FAIR Research Object Assessment: A landscape analysis*

Esteban González[1 [0000−0003−4112−6825]], Daniel Garijo[1 [0000−0003−0454−7145]], Oscar Corcho[1 [0000−0002−9260−0753]] Raul Palma[2 [0000−0003−4289−4922]], Malgorzata Wolniewicz[2 [0000−0003−2388−0744]], and Aron Rynkiewicz[2,3 [0000−0002−0528−7544]]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid
{egonzalez, dgarijo, ocorcho}@fi.upm.es
[2] Poznan Supercomputing and Networking Center. Noskowskiego 12/14, 61-704 Poznan, Poland {rpalma,gosiaw,arynkiewicz}@man.poznan.pl
[3] Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland

Abstract. Research Objects (ROs) are becoming a popular means to capture the context and research artefacts associated with a research investigation in both human-readable and machine-readable formats. However, it is unclear how well ROs themselves adhere to the FAIR (findable, accessible, interoperable, and reusable) principles. In this work, we describe a comprehensive analysis of the FAIR assessment of more than 2500 ROs across multiple disciplines. Our work integrates FAIROs, our existing RO evaluation service, in the ROHub platform. We discuss the challenges of calculating the FAIR assessment of aggregations of resources, and how we supplement the FAIROs tests with information from the RO-Crate descriptor file generated by ROHub.

Keywords: Research Object · Metadata · FAIR · analysis.

1 Introduction

Research Objects (ROs) [2] enable researchers to aggregate diverse research outputs, such as datasets, publications, and software, into a *digital object*. ROs not only provide a consolidated access point to these related resources but also incorporate human and machine-readable metadata and contextual information, improving the findability, understanding, and reuse of scientific results [12].

* This work has been funded by the European Commission within the H2020 Programme in the context of the project RELIANCE under grant agreement no. 101017501, and the project FAIR IMPACT (Horizon Europe) under grant agreement no. 101057344, and by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with Universidad Politécnica de Madrid in the line Support for R&D projects for Beatriz Galindo researchers, in the context of the V PRICIT (Regional Programme of Research and Technological Innovation) and the call Research Grants for Young Investigators from Universidad Politécnica de Madrid

Despite the potential for ROs to enhance the impact of research artefacts, few studies have assessed their adherence to the Findable, Accessible, Interoperable, and Reusable (FAIR) principles [13]. Our previous work [7] introduced FAIROs, a tool for assessing ROs based on existing approaches for assessing dataset [3] and ontology [6] FAIRness. In this paper, we complemented FAIROs with an analysis of the Research Object description file provided by the ROHub platform, a repository of over 2,500 ROs spanning multiple disciplines. We transformed FAIROs into a service and integrated it into the ROHub platform, providing a comprehensive assessment of the platform's FAIRness landscape. We also present preliminary results of the integration of the service into the platform's graphical user interface. By integrating FAIROs with ROHub, we offer researchers an efficient way to evaluate the FAIRness of their ROs.

The rest of the paper is structured as follows. Section 2 introduces the main concepts we build on for our approach. Section 3 describes how we have extended ROHub with FAIROs, while Section 4 shows the results of our analysis, further discussing them in Section 5. Section 6 summarizes related work and Section 7 concludes the paper.

2 Background

Our approach relies on the Research Object Crate specification (RO-Crate) [12] and ROHub [11], an online platform that collects and enriches Research Objects from different domains. We further describe them below.

2.1 RO-Crate

RO-crate[4] [12] defines a lightweight approach for packaging research artifacts and their relations, annotations and provenance, in a machine-readable format.

RO-Crates extend Schema.org[8], and are usually serialized using JSON-LD. The root directory of an RO-Crate consists of files including a ro-crate-metadata.json file (which describes the RO), payload files and directories. The ro-crate-metadata.json file is mandatory, while the other files and directories are optional. A ro-crate-metadata.json file should include:

- A **RO-Crate Metadata File Descriptor**, describing the RO metadata file,
- A **Root Data Entity**, which stands for the RO-Crate itself, and its relationship with other data and contextual entities,
- optional **Data Entities**, i.e., files, directories and web resources described in the RO,
- optional **Contextual Entities**, which include people, organizations, contact data, publications, publisher, etc.

[4] https://w3id.org/ro/crate

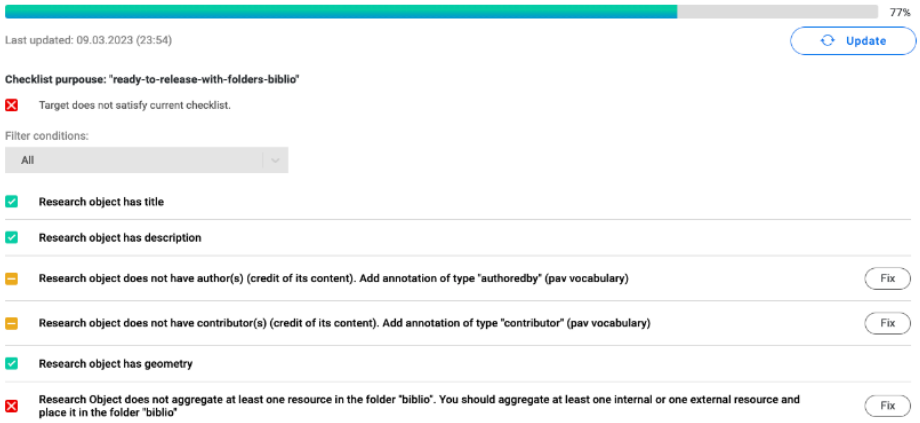

Fig. 1: A snippet of the checklist assessment in ROHub, for a given RO. Each passed check appears in green, optional checks in yellow and failed checks in red.

2.2 ROHub

ROHub [11][5] is an online platform to search, define, enrich and store ROs, using the RO-Crate specification.Resources stored in the platform are provided with permanent identifiers using the w3id service.[6]

ROHub aims to produce high quality Research Objects according to what the research communities producing and reusing them consider relevant, typically by assessing a number of quality dimensions, such as accuracy, completeness, or availability. The platform draw upon the idea of checklists, a well-established tool for ensuring safety, quality and consistency in complex operations, such as manufacturing or critical care. A checklist explicitly defines a list of requirements that must be fulfilled or assessed for a given task.

ROHub designed checklists based on the feedback from its user communities. Common checks for all ROs include having a title, description, creator, publisher, keyword(s), research area and a sketch (an overview how the different elements of the RO relate with each other). In addition, ROs should aggregate at least one resource, which should have a type (i.e., dataset, workflow, etc.)

Figure 1 shows an example of checklist in ROHub, where some checks have passed (green checks) and some tests failed (red checks). ROHub indicates a degree of metadata completeness (green bar on the top of Figure1), and highlights which fields should be better described. Checklists allow researchers assessing the metadata quality of their Research Objects, but they do not analyze the adherence of ROs against the FAIR principles.

[5] https://reliance.rohub.org/
[6] https://w3id.org/

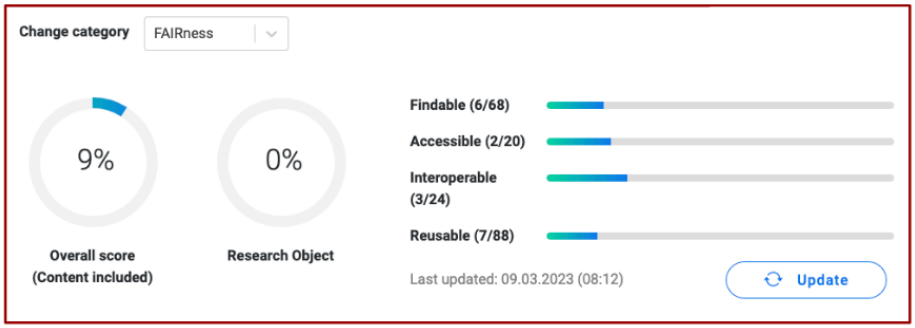

Fig. 2: Prototype implementation overview of the FAIROs service in ROHub.

3 Extending ROHub with FAIR RO assessment

We have integrated the FAIR Research Object Assessment service (FAIROs) [7] in ROHub. Given a RO, FAIROs checks a set of tests against each FAIR principle and returns a detailed explanation of the results, including a description of the test that was performed. FAIROs takes into account both the RO itself and its associated resources, going beyond the assessment of the RO-Crate metadata file.

In particular, FAIROs builds on external services such as F-UJI [3] for datasets and FOOPS [6] for ontologies. For software, FAIROs includes a custom module based on the software metadata extraction framework for code repositories [10]. Each module has its own set of tests for each FAIR principle, based on existing recommendations from the community.

FAIROs collects all test results and 1) enriches and complements them with the information present in the RO-Crate metadata file, 2) integrates and harmonizes the results of different tools/RO-Crate file and 3) indicates, for each principle the number of tests that have passed.

FAIROs is open source,[7] with a public API available online.[8] Figure 2 shows an overview of the integration of FAIROs in ROHub. An overall score groups the scores of all the elements in the RO, while the RO FAIRness is measured separately. For each principle, all performed and passed tests are shown.

4 RO FAIR assessment: A landscape analysis

In this section, we present the results of our FAIR assessment analysis over more than 2500 ROs from ROHub. We discuss first the nature and domain of these ROs in Section 4.1, while Section 4.2 discusses the obtained results.

[7] https://github.com/oeg-upm/FAIR-Research-Object
[8] https://w3id.org/FAIROS/api

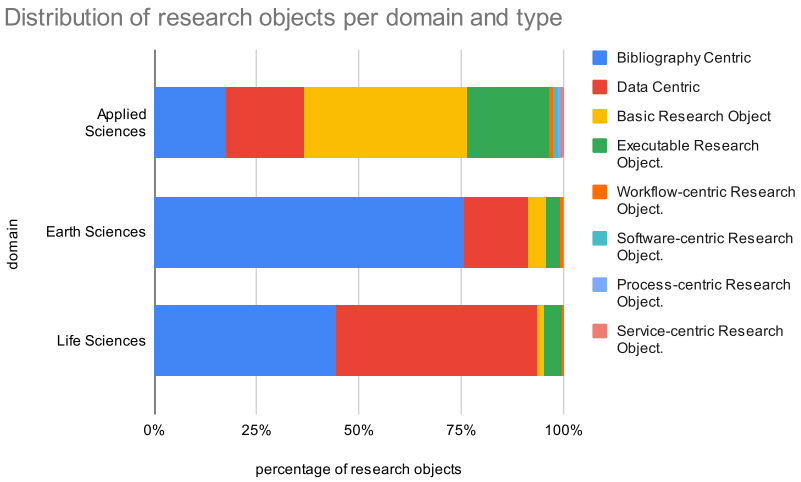

Fig. 3: Distribution of analyzed ROs per domain and type.

4.1 RO types and domain

We have analyzed ROs extracted from the ROHub platform as of March, 2023. ROs are distributed in different categories: Applied Sciences, Earth Sciences and Life Sciences.

Research Objects in ROHub can belong to different types [5], which represents an internal distribution of resources (folders and files) inside the RO. The following RO types are defined:

. Bibliography-centric RO: includes bibliographical reference documents, manuals and other kind of material which support a researcher.
. Data-centric RO: includes datasets of a researcher, including raw and processed data.
. Executable RO: includes scripts or applications needed to run experiments.
. Workflow-centric RO: includes scientific workflow specifications and their respective executions.
. Software-centric RO: includes the source code, datasets and documentation needed to run a software component.
. Process-centric RO: focused on scientific processes (e.g., water modeling)
. Service-centric RO: includes detailed descriptions of services.
. Basic RO: used to describe any resource not covered by the other types.

Figure 3 depicts the RO type distribution per domain. In Earth Sciences, most of the ROs are bibliography-centric. Basic RO is the most prevalent type in the Applied Science domain, while in the Life Science domain, both bibliography-centric and basic-centric ROs are equally distributed. Executable-centric ROs

Table 1: Characteristics of the analyzed ROs per domain

Metric	Applied Sciences	Earth Sciences	Life Sciences
Resources(0-50)	100%	99,67%	100%
Resources(51-99)	0%	0,28%	0%
Resources($\geq$ 100)	0%	0,05%	0%
Completeness (0-50)	15,54%	21,96%	25,66%
Completeness($\geq$ 51)	84,46%	78,04%	74,34%
Downloads(0-50)	97,97%	99,86%	99,87%
Downloads(51-99)	2,03%	0,09%	0%
Downloads($\geq$ 100)	0%	0,05%	0,13%
Views(0-50)	97,97%	99,95%	100%
Views(51-99)	2,03%	0,05%	0%
Views($\geq$ 100)	0%	0%	0%

only have a significant impact in the Applied Science domain. The remaining RO types have a minimal impact in all domains.

ROs in ROHub can be generated in different ways: 1) Imported, where users import a RO described by a RO-Crate file to the platform; 2) Automated, where the RO is created using the API; and 3) Manual, where users created the RO by hand. The percentage of ROs created manually varies significantly between research areas. In Applied Science, 99% of the Research Objects were created manually, in Life Sciences, this figure is 59%, and in Earth Sciences, only 23% were created manually. The percentage of manual creation of ROs varies significantly depending on the type of RO. more than 94% of Data-centric and Basic ROs were created manually, while 52% of Workflows-centric ROs were created manually. Only 4% of bibliography objects were created manually.

Table 1 drills down on the characteristics of the analyzed ROs in terms of their number of resources, completeness score, number of downloads and number of views. The majority of the ROs have less than 50 resources, which suggests that researchers were selective in choosing the resources they wanted to include. The completeness scores of the ROs are generally high, especially in the applied sciences. This may be due to the fact that the platform allows userts to easily identify and address any gaps in completeness, making it easier to improve the overall quality of their ROs. Finally, we do not observe a high number of downloads and views in the analyzed ROs, probably due to thir short lifetime (many have not yet been cited in publications).

The number of Research Objects in ROHub from Earth Sciences is higher than in other disciplines. This is due to the scope of the ROHub platform, initially designed for use cases that are related with that domain (e.g., through European projects like RELIANCE[9]).

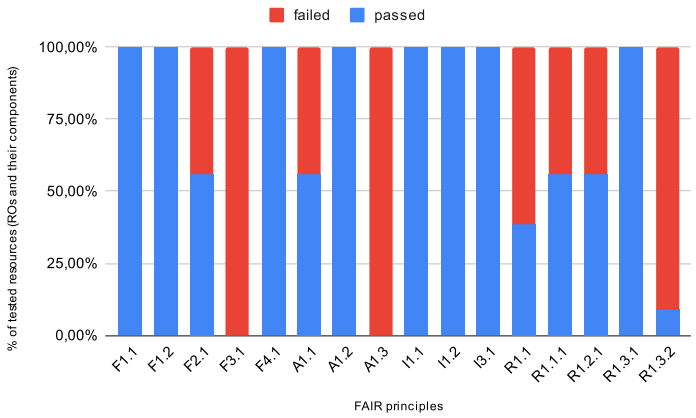

Fig. 4: Percentage of resources (i.e., ROs and their associated components) with passed or failed tests for each FAIR principle.

4.2 FAIRness assessments results

Figure 4 shows the percentage of resources (i.e., ROs and their respective components) that pass all the tests associated with each principle in the analyzed ROs. For some principles, all the test pass, as explained below:

- F1.1: Data is assigned a globally unique identifier. ROHub generates by default persistent identifiers based on the w3id service (https://www.w3id.org)
- F1.2: Data is assigned a persistent identifier. Same reason as F1.1.
- F4.1: Metadata is offered in such a way that it can be retrieved programmatically. ROHub describes ROs through a public API with content negotiation, which returns a machine readable format (JSON-LD).
- A1.2: Metadata is accessible through a standardized communication protocol. ROHub uses standard web mechanisms to return RO-Crate metadada.
- I1.1: Metadata is represented using a formal knowledge representation language. ROHub uses JSON-LD, which represents knowledge using the RDF standard.
- I1.2: Metadata uses semantic resources. RO-Crate relies on Schema.org, a popular vocabulary to describe resources on the Web.
- I3.1: Metadata includes links between the data and its related entities. RO-Crates include the relation *hasPart* to link resources to the the main RO.
- R1.3.1: Metadata follows a standard recommended by the target research community of the data. All analyzed ROs follow the RO-Crate specification, currently recommended by the Research Object community.

Interestingly, all ROs fail the tests associated with two FAIR principles:

[9] https://www.reliance-project.eu/

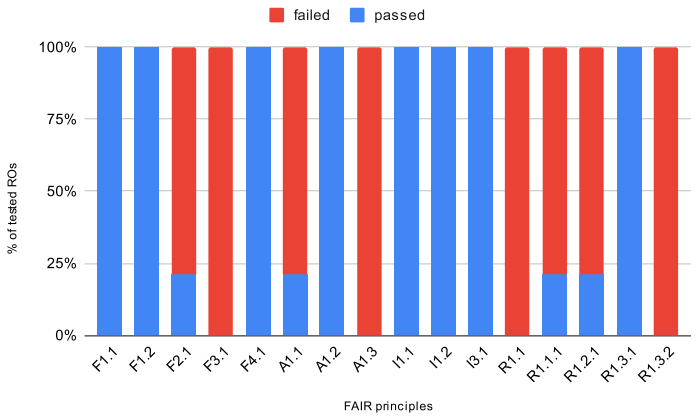

Fig. 5: Percentage of ROs passing all tests (i.e., all their associated resources also pass the tests) for each principle.

. F3.1: Metadata includes the identifier of the data it describes. RO-Crate file generated by the platform ROHub does not contain a link to its corresponding data file.
. A1.3: Data is accessible through a standardized communication protocol. The RO-Crates analyzed do not contain a download link for their associated resources.

However, it is worth noting that the download information of both ROs and their resources is available through their HTML representation in ROHub.

For the rest of the principles, the analyzed ROs and their components show similar results ($40\% - 60\%$ pass all tests). This is due to three main reasons: i) many of the tests associated with FAIR principles depend on metadata provided by the users, which may be absent (e.g., keywords, summary, copyrightHolder), ii) missing metadata fields that may be automatically completed by ROHub such as publicationDate, and iii) missing metadata fields for RO resources such as encodingFormat and contentType, needed to interpret resources correctly.

Figure 5 goes a step further by showing distribution of the principles covered taking into account if both the RO and its included resources pass all tests for each principle. As expected, the results show that the percentage of Research Objects that not cover entirely a FAIR principle are quite similar. This finding is consistent with the previous analysis in Figure 4, where we considered all RO components, potentially leading to varied percentages with a certain margin. Overall, Figure 5 provides a more refined perspective on the extent to which ROs successfully incorporate FAIR principles.

Our next step was to explore if ROs with a higher completeness score (i.e., a filled metadata checklist) had a better impact in their FAIR assessment. We analyzed only ROs with a completeness greater or equal to 50 (see Table 2).

Table 2: Percentage of ROs passing FAIR assessment tests by principle, based on their metadata completeness score (checklists)

Completeness	F2.1	A1.1	R1.1	R1.1.1	R1.2.1	R1.3.2
(0-49)	32,30%	32,30%	25,05%	32,30%	32,30%	1,78%
(50-100)	60,20%	60,20%	41,07%	60,20%	60,20%	10,25%
(0-100)	56,05%	56,05%	38,68%	56,05%	56,05%	8,99%

There is a slight improvement in the F2.1, A1.1, R1.1, R1.1.1, R1.2.1 and R1.3.2 principles, but we do not appreciate a significant improvement in the overall assessment. This is due to the fact that completeness is related with the RO metadata and not with the metadata of individual resources within the RO.

Finally, we analyzed ROs by type, detecting an improvement in the number of tests passed in the Data Centric Research Objects. In particular in:

. F2.1: Metadata includes descriptive elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support findability.
. A1.1: Metadata contains access level and access conditions of the data.
. R1.1.1: Metadata includes license information.
. R1.2.1: Metadata includes provenance information.

These ROs are based on a template, and their authors tend to include more information about licenses and provenance than other types of ROs. We believe this may be due to the domain and nature of the ROs (provenance and licensing is seen as key when reusing datasets).

5 Discussion

Do ROs (and following the RO-Crate specification) help align research artifacts with the FAIR principles? Our results have shown that passing the tests associated with all FAIR principles is conditioned by two factors: i) the mechanism by which a platform or repository provides persistent identifiers, appropriate metadata schemas, etc; and ii) the will of users to complete the metadata fields.

A platform like ROHub provides researchers with the means to create FAIR ROs, having persistent identifiers associated with each resource of the RO as well as describing semantically the RO and following a community-approved specification (RO-Crate). RO-Crates may also contain metadata from the resources they describe, which complement the tests run by external tools within FAIRness assessment services within FAIROs.

Figure 6 shows the impact of RO-Crate and ROHub when assessing the FAIRness of Research Objects, with nine of sixteen principles affected. The percentage of components assessed with external tools such as F-UJI are represented in blue; the assessment passed just by the analysis of the RO-Crate file is represented in greed and those resources which did not pass any FAIR assessment tests are shown in red.

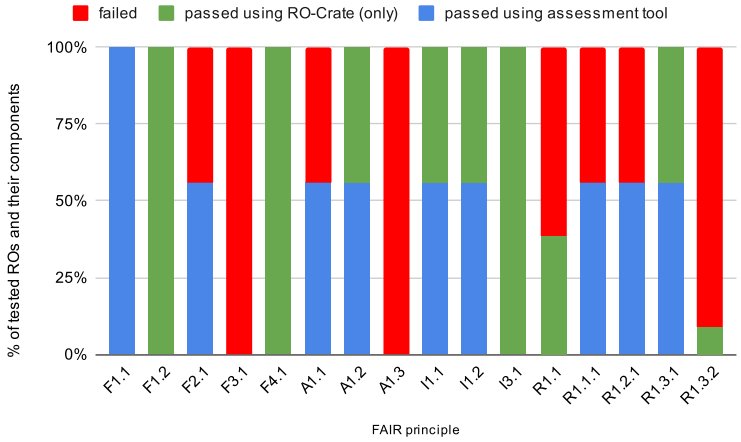

Fig. 6: Distribution of results per FAIR principle and tool used for the assessment

To increment the coverage of FAIR principles in ROs it is important to provide users with guidance on how to improve the FAIRness of their resources. Tools such as FAIROs can assist users in achieving this by highlighting the necessary information that needs to be included in the RO to improve its FAIRness. Additionally, platforms like ROHub have built-in mechanisms to automatically or manually include missing metadata fields, or make them mandatory to ensure they are completed. The aim is to make it as easy as possible for users to ensure their data meets the required FAIR principles.

Finally, in this work we have presented two mechanisms to measure the metadata quality of a RO: through completeness indicator (based on metadata checklists In ROHub) and through a FAIR assessment service (FAIROs). Completeness is an indicator used to analyze the metadata that is present and missing in a RO, and is meant to be provided by users. FAIR assessment implies not only the metadata present in a RO, but the implementation of mechanism to make a RO more Findable, Accesible, Interoperable and Reusable. In order to achieve FAIRness, we believe the platform/repository must provide the mechanisms to achieve these principles, e.g., with persistent identifiers, machine readable formats, etc.

6 Related Work

While a number of tools have been proposed to measure research artifact FAIRness [13, 7, 6, 3], there are barely any studies on FAIR Research Object fariness assessment. In our previous work [7] we introduced the FAIROs service, testing it against 160 workflow-centric ROs. In this work, we have integrated FAIROs in ROHub, extending the analysis to thousands of ROs in different scientific domains, refining both FAIROs and ROHub in the process.

The discussion on FAIR adoption is still open within the scientific community. The RDA Working Group FAIR Data Maturity Model is working in a model to compare different approaches of FAIR assessment. As a result, a collection of indicators (and their importance) has been made available, separated by FAIR principle [1]. However, the implementation of each indicator still depends on the interpretation of tool developers and their application domain (e.g., describing a resource with "rich metadata" may have different requirements for different types of artifacts or application domains).

The FAIRSFAIR initiative[10] also defined a collection of metrics to assess data FAIRness [4]. These metrics are based on the indicators defined by the RDA group, but with more detail. For example, rich metadata is defined as including fields like creator, title, publisher, etc. F-UJI uses these metrics [9], given an score to each test and a total score for each principle. These metrics are present in FAIROs through F-UJI to assess the datasets included in ROs.

Another approach for FAIR assessment uses a community-driven framework to assess digital objects [13]. The authors claim that some FAIR behaviours may be consider 'universal' but may be complemented with additional behaviours depending of the resource and the domain. Using description files like RO-Crate may play a role complementing these behaviours.

7 Conclusions and Future work

In this paper we described a comprehensive FAIR assessment analysis over a diverse collection of ROs published on the ROHub platform. Similar FAIR assessment results were observed in our analysis, regardless of the type, views/-downloads and completeness of the ROs. We believe this is attributed to two key factors: first, the implementation of FAIR principles is heavily reliant on the mechanisms incorporated within the RO-Hub platform (e.g., provision of persitent identifiers, including machine-readable information, etc.), and second, the authors of ROs do not consistently add supplemental metadata. We believe that integrating FAIROs as part of ROHub will help users becoming aware of some of this issue, helping them better align against the FAIR principles. However, additional studies are needed to track the impact of FAIROs in ROHub.

As shown in our analysis, using a specification like RO-Crate to describe Research Objects improves their FAIRness, especially when combined with RO-Hub. In fact, thanks to the integration and enrichment effort we have identified areas of improvement for ROHub which are part of ongoing work, such as including key metadata in the RO-Crates generated by ROHub (e.g., pointer to the data and RO download URLs, publisher date, etc.).

As part of our future work, we plan to improve the FAIROs module in charge of the research software assessment based on ongoing discussions within the community, such as the Research Data Alliance Research Software Workshop.[11]

[10] https://www.fairsfair.eu/

[11] https://fair-impact.eu/events/fairimpact-events/research-software-workshop-guidelines-and-metrics-metadata-curation

References

1. Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., Russell, K., Stall, S.: The fair data maturity model: An approach to harmonise fair assessments. (2020)
2. Bechhofer, S., De Roure, D., Gamble, M., Goble, C., Buchan, I.: Research objects: Towards exchange and reuse of digital knowledge. Nature Precedings pp. 1–1 (2010). https://doi.org/10.1038/npre.2010.4626.1
3. Devaraju, A., Huber, R.: F-UJI - An Automated FAIR Data Assessment Tool (Oct 2020). https://doi.org/10.5281/zenodo.4063720
4. Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R., Herterich, P., de Vries, J., Akerman, V., L'Hours, H., Davidson, J., Diepenbroek, M.: From conceptualization to implementation: Fair assessment of research data objects. Data Science Journal **20**(1), 1–14 (2021)
5. Fouilloux, A., Foglini, F., Trasatti, E.: FAIR Research Objects for realizing Open Science with RELIANCE EOSC project. Research Ideas and Outcomes **8**, e93940 (2022). https://doi.org/10.3897/rio.8.e93940
6. Garijo, D., Corcho, O., Poveda-Villalón, M.: FOOPS!: An ontology pitfall scanner for the fair principles. International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks **2980** (2021)
7. González, E., Benítez, A., Garijo, D.: FAIROs: Towards FAIR Assessment in Research Objects. In: Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G.M., Golub, K., Ferro, N., Poggi, A. (eds.) Linking Theory and Practice of Digital Libraries. pp. 68–80. Springer International Publishing, Cham (2022)
8. Guha, R.V., Brickley, D., Macbeth, S.: Schema. org: evolution of structured data on the web. Communications of the ACM **59**(2), 44–51 (2016)
9. Koers, H., Bangert, D., Hermans, E., van Horik, R., de Jong, M., Mokrane, M.: Recommendations for services in a fair data ecosystem. Patterns **1**(5), 100058 (2020)
10. Mao, A., Garijo, D., Fakhraei, S.: Somef: A framework for capturing scientific software metadata from its documentation. In: 2019 IEEE International Conference on Big Data (Big Data). pp. 3032–3037 (2019). https://doi.org/10.1109/BigData47090.2019.9006447
11. Palma, R., Hołubowicz, P., Corcho, O., Gómez-Pérez, J.M., Mazurek, C.: Rohub—a digital library of research objects supporting scientists towards reproducible science. In: Semantic Web Evaluation Challenge: SemWebEval 2014 at ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers. pp. 77–82 (2014). https://doi.org/10.1007/978-3-319-12024-9_9
12. Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L.J., Coppens, F., Fernández, J.M., Garijo, D., Grüning, B., La Rosa, M., Leo, S., et al.: Packaging research artefacts with ro-crate. Data Science **Pre-press**, 1–42 (2022). https://doi.org/https://doi.org/10.3233/DS-210053
13. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)