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

D2.9 Second reference implementation of the data repositories features and client application

Work Package	WP2
Lead Author (Org)	Claudia Behnke (SURF)
Contributing Author(s) (Org)	Kees Burger (LUMC), Yann Le Franc, (e-SDF), Pekka Järveläinen, Jessica Parland-von Essen (CSC), Vyacheslav Tykhonov (DANS/EOSC Synergy)
Due Date	28.02.2022
Date	21.02.2022
Version	1.0
DOI	10.5281/zenodo.6204055

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 document is the third 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
0.1	07.12.2021	Claudia Behnke	First draft created
0.7	02.02.2022	Kees Burger, Yann Le Franc, Pekka Järveläinen, Jessica Parland-von Essen, Vyacheslav Tykhonov, Claudia Behnke	Draft for internal review
0.8	02.02.22	Robert Huber, Hervé L'Hours	Internal Review
0.9	21.02.22	Yann Le Franc, Pekka Järveläinen, Jessica Parland-von Essen, Vyacheslav Tykhonov, Claudia Behnke, Rob Hooft, Kees Burger	Incorporating remarks of the reviewers
1.0	21.02.22	FAIRsFAIR Project Coordinaiton	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
DDI-CDI	Data Documentation Initiative - Cross-Domain Integration
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 third report from Task 2.3 of the FAIRsFAIR project. It describes developments of the prototype based on the DCAT2 data model, that were implemented in the last year of the project, addressing issues that have been identified earlier in the project. Also discussed are the results of a workshop between the groups involved in this task and three organisations, showing progress towards their implementation of the FAIR data point architecture.

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 describes the FAIR Data point and its FAIRsFAIR T2.3 'reference implementation' covering distinct features, like interoperability and machine actionability of (meta)data, to enhance the FAIRness of repositories, first specified during the workshop "Building the data landscape of the future: FAIR Semantics and FAIR Repositories" (22nd October 2019, Espoo, Finland)[1] and summarised in the previous deliverable of the task (Behnke et al., 2020).

The basic architecture of the FAIR data point and the FAIRsFAIR reference implementation is described in D2.6[2] (chapter 2) and will not be repeated in this deliverable. If you are not familiar with the FAIR Data Point, that chapter is essential background for the information in this deliverable.

The deliverable will include updates and enhancements compared to the last deliverable[2] (see Chapter 3) of the task and results of a workshop with participating repositories (see chapter 4)[3].

Regarding the approach:

> The FAIR data point is built upon semantic web technologies and follows a Linked Data approach. Some aspects of this document may be more challenging to readers who are not familiar with Linked Data[4].

1.1. What is FAIR about this approach?

One key aspect of making repositories FAIR enabling is exposing metadata in a more FAIR manner. Metadata are mentioned in most of the FAIR principles[5]. During the work of T2.3, the task assumed that the digital object stored in repositories is already as FAIR as possible [6]. 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. It is this machine actionability of metadata that has been the focus of this task.

[1] https://www.fairsfair.eu/events/building-data-landscape-future-fair-semantics-and-fair-repositories
[2] https://zenodo.org/record/5362027#.Ygoc9O7ML0o First reference implementation of the data repositories features
[3] https://zenodo.org/record/5795809#.Yek3rljML0o
[4] https://www.w3.org/DesignIssues/LinkedData.html
[5] https://doi.org/10.1038/sdata.2016.18
[6] This excludes the extent to which the digital object itself is Interoperable: Findability, Accessibility, and Reusability are elements that can primarily be addressed by rich metadata, but interoperability characteristics of the digital object itself requires compliance with community standards for schema and semantics.

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

2. The FAIR Data Point and its FAIRsFAIR Reference Implementation

To enable the FAIR exchange of metadata between repositories and the exchange of metadata between repositories and registries or catalogues, it is really helpful that the metadata satisfies at least these criteria: (a) metadata fields are using common definitions; (b) metadata fields are put together in common schemas and (c) metadata schemas are queryable through common APIs and available in common formats. The definition of the FAIR Data Point[7], based on general standards like DCAT, addresses these three points. As described before, there are two approaches in which existing services can make use of the FAIR Data point: The first one is to implement a FAIR Data Point in their own software, simply exposing existing metadata in an additional way. The second one is to deploy the FAIRsFAIR reference implementation.

In the previous deliverable (see box on the next pages), we described the user interface of the FAIRsFAIR reference implementation. However, a human would not add all the metadata records by hand in a production environment.

The ingestion of the existing metadata into an instance of the reference implementation requires three steps, in the database world known as “Extract, Transform, Load” or ETL:

1. The metadata schema of the dataset, but also of the repository, needs to be created in a FAIR data point compliant DCAT format. If the metadata format is identical to the DCAT format (See Figure 2, taken from D2.6) this step can be skipped. In D2.6, the task explained how users could extend the SHACL (Shapes Constraint Language (SHACL)) shapes, which are used to validate the content against the constraints implied by the RDF schema.
2. The existing metadata needs to be mapped to the (newly created) DCAT format, meaning that for all existing fields in the repository a respective field in DCAT needs to be found. This is a step that needs to be done by the repository. Chapter 3.2 and 4.2 show two alternative approaches where fields can be semi-automatically mapped.
3. Once this is done a repository data steward can load the metadata in a FAIR Data point configured with the new metadata schemas. The FAIR Data Point will validate the new metadata based on the metadata schemas, and when all validation steps are passed it will expose the enriched metadata with increased FAIRness. The metadata resources can be configured to require authentication and authorization. Some endpoints are available for all users, while others require an API token for authorisation.

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

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 documentation of FAIRdata[8] point also provides an extensive description of how to create, delete and modify those resource definitions. https://fairdatapoint.readthedocs.io/en/latest/usage/api-usage.html.

1. The FAIRsFAIR Reference Implementation (from D2.6)

The root of the FAIRsFAIR reference implementation is a FAIR Data Point[9] (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 openly available and well documented[10].

There are two different ways of reaching these goals. The first one is the deployment of a stand-alone application[5]. It is also possible to extend existing appliances using only the specifications[11]. 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.

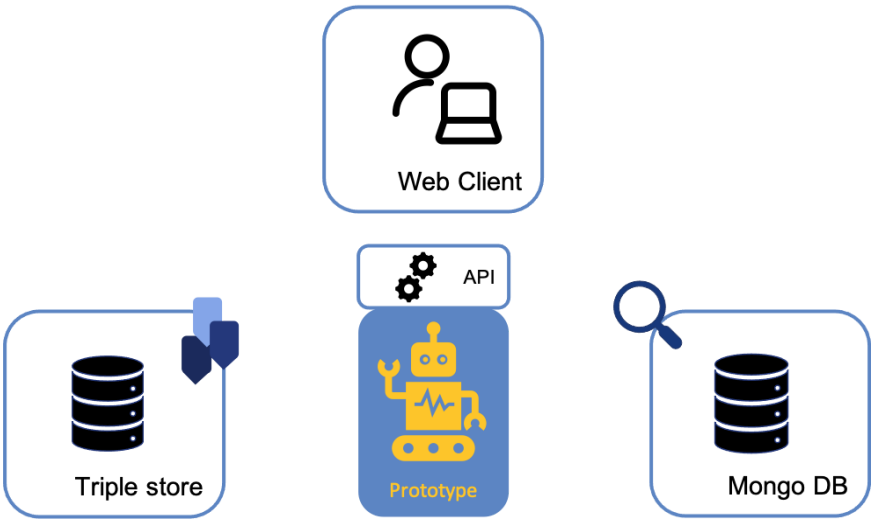


Figure 1. The reference implementation in a nutshell

[8] https://fairdatapoint.readthedocs.io/en/latest/usage/usage.html#resource-definitions
[9] https://www.fairdatapoint.org/
[10] https://github.com/FAIRDataTeam/FAIRDataPoint
[11] 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

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). In the document's appendix, we show 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: metadata interaction for reading and writing metadata, metadata model interaction for custom model definitions, and additional client features. It offers points of interaction with the metadata for developers of third party systems. 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*)[12], focusing on the most popular choices (*blazegraph, allegrograph*). The maintainer of the deployment is free to choose a specific triplestore 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 internal data is stored in a document store, where implementation choice for the prototype has been MongoDB[13]. The user credentials, access-control, and custom resource definitions are only available through the APIs with proper authorisation (see Figure 9.2).

The main components are published in Docker[14] 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[15]. This work was carried out on the Dutch national e-infrastructure with the support of SURF[16] Cooperative.

[12] https://graphdb.ontotext.com, https://blazegraph.com, https://allegrograph.com, https://rdf4j.org/documentation/reference/configuration/#memory-store
[13] https://www.mongodb.com
[14] https://www.docker.com
[15] https://github.com/FAIRDataTeam/FAIRDataPoint, https://github.com/fairdatateam/FAIRDataPoint-client
[16] 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

The DCAT2 Data Model (from D2.6)

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). Even though DCAT has a native namespace, it makes extensive use of other vocabularies, particularly Dublin Core ("DublinCore," 2020).

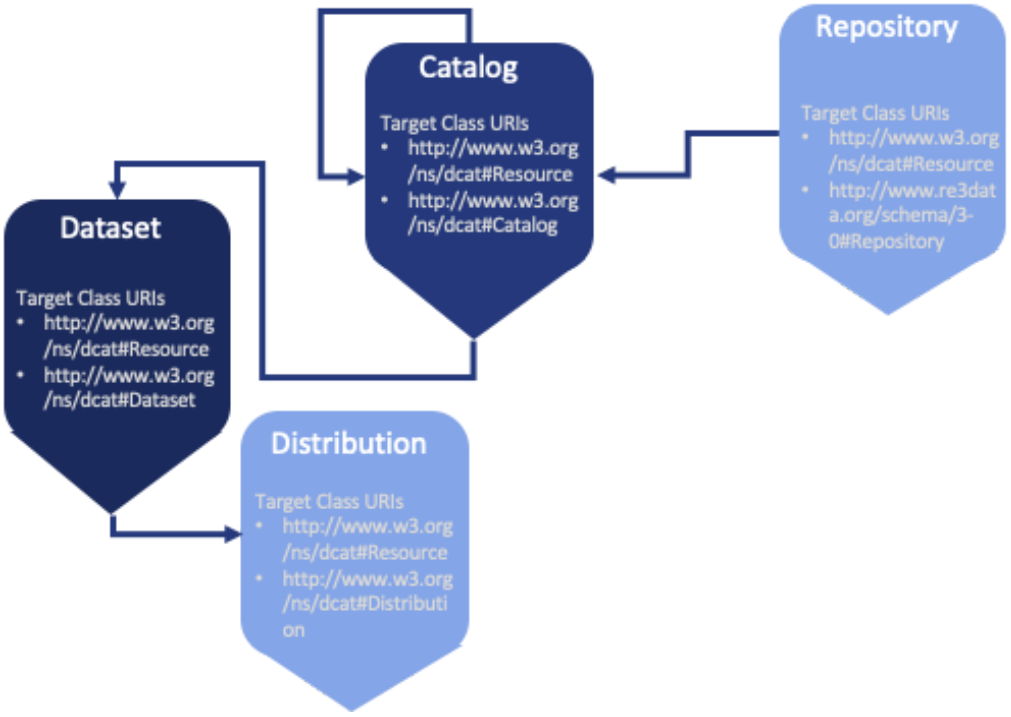


Figure 2. Simplified diagram of the data model for the default types. The error points to the child ressource, meaning that catalogues are child ressources of repositories, catalogues can be child resources of themselves or contain datasets and a dataset can have multiple distributions.

In this model, a dataset is defined as "a collection of data, published or curated by a single agent" [17]. 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.

[17] 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

3. Using the FAIR Data Point reference implementation

In the course of the project the setup reference implementation was not only tested by the participating repositories but also used within the project. The first one will be described in the community update, and the second in the following part.

3.1. Updates since last deliverable

Since the last deliverable, the FAIR Data Point underwent the following improvements:

- FDP Index communication made more configurable
- Metadata schemas (SHACL shapes) are publishable and shareable for reuse in other FDPs
- Improved features for FDP admin users
- Dublin Core's "conforms to" properties are made more explicit and resolvable

3.2. The FAIRsFAIR Reference Implementation and F-UJI

The FAIRsFAIR project developed a FAIR assessment service called F-UJI[18] based on REST. F-UJI reads in a given URL or a DOI and evaluates 17 core metrics on the resource that it identifies. These metrics were defined by FAIRsFAIR to evaluate the FAIRness of research data objects in Trustworthy Digital Repositories (TDRs)[19]. The results of a F-UJI evaluation are given as a total percentage, and, more importantly, the report also contains detailed information on the individual criteria, which can assist users to make resources more FAIR.

For a simple initial assessment of how a FAIR Data point an contribute to a FAIR ecosystem, the team evaluated an example dataset using F-UJI by pointing it directly to the DOI, and by pointing it to the URL for the same dataset described in our own deployment of a FAIR data point.

The first results were disappointing: the F-UJI score of the dataset exposed via the FAIR data point was much lower than those from the original dataset. A careful analysis of the report quickly identified a few points where the FAIR Data point could implement more standard ways of exposing information, and a number of other issues that were due to F-UJI not understanding/implementing all the information given in the DCAT metadata. The following observations were made:

[18] https://www.fairsfair.eu/f-uji-automated-fair-data-assessment-tool
[19] https://zenodo.org/record/4081213#.Ye595_XML0o

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

- Purl URL was not handled correctly: purl.org/fairdatapoint/… is passed but is resolved to app.fairdatapoint.org/.
- Content negotiation on the FAIR data point provided formats (TTL, RDF XML, JSON-LD) do not work properly.
- Access rights seemed not to be adequately recognised. F-UJI recognized a limited number of vocabularies for access rights descriptions.
- The license was not resolved. At least the purl.org/NET/rdflicense version was not recognised.

After a single meeting with the F-UJI team, Issues were created on their git repository and quickly resolved. In parallel, fixes were made to the FAIR Data point reference implementation. These changes together resulted in a much better score for the data set in FAIR data point.
The technical details can be found here https://github.com/pangaea-data-publisher/fuji/issues.

Our experience with this process shows that metadata interoperability is not a simple matter, even when standards are strictly adhered to. Analysis of the suggestions made by a FAIR evaluator, in this case F-UJI, and implementing improvements to the implementation of the standards (here performed in the implementation of the FAIR Data Point, but also through improvements of the DCAT interpretation in F-UJI), can really make a difference in the FAIR availability of 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

Status in June 2021:

NARCIS metadata in Dublin Core (dc) format

Resource PID/URL:	doi:10.34894/QOH33L
Metric Version:	metrics_v0.4
Metric Specification:	https://doi.org/10.5281/zenodo.4081213
Software version:	v1.1.2

Summary:

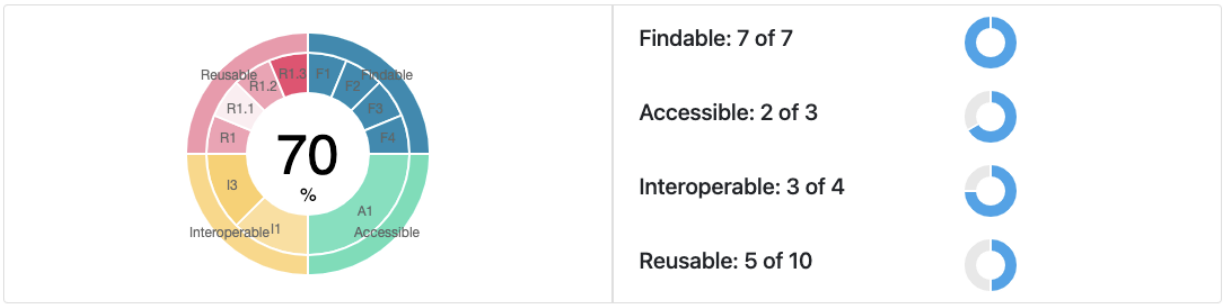


Fig 1: F-UJI results when evaluating the original dataset (doi:10.34894/QOH33L) in June2021

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

NARCIS PORTAL metadata in Dublin Core (dc) format

Resource PID/URL:	https://fairsfair.fair-dtls.surf-hosted.nl/dataset/fb33948b-8656-44aa-9bfc-4acd39d0784c
Metric Version:	metrics_v0.4
Metric Specification:	https://doi.org/10.5281/zenodo.4081213
Software version:	v1.1.2

Summary:

Fig 2: F-UJI results when evaluating the metadata exposed from the FAIR data point, before implementing improvements on both tools..

Status today

FAIR level:	advanced
Resource PID/URL:	http://purl.org/fairsfair/2.3/dataset/9cb3f3e6-f17c-4109-a82e-d3b39cab603c
DataCite support:	enabled
Metric Version:	metrics_v0.4
Metric Specification:	https://doi.org/10.5281/zenodo.4081213
Software version:	v1.3.8
Download saved assessment results:	{JSON}
Save and share assessment results:	Saved assessments: • 2021-10-13 advanced

Showing cached or saved response. Click here to rerun the assessment

Summary:

	Score earned:	Fair level:
Findable:	3.5 of 7	moderate
Accessible:	3 of 3	advanced
Interoperable:	3 of 4	advanced
Reusable:	6 of 10	moderate

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

https://www.f-uji.net/view/36

3.3. FAIR Data Point for Semantic Artefacts

In the context of the FAIRsFAIR Task T2.2., community-driven recommendations to make semantic artefacts (i.e. controlled vocabulary, code list, thesauri, ontologies,...) FAIR[20] have been proposed. One of the mandatory recommendations (P-Rec3) states that semantic artefacts and their content should be documented according to a common set of minimum metadata. The team of Task 2.2 worked together with experts from various communities to define an initial version of the minimum metadata based on the DCAT model. To collect a broader set of feedback from the communities, a workshop was held last year. During this workshop, the alignment with the DCAT model has been presented together with the integration with MOD, the Metadata for Ontology Description and Publication Ontology[21,22]. In the second part of the workshop, participants were asked to vote to define the mandatory, recommended and optional metadata fields that should compose a DCAT Application Profile for semantic artefacts. Our process, the voting results and the minimum metadata schema are described in the deliverable D2.8.

To support the definition of this metadata schema, we created a simple use-case in which an ontologist, a knowledge engineer, is building a domain ontology for a specific application and needs to find existing relevant ontologies for his work. At the moment, this can be done, but it requires searching in various places (GitHub repositories, webpages, semantic repositories, registries, …), which is cumbersome and time consuming. Therefore, there is a clear need for a unique central interface to search for ontologies across multiple sources similar to the discontinued Swoogle[23].

To address this need, we defined a Proof of Concept (PoC) implementation of a FAIR Semantic Space, which would leverage the F2DS technology developed in the EOSC Pillar project[24] to publish the metadata content of semantic repositories in an instance of the FAIR Data Point reference implementation. This content would be aligned with the minimum metadata schema developed in T2.2. This PoC offers a unique opportunity to test and improve the minimum metadata schema and to provide a PoC implementation of a search engine for semantic artefacts.

The FAIR Data Point reference implementation offers a unique platform to support the FAIRification of existing repository content. This added value has been identified within EOSC-Pillar, one of our collaborating projects. The FAIR Data Point is at the centre of a

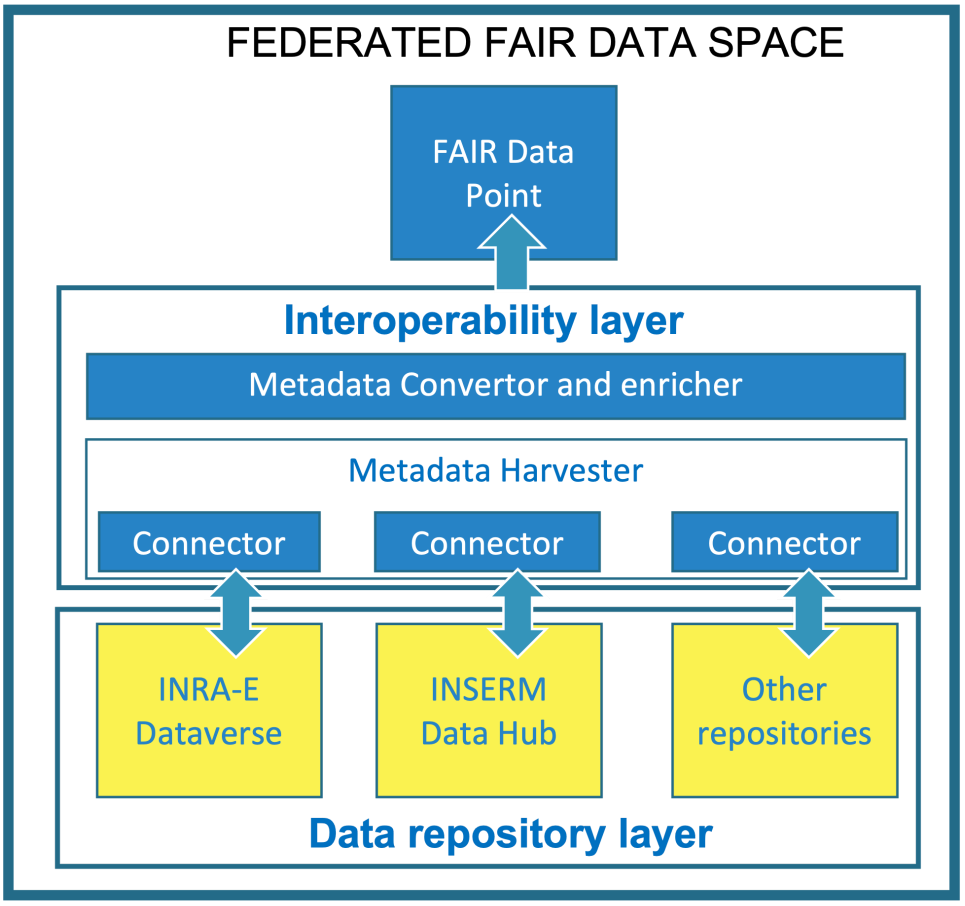
[20] https://zenodo.org/record/4314321
[21] Biswanath Dutta, Anne Toulet, Vincent Emonet, Clement Jonquet. New Generation Metadata vocabulary for Ontology Description and Publication. *MTSR: Metadata and Semantics Research Conference*, Nov 2017, Tallinn, Estonia. pp.173-185, ⟨10.1007/978-3-319-70863-8_17⟩. ⟨lirmm-01605783⟩
[22] https://www.isibang.ac.in/ns/mod/
[23] https://ebiquity.umbc.edu/project/html/id/53/Swoogle
[24] 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

service architecture called the Federated FAIR Data Space (F2DS)(Candela et al., 2021[25]). This service architecture allows repository owners to register their resource in the data space, document and annotate their APIs using the SmartAPI[26] approach. This step enables the creation of a harmonised set of machine-readable API descriptions, which is then used by the metadata harvester service (see figure 11 below) that will fetch the repository metadata from the APIs without any consideration for the data model involved. The collected metadata must then be mapped by the repository owner to the DCAT model or any DCAT Application Profile such as the GeoDCAT-AP model. To build such mappings, users are provided with a user interface showing on one side the DCAT model and on the other side the metadata elements from the repository to be mapped. Once the mappings are defined, a conversion service transforms (in the back end) the metadata into the DCAT model and publishes it in a central FAIR Data Point. The FAIR Data Point is then connected to the EOSC Pillar D4Science[27] catalogue which provides an extended user interface to search and retrieve datasets from the FAIR Data Point. This service offers a unique way to support the FAIRification of data repositories without requiring them to make major changes. The different services composing the F2DS are available online[28] and can be deployed in a kubernetes infrastructure. The general architecture diagram of the F2DS is shown below (courtesy of EOSC Pillar).

[25] Candela, Leonardo, Frosini, Luca, Mangiacrapa, Francesco, Rouchon, Olivier, Le Franc, Yann, & Toulemonde, Baptiste. (2021). EOSC-Pillar D5.6 FAIR Research Data Management Tool Set Update. Zenodo. https://doi.org/10.5281/zenodo.5720411
[26] https://smart-api.info/
[27] https://eosc-pillar.d4science.org/catalogue-eoscpillar
[28] https://www.eosc-pillar.eu/establishing-fair-data-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



Concept, Design: Y. Le Franc, N. Cazenave

Figure 11: Simplified architecture diagram of the F2DS. At the bottom layer, data repositories are connected to an interoperability layer which leverages a common machine-readable API description (Connector). These descriptions are used by the Metadata Harvester to harvest the repository's metadata. Provided that a mapping template is provided, the metadata can then be converted into a common metadata model providing more interoperability between the content of heterogeneous and distributed resources. Once the conversion is done, the resulting metadata is then published in a FAIR Data Point reference implementation instance. Semantic enrichment, i.e. the use of semantic artefacts (ontologies, controlled vocabularies, thesaurus,..) to standardise the metadata descriptions and making interoperable (FAIR principle I2), should be provided by the semantic enricher service. This additional service should provide 1) access to existing domain ontologies and 2) means to replace free text keywords by the corresponding concept URI from these domain ontologies. It will be implemented in later versions of the F2DS.

We, therefore, worked together with the EOSC Pillar team to deploy an instance of the F2DS that will be used to populate the FAIR Data Point with the content of several identified

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

semantic repositories, e.g. Bioportal[29], AgroPortal[30], the NERC Vocabulary server[31], MatPortal[32], IndustryPortal[33], ….

For this PoC, we decided to start with three ontology repositories: Bioportal, Agroportal and the NERC Vocabulary service.

The F2DS offers a coherent workflow for registering repositories into the data space. The registration workflow uses dedicated and connected User Interfaces supporting:

1. the capture of the metadata describing the repository they want to register,
2. the capture of the information regarding the API and the creation of a SmartAPI description for the repository API,
3. the triggering of the smartHarvester service which will collect the repository metadata using the SmartAPI descriptio,
4. the creation, the testing and saving of the mapping scheme between the DCAT model and the repository metadata model,
5. the triggering of the conversion service to transform the repository metadata into the Semantic DCAT Application Profile and to publish on the FAIR Data Point.

For this PoC, we tailored the interface to provide the minimum metadata model proposed by task T2.2. and resolved a few technical issues. The configuration, debugging, and tailoring of the federated FAIR Data Space is now complete.

We are now working on creating the different API descriptions for the three target repositories, the mapping templates, and the converted metadata publication into a FAIR Data Point instance. These API descriptions and mapping files will be made available on GitHub together with the machine-readable version of the minimum metadata schema proposed by T2.2[34]. The FAIR Data Point will be made openly available for testing, and the link to the resource will be added to the GitHub repository.

4. Community Uptake - Results of a 2021-10 workshop

In October 2021, the task held a hackathon[35,36] where three dedicated stakeholders participated, partly in person and partly through remote connections.

Two of the participating stakeholders were repository operators (B2Share and Dataverse) the third was a registry (Re3data). The two repositories tried two different approaches: While the B2SHARE team tried to extend their functionalities to implement the FAIR Data

[29] https://bioportal.bioontology.org/
[30] http://agroportal.lirmm.fr/
[31] http://vocab.nerc.ac.uk/
[32] https://matportal.org/
[33] http://industryportal.enit.fr/
[34] https://github.com/FAIRsFAIR/SemanticDCAT-AP
[35] https://www.fairsfair.eu/events/make-your-repository-fair-enabling
[36] https://zenodo.org/record/5795809#.YffykfXML0o

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

Point protocol in their own code, Dataverse connected to the existing FAIRsFAIR reference implementation. Re3data investigated the possibilities to harvest metadata from repositories through the FAIR Data Point protocols and standards.

The hackathon results were published in the FAIRsFAIR git repository[37]. In the next part, we will present the achievements of the different working groups. To maintain readability we have avoided including code in this text.

1. B2SHARE

B2SHARE is a user-friendly, reliable and trustworthy way for researchers, scientific communities and citizen scientists to store, publish and share research data in a FAIR way. B2SHARE is a solution that facilitates research data storage, guarantees long-term persistence of data and allows data, results or ideas to be shared worldwide. B2SHARE supports community domains with metadata extensions, access rules and publishing workflows. EUDAT offers communities and organisations customised instances and/or access to repositories supporting large datasets.[38]

During the workshop, a FDP instance has installed and set up[39]. This remains available at the time of writing, but there is no intention of maintaining this in the future. During the workshop, one dataset https://fmi.b2share.csc.fi/records/e13c7cf71118462087c513e2407d88f2 was manually described to FDP: http://193.166.24.131:5000/dataset/3b368e6c-cd9b-4684-a84d-a662bfa9969f. There were special problems that make difficult Eudat B2Shares using the FAIRdatapoint: B2Shares are tailored by customer and each customer has own instance which has its own metadata model. The relevant metadata requires DDI-CDI SHACL extension to DCAT2.

The effort needed to automate metadata transfer is beyond the scope of a single workshop.

F-UJI tool is useful to pick low-hanging fruit and communicate with researchers.

At the moment, the technical challenges do not balance the benefits, but with increased demand from the customers, the effort can be taken up again.

EUDAT now knows that when/if a FDP is required, it can be implemented with the new code (not the old EUDAT FDP)

2. Dataverse

Dataverse is an Open Source data repository originally developed by Harvard IQSS. It's being widely used around the world for organising, managing, and showcasing datasets, and has a well established and active community. [40]

[37] https://github.com/FAIRsFAIR/repository-octoberfest-2.3
[38] https://b2share.eudat.eu/
[39] https://github.com/EUDAT-B2SHARE/B2SHARE-FAIR/tree/7ea00e0492a0b85510155c4b4400299a9101bc86
[40] https://dataverse.org/

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

At the workshop, the national instances of dataverse from the Netherlands[41] and from Norway[42] participated. The integration of Dataverse with FAIR Data Point (FDP) on the metadata level can bring a lot of advantages to the interoperability of data repositories that use Dataverse as the underlying technology. For example, FDP can be deployed as an infrastructure plugin in the "Archive in a box"[43] and integrated with Dataverse in order to add DCAT2 support with possibility to query its metadata with native SPARQL queries. The FDP standard, implementing DCAT2, incorporates different levels of metadata: it keeps metadata about the FDP itself, its catalogues, and datasets inside each catalogue. If a data provider reports when there is metadata change at the catalogue level to another FAIR data point that indexes it in that catalogue, it can copy the metadata in the standardised way and make it searchable at this level. It becomes possible as FDP uses DCAT2 as a vocabulary and Shapes Constraint Language (SHACL) as a basic mechanism for metadata verification. In that way, it's becoming possible to aggregate all available information from multiple locations and from different data providers. The only condition is to get it exposed as DCAT2, and that's really the purpose of the existence of FAIR Data Points.

During the workshop, the script was created to convert Dataverse metadata fields from citation block[44] to SHACL shapes required for FAIR Data Point metadata. It allowed testing the interface between Dataverse API delivering metadata in JSON format and the FDP based on DCAT2. As a result, a few selected datasets from the Dataverse network were successfully uploaded and deposited in the appropriate collections created in the SURF hosted sandbox of the FAIR Data Point registry.

There are some limitations of the selected approach that could be improved:

1. all available Dataverse metadata schemes should be converted to SHACL shapes separately and verified manually; in the current implementation, term URI fields are missing for the legacy reasons. However there is ongoing work on the ontologies and controlled vocabularies in Dataverse[45] and this issue with linking fields to their appropriate term URIs should be fixed soon. The new experimental Dataset Semantic Metadata API was introduced in the Dataverse release 5.6[46] (August 2021) and intended to support OAI_ORE[47] metadata import/export using json-ld. According to its Notational Conventions[48], the common term URI means both IRI [RFC3987] and URI [RFC3986] and json-ld natively supports IRIs without any special measures.
2. missing predicates can be mapped based on the metadatablock_id column available in the Dataverse metadata schema.

[41] https://dataverse.nl/
[42] https://dataverse.no/
[43] http://github.com/IQSS/dataverse-docker
[44] https://guides.dataverse.org/en/latest/admin/metadatacustomization.html
[45] https://zenodo.org/record/5845540#.YgpvDi-B1ZI
[46] http://github.com/IQSS/dataverse/releases/tag/v5.6
[47] https://www.openarchives.org/ore/
[48] https://www.openarchives.org/ore/0.9/jsonld

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 general, this data integration task is beneficial for the whole Dataverse network as it can potentially increase the interoperability level for all available datasets, as soon as metadata will be exposed by FAIR Data Point in RDF format using standard vocabularies like DCAT and DCAT2.

3. Re3data

Re3data is a global registry of research data repositories that covers research data repositories from different academic disciplines. It includes repositories that enable permanent storage of and access to data sets to researchers, funding bodies, publishers, and scholarly institutions. re3data promotes a culture of sharing, increased access and better visibility of research data. The registry has gone live in autumn 2012 and has been funded by the German Research Foundation (DFG).[49]

The last participant at the hackathon was not a repository but the registry Re3data. While a repository would use a FAIR data point to expose its (meta) data, a registry could harvest FAIR metadata from repositories. Therefore this group tried to understand the technical requirements and align their metadata models. This alignment will allow a machine arriving at the registry to follow the metadata to the repositories and harvest all of them.
The next steps will be mappings/resolving the different metadata vocabularies into compatible formats. During the workshop, the following points were discussed.

- Introduction into FAIR Data Point, discussion about DCAT adaption and concepts, as well as future developments
- Looking at current implementations and prototypes
- Observe the current re3data software architecture/code and discuss potential solutions for implementing FAIR DP including update mechanisms for repositories and properties by external parties

One indirect outcome of the workshop is that I incorporated FDP as a good example of machine-actionable implementation into the "NRW Zertifikatskurs FDM"[50] which is intended to introduce Librarians, Administration and other staff at the university to research data management.

5. Conclusion and future plans

All developments are stored in the public git repository[51] of the FAIR data point team. This repository will remain after the end of the FAIRsFAIR project and contains all developments that were done in the course of this project. The deployed reference implementation will be decommissioned since it can be rebuilt if needed.

[49] https://www.re3data.org/
[50] https://www.th-koeln.de/weiterbildung/zertifikatskurs-forschungsdatenmanagement_82048.php
[51] https://github.com/FAIRDataTeam/FAIRDataPoint

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

While the initial plan was to have 6 to 12 repositories participating as active developers and testers during the project, it became clear that this was not possible. We believe that there are two main reasons:

1. The repository members underestimated the technical knowledge needed to use the FAIR Data Point. While the deployment of a reference implementation works relatively intuitively, mapping existing metadata into DCAT2, especially the SHACL shapes, need a rather deep technical and semantic knowledge. Extensions of DCAT2 and the creation of the corresponding SHACL shapes were only managed with a lot of help from FAIR data point developers. The group around dataverse (See chapter 4.2) and the F2DS (See chapter 3.2) are working on solutions that will allow non-technical users to map their metadata to DCAT2 without a more profound knowledge of SHACL.
2. While the DCAT2 model allows flexibility, domain communities demand an agreement on a minimum metadata schema. However, these minimum metadata schemas do not exist cross-domain, making interoperability between domains challenging.

The end of the FAIRsFAIR projects does not imply the end of the development of the FPD but will close one chapter. Many other existing and upcoming initiatives will continue developing the technology and the metadata models. The work done in this task will remain as a piece in a large complex puzzle.

> "The development of FAIR can be compared to the World Wide Web development. While only technical experts could create content initially, after a few decennia, many people hosted websites themselves. Nowadays, buying websites is an out-of-the-box service by larger companies"
>
> Rob Hooft, DTL
> (during the Sustainability meeting of FAIRsFAIR January 2021)

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