FAIRSFAIR
Fostering Fair Data Practices in Europe

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

D2.3 Set of FAIR data repositories features

Work Package	WP2 - FAIR practices: semantics, interoperability and services
Lead Author(s) (Org)	Claudia Behnke (SURFsara), Christine Staiger (DTL)
Contributing Author(s) (Org)	Jessica Parland-von Essen (CSC), Leah Riungu-Kalliosaari (CSC), Yann Le Franc (e-SDF), Luiz Bonino (GO FAIR), Gerard Coen (DANS)
Due Date	28.02.2020
Date	29.01.2020
Version	1.0 DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION
DOI	10.5281/zenodo.3631528

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)

Abstract

This is the first set of guidelines developed by the FAIRsFAIR project to enable features for repositories which allow them not only to host FAIR digital objects, but also to be FAIR themselves.

Versioning and contribution history

Version	Date	Authors	Notes
0.1	06.12.2019	Claudia Behnke (SURFsara), Christine Staiger (DTL), Jessica Parland-von Essen (CSC), Leah Riungu-Kalliosaari (CSC), Yann Le Franc (e-SDF), Luiz Bonino (GO FAIR), Gerard Coen (DANS)	Contribution via materials prepared for workshops and in team documents
0.2	06.12.2019	Claudia Behnke (SURFsara), Christine Staiger (DTL)	First draft
0.3	31.01.2020	Yann Le Franc (e-SDF), Luiz Bonino (GOFAIR) Claudia Behnke (SURFsara)	Ready for internal review
0.4	14.02.2020	Claudia Engelhardt (UGOE), Gabin Kayumbi Kabeya (STFC)	Internal review
1.0	21.02.2020	Claudia Behnke (SURFsara)	Inclusion of comments from the internal review

Disclaimer

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

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

Executive Summary

This report presents the results of the first year of Task 2.3[1] from the FAIRsFAIR project. It gives guidelines to enable features for repositories which allow them not only to host FAIR digital objects, but also to be FAIR themselves. The recommendations were collected in the workshop "Building the data landscape of the future: FAIR Semantics and FAIR Repositories" (22 October 2019, Espoo Finland) that was hosted by this task together with the FAIRsFAIR task 2.2. It derived input from more than 70 participants from 6 communities: the European Life Sciences Infrastructure for Biological Information (ELIXIR), the European Incoherent Scatter Scientific Association (EISCAT), the Social Sciences and Humanities (SSH), the Integrated Carbon Observation System (ICOS), the European network of Long-Term Ecosystem Research sites (eLTER), and the Data Publisher for Earth & Environmental Science **(**Pangea). The background of participants lied in infrastructures, research and libraries.

[1] This is one of the four tasks under FAIRsFAIR Work Package 2, see https://www.fairsfair.eu/fair-practices-semantics-interoperability-and-services

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

Table of Contents

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

From FAIR data principles to FAIR repositories

The FAIR principles (Wilkinson et al., 2016) give guidance for scientific data management and stewardship and are relevant to all stakeholders in the current digital ecosystem. They aim at improving findability, accessibility, interoperability and reusability of data and define expected behaviours for metadata, data and some supporting infrastructure elements such as search engines, identification systems, communication protocols, languages for knowledge representation and vocabularies. However, the authors formulated those behaviours as high-level guidelines, that require further interpretation and definition.

The FAIR principles are explicitly targeted at both metadata and data, with data here being regarded as any digital resource, asset or object (e.g., APIs, workflows, ontologies, models, and others). However, digital objects can not be made FAIR without supporting infrastructure services that are FAIR themselves. In this report, we focus on the "FAIRification" of repositories and give the first set of features which allows them not only to host FAIR digital objects but also to be FAIR themselves. Other tasks of FAIRsFAIR are focusing on an assessment of "FAIRness of services" in a broader sense.[2]

FAIR Digital Objects

According to RDA and the FAIR data principles (Wittenburg et al., 2018) a (FAIR) Digital Object, as shown in Figure 1, consists of:

- Bitstream(s) of the data, (e.g. files, pictures)
- Metadata information (descriptive, provenance, operational metadata) that describes attributes of the data and helps to extract information from the data
- A persistent identifier (PID) to unambiguously identify the data and parts of it.

Over time, the stakeholders learned that three different layers influence the FAIRness of digital objects: Machines, human beings and legal aspects of the data. Depending on the use case and purpose of the data, it has to be decided which parts of a digital object need to be made FAIR for machines, which improve the FAIRness and interoperability with human brains and which are ensuring, for example, legal reusability. In the lists of features at the end of this document, it is indicated if the point in question is addressing machines, humans, legal aspects or a combination of those.

[2] To be published under FAIRsFAIR Milestone 2.7 1st March 2020

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

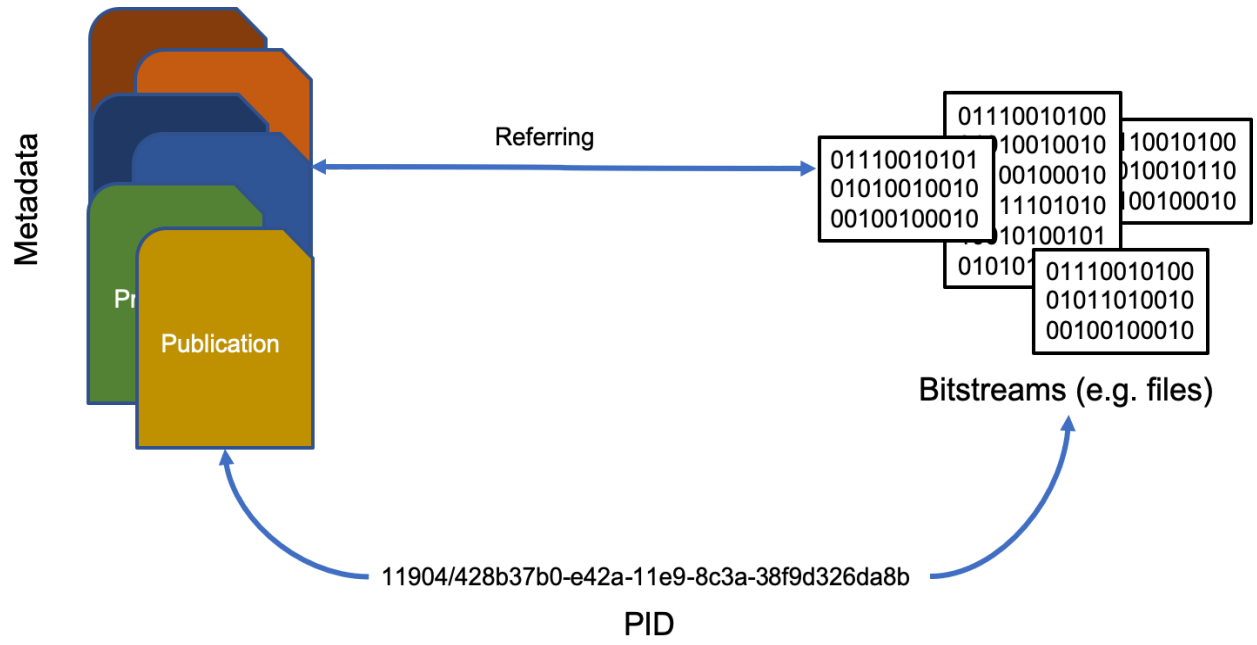

Figure 1: FAIR Digital objects - the holy trinity of data management.

Data infrastructures provide several services targeted at all or specific parts of the digital object, e.g. metadata services are specifically tailored to the handling of metadata (creation, linking, updating, etc.). Repositories deal with all aspects of digital objects (metadata, bitstreams and identifiers) and synchronise them throughout the lifetime of the data and sometimes beyond. Especially in the context of a repository, it is essential to note that FAIR only gives recommendations on the state of an existing digital object, but not about the duration of its existence. It is, therefore, crucial for data repositories to define what FAIR means for digital objects that once existed but have expired over time and align this definition with principle A2.

"FAIRification" of repositories

The Research Data Alliance has defined it as "a searchable and queryable interfacing entity that can store, manage, maintain and curate Data/Digital Objects. A data repository provides a service for human and machine to make data discoverable/searchable through the collection(s) of metadata."[3] In practice, there is much variation within and between repositories on how the questions of interoperability, machine-accessibility and, to some extent, reusability are addressed. We will, in this context, define a repository as a service that stores and gives access

[3] https://smw-rda.esc.rzg.mpg.de/index.php/Repository

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

(with needed restrictions) to research data and metadata, is searchable and offers persistent identifiers.

Data repositories play an essential role in scientific data management since their core business is to keep research data safe. They can provide functionalities in such a way that a higher level of FAIRness can be guaranteed to their users. At the moment, there is no convergence on what these functionalities could be. When describing a repository in a FAIR context, some of the principles themselves contain specific criteria that can be considered as requirements:

- Globally unique and persistent identifiers (F1)
- Structured metadata (F2, F3, A1, I1-I3, R1-R3)
- Index or other search functionality (F4)
- Metadata are accessible, even when the data is no longer available, for example via tombstone pages (A2)

Wu et al. (2019) propose the following recommendations for data repositories which focus in particular on metadata and data findability:

1. Provide a range of query interfaces to accommodate various data search behaviours.
2. Provide multiple access points to find data.
3. Make it easier for researchers to judge the relevance, accessibility and reusability of a data collection from a search summary.
4. Make individual metadata records readable and analysable.
5. Enable sharing and downloading of bibliographic references.
6. Expose data usage statistics.
7. Strive for consistency with other repositories.
8. Identify and aggregate metadata records that describe the same digital object.
9. Make metadata records easily indexable and searchable by major web search engines.
10. Follow API search standards and community adopted vocabularies for interoperability.

Currently, researchers are facing a plethora of different repositories (see Figure 2). All come with different graphical user interfaces and application programming interfaces (APIs). Hence, data and metadata are represented in various ways which are not easily combinable or even interoperable. Moreover, repositories hardly agree on schemes and forms to capture metadata. For example, some repositories might ask for the author, others for the depositor, distributor or publisher, but all could mean the same, i.e. the person that added the dataset.

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

Here we are faced with two problems. First, semantics are not always clear from those keywords, causing confusion when uploading data. Semantic interoperability is further addressed in the reports FAIR requirements for persistence and interoperability 2019 and FAIR Semantics: First recommendations[4].

Secondly, the same semantics can be denoted by different keywords which makes it necessary first to harmonise query results across several repositories, before the results become available. Furthermore, when querying several repositories at the same time, gathering information on query results and combining data sets, researchers face the problem of mapping different (meta) data models.

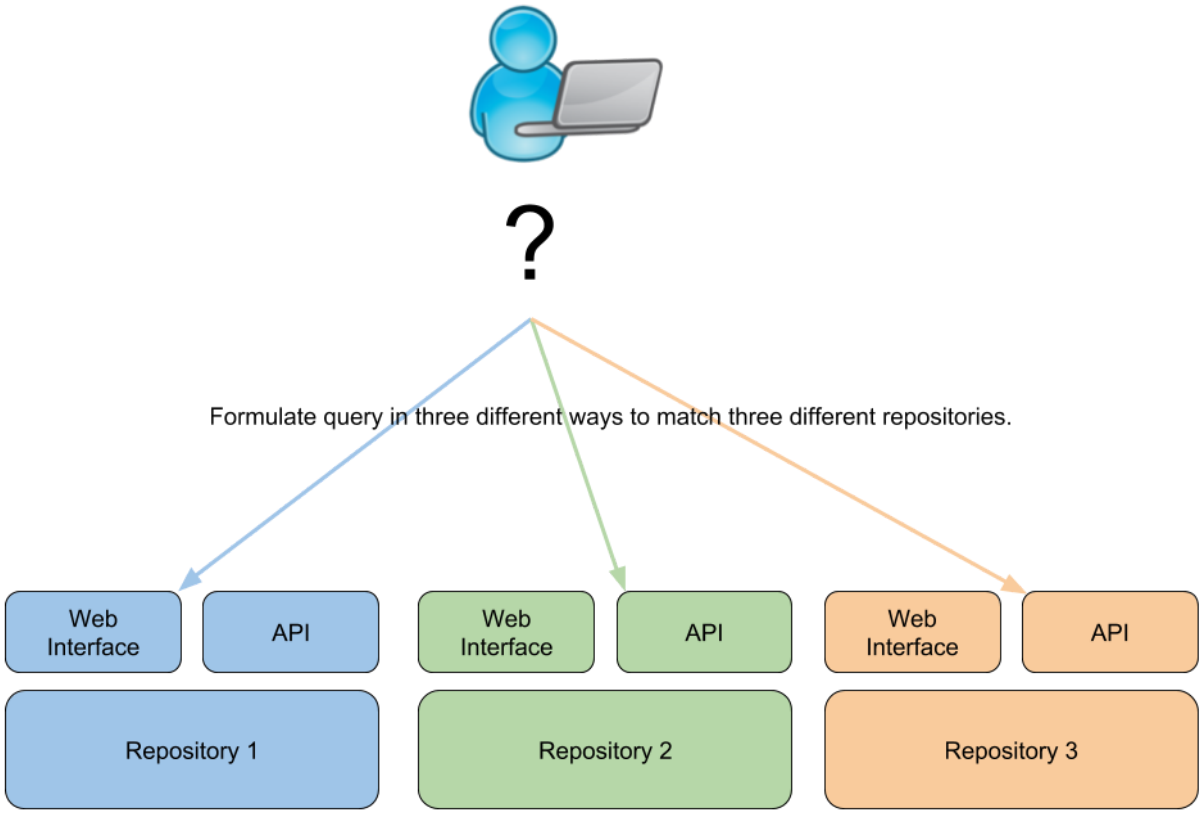

Figure 2: A user who tries to query several repositories at the same time with the same query.

Hence users querying a number of repositories at the same time without the help of an aggregator need to be aware and knowledgeable of many details of APIs, metadata schemes, and how repositories implement them, which hinders reuse of data. Repositories usually facilitate secure data (file) storage and capturing of metadata (information about the data).

[4] To be published as FAIRsFAIR Deliverable 2.2 on 1st March2020

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

Metadata aggregators do not store data themselves but try to harmonise metadata across repositories and offer extended search functionality.

As shown in Figure 3 below, metadata aggregators try to solve the problem of metadata harmonisation between repositories. Harvested metadata from different repositories can be searched by users simultaneously and hence, data from different sources can be linked. However, data which is not stored in the harvested repositories or data which is represented by other metadata aggregators is still not reachable. Hence metadata integration through metadata aggregators only provides a part of the solution and tends to elevate the general problem of (meta)data integration to the next level of the (FAIR) data infrastructure.

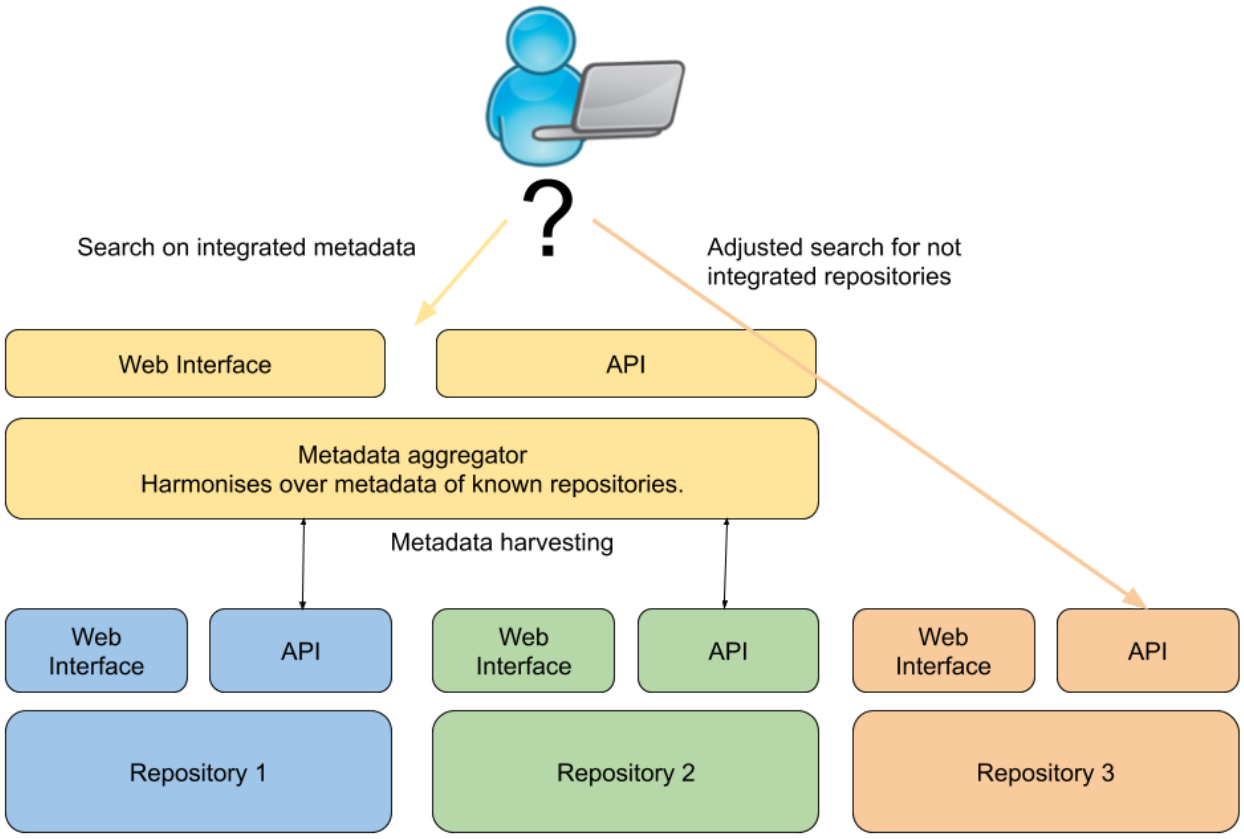

Figure 3: When repositories' metadata is only captured by integrated platforms and not by others, the problem of non-interoperable repositories is only shifted to another layer of the data infrastructure.

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

The goal of FAIRsFAIR is to alleviate these problems and seek harmonisation between repositories. This is achieved by offering a generic and abstracting interoperability layer, which improves the interoperability between repositories themselves to access data and information.

This is the first step towards the “FAIRification” of repositories by extending repositories interfaces to comply with FAIR practical guidelines, e.g. on the use of FAIR supportive standards or the FAIR Data Point specifications proposed by GO FAIR, defining common metadata representation formats, etc. In Figure 1, the arrow labelled "Referring" is what this task would like to minimise since it implicates a lot of manual updates if something changes. This is future work and partially implemented in the prototype, which is the scope of future deliverables (see Future plans). In the following, we would like to use the PID and some verb or extension to refer to elements of the data object.

The process to gather the requirements for FAIR repositories

To define the features which compose a FAIR data repository, Task 2.3 (together with Task 2.2) conducted the workshop “Building the data landscape of the future: FAIR Semantics and FAIR Repositories” (22nd October 2019, Espoo, Finland)[5]. It derived input from more than 70 participants from 6 communities: the European Life Sciences Infrastructure for Biological Information (ELIXIR), the European Incoherent Scatter Scientific Association (EISCAT), the Social Sciences and Humanities (SSH), the Integrated Carbon Observation System (ICOS), the European network of Long-Term Ecosystem Research sites (eLTER), and the Data Publisher for Earth & Environmental Science **(**Pangea), including representatives from the registries FAIRsharing and re3data.org. The background of participants lied in infrastructures, research and libraries.

The discussions and knowledge exchange between the participants and members of the FAIRsFAIR project led to a set of recommendations and technical requirements as well as suggestions which were not directly related to FAIR but nevertheless improve the reusability of data stored in data repositories.

In the following, those recommendations are converted into requirements on an organisational or a technical level. Some of the recommendations affect both aspects. Furthermore, the connection of the individual recommendations to the corresponding FAIR principle(s) is indicated in parentheses after each recommendation.

[5] https://www.fairsfair.eu/events/building-data-landscape-future-fair-semantics-and-fair-repositories

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

Organisational requirements

Here we list requirements which can not be implemented on the technical level of data repositories but are targeted at service level agreements, further agreements between users and repositories or communities and data providers.

- The repository itself should have a PID (FA)
- The repository needs to be listed in registries of repositories (F)
- Explicit data deletion policy - explicit roles and responsibilities (I)
- Different access policies for different versions of the data (A)
- Technical support for predefined file formats (I)
- Reuse of community standards and ontologies from public registries (FI)
- Use of PIDs as the manifestation of a data policy (I)
- Only mint one PID per digital object, collection or what one wants to identify (IR)
- Explicit data policies (like versioning and dynamic data) and PID policies in human and machine interoperable way (FAIR)
- Documentation of interfaces and APIs (FAIR)

Technical requirements

Below we provide a list of technical features which will improve the FAIRness of data repositories. Currently, those features do not suggest any specific implementation or technology. However, if implemented, they should nevertheless improve the interoperability between data repositories.

- Metadata for digital objects:
    - The repository should provide metadata in different formats, which can be harvested by different search engines (I)
    - Metadata should be provided as RDF, including JSON-LD. Based on these machines can provide human-friendly presentations/visualisations by resolving the URIs and retrieving the human-readable labels (I)
    - Providing metadata at the level of files, variables, attributes, individual cells, granularity to be decided by the repository (I)
    - Gather provenance metadata on digital objects and files upon upload (IR)
    - Provide masks and ways to quickly upload metadata (I)

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

  - Demand fine-grained metadata from data providers (FI)
  - Implement community standards (FI)
  - Automatic ontology suggestions and lookup (IF)
  - Landing pages should be machine-interpretable or implement content negotiation, have metadata in different formats (FI)
  - HTTP header should contain technical metadata about the DO (FI)
- Machine-readable and interpretable metadata about repository itself (I)
- Expose (Meta) Data Model (in machine-readable form) (I)
- PID policies
  - PID for each digital object or file (I)
  - Use global persistent identifiers (I)
  - The target of PID should be inferable by machines from PID metadata itself, employ PID information types or Linked Data type (I)
- Data object and file requirements
  - Connect compute infrastructures and data repositories (to avoid commuting data) (I)
  - Subsetting of data (I)
  - Technical support for predefined file formats (including complex data formats like netCDF), with a preference for open file formats (FI)
- Machine-readable license (R)
- The repository should provide a search interface or be linked to aggregating services that enable findability (F)

Not directly linked to FAIR

During the discussions, we also identified requirements which indirectly affect FAIR. Although those features fall out of the scope of the report, we would like to mention them here as recommendations.

- The repository should:
  - Support dynamic data sets (f.e. time series data)
  - Sent notifications to the creator if similar data appears elsewhere
  - Publication tracker for associated datasets
  - Have clear Service Level Agreements
  - Allow citation of reuse of partial data or single elements of datasets
  - Have downloadable citations (BibTeX) that point to the data

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

    - Variety of access restrictions
    - Tombstone procedure
- The repository search interface should have high usability.
- Repository staff should:
    - Provide training on APIs
    - "Spend time being a researcher to better understand the challenges they have making data available in a way that supports findability."

Future Plans

The authors of this article are responsible for delivering a set of specifications for the implementation of the identified features to improve the support of the FAIR principles by repositories. Based on these specifications, the task is also responsible for delivering a reference implementation aiming at helping repository application developers and client application developers to implement the specifications.

With these goals, the progress of the work will happen in these following steps:

- Initial specifications (Spec beta 1) of FAIR repository focusing on metadata content (Q1-2020);
- Workshop on common metadata interfaces for FAIR repositories (Q2-2020);
- The initial reference implementation (Impl beta 1) based on Spec beta 1;
- The second version (Spec beta 2) of the specifications including new elements such as the recommendations for semantic metadata schema for semantic artefacts repositories (Q3-2020);
- The second version of the reference implementation (Impl beta 2) based Spec beta 2 (Q4-2020);
- Workshop to engage with repositories to work on how they can implement the specifications in their solutions (Q1/Q2-2021);
- Evaluation of the discussions of the previous workshop and adjustment and release of the final version of the specifications (Spec v.1) (Q3-2021);
- Release of the final version of the reference implementation (Impl v.1)

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

References

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018 https://doi.org/10.1038/sdata.2016.18

Wittenburg, P., Strawn, G., Mons, B., Boninho, L., Schultes, E., 2018. Digital Objects as Drivers towards Convergence in Data Infrastructures. https://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11

Wu, M., Psomopoulos, F., Khalsa, S.J., Waard, A. de, 2019. Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. Data Sci. J. 18, 3. https://doi.org/10.5334/dsj-2019-003

Lehväslaiho, H., Parland-von Essen, J., Behnke, C., Laine, H., Riungu-Kalliosaari, L., Le Franc, Y., & Staiger, C. (2019). D2.1 Report on FAIR requirements for persistence and interoperability 2019 (Version v1.0 Draft). FAIRsFAIR. https://doi.org/10.5281/zenodo.3557381

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