EOSC-Pillar
Coordination and Harmonisation of National &Thematic Initiatives to support EOSC

D5.1 FAIR Research Data Management Tool Set

Lead Partner:	CINES
Authors	N. Cazenave (CINES), L. Candela (CNR-ISTI), L. Berberi (KIT), Jos van Wezel (KIT), A. Hashibon (Fraunhofer), Yann Le Franc (CINES)
Version:	2.0
Status:	Final version
Dissemination Level:	Public
Document Link:	Portal users: https://repository.eosc-pillar.eu/index.php/f/35824 Public URL (PDF): https://repository.eosc-pillar.eu/index.php/s/MBgkN5kexKMbMoq

Deliverable Abstract

This document is accompanying the delivery of the bundle of service instance(s) that are the output of T5.1 and T5.2 activities. It provides a short summary of the work and the list of services and how to access them. The tool-set aims at offering solutions for Research Data Management promoting the implementation of FAIR principles and practices. This tool-set is updated every 6 months after the initial release. A major release is planned in June 2021 (PM 24).

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

COPYRIGHT NOTICE

DELIVERY SLIP

	Name	***Partner/Activity***	***Date***
From:	N. Cazenave L. Candela	CINES CNR-ISTI	
Moderated by:	Y. Le Franc	CINES	
Reviewed by:			
Approved by:			

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 from separated data sources and repositories with the aim to give access to them according to the FAIR principles.
FDP	see FAIR Data Point
FFDS	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

Contents

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 will provide 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 (Wilkinson et al, 2016). EOSC-Pillar performs a pilot into the federation of data sources presenting solutions that differ depending on the requirements of the use cases promoted by researchers.

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 should provide 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 accessed and reuse 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 save working time because it masks the different access methods of different sources and offers a single access point through user and programming interfaces (UI and API), but given the proper combination of selected search criteria and content of one or more data-sets is explored in unison it can truly deliver new insights.

Prerequisite to the unification 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 format 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 discussed 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. Several (components of) the 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 using for example the framework[2] developed in the Deep Hybrid DataCloud project and its predecessors. Most important, the presented solutions offer the best possible adaptations to the specific area i.e. use-case, they are developed for.

The tools presented implement the EOSC-Pillar Federated FAIR Data Space (F2DS), i.e. 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] available in the EOSC marketplace

2 Implementations 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. Because of this it was decided to follow a pragmatic and exploratory-oriented approach for the first release of the tool-set leading to the development and testing of two independent solutions along side. Such an approach promotes discussion and exchange between two possible implementations sharing some commonalities and technologies yet proposing slightly diverse work flows, technical approaches and delivery strategies. This target driven optimisation in two lines of development is to be evaluated at PM24 at which further development of either one of the solutions must be halted for reasons of efficiency. Even without ongoing development the service will function, enhance FAIRness of the presented dataset and continue to be useful for specific use cases. The two proposed solutions are not completely disjoint. Rather, it is likely that during their operation and exploitation they will be assembled into a single entry access.

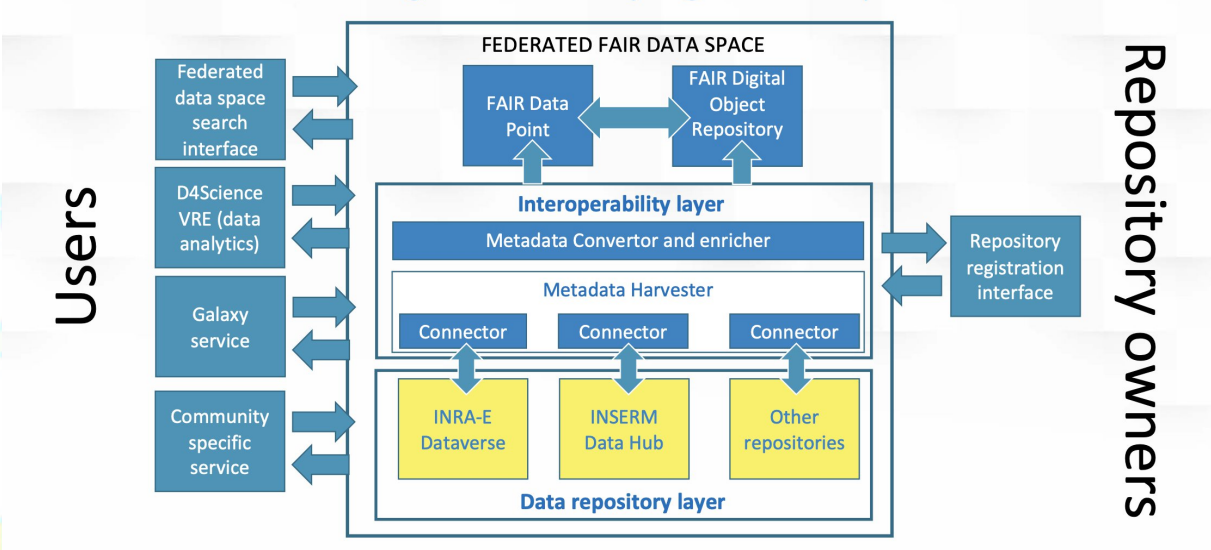
The two implementations of the F2DS for EOSC-Pillar are the Generic F2DS (g-F2DS) and the D4Science-based F2DS. The D4Science data space is based on an existing service whereas the g-F2DS implementation is built from the ground up by integrating existing services for FAIRifying data and the latest of technologies and protocols. While both offer the researcher a shared data space, there are some fundamental technical differences between the generic F2DS and D4Science based solutions. E.g. the g-F2DS approach uses 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[3]). D4Science is based on the CKAN[4] technology which is also used as the underlying framework of the EUDAT B2FIND[5] service and offers the possibility to harvest metadata using various formats (CSW[6], OAI-PMH[7],...). Both approaches offer capabilities to search datasets either through the rich search interface of D4Science solution, or the SPARQL [8] engine built-in the g-F2DS.

2.1 Generic F2DS solution

The g-F2DS solution 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 chosen to take advantage of container technologies and deployments on kubernetes[9].

[3] https://www.w3.org/TR/vocab-dcat-2/
[4] https://ckan.org/
[5] http://b2find.eudat.eu/
[6] https://www.ogc.org/standards/cat
[7] https://www.openarchives.org/pmh/
[8] https://www.w3.org/TR/sparql11-query/
[9] https://kubernetes.io

The solution is based on the Fair Data Point technology[10] (FDP) linked to an access API registry (OpenAPI[11]/smartAPI[12] technology). The generic architecture of this solution is shown below.



Architecture of the g-F2DS

The g-F2DS solution 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 within the F2DS and another interface to access the federated space for searching and using the data.

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 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 (changes, deletions, modifications, additions, ...). Next, the *metadata converter and enricher* parses, converts all the metadata into the DCAT model and serialises it as RDF[13]. It uses pre-defined elements (catalogs/datasets/distributions) of the FDP. Thus all metadata will be described according to the same data model, which promotes interoperability and reuse of data.

[10] https://github.com/FAIRDataTeam/FAIRDataPoint-Spec
[11] https://swagger.io/specification/
[12] https://smart-api.info/
[13] https://www.w3.org/RDF/

The heart of the F2DS is composed of a metadata storage called **FAIR data point (FDP).** FDP is developed within the GOFAIR initiative[14] and in the FAIRsFAIR project[15]. 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.

In this version of the tool set, we have been focusing on the metadata harvesting. The metadata enricher should be added in a later version of the service bundle and should be built using the FAIRifier[16] service which allows metadata enrichment with ontologies (leveraging the work of T5.5).

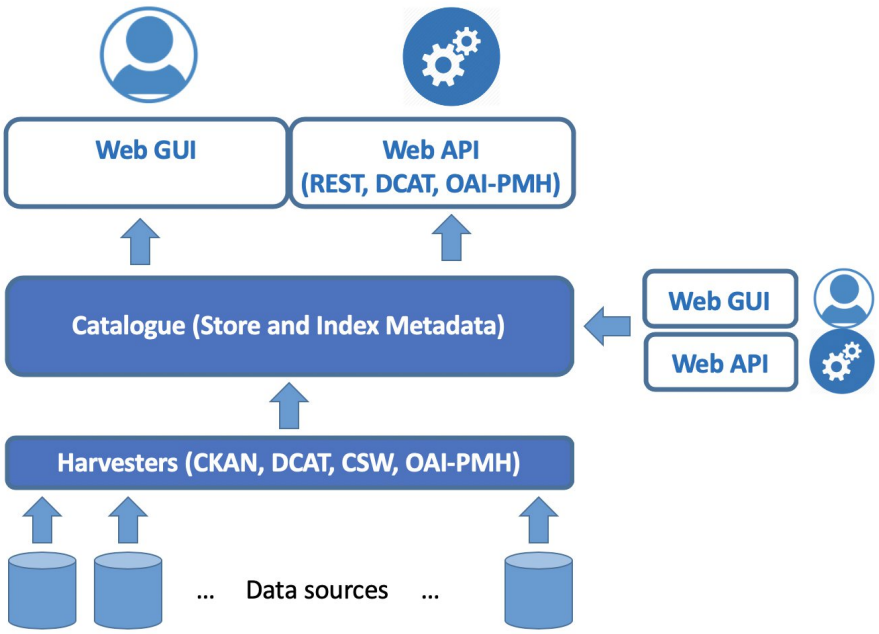
During the course of the EOSC-Pillar project, the FDP graph database is planned to be connected with the FAIR Digital Object Framework (FDO-F)[17].

Finally, this solution is being tested with the repositories from INRA-E (T6.3) and IRD Data Terra (T6.2) and should integrate more datasets in the upcoming releases.

2.2 D4Science solution

The second technical solution to build the federated FAIR data space is characterised by the exploitation of the D4Science service (Assante et al. 2019a, 2019b) offering and operational settings. In particular, it has a catalogue as the key component that is capable to harvest items from data sources as well as to enable users to publish new items according to user/community defined profiles. Catalogue instances are integrated into working environment that is specifically created to serve the needs of their designated communities. This means that every working environment can be customised 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.

[14] https://www.go-fair.org/how-to-go-fair/fair-data-point/
[15] https://www.fairsfair.eu/
[16] https://github.com/FAIRDataTeam/FAIRifier
[17] https://www.go-fair.org/today/fair-digital-framework/



Architecture of the D4Science based F2DS

The ***EOSC-Pillar Research Data Catalogue*** has been created as instance of these tools. It is a virtual research environment and proof of concept of a working environment facilitating the development of the overall F2DS. 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. The catalogue is planned to be populated with datasets and any other digital artifacts stemming from the EOSC-Pillar use cases represented in WP6. Currently the catalogue contains more than 80K items, mainly resulting from harvesting content from the INRAE Data Repository (data.inrae.fr) via the OAI-PMH protocol and the IFREMER Catalogue (https://sextant.ifremer.fr/geonetwork) via the CSW protocol.

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.:

- Nicolas Cazenave (cazenave@cines.fr) for CINES hosted services,
- Leonardo Candela (leonardo.candela@d4science.org) for the D4Science hosted services, and
- Lisana Berberi (lisana.berberi@kit.edu) for the service at KIT.

Name of the service	Hosted by	Access URL	Description
FAIR Data Point API	CINES	http://ffds.eosc-pillar.eu	API to access programmatically the content of the FAIR Data Point
FAIR Data Point User Interface	CINES	http://ffds.eosc-pillar.eu/fdp	User Interface to access the content of the FAIR Data Point
FFDS Register front	CINES	http://ffds.eosc-pillar.eu/front	User Interface for repository registration
FAIR Data Point SPARQL search	CINES	http://ffds.eosc-pillar.eu/blazegraph	Explore metadata in FDP with SparQL query language
D4Science data catalogue	C N R - I S T I (by D4Science)	https://eosc-pillar.d4science.org/group/eoscpillarresdatactlg	The virtual research environment is available at (for authorized users)
D4Science data catalogue- public use	C N R - I S T I (by D4Science)	https://eosc-pillar.d4science.org/web/eoscpillarresdatactlg/catalogue	The publicly available version of the catalogue

CORDRA instance	KIT	http://fc05f45d-4bab-441b-8d8b-b96a523576d6.ka.bw-cloud-instance.org:8080/	Repository compliant with Digital Object specification to be updated with upcoming FAIR Digital Object specification

4 Concluding remarks

This deliverable describes an initial tool-set supporting the implementation of 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 implementations have both ingested metadata from diverse data sources and communities. Further data-sets will be added in the coming months to showcase the benefits to other communities as well. Each of the implementations will thereby equally evaluated in the field. Ultimately the implementations will be turned into real services and be included in the local or EOSC wide services catalogue.

This first bundle of services focuses specifically on the services involved in creating a federated data space. In the next iteration, facilities for data enrichment will be added as well as changes resulting from the feedback collected from the use cases will be incorporated.

As dictated in the Description of Work, another major release will be done in Project Month 24.

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

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