

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

M4.8 INTRODUCE ADDITIONAL COMPONENTS AND PRACTICES TO METADATA SCHEMA AND ALIGN THEM WITH FAIR DATA PRACTICES

Work Package	WP4, FAIR Certification
Lead Author (Org)	Sarala Wimalaratne (DataCite), Robert Ulrich (DataCite-Re3data)
Contributing Author(s) (Org)	Margarita Trofimenko (DataCite-Re3data), Ilona von Stein (DANS-KNAW), Mustapha Mokrane (DANS-KNAW), Herve L'Hours (UKDA), Joy Davidson (UEDIN-DCC), Patricia Herterich (UEDIN-DCC)
Due Date	28.02.2021
Date	03.03.2021
Version	1.0
DOI	https://doi.org/10.5281/zenodo.4590298

<input checked="" type="checkbox"/>	PU: Public
<input type="checkbox"/>	PP: Restricted to other programme participants (including the Commission)
<input type="checkbox"/>	RE: Restricted to a group specified by the consortium (including the Commission)
<input type="checkbox"/>	CO: Confidential, only for members of the consortium (including the Commission)

Table of contents

1. Summary	4
2. Repository Search	5
2.1. Repository Finder	5
2.2. DataCite Commons	6
2.3. Repositories in DataCite Commons	6
2.4. Repository Search in DataCite Commons	7
2.5. Stakeholders	8
2.6. Repository Identifiers	8
2.7. User stories	8
2.7.1. MUST	9
2.7.2. SHOULD	9
2.7.3. COULD	9
2.8. Wireframe	10
2.8.1. Search for a repository in Commons	10
2.8.2. Search output	10
2.8.3. View repository information	11
3. Support of integration of research data repositories in Commons	12
3.1. Metadata schema updates	12
3.1.1. Certification information	12
3.1.2. API description	13
3.1.3. Community profile property	14
3.2. API	15
3.3. Editorial Board support	16
4. Discussion	17



FAIRsFAIR
Fostering Fair Data Practices in Europe

Abbreviations and Acronyms

API	Application Programming Interface
AGU	American Geophysical Union
FAIR	Findable, Accessible, Interoperable, Reusable
FsF	FAIRsFAIR
RDF	Resource Description Framework
PID	Persistent Identifier

1. Summary

This document describes the integration of the Repository Finder into DataCite Commons to enable researchers identify FAIR enabling repositories to either deposit their research data or find research output, organizations or researchers related to the repository. First the specific user stories describe the requirements by the research community and are then translated into the product specification for advancing DataCite Commons and accompanying measures by re3data as data source.

The Repository Finder started as an first effort to make FAIR enabling infrastructures visible. It was developed in the Enabling FAIR data project driven by a coalition of groups representing the international Earth and space science community (COPDESS), convened by the American Geophysical Union (AGU). For FAIRsFAIR, the roadmap was set in Milestone 4.7¹ to advance this service to a broader audience, incorporating the requirements and findings of related works in FAIRsFAIR. As DataCite is merging its services to provide an improved overall utility to its customers, this will advance the repository finder from a simple search to become a part of the PID-graph. Thus it will link repository metadata with other entities of the research landscape. In addition the integration within DataCite Commons will also foster easier maintenance and ensure sustainability as well as future developments beyond the scope of the FAIRsFAIR project.

Drafting the transition to DataCite Commons and implementation of the related changes to the metadata schema and the API of re3data, this can only be initial work towards greater FAIRness, as there are still many open questions like community agreed PIDs for repositories, final implementation of a “FAIR certification” or integration with other catalogues describing repositories as FAIRsharing, RRID, CatRIS etc. DataCite will deploy the proposed advancements by 4th Quarter of 2021 and foster future works on these topics and evaluate reasonable developments to be integrated in the time to come.

¹ <https://10.5281/zenodo.4590336>

2. Repository Search

2.1. Repository Finder

Repository Finder² allows researchers to search for repositories in which to deposit their data. The tool was first developed in the Enabling FAIR data project led by the American Geographical Union³. It relies on re3data as a data source and provides an interface with predefined recommended filters to look up repositories.

For the first iteration of FAIRsFAIR work, DataCite extended the Repository Finder tool to enable users to search using 3 criteria (see Figure 1):

- The repository provides open access to its data
- The repository uses persistent identifiers
- The repository is certified

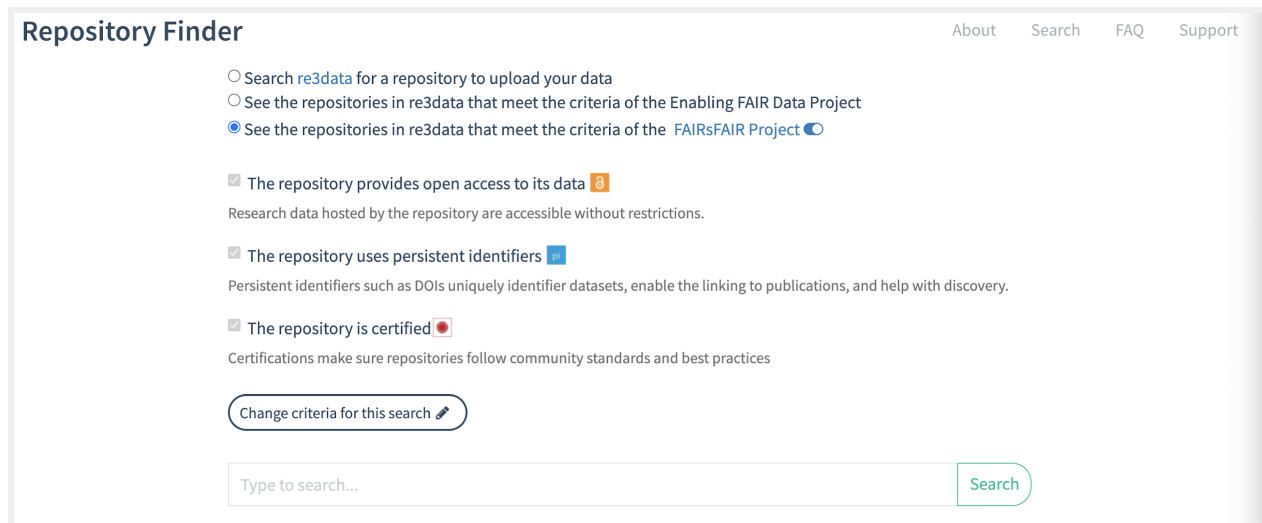


Figure 1: Repository Finder Tool

User feedback and discussions with Enabling FAIR data and FAIRsFAIR participants clarified that a more integrated system with links to relevant research outputs (e.g. publications, datasets, software), people and organisations would benefit the users. More targeted queries can be done via repository catalogues such as re3data. This has led to the consolidation effort of Repository Finder tool into DataCite Commons which is DataCite's integrated discovery service

² <https://repositoryfinder.datacite.org/>

³ <http://www.copdess.org/enabling-fair-data-project/>

for persistent identifiers (PIDs). The Repository Finder tool will be retired by the end of the year with the new support for repository search transitioned in DataCite Commons.

2.2. DataCite Commons

DataCite Commons⁴ is a discovery service that enables simple searches while giving users a comprehensive overview of connections between entities in the research landscape. Users can search for connections between research outputs (DOI), people (ORCID⁵) and organisations (ROR⁶). This connected graph is referred to as a PID graph. At the moment, DataCite commons consists of all of DataCite DOIs, ORCIDs and RORs and part of Crossref DOIs (see Figure 2). This year DataCite Commons will continue to add CrossRef DOIs and their connections.

Data Sources			
The following main data sources are used in DataCite Commons for a total of currently 42,466,919 records:			
DataCite	Crossref	ORCID	ROR
21,603,705 Works 100% of identifiers and metadata.	9,976,797 Works 8.23% of identifiers and metadata. Import is ongoing.	10,786,808 People 100% of identifiers. Personal and employment metadata.	99,609 Organizations 100% of identifiers and metadata.
Additional information comes from these data sources:			
<ul style="list-style-type: none"> Wikidata: inception year, geolocation and Twitter account for organizations Unpaywall: download link for Open Access content via Crossref 			

Figure 2: DataCite Commons content at February 2021

2.3. Repositories in DataCite Commons

DataCite members create DOIs within their managed repositories. DataCite currently stores over 2000 repositories with links to re3data where they exist. DataCite Commons can be used to search for DOIs from a specific repository by using DataCite internal repository identifiers (see Figure 3).

⁴ <https://commons.datacite.org/>

⁵ <https://orcid.org/>

⁶ <https://ror.org/>

DataCite Commons

client.uid:bl.cam

Pages
Support
Sign In

Works
People
Organizations

61,648 Works

Publication Year

☐ 2021
1,597

☐ 2020
12,098

☐ 2019
11,212

☐ 2018
12,121

☐ 2017
9,055

☐ 2016
5,639

☐ 2015
1,635

☐ 2014
1,148

☐ 2013
721

☐ 2012
643

☐ 2011
575

☐ 2010
469

Work Type

☐ Text
49,242

☐ Image
4,261

☐ Collection
3,746

☐ Audiovisual
2,515

☐ Dataset
1,824

☐ Software
24

☐ Interactive Resource
6

License

☐ CC-BY-4.0
8,212

☐ CC-BY-NC-ND-4.0
3,356

☐ CC-BY-NC-4.0
864

Tracer populations in the local group

Laura Louise Watkins

Thesis published 2011 in University of Cambridge

So often in astronomy, an object is not considered for its individual merits, but for what we may learn from its properties regarding some larger population. The existence of dark matter is a prime example of this; we cannot see it directly but we can infer its presence by noting its effects on the stars orbiting within its potential. This thesis describes how various sets of tracer populations can be used to probe the properties of a variety of galaxies in the Local Group. I begin by describing the extraction of a variable catalogue from the Sloan Digital Sky Survey Stripe 82 dataset and then use the catalogue to select a high-quality set of RR Lyrae stars. Analysing the distribution of the RR Lyrae reveals three significant substructures in the Milky Way halo: the Hercules-Aquila Cloud and the Sagittarius Stream, which were already known to exist, and the Pisces Overdensity, which was previously undetected. It is a faint, extended structure found at ~80 kpc and is of unknown origin. Altogether, I find that nearly 80% of the RR Lyrae are associated with substructures, consistent with the theory that galaxy halos are predominantly, or even entirely, made up from disrupted satellites. I also investigate the density distribution of RR Lyrae in the halo, finding that it is best fit by a broken-power-law model, in good agreement with previous work. I go on to develop a set of tracer mass estimators that build on previous work which make use of actual (and not projected) distance and proper motion data, reflecting the amount and quality of data now available to us. I show that proper motion data is, in theory, very useful and can greatly increase the accuracy of the mass estimates; in practice, however, current analysis is hampered by the large errors inherent in the proper motion data. The results are also subject to mass-anisotropy degeneracy, which current data is not yet able to break. Nevertheless, I am able to estimate the mass of the Milky Way to be $M = 2.7 \pm 0.5 \times 10^{12} \text{ Msun}$ and the mass of M31 to be $M = 1.5 \pm 0.4 \times 10^{12} \text{ Msun}$. Andromeda XII and Andromeda XIV are two M31 satellites that have been dubbed "extreme" and are thought to be on first infall into the M31 system. I modify the classical Timing Argument so that it can be applied to two external galaxies and then apply it to M31 and each of And XII and And XIV in turn to investigate the properties of their orbits. I then run a series of Monte Carlo simulations to investigate how likely su...

Other Identifiers

other: [PhD.34274](#)

uri: <http://www.dspace.cam.ac.uk/handle/1810/240582>

DOI registered May 23, 2016 via DataCite.

Text
English

<https://doi.org/10.17863/cam.3>

Figure 3: Searching for DOIs in Apollo (University of Cambridge's institutional repository) using DataCite Commons.

2.4. Repository Search in DataCite Commons

As part of FAIRsFAIR work, DataCite Commons will be extended to support:

- Separate repository search tab
- Showing of links to works/organizations/people
- Display of re3data repository information where they exist

The rest of this section details the Product Specification for improving DataCite Commons to deliver *D4.7 Tools for finding and selecting certified repositories for researchers and other stakeholders* towards the end of the project.

2.5. Stakeholders

The different types of stakeholders who might be interested in improving repository metadata or searching, include:

- Researchers
- Infrastructure and Service providers
 - Data repository providers
 - Data services (e.g. PID services, Registries)
- Journals
- Funder

2.6. Repository Identifiers

Currently there is no community agreed PID for repositories in the way that ORCID identifies people. There are several repository catalogues such as re3data, FAIRSharing, COAR, RRID, OpenDOAR etc. These repository catalogues have their own internal identifiers to identify repositories and in some cases assign DOIs for their metadata records e.g.: <http://doi.org/10.17616/R3SW4D> points to Apollo re3data record. COREF⁷ is an ongoing effort which is looking at recommendations for PIDs for repositories.

The PID graph essentially connects PIDs together. As there are no PIDs for repositories, in order to support repository search in DataCite commons an initial pragmatic solution will be used until there is a community agreed PID for repositories. To support this, DataCite will auto generate local identifiers (LIDs) for repositories so that DataCite Commons can use these LIDs in our PID graph. Once there is consensus around community PID for repositories, DataCite will adopt the new PID in Commons.

2.7. User stories

The following list of user stories will be addressed by DataCite Commons with progress through the year. The stories are prioritised following the MoSCoW method. The “Must” user stories will be addressed to satisfy FsF project requirements and deliver D4.7 while ensuring the transition of the Repository Finder in DataCite Commons. The “Should” user stories will be addressed on a best effort basis to support other work within FsF and the needs of important stakeholders. The “Could” user stories are part of additional functionalities for future improvements.

⁷ <https://blog.datacite.org/german-research-foundation-to-fund-new-services-of-re3data/>

2.7.1. MUST

1. As a user, I would like to search for repositories by their names. For example:
 - a. A researcher searching for a repository to deposit data and obtain a DOI
 - b. A librarian helping a student to find a repository
2. As a user, I would like to see repositories filtered by organization. For example
 - a. A researcher searching for repositories linked to their organization
 - b. A repository manager of an organisation making sure their repositories are available for discovery
3. As a user, I would like to see basic repository metadata eg: name, description, repository URL, disciplines, certifications, access conditions, licensing, persistent identifiers used, repository type, link to re3data
4. As a user I would like to filter repositories by their capabilities. For example:
 - a. A researcher searching for repositories that allow data deposits for specific formats, access controls etc
 - b. As a user, I would like to search for trustworthy (e.g. CoreTrustSeal certified) repositories
 - c. As a user, I would like to search for FAIR-enabling repositories
5. As a user, I would like to assess how up-to-date is the information about a repository
6. As a user, I would like to find community recommended repositories (e.g. by COPDESS/American Geophysical Union)

2.7.2. SHOULD

7. As a user, I would like to search for digital objects referenced by DOIs by filtering for repositories. For example:
 - a. Discovery objects held within the repository by tools like F-UJI⁸.
8. As a user, I would like to see related outputs, e.g. publications, data, software, for a given repository. For example:
 - a. A repository manager creating reports
 - b. A funder looking for impact of the objects stored in a repository they are funding

2.7.3. COULD

1. As a researcher, I would like to find repositories that meet certain criteria

⁸ M4.9 Report on Fair Data Assessment Mechanisms to Develop Pragmatic Concepts for Fairness Evaluation at the Dataset Level <https://doi.org/10.5281/zenodo.4118404>

2. As a user I would like to filter repositories by their self-declared long term preservation mission and sustainability.

2.8. Wireframe

DataCite has put together a set of wireframes⁹ to illustrate the user interfaces that will be supported by DataCite Commons.

2.8.1. Search for a repository in Commons

A new tab will be introduced to search for Repositories (see Figure 4). This will enable users to search for Repositories in the PID Graph.



Figure 4: Search for a repository in Commons

2.8.2. Search output

Search output will list relevant repositories (See Figure 5). Each search result will have a repository name and additional links to the repository homepage and re3data. DataCite Commons will also support a selected number of search facets to support the filtering options mentioned in the user stories, for example Certification, Community Profile, Subjects and PIDs to filter the search output.

⁹

<https://www.figma.com/proto/8xRL8uhipz0UYNePwfizHy/Repository-Search?node-id=2%3A5707&scaling=min-zoom>

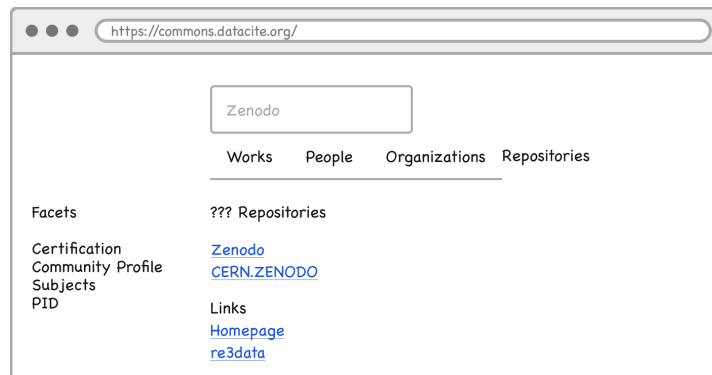


Figure 5: Search output

2.8.3. View repository information

Users can select a repository in the search list to see more metadata and associated connections (see Figure 6). The 'View Repository Information page' will show metadata stored in the DataCite metadata store as well as information retrieved from re3data where the links exist. In addition, users will be able to explore DOIs associated with the repository, potentially including usage and engagement information such as citations, views and downloads cumulative for the selected repository.

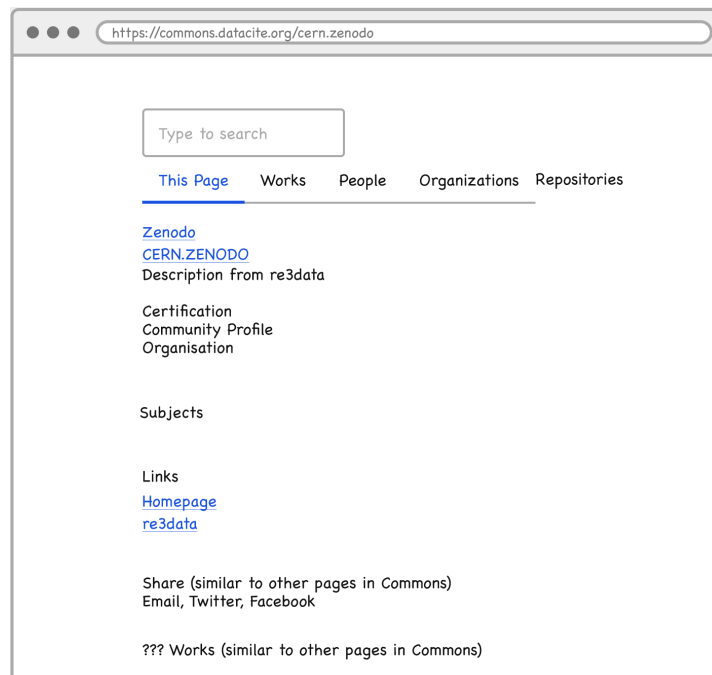


Figure 6: View repository information

3. Support of integration of research data repositories in Commons

DataCite Commons, as the resource discovery system, is going to integrate the repository search and linkage with other entities in the PID graph. This functionality has to be supported by re3data as the backend and require further changes and adoptions regarding metadata schema, the interfaces and current data stored in re3data.

3.1. Metadata schema updates

3.1.1. Certification information

Based on the input and suggestions by members of the CoreTrustSeal, the current description of repository certification in the re3data metadata schema will be extended. Those changes will improve the current data linking to the certificate for proof, showing the expiry date to highlight up to date information as well linking to badges that can be displayed in Commons and re3data, easily recognized by researchers.

#	Property	Description	O/C	Vocabulary/ Values
34	certificate	Wrapper element	0-n	
34.1	certificateName	The certificate, seal or standard the research data repository complies with.		Controlled vocabulary: CoreTrustSeal CoreTrustSeal+FAIR ...
34.2	certificateUrl	URL linking to details about certification. If the Id is blank, URL is displayed.	1	URL

34.3	certificateWidget	URL to support for displaying widgets (badges)	0-1	URL
34.4	validUntil	The date at which certification expires	0-1	DateTime

Example XML

```

<repository>
...
  <certificate>
    <certificateName>CoreTrustSeal+FAIR</certificateName>
    <certificateUrl>//cert.example/cert.pfd</certificateUrl>

    <certificateWidget>cert.example/badge.jpg</certificateWidget>
    <validUntil>2024-01-01</validUntil>
  </certificate>
...
</repository>

```

3.1.2. API description

As FsF work package 2 follows a modern linked data approach and fosters the implementation of the FAIR Data Points within repositories, those shall be linked within the current API description. Even so re3data currently does not model its information in RDF, this will be a first step towards linked open data and improves findability from registry information down to the datasets offered by the repository. As the implementation described in deliverable 2.6¹⁰ is reusing existing RDF vocabularies, repositories exposing its catalogue via plain DCAT will be supported too.

#	Property	Description	O/C	Vocabulary/ Values
28	api	Wrapper element	0-n	
28.1	apiType	The type of the API	1	Controlled vocabulary: FAIR Data Point DCAT ...

¹⁰ D2.6 First reference implementation of the data repositories features <https://doi.org/10.5281/zenodo.4501201>

28.2	apiUrl	The URL of the API	1	IRI resolving to the catalogue
28.3	apiDocumentation	A link referring to the API documentation, a website that states its availability and other information for using the API	1	URL

Example XML

```

<repository>
...
  <api>
    <apiType>FAIR Data Point</apiType>
    <apiUrl>api.example/catalogue</apiUrl>
    <apiDocumentation>api.example/doc</apiDocumentation>
  </api>
...
</repository>

```

3.1.3. Community profile property

The planned integration of the repository finder into DataCite Commons will preserve current functionality to researchers, which is an easy selection of repositories. As a pilot, the set of FAIR enabling repositories in the earth, space and environmental sciences have been implemented in the Repository Finder based on COPDESS community developed criteria. Currently those profiles are not tracked in the metadata, still relevant to researchers looking for repositories. In addition to making FAIR-enabling repositories identifiable, this will also be available to identify infrastructures recommend by disciplinary communities, publishers, funders or from other networks like DARIAH-EU, which developed a data deposit recommender¹¹ for repositories in arts and humanities.

#	Property	Description	O/C	Vocabulary/ Values
xx	profile	Wrapper element for the community profile	0-n	

¹¹ <https://ddrs-dev.dariah.eu/ddrs/about>



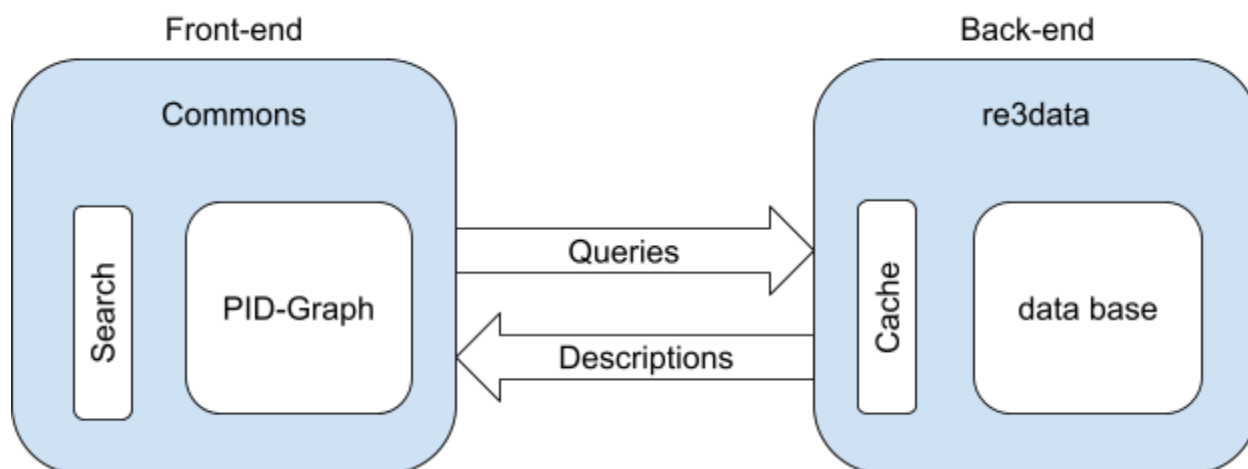
xx.1	profileTitle	Name to describe the set of repositories	1	Example: FAIR-enabling COPDESS/AGU ...
xx.2	profileUrl	Link to the profile description providing further information	1	URL

Example XML

```
<repository>
...
<profile>
<profileTitle>FAIR-enabling</profileTitle>
<profileUrl>fair-community.example/profile-description</profile
Url>
</profile>
...
</repository>
```

3.2. API

Drafting the architecture to connect different services, for the data exchange data, this needs to be reflected in the different APIs. To enable the transition from Repository Finder to functionality within DataCite Commons, re3data will implement the necessary interface to integrate into DataCite Commons providing essential repository description as well as the required facets, properties and links to be shown and filtered.



Following the best practices of the scientific community, re3data has implemented and established a review process. Every repository record ingest and change request is revised and reviewed by two members of an international Editorial Board¹². As this process ensures the quality of the re3data entries, information updates may be delayed. As a start for identifying different data providers and implementing an authentication and authorization concept later on, the current RESTful-API of re3data will be extended in collaboration with CoreTrustSeal to update certification information automatically.

The current mapping of the re3data metadata schema to RDF in FAIR Data Point will be further examined with the goal to make the resources using the re3data namespace, respectively the IRIs, resolvable. This is not required for the DataCite Commons integration but to support the efforts in FsF Work Package 2.

3.3. Editorial Board support

Within work package 4 of FAIRSFAR, 10 repositories are supported regarding CoreTrustSeal certification and an additional 12 are supported in work package 2 regarding the FAIR Data Point implementation. As 3 of the supported repositories are still missing in the re3data database, those will be indexed as the other entries will be checked by the Editorial Board and updated, if necessary.

Furthermore links in DataCite Commons to re3data entries will be analyzed by the Editorial Board for updates and extensions, thus improving data discovery and linkage in the PID-Graph,

¹² <https://www.re3data.org/editorialboard>



which has been used by the F-UJI tool¹³. Also the re3data team has started to index RORs for repository institutions and will incorporate ORCIDs into future schema updates to improve findability and discovery within the PID-Graph.

Together with FsF work package 3 and FAIRsharing, a handout on repository registries will be developed, providing guidance for repository managers to improve their records and therefore overall data quality of the registry entries.

4. Discussion

The drafted architecture and implementation plan to connect existing infrastructures regarding FAIR enabling and trustworthy repositories will be sustained by DataCite in its services Commons and re3data. It will be improved by adopting further developments on FAIR Data practices beyond the time horizon of FsF. Future extensions are also required, as needs and shortcomings have been exposed by related work, e.g. FAIR Data Point, F-UJI tools. As the FsF project started necessary work to roll out FAIR into practice, for now the planned implementations can only reflect those initial efforts as well as foster services building up on and starting the next iteration towards supporting researchers and pushing scientific work.

¹³ <https://www.fairsfair.eu/f-uji-automated-fair-data-assessment-tool>