D5.6 FAIR Research Data Management Tool Set Update

Lead Partner:	CINES
Authors	L. Candela, L. Frosini, F. Mangiacrapa (CNR-ISTI), O.Rouchon, B.Toulemonde, Y. Le Franc (CINES)
Version:	1.0
Status:	submitted
Dissemination Level:	Public
Document Link:	https://repository.eosc-pillar.eu/index.php/s/atEpqqBa28HnKMJ

Deliverable Abstract

This document is an update of D5.1, a report accompanying the delivery of the bundle of service instance(s) resulting from T5.1 and T5.2 activities. It provides a short summary of the work, the list of services and how to access them as also described in D5.2. The tool set aims at offering solutions for Research Data Management promoting the implementation of FAIR principles and practices.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 857650.

DELIVERY SLIP

Date	*Name*	*Partner/Activity*	*Date*
From: 31/07/2021	O. Rouchon L. Candela	CINES CNR	30/09/2021
Moderated by:	P. von Hartrott	Fraunhofer IWM	
Reviewed by:			
Approved by:			

DOCUMENT LOG

Issue	*Date*	*Comment*	*Author*
v0.1	24/09/2021	First draft	L. Candela (0000-0002-7279-2727), O.Rouchon (0000-0002-1816-5546)
v0.2	27/09/2021	Review	L. Berberi, M. Vernet
v0.3	28/09/2021	Updated with reviewers feedback	O.Rouchon L.Candela
v1.0	30/09/2021	Ready for submission	F. Galeazzi

TERMINOLOGY

https://eosc-portal.eu/glossary

Terminology/Acronym	*Definition*
API	Application Programming Interface
FAIR Data Point	A software enabling the implementation of a metadata repository providing access to metadata according to the FAIR principles.
Federated FAIR Data Space	A unifying data space aggregating datasets scattered across several data sources and repositories with the aim to give access to them according to the FAIR principles.
FDP	See FAIR Data Point
FFDS or F2DS	See Federated FAIR Data Space
Virtual Research Environment	A web-based working environment conceived to provide a community of practice with services and data of interest;
VRE	See Virtual Research Environment

Table of Contents

Executive summary 4
1 Introduction 5
2 Implementation and initial results 6
2.1 F2DS-Metadata Repository 7
2.2 F2DS Data Catalogue Service 10
3 Access to the bundle of services 14
4 Concluding remarks 15
References 16

Executive summary

The goal of the EOSC-Pillar WP5 “The Data layer: establishing FAIR data services at the national and transnational level”, is to create the settings for an effective sharing, exploitation and reuse of data across initiatives and communities partaking to EOSC-Pillar and beyond. To pursue this challenging goal, the project leverages and builds upon results from previous and ongoing projects as well as on the experience of the partners in the project. This combined expertise provides data providers and data consumers with a dedicated set of services (and accompanying training) supporting the creation of a data space where multiple datasets from separate locations are virtually joined by combining their metadata and are subsequently published in accordance with the FAIR principles. EOSC-Pillar takes advantage of several scientific uses cases ran as part of the project to aggregate research data from heterogeneous sources into a federated environment, also known as the Federated FAIR Data space, which has been developed and deployed in WP5.

1 Introduction

The aim of T5.1 and T5.2 is to provide a set of services for creating a Federated FAIR data space (F2DS). This F2DS provides tools for data producers to make their data more compliant with the FAIR principles and any other specific policies (T5.1) and to integrate them with other data coming from multiple disciplines that could then be accessed and reused by data consumers through dedicated interfaces (T5.2).

The advantage of such federated FAIR data space as piloted in EOSC-Pillar, which implementation involves development as well as support and collaboration with domain scientists, is manifold. With a F2DS, researchers will be able to search, find and retrieve data using a single access point and tool set. Not only does this saves working time because it masks the various access methods of different sources and offers a single access point through user and programming interfaces (UI and API), but it can truly deliver new insights, given the proper combination of selected search criteria and content of one or more data-sets is explored in unison.

Prerequisite to the aggregation of (meta)data sources is, on one hand, the normalisation through the use of existing common standards, starting from the access method and authentication, through authorisation and metadata and data formats. This normalisation, i.e. the adoption of such common standards and formats, is the only path to improve the FAIRness of the data. Although it is considered the most straightforward approach for unification and shows many results, the extent of the scientific domains, the dynamics of data collection over time and ultimately the rapidly changing information technology makes normalisation a challenge. On the other hand, technology is needed that can adapt to the changing and developing data types and at the same time presents a stable and flexible connection to the data sources for automatic parsing and data exploration. This is one of the goals in the implementation of, for example, the FAIR data point[1] software.

The solutions delivered in WP5 and described in this document consist of a variety of tools and services. Some tools are optimised for either solution, some maybe used in another context or are offered as an independent service. Components of several services are able to use existing resource offerings available in EOSC, e.g. virtual computing and data services. Eventually, many of the WP5 services can be easily deployed taking advantage of mature technologies such as Docker and Kubernetes. Furthermore, WP5 services could also be offered by the “as-a-Service” delivery model.

Most important, the presented solution, which integrates two component tools, offers the best possible adaptations to the specific domains , i.e. use-cases, they are developed for. It implements the EOSC-Pillar Federated FAIR Data Space (F2DS), a unifying data space that is built by aggregating and enriching datasets from a set of multidisciplinary repositories, i.e. data sources, with the aim to facilitate data discovery and re-use. Although datasets are the primary focus of the resulting data space, other items are managed including repositories and data sources, APIs, metadata schemas and ontologies.

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

2 Implementation and initial results

The implementation of the envisaged Federated FAIR Data Space concept is leveraging on existing tools and services that have been developed for similar requirements and have matured independently. Initially, it had been decided to follow an exploratory-oriented approach for the first release of the tool-set leading to the development and testing of two independent, alongside solutions. This approach was documented in D5.1 (Cazenave et al. 2020). The rationale behind that was to promote discussion and exchange between two possible implementations sharing some commonalities and technologies yet proposing slightly diverse workflows, technical approaches and delivery strategies. It became quickly apparent that the two proposed solutions were not completely disjoint, as they had complementary features that could fit together nicely. Thus, it has been decided to integrate the two tools as shown in Figure 1 – namely Metadata Repository (based on FDP) and Data Catalogue (based on D4Science) into one single solution – the Federated FAIR Data Space (F2DS). This was reflected in D5.2 (Candela et al. 2021).

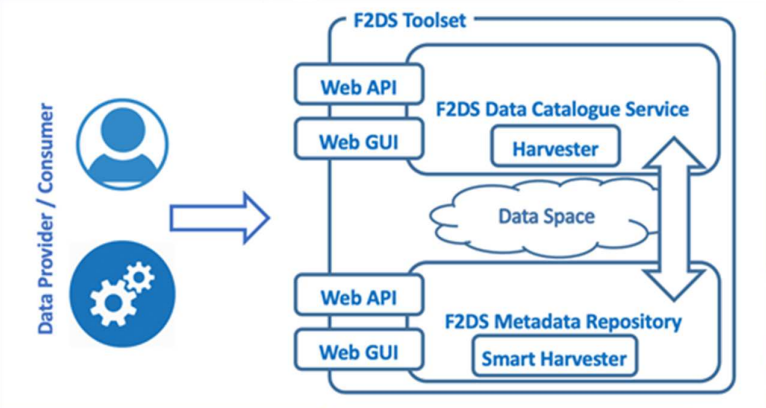

Figure 1. F2DS Overall Architecture

The Metadata Repository is built from scratch by integrating existing services for FAIRifying data and the latest of technologies and protocols, while the Data Catalogue is based on an existing service. The Metadata Repository implements a common API description and a simple metadata mapping to automatically and intelligently harvest, convert and publish metadata describing data sets in a single format (DCAT[2]). The Data Catalogue harvests metadata from the Metadata Repository with a dedicated harvester and makes them publicly available on a portal based on the CKAN[3] technology which is also used as the underlying framework of the EUDAT B2FIND[4] service. Such an integration will save duplicate effort, e.g. developing capabilities to search datasets in the Metadata Repository. Moreover, the Data Catalogue is designed and developed to facilitate the integration into virtual research environments, i.e. working environments offering data analytics and other collaboration oriented services to facilitate the exploitation of the data discovered and

[2] https://www.w3.org/TR/vocab-dcat-2/

[3] https://ckan.org/

[4] http://b2find.eudat.eu/

accessed by the catalogue. The high level workflow, which describes the different steps required to make datasets available in F2DS, is shown in Figure 2.

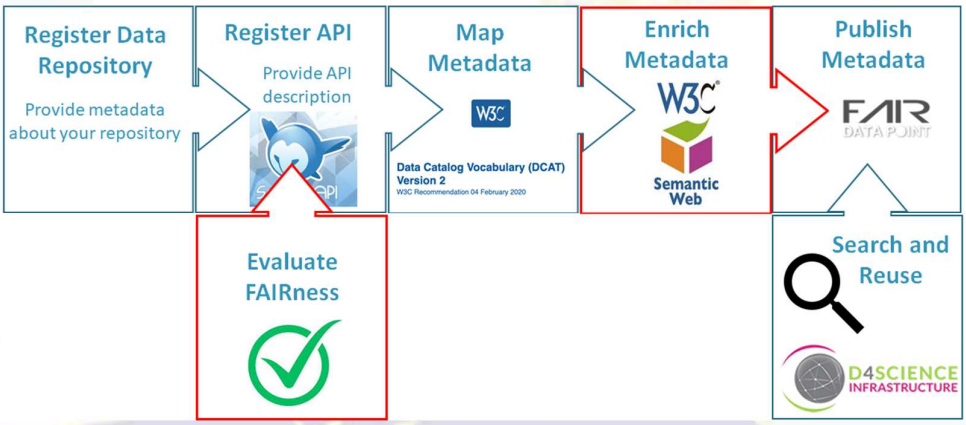

Figure 2. FFDS High-level Workflow

2.1 F2DS-Metadata Repository

The F2DS-Metadata Repository has been designed to use existing state-of-the-art tools developed by various EU stakeholders for FAIRifying data and to integrate them together into a coherent, scalable and innovative solution that can be easily deployed on any cloud infrastructure. To achieve this objective, it was therefore decided to take advantage of container technologies and deployments on Kubernetes[5].

The solution is based on the Fair Data Point technology[6] (FDP) linked to an access API registry (OpenAPI[7] technology). The generic architecture of this solution is shown below in Figure 3.

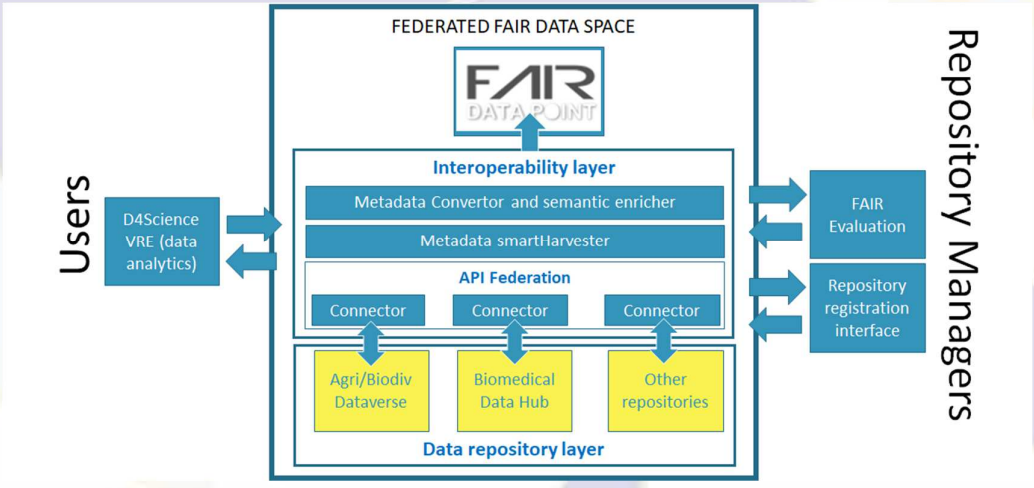

Figure 3. Architecture of the F2DS-Metadata Repository

[5] https://kubernetes.io

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

[7] https://swagger.io/specification/

The F2DS Metadata Repository includes two key layers: the data repository layer and the interoperability layer and offers two specific interfaces: one for data producers to register their repositories (Figure 4) within the F2DS and another interface to access the federated space for searching and using the data (Figure 9).

The first layer is the Data Repository layer which contains the metadata description of the different repositories (which will be made compliant with the EOSC Portal Service Description Template, or "profile") as well as technical information to access the repository content and the description of each repository's API in a common format.

The Data Repository Layer is then connected to the Interoperability layer which is based on two components: the metadata harvester and the metadata converter and enricher. The metadata harvester, also called smartHarvester, uses the API descriptions (Figure 4) to automatically build a dedicated client and the appropriate queries to gather the metadata stored in each repository: these clients make automatic links and regularly test the state of the metadata (additions, changes, deletions, etc.).

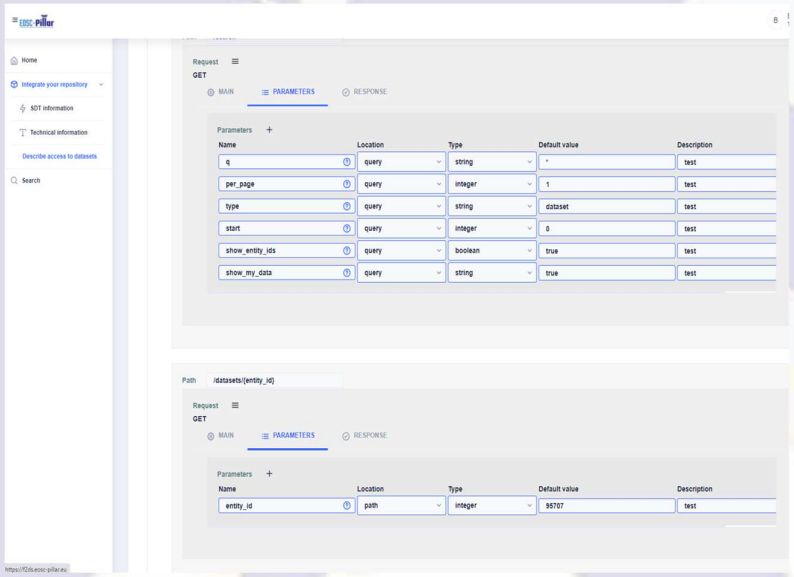

Figure 4. F2DS-Metadata Repository user interface: API description

Next, the metadata converter and enricher parses, transforms all the metadata into the DCAT model and serialises it as RDF[8] (see Figure 5). It uses pre-defined elements (catalogs/datasets/distributions) of the FDP, which have been specified in the DCAT standard vocabulary. Thus, all metadata will be described according to the same data model, which promotes interoperability and reuse of data.

[8] https://www.w3.org/RDF/

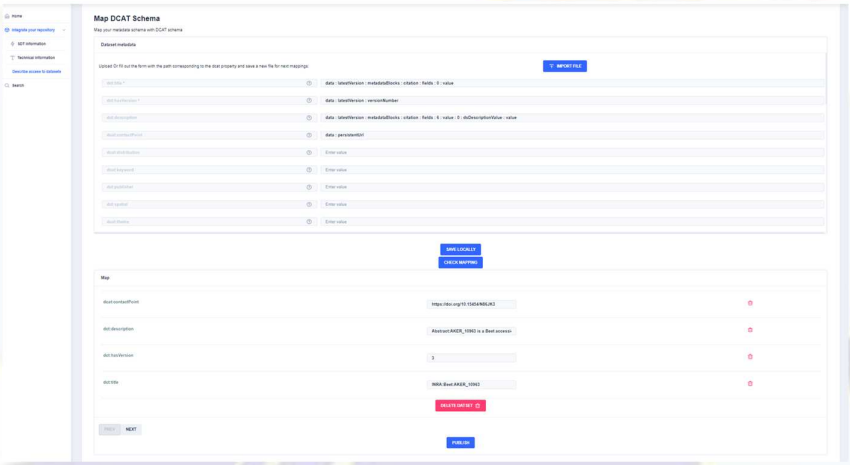

Figure 5. F2DS-Metadata Repository user interface: mapping metadata to DCAT

The heart of the F2DS is composed of a metadata storage called FAIR data point (FDP). FDP is developed within the GOFAIR initiative[9] and in the FAIRsFAIR project[10]. On one side, this service enables data owners to expose datasets in a FAIR manner and on the other side, allows data users to discover properties of the datasets through the exposed metadata and to access the data for download depending on the license condition.

This solution is being tested from a functional standpoint with the repositories from IRD Data Terra (T6.2), INRAE (T6.3) and INSERM (T6.6) as part of the scientific use cases from WP6, and should integrate more datasets in the upcoming releases.

In this version of the tool set, the focus has been set on the metadata harvesting and the DCAT mapping in order to publish the metadata to the FDP. The metadata enricher should be added in a later version of the service bundle and should be built leveraging the work of T5.5, e.g. using for example the FAIRifier[11] or OpenRefine[12] services which allows metadata enrichment with ontologies once it is confirmed which one is still the tool of choice.

During the remaining lifetime of the EOSC-Pillar project, a study will be carried out to evaluate the feasibility of the integration of F-UJI[13] (an automated tool developed as part of the FAIRsFAIR project which allows a programmatic assessment of the FAIRness of research datasets), and the connection of the FDP graph database with the FAIR Digital Object Framework (FDO-F)[14]. The outcome of this analysis, may possibly lead to the decision to implement these new features.

[9] https://www.go-fair.org/how-to-go-fair/fair-data-point/

[10] https://www.fairsfair.eu/

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

[12] https://github.com/FAIRDataTeam/Openrefine-metadata-extension

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

[14] https://www.go-fair.org/today/fair-digital-framework/

2.2 F2DS Data Catalogue Service

The F2DS Data Catalogue Service is developed by relying on the D4Science (Assante et al. 2019a, 2019b) offering and operational settings.

In particular, the F2DS Data Catalogue component is an instance of the D4Science catalogue extended and configured to serve the needs arising in the EOSC-Pillar project. The D4Science catalogue service is based on the gCube open source technology, that builds upon the CKAN technology for implementing its catalogue component. Two distinguishing features of D4Science catalogue service instances are: (i) the support for the definition of specific metadata profiles for the catalogue items, and (ii) the integration by design with Virtual Research Environments thus to provide the specific community served by the VRE with a custom view of the data space of interest.

The architecture of the F2DS Data Catalogue component is depicted in Figure 4. This picture describes (a) how the catalogue can be populated by using both contents from existing data sources via harvesters or contents uploaded directly in it via a GUI or an API[15, 16], and (b) how catalogue content is made available for data consumers via a GUI as well as a set of API and standards including DCAT.

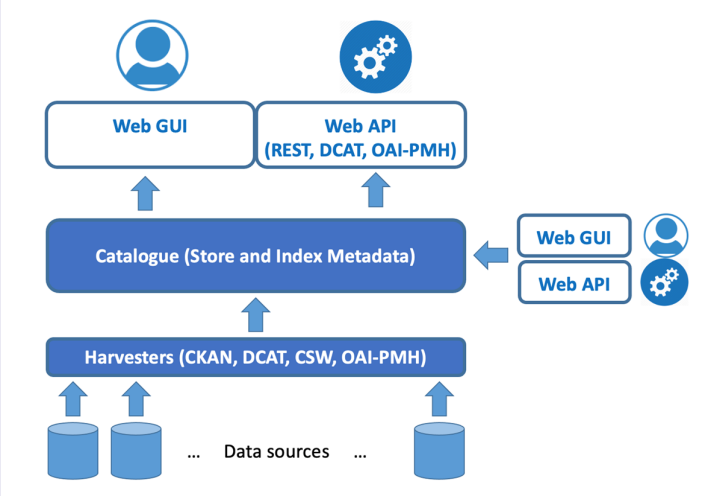

Figure 6. F2DS Catalogue Service Architecture

The F2DS Data Catalogue has been equipped with a specific harvester that systematically collects the contents aggregated by the F2DS Metadata Repository. In particular, this harvester relies on the DCAT API implementation offered by the F2DS Metadata Repository and map all the catalogues and datasets into its own data structures to offer search and browse facilities. Each Catalogue instance is integrated into a working environment that is specifically created to serve the needs of a designated community (Figure 5). This means that every working environment can be customised

[15] https://wiki.gcube-system.org/index.php?title=GCat_Service

[16] The facility to populate the catalogue directly complements the offerings facilitating the population of the F2DS as a whole. In fact, the F2DS repository offers a mechanism focusing on contents integrated into an existing repository while the functionality to publish datasets directly via the catalogue is for cases that are not covered by an existing repository.

by selecting the data space and the set of tools to be made available. Hence researchers use this custom environment for further developing and consuming the specific catalogue content.

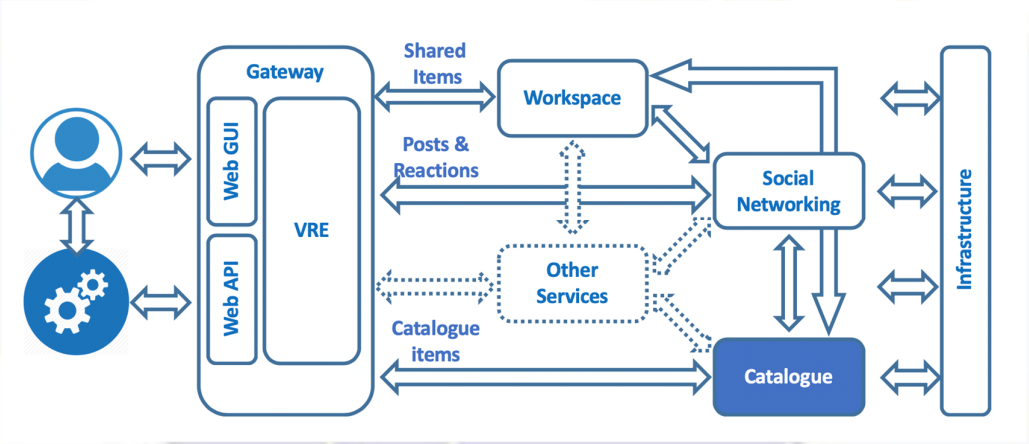

Figure 7. F2DS Data Catalogue integrated into a Virtual Research Environment

The EOSC-Pillar Research Data Catalogue Virtual Research Environment[17] has been created by instantiating these tools. It is a virtual research environment and proof of concept of a working environment facilitating the development of the overall F2DS and showcasing it. The environment offers the basic services enabling researchers to collaborate and share material (e.g. in a social networking area and in a workspace for storing files of interest) and publish data in a catalogue (enacting authorized users to publish new items and manage the published items). Instances of other tools can be added depending on forthcoming needs.

Figure 8 displays the welcome page of the VRE that has been instantiated by using the social networking facility. It also displays the menu with the services offered by this VRE where the F2DS Metadata Repository and Catalogue are accompanied with other services including FAIRness tools like F-UJI.

[17] https://eosc-pillar.d4science.org/web/eoscpillarresdatactlg

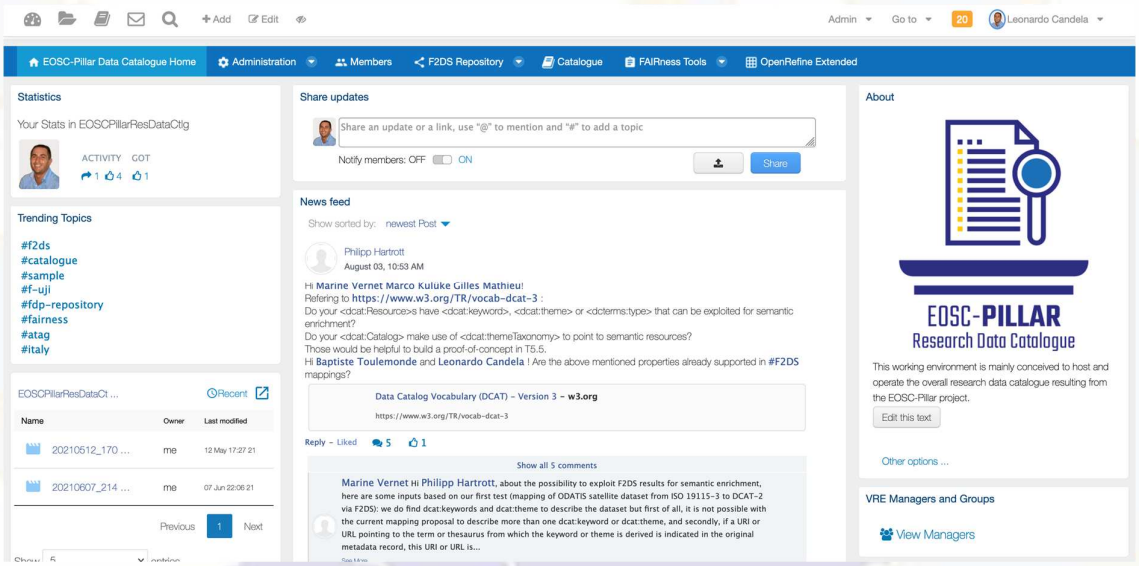

Figure 8. EOSC-Pillar Research Data Catalogue VRE Welcome page

Figures 9 and 10 display the F2DS Data Catalogue GUI. In particular, Figure 9 displays the Search and Browse view while Figure 10 displays how an item collected from the F2DS Metadata Repository is presented to its users.

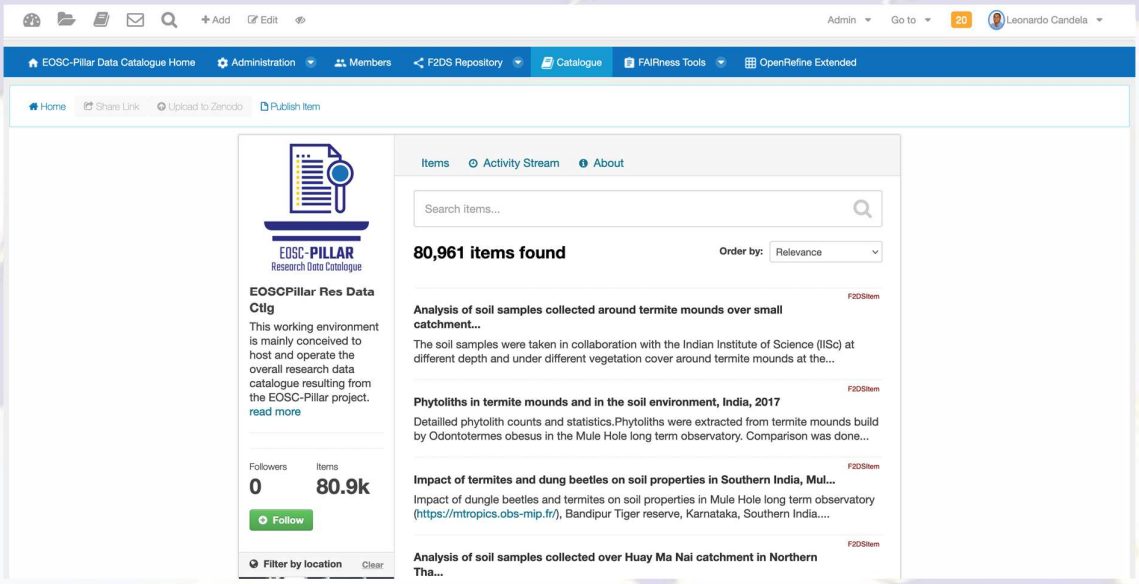

Figure 9. F2DS Catalogue GUI: Search & Browse View

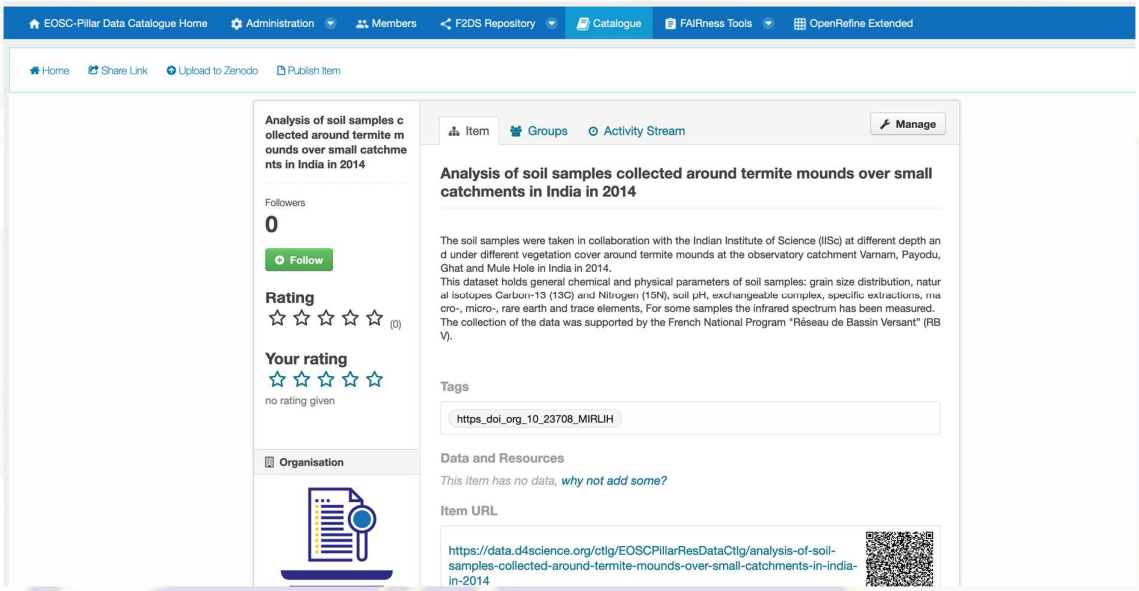

Figure SEQ Figure * Arabic 10. F2DS Catalogue GUI: Item View

3 Access to the bundle of services

To ease the access to the current state of work, we provide in the table below the list of services together with the partners hosting the services, the access URL and a short description.

DISCLAIMER

Please be aware that these services might not be available continuously because they are periodically updated. In the case where links are not working please contact by email the related contact person to obtain information on the status of the service i.e.:

- Olivier Rouchon (olivier.rouchon@cines.fr) for CINES hosted services
- Leonardo Candela (leonardo.candela@d4science.org) for the D4Science hosted services

Name of the service	Hosted by	Access URL	Description
FAIR Data Point API	CINES	http://f2ds.eosc-pillar.eu	API to access programmatically the content of the FAIR Data Point
FAIR Data Point User Interface	CINES	http://f2ds.eosc-pillar.eu/app	User Interface to access the content of the FAIR Data Point
FFDS Registration Interface	CINES	http://f2ds.eosc-pillar.eu/dashboard	User Interface for repository registration
FAIR Data Point SPARQL search	CINES	http://f2ds.eosc-pillar.eu/blazegraph	Explore metadata in FDP with SparQL query language
D4Science data catalogue	CNR-ISTI (by D4Science)	https://eosc-pillar.d4science.org/group/eoscpillarresdatactlg	The virtual research environment is available at (for authorized users)
D4Science data catalogue- public use	CNR-INSTI (by D4Science)	https://eosc-pillar.d4science.org/web/eoscpillarresdatactlg/catalogue	The publicly available version of the catalogue

4 Concluding remarks

This deliverable describes the tool-set implementing the Federated FAIR Data Space (F2DS) concept. In particular, the concept of F2DS has been formulated as a unifying space of datasets (and other typologies of items) stemming from several data repositories and data sources as well as from several communities. The aim of the Federated FAIR Data Space is to make heterogeneous and distributed datasets compliant to the FAIR principles.

These tools are being built in collaboration with the various communities involved in the project (WP6 use-cases) thus collecting their feedback to consolidate and extend the service offering of the EOSC-Pillar Federated FAIR Data Space.

The current implementation has now ingested metadata from diverse data sources and communities. Further data-sets will be added as the project moves forward to showcase the benefits to other communities as well. The development is not yet complete, and further releases - at least for the F2DS-Metadata Repository - are expected before the end of the project, which will include enhancement and new features. Ultimately the implementation will be turned into a genuine service and potentially included in the EOSC-pillar National service registry or EOSC wide service catalogue.

References

Assante, M. et al. (2019a) Enacting open science by D4Science. Future Gener. Comput. Syst. 101: 555-563 DOI: 10.1016/j.future.2019.05.063

Assante, M. et al. (2019b) The gCube system: Delivering Virtual Research Environments as-a-Service. Future Gener. Comput. Syst. 95: 445-453 DOI: 10.1016/j.future.2018.10.035

Candela, L., Cazenave, N., Frosini, L., Le Fran, Y., Mangiacrapa, F. (2021) D5.2 FAIR Research Data Management Workbench Operation Report. EOSC-Pillar Deliverable D5.2 https://doi.org/10.5281/zenodo.5513795

Cazenave, N., Candela, L., Berberi, L., van Wezel, J., Hashibon, A., Le Franc, Y. (2020) D5.1 FAIR Research Data Management Tool Set. EOSC-Pillar Deliverable D5.1 https://doi.org/10.5281/zenodo.4283400

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). DOI: 10.1038/sdata.2016.18