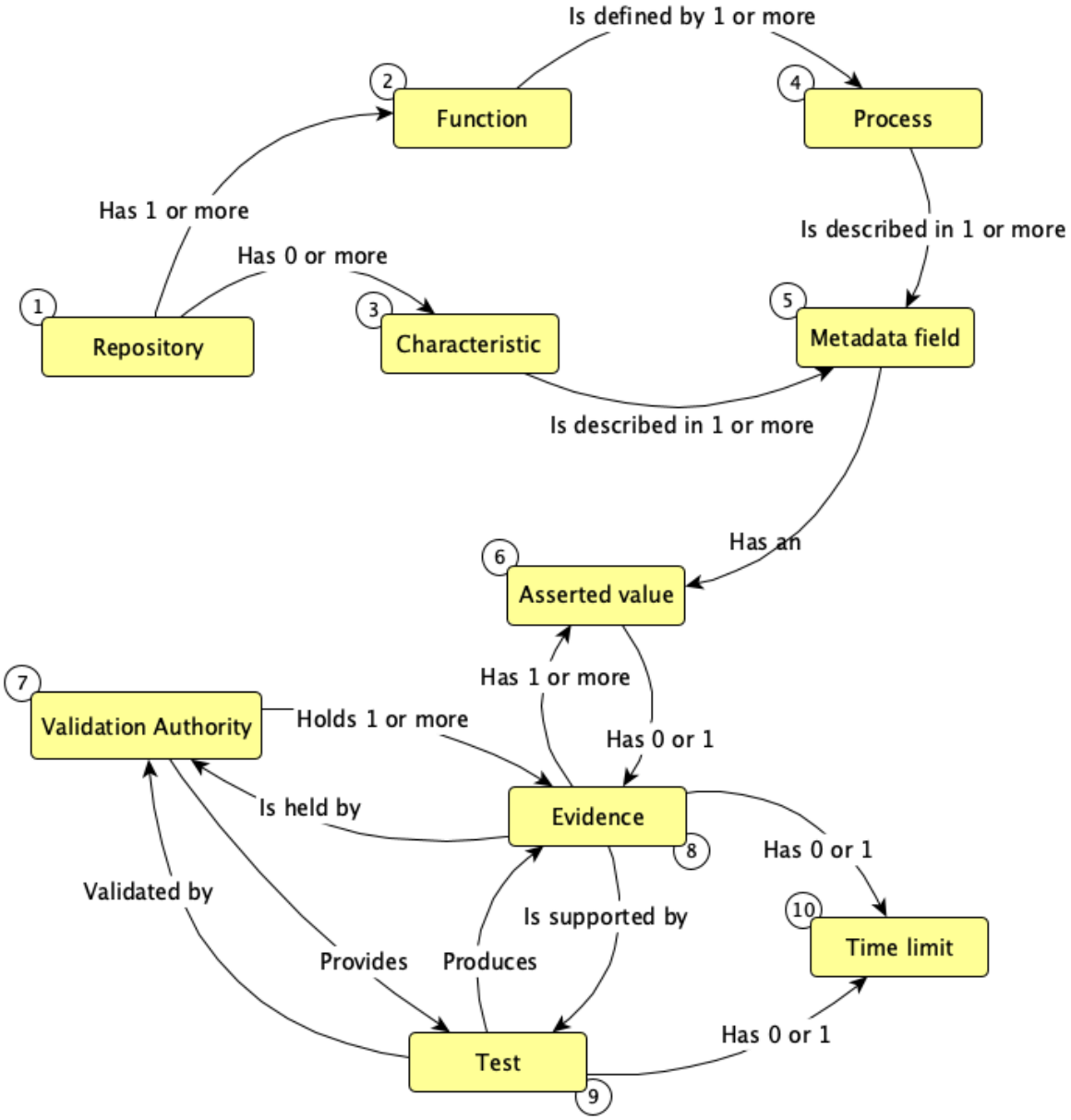

Figure 3 - A high-level representation of the relationships between entities involved in trustworthiness of a repository.

In Figure 3 a **Repository** [1] has both **functions** [2], such as dataset ingest or long term preservation, and **characteristics** [3], such as contact information or organisational name. Functions [2] can consist of one or more **processes** [4] which can be performed by human

Funded by the European Union

actors or machine agents or a combination of both. Characteristics [3], functions [2] and processes [4] are described in one or more **metadata fields** [5], such as a characteristic of the repository, e.g. dcterms:license, or as a an aspect of the process such as dcterms:dateAccepted. Each metadata field [5] has an associated **asserted value** [6], such as a free-text description, terms selected from a controlled vocabulary, or links to resources with the relevant information. Different validation actions can be undertaken based on the asserted value [6]. The assertion [6] can be accepted as is, without further validation or checking, but there are also possibilities for human and/or machine validation, based on the **evidence** [8] that is presented. For example, if a link is provided in the assertion [6], a machine could directly validate the presence and availability of the link. A human could then take the next step and check the content and quality of the evidence [8] that the link provides. If a **validation authority** [7] has been established for the type of information exposed, they can be referred to for validation of the evidence [8]. They can also give insights into the **tests** [9] that have produced the evidence [8], so that human actors and machine agents can understand how a test [9] was performed and how this has created the evidence [8] that supports the assertion [6] made. The evidence [8] that validates an assertion value [6] may have a **time limit** [10]. In the example of an automated FAIR assessment of a digital object, the assessment tool is the validation authority [7] which provides information on how the tests [9] are performed for the FAIR principles and the metric based scores given [8]. The tests [9] in this instance may also have a time limit [10] since the information [6] about the digital object may have changed and the scores achieved [8] may have improved. With CoreTrustSeal as an example of a validation authority [7], the evidence [8] provided for the certification process has a time limit [10] of three years at which point the repository [1] must reapply.

2.3 Scoping

Connecting the vast landscape of scientific information and making sure it is done in a way that is most relevant to the community brings the risk of infinite scope. Though the general intent is for the model to be generic and widely applicable, work will still need to be done to ensure applicability and mapping to make sure the model indeed becomes inclusive. To ensure feasibility of the work, a scoping exercise is needed to put the necessary boundaries in place. This section will report on the current boundaries identified.

Object types - The model addresses digital objects deposited in a repository. Some information is most relevant to be exposed at the object level, such as the FAIR assessment results for that particular object.The current scope of objects is focused on data, but in later iterations of the guidelines the scope will be broadened to include research software. Semantic artefacts fit into the generic and flexible model, but their unique characteristics will not be explored and the guidelines will not be tailored to include their specific needs.

Repository types - Similarly to the object types, the model will be designed to be generic and flexible, and therefore likely applicable to many of the relevant stakeholders. The

Funded by the European Union

extensive crosswalk and mapping made by UK Data Service[20] was used to identify overarching activities and functions to focus the model on (see our previous publication[1] for a more extensive overview of the crosswalk exercise and included criteria). Our main focus is on repositories exposing their information in a way that registries, assessment tools, and validation authorities can usefully interact with.

Guideline specificity - As the guidelines will be iteratively developed over the course of the project, the specificity of the guidelines will increase over time. The current first presentation of the guidelines will not yet adhere to a level of detail that makes them directly implementable, but rather present the overall description in a narrative that can invite community feedback for next iterations.

As the guidelines advance and the prototype is developed through multiple iterations, informed by community testing and input, the boundaries set here will be reviewed and revised as necessary. Based on the feedback received and lessons learned through the prototype testing, amendments to the guidelines and prototype will be carried out where these are deemed most useful to the community, while also considering the feasibility of implementation within the context of the current project. Any resulting changes in the scope of the work will be communicated in next iterations of the guidelines.

2.4 Future developments

This section describes some of the future developments envisioned for the guidelines and prototype with the level of specificity that is currently possible.

2.4.1 Identifying additional use cases and added values

Aside from curating the list of guidelines, attention will also be given specifically to the identification of the added values of the model and the relevance for the community to adopt it. Collecting and describing examples of how the model can provide added value to a certain context or stakeholder will make the work come alive and give it the practical implications it needs to be sustainable and relevant. Some initial descriptions of added value are presented here:

- **Reporting on repository FAIR outcomes**

An object's FAIR score is partially dependent on the many FAIR-enabling activities undertaken by a repository. Because of this, FAIR assessment scores across many objects held in the same repository are often similar or identical. A repository could execute a FAIR assessment using an (automated) assessment tool on their entire holdings, which provides insights into their overall scores, allowing to identify areas of strength and scores that could be improved. Such insights can be reported at the organisational level and communicate to potential depositors what kinds of scores are achievable for digital objects deposited with a

[20] L'Hours, H., & Bell, D. (2023). Repository & (meta)data Services Functions & Activities: Crosswalk (v01.00). Zenodo. https://doi.org/10.5281/zenodo.7690658

Funded by the European Union

particular repository. This 'baseline'[21] provides information about the current level of FAIRness outcomes. From here it is possible to identify where there is 'room for growth' whether distilled from considering the information that is declared about FAIRness on the organisational level, or by the depositor for their specific object.

- **Reporting on changes in FAIRness of an object over time**

By exposing information about the FAIR assessment outcome at the object level it is possible to obtain details about the metrics and methods used by each version of the tool, which is especially important when considering domain-specific assessments[22]. Since the use of FAIR assessment tools is intended to provide a starting point from which improvements can be made, it is desirable that multiple assessments can be done for one object over time, and that this history of scores and their changes can be exposed for the object. This can give important insights into the efforts made by the object owner or the repository in improving FAIRness.

FAIR assessments are not static and standalone. Some metrics, particularly those associated with technical implementations (file formats, metadata schemas) and community needs (specific software applications, disciplinary semantic artefacts) will evolve over time. Objects that are not being actively preserved[23] to address changes to technology (file format migration) or the community (ontology updates) risk a deteriorating FAIR score over time. This can also be reflected in a continuous history of FAIR assessments for an object.

- **Generating FAIR Implementation Profiles**

A FAIR Implementation Profile (FIP)[24] is a report made by an organisation or wider community of practice detailing specific choices and implementations made related to the FAIR principles. FIPs are intended to be openly available to peruse and consult by others, so that community practices and standards can be more easily discerned and implemented by others to improve their own FAIR-enabling activities. The aim of FIPs is to communicate FAIR practices openly and widely, and to help accelerate convergence of standards. If a repository were to expose information about their FAIR-enabling practices using the guidelines and model, there is a potential to harvest this information and generate a FIP from it. By establishing a direct and ongoing connection, such a FIP could be automatically updated if the information exposed changes (ideally with a version history included). Moreover, if communities start setting requirements for exposing information, FIPs could be created for a wider and more informative unit of the landscape. For example, a CESSDA FIP could be created based on the information exposed by all their Service Providers. This could also give insight into which practices are less standardised in the community, which can point focus to areas that need more attention.

[21] 'baseline' FAIR score here indicates the FAIR score a repository can provide for an object with their standard FAIR-enabling practices, without yet considering the influence the object creator can exert over that score. This concept comes from the observation that there is often only little variation in FAIR scores within one repository, indicating that the repository practices are responsible for a considerable part of the FAIR score of any one object.
See section 5.3.2. of: Pijus Krūminas, Joy Davidson, Ingrid Dillo, Carmela Asero, Jonas Antanavičius, Peter Doorn, Aurinta Garbašauskaitė, Marjan Grootveld, Laurence Horton, Žilvinas Martinaitis, Adriana Rantcheva, Wilko Steinhoff, & Maaike Verburg. (2022). European Research Data Landscape Study Report (deliverables 3.2, 4.2, 5.2). Zenodo. https://doi.org/10.5281/zenodo.7351121
[22] Robert Huber, Maaike Verburg, Mike Priddy, Hervé L'Hours, Joy Davidson, & Hannah Mihai. (2023). D5.1 Implementing metrics for automated FAIR digital objects assessment in a disciplinary context (V1.0). Zenodo. https://doi.org/10.5281/zenodo.7784119
[23] L'Hours, H., Kleemola, M., von Stein, I., van Horik, R., Herterich, P., Davidson, J., Rouchon, O., Mokrane, M., & Huber, R. (2022). FAIR + Time: Preservation for a Designated Community (02.00). Zenodo. https://doi.org/10.5281/zenodo.5797776
[24] https://www.go-fair.org/how-to-go-fair/fair-implementation-profile/

Funded by the European Union

- **Including repository information directly into Data Management Planning tools**

Some funders, including the European Commission, require grantees to deposit selected research data with trustworthy repositories[25]. Formal certification by a validation authority is a good source of evidence about a given service's overall trustworthiness and its functions. As emphasised before, not just certification, but general transparency about the repository service is just as important to help users to make an informed decision about where they deposit their data and whether they have trust in the service. However, it can be hard to know what to look for to inform a decision on whether a repository is trustworthy and, as such, users will need guidance and support to help them make an informed choice. There is potential to incorporate guidance into existing questions about data storage, curation, preservation, and access that are included in most data management planning (DMP) tools. By incorporating guidance into the DMP tools directly, users will be better supported to make informed choices about data deposits from the earliest stage of their research.

- **Harvesting of repository information requirements to support assessment and certification**

The exposure of information following the proposed guidelines and model could be used to aid the completion of applications for assessment by a certification body (validation authority). This would simplify the process for the applicant, support partial automation and address the issue of relevant information being buried somewhere on a website or absent altogether. This all reduces administrative overhead, improves interoperability, and benefits the wider scope of research. For the specific example of CoreTrustSeal, a repository could make sure to expose metadata related to each of the 16 requirements. CoreTrustSeal could then directly harvest and check for presence (machine validation) and quality (human validation). After obtaining the certification, the CoreTrustSeal badge could then be displayed by the repository, linking directly to the validation of their certification on CoreTrustSeal's side (e.g. the list of currently certified repositories, or the approved version of the application submitted by the repository).

- **Increased ease of executing landscape analyses**

The added value of wide uptake of this method of information exposure is of course that more and more stakeholders can come together and understand each other better. What it also supports is the ability to zoom out to get a view of the landscape as a whole and identify strengths, synergies, overlaps, and gaps in services and information. This kind of information could be of great value to funders and policy makers, to identify new areas of focus that the whole scientific community (or specific subsets of it) could benefit from the most. For an umbrella initiative such as the EOSC, this could also provide useful information to accelerate the realisation of their vision.

Our outreach activities and calls for community-input will focus on collecting more of such examples to present with the prototype and the final recommendations. These activities will also focus on community and domain-specific topics, to identify a wide perspective. Thus, as

[25] https://open-research-europe.ec.europa.eu/for-authors/data-guidelines

Funded by the European Union

this list may extend, it is important to note that the efforts and resources available to the project will limit the number of use cases and examples that can be practically explored. This will result in an overview of added values that contains some explored cases and some cases that remain to be investigated in the future.

2.4.2 Evaluation through the prototype

The prototype complements the guidelines and plays a crucial role in bringing the draft ideas to life. The prototype is envisioned to be a small sample setup, with different infrastructures and services involved. It will serve as a demonstrator and showcase the potential of the concepts. By providing hands-on experience for the involved project partners and the support action participants, the prototype will allow for experimentation, feedback, and refinement in an iterative approach, leading to a clearer vision of the overall concept and refinement of specific guidelines. This covers all aspects, from technical integration to use cases building on top.

This proof of concept allows for early identification of potential flaws, gaps, or improvements. When linking services and breaking up information silos, the subsequent development can be adjusted, given the feedback and experience gathered. The goal is not to define or build new standards, services, or technologies. The test setup will start with the embedded organisations in the dedicated FAIR-IMPACT task and will be extended by external data services through the FAIR-IMPACT open calls to gather added value and identify future potential. With a positive evaluation of the prototype results, they are thought to be incorporated into the production services to sustain them.

Funded by
the European Union

3 First iteration of the guidelines on exposing information

This section presents the first iteration of the guidelines, as the result of the methodology and considerations presented in the rest of the document. As detailed in the scoping (Section 2.3), the specificity of the guidelines is currently focused on human interpretation of the rationale of each guideline. To this end, each guideline is currently presented with a name and description. In later iterations, more information will be added to each guideline to make them more focused on the implementation of the guidelines as will be explored through the prototype. The tables presented here will then be expanded with more relevant information on how the guidelines can be followed, implemented, and what that could look like in certain scenarios.

Table of contents - Guidelines

FIELD	DESCRIPTION
Name	**Relevant information should be exposed to achieve transparency.**
Description	Exposing information on organisations, services, and objects, as well as their functions and characteristics, implies the precise descriptions of the

Funded by the European Union

	particular resources. For example, a dataset should be recognisable as a dataset by the information consumer and a repository should identify itself as a repository and data catalogue. The denotation should be as detailed as possible, yet referencing superclasses (e.g. 'Resource' for a dataset, or 'DataService' for a catalogue, depending on the standard used). This will enable machine agents to select and process the information in specific as well as in generic use cases. It also helps human actors to recognize and understand the information.

FIELD	DESCRIPTION
Name	**Information about the functions and characteristics of repositories and objects should be expressed in line with defined standards and criteria.**
Description	Transparent information exposed should take account of, and map to, existing standards and criteria. While perfect alignment may not be possible, defining and documenting how the information structure and content relates to existing efforts will minimise divergence and maximise interoperability. Designing the presentation of information should take account of objects at a generic level (e.g. Dublin Core[26] or DataCite[27]) and at disciplinary level (e.g. DDI[28] for the social sciences), and for repositories at a functional level (e.g. CoreTrustSeal) or a more granular level (e.g. DRAWG[29]). The design should also map to the needs of potential consumers of information about repositories (e.g. re3data[30]) or objects (e.g. F-UJI[31]) is also important.

FIELD	DESCRIPTION
Name	**Information should be exposed by, and/or provide references to, an originating source.**
Description	As observed in the current landscape, much of the relevant information that needs to be available in a transparent way to inform trust is currently available at multiple third party service providers. Registration, identification, and aggregation adds value to the research landscape and in most cases is a necessity (e.g. PIDs). Despite the complexity to navigate multiple services, merging all information with the same up-to-dateness, it also requires additional trust in the intermediary services to inform decisions. This guideline therefore specifies that an information provider should be placed in control of its own information and permit multiple other organisations to consume it for a variety of uses. It is also of

[26] https://www.dublincore.org/
[27] https://datacite.org/
[28] https://ddialliance.org/
[29] The RDA Working Group 'Data Repository Attributes Working Group'
https://www.rd-alliance.org/plenaries/rda-20th-plenary-meeting-gothenburg-hybrid/data-repository-attributes-working-group-drawg
[30] https://www.re3data.org/
[31] https://www.f-uji.net/

Funded by the European Union

	importance that when the information is exposed in different places, the information is consistent across all locations. This should be possible if the information can be validated using the same validation actions, regardless of the location of exposure.

FIELD	DESCRIPTION
Name	**Clarity should be provided on how information should be expressed to support humans and machines.**
Description	Defining how different assertions should be presented requires a balance between flexibility and providing enough guidance to ensure uniformity. For example, the following assertion types could be considered: • **Free-text assertions**[32]: statements in response to descriptive criteria about the required information. • **Controlled assertions**[32]: selections from ontologies or controlled vocabularies. • **Identification**: Expose and reference PIDs with your (meta)data • **Evidence artefacts**[32]: links to resources containing the asserted information. Information consumers or communities could work to specify the scope of content of free-text assertions, the ontologies and controlled vocabularies used for the controlled assertions, or acceptable links to use for evidence artefacts. Persistent identification, resolution and associated metadata are essential foundations for this guideline. Organisations, such as DataCite[33] or ORCiD[34], enable sustainability and availability far into the future. They provide unique references, deduplication and in general improve findability and reuse within and between other services beyond a specific technology stack. To facilitate meaningful validation actions, supporting information and documents should be linked and exposed (see Figure 3). As an example, when a repository claims certification, it should provide a link to the certificate at the certification authorities' site. This enables human actors and machine agents to consider assertions as a kind of evidence for other assertions and validate accordingly.

FIELD	DESCRIPTION

[32] This assertion type reflects the currently available response types to CoreTrustSeal applications.
[33] https://datacite.org/
[34] https://orcid.org/

Funded by the European Union

Name	**When an assertion can be validated, the possible validation action(s) should be defined.**
Description	Such specified validation actions ensure uniformity in the interactions with human agents and/or machine agents can carry out and expect to be carried out with regards to their exposed information. For example, the following validation actions could be considered: • **Acceptance of assertion**: assertion is accepted without further validation. • **Direct machine-actionable validation**: given that the assertion is machine-testable, the information presented is automatically validated in an established process. • **Machine-actionable validation through a third party**: given a third party can be pointed to as the authority on the information asserted, they are called on to validate the information automatically through an established process. • **Validation through human action**: given the assertion is not machine-testable, the consumer must take manual steps to validate the information presented. • **Validation through a mixture of human and machine action**: given that the assertion is machine-testable, the choice may be made to also validate the information manually to ensure the content or quality of the supplied information (i.e. the machine tests the information is available, the human validates the content). Choices made in assertion actions can be based on the purposes of the information consumer, or the level of information made available by the provider. To achieve certain use cases for the model, choices will need to be made in the validation actions that will be desired and sufficient to reach specific goals. Technically, validation mechanisms should be provided to allow users to easily identify and verify trustworthiness or FAIR certificates or ratings. This should be done on the basis of tamper-proof badges or seals, such as OpenBadge[35].

FIELD	DESCRIPTION
Name	**Appropriate standards should be used to expose dataset metadata, FAIR assessment results, and catalogue information towards harvesters and discovery services.**
Description	To enable exploration of (meta)data services, a standard should be used that is able to describe a repository, the digital objects it holds, and their

[35] https://openbadges.org/

	context properly, such as schema.org[36] or DCAT[37,38]. For example, DCAT is an established and matured standard and recommended by the World Wide Web Consortium (W3C). It is an RDF[39] vocabulary and designed to foster interoperability between data catalogues on the web. It allows for tailored profiles and yet persisting flexibility and interoperability of the semantic web. DCAT already acknowledges the research data landscape in its specification, e.g. PIDs[40]. With a broad and active community it also bridges to a wider audience like the Open (Governmental) Data-Community[41,42,43] or subject specific organisations like the OGC[44]. Thus the implementation of DCAT will increase consistency and machine-actionability for the repository and its datasets[45]. It will also enable exposing related (meta)data like data quality and linking to other Linked (Open) Data resources. To standardise outputs of FAIR metrics and associated assessment results the use of Data Quality Vocabulary (DQV)[46] is recommended since this may be used to embed FAIR assessment results within the metadata of assessed data sets via DCAT as recommended by the W3C 'Data on the Web Best Practices'[47]. This standard further allows to include a minimum set of metadata required to reproduce FAIR assessments which are: test date, assessment target, metric used, and name and software version of the testing tool.

FIELD	DESCRIPTION
Name	**The levels of care offered by repositories and received by digital objects should be expressed.**
Description	Beyond basic information about retention periods, repositories should expose information about the different levels of curation and preservation[48] they provide across their digital object collections. At the digital object level it should be clear what levels of retention, curation and preservation are in place, and how and when these might change. Supporting information would include appraisal and selection criteria, re-appraisal schedules, preservation plans etc.

[36] https://schema.org/
[37] https://www.w3.org/TR/vocab-dcat-2/
[38] https://www.w3.org/TR/vocab-dcat-3/
[39] https://www.w3.org/RDF/
[40] https://www.w3.org/TR/vocab-dcat-3/#dereferenceable-identifiers
[41] https://joinup.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-data-portals-europe/about
[42] https://data.europa.eu/en
[43] https://www.dcat-ap.ch/
[44] https://www.ogc.org/press-release/ogc-forms-new-geodcat-standards-working-group/
[45] FAIR-DataPoint is using a basic DCAT-Profile as a wrapper to make existing research data repositories discoverable.
[46] https://www.w3.org/TR/vocab-dqv
[47] https://www.w3.org/TR/vocab-dcat-3/#quality-information
[48] CoreTrustSeal Standards and Certification Board. (2023). Curation & Preservation Levels: CoreTrustSeal Discussion Paper (v02.00). Zenodo. https://doi.org/10.5281/zenodo.8083359

Funded by the European Union

FIELD	DESCRIPTION
Name	**Multiple calibrated FAIR assessment tools should be used, embedded in a holistic FAIR consultation process that supports contextual understanding.**
Description	Since there are several FAIR evaluation tools, each of which evaluates the various FAIR implementation options somewhat differently, multiple evaluation tools should be used which should be calibrated against a selection of FAIR benchmarking standard datasets, such as the currently prepared set of benchmarks for FAIR signposting[49]. Because these tools usually focus on machine-readable FAIR implementations, it is also of importance to focus separate attention to human-friendly FAIR implementation. Therefore, FAIR evaluations should always be part of an intensive consultation process, supporting a holistic understanding of FAIR and its context. This process should already start with the selection of a representative set of datasets to be studied and subsequently help to both interpret the machine-aware FAIR implementations and match them with the human-friendly ones.

[49] https://s11.no/2022/a2a-fair-metrics/

Funded by the European Union