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

M4.7 IMPROVED DESCRIPTION OF DATA REPOSITORIES

Work Package	WP4, FAIR-Certification
Lead Author (Org)	Sarala Wimalaratne (DataCite) and Robert Ulrich (DataCite-Re3data)
Contributing Author(s) (Org)	
Due Date	31.08.2020
Date	17.09.2020
Version	1.0
DOI	https://doi.org/10.5281/zenodo.4590336

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)

Summary

This document describes the work that is being done to improve re3data registry to enable researchers identify FAIR enabling repositories to find relevant datasets or deposit their research data. We are taking input from multiple efforts such as CoreTrustSeal plus FAIR alignment of repository practice, FAIR data repositories features and FAIR Object Assessment within FAIRsFAIR to identify specific properties within re3data schema that can be used or extended to capture criteria for FAIR enabling repositories. In parallel, we are also developing APIs to provide access to the re3data metadata and a search application to discover FAIR enabling repositories. The outcome of this is an improved description of repositories, hence enhanced metadata. This work will set the foundation for streamlining assessment and certification through improved organizational and data collection metadata.

Background

re3data

re3data is a registry of research data repositories (RDR) from different research disciplines. It stores metadata of repositories which provides access to research datasets to the scholarly community. Since 2016, re3data is a service of DataCite promoting a culture of sharing, increased access and better visibility of research data. DataCite is a global non-for-profit organization that is actively involved in several initiatives to improve the availability and citation of research data.

The re3data.org metadata schema has an extensive set of metadata properties describing a research data repository such as its general scope, content and infrastructure as well as its compliance with technical, quality and metadata standards. The schema includes required metadata properties and optional properties providing additional information.

The schema serves the purpose of:
- recommending a standard for describing a research data repository;
- providing the basis for identifying and referencing research data repositories in the research data landscape;
- helping data repositories to be visible and as a recommendation towards shared standards and practices.

As part of the FAIRsFAIR work we are in the process of selecting relevant metadata properties to identify repositories as 'FAIR Enabling' as well as introducing new metadata properties where needed (Table 1). Such use cases are collected by the team, the editorial board and the re3data working group. They are then incorporated into a new metadata schema draft and published as a RFC-Version to collect final feedback from the community before releasing a new version.

Repository Finder

Repository Finder allows researchers to search for repositories in which to deposit their data. The tool has been developed in the Enabling FAIR data project led by the American Geographical Union. It provides an easy to use interface to lookup recommended repositories. It relies on re3data as a data source but provides a simpler interface with predefined recommended filters to look up repositories.

For the first iteration we have enabled users to search using 3 criteria:
- The repository provides open access to its data
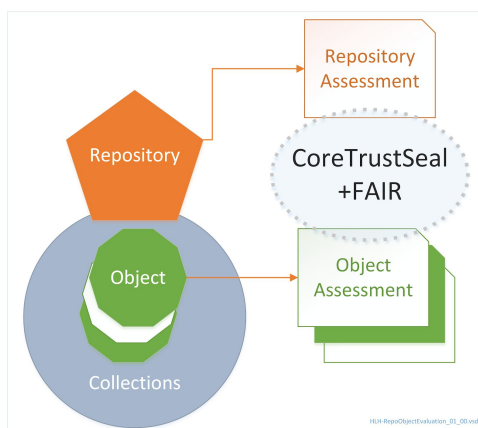- The repository uses persistent identifiers
- The repository is certified

FAIRsFAIR work alignment

A number of parallel work has been carried out to identify resources as FAIR at the data level and repository level. We will engage with this work and outcomes will be integrated into future iterations.

CoreTrustSeal plus FAIR

As part of WP4, task 4.1 work is being carried out to align CoreTrustSeal trustworthy data repository requirements and the FAIR data principles (CoreTrustSeal plus FAIR). This includes support to inform repositories seeking to enable FAIR data and metadata for the long-term. re3data provides information about CoreTrustSeal-certification and is updating its requirements periodically. In the project timeframe, there is no formal process of FAIR enabled certification through CoreTrustSeal, although recommendations for integration are being shared and discussed with the CoreTrustSeal Board. Re3data will engage with this work and outcomes will be integrated into the metadata.

FAIR data repository features

FAIR data repositories feature deliverable by FAIRsFAIR task 2.3 provides first recommendations to enable features for repositories which allow them to host FAIR digital objects, and constitute FAIR-aligned infrastructure themselves. Following these features, the table below shows the mapping between re3data metadata properties to the FAIR data repository features.

Addressing the repositories level directly, re3data as a registry cannot capture all of the listed FAIR data repository features within its metadata as they would be too extensive for the indexing processes and a core repository description. For example, *"Providing metadata at the level of files, variables, attributes, individual cells, granularity to be decided by repository (I)"* is commonly not visible on the repository website and would require an intensive communication process with the repository owner. Thus, it is not feasible for the given resources in the re3data

Editorial Board to go into such granularity. In addition to that, indexing of details requires a broad knowledge on the level of the field, which cannot be covered by one person in the staff. Also subject related properties are difficult to capture with a generic background. Thus, the re3data metadata is tradeoff in the level of detail it covers. Properties that in our experience cannot be mapped, are grayed out in the table below.

	FAIR data repositories features	re3data metadata properties
	Policies	
1	The Repository itself should have a PID (FA)	identifier/DOI/repositoryidentifier
2	The Repository needs to be listed in registries of repositories (F)	If it is in re3data this property is full-filled
3	Explicit data deletion policy - explicit roles and responsibilities (I)	policy
4	Different access policies for different versions of the data (A)	policy
5	Technical support for predefined file formats (I)	contentType
6	Reuse community standards and ontologies from public registries (FI)	metadataStandard
7	Use PIDs as manifestation of a data policy (I)	pidSystem
8	Only mint one PID per data object, collection or what one wants to identify (IR)	pidSystem
9	Explicit data policies (like versioning and dynamic data) and PID policies in human and machine interoperable way (FAIR)	policy
10	Documentation of interfaces and APIs (FAIR)	api
	Technical requirements	
11	Metadata for data objects	metadataStandard or pidSystem in case of DOI
12	The Repository should provide metadata in different formats, which can be harvested by different search engines (I)	pidSystem
13	Providing metadata at the level of files, variables, attributes, individual cells, granularity to be decided by repository (I)	
14	Gather provenance metadata on data objects and files upon upload (IR)	
15	Provide masks and ways to easily upload metadata (I)	

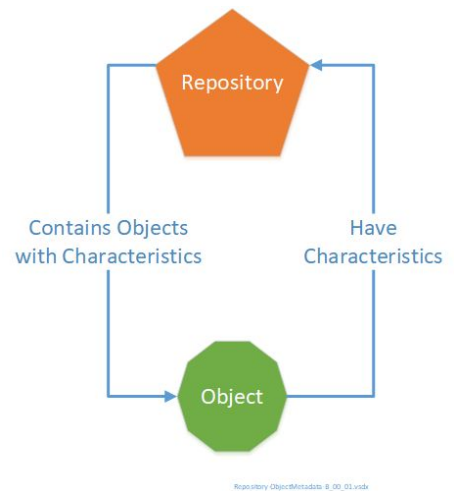| 16	Demand fine grained metadata (FI)	
17	Implements community standards (FI)	metadataStandard
18	Automatic ontology suggestions and lookup (Reference to Task 2.2) (IF)	
19	Landing pages should be machine interpretable or implement content negotiation, have metadata in different formats (FI)	api
20	HTTP header should contain technical metadata about the DO (FI)	
21	Machine readable and interpretable metadata about repository itself (I)	If it is in re3data this property is full-filled
22	Expose Data Model (in machine readable form) (I)	
23	PID policies	policy or pidSystem
24	PID for each data object or file (I)	pidSystem
25	Use global persistent identifiers (I)	pidSystem
26	Target of PID should be inferable by machines from PID metadata itself, employ PID information types or Linked data type (I)	
27	Tombstone procedure (FR)	policy
28	Data object and file requirements	
29	Bring compute to data (to avoid commuting data) (I)	
30	Subsetting of data (I)	
31	Technical support for predefined file formats (including complex data formats like netCDF), hereby prefer open file formats (FI)	api
32	Machine readable license (R)	databaseLicense
33	Repository should provide a search interface or be linked to aggregating services that enable findability (F)	If it is in re3data this property is full-filled
	Not directly linked to FAIR or repositories	
34	Repository should offer good search interface	If it is in re3data this property is full-filled
35	Support for dynamic data sets (f.e. time series data)	
36	Notification of creator if similar data appears	
37	Publication tracker for associated datasets	enhancedPublication

38	Repository staff should spend time being a researcher to better understand the challenges they have making data available in a way that supports findability	
39	Provide training on APIs	
40	Variety of access restrictions	dataAccess/databaseAccess
41	Clear SLAs	policy
42	Downloadable citations (bibtex) THAT POINT TO to the data	pidSystem(DOI)
43	Citation of re-use of partial data or single elements of data-set	enhancedPublication

Table 1: Mapping between FAIR Data Repositories features and re3data metadata

FAIR Object Assessment

FAIRsFAIR task 4.5 provides - inline with other initiatives and through several iterations - a set of core metrics to measure the extent to which research digital objects are FAIR. In parallel, it runs pilots to support the assessment of FAIR data within trustworthy repositories. As repositories state to support FAIR data objects, an object level assessment for FAIRness will provide proof and automated verification of certain aspect of FAIR enabling properties. Other properties relevant to FAIR, such as PID curation, long time archiving etc. need to be addressed via other methods. The preliminary work resulting in the report on FAIR data assessment mechanisms shows the metadata about the repository is an enabler for automated assessment. DataCite supported those efforts with providing links from DOIs to the repository descriptions, still the prototype revealed shortcomings in machine-2-machine communication, e.g. no complete coverage of the repositories APIs.

FAIR Enabling Repository Criteria

As the work packages address FAIRness from different perspectives, re3data metadata is able to cover and relate to many criteria proposed by the different work packages. Still some properties have to be refined and as a registry with a review process, it is not possible to curate

the information on fair alignment in full detail. Certain aspects have to be covered by the repository itself or other entities like certification authorities.

Metadata schema

As for the metadata schema the following fields have been identified to indicate FAIR-aligned infrastructure. They are currently defined for human readability and a manual editorial process. That imposes limitations to automated processing of property values, due to broad semantics or missing specifications on API details etc., as well as filtering for different levels of FAIR-alignment.

Property -Id	Property	Definition	Comment
1	identifiers	The identifiers provided by DataCite re3data	Persistent identifier to reference the FAIR-aligned repository in other services, systems and tools
4	repositoryUrl	The URL of the RDR.	
5	repositoryIdentifier	An identifier provisioned for the website of the RDR (wrapper element).	Identifiers that have been given by external bodies, e.g. community specific services. Enables discovery of subject specific information
15	contentType	All types of resources available in the RDR.	Needs to be extended to provide content and files types as machine readable information
19	policy	Policies providing information concerning the usage of the RDR (wrapper element).	Needs to be extended to reference policies covering FAIR and related community standards
20	databaseAccess	The access regulation to the RDR (wrapper element).	
21	databaseLicense	The database license of the RDR (wrapper element).	
22	dataAccess	The access regulation to the research data provided by the RDR (wrapper element).	
23	dataLicense	The license of the research data, existing in the RDR (wrapper element).	
28	api	The API supported by the RDR (wrapper element).	Needs to refined for automated discovery, e.g. link to FAIR data points or DCAT-catalogues

29	pidSystem	The persistent identifier system that is used by the RDR	
34	enhancedPublication	The RDR offers the interlinking between publications and data.	
36	certificate	The certificate, seal or standard the RDR complies with	Needs to reflect possible FAIR-Amendments of certificates
37	metadataStandard	The metadata standard the RDR complies with (wrapper element).	Needs to be extended to support automated discovery and assessment as well as reflecting community standards
NEW	Trust/Assessment results	Verification and results of (automated), assessments, e.g on object, repository or service levels.	Currently assessment and verification is not supported in the re3data metadata and needs to added.

Table 2: re3data metadata properties related to FAIR-enabling repository criteria

With output of the other FAIRsFAIR work packages, the schema will be extended to support the given requirements and improve machine readability and description of FAIR-enabling properties. At present a questionnaire is opened for the re3data stakeholders, where members of FAIRsFAIR next to others will contribute the re3data service model resulting in metadata schema revision adopting the current developments in the research data landscape.

Technical Improvements

APIs

With the changes in the upcoming metadata schema, the re3data API will be updated, e.g. to get to the object and repository level information as well as FAIR certificates (if any). We will also work on better alignment with the DataCite-API to support FAIRsFAIR object assessment. It is planned to extend the platform so that the trust-related information, e.g. CTS-certification or FAIR object assessment result, can be retrieved automatically from external platforms, for example the CoreTrustSeal web service, and kept uptodate.

DataCite Commons

DataCite is building DataCite Commons as part of the FREYA project, another European Open Science Cloud (EOSC) related project, funded by the European Commission. DataCite Commons is a discovery service that enables simple searches using a single PID such as DOI, ORCID or ROR while giving users a comprehensive overview of connections between entities in the research landscape. In the next iteration of FAIRsFAIR development, we will support search

for repositories within DataCite Commons with links to organisation (ROR), people (ORCID) and works (DOIs) where possible. In addition we will indicate FAIR-alignment via extending the re3data badges.

Discussion

Given the current input, gaps in the repository descriptions have been identified and will be addressed in the metadata schema and DataCite services. Still as a registry with a specific scope, not all facets of FAIR-enabling can be covered soley. Instead it is necessary to link necessary resources and build a network to foster the required services and automation regarding FAIRness. This is also reflected in the recently started and ongoing alignment of the working groups as a necessary step towards broad implementation beyond prototypes and pilot repositories. Further adjustments are expected as users give feedback on the first implementations, adopting and profiting from FAIR data and the related infrastructures.