FAIRness Characterization Questionnaire

Leonardo Guerreiro Azevedo, Julio Tesolin, Gabriel Banaggia, Renato Cerqueira
IBM Research – Brazil
lga@br.ibm.com, {julio.tesolin,gbanaggia}@ibm.com, rcerq@br.ibm.com

Abstract

The FAIR principles guidelines aim to enhance the discovery and usage of digital objects by humans and computational agents. They are formulated at a high level and, as such, are interpreted and implemented in different ways by communities of practice. However, practi cal approaches outlining FAIR-related characteristics of digital objects are few and far between. There are many proposals of metrics, questionnaires, and tools for manual, automated, and semi-automated FAIRness assessment. Several of these are domain-agnostic and do not consider scientific communities' needs which are varied and require specific implementations for better estimation. This work presents an improved questionnaire for digital object FAIRness charac terization. Our goal is not to present a FAIRness grade for digital objects, but to outline their current status, at any point of their data life cycle.

1 The FAIR questionnaire

Research data can benefit from improvements to its infrastructure to support its use in investigations and the accrual of newly generated information and knowledge [8]. Reproducibility and completeness could be better, and descriptions of methods or computer code used for data analysis could be more adequate [6].

Enhancing knowledge discovery for human and computational agents is a challenge for data intensive science, involving the access, integration, and analysis of task-appropriate scientific data [8]. Aiming to tackle it, Wilkinson *et al.* [8] published the FAIR principles, a set of 15 recommendations for improving the Findability, Accessibility, Interoperability, and Reusability of digital objects presented in Table 1 [5]. The FAIR principles are domain-independent and aim to facilitate the reuse of data by humans and machines [7]. They are suggested as a prerequisite for proper data management and stewardship [8].

Several approaches for that assessment have been proposed, such as RDA's FAIR Data Maturity Model [4], a FAIR metrics framework [9], or questionnaires and tools for manual, automated, and semi-automated FAIRness assessment [1][3]. Several of these are domain-agnostic, *i.e.*, they try to cover a broad list of distinct requirements, which is a big challenge. Scientific communities' needs, however, are tremendously varied and require specific implementations for better assessment [2].

Our goal is to create a questionnaire whose answers characterize the current FAIRness status of a digital object. The questionnaire's initial questions are used to contextualize the domain and define the digital object that will be characterized as presented in Table 2. After understanding the assessment context, we begin with questions about each principle presented in Tables 3, 4, 5, and 6.

Table 1: FAIR principles.

Findability	F1. (meta)data are assigned a globally unique and eternally persistent identifier. F2. data are described with rich metadata (defined by R1 below). F3. metadata clearly and explicitly include the identifier of the data it describes. F4. (meta)data are registered or indexed in a searchable resource.
Accessibility	A1. (meta)data are retrievable by their identifier using a standardized communi cations protocol. A1.1. the protocol is open, free, and universally implementable. A1.2. the protocol allows for an authentication and authorization procedure, where necessary. A2. metadata are accessible, even when the data are no longer available.
Interoperability	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles. I3. (meta)data include qualified references to other (meta)data.
Reusability	R1. meta(data) are richly described with a plurality of accurate and relevant attributes. R1.1. (meta)data are released with a clear and accessible data usage license. R1.2. (meta)data are associated with detailed provenance. R1.3. (meta)data meet domain-relevant community standards.

References

[1] fairsharing. Fairassist.org. https://fairassist.org/. Accessed in 2023–03-06.

[2] Directorate-General for Research and Innovation (European Commission), EOSC Executive Board, Jan Magnus Aronsen, Oya Beyan, Natalie Harrower, András Holl, Rob W. W. Hooft, Pedro Principe, Ana Slavec, Sarah Jones, and Françoise Genova. *Recommendations on FAIR metrics for EOSC*. Publications Office of the European Union, LU, 2021.

[3] GO FAIR Initiative. Fair implementation profile (fip) mini-questionnaire. https://bit.ly/ yourFIP. Accessed in 2023–02-02.

[4] RDA FAIR Data Maturity Model Working Group. FAIR Data Maturity Model: specification and guidelines. https://zenodo.org/record/3909563, June 2020. Accessed in 2023–01-27.

[5] Annika Jacobsen, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, Mercè Crosas, Michel Dumontier, Chris T. Evelo, Carole Goble, Giancarlo Guizzardi, Karsten Kryger Hansen, Ali Hasnain, Kristina Hettne, Jaap Heringa, Rob W.W. Hooft, Melanie Imming, Keith G. Jeffery, Rajaram Kaliyaperumal, Martijn G. Kersloot, Christine R. Kirkpatrick, Tobias Kuhn, Ignasi Labastida, Barbara Magagna, Peter McQuilton, Natalie Meyers, Annalisa Montesanti, Mirjam van Reisen, Philippe Rocca-Serra, Robert Pergl, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Juliane Schneider, George Strawn, Mark Thompson, Andra Waagmeester, Tobias Weigel, Mark D. Wilkinson, Egon L. Willighagen, Peter Wittenburg, Marco Roos, Barend Mons, and Erik Schultes. FAIR Principles: Interpretations and Implementation Considerations. *Data Intelligence*, 2(1-2):10–29, 01 2020.

[6] Dominique G. Roche, Loeske E. B. Kruuk, Robert Lanfear, and Sandra A. Binning. Public data archiving in ecology and evolution: How well are we doing? *PLOS Biology*, 13(11):e1002295, November 2015. Publisher: Public Library of Science.

[7] Cassia Trojahn, Mouna Kamel, Amina Annane, Nathalie Aussenac-Gilles, and Bao Long Nguyen. A FAIR core semantic metadata model for FAIR multidimensional tabular datasets. In Oscar Corcho, Laura Hollink, Oliver Kutz, Nicolas Troquard, and Fajar J. Ekaputra, editors, *Knowledge Engineering and Knowledge Management, Lecture Notes in Computer Science, pages 174–181, Cham, 2022. Springer International Publishing.*

[8] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Ax ton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. Scien tific data, 3(1):1–9, 2016.

[9] Mark D. Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. A design framework and exemplar metrics for FAIRness. Scientific Data, 5(1):180118, June 2018. Number: 1 Publisher: Nature Publishing Group.

Table 2: Questionnaire to understand the FAIRness assessment context.

Q.ID	Questions
Q1	What is the reseach community evaluating the digital object?
Q2	Which digital object will be evaluated in this assessment? This question concerns possibilities like: - Evaluate the dataset as the digital object and, consequently, the data and metadatada about the dataset independently if it is bulked (with data instances). *E.g.*, the evaluation targets the dataset schema and strategies to accomplish FAIR. - Evaluate each instance of the dataset individually as the digital object. *E.g.*, evaluate the data instance schema, its properties values, its metadata, and the strategies undertaken to accomplish FAIR for the instance. - Evaluate the dataset considering that its FAIRness is computed from the FAIRness of a set of the dataset instances, *i.e.*, the dataset FAIRness is derived from the instances FAIRness (which is computed as presented in the previous item).

Table 3: Questionnaire to assess Findability.

Princ.	Q.ID	Question
F1	Q3	What is the main identifier (ID) of the data (*i.e.*, data of the digital object under evaluation)?
	Q4	Are there other attributes able to identify the data? If so, what are them?
	Q5	Is the data identifier globally unique or unique in the dataset domain or for a specific context?
	Q6	Is the data ID persistent?
	Q7	Is the data ID resolvable, *e.g.*, to a landing page?

	Q8	Are there metadata describing the data?
	Q9	Do the metadata have a distinct ID from the data? *(If the data and metadata have distinct IDs)*:
	Q10	What is the identifier of the metadata?
	Q11	Is the metadata ID globally unique or unique in the database domain or for a specific context?
	Q12	Is the metadata ID persistent?
	Q13	Is the metadata ID resolvable, *e.g.*, to a landing page?
F2	Q14	Which metadata schemas, if any, are used to describe the data?
	Q15	What kinds of metadata (*e.g.*, descriptive, administrative and structural) are used to describe the data?
	Q16	Which of these metadata schemas are domain specific and which are domain agnostic?
F3	Q17	Does the metadata include the identifier of the data it describes?
	Q18	What technology does link metadata to the data (and vice-versa)?
	Q19	How are the metadata and data linked?

F4	Q20	Is metadata registered or indexed in a searchable resource?
	Q21	Which searchable resource is used to register or index the metadata?
	Q22	Which technology is used to make your metadata available (or indexed)?
	Q23	How is the metadata available or indexed? (*E.g.*, as a static web page, in a database, JSON returned from an API call)
	Q24	Is data registered or indexed in a searchable resource?
	Q25	Which searchable resource is used to register or index the data?
	Q26	Which technology is used to make the data available (or indexed)?
	Q27	How is the data available or indexed? (*E.g.*, as a static web page, in a database, JSON returned from an API call)

Table 4: Questionnaire to assess Accessibility.

Princ.	Q.ID	Question
A1	Q28	Which communication protocols are used to access the metadata?
	Q29	Which communication protocol are used to access the data?
A1.1	Q30	Is the protocol used to access the metadata standardized, open, free, and universally implementable?
	Q31	Is the protocol used to access the data standardized, open, free, and universally implementable?

A1.2	Q32	What security mechanisms are used for metadata access, such as ones used for authentication and authorization, and access conditions and access levels?
	Q33	What the security mechanisms are used for data access, such as ones used for authentication and authorization, and access conditions and access levels?
	Q34	What security information is provided in the metadata that allows one to access the data manually or through a client application?
A2	Q35	Are data and metadata independently stored?
	Q36	What is the metadata longevity plan?
	Q37	What is the data longevity plan, if any?

Table 5: Questionnaire to assess Interoperability.

Princ.	Q.ID	Question
I1	Q38	What is the knowledge representation used for metadata? *E.g.*, Relational, Document, Key Value, Graph, Object, Hierarchical, Net work.
	Q39	Is the knowledge representation used for metadata formal, accessible, shared, and broadly applicable?
	Q40	In what format the knowledge representation used for metadata is provided? *E.g.*, eXtensible Markup Language (XML), Turtle (TTL), JSON, JSON-LD, CSV, BLOB, CLOB.

	Q41	Is the format used for knowledge representation of metadata formal, accessible, shared, and broadly applicable?
	Q42	What is the knowledge representation used for data? *E.g.*, Relational, Document, Key Value, Graph, Object, Hierarchical, Net work.
	Q43	Is the knowledge representation used for data formal, accessible, shared, and broadly applicable?
	Q44	In what format the knowledge representation used for data is provided? *E.g.*, eXtensible Markup Language (XML), Turtle (TTL), JSON, JSON-LD, CSV, BLOB, CLOB.
	Q45	Is the format used for knowledge representation of data formal, accessible, shared, and broadly applicable?
I2	Q46	Which structured vocabularies are used for metadata? *Structured vocabularies range from simple taxonomies or thesauri (e.g., in SKOS) to ontologies in OWL available in public accessible registry.*
	Q47	Are the vocabularies used for metadata FAIR in their own right?
	Q48	Which structured vocabularies are used for data? *Structured vocabularies range from simple taxonomies or thesauri (e.g., in SKOS) to ontologies in OWL available in public accessible registry.*

	Q49	Are the used vocabularies for data FAIR?
I3	Q50	Which qualified references the metadata include to other data or metadata? *Qualified references means any external metadata used to enrich the infor mation.*
	Q51	Which qualified references the data include to other data or metadata? *Qualified references means any external metadata used to enrich the information.*

Table 6: Questionnaire to assess Reusability.

Princ.	Q.ID	Question
R1	Q52	What are the relevant metadata attributes?
	Q53	What is the required accuracy of each metadata attribute, if any?
	Q54	What are the relevant data attributes?
	Q55	What is the required accuracy of each data attribute, if any?
R1.1	Q56	Which usage license is used for metadata?
	Q57	Is the metadata usage license clear?
	Q58	Is the metadata usage license accessible?
	Q59	Which data usage license is used for data?

	Q60	Is the data usage license clear?
	Q61	Is the data usage license accessible?
R1.2	Q62	Which metadata schemas do you use for describing provenance of the meta data?
	Q63	Which metadata schemas do you use for describing provenance of the data?
	Q64	What are the attributes used for data provenance?
R1.3	Q65	What are the domain-relevant community standards for metadata? *Domain-relevant community standards mean minimum information stan dards, well-established and sustainable file formats, common types for in formation, use of template and standardized vocabularies and ontologies etc.*
	Q66	Does the metadata under assessment meet these domain-relevant community standards?
	Q67	What are the domain-relevant community standards for data? *Domain-relevant community standards mean minimum information stan dards, well-established and sustainable file formats, common types for in formation, use of template and standardized vocabularies and ontologies etc.*
	Q68	Does the data under assessment meet these domain-relevant community stan dards?