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

D2.6 First reference implementation of the data repositories features

Work Package	WP2
Lead Author (Org)	Claudia Behnke (SURF)
Contributing Author(s) (Org)	Kees Burger (LUMC), Yann Le Franc, Wim Hugo (e-SDF), Pekka Järveläinen, Jessica Parland-von Essen (CSC), Gerard Coen (DANS)
Due Date	28.02.2021
Date	26.02.2021
Version	1.0
DOI	https://doi.org/10.5281/zenodo.4501201

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)

FAIRSFAIR
Fostering Fair Data Practices in Europe

Abstract

This document is the second report from Task 2.3 of the FAIRsFAIR project. It demonstrates how the requirements for FAIR enhancing repositories (Behnke et al., 2020) are put into practice by building a prototype based on the DCAT2 data model. The reference implementation's technical details, the used data model and custom extension, and the community uptake are discussed.

Versioning and contribution history

Version	Date	Authors	Notes
0.6	20.01.2021	Claudia Behnke & Kees Burger	First draft created
0.7	31.01.2021	Claudia Behnke (SURF), Kees Burger (LUMC), Yann Le Franc (eSDF), Wim Hugo (eSDF), Pekka Järveläinen (CSC), Jessica Parland-von Essen (CSC)	Draft for internal review
0.8	16.02.21	Robert Ulrich (KIT), Herve L'Hours (UKDA)	Internal Review
0.9	18.02.21	Claudia Behnke (SURF), Gerard Coen (DANS)	Incorporating remarks of the reviewers
1.0	26.02.2021	Claudia Behnke (SURF), Gerard Coen (DANS)	Final editing

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

Abbreviations and Acronyms

API	Application Programming Interface
DCAT	Data Catalog Vocabulary
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
FDP	FAIR Data Point
HEIs	Higher Education Institutions
IRI	Internationalized Resource Identifier
JSON-LD	JavaScript Object Notation for Linked Data
RDF	Resource Description Framework
SHACL	Shapes Constraint Language
SPARQL	SPARQL Protocol and RDF Query Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

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 document is the second report from Task 2.2 of the FAIRsFAIR project. It demonstrates how the features for FAIR enabling repositories (Behnke et al., 2020) are put into practice by building a FAIR Data Point[1] (FDP) prototype based on the DCAT2 data model, which is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web.

The root of the FAIRsFAIR reference implementation is a FAIR Data Point. In general, it serves three goals:

1. It allows a repository or any other holder or *editor*[2] of data to expose metadata in a FAIR manner, with a strong focus on the F, A, and R as described in the FAIR Principles (Wilkinson et al., 2016)
2. A consumer or *viewer* of the data can discover information that is stored in it.
3. It is optimised to interact with humans and machines.

The first chapter gives a recap on the motivation and maps requirements from Behnke et al. (2020) to the prototype. The second chapter explains the reference implementation's technical details, while the third describes the DCAT2 data model. In the fourth chapter, the authors analyse the community uptake, and in the fifth and last chapter, an outlook is presented.

[1] https://www.fairdatapoint.org/

[2] Italic words are names used in the FAIR Data Point documentation

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

1. Motivation - Expose Metadata in a more FAIR manner

This deliverable is describing the FAIRsFAIR T2.3 'reference implementation' covering distinct features to enhance the FAIRness of repositories, derived in the workshop "Building the data landscape of the future: FAIR Semantics and FAIR Repositories" (22nd October 2019, Espoo, Finland)[3] and summarised in the previous deliverable of the task (Behnke et al., 2020). Those will be discussed in Chapter 1.2 and 1.3.

Regarding the approach:

> The FAIR data point is built upon semantic web technologies and follows a Linked Data approach. Readers who are not familiar with Linked Data[4] may struggle through the more technical parts of the document. The lack of skills regarding these technologies, and the team's approach in helping to bridge the current knowledge gap, are outlined in sections 4 and 5.

1.1. What is FAIR about this approach?

One key aspect of making repositories FAIR enabling is exposing metadata in a more FAIR manner. The workshop participants assumed that the digital object itself, stored in the repository, is already as FAIR as possible[5]. Meaning, for example, is "described with rich metadata" (F2) and "(Meta)data are registered or indexed in a searchable resource" (F4,(Wilkinson et al., 2016)). However, even those data sets are not necessarily findable for machines since the metadata might not be presented in a machine-actionable way. Even if the metadata is machine-actionable, semantic interoperability might not be straightforward. During a digital workshop[6], the authors of this deliverable asked the participants if their metadata is enhanced with semantic lookup, meaning that their metadata uses semantic artefacts. Only 30% of the participants answered with "Yes", while another 30% said "No", and 40% were unsure, as shown in Figure 1.

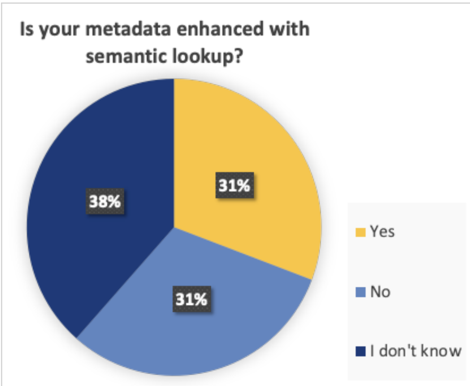

Figure 1. Answer by participants

After the publication of the previous deliverable (Behnke et al., 2020), the task validated the requirements based on technical constraints of the underlying technology and prioritised them based on the expected impact on the participating repositories. This work led to a list of

[3] https://www.fairsfair.eu/events/building-data-landscape-future-fair-semantics-and-fair-repositories
[4] https://www.w3.org/DesignIssues/LinkedData.html
[5] This excludes the extent to which the digital object itself is Interoperable: Findability, Accessibility, and Reusability are elements that can largely be addressed by rich metadata, but interoperability characteristics of the digital object itself requires compliance with community standards for schema and semantics.
[6] https://www.fairsfair.eu/events/workshop-common-metadata-interfaces-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

requirements (see appendix 7.1), presented in a web seminar (Claudia Behnke and Hylke Koers, 2020).

1.2. Requirements that are realised by the prototype

The following part summarises the requirements that the prototype will address and solve:

Expose (Meta) Data Model (in machine-readable form): The prototype will expose digital objects' metadata in a machine-readable and actionable format, allowing harvesters to find the data.

Machine-readable and interpretable metadata about the repository itself: The model used in the prototype stores the repository information as a built-in class. This metadata is aligned with the values required by re3data (Rücknagel et al., 2015).

Metadata should be provided as RDF[7], including JSON-LD. (Based on these machines can give human-friendly presentations/visualisations by resolving the URIs and retrieving the human-readable labels): The metadata is encoded in SHACL ("Shapes Constraint Language (SHACL)," n.d.), validating a set of conditions against RDF graphs. The metadata is exposed as RDF (in an abstract sense), with the preferred (concrete) serialisations of turtle and JSON-LD.

1.3. Requirements that are partially realised by the prototype

The reference implementation can partially satisfy some of the requirements since they need a further uptake in the communities and some implementation work from the repository itself.

Reuse community standards and ontologies from public registries: The prototype allows for adding any community standard by extending the default model. Due to the variety of existing community standards, not all of them are available by default. Furthermore, the model can be semantically enhanced, meaning relations between metadata fields can be defined. Within the FAIRsFAIR project, recommendations for interoperability of semantic artefacts (Hugo et al., 2020) will be the baseline.

Documentation of interfaces and APIs: The prototype provides one well-defined and documented API. Repositories are free to expose their data using any other API, which is not part of the project's scope.

The repository should provide metadata in a community-endorsed schema, which appropriate search engines can harvest: Even though the metadata exposed via the reference implementation is machine-actionable, it is provided in only one schema (in this case, Linked Data). This is maximising the reuse of existing vocabularies. Repositories are free to expose their data using any other metadata schema (this is beyond the project's scope).

The repository needs to be listed in registries of repositories: The repository can apply to any registry by itself. Registries such as re3data require that the mode of access to data and

[7]The wording was taken from participants at the Workshop "Building the data landscape of the future: FAIR Semantics and FAIR Repositories". The authors of this documents also endorse other formats, which serve the same purpose, like RDF(S) or OWL.

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

repositories is clearly explained. Since the FDP provides a harmonised approach to exposing metadata, it can facilitate repositories to meet criteria such as this more easily. FAIRsFAIR actively engages with CoreTrustSeal to combine the FAIR principles with requirements to become a trustworthy digital repository (L'Hours et al., 2020).

Machine-readable license: Not all licenses are machine-actionable. Communities should be encouraged to implement a narrow set of standardised, machine-actionable licenses, of which the Creative Commons license portfolio[8] is a practical resource for Open Data. The reference implementation uses the purl.org/NET/rdflicense list. It includes most versions of the most common licenses in machine-actionable format.

The repository should provide a search interface or be linked to aggregating services that enable findability: It is possible to connect repositories that have implemented the prototype to aggregating and query services, for example, by exposing the metadata through a SPARQL endpoint. Being listed in repository registries like re3data and resource discovery services like DataCite Commons, provide visibility, and therefore findability is improved[9].

2. The FAIRsFAIR Reference Implementation

The root of the FAIRsFAIR reference implementation is a FAIR Data Point[10] (FDP). In general, it serves three goals:

1. It allows a repository or any other holder or *editor* of data to expose metadata in a FAIR manner, with a strong focus on the F, A, and R as described in the FAIR Principles (Wilkinson et al., 2016)
2. A consumer or *viewer* of the data can discover information that is stored in it.
3. It is optimised to interact with humans and machines.

Furthermore, this technology is available as free and open-source software and documented[11]. The FAIR data point is by decision built upon semantic web technologies.

There are two different ways of reaching these goals. The first one is the deployment of a stand-alone application[5]. Second, it is also possible to extend existing applications using the guidelines provided by the FDP specifications[12]. FAIRsFAIR is using the first approach and a reference implementation where the metadata models are based on DCAT2. It consists of three main components (as shown in Figure 2.): the web client, the server, its APIs, and the triplestore.

[8] https://creativecommons.org/choose/results-one?q_1=2&q_1=1&field_commercial=n&field_derivatives=n&field_jurisdiction=&field_format=&field_worktitle=Leticia+Saito+-+Portfolio&field_attribute_to_name=Leticia+Saito&field_attribute_to_url=www.leticiasaito.com&field_sourceurl=&field_morepermissionsurl=www.leticiasaito.com&lang=en_US&n_questions=3

[9] Implementation work is scheduled in Task 4.4 of FAIRsFAIR and is going to reference FAIR DataPoint APIs based on this deliverable.

[10] https://www.fairdatapoint.org/

[11] https://github.com/FAIRDataTeam/FAIRDataPoint

[12] https://github.com/FAIRDataTeam/FAIRDataPoint-Spec

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

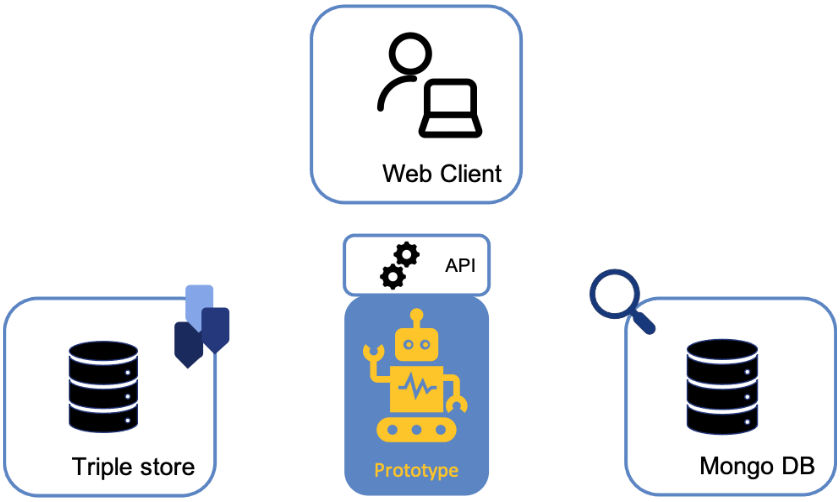

Figure 2. The reference implementation in a nutshell

The web client is the primary interaction point for data stewards and non-technical consumers of the metadata. It offers user management and access control for admin-level users and extensive metadata editing capabilities for data stewards.

The web client of the reference implementation is hosted here: https://fairsfair.fair-dtls.surf-hosted.nl/. A login is required and can be requested from the authors of this document (see Figure 9.1). The document's appendix shows some examples of the web interface (Figure 11 to Figure 15).

The metadata models are defined by customisable definitions written in the SHACL language. The web client allows an admin user to extend existing metadata models and create new metadata types to allow for custom resource descriptions.

The server provides for the web client's functionality through several APIs[13]: metadata interaction for reading and writing metadata, metadata model interaction for custom model definitions, and additional client features. Since the reference implementation is designed as a middleware layer, it offers points of interaction with the metadata of any metadata consuming application. Furthermore, it provides housekeeping features tailored towards the web client, allowing for a better user experience. These features include dynamic metadata schema management, user management, and access control management. API key mechanisms secure the core metadata APIs for writing and updating.

The server stores its data in two separate databases. The primary metadata content is stored in a triplestore. The reference implementation supports several triplestore implementations (*graphdb, blazegraph, allegrograph, in-memory store*)[14], focusing on the most popular choices (*blazegraph, allegrograph*). The maintainer of the deployment is free to choose a specific triplestore

[13] API documentation of the FAIRsFAIR reference implentation is available https://fairsfair.fair-dtls.surf-hosted.nl/swagger-ui.html

[14] https://graphdb.ontotext.com, https://blazegraph.com, https://allegrograph.com, https://rdf4j.org/documentation/reference/configuration/#memory-store

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

implementation. By default, the triplestore content is reachable through the server's APIs, but the maintainer can also expose it via a SPARQL endpoint ("SPARQL 1.1 Overview," n.d.) to the public.

The server's private data is stored in a document store, where the implementation choice for the prototype has been MongoDB[15]. The user credentials, access-control, and custom resource definitions (see Figure 9.2) are only available through the APIs with proper authorisation.

The main components are published in Docker[16] images to allow for convenient deployment. The deployment procedures are described in the documentation pages ("FAIR Data Point 1.6.0 documentation," n.d.) of the reference implementation. The reference implementation is developed as an open-source project, and all its components are available in repositories on GitHub[17]. This work was carried out on the Dutch national e-infrastructure with the support of SURF[18] Cooperative.

3. The DCAT2 Data Model

The metadata exposed in the reference implementation is structured using the DCAT2 model ("Data Catalog Vocabulary (DCAT)," 2020). The Data Catalog vocabulary (DCAT) provides a model to describe datasets (see Figure 3). As DCAT has a native namespace, it makes, like most other semantic web vocabularies, extensive use from different vocabularies, particularly Dublin Core ("DublinCore," 2020), understood by many systems. This will foster integration and reuse in systems that aren't even aware of research data, e.g. search engines, reporting tools etc.

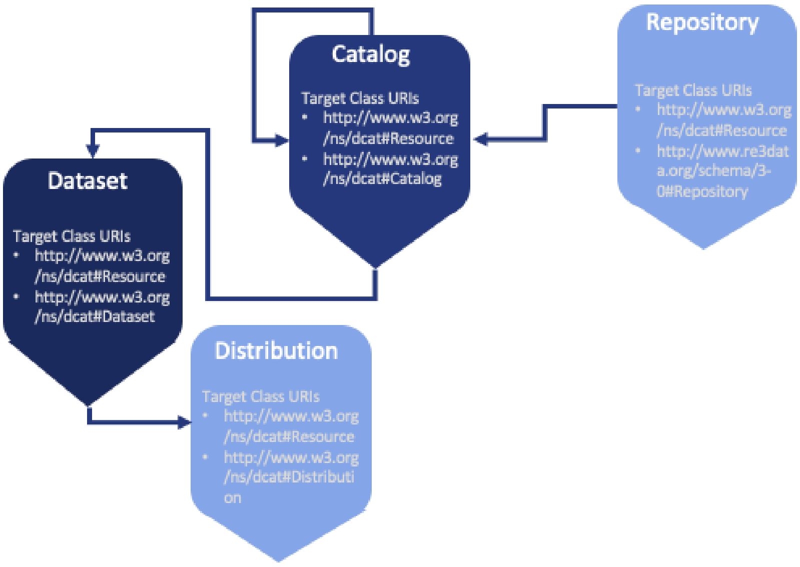

Figure 3. Simplified diagram of the data model for the default types. The arrow points to the child resource, meaning that catalogues are child resources of repositories, catalogues can be child resources of themselves or contain datasets, and a dataset can have multiple distributions.

15 https://www.mongodb.com
16 https://www.docker.com
17 https://github.com/FAIRDataTeam/FAIRDataPoint, https://github.com/fairdatateam/FAIRDataPoint-client
18 https://www.surf.nl

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

In this model, a dataset is defined as “a collection of data, published or curated by a single agent”[19]. At its core, it introduces concepts like the *catalogue*, which allows for an arbitrary grouping of *datasets*, and the *distribution*, which describes a particular distribution format of a *dataset* and allows for different file formats for the same dataset. The *dataset* is an abstract description of a digital object’s metadata, meaning it does not contain the data itself, even though the name might suggest that it does.

DCAT2 provides several improvements to the previous version that was used in the reference implementation. Firstly, it introduces a new type: *resource*. The *resource* acts as a base class for the core DCAT types, which means that most of the existing *definitions* inherit properties from it. Secondly, a *catalogue* can now not only store *datasets* but any type, including *catalogues* themselves, which increases the flexibility and allows for more use cases to be modelled. Besides the DCAT base classes, the reference implementation also adds a *repository* metadata layer on top of the *catalogue* based on the re3data model (Rücknagel et al., 2015).

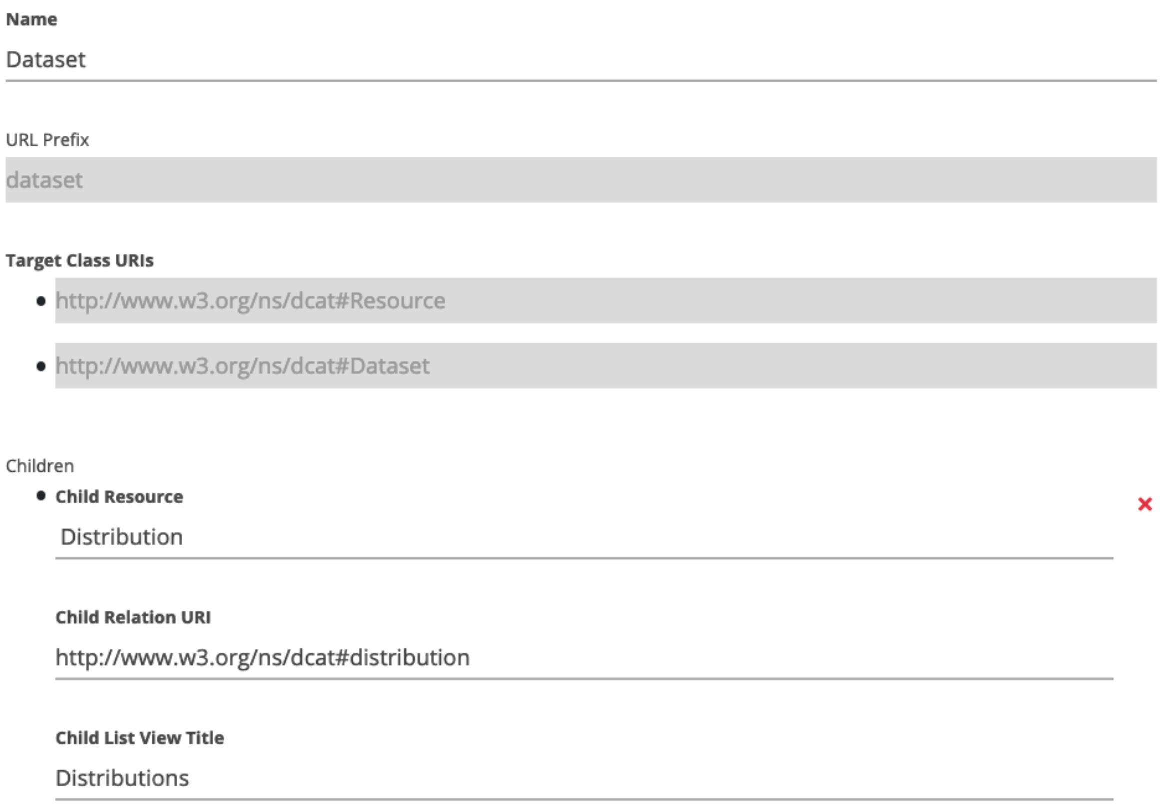

Figure 4. Example of the definition of the dataset type. The URL prefix binds this resource to the /dataset/url prefix. The target class URI binds the DCAT definitions of those resource types to the corresponding SHACL validation shape. Moreover, the child resources are here limited to distribution since a dataset can only contain distributions.

[19] https://www.w3.org/TR/2020/SPSD-vocab-dcat-20200204/#class-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

Furthermore, the *resource* acts as an extension point to introduce custom types into the DCAT model. While the *dataset*'s custom metadata is limited to few properties, an extension allows adding community-defined metadata. Custom extensions can be added using the web application, as seen in Figure 3. The *resource definition* allows the reference implementation to capture some housekeeping data for custom extensions and gives access to metadata's default types since it inherits from the *resource*.

The properties and structure of an extension are captured in a SHACL shape. A shape describes the properties of an extension and constrains them. Those constraints are used to validate metadata against standard rules, such as existing target URIs (see Figure 4).

Furthermore, the shapes determine how properties are exposed in the user interface, where the resources definition links the structure of the application to the expected behaviour. For example, a specific type is linked to a particular URL pattern, and validation constrains it to a specific type. Figure 5 shows such validation in the prototype.

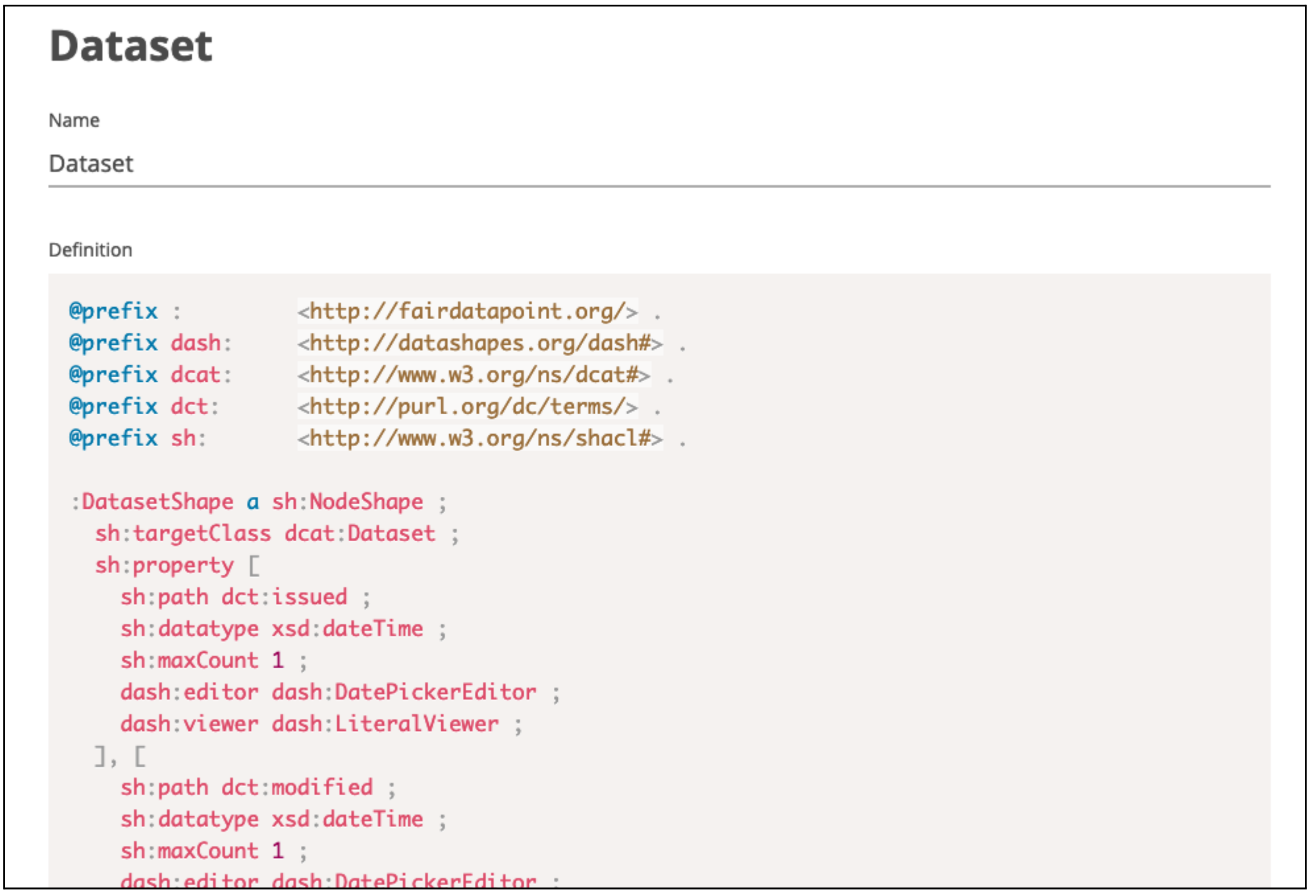

Figure 5. The SHACL shape definition for the "dataset" type, as seen in the user interface. The resource definition of the "dataset" type is matching against the dcat:dataset type and the defined properties. For example, the property "issued" is defined by the "dct:issued" type from Dublin Core. The "dash:"- prefixed definitions instruct the user interface on displaying the properties for editors and viewers.

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 6. Example of two validations in the user interface of an editor. The metadata type of "rights" requires a valid IRI, which is not fulfilled by the word "cat". The metadata of "issued" expects a date as input, which opens a user interface that only allows selecting date-typed metadata.

4. Community Uptake

The reference implementation is up and running since August 2020 and was presented in a web-seminar (Claudia Behnke and Hylke Koers, 2020). It was further explored together with the participating repositories[20] in a digital workshop[21]. In each event, about 60 people registered, and 30 people participated in a questionnaire.

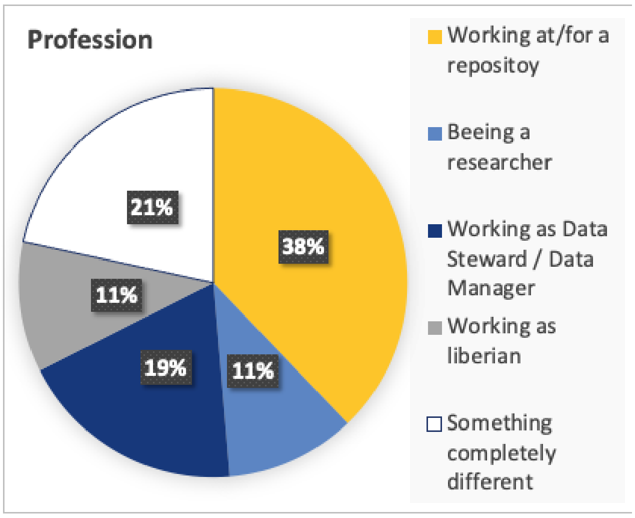

Figure 7.1

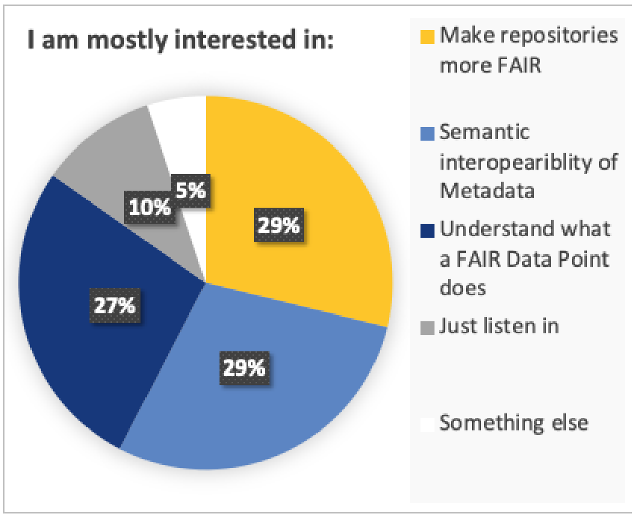

Figure 7.2

Participants professions (Fig. 7.1) and motivation (Fig. 7.2) to participate in the digital workshop[14].

At the publication date of this deliverable, about 30 users have requested and gotten access to the reference implementation. However, the prototype's expected uptake and usage are still not at the level the authors had hoped. It seems that there are skill- and knowledge gaps on the side of the repositories and potentially among data stewards, which are hindering the uptake and use of the FDP. The team has identified, in particular, a knowledge gap related to both DCAT and SHACL as driving factors here. The challenge facing the team to guide both decision-makers and technical staff

[20] https://www.fairsfair.eu/application-results-open-call-data-repositories
[21] https://www.fairsfair.eu/events/workshop-common-metadata-interfaces-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

towards adopting and implementing the FDP is highlighted from the questionnaire results. The majority of those engaging with the work positioned themselves close to 'novice' regarding an understanding of the underlying technologies involved. (See Figure 5.2).

Even though a large proportion of the participants were interested in semantic interoperability and making their repository more FAIR, few were familiar with DCAT or DCAT2. Most of the participants selected "I have no idea"[22] when asked about their knowledge of these topics. The stated familiarity with SHACL was even lower. This knowledge gap hinders the uptake of the prototype and leaves the participants as passive observers.

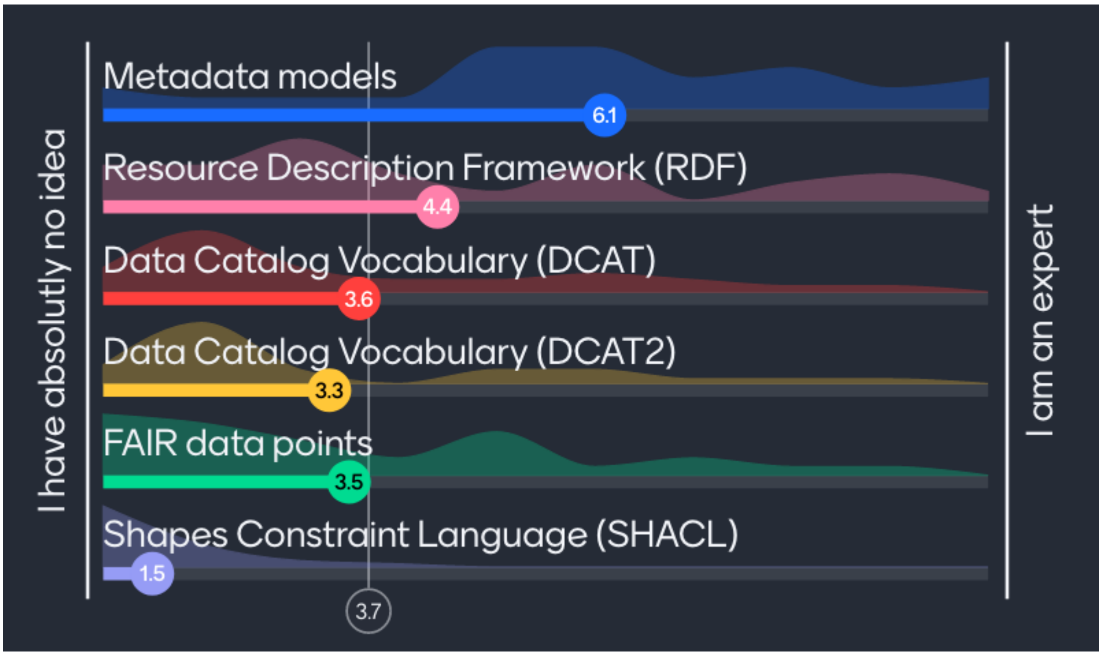

Figure 8. Self-assessment of the existing knowledge from the digital workshop[14]

We will continue to communicate the potential benefits of Linked Data and data catalog vocabularies such as DCAT and DCAT2 in delivering semantically interoperable FAIR enabling digital repositories. FAIRsFAIR held hands-on training on adding metadata shapes to existing metadata models so that they could expose elements X and Y about their repository or objects in a machine-actionable way. The default FDP resources were extended to describe geospatial datasets using a dataset from NBN Atlas[23], one of the developer repositories. That led to the PHAIDRA[24] repository uptake, which then also created a cultural heritage custom extension (See Figure 8). The authors hope that this approach will encourage more repositories to interact with the prototype to test the application better.

[22] on a scale from 1 to 10, where 1 represents "I have no idea" and 10 was "I am an expert".
[23] https://nbnatlas.org/
[24] https://phaidra.cab.unipd.it/

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

Furthermore, there is a close collaboration with the Task 2.2 of the FAIRsFAIR project, which plans to define a minimum metadata schema for semantic artefacts (Hugo et al., 2020) and may be able to test this in the prototype.

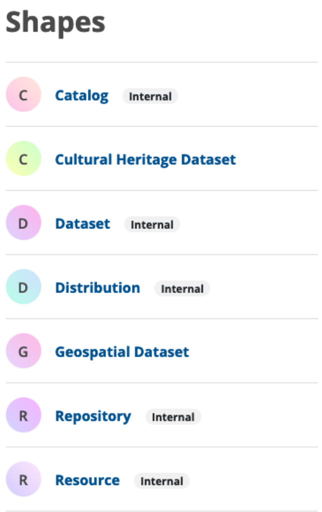

Figure 9.1

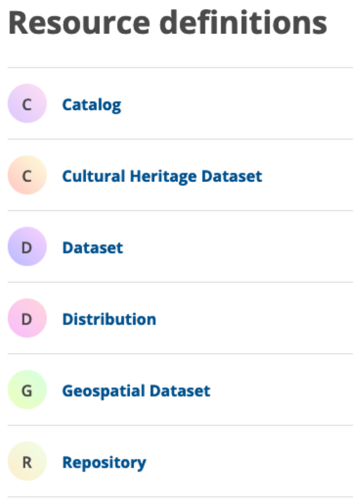

Figure 9.2

Figure 7. Overview of all shapes (9.1) and resource definitions (9.2) currently available in the reference implementation. Shapes labelled "internal" are form base types, and shapes without a label are custom extensions. The internal types are not editable since the behaviour of the application depends on them.

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

5. Outlook and Future Plans

In the last year of the FAIRsFAIR project, the authors aim to increase the community uptake, and collaborate with other projects also using FDPs and focusing on exposing metadata, for example the INFRAEOSC-05B project EOSC Pillar[25]. Furthermore, the semantic interoperability of the metadata shall be further addressed together with Task 2.2 of the FAIRsFAIR project.

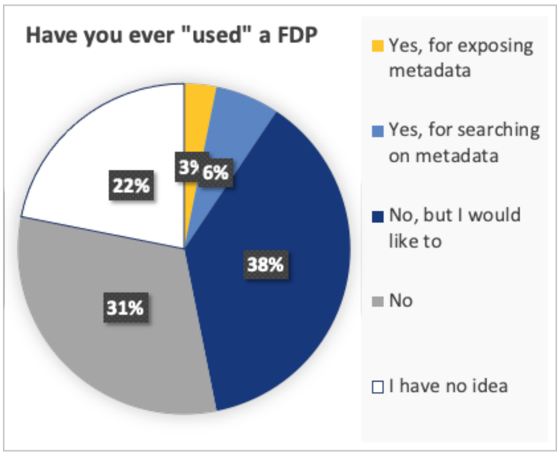

Figure 10. Existing experience with FDPs

More hands-on training, combined with targeting known repository owners and clarifying their needs, will be undertaken in order to stimulate uptake. Furthermore, the team has plans to improve the user interface, which will make the reference implementation even more intuitive to use.

In the long term, together with other work packages of FAIRsFAIR, this shall reduce the knowledge gap hindering the technical implementation of tools like the FDP, which will support the adoption of solutions for enabling FAIR in repositories. Skills development in this area should also aid the uptake of the reference implementation, which will drive the virtuous circle of testing and usability improvements that is a key aspect of the development of the FDP. The team is confident that a need and appetite for implementing the FDP exists based on the survey results[14], where about 40% of respondents stated that they are interested in using a FAIR data point (see Figure 10).

The work will progress according to the following milestones:

- 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) (Q4 2021/Q1 2022).

[25] https://www.eosc-pillar.eu/

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

6. Bibliography

Behnke, C., Bonino, L., Coen, G., Le Franc, Y., Parland-von Essen, J., Riungu-Kalliosaari, L., Staiger, C., 2020. D2.3 Set of FAIR data repositories features. https://doi.org/10.5281/zenodo.3631528

Claudia Behnke, Hylke Koers, 2020. FAIRification of Services – Two Examples | 2nd FAIRsFAIR webinar. https://doi.org/10.5281/zenodo.3747659

Data Catalog Vocabulary (DCAT) [WWW Document], 2020. URL https://www.w3.org/ns/dcat (accessed 7.30.20).

DublinCore [WWW Document], 2020. URL https://www.dublincore.org/ (accessed 7.30.20).

FAIR Data Point 1.6.0 documentation [WWW Document], n.d. URL https://fairdatapoint.readthedocs.io/ (accessed 1.26.21).

Hugo, W., Le Franc, Y., Coen, G., Parland-von Essen, J., Bonino, L., 2020. D2.5 FAIR Semantics Recommendations Second Iteration. https://doi.org/10.5281/zenodo.4314321

L'Hours, H., Stein, I.V., Huigen, F., Devaraju, A., Mokrane, M., Davidson, J., Vries, J.D., Herterich, P., Cepinskas, L., Huber, R., 2020. M4.2 Draft Maturity Model Based on Extensions and-or Additions to CoreTrustSeal Requirements. https://doi.org/10.5281/ZENODO.4003598

Rücknagel, J., Vierkant, P., Ulrich, R., Kloska, G., Schnepf, E., Fichtmüller, D., Reuter, E., Semrau, A., Kindling, M., Pampel, H., Witt, M., Fritze, F., Van De Sandt, S., Klump, J., Goebelbecker, H.-J., Skarupianski, M., Bertelmann, R., Schirmbacher, P., Scholze, F., Kramer, C., Fuchs, C., Spier, S., Kirchhoff, A., 2015. Metadata Schema for the Description of Research Data Repositories. Deutsches GeoForschungsZentrum GFZ. https://doi.org/10.2312/RE3.008

Shapes Constraint Language (SHACL) [WWW Document], n.d. URL https://www.w3.org/TR/shacl/ (accessed 1.26.21).

SPARQL 1.1 Overview [WWW Document], n.d. URL https://www.w3.org/TR/sparql11-overview/ (accessed 1.26.21).

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

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

7. Appendix:

7.1. Requirements for FAIR repositories coming from FAIRsFAIR D2.3

FAIRsFAIR requirements for FAIR repositories

Organisational requirements

List of the 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 a human and machine interoperable way (FAIR)
- Documentation of interfaces and APIs (FAIR)

Technical requirements

Below is a list of technical features which are proposed to improve the FAIRness of data repositories. Currently, those features do not suggest any specific implementation or technology. However, if implemented, they are expected to 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)

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

- 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)
- Demand fine-grained metadata from data providers (FI)
- Implement community standards (FI)
- Automatic ontology suggestions and lookup (FI)
- 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 development of these requirements by the FAIRsFAIR team, requirements were identified, which indirectly affect FAIR. Although those features do not directly contribute to enabling FAIR in repositories, they are mentioned below for completeness.

- Depending on the nature of the repository and the types of data that it houses, the repository should:
  - Support dynamic data sets (e.g. time-series data)
  - Sent notifications to the creator if similar data appears elsewhere

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

- ○ 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 (e.g. RIS, BibTeX) that point to the data
  - ○ Variety of access restrictions
  - ○ Tombstone[26] 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."

7.2. Some examples from the reference implementation

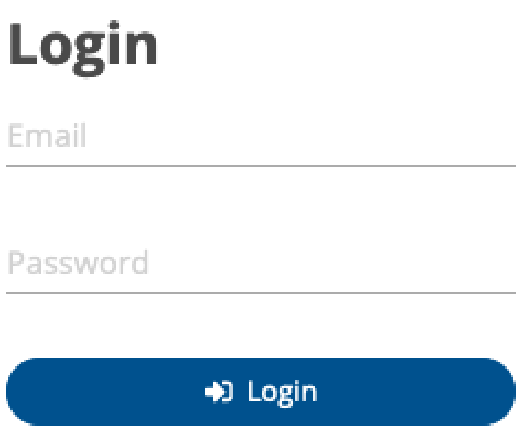

Figure 11.1

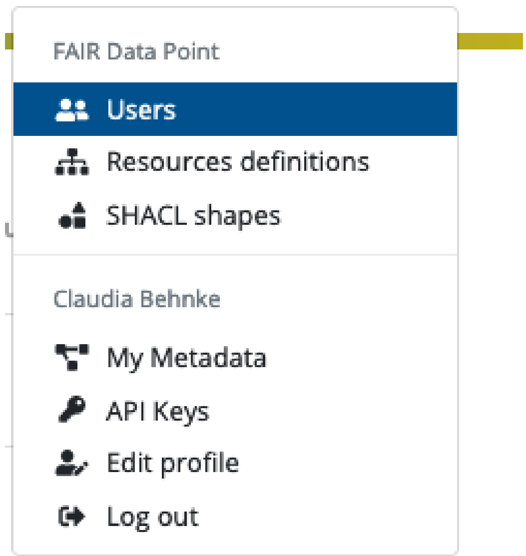

Figure 11.2

Figure 11: Login interface (11.1) and user options for admin-level users (11.2)

[26] A "tombstone page" is a special type of landing page describing an item which has been removed or made inaccessible from a collection [R33].

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

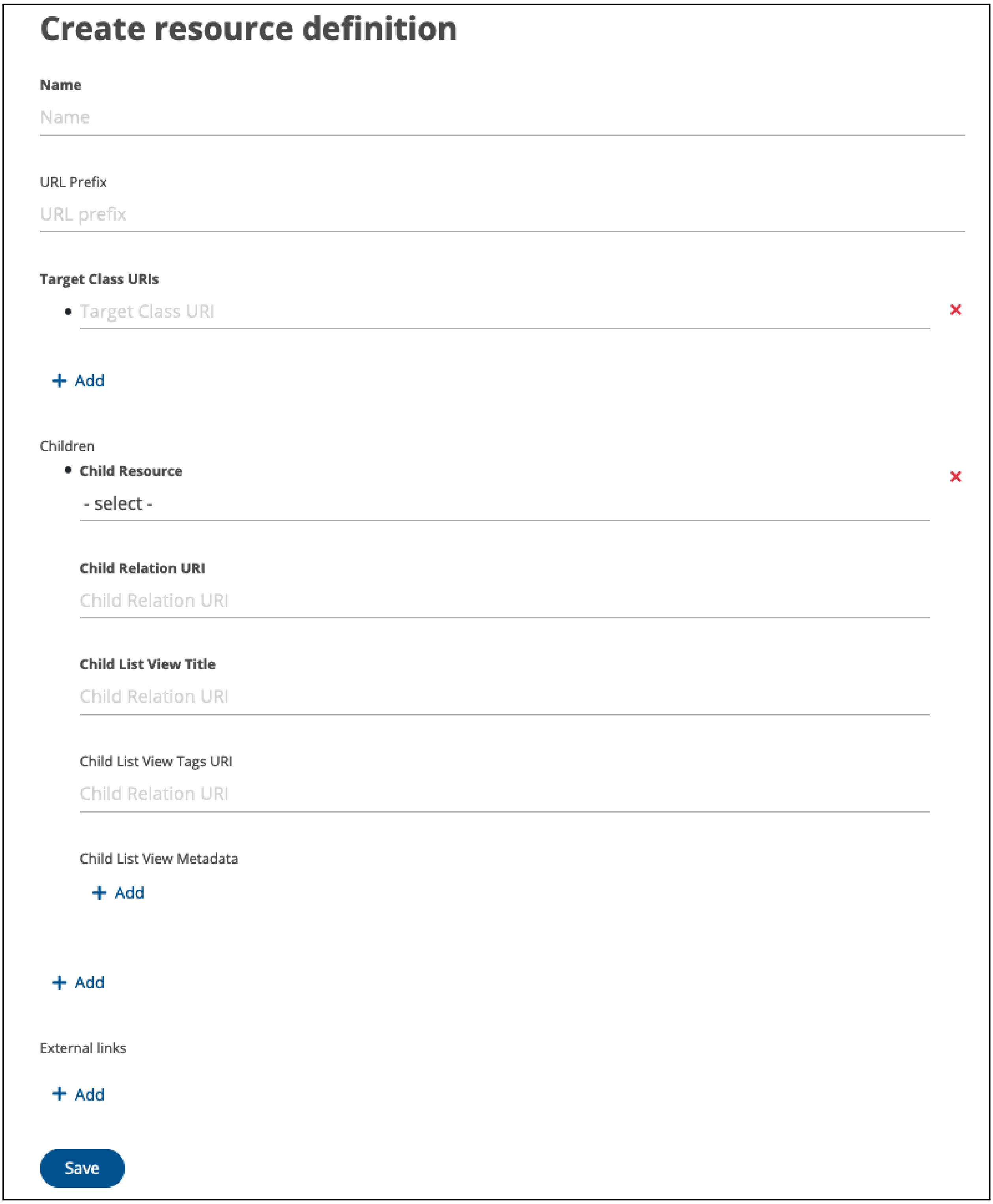

Figure 12. The user interface to add a custom resource definition in the web client

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

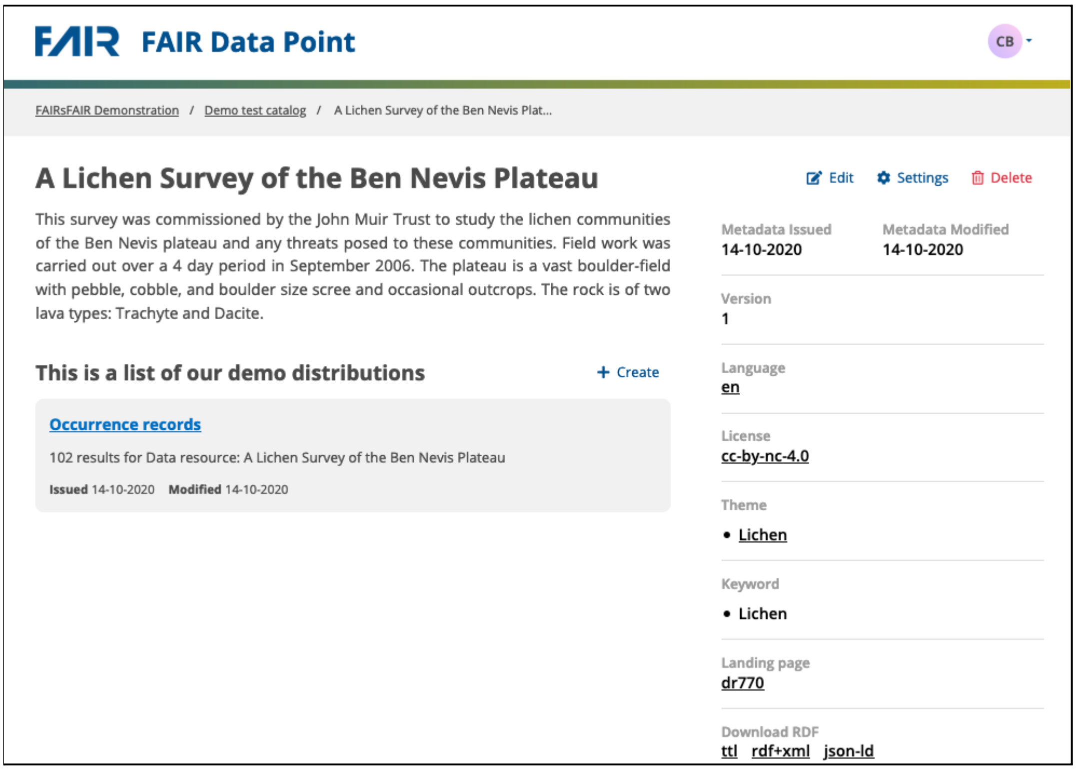

Figure 13. The user interface for a dataset

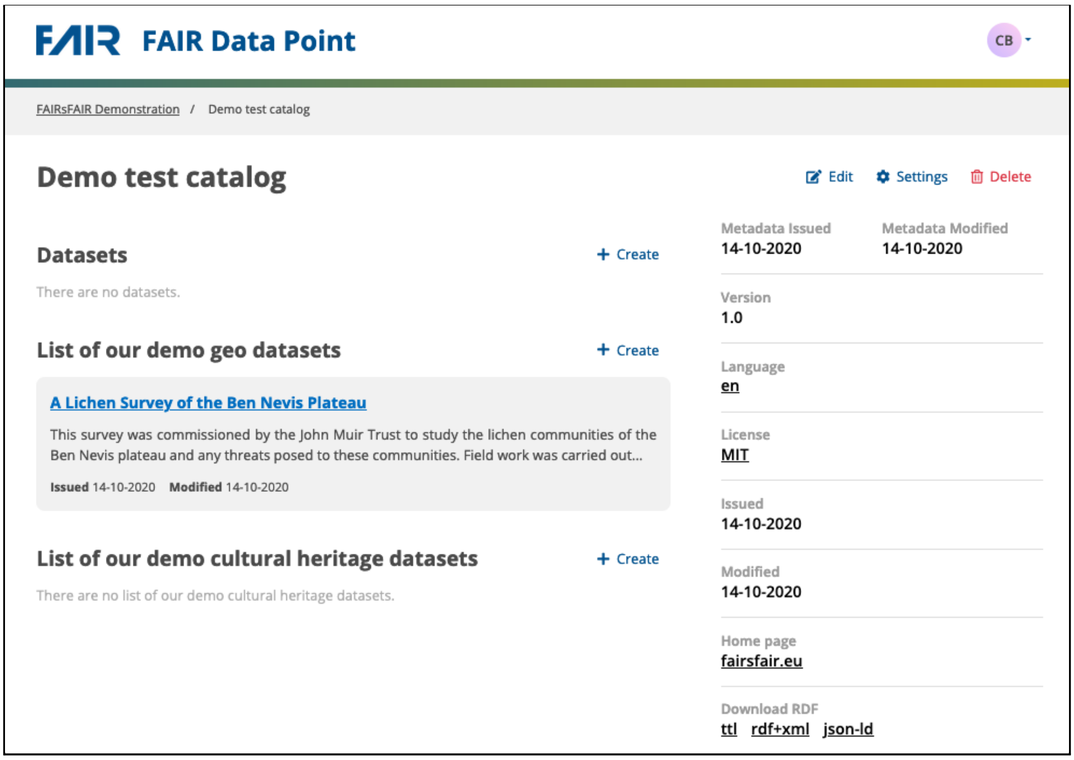

Figure 14. The user interface for a catalogue

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

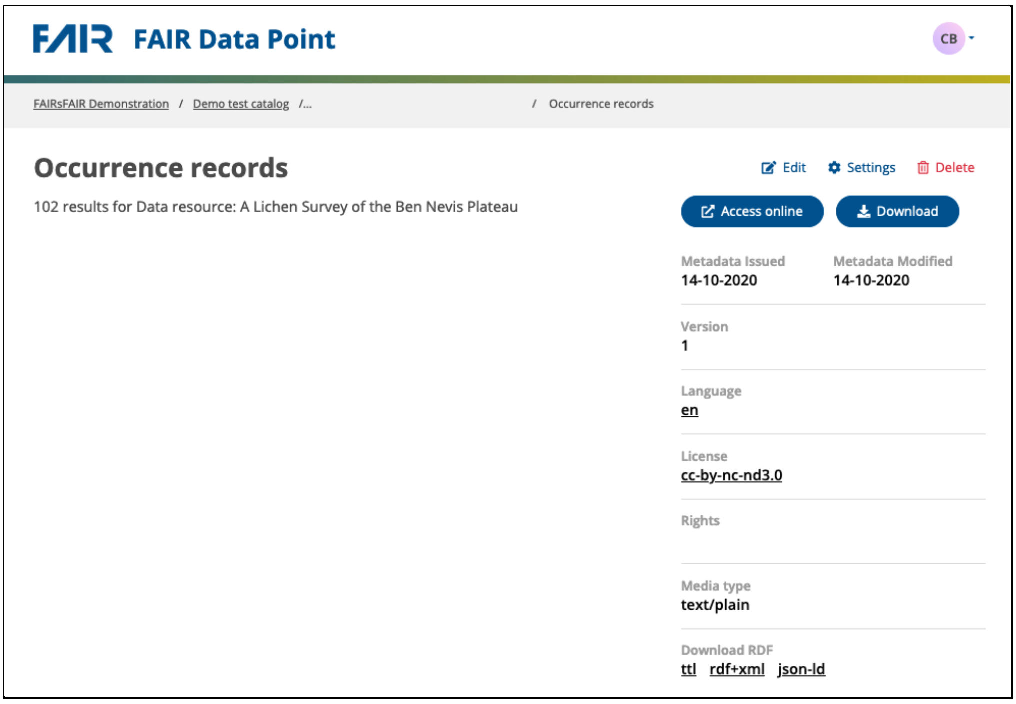

Figure 15. The user interface for a distribution

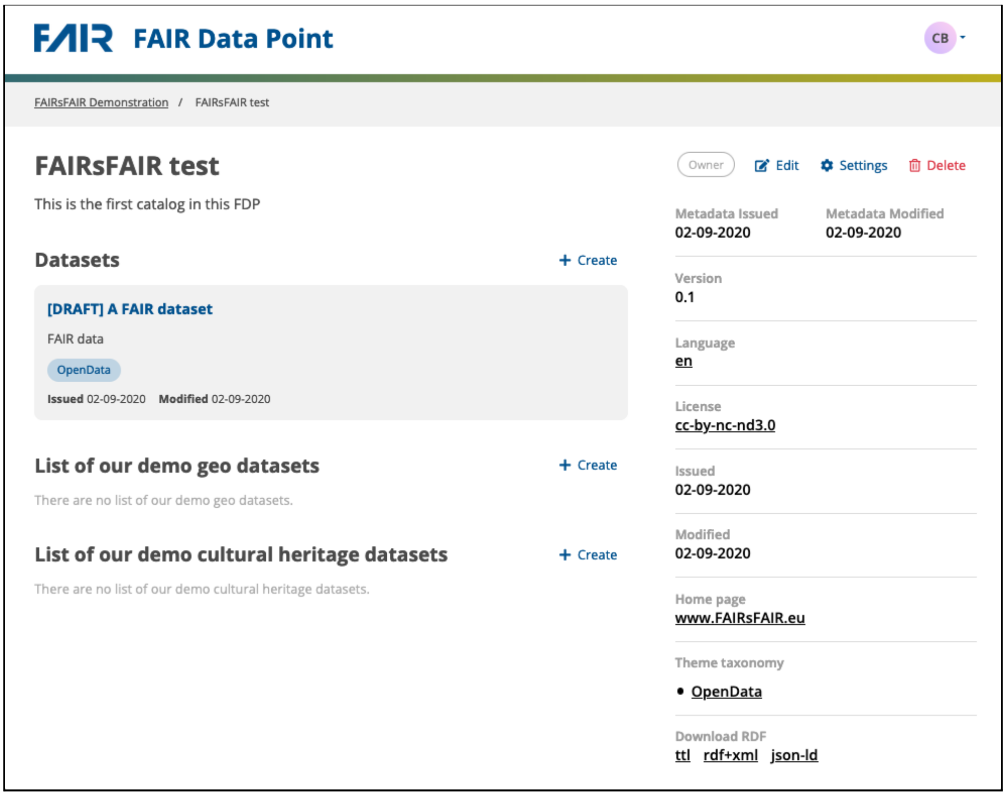

Figure 16. The user interface for a catalogue

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

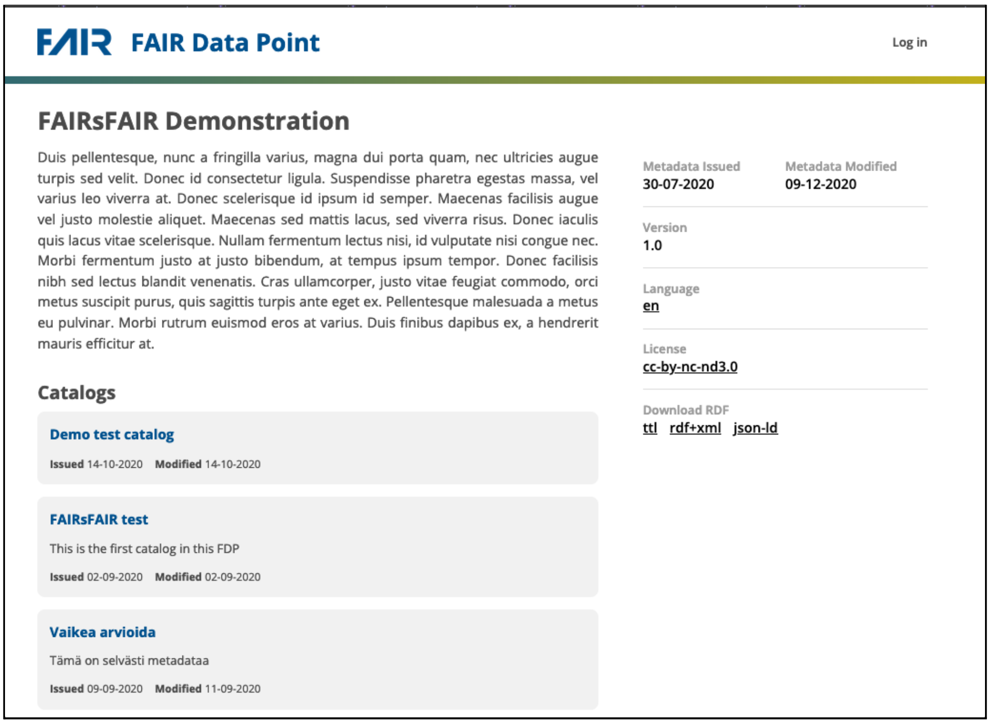

Figure 17. The user interface on the highest level of the application

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