This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

Call: NFRP-2018
(Nuclear Fission, Fusion and Radiation Protection Research)
Topic: NFRP-2018-11
Type of action: CSA

Project:
“Fair4Fusion – open access for fusion data in Europe”

Blueprint architecture for a Fusion Open Data Framework

Version	Version 1.6
Type	Report
Dissemination level	Public
Lead Beneficiary	PSNC
Date	24.05.2022

Authors:

PSNC: Marcin Płóciennik, Bartosz Bosak, Raul Palma, Michal Owsiak, Michał Urbaniak, Piotr Grabowski
UKAEA: Shaun de Witt, George Gibbons, Nathan Cummings
NCSRD: Iraklis Klampanos, Iris Xenaki, Andreas Ikonomopoulos, Stasinos Konstantopoulos, Vangelis Karkaletsis
Chalmers: Pär Strand
CEA: Frédéric Imbeaux
MPIPP: David Coster
EPFL: Joan Decker, Yves Martin, Olivier Sauter

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

Abbreviations, terms and definitions

Acronym	Description
EU	European
FAIR principles	FAIR is an acronym for Findable, Accessible, Interoperable, Reusable. These are recommended principles towards Open Science. See [1] for a detailed description of these principles.
IMAS	ITER Integrated Modelling and Analysis Suite. This suite of interoperable analysis codes, sponsored by the ITER Organization, is based on a machine-generic ontology, the Data Dictionary. A useful reference explaining the underlying principles of the Data Dictionary is [2].
AAI	Authentication and Authorisation Infrastructure that simplifies access to online resources through the use of a standard authentication procedure.
Open Data	Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. See [3].
Data	In this report, we address experimental data, which encompasses machine description, calibration information, raw data acquired during an experiment and the data processed from those. In addition we also address simulation data.
Metadata	In this report, we define the metadata as a subset of physical data that are made searchable in order to do Data Mining and/or to find plasma discharges of interest.
Annotation	The information inserted by users and associated with the metadata.
Experiment	An experimental magnetic fusion device, operated for research purposes: tokamak, stellarator, …

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

1. Executive summary

The overall objective of the Fair4Fusion project is to demonstrate the impact of making data from fusion devices more easily findable and accessible. The main focus towards achieving this goal is to improve FAIRness of the fusion data to make scientific analysis interoperable across multiple fusion experiments. This blueprint report aims for a long term architecture for the implementation of a Fusion Open Data Framework.

By making data from different fusion experiments more readily available and accessible through common interfaces we increase the possibility of broadened collaborations on the European level and thus help facilitate new scientific results and enhanced impact. With the FAIR approach extended to also cover simulation and modelling results we are bringing together the elements needed to form a broadened research arena for the European fusion community where each individual researcher and/or research group can contribute more efficiently to the joint research programme. This arena will also maximise the exploitation of data, publications, software and other research artefacts.

We present this blueprint for the benefit of the joint European research programme as well as the international devices and collaborations that extend it, in particular ITER and JT60-SA. As the implementation demands a certain level of coherence and integration within the current programme, the document is targeted toward the EUROfusion programme manager, members of the General Assembly, and EC representatives for implementation on joint experiments and modelling activities. As a significant fraction of European fusion research is done in joint collaboration with domestic programmes, the support and commitment from administrative, scientific and technical leadership of the individual experiments and programs is needed for a successful implementation and we are therefore aiming this blueprint directly towards them as well. Finally, with new publicly funded devices coming online in the coming decade, we see that there would be mutual benefits from adopting the FAIR philosophy and the technical implementation promoted here also in these devices and we are presenting the blueprint also in this context.

Currently, largely for historical reasons, almost all experiments are using their own tools to manage and store measured and processed data as well as their own ontology. Thus, very similar functionalities (data storage, data access, data model documentation, cataloguing and browsing of metadata) are often provided differently depending on experiment.

We have collected a number of user stories about searching for and accessing data and/or metadata, as well as some of the wishes from data providers. These use cases present the different perspectives of members of the general public, EUROfusion researchers and data providers that are the main target users of analysed scenarios. The basic requirements and user stories have been transformed into a list of functionalities to be fulfilled. Those functionalities in general have been grouped in several categories: search, visualisation and accessing outputs, report generation, user annotation, curation management, metadata management, subscriptions and notifications, versioning and provenance, authentication, authorization/access restrictions, accounting, licensing. Subsequently, the collection of functionalities has been used as the basis for the iterative process of architecture design. In the first step, the very generic concept of the architecture has been materialised and presented to the project community. Once this basic picture had been evaluated we were able to develop a more detailed architecture that was the subject of further improvements.

We are assuming the use of the IMAS Data Dictionary as a standard ontology for making data and metadata interoperable across the various EU experiments, for the following reasons:

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

- It is designed as a machine-generic ontology, capable of covering all experiment subsystems and plasma physics, and is extensible;
- It is the only ontology standard that has been elaborated in the fusion community (with the exception of the “CPO” data model [4], which can be considered as the predecessor of the IMAS Data Dictionary);
- It represents simulation and experimental data with the same data structures, enabling direct comparisons;
- It provides the possibility to store and easily access complete information about a subsystem (e.g. machine description, calibration coefficients, as well as the more usual raw and processed signals), while such information may be sometimes difficult to find in present experiment databases (if present at all in the database);
- It comes with Remote Data Access methods and a database organisation. Although these features are beyond the primary aspect of ontology and thus are optional technologies, they are also useful in the context of this blueprint architecture;
- It is already used by a number of EUROfusion Work Packages (WPCD, WPISA), projects (EUROfusion databases) and even an experiment (WEST);
- It is the standard ontology for ITER scientific exploitation;
- Even if managed and owned by the ITER Organization (IO), EU labs have access to it and EUROfusion has already a formal collaboration with the IO on development and usage of IMAS.

The resulting architecture of the system, presented in its simplistic form in Figure 1, consists of 3 main building blocks, namely *Metadata Ingests*, *Central Fair4Fusion Services* and *Search and Access Services*. Metadata Ingests are the entry point to the system for the metadata provided by Data Repositories associated with experiments. In the proposed design, Metadata Ingests stay within the administration of particular data repositories, thus the data repositories themselves can filter or amend data before they decide to expose it to the rest of the system. From Metadata Ingests, the metadata is transferred to the next block of the system, i.e. Central Fair4Fusion Services. The Core Metadata Services, being the heart of this block and the entire system in general, natively operate on the IMAS data format, but thanks to the translation components can accept different formats of metadata as input. Central Fair4Fusion Services provide supplementary functionality for specification of data that is not strictly tied to experiments, such as user-level annotations or citations. The last main block of the system is a set of Search and Access Services. It contains all user-oriented client tools that integrate with the Central Fair4Fusion Services. At this level of the system, key importance is given to the Web Portal that is expected to offer an extensive set of functionalities for searching, filtering or displaying metadata and data managed within the system.

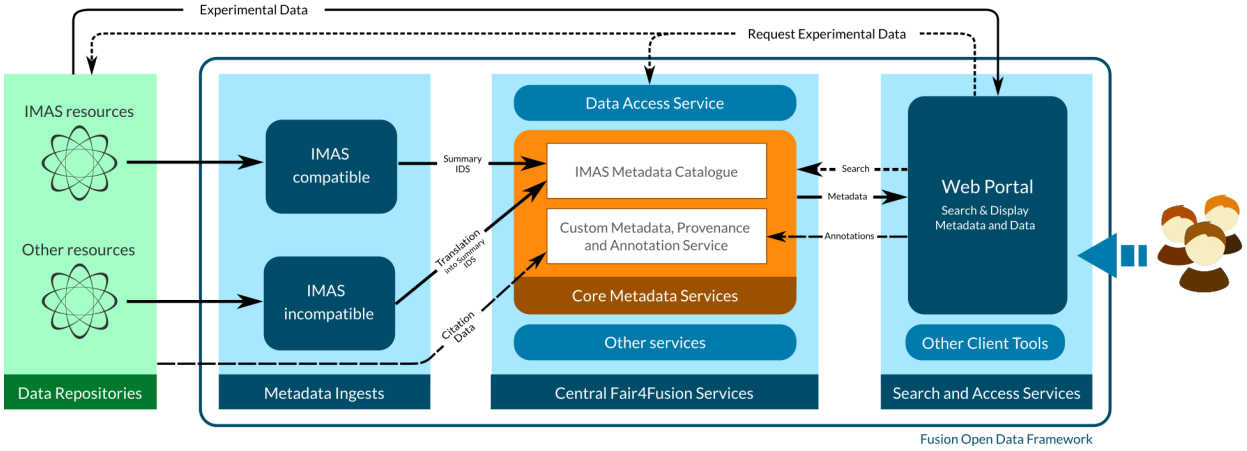


Figure 1. High-level overview of the architecture for a Fusion Open Data Framework

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

2. Introduction

2.1 Objectives

The overall objective of the Fair4Fusion project is to demonstrate the making of European funded data more widely available to the fusion community, other science communities, funding bodies, and the public at large in order to maximise the impact of the research. This means ensuring that the appropriate data is identified and given a correct classification (e.g. open, embargoed, restricted or closed, including appropriate licensing), providing a means of discovering the data and understanding its scientific content, providing methods for accessing the data, ensuring data (and metadata) quality and consistency, enabling secure access when required, etc. The key underpinnings of open data are excellent **data management policies** and adherence to **FAIR principles.**

FAIRness of the data. A key objective for improving the FAIRness of the fusion data would be to provide to the EU fusion community a way **to make scientific analysis interoperable across multiple fusion experiments, increasing the potential for new discoveries**. The benefits are to be found not only for the usual manual database queries but would also enable the use of new methods of research with Data Mining and Machine Learning techniques at an unprecedented scale.

Necessity of open data. The plethora of information collected is generally fine for experienced users to navigate, but in order to obtain the maximum benefit from open data, it is important to understand what are the primary information sources users actually want access to, and based on policies, how access can be granted to each level. Within the fusion community it is important to assess each data set in terms of '*as open as possible, as closed as necessary*'.

This **blueprint architecture** aims to provide a long term architecture for a **Fusion Open Data Framework** implementation. This blueprint architecture presents:

- the reference architecture,
- recommendation of the best technical approaches for providing easy discoverability and access to data,
- recommendations on standards, achieving interoperability, type and granularity of metadata and persistent identifiers to expose,
- an investigation on the use of metadata annotations to allow enrichment and enhancement of the semantics of the exposed metadata.

To ensure good coverage of the requirements, policies as well to increase the possible uptake, implementation and impact of this architecture, the project involves representatives of all the major European tokamaks; CEA operates the WEST tokamak, MPIPP operates ASDEX Upgrade, EPFL operates TCV and UKAEA operates both the MAST tokamak and, on behalf of EUROfusion, JET.

2.2 Scope

This document is the next update to the blueprint architecture that was initially published in Month 12 of the project as an early draft. Its final version is foreseen at the end of the project at month 30.

The scope of this document is a description of the target architecture of the Fusion Open Data Framework system for the aggregation and management of metadata coming from distributed fusion resources. It should be stressed that such aspects as advanced data management, advanced data processing or

metadata management at particular experimental sites are out of the scope of the Fair4Fusion project and won't be addressed in this document.

2.3 Document organisation

Section 3 of the document presents the Fusion community and experiments background, as well as introduces the FAIR data concept and describes the basic description of experimental fusion data. Section 4 discusses in detail the current state of the art - so the starting point as well as existing obstacles in terms of policies, data access and FAIRness. It also introduces existing standards and ontologies used to describe the metadata. Section 5 provides categorization of the user groups, their roles and possible access policies, and describes the leading user stories and their requirements. It is summarised with the list of the required functionalities that the blueprint architecture should aim for. The following section introduces the policy recommendation for the architecture and the baseline architecture with the components description and the relationship between them, as well as describes the protocols and standards. Section 7 concludes with the summary and the next steps.

3. Background

3.1 Fusion community

The fusion 'community' within Europe can trace its history back to the 1958 signing of the Euratom Treaty ("The Treaty establishing the European Atomic Energy Community") and still stands as an independent entity, although a part of the Treaties of the European Union. Currently 30 research organisations, and behind them about 150 affiliated entities including universities and companies, from 25 European Union member states plus the United Kingdom, Switzerland and Ukraine are members of the EUROfusion consortium [5] that represents the collaborative spirit of the European fusion research landscape by supporting and funding fusion research activities on behalf of the European Commission's Euratom programme.

There are 18 experimental fusion devices at a number of sites across Europe producing tens to hundreds of terabytes of experimental data per year. Beyond that, many universities and academic institutes work on materials science, plasma physics, nuclear physics, technology, laser physics, robotics and instrumentation related to the development, evolution and operation of fusion devices, and modelling codes can provide additional tens to hundreds of terabytes. The next large-scale fusion experiment, ITER, is projected to produce up to 2PB of data per day when fully operational.

The fusion community is a long established one with a legacy of security being at the forefront of its work. This history means that many data management processes are now well established and have led to successful and safe operation of tokamaks and quality science and engineering produced over many decades. Data management, while adhering to the rules established at the time, was delegated to local site operations which has led to a significant divergence in data stewardship between different tokamak sites across Europe and beyond, including different formats, metadata schemas, data reduction process and nomenclature. Indeed, even security is currently delegated to sites, with different experiments operating different policies for accessing the data. However, having such long established and successful methods also means that any change in these site policies should have negligible, or no, impact on current operations but should be seen as an 'added value' operation outside the normal scope.

The European fusion community has become increasingly collaborative over the last few decades with more experimental devices becoming available for broader groups of researchers thanks to investments made by EUROfusion and its predecessors. The diversity of devices is a great strength of the programme, but as each facility largely has developed their own data technologies, philosophies and access methodologies it has in some cases also presented challenges in sharing data even between collaborating scientists. Opening the data up and making them more easily available on a pan-European basis is a key ingredient in exploiting the investments in the research infrastructures made so far. Across Europe there has been a move to make publicly funded research data more open and accessible based on the G8 open data charter signed in June 2013. This effort is happening both nationally [6],[7] and across national borders [8]. An effort towards an Open Data environment for European fusion research can spearhead also efforts to be made in ITER.

3.2 Fusion experiments

The Joint European Torus (JET)

JET is currently the world's largest nuclear fusion experiment and has been operational servicing the fusion community since 1983. It holds several records in terms of progress towards sustainable fusion and has undergone many enhancements over its lifetime, from testing new diagnostic methods, through complete changes to the plasma wall material through being run with different fuels and different methods of plasma heating. It holds a large number of records for fusion energy devices, including the highest Q value recorded (the ratio of power in to power out), the highest fusion power output and the highest plasma current. In support of ITER operations, JET is embarking on the first experimental campaign using Deuterium-Tritium (DT) as a fuel source since 1997. JET is currently the only tokamak capable of running DT plasmas.

WEST

The WEST tokamak is operated by CEA in Cadarache, France, close to ITER. WEST provides an integrated platform for testing the ITER divertor components under combined heat and particle loads in a tokamak environment. It will allow assessing the power handling capabilities and the lifetime of ITER high heat flux tungsten divertor technology under ITER-relevant power loads (10–20 MW m^{-2}), particle fluence (~10^{27} D m^{-2}) and time scales (above 100 s). In order to fulfil its scientific objectives, WEST is equipped with upper and lower divertor coils, W coated upper divertor, baffle, inner bumper and with a flexible lower divertor made of twelve 30° sectors where the ITER-like W monoblocks are being installed. The additional heating and current drive power is provided by high frequency heating systems, namely ion cyclotron resonance heating (ICRH) and lower hybrid current drive (LHCD), delivering up to 9 MW of ICRH power and 7 MW of LHCD power.

ASDEX Upgrade

The ASDEX Upgrade tokamak is sited at the Max Planck Institute for Plasma Physics in Garching, Germany and started operation in 1991. It is designed to operate with plasma currents up to 1.6 MA and a toroidal field of up to 3.1 T, though typical discharges are operated with 1 MA and 2.5 T and a pulse length of up to 10s. 20MW of neutral beam injection (NBI), 6 MW of ion cyclotron heating (ICRH) and 8 MW of electron cyclotron heating (ECH) are available. Over the nearly 30 years of operation it has performed 40000 plasma discharges.

TCV

TCV is a medium sized tokamak located at the Swiss Plasma Center of EPFL, in Lausanne, Switzerland. It's main specificity is a strong capability of plasma shaping via a series of 16 poloidal field (PF) coils placed on both sides of the highly elongated, rectangular, vacuum vessel cross section. It allows a wide coverage

of the traditional plasma shaping parameters such as elongation and triangularity, as well as developing new plasma configurations such as snowflake or super-X divertors. In addition, a highly flexible Electron Cyclotron Heating (ECH) system allows heating of predefined plasma layers, and it's combination with a powerful Neutral Beam Injection (NBI) system enables a wide range of plasma electron to ion temperature ratio.

ITER

ITER ("The Way" in Latin) is one of the most ambitious energy projects in the world today.
In southern France, 35 nations are collaborating to build the world's largest tokamak, a magnetic fusion device that has been designed to prove the feasibility of fusion as a large-scale and carbon-free source of energy. ITER will be the first magnetic confinement fusion device to produce net energy. ITER will be the first fusion device to maintain fusion for long periods of time. And ITER will be the first fusion device to test the integrated technologies, materials, and physics regimes necessary for the commercial production of fusion-based electricity.

MAST

The Mega-Amp Spherical Tokamak (MAST) and it's upgraded configuration (MAST-U) are non-traditional devices allowing more compact configurations with a smaller central core. This configuration is of interest because theory demonstrates it should be less prone to instabilities and production costs should be reduced. MAST represents the UK's national contribution to the MST (Medium Scale Tokamak) program and was first operational in 1999. Since 2013 it has undergone significant refurbishment to increase the heating power, plasma current, magnetic fields and pulse length, Importantly MAST-U has installed a novel divertor known as the Super-X divertor which will reduce the heat load by a factor of 10, overcoming one of the issues around commercial fusion where the divertor would be required to handle very high heat loads with normal configurations.

3.3 Fusion data

The community as a whole creates a wide range of data from experiments covering a range of parameters of interest both for physics and engineering purposes from a wide range of sensors, as well as from a variety of modelling activities. From these diagnostic measurements, a wide range of physics information related to the plasma and vessel itself are derived. In addition, calibration requires data regarding the experimental and sensor configuration in order to convert the raw data into scientific information. Typically, both the raw data, the calibration information and the calibrated science products are stored at full temporal and spatial resolution, but also summary products are created which present an easily understandable summary of the main subjects of interest either at low resolution or simply average values over the time series. Largely for historical reasons, almost all experiments are using their own tools to manage and store measured and processed data as well as their own ontology. Thus, very similar functionalities (data storage, data access, data model documentation, cataloguing and browsing of metadata) are often provided differently depending on experiment. Modelling data is more varied and initial proposals include the creation of a Long Term Simulation Storage facility, the storage of metadata in the SUMMARY IDS and the use of IDSs to store simulation results. Other fusion data will need to be brought into the FAIR process at a later stage outside of FAIR4fusion.

3.4 FAIR

The FAIR principles [9] are 15 guidelines to ensure that any data generated is Findable, Accessible, Interoperable and Reusable. FAIR provides a framework for easing discovery of data, encouraging suitable

licensing and ensuring that data (or information about the data) can persist over time spans of ten years or more as well as ensuring suitable Authentication and Authorization processes are in place. For data to be FAIR there are 15 policies which should be adhered to, and most of these relate in some way to either metadata, persistent identifiers and licensing. However, there have been many nuances and interpretations of these, notably from the Research Data

Alliance Working Group on Fair Data Maturity Model [10] and the EOSC Secretariat FAIR Working Group recommendations on FAIR metrics for EOSC [11], which add a level of complexity and clarity. Typically, at a minimum, this means that FAIR data requires a well-defined, and preferably machine readable, metadata schema with persistent metadata objects (such that the metadata can exist beyond the lifetime of any data it is associated with), clear rules and protocols for allowing access to the data (including licensing information and restriction on usage), a globally unique and resolvable persistent identifier at an appropriate granularity and standards based methods of presenting the data either through suitable APIs and/or using common formats. Often supporting this is a well-defined provenance schema, to increase the trust in the data, and a data dictionary or ontology service to support cross disciplinary usage of the data.

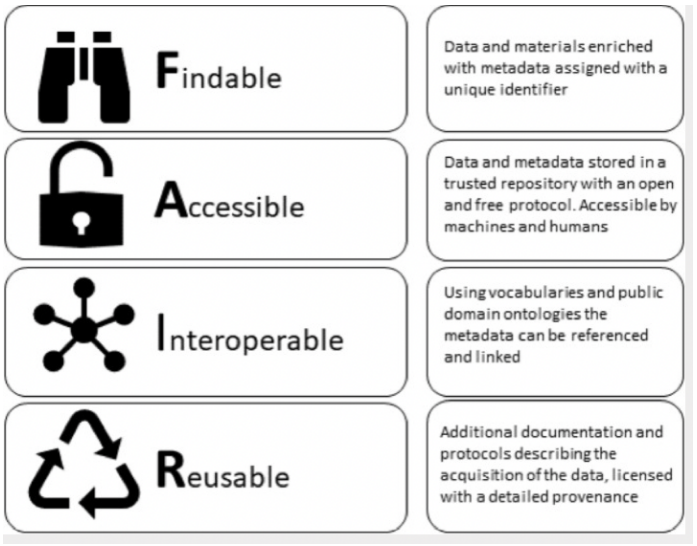


Figure 2. Schematic representation of FAIR data principles [11]

One of the goals of this project is to demonstrate to the community the advantages of FAIR data, not necessarily in a global context but making data “community FAIR”. This is achieved by taking samples of existing data, services and tools and applying FAIR principles (see Figure 2), making the data from a range of devices discoverable and accessible in an interoperable way from any site and by any authorised user.

Within the scope of this project the goal is to promote “Community FAIRness” by demonstrating that many of the tools and services required to support FAIR principles already exist, and can be implemented with no or minimal changes to existing local processes - with an aim for these to be additional steps in the process rather than disruptive changes. In addition, the project has the goal of promoting a more general open approach to sharing data across community sites and to the wider science community. Fusion has received much attention as a future power source due to its “green-ness” and it is important that the community can demonstrate that their work is advancing this goal not only to funders, but to the wider public to ensure support can be built up from public opinion.

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

4. Current state of the art

The detailed analysis of the current state of the art was performed at the beginning of the project and is detailed in the deliverable D2.1 [12]; in this section we present highlights of these findings.

4.1 Policies

All European tokamak and stellarator experiments grant access to their measured and processed data on an individual basis to collaborators who are formally identified as members of the experiment's team. In some cases (e.g. W-7X), researchers are required to sign a data access user agreement to become part of the experiment team. An individual computer and data access account is created, with password protection allowing authentication of the user as part of the experiment's team. Technically, the authentication is done by various means, e.g. JET uses a multi factor authentication with SecurID key, WEST implements IP address filtering in addition to password protection. AAI solutions for simplifying the authentication of researchers across various experimental sites are currently being investigated by EUROfusion and their usage may start to develop in the near future.

Once a researcher is authorized for a given experiment, he/she has access to **all** measured data and processed data (Plasma Reconstruction Chain, PRC) of that experiment. No experiment has implemented access rules that would depend on the type of collaboration or funding under which a particular set of pulses would have been produced. Data has some degree of FAIRness at the level of a given experiment, but EU experiments are presently not interoperable, which prevents exploitation of results of the EU fusion experiments at their full potential.

Formal Data Management Plans (DMPs) have not been established by any EU funded experiment yet, although some experiments (W-7X, MAST-U) have a formal Data Management Policy dealing with data access, sharing and usage in publication, aspects which are usually part of a Data Management Plan.

Even when they don't have a formal Data Management Policy in place, all experiments have established similar rules for using data in a publication, based on a formal publication clearance procedure.

Among the European experiments, only MAST-U has presently an active Open Data policy: by default a 3 years embargo is applied before public release of data, while "immediate" openness is applied for data related to a publication: "free access to all data behind published papers must be granted in a timely manner". However, currently this only applies to the one device, although TCV is in the process of developing an open data policy.

4.2 Data access and existing ontologies

The management and storage of generated raw and processed data is realised differently by each of the experiments. Often, to fulfil typical functionalities such as data storage, data access, data model documentation, cataloguing and browsing of metadata, the experiments use their own tools as well as their own ontologies. Although there have been some standardisation works initiated, there is still a lack of commonly accepted and implemented solutions. The current state in this area across the Fusion community is outlined below.

- Recent work on standardisation has been driven by ITER, the next generation of tokamak devices. With the support of EUROfusion and in the frame of the ITER Integrated Modelling and Analysis Suite (IMAS), a device-neutral ontology known as the IMAS Data Dictionary has been developed. While still not widely adopted as a native format, work has been ongoing into allowing access to

data using IMAS Data Dictionary naming conventions and providing mappings between local naming conventions and the Interface Data Structures (IDS), which are high level structured objects defined in the IMAS Data Dictionary.

- WEST made all its processed data and part of the measured data accessible via IMAS. Data access is mostly done via APIs allowing retrieving experimental data from various programming languages typically used at the experiment site (C, Fortran, Python, in some cases Matlab and IDL as well). The IMAS API uses similar principles, although it offers the possibility to access data at a broader granularity, namely at the level of the defined Interface Data Structures. These structured data objects contain potentially all information corresponding to an experimental subsystem such as a diagnostic, or a heating & current drive system. The IMAS ontology provides the possibility to store and easily access complete information about a subsystem (e.g. machine description, calibration coefficients, as well as the more usual raw and processed signals), while such information may be sometimes difficult to find in present experiment databases (if present at all in the database). As explained above, the WEST experiment already makes use of the IMAS ontology and access methods, thus exploiting the above feature.
- TCV is also using a similar approach, storing exhaustive information about experimental subsystems in structured MDS+ trees [13].
- In some experiments, a few different APIs must be used depending on the nature of the data, e.g. JPF (JET Pulse File) and PPF (Processed Pulse File) for respectively raw and processed data at JET. W-7X uses another system, namely a web-service based API serving data to users in JSON format. Data is uniquely addressed via a URL.
- Remote data access is often provided via the MDS+ technology used as a client/server architecture on top of the native database (AUG, TCV, JET). AUG also uses Andrew File System (AFS) for remote data access. The Unified Data Access (UDA) technology starts to spread outside UK to do the same thing (MAST, WEST and potentially ITER in conjunction with IMAS). This technology can be used stand-alone but has been coupled to IMAS to provide it with remote data access. On W-7X, no remote data access is allowed, one has to connect to W-7X using a VPN connection to carry out off-site analysis.

4.3 FAIRness of experimental and processed data

Present practices related to experimental and processed data with FAIR Principles are:

- Findable:
  - All experiments have a metadata catalogue with 0D/1D quantities (time traces) and tools to browse it and formulate queries.
  - However each experiment has its own tool, capable of finding only the data of that experiment.
  - There is no central metadata catalogue that would allow multi-machine searches, apart from the International Databases (as maintained by, for example, the various ITPA groups [14])
  - Although there is a community way of identifying data sets, this is not a persistent identifier
- Accessible (via authentication, so not open), for fusion researchers having an official link to an experiment, using access methods specific to that experiment
  - Currently data is stored at each site in indexed repositories and there is widespread adoption of the MDSplus open protocol.
  - There is a lack of common metadata standards and vocabularies
  - There is no common Authorization and Authentication System, meaning data access methods are site dependent
- Not Interoperable between various experiments because each one is using its own ontology (both for data and metadata)

- Reusable,
    - For fusion researchers having an official link to an experiment and being able to read provenance data and the experiment-specific data documentation.
    - Although sharing of data is done through Acceptable Use Policies, this is not the same as a licence
    - Although some provenance is captured it is often incomplete and not accessible to external users via the protocols available
    - A major limitation of reusability for some applications (e.g. synthetic diagnostics) is the fact that machine descriptions and calibration data are sometimes not recorded in the local experiment's database.
    - Researchers seldom seek experimental raw data. Instead, they use data processed by numerical codes with varying levels of complexity and interdependence. Lack of code version control and/or corresponding processed data versioning can considerably reduce data reusability.

Figure 3: Current status of the FAIRness in the EU fusion community

In summary, when considering a single experiment, its data has already some degree of FAIRness in the context of that experiment. But when considering the whole potential dataset coming from the various fusion experiments, the EU fusion community has no simple means to exploit it in a FAIR way. We can summarise the current status of the FAIRness in FIGURE 3.

For non-experimental data the situation is worse. In terms of volume, the largest contributor is data from modelling which is (mostly) produced by individuals or small groups without the support of dedicated computer professionals (in contrast to the experiments). This means that the data is (largely) not findable, accessible, interoperable or reusable. Code developers have expressed varying levels of support for remedying this, but will need support. The provision of a Long Term Simulation Storage Facility as recommended by the Gateway Expert Group and covered in D2.4 would significantly improve the FAIRness of the modelling data.

Other forms of non-experimental data (e.g engineering drawings, material databases) have not been examined by FAIR4fusion but will need to be considered by EUROfusion in their Data Management Plan.

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

4.4 ITER Simulation Database

In addition to the work presented within this document, ITER has been pursuing a development to support the cataloguing of data derived from simulations, named *SimDB*. While these simulations have been written out as Interface Data Structures, they have an associated yaml file which contains additional information which is not present in the IDS, and it is primarily this file that is used in the cataloguing. The primary use case is for each site to deploy one or more SimDB instances. Only metadata is ingested into the SimDB database; data remains on the underlying mounted file system and can be accessed directly. Rather than using a persistent identifier to identify data sets, a Universally Unique Identifier (UUID) is supplied instead. While the schema could be altered at a later date to add a PID (or replace the UUID), that is not currently planned. Figure 4 below shows a sample of the information which can be stored in SimDB.

```
[hollocj@sdcc-login02 simdb]$ simdb simulation info tutorial
uuid:      7b10aab0-7f83-11ec-b1ea-9440c9769e5c
alias:     tutorial
metadata:
  status: not validated
  ids: [core_profiles, core_sources, core_transport,
dataset_description, edge_profiles, edge_sources, edge_transport,
equilibrium, summary]
  summary.code.commit: 8a92fee6f0efd05cee1396a9997c976e3858a8a3
  summary.code.name: JINTRAC_JETTO
  summary.code.repository: /home/sim/jetto/source/v191219_imas
  summary.code.version: Release-v191219
  ...
inputs:
```

Figure 4: Sample of Information Stored in SimDB

Each simulation can be identified by either the UUID or an alias and some provenance information is supplied in the YAML, which can also be used to initiate simulation runs. A list of IDS contained in this simulation is also provided.

While primarily designed as a command line tool, a user interface is under development which is quite similar in look and feel to the demonstrator dashboard, allowing selection of various metadata elements by name or by range and allowing the viewing of a single signal or a comparison of the same signal from multiple simulations (figure 5) and allows the download of that data as a CSV file. However, the underlying technologies are quite different and the assumptions about where data will be stored are similarly different.

One of the most significant differences between the two concepts is how data is discovered at each site. The FAIR4Fusion demonstrator is built around the assumption that the data is held and managed at sites using existing site specific tools, and sites would then make metadata available to the dashboard for cataloguing. SimDB assumes it is run locally with direct access to the long term storage at each site. While currently under development, querying of remote sites behaves as a more peer-to-peer relationship within the SimDB architecture. In addition, currently SimDB does not provide a data access mechanism, relying on direct access to files through mounted file systems. Being built around simulations it is also quite specific around what metadata is provided such as arguments to simulations, location of input files, etc.

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

It is also important to note that the SimDB product has been optimised for simulations while the FAIR4Fusion dashboard has concentrated on experimental data, While both assume the the data is contained within Interface Data Structures, their contents are significantly different. Even something as simple as the shot/pass relationship is quite different; in experimental data each pass represents a new version of the data in a shot, while in simulations each pass may be equally valid and versioning needs to be handled differently.

The current project has been working with the SimDB development team to look for avenues of collaboration. Some work was undertaken to try and use the current dashboard as a front end to SimDB. However, the difference between the two architectures and the tight coupling of the demonstrator to the Summary IDS made this difficult. While there was cross fertilisation of ideas, a closer integration could not be possible. However, discussions are still progressing around supplying some information from SimDB to the FAIR4Fusion dashboard.

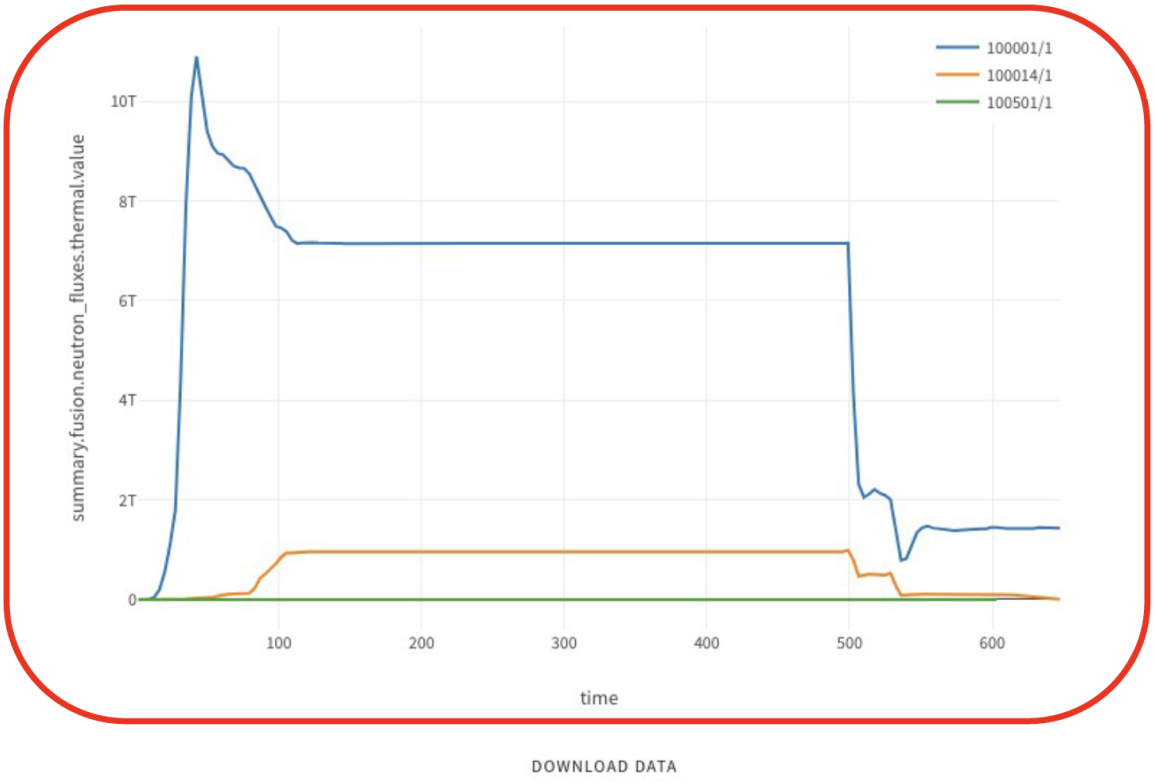


Figure 5: SimDB Prototype UI Visualisation

5. Requirements

The process of collection of the requirements for the system required intensive cooperation between all work packages and iterative fine-tuning. In this section we aim to present all finally identified requirements in a condensed and clear form. For those who need more detailed information we refer to the D3.1 [15].

5.1 Users and access levels

The target audience of the blueprint architecture proposed by the Fair4Fusion project will be a diversified community of users. Some of the users will come from the EUROfusion consortium and some will come from the associated projects or from the general public. Some will have a broad expertise about a particular experiment and will look for detailed information about shots generated in that environment and some will just look for an overview over all experiments. Ultimately we can also distinguish between human users interacting with the system in a classical way and automated tools that will take advantage of machine

readable data. It is hardly possible to precisely define all categories of the users. This leads to quite an extensive set of requirements and interface design decisions that need to be incorporated into Fusion Open Data Framework. Therefore our goal is to make the blueprint architecture generic and extendable.

As a starting point, we have identified six basic user categories that are the main target of Fusion Open Data Framework:

1. The general public
2. Funding agencies
3. External collaborators (defined as researchers not covered by EUROfusion agreements)
4. General EUROfusion collaborators
5. Internal (to the experiment) scientists
6. Data Provider/Manager

Non-fusion researchers would sit in category 1, 3, 4 or 5 depending on their relationship with the experiments or EUROfusion. These categories of users map to different access-levels to the data stored in the system. As examples,

- category 5 might have access to all of the data associated with their experiment, but only to a subset of the data available on other experiments
- category 4 would have access to all data whose collection was funded by EUROfusion
- category 3 might have access to very detailed data, but only after any embargo period has expired
- category 1 might have access to less detailed data after the expiry of any embargo period

In order to adjust the system views to specific categories of users and ensure its good ergonomics in accordance with particular permissions, preferences or expertise of users, the developed solution might need to be based on a multi-faceted logic that takes into account the following aspects as a minimum:

- information if a user is authenticated or not,
- user's category,
- user's expertise level.

The goal is to present the interface and data based on the cross-section of all collected information from this set. It means that the views should be different for each of the following example usage scenarios:

- Non-authenticated user with the basic expertise level
- Authenticated user of category 6 (Data provider) with the advanced expertise level
- Authenticated user of category 6 (Data provider) with the moderate expertise level

The exact access rights, and any limitations as to what level of data is to be provided, is likely to evolve as a result of interactions within this project, with the experiments, with the funding agencies and the development of attitudes to open-data, and could be clarified in the final version of the blueprint.

5.2 Leading user stories

We have collected a number of user stories about searching for and accessing data and/or metadata, as well as some of the wishes from the data providers. Those use cases present the different perspectives of members of the general public, EUROfusion researchers and data providers that are the main target users of analyzed scenarios. More details of these user stories can be found in the Fair4Fusion project deliverable D2.3 [16], but are summarized below.

The general public requests fall into two broad categories: queries that are motivated by recent press releases about breakthroughs in fusion research where a member of the general public might want to compare EU tokamaks with regard to the metrics presented in that publication; and queries that attempt to

ascertain whether the fusion devices are making progress towards the goal of energy production and are making good use of their resources.

Fusion researchers, whether from EUROfusion or internationally, tend to have specific queries about the data wanting, first, to find the relevant discharges that meet criteria they have in mind, and second, to then obtain the data they need for their analysis. In the D2.3 deliverable mentioned above some specific examples are presented for both of these. One example is “As a researcher, I want to compare the time traces of H-98 for different shots from different machines.” To provide some background, “H-98” is the ratio of the energy confinement time to that provided by a scaling derived from a large database of tokamak data, is a good measure for the “quality” of a plasma discharge, and is readily comparable across various devices. The comparison is expected to be in the form of a plot, with the ability to download the underlying data.

Additional input is supplied by the data providers: providing details of current access methods; expressing the desire to ensure that the data provision will not incur legal liabilities, excessive costs or impact the operation of the facilities; and expressing the desire for feedback about the use of the provided data.

5.3 Identified Client Interactions

Figure 6 below shows the client interactions needed based on the given requirements. Mapping these to the basic user categories identified in section 5.1, the *user* is equivalent to a member of the general public and external researchers who can perform basic searches on a limited set of physical parameters which are of most interest to the public, are able to perform simple plotting and are able to download Summary IDS information but who may have more limited access to more detailed data dependent on site policies. Fusion workflows, including those run on specific machines or making use of AI/ML technologies are also represented as clients of the system.

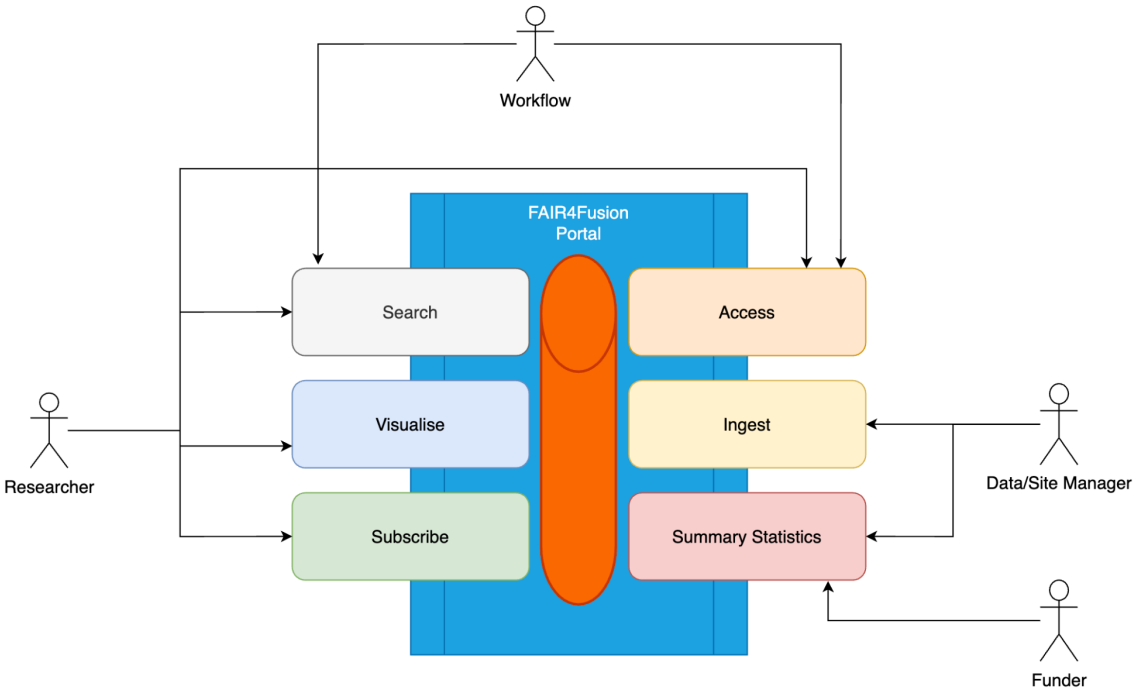


Figure 6: User Interactions with Proposed System based on Requirements

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

5.3 Required Functionalities

In order to develop the Fair4Fusion blueprint architecture the basic requirements and user stories have been transformed into a list of more technically informative functionalities and grouped into eleven sections (F1 - F11) as presented below.

- **F1 Search**
  - Free text search on an entire set of stored metadata or/and created indexes
    - Define vocabulary type searches - using controlled vocabulary
    - Optional semantification of the data
  - Faceted search over a set of predefined parameters supporting complex aggregated search queries
  - Ranged queries over continuous parameters
    - Support for defining time-spans and ranges
    - Possibility to query for ranged parameters (including time) stored inside a shot
- **F2 Visualisation and accessing outputs**
  - F2.1 Preparation of metadata and data for Visualisation
    - Gathering data from one or many experiments
    - Export of the data to common formats (plugins for transformation)
    - Request for more data of the shot that was found
  - F2.2 Visualisation in Portal
    - Plot 1D using metadata, Plot 2D, etc through data access
    - Plugins that can render this data,
    - Compare data from single experiment
    - Compare data from multiple experiments
  - F2.3 Download of data from experiments based on search results
    - CSV file with basic fields
    - Get plots results in different formats e.g. png/jpeg
    - Download the data related
- **F3 Report generation (output metadata resulting from the Search)**
  - Selection of parameters/statistics to include (e.g. output fields)
  - Support various formats
  - Customising output format where applicable
  - Sorting results
- **F4 User Annotation Curation Management**
  - Ability to associate annotations with experimental metadata
  - Public and private annotation metadata scopes, at different granularity levels
  - (Semi) automatic metadata enrichment, including capability to carry out text mining and/or natural language processing (NLP)
  - Diagnostic annotations from experiments and quality assessment of experiments/shots (description) based on available metadata coming from users
  - Development of the fusion controlled vocabulary (tags in Summary IDS) or ontology
- **F5 Metadata Management**
  - Internally derived metadata, IDS summary, Other data from experiments, not in IDS, Associations of post-harvesting metadata (linked in most cases to Provenance) and associations between related resources
  - Interface for metadata specification and management about different resources involved in experiment

  - Metadata information about publications, devices, scientists, etc. associated to, e.g., discharge/experiment
  - Categories (topics) - scientific justifications for campaigns
  - List of possible extensions dependent on GDPR (we might limit the exposed information depending on the cases)
  - Aggregation of metadata from associated resources, enabling their access through a single information unit
  - PID management
  - Discharge success assessment and reliability information based on pre-defined criteria matched with available metadata
  - Frequency of updates - keeping metadata consistent with experimental data
- **F6 Subscriptions and notifications**
  - Registering for updates on metadata
  - Various forms of notifications (e.g. email, XMPP)
- **F7 Versioning and provenance**
  - Capturing provenance history of the metadata being provided
  - Capability to generate snapshots of experiment that can be shared/cited
  - Towards distributed provenance, provenance chain: capability to keep track of derived/new lines of work (what publications came from the data downloads, maintain provenance, include initial provenance)
  - Time span on which the dataset is Valid, trace version updates - some provenance
- **F8 Authentication**
  - Users might need to be authenticated
- **F9 Authorization/Access restrictions**
  - Different roles and granularity of access according to categories of users
  - Private, Group and Public levels of access
  - Taking account of local policies, e.g. embargo periods
- **F10 Accounting**
  - Ability to collect and present accounting information. Requested functionalities / queries depend on users needs, e.g.:
    - The number of user requests per specific collection
    - The size of data accessed per specific data collection or experiment
    - Who and when accessed particular data
- **F11 Licensing**
  - The data and metadata should be properly licensed
  - The license information should be clearly visible in Portal

5.4 Policies recommendations

Several policies recommendations for architecture have been identified (whole analysis in deliverable D2.1 [12], here we only focus on highlights). Towards a higher compliance with the FAIR and Open data principles the following policies and practices are recommended:

- Findable:
  - A central metadata catalogue should be accessible and searchable (through a Web Portal), gathering data from multiple experiments.
  - This system shall enable the creation of persistent identifiers both for data and metadata.
  - To make metadata catalogue open to the public without any embargo period
- Accessible:

- Provide a single method to access data across multiple experiments, open to the EU fusion researcher community (or restricted to the collaborators of the experiment) and after some embargo period make it accessible even to the public (in some simplified form).
- Make use of the IMAS Access Layer

- Interoperable between various experiments (both data and metadata) by using a standard ontology (IMAS).
  - This means mapping local ontologies to the IMAS data dictionary at some stage, before exposing it to users/public.
- Reusable,
  - by making the access to the experiment documentation more systematic (e.g. machine description) and more open to the public
  - Also by increasing (when needed) the amount of provenance information contained within the data.

6. Architecture

The general idea of Fusion Open Data Framework can be depicted in a way presented in Figure 7. As it can be seen, the Data Repositories, typically associated with fusion experiments, publish Metadata and Citation Data to the system, which collects them and exposes them to the Clients. Clients can Search over this data, request Experimental Data as well as add their own Annotations to the system.

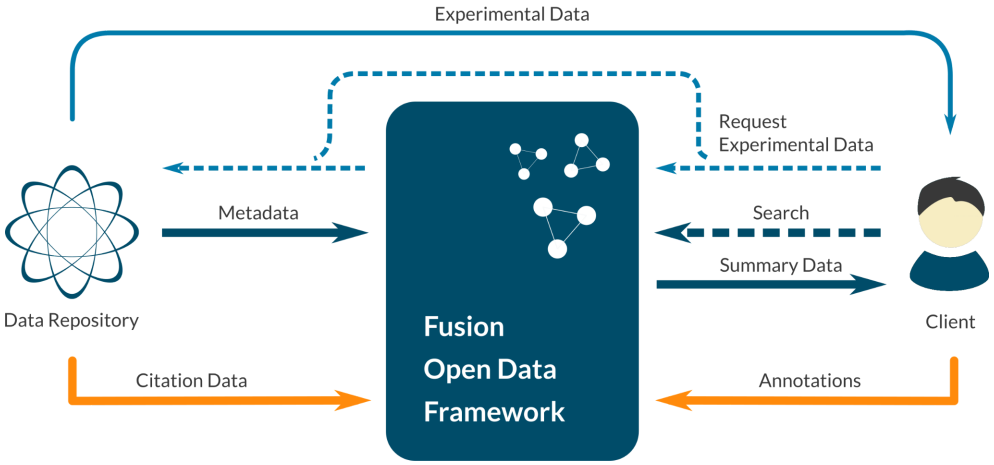


Figure 7: The general idea of the Fusion Open Data Framework

These are the main assumptions of the target system that have been used as a starting point for the architecture development. Then, based on collected requirements and motivations of the Fusion community, addressing open-data principles behind the FAIR requirements and utilising the outcomes of the technology survey conducted within task T3.2, we have managed to create an initial version of the blueprint for of FAIR service for European Funded fusion data, which is detailed in the rest of this section.

In order to make the concept easier to understand for the readers, firstly we will explain the baseline architectural assumptions based on the high-level diagram. Next, we will present a more complete picture of the system, with an extended view on the Fair4Fusion demonstrators with particular focus on the user facing tools. We will describe the role and functionality of particular components, the core relationships within the system as well as standards and protocols that are representative in the matter of the proposed architecture.

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

6.1 Baseline architecture

The high-level diagram of the Fair4Fusion blueprint architecture is presented in Figure 8.

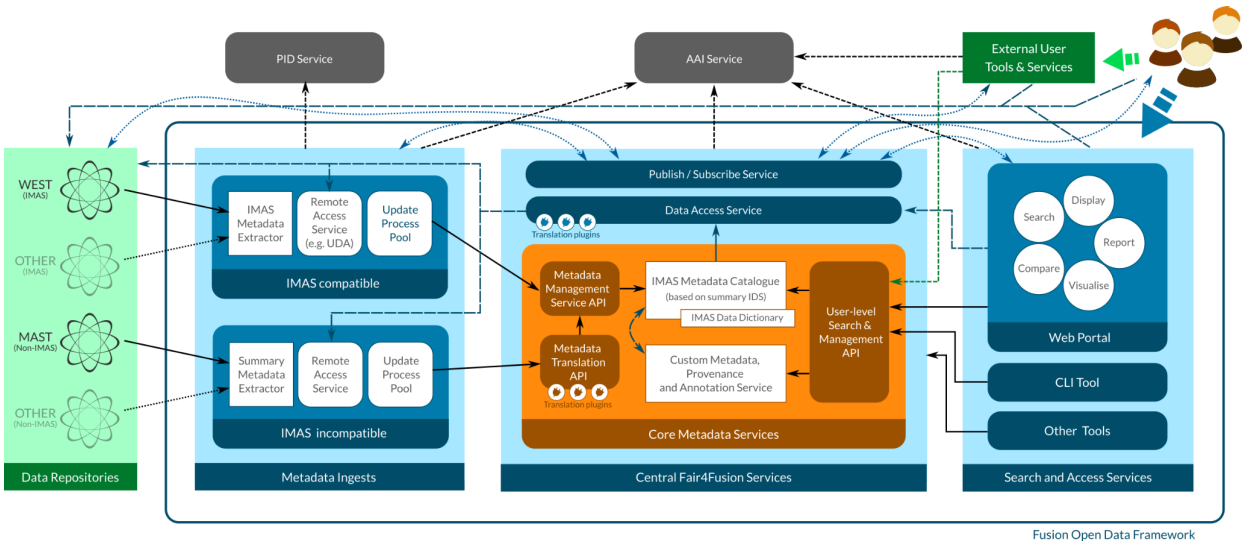


Figure 8: Architectural overview of the Fusion Open Data Framework system

On the most general architectural level, the whole Fair4Fusion environment can be divided into four main parts, i.e. Fusion Experiments, Metadata Ingests, Central Fair4Fusion Services and various User-level tools and services. While the last three parts constitute the integral content of the blueprint architecture being developed, the first part, i.e. Fusion Experiments, should be considered an external element. Below we outline the role of all these core parts as well as the role of a few supplementary components.

Data Repositories

As described in section 3 and 4, the Data Repositories are mainly, but not exclusively exposed by EUROfusion experiments spread over several European countries. For many years these experiments have been managed by different institutions as separate islands. This has led to creation of custom software and diversified data repositories that are not interoperable and can't be simply reused at scale. Furthermore, the data repositories associated with the experiments are governed by strict administration policies that lead to the practical impossibility of altering the technological environment of any of those. Therefore our only feasible decision aimed to bring together data from many data repositories is to treat them as much as possible as black-boxes and integrate them on a higher conceptual level. However we are encouraging experimental sites to enrich the data repositories, to be compliant with the FAIR best practices (e.g. to extend handling of the provenance data)

Metadata Ingests

The metadata from experiments are fed into the Fusion Open Data Framework system through the Metadata Ingests. The role of ingests is to transform metadata to the form which can be published to the Core Metadata Services and its users. What is crucial and should be stressed, technically the ingests are still placed in the administrative domains of the specific Data Repositories, which ensures confidentiality of the data until it is published. It means that all data that shouldn't be published (e.g. due to embargo period) can be witheld at this stage. Depending on the type of source data coming to the system, we can distinguish ingests operating on IMAS data and ingests operating on non-IMAS data.

Central Fair4Fusion Services

Central F4F Services are a basic service layer of the proposed system. A key role is played here by a set of software components marked as **Core Metadata Services**. The aim of these services is to collect metadata from diversified sources and provide users a unified way of searching and accessing this metadata.

The metadata coming from experiments will be the first and foremost type of data handled by the system. It will be stored in a homogenised form of IMAS format in the **IMAS Metadata Catalogue**. All experiments natively supporting IMAS will be able to directly use the **Metadata Management Service API** for pushing metadata. Other experiments that do not support IMAS, will need to use the **Metadata Translation API** and plugins that will automatically translate specific formats to IMAS.

In order to support FAIR open data and user-centric scenarios and separate it from the core experiment metadata management, the architecture proposes the **Custom Metadata and Annotation Service** as an additional unit. This service will be employed for the management of data pieces external to IMAS, such as references to publications, data lifecycle status (i.e. whether the data is valid or has been replaced by another dataset), and user annotations.

With the focus on usability, all the data managed by the Core Metadata Services is going to be exposed to the external world via a single endpoint which will implement the **User-level Search and Management API**.

In some collected usage scenarios, the clients of the system need to access not only metadata, but also experimental data stored at particular resources. Although implementation of this functionality is not considered as the core part of the system, we have analysed several possible ways of providing such a service. In the most basic scenario the data can be accessed practically without any interaction with the Fair4Fusion services, only based on previously generated PIDs. This way of accessing data will require many manual interactions and therefore it won't be very efficient. We argue that some assistance from the Fair4Fusion services is a better choice and thus we propose the **Data Access Service** as a moderator in accessing the physical data when a user wants to get it from the experiment.
The Central Fair4Fusion Services will be complemented by the **Publish / Subscribe Service**. Its role will be to enable asynchronous notification exchange across the system. Among other scenarios it will be employed to inform subscribed Core Metadata Services as well as users about updates made within the observed data collections.

Search and Access Services
The user's access to the Fusion Open Data Framework system will be enabled primarily through a set of dedicated software components grouped in Search and Access Services. It is expected that the **Fusion Open Data Framework Web Portal** will be the main entry-point to the system. With this component users will be able to search for various kinds of metadata, visualise discharges, compare shots, generate reports and so on, as well as define and manage custom annotations. It is expected that other types of client components, such as **Command Line** tools, will be developed or/and integrated with the API in the future.

External User Tools and Services
Since the data stored within the Fusion Open Data Framework system can be of value for further processing or more in-depth analysis, e.g. using Data Analytics Frameworks, the User-level Search and Management API will be accessible for external clients and services.

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

PID

In order to guarantee that data generated by experiments are unique and can be referenced during its whole lifetime, the system will utilise Persistent Identifiers technology, such as DOI or ePIC, to register the data globally.

AAI

The system will be complemented by a common Federated Authorization and Authentication Infrastructure(AAI) that follows the blueprint design proposed by the AARC project. The core service of the Federated AAI, is an AAI Proxy solution, managed centrally in the community, that decouples Identity Management from the services and the respective hosting sites. It should allow at least one of the supported protocols for enabling federated authentication like SAML, OIDC, OAuth2. Since many of the scientists come from universities and research institutes that are part of their national identity federation they should be able to authenticate using their home Identity Providers (IdPs) - institutional accounts to gain access to the services.

6.2 Architectural components

The architecture overview presented in the previous section can be moved into a more detailed description. In this section we are going to provide extended information about all components that have already been identified as an integral part of the Fusion Open Data Framework system.

6.2.1 Detailed architecture scheme

Before we start describing the particular components of the system, let us demonstrate the state of art diagram in Figure 9.

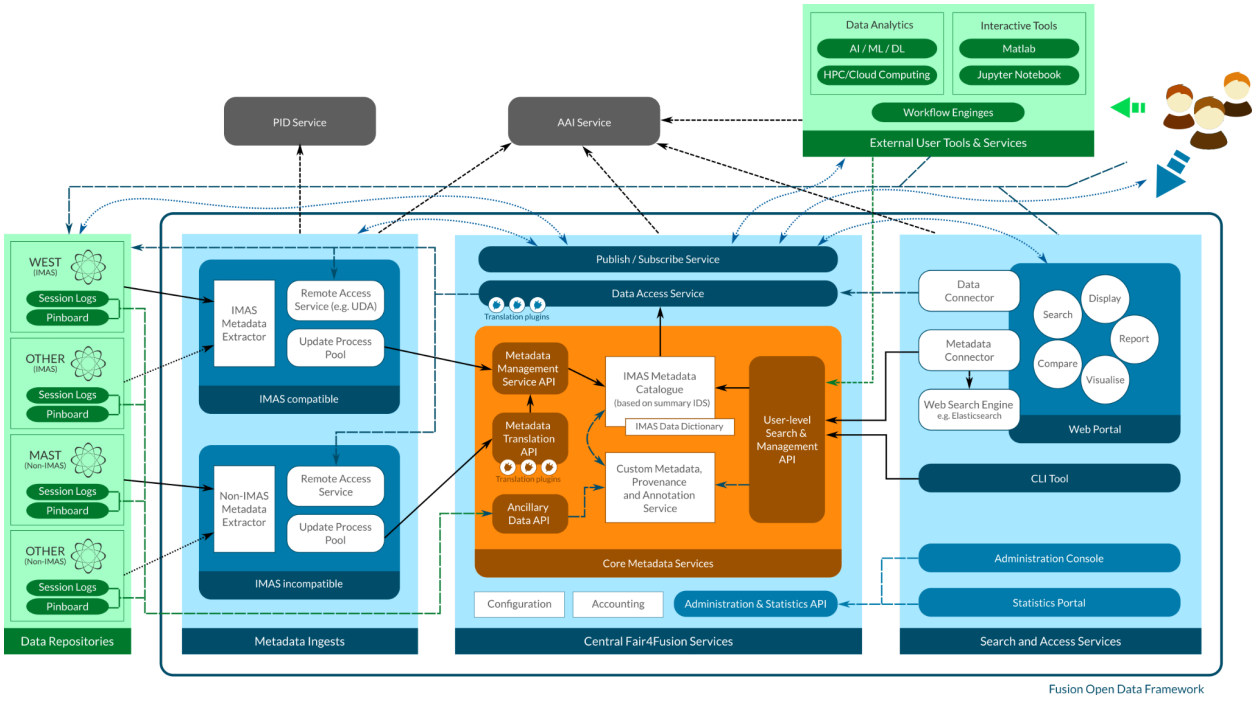


Figure 9: Architectural overview of the Fusion Open Data Framework with the detailed specification of individual components.

6.2.2 Data Repository side components

Data Repositories already using IMAS format

The metadata related to shots stored in these Data Repositories can be directly, without conversion, transferred to Central Metadata Services in the IMAS format. In this case the Data Repository has a local metadata catalogue similar to the central one.

As an example in the case of WEST, it's already populated directly by the intershot Plasma Reconstruction Chain (PRC), which generates a Summary IDS filled with a few time slices (corresponding to identified plateau phases of the pulse), then this Summary IDS is fed to the local metadata catalogue (a relational database that is a core element of the Catalogue Query Tool developed in EUROfusion). Two main strategies can be foreseen here:
1) a direct synchronisation between the local and central metadata stores;
2) the WEST PRC could be modified to also populate the central metadata catalogue. The former strategy appears safer, since it allows coping with possible local changes of the local metadata catalogue that would occur outside of the PRC.

Data Repositories not using IMAS format

The metadata from these repositories needs to be mapped to IMAS DD common ontology in order to be processed by Central Metadata Services.

6.2.3 Metadata Ingests

The Fair4Fusion Metadata Management services needs to ingest metadata from the Data Repositories. Because sites are mandated to maintain control of their data, and in order to minimise disruption to existing security and access processes, it has been agreed that sites will push metadata for data which is "open" (that is either open to the community or to the wider public) to the FAIR4Fusion portal. Sites can elect to push metadata from restricted datasets to the portal to advertise its existence, but the control of what appears in the portal remains at the site level. While this goes away from the OAI-PMH standard which works on a highly efficient pull model, the community and data providers in particular consider this will cause more disruption to existing processes and would be more likely to lead to accidental data leaks.

A further point for consideration is the format in which the metadata is passed. Currently there is no implementation for this at any sites, with each site being responsible for not only it's own metadata but also for any portal which uses this metadata for search and retrieval purposes. Currently WEST supports supplying data in IDS format natively, while for other sites there will need to be a mapping between site specific signals and IDS parameters. This metadata could come in the form of XML or JSON and be translated, or a future Fusion Open Data Framework implementation could supply each site the tools they need to convert the relevant metadata to IDS. While within the architectural diagram presented we have shown this as a service, it has not been finalised yet. The main issue is that the data from different sites might not necessarily be comparable, meaning that the mapping between site metadata and IDS definitions might not be straightforward.

For the moment we have identified 3 necessary components: Metadata Extractor - that is responsible for extracting/creating the metadata based on the data, Remote Access Service - providing access for the summary data in a pull mode, and Update Process Pool as a channel updating the metadata information in a push mode.

6.2.4 Fair4Fusion Core Metadata Services

The core services responsible for management of metadata, including the metadata available in IDS Summary, but also supplementary information such as references to publications or user-defined annotations.

IMAS Metadata Catalogue

The central service that integrates IMAS metadata (Summary IDS) coming from different Data Repositories. It is accessible with two APIs: Metadata Management API for population of metadata and Search API for integration with user-level clients. The presence of this service in the architecture is obligatory.

Custom Metadata, Provenance and Annotation Service

A supplementary service or, in an alternative implementation, a module of IMAS Metadata Catalogue for the management of data not available in Summary IDS. It allows for storing various kinds of information that is not present in IMAS Metadata Catalogue in explicit form nor can be easily inferred from the metadata present in that service. In particular this service can store information related to publications, provenance or workflows as well as various kinds of annotations specified by users after the initial metadata submission. The final functionality of this component will depend on the target scope of the Summary IDS and the functionality of both IMAS Metadata Catalogue and Portal.

A provenance service that takes as input the user's data set and extracts the relevant provenance data from it. Provenance capture from IDS data can be achieved by encouraging data providers to fill the relevant fields present in all IDS objects with machine-readable input. The python package, 'fusionprov' (available on the python package index, PyPi) [17] demonstrates a way of extracting relevant information from an IDS and collates it into a provenance document that complies with the W3C-PROV [18] standard. Data providers should be encouraged to also provide such provenance information in any other format that they publish/expose data. A demonstration of this for MAST and MAST-U data is also available in the 'fusionprov' python package, which takes as input a MAST signal and generates a provenance document by collecting the relevant information from various logs and files in the MAST data archive.

It would be up to sites to choose whether to implement and host a middleware service that generates and serves provenance data on-the-fly, or to build the curation of provenance information into their signal processing chain, storing the provenance as metadata. Data providers wishing to contribute a module for their site to the 'fusionprov' package are welcome to do so, provided that any external dependencies for accessing the data are kept to a minimum.

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

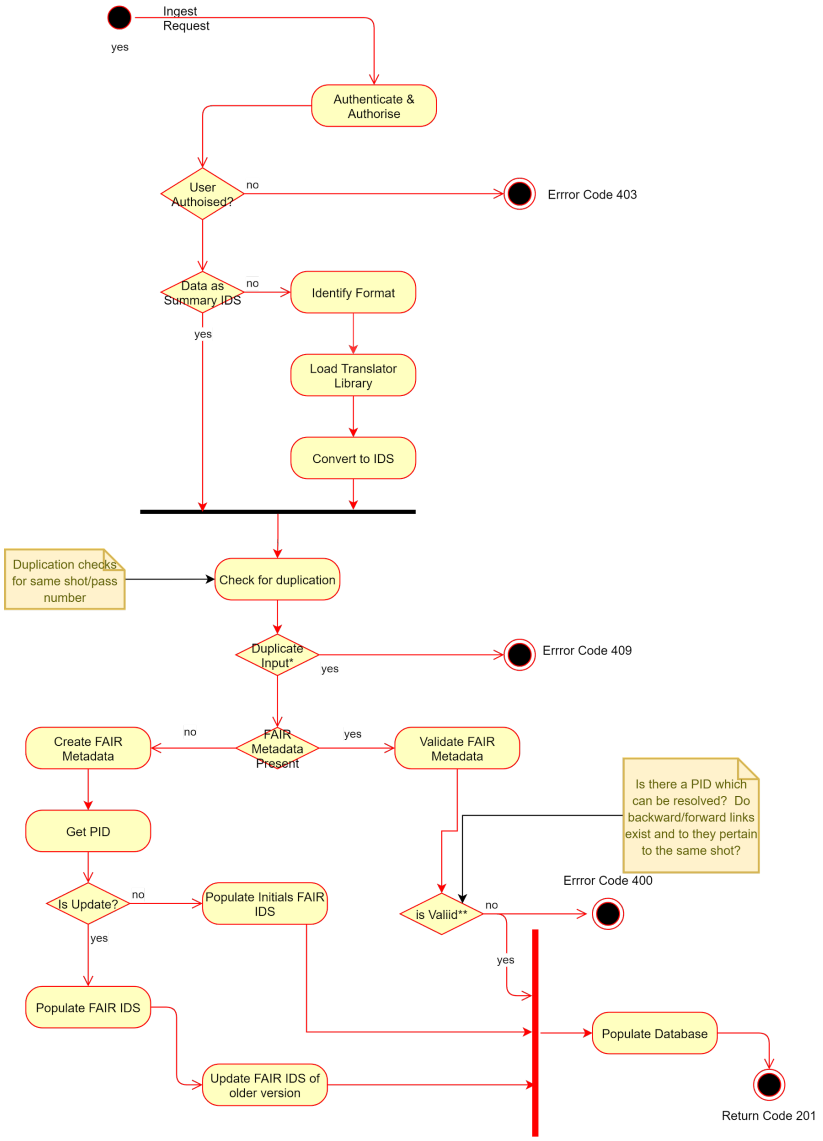


Figure 10: Activity diagram for the conversion of custom metadata into the IMAS format

Metadata Management API

The main access point for metadata produced by experiments. Since the Central Metadata services internally store data in the IMAS data structures, the experiments using IMAS format could use this API to publish new metadata to the system as well as to update existing one, it would in addition require a synchronisation mechanism between local and remote IMAS repositories. All experiments that don't generate IMAS metadata need to use Metadata Translation API that will perform a conversion from custom metadata format into the IMAS format. This is shown in Figure 10.

In this, it is assumed that only authorised data generators are able to upload information, such as an experiment representative or local data manager. As noted, the process for checking duplication relies on the existing experiment/shot/pass identifier and if these are duplicated then the import is rejected. If replacement using the same identifier is required, existing information will need to be removed by the data owner. The workflow also allows for either the site to provide the FAIR IDS structure, or for it to be created on behalf of the site by the portal. Since the FAIR IDS is unique in being mutable, many sites might initially prefer the creation to be done externally.

Metadata Translation API & Translators

This component will play a role of converter from non-IMAS formats produced by certain experiments to IMAS format natively supported by the system. Each data providing site will be decoupled from the schema and technology used by the Central Metadata services - it will only need to make use of the API.

This API will make use of one, or more, translator modules with a shot-summary metadata object in a site specific format *A* and will generate its IMAS-summary equivalent *B*. *B* will then be indexable by Central Metadata Services and subsequently searchable via its graphical user interface. Translators will be software modules that will implement translation of a site-specific shot-summary into the commonly agreed IMAS format. They will make use of existing IMAS technology as well as the corresponding Data Dictionary and related schemas.

User-level Search & Management API

An extensive API for querying the system from user-level tools, i.e. Web Interface and possible CLI. This RESTful API will provide calls for indexing requests to the search engine, as well as for evaluating user queries initiated primarily on the graphical user interface. The parameters, and therefore functionality, of the indexing and searching functionality will depend on the details of the user stories chosen to be implemented as part of the Central Metadata Services. This API will also support data manipulation operations (e.g. add, update, delete), particularly on the data being in administration of Custom Metadata, Provenance and Annotation Service. As example, in Figure 11, we demonstrate two activity diagrams outlining the procedures of creation and update of annotations by users, which are exposed via the API.

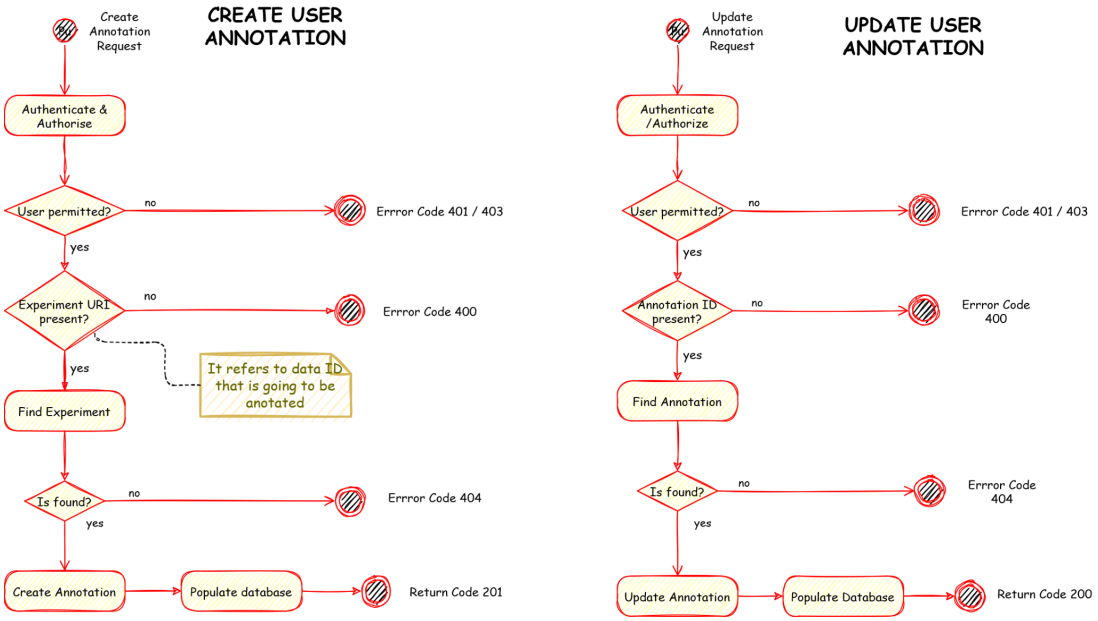


Figure 11: Activity diagrams for the creation and update of a users' annotation operations which are implemented within User-level Search and Management API

Ancillary Data API

In this context ancillary data represents data which is not primary data (primary being for example experimental or modelling data). Examples could be published papers, machine configuration information or diagnostic calibration information. In these examples, ancillary data represent either secondary products or required information which was used during the processing and analysis of data, but which was not obtained directly from the primary experiment.

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

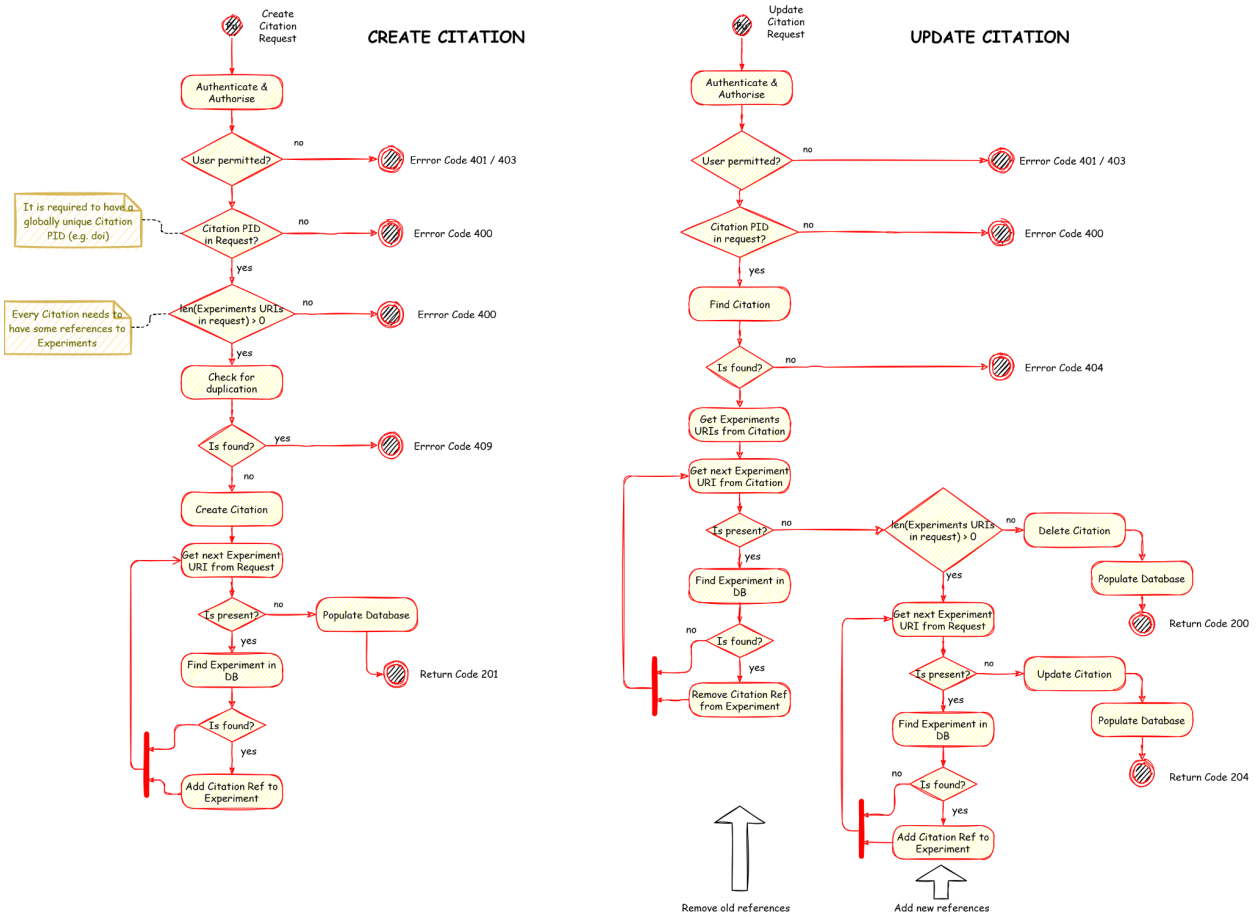


Figure 12: Activity diagrams for the creation and update of a citation operations which are implemented within Ancillary Data API

In order to meet the requirements, particularly related to funders and local site administrators, there is a requirement to allow access to information related to data referenced within publications and also for provenance information to be queryable and retrievable, and these are the types of information which is referred to in this document as *ancillary data*. For provenance information, most sites typically hold this information in session logs with a combination of automated, semi-automated or manual data capture of information related to both the pulse and diagnostic configurations and the processing chain converting raw data to physically meaningful parameters. Most sites also make use of a 'pinboard' mechanism or similar or journal clearance but there is currently no specific information about which datasets have been used for publication. However for a paper using experimental data there is need for some clearance form the relevant research officer.

In order to ensure a sufficient level of privacy and give a full control over the data items that will be exposed to the Fair4Fusion, the procedure of acquiring ancillary data has been based on a push model and a dedicated Ancillary Data API that can be accessed remotely by sites. In this way, a site can decide its own what and when to publish. Baseline activity diagrams for the citation creation and update, that demonstrate typical usage scenarios of Ancillary Data API, are presented in Figure 12.

6.2.5 Fair4Fusion Central Services

Data Access Service

Once a particular dataset has been selected as a result of queries on metadata, this service will enable automated client access to the corresponding data. This service is the final gate for users of the F4F services to the original data that is referenced (by metadata) in the F4F portal. This data will not reside

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

within the F4F services but at particular sites (e.g. experiment sites), therefore remote connections will have to be open to transfer data on the fly (upon user request). Of course, depending on local policy, the requested data will not necessarily be open and accessible directly via the F4F portal. In such a case, the minimal requirement is that the data access service will provide the instructions for accessing the particular dataset, assuming the user has the credentials to access the needed resources (e.g. the cluster of the targeted experiment). A more convenient solution (still in case of non-open data) would be to embed an authentication mechanism in the data access service, so that the user of the F4F service can authenticate themselves towards the data server.

Data access needs to be separated into two distinct workflows; authorised data access by a member of the community and public data access. Authorised users can gain access to full resolution experimental data through the existing UDA or MDSplus protocol. This requires HTTP redirection based on the existing standards and is supported by both of the mentioned protocols. Note that currently this represents accessing a single IDS from a single dataset; bulk access has not been considered since neither is currently supported by the underlying protocols. The exemplary activity diagram for this workflow is shown in Figure 13.

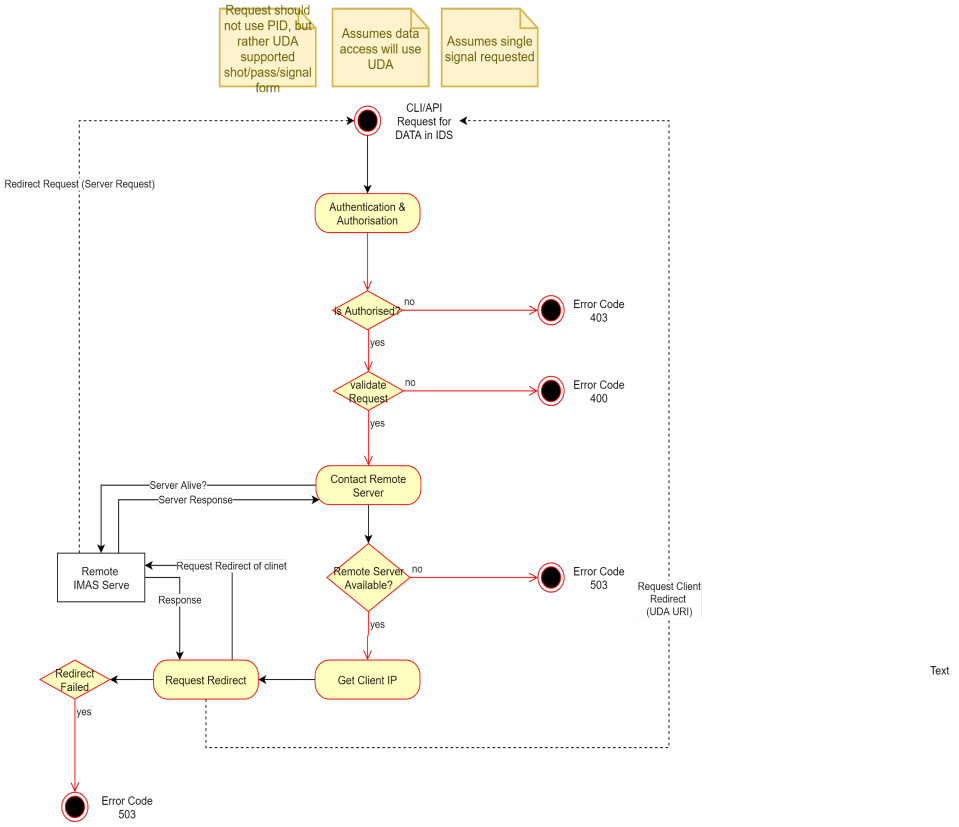


Figure 13: Authorised Data Access Workflow

Furhtermoredata needs not only to be fetched from a remote site, but also to be mapped on the fly to the IMAS format, so that it appears in the same data structures / ontology whatever the particular experiment it comes from. Not all experiments (or simulation stores) will provide the same coverage in terms of IMAS format mapping: it's expected that some experiments will start by mapping data from a few diagnostics and progressively extend the perimeter of these mapping to other sub-systems of the experiment. Therefore the tool will indicate to the user what IMAS data objects (Interface Data Structures in IMAS terminology) are available for a given experiment.

Data access and manipulation by the user must be flexible, therefore a command line type interface (e.g. Jupyter notebook) will be available to the user to type/formulate access and data processing or visualisation commands, using the regular IMAS Access Layer API.

This service thus closes the loop for the researcher who, after having searched for datasets of interest via the querying services, can finally access and manipulate the original data corresponding to the selected datasets.

Publish Subscribe Service

The role of this service will be provisioning of system-wide asynchronous communication between Fair4Fusion components. In particular, the service will be used for distribution of notifications about changes in specific data collections. The service will support registration of notification consumers being system's users, but also software components.

Configuration, Accounting and Administration & Statistics API

Configuration and Accounting are approximate names for all components that will assist in regular administrative tasks, such as configuration of Fair4Fusion services or collection and management of accounting information. It is expected that administrators of the system will be able to use Administration & Statistics API to access these elements of the system.

6.2.6 Search and Access Services

The set of user-level tools for accessing the system. The Web Interface access needs to be provided. In addition the system includes the command-line interface (CLI) and REST APIs for the machine.

However, there may be interest from outside the community to gain access to some information, for instance in higher education as a learning aid, or from citizen scientists wanting to understand more about the data generated by fusion devices, or even commercial organisations. Since we recommend the use of a Non-Commercial license (subsection 8.1) for full data, we propose to allow information to be made available in standard formats such as CSV, JSON or XML from the portal directly or via a command line interface. This workflow is shown in Figure 14 below.

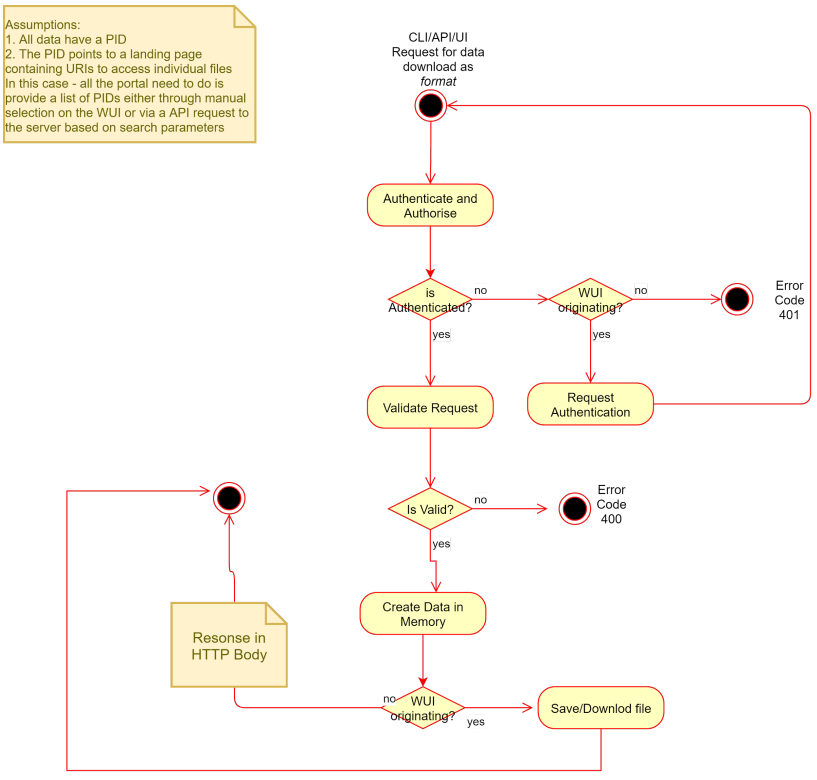


Figure 14. Access Workflow for publicly available data

Note that authentication is still required to allow data providers to understand who has accessed data. The precise mechanism will depend on the AAI implementation, but it is anticipated that the user of social media or EDUGain identifiers will suffice. In this case, additionally there is different behaviour in the case of trying to download from the WebUI or through the command line; if using the CLI then it will be necessary to generate an access token (not shown) before use.

Web Portal

The primary access tool for the users. It will allow for searching, filtering, displaying, comparison and management of metadata. Based on searching it will also allow for retrieval of data (download) directly from experiments. The Web Portal will be implemented in modern software development technologies with a split between front-end and back-end. The front-end will be responsible for the graphical presentation and user-interactions, while the back-end will perform tasks related to data access and interactions with other services. Conceptually, the latter will consist of the three main elements: Metadata Connector, Data Connector and Web Search Engine.

Metadata Connector
This first element of the backend of Web Portal will be responsible for accessing Fair4Fusion services in order to resolve queries and retrieve metadata.

Data Connector
This backend component will integrate with Data services in order to enable access to experimental data stored on individual sites. The usage of this component in a system is optional, but can streamline data access.

Web Search Engine
In order to improve the performance of the system Web Portal will use a state-of-the-art Web Search technology such as elasticsearch or memcached. With this component, the data accessed frequently via Metadata Connector will be indexed and/or cached for further usage.

CLI Tool

Command-line interfaces remain a preferred way of accessing software systems for many scenarios and users, particularly for integration into workflows. To support this way of interaction with the Fair4Fusion system, a dedicated tool needs to be provided to the Fusion community as an alternative to the basic part of the functionality offered by Web Portal.

Administrative Console & Statistics Portal

Client application for the administration and statistics services for configuration of Fair4Fusion services or collection and management of accounting information, as well as different statistical information as requested in several user stories requirements.

6.2.7 External User Tools and Services

Workflow Engines
Several workflow systems are used within the fusion community including tools such as Kepler, MUSCLE2, MUSCLE3, but also shell scripts/Python workflows. While those tools are not part of the Fusion Open Data Framework, they should be considered as full-fledged clients of the Central Services, since searching and retrieving the data, as well as storing the results and relevant metadata are inherent part of the scientific workflows lifecycle.

Interactive Tools: Matlab, Jupyter Notebook
Several applications and tools like Matlab, Jupyter Notebook that support the scientists in their research are another example of the clients for Central Fair4Fusion Services, searching and retrieving the data. Those tools are not part of the Fusion Open Data Framework.

Data Analysing Frameworks - Feature Extraction and Data Mining
A Fusion Open Data Framework implementation could also provide an interface where data mining across the different tokamaks could occur. A typical pipeline for this would be

- Search across all the machines of interest for shots meeting the desired criteria
- For each shot found
    - Find suitable time-points in the shot
    - For each such time-point
        - Gather the data that is needed
        - Extract the desired feature(s)
        - Store these features in some form

Artificial Intelligence, Machine Learning, Deep Learning, HPC and Cloud Processing
Fusion community is making usage of many available computational infrastructures, such as HPC systems in particular dedicated computing infrastructure for fusion - currently Marconi@CINECA, those provided by PRACE [19], EoCoE [20], and will plan to use the EuroHPC resources; besides it is making use of the cloud infrastructures like EOSC. Searching for the correct input data or feature extraction/data mining, storing the results of analysis and related metadata is a part of the process of processing scientific applications. Collected use cases assume the use of technologies like ML/DL/AI where the search interface for the data is an important feature.

6.2.8 Authentication and authorisation

The system will be complemented by common Federated Authorization Authentication Infrastructure based on latest technologies, following the AARC blueprint architecture [21], such as eduTeams [22] (and related technologies), enabling easy and safe integration between components. Using one of the supported protocols for enabling federated authentication (e.g. SAML, OIDC, OAuth2), users will be able to use one account and access all the services available to the whole community. Since most of the scientists come from universities and research institutes that are part of their national identity federation; through that, in eduGAIN, users will be able to authenticate using their institutional accounts to gain access to the services.

Latest EUROfusion efforts to establish EUROfusion AAI Proxy for fusion community in Europe can be leveraged but since the current focus is not on data/metadata access the AAI services should be extended to support the data access policies.

6.3 Technology candidates for the Fusion Open Data Framework components

In order to advocate the proposed architecture and justify its realisation, within this section we present a mapping of technological solutions readily applicable for implementation of Fusion Open Data Framework components. The presented mapping is a result of both the ongoing state of the art analysis aimed at juxtaposing existing solutions with the general Fusion Open Data Framework assumptions and the survey performed by the project to point out possible technologies for fulfilling defined Fusion Open Data Framework requirements. In regards to the former, we have already analysed several existing research

infrastructures handling large data sets like ICOS, wLCG, EOSC, EUROPEANA, CLARIN, IVOA. Based on the findings, the specificity of the Fusion community, i.e. decentralised experimental devices having their own data, procedures and software, brings the Fusion Open Data Framework system near the research infrastructures which have been built around existing data sets (e.g. EUROPEANA, CLARIN, IVOA). By analysing these infrastructures in the first place, we were able to learn not only from the technological choices, but also from the need to harmonise individual repositories to allow easier access to an increased range of community users. The separate extensive technology survey allowed us to compile a summary of proven technological solutions applicable for the core requirements of the project. For the detailed outcomes of this survey we refer to D3.1

The resulting mapping of technological candidates for the implementation of the Fusion Open Data Framework based on blueprint architecture is presented in Table 1. It should be noted that some of the components have been already incorporated into the Fair4Fusion Demonstrators [D5.2 Data Platform Implemented and Documented]. It is expected that further analysis will allow to fine-tune this set and evaluate applicability of the components for the target system.

Table 1. Mapping between F4Fcomponents and technologies

F4F component	Technology candidates
PID Service	DOI, ePIC
AAI Service	Eurofusion AAI; KeyCloak(for IdPs), eduTeam (for AAI Proxy) - internally using Perun, sUnity IDM, Perun (alternative technologies: EGI CheckIn, B2Access (based on Unity IDM), Indigo IAM)
IMAS MetadataCatalogue	noSQL and SQL database systems (e.g solution in community CatalogueQT)
Custom Metadata, Provenance and Annotation Service	Graph-databases / triple-stores: Virtuoso, GraphDB, Neo4J; Custom metadata databases, e.g.: ROHub
Metadata Management Service API	REST API
Metadata Translation API	REST API
Auxiliary Data API	REST API
User level Search & Management API	REST API
Web Portal Interface Frontend	ReactFX, Angular, AngularJS, jQuery, Bootstrap, Django
Web Portal Visualisation Modules	Kibana, Grafana, Tableau, Splunk, Cyclotron, matplotlib, plotly.js, seaborn, bokeh
Web Portal Backend	REST API (possible implementation in Python / Django, Node.js, JavaEE)
Web Search Engine	Lucerne, ElasticSearch, Solr
CLI Tool	Python, Bash, Perl, cURL, etc.

Publish/Subscribe Service	Redis, RabbitMQ, Apache Kafka, Dapr
Data Access Service	Fusion related technologies: UDA/MDS+(data access) EOSC ecosystem: OneData, EUDAT/B2SHARE (data sharing) CERN: Invenio, EOS, CS3MESH (sync & share mesh technology for federation of distributed on-premise sync&share system such as ownCloud, NextCloud, Seafile and Cubbit) Protocols: Amazon S3, POSIX, Network File systems (NFS, web based - WebDAV)
Administration and Statistics Services	REST API (possible implementation in Python / Django, Node.js, JavaEE)
Administration Console	ReactFX, Angular, AngularJS, jQuery, Bootstrap
Statistics Portal	ReactFX, Angular, AngularJS, jQuery, Bootstrap Kibana, Grafana, Tableau, Splunk, Cyclotron, matplotlib, seaborn, bokeh

6.4 Relationship between components and services

6.4.1 Metadata Conversion

As discussed in section 4.2, at the moment only WEST directly outputs part of its data and metadata in the IMAS format. Any metadata we get from the other experiments will have to be converted to IDS. (Semi-)automatic tools for facilitating the mapping of different standards to IDS are necessary, and will need to be developed.

6.4.2 Retrieving Metadata from Sites - Push vs Pull Models

There are careful considerations as to whether metadata should be pushed from a site to a central aggregator or pulled by an aggregator from the experiment site. The pull model, where the aggregator pulls information from the site hosting the data can make for a more reliable service since transient events can be better dealt with and accidental Denial of Service events between the aggregator and site can be controlled. However, it would potentially mean sites having to modify their existing metadata infrastructures in the case where data is a mix of commercially sensitive and more open data which it is unlikely sites would accept. The alternative, where sites push data to a central aggregator is also not without cost to the sites since this push service would become an additional production service which would need monitoring. However, it does give sites more freedom as to when metadata can be pushed to the central aggregator, doing this during the evening so as not to interfere with ongoing operations.
The choice between pull and push models also impacts the management of data updates: with push methods, remote sites are responsible for updating the central aggregator following, for instance, data reprocessing; with pull methods the central aggregator must either scan the remote data to search for updates, or be somehow notified of an update.

If the Universal Data Access layer of IMAS is used to gather data from sites before conversion to Summary IDS, this may be more easily done by making pull requests, while for data sets already adhering to the IMAS standards, either push or pull would be possible.

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

6.5 Standards and protocols

6.5.1 The Interface Data Structure

The IMAS Data Dictionary is one of the standards promoted by ITER. Within the IMAS Data Dictionary, some structures are marked as Interface Data Structure (IDS). An IDS is an entry point of the Data Dictionary that can be used as a single entity to be used by a user. Examples are the full description of a tokamak subsystem (diagnostic, heating system, …) or an abstract physical concept (equilibrium, set of core plasma profiles, wave propagation, …). This concept allows tracing of data provenance and allows a simple transfer of large numbers of variables between loosely or tightly coupled applications. The IDS thereby defines standardized interface points between IMAS physics components.

Although fully open to the fusion scientific community of the ITER members, the IMAS Data Dictionary is presently not open to the general public. After discussions at the working level with our colleagues from the ITER Organization, there should be no obstacle for making the IMAS Data Dictionary open source. Therefore we recommend that EUROfusion or the European Commission requests this from the ITER Organization in the near future. It would be interesting to push at the same time for making other components of the IMAS core infrastructure open source as well (e.g. the Access Layer), although it's not a requirement for making EU data open with the IMAS technologies.

6.5.2 IDS Summary Metadata

Within the IMAS Data Dictionary, the Summary IDS is the placeholder for physical metadata summarizing an experiment or a simulation. It contains time traces of several global, local or space-averaged physical quantities that physicists typically use to search plasma experiments of interest. In addition to the value of each quantity, there are also placeholders for error bars and provenance information (a simple string so far). Being defined in a machine-generic way and usable for both experiments and simulations, we propose to use this ontology as the standard for metadata for making European fusion experiments and simulation data FAIR.

An informal study was carried out to see how the individual experiments allowed users to search through their metadata. A total of four experiments were surveyed (WEST, JET, MAST-U and ASDEX Upgrade) and each term mapped onto the Summary IDS.

Each experiment's searchable metadata mainly focused on the physics summary parameters such as the plasma current for a shot and there was little focus on more generic metadata. This meant the study soon morphed into a comparison of these physical parameters. A common set of these terms (which were made searchable by each experiment) was then formulated although there was no guarantee that the values were measured in the same way. Continuing the plasma current example this can be taken when the shot is in the flat top phase but it is likely that each experiment has subtly different definitions of this. In fact the method of measurement may not even be the same. This is not an issue though, since information on how the data was obtained can be added in the "source" node attached to each "value" node in the Summary IDS.

6.5.3 Extending IDSs with more FAIR information

The Summary IDS provides a large coverage of the physics quantities that can be captured in fusion experiments but does not contain more generic documentation that will help make the data more findable and accessible to non-fusion users, including funders, other researchers and the general public. We have thus decided to extend the Data Dictionary with additional FAIR information. A dedicated Dataset_fair IDS has been created as a placeholder for FAIR metadata that is not immutable but will evolve during the lifetime of the dataset, such as validity of the dataset, licensing, references attached to the dataset.
Based on the requirements we have selected a number of Dublin Core Elements to put in this new IDS. Dublin Core have curated a list of generic metadata terms known as DCMI Metadata Terms (superseded qualified Dublin Core in 2008) based on the smaller Dublin Core Metadata Element Set (DCMES). Whilst, DCMI only has two compulsory terms it is understood that by using the generic terms provided by DCMI we will improve the interoperability of the metadata schema with other schemas. As a generic schema not all DCMI terms apply to fusion but by comparing the DCMI terms and the Dataset_fair IDS a subset of DCMI can be selected to improve the FAIRness of the proposed fusion metadata schema.

In addition, we have extended the ids_properties structure of all IDSs with a new structure to record the provenance of the data stored in the IDS. This structure allows choosing the granularity of the provenance information recorded, from the global IDS level to substructures or even leaf level. With this extension, the provenance can be documented by Data Processing Chains directly in the IDSs they produce. The whole provenance scheme for any IDS data node can then be reconstructed recursively. This was missing before since provenance information was only captured at the level of Summary IDS.

7. Evaluation of technologies

In addition to the architecture of Fusion Open Data Framework described in the previous section, the other essential goal of the project is recognition and evaluation of technologies suitable for the implementation of the final system. Two system demonstrators developed within the project to practically evaluate promising technological solutions are described. The final part of this section contains a discussion about discovered gaps and lessons learned during the process of development of the system's architecture and prototypes..

7.1 Methodology

In order to discover and evaluate technologies needed to fulfil all requirements defined to build a fully functional system for FAIR-compliant management of data produced by the Fusion community, a multi-step process has been proposed. Its general overview is presented in Figure 15. On a high level the flow can be seen as quite typical as it comes down to the collection of use cases, definition of requirements, implementation of prototype and analysis; however a few elements need to be explained in more detail. First of all, it was particularly important to ensure that the requirements defined by the fusion researchers are well-defined and understood. Taking into account the complexity of the fusion community, this part of the process has been split into two phases. Firstly, the user stories produced by the researchers have been merged with the general FAIR requirements and then translated with assistance of IT specialists into an extensive set of purely technical requirements. By this splitting, the process of collection requirements has been made more organised and relatively smooth.

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

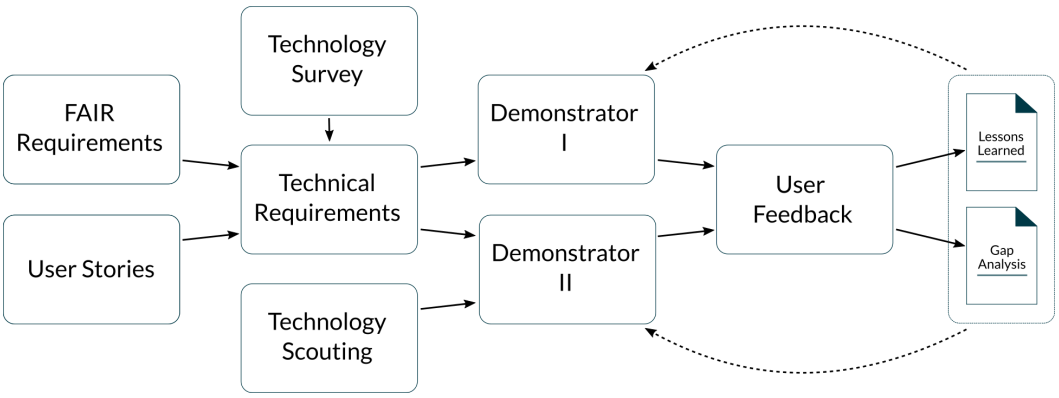


Figure 15: Technology evaluation strategy in the Fair4Fusion project

The next specificity of the employed workflow is the development of two demonstrators. This topic will be discussed in more detail in the next section, but a few aspects need to be outlined. As it is presented in Figure 15, both demonstrators respond to Technical Requirements collected on a basis of User Stories and generic Technology Survey, but only *Demonstrator II* depends on the Technology Scouting task. The general idea is here to ground *Demonstrator I* in well recognized software components already available in the fusion community (e.g. IMAS data structures) and allow *Demonstrator II* to test external technologies that could be potentially useful for the target system but may require too much effort to be integrated with the existing Fusion solutions in the scope of the project.
Last but not least, it is important to underline that the analysis - led again by fusion researchers - is not a simple one-shot task but rather a continuous process bi-directionally coupled with new developments in *Demonstrator I*. It allowed us to iteratively update results of analysis throughout the project's duration, thus it plays a significant role in fine-tuning of the blueprint architecture.

7.2 Demonstrators

The FAIR4Fusion provides two demonstrators: The role of Demonstrator I is to provide basic functionality in a production environment for real data, while the role of Demonstrator II is to evaluate new solutions that may be useful in the longer term Fusion Open Data Framework Implementation. Although demonstrator roles are different, they share common goals, e.g. to be usable and respond to basic requirements of the Fusion community and to be useful for developing new ideas and validating existing approaches. The two demonstrators are summarised below:

7.2.1 Demonstrator I

The focus of Demonstrator I is on reusing as much of the present Fusion technological ecosystem as possible. The main assumption of this solution are as follows:

- Metadata ingestion from various sites via IMAS Access layer
- Data access from IMAS based data sources (MDSPlus files, UDA client) and from various locations (local data files, remote data)
- Using the demonstrator as a testbed for de-facto standards, such as the Summary IDS and Dataset FAIR (a new, dedicated IDS for FAIR project) defined as part of the ITER Physics Data Model (PDM)
- Modularity of the solution
- Utility both inside Gateway and outside the Gateway (based on Docker)

Implementation

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

Demonstrator I is based on loosely coupled components that are responsible for reading, storing and presenting data to the user as shown in Figure 16, in particular:

- Catalog QT Database - stores metadata based on IMAS Data Dictionary
- Catalog QT Web Services - provide clients with metadata available inside Catalog QT Database
- Update Process - reads metadata from IMAS based pulse files (summary IDS) - also via UDA - and stores it inside Catalog QT Database
- Demonstrator Dashboard - the main user interface to the system. Initiall idea of dashboard was inspired by JET Dashboard, however, final solution was developed using React JS technology in a Single-page application model. Data access is realised using Catalog QT Web Services. It allows users to: filter, browse and compare metadata stored inside Catalog QT.

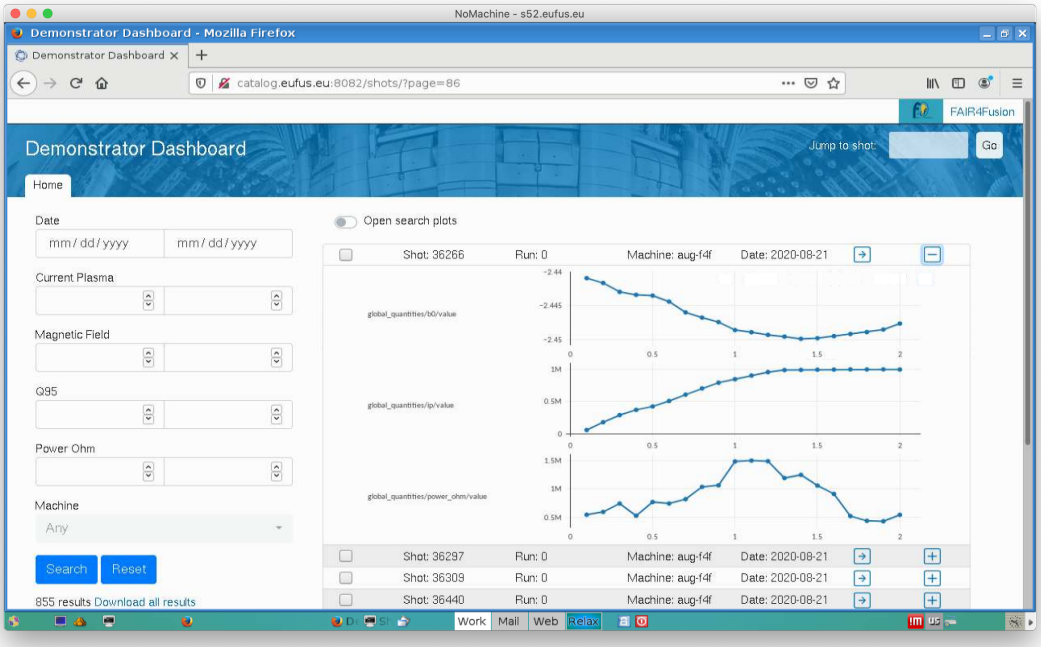
Each component is an independent piece of software so it can be flexibly deployed on one or many machines according to particular needs. For the user's convenience, a solution where all the components are integrated inside a single Docker container is provided at GitHub [23]

Demonstrator I has been also pre deployed on the EUROfusion Gateway environment where it can be easily tested in a realistic environment. Detailed description of how to setup, run and test the Docker based installation is provided directly inside the repository [24]

Functionality

The following functionalities are already available with Demonstrator I:

- Browsing all the entries stored inside Catalog QT.
- Filtering over machine (resource)
- Filtering entries by specifying values (minimum and maximum) of variables. At the moment the following variables are supported: Plasma Current, Magnetic Field, Q95 and Power [Ohm].
- Dedicated view for presentation of all terms stored in a Summary IDS, with a possibility to extend it with other data available in IDS
- Different presentation forms of data: pure text, lists / tables and charts
- Comparing variables of multiple entries (from one or many experiments) by displaying graphs
- Data ingestion based on IMAS Access Layer. Possibility to access data stored locally (MDSPlus files) and remotely (UDA plugins). Design allowing for other data sources to be easily handled in the future.



This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

Figure 16. Example screenshot from Demonstrator I: a view presenting list of pulses with visible filtering pane and charts for basic quantities of selected shot: b0, ip and power_ohm

7.2.2 Demonstrator II

Demonstrator II is focused on exploring alternative and additional technologies that will be required or may improve usability of the Fusion Open Data Framework once it is released. Since Demonstrator II is not tied to a concrete set of technologies, but rather its idea encourages to explore new possibilities, the implementation has been started from scratch based on generic and popular solutions. The implementation comprises a backend that executes computational experiments and a frontend for visualizing shots and experiment results, exploring the following elements:

- Alternative metadata/Summary IDS representations
- New solutions for visualisation and analytics, e.g. Python/bokeh
- The use of search technologies, e.g. Solr or ElasticSearch
- Open science paradigms, including PID management, provenance, publication linkages
- User defined annotations
- User management, authentication & authorization, e.g. Keycloak
- Containerization technologies, especially in the context of FAIR sharing of complete computational experiments, besides experiment data
- Cloud computing for distributed execution of computational experiments

Backend

Containerization is a highly flexible way to publish specific, reproducible execution environments and deploy software on modern computational infrastructures. Most e-Science infrastructures, notably including the EUROfusion Gateway [25], have endorsed it as a medium for packaging pre-built, pre-configured, and ready to execute software in a way that allows automatic deployment on Cloud-computing infrastructures. In our prototype, computational experiments are defined as compositions of containers, where each container provides an elementary tool. As an example, Figure 17 gives as an example the definition of a pipeline that transforms data from the format they are stored in into the format expected by the second step, then executes a tool that evaluates a similarity metric, and then selects the most similar shots based on this metric. The pipeline definition refers to specific images served by an image registry, so the pipeline can be reproduced and yield identical results on different installations of the system, relying on container technology for portable software packaging.

```
metadata:
  name: experiment-ip-similarity
spec:
  comparisonType: ip
  work:
  -    image: registry.gitlab.com/fair4fusion/pipelines/dtw-transform
       name: transform
  -    image: registry.gitlab.com/fair-for-fusion/tools/dtw
       name: dtw
  -    image: registry.gitlab.com/fair-for-fusion/tools/dtw_collect_results
       name: collect-results
  -    image: registry.gitlab.com/fair-for-fusion/tools/filter_similar
       name: filter-similar
       workLabel: reduce
```

Figure 17: Listing of an example pipeline definition.

The implementation also experiments with scalability in Cloud computing (Figure 18), and we have developed a custom Kubernetes *operator* [26] for deploying pipelines defined using yaml notation. The operator distributes the processing among the computation nodes that are available, but is aware (via the

relevant label in the yaml) of *reduction* steps where all intermediate results need to be collected; in our example computing distances can be parallelized but finding the five most-similar shots needs to first have all distances collected.

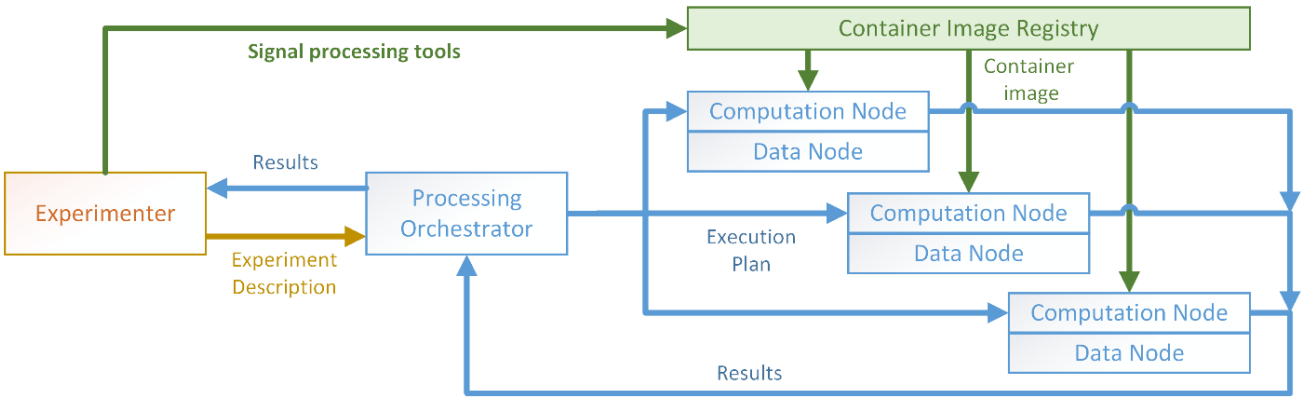


Figure 18: Deploying containers that distribute a computational experiment and collect results.

Frontend

The frontend (see Figure 19) is a Django application that offers:

- Different views for displaying list of pulses, machines, PIDs and users
- Dedicated view for displaying pulse details that present basic information about pulse, annotations, PID as well as charts for time series of selected variables.
- Dedicated view for displaying analytical / statistical details of the pulse variables
- Possibility to download pulse data in a JSON format as well as generated charts as images
- Filters to enable searching of the pulse data
- Support for declaration and management of user-defined annotations
- Support for different roles of users. Currently two roles are supported: administrator and regular user
- Data ingestion through RESTful API. Currently the demonstrator supports data ingestion from the AUG machine where it receives two CSV files (summary.csv and summary-time.csv) and creates the pulses. Since the files are large and the request is time consuming, the API creates a celery task in a Rabbitmq Server and the ingestion is executed asynchronously

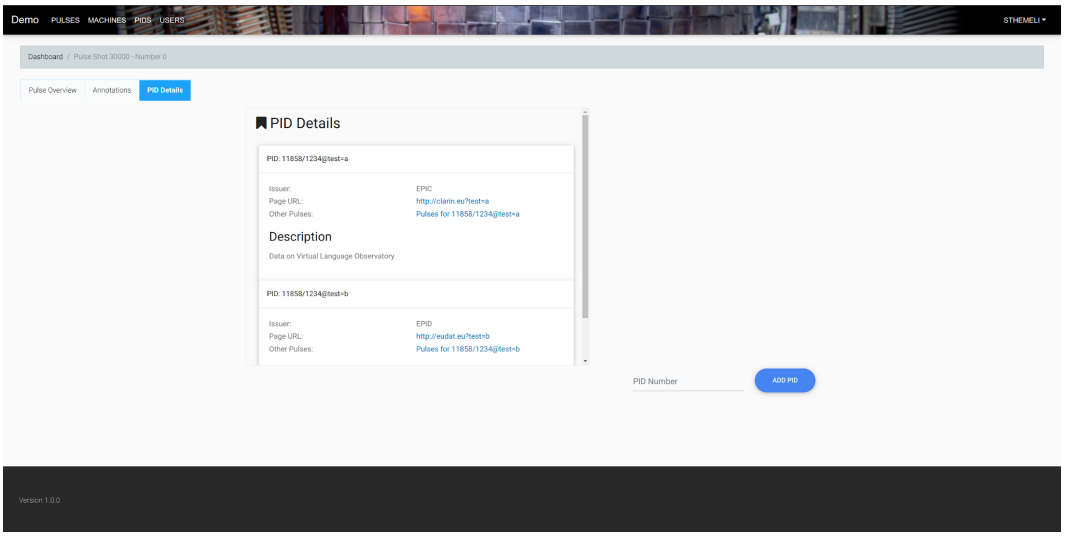


Figure 19: Example screenshot from Demonstrator II: PID Details tab in the Pulse view

7.3 Lessons learned

The process of development of Fair4Fusion blueprint architecture and two demonstrators revealed several topics that could not be sufficiently covered within the timeframe of the project, but they need to be discussed in more detail before the actual implementation of Fusion Open Data Framework.

7.3.1 Lack of common procedures and community-wide solutions

The limitation of currently available policies, lack of formal procedures as well as shortage of software systems that are common for the whole EUROfusion community cause significant difficulties in the Fusion Open Data Framework Implementation. The first item worth emphasising is the absence of a community-wide standard for distinguishing between EUROfusion (global) and national (local) data. Only EUROfusion data should by default be made open to the international community, but since there is no explicit distinguishing procedure, the ambiguities are unavoidable. The next limitations are diversified restrictions and physical data access mechanisms present on individual experimental resources. In a far-reaching perspective, all resources should present data compatible with IMAS, but at this maturity level of the EUROfusion environment, remote access to data spread over all resources may be realised only in an inefficient and complicated way using dedicated translation services.
Taking it a step further, there are currently two key elements missing in EUROfusion that would importantly streamline the Fusion Open Data Framework implementation and possibly a number of other systems. These are common Authentication & Authorization Infrastructure (AAI) and worked-out solutions or/and procedures for handling Persistent Identifiers (PIDs) for data generated within the community. Although the former element is currently being developed and it is foreseen to be available in a relatively short time, the lack of the later element needs to be stressed and a community-wide debate about it should be encouraged.

7.3.2 Open questions for software implementation

The main goals of development of two demonstrators were to empirically recognise limitations in a system's conception and, based on collected feedback, to figure out optimal solutions for the system to enhance its usability. The first discovered issue in this field is related to processing of large data on both the service part and the interface part of the system. The particular analysis is required to figure out a way of efficient presentation and comparison of data stored in pulse files that in a raw form is far too large to be handled by generic techniques. Demonstrator II is based on Cloud technologies for scalable computing in order to address this issue. Preliminary experiments carried out in the framework of WP5 have validated that Demonstrator II takes good advantage of the scalability offered by Kubernetes and, given adequate computational resources, will be able to process large-scale pulse data. Further works are also required to design the interface adaptable to different kinds of users and their expectations. Nevertheless additional analysis should be carried out to determine whether development of other levels is not essential and shouldn't be performed for specific groups of stakeholders.

8. Licensing

Clear data licensing is important regardless of whether data is made open or not. It tells users what they can and more importantly cannot do with the data to which they have access. Currently the methods of licensing vary quite a lot. In some cases data is made available to collaborators in projects through declaring it as background IP. In others, formal acceptable use policies need to be signed either by an individual or an organisation (legal entity). In some cases, even on-site users are unclear whether data can

be copied to an off-site location to allow more complex analysis. Thus there is often no single license model even within a site, which makes controlling data access difficult and is a barrier to scientific and engineering studies. Most experimental sites across the community do have some common themes in allowing users access to data:

- If data is used in a publication, the originating site must have an opportunity to review, augment, correct or reject a publication. There is a so-called “clearance process” imposed by all present experiments.
    - Some sites have an exception for small ad-hoc meetings where preliminary results are presented providing it is clear that the data is preliminary and/or unvalidated
- If a publication uses data from a diagnostic provided by a collaborating institute, they must also have the opportunity to review and comment prior to publication
- There is typically a need to cite the home institute in some way through either a citation or acknowledgement

Typically this is required because the interpretation of the data often requires a degree of expertise beyond the expertise of most researchers and it is relatively easy to make invalid assumptions. Much of this could be improved with improvements in metadata and documentation, and making more systems engineering metadata available, but this represents the current state.

In terms of FAIR it is recommended that this license be machine readable, but this is not mandatory, and should be as permissive as possible where data is publicly funded. That said, if data is initially licensed as permissively as possible, it can be difficult or impossible to relicense it under more restrictive terms, while a more restrictive initial license does not generally prevent relicensing to a more open license (excepting conditions for exclusive use by the licensor for a fixed period of time). The EU recommends data be licensed with one of the more permissive Create Commons licenses; either CC-BY or CC-BY-SA. A fuller description of the Creative Commons license is given in Annex A, together with a summary of licenses used by other large experimental communities. As a good general guide to different license types, refer to the Digital Curation Centre guide to research data licensing [27]. A fuller legal guide on open data licensing can be found in the article by Jyh-An Lee [28].

8.1 Recommendations

We recommend to follow the example of other major European projects [29],[30], and EU guidelines, and use Creative Common licenses [31] for opening validated fusion data and metadata to the public and wider research community. An embargo period of a few years (24 months) for data to give sufficient time for the Institutes running the experiment or collaborating with it to exploit the data first. In addition, data used for publication should be released co-incident or as soon as practicable with the recommended license. The release of data assumes that no restrictions have been placed on it for strategic, commercial or security reasons. The precise combination of CC flavour to be chosen is left to each data owner (experiments, modellers,...), but after discussion with legal experts of labs involved in Fair4Fusion we recommend using CC-BY-NC-SA [32]. As summarised on the Creative Commons site, “this license allows users to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator. If you remix, adapt, or build upon the material, you must license the modified material under identical terms”. In addition, the definitions for non commercial given under the license is “*...not primarily intended for or directed towards commercial advantage or monetary compensation*.” [33], with the use of the word **primary** as a recognition of the fact that no activity is completely disconnected from commercial activity. We note that MAST-U has decided to adopt this license, and we think that this is a good trade-off in terms of openness and fair usage of the data produced by a huge effort based on public funds, for the following reasons:

- BY (attribution): requires citing the data creator. Citation of the authors of a work or of a dataset is a common practice in scientific research, thus the usage of the BY option is quite natural. We further recommend that the citation be based on the institute name rather than an individual to recognise the contributions made by the many individuals involved in creation of the data.
- SA (share-alike): requires that, if derived data is produced and exploited, it has to be published with the same license as the original data. This would also allow the experiments to benefit from any adaptation of their own data under the same license.This appears as a relatively light and fair requirement.
- NC (non-commercial): prevents commercial usage of the data and of its adaptations. This appears as a fair restriction in view of the large public investment made for the experiments, which produces data that is then made available for free. Having a third party making commercial use of that data is seen by us as unfair, in particular in a context in which it is more and more difficult for experiments to be sufficiently funded by their public subsidies and they are encouraged to obtain additional funding via contracts. This leaves the possibility for public Institutes to benefit from a potential commercial exploitation of their data with e.g. a private partner, if there is such an opportunity. Without NC, instead, it's likely that a private entity would make a commercial use of the data without associating the public Institute to its benefits.
    - Note that NC doesn't prevent using the data to help produce additional work that can be commercially exploited, e.g. someone can use the data to build or verify a model, and then exploit the model commercially. Therefore, in our view, the NC flavour doesn't go against the idea of fostering the global research effort by associating more partners into it via open data.

Further details on the implications of the non-commercial license rider are available on the creative commons wiki page [34], but are repeated in Annex A for posterity.

It should also be noted that if different data producers use different licenses it will create difficulty when trying to combine data for the benefit of the community. For example, the pedestal and disruption databases would then need different licenses for different data providers and cross-experimental AI work could be constrained by the different licenses. Indeed, it is possible that in the worst case, data sets from different experiments could not be used together due to the licenses.

For strategic, commercial aspects, this license does not preclude any site from partnering with industry and sharing data with them under a bespoke license or even charging for data access requested by commercial entities, but that would be a specific agreement between the data owner and the data supplier and would not be through the mechanisms proposed in this blueprint.

9. Costs

As a part of the project, we have produced an estimate of the costs of making experimental data FAIR and open, and this is described in this section. While it is clear that some additional effort will need to be made by data generating facilities, the scope of this blueprint, and the project itself, has been to reduce this cost as much as possible, centralising as much as possible while still ensuring data is hosted outside of any portal. In this way, data producers are in charge of what should be made available to the community at large or the general public.For the rest of this section we look at costs split between central services, data producer services and optional services which, while not strictly necessary, will create either an enhanced user experience or improve the accessibility of the data.

9.1 Major Assumptions

A number of assumptions have been built into the cost model. These assumptions have been part of the basis for this work and from some knowledge of the future direction of data management across EUROfusion. The main assumptions, which are or will be costed elsewhere are:

- The costing provided here is the total cost for a full FAIR implementation, in practice this can be broken down in stages with costs related to different scenarios.
- The central services and portal will be hosted on a EUROfusion gateway machine, which is subject to ongoing discussions in relation to FP9 funding. If this were not the case then an additional call would need to go out for a hosting site, which would potentially need to be whitelisted by all data producers.
- The AAI system developed under FP8 will be supported, deployed and adopted at all sites during FP9 which is currently foreseen in the FP9 proposal under the EUROfusion PMO. Without such a community adopted system, significant additional effort would need to be invested into a one-off security system for the portal.
- There is a populated IMAS installation at each site, with data available as IDS structures, including a Summary IDS (prerequisite).
- Increase in network bandwidth. Based on work in other similar communities, open data use cases could require about 10% additional bandwidth capacity per annum above existing and foreseen usage.
- Public access to full resolution data will not be permitted unless specifically requested by the data provider (in this case, public means not associated with EUROfusion). Public access to Summary IDS data will be done directly from the portal, not impacting data producer sites
- Initially, access to full resolution data will follow existing procedures. If a common policy can be adopted, access to EUROfusion partners will be via the AAI system
- Backing up the service and associated metadata will be the responsibility of the site hosting the central service
- Reuse and adoption of FAIR4Fusion components is encouraged to reduce costs

In the costing exercises we have been evaluating different scenarios of FAIR compliance:

Scenario A: making metadata only available and searchable using IMAS data subsets for interoperable definitions of quantities [F,(I)]

Scenario B: adds to Scenario A by allowing a subset of the data to be accessed using common tools (ro example UDA). Facilities are responsible for the access level and qualification of data through the data mappings [F,A,I,(R)]

Scenario C: builds on the previous stages and allows for enhanced data provenance and referencing through PID's [F,A,I,R]

Scenario D: adds a lightweight layer for open access to non-embargoed metadata and where allowed by the facilities also data access for export in human readable formats (CSV files) [F,A,I,R] and open.

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

9.2 Benefits

9.2.1 Non financial benefits

In terms of costs there are a number of benefits. The vast majority of these are accrued over time and are somewhat dependent on new researchers who will make use of experimental data. While cost savings are difficult to estimate, there are intangible benefits - where a direct cost can not be made and there benefit must be weighed against cost on a case by case and site by site basis. Amongst these intangible benefits:

- By providing more access to data the community can enthuse a new generations of researchers in secondary education,
- More opportunities for outreach for example “Bring your own data” coding challenges which has the opportunity to improve codes used in simulation and research,
- Better ability to validate simulation by allowing simpler comparison with a wider range of devices and configurations,
- Closer collaborations between experiments may be needed to more rapidly make progress towards realisation for commercial fusion
- Possible improved collaboration with industry leading to faster technology breakthrough, for example in materials science,
- Collaboration on algorithms and infrastructure with other communities

9.2.2 Financial Benefits

The identified financial benefits come from

1. Reduced training costs for new generation of scientists
2. Common software infrastructure means support and development can be shared across sites
3. A common security infrastructure means practises can be shared across organisational boundaries

9.3 Centrally Managed Services

The central services identified are shown in Figure 20. Items shown in green are either developed or in development within the project while those in white will need to be developed or deployed. It should be noted that even for components developed within this project, additional effort will be required to extend functionality and production harden the services as more data sources are integrated, more users access the service and to support full support to sites and users.

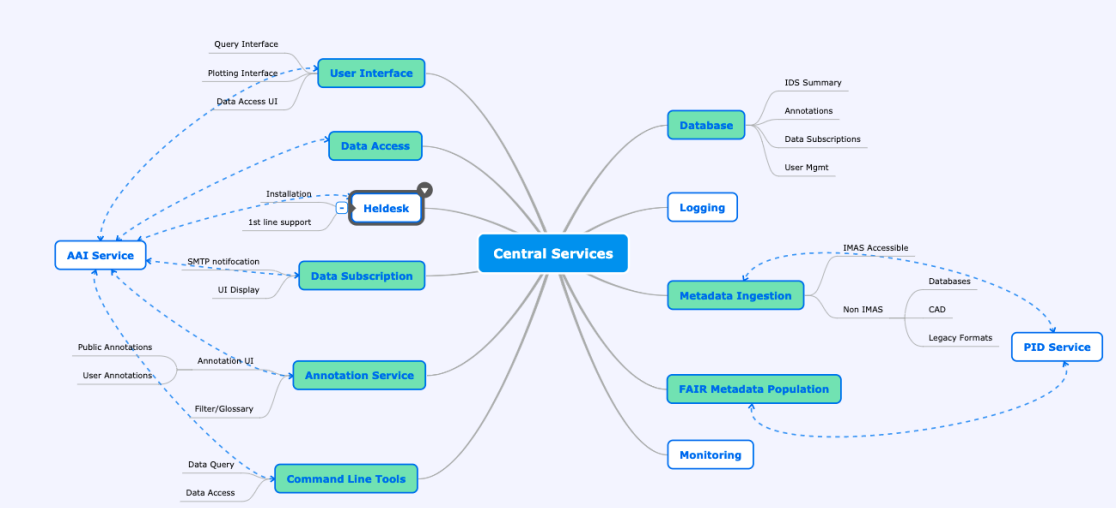


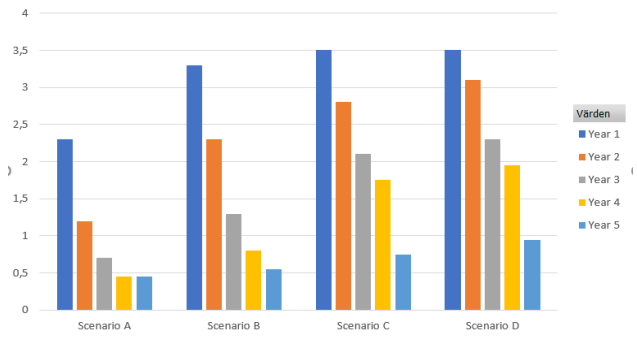
Figure 20: Centrally Managed Services

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

Of those services not addressed within the current project, the AAI has already been discussed as being out of scope, helpdesk and monitoring software can either be procured, open source versions can be adopted, or can be integrated into the hosting sites existing services for this, with each option having a different price. Also note that the services shown in this diagram represent a minimal set and do not offer high availability. This is discussed in a later section. Based on the consortium's experience in the development of e-Infrastructures and discussions with similar organisations, a summary of our estimate of costs for supporting the central services are shown for the different scenarios below in Table 2.

Table 2. Direct Costs for supporting Central Services

	Total (5yrs) PPY	Average/yr PPY
Scenario A	5,1	1,02
Scenario B	8,25	1,65
Scenario C	10,9	2,18
Scenario D	11,8	2,36



9.4 Site Services

The identified site services are shown in Figure 21 below. Unlike the central services, these have only been explored at a high level within the current project. However, most data producing sites will already have these services in place, but may need adaptation to, for instance, comply with FAIR guidelines or provide metadata as a Summary IDS.

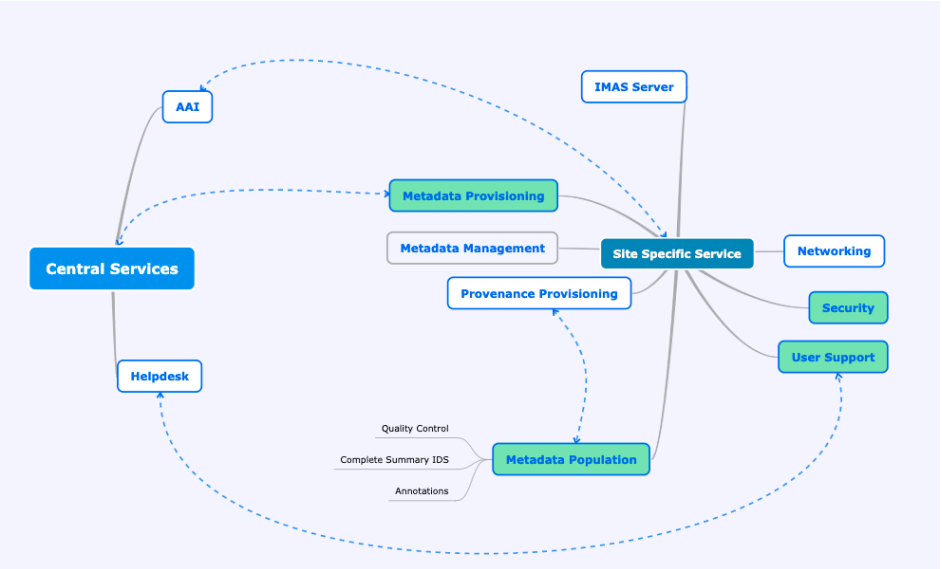


Figure 21: Site Managed Services

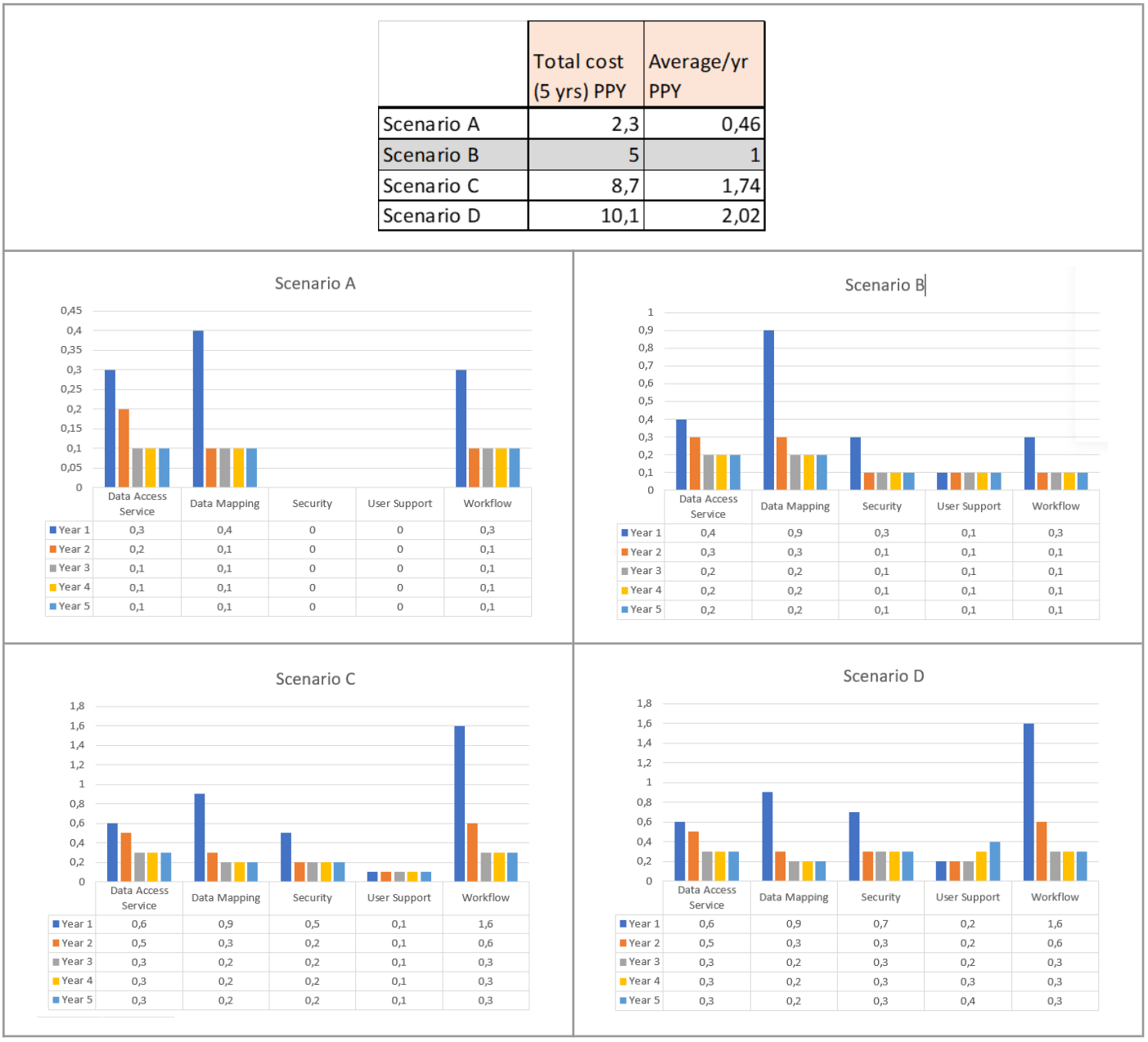
There are also options for levels of information supplied. For example, provenance information can be scraped from existing logging (the approach we have taken within this project), or could be integrated into the processing workflows which, with more effort, will provide a more interoperable and complete provenance model. While data access follows existing rules, additional effort will be required if access is made available from a wider audience and processes and procedures will need to be put in place to

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

support this. We suggest a progressive growth of the perimeter of data that is made available to the central services for a given experiment, and a progressive improvement of the provenance capture, so that the additional effort remains acceptable: start with a restricted set of Summary data and expand over the years to a more complete coverage of the experiment's data, with augmented provenance information. Note that this effort will leverage other FP9 EUROfusion activities (e.g. the EUROfusion Databases, code validation on experimental data in the TSVVs), so it doesn't represent pure additional effort but will create synergies. As above, the estimated costs of supporting open access are shown in the summary table below (Table 3). The figures below represent the maximum estimated costs required at each site. We also acknowledge that many local Data Management Plans now provision a dedicated budget so that data providers can make their data more fair/open. Such resources may help cover some of the discussed implementation costs.

Table 3. Costs for supporting Sites Services (at each site)

	Total cost (5 yrs) PPY	Average/yr PPY
Scenario A	2,3	0,46
Scenario B	5	1
Scenario C	8,7	1,74
Scenario D	10,1	2,02



This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

9.5 Including non-experimental and non-IDS data

While we anticipate that at least the Summary IDS will be made available to the central services from the large fusion facilities, modelling, simulation and engineering data could come in a number of different formats which may or may require additional mapping to the IDS structure. Most of the effort for incorporating this type of data will fall within the central services, with a small additional cost for the data producers in terms of secondary support related to the data. While this is not a technical barrier the amount of effort is heavily dependent on the source and the effort required to map the source to IDS parameters.

9.6 High Availability Service

As stated previously, the deployment suggested in 9.2 represents a single instance with no redundancy. While information contained within the central services can be backed up and any data service restored from backup, this is likely to be an effortful process meaning a hardware failure could result in the service being down for significant time (order of a few days). If the service proves popular and more people rely on it for dataset identification and access, thought must be given to making the service highly available so that a service interruption on one instance should allow requests to be redirected to a secondary service running at the same or different site, with minimal disruption to users. While much of this functionality can be achieved through typical database features common in most modern databases, some additional effort is required in terms of monitoring, automated failover and possibly additional security implications for sites. This is shown diagrammatically in Figure 22, and a summary costing is shown in Table 4. Note that it has been assumed that the cost difference between hosting both the primary and mirror service within the same organisation and being hosted by different organisations is believed to be small. The main consumer of effort will be in the policy concerning how changes will be made to each system and development of procedures to verify successful metadata replication. We have assumed the instance can share the same monitoring and helpdesk is available at each site as well.

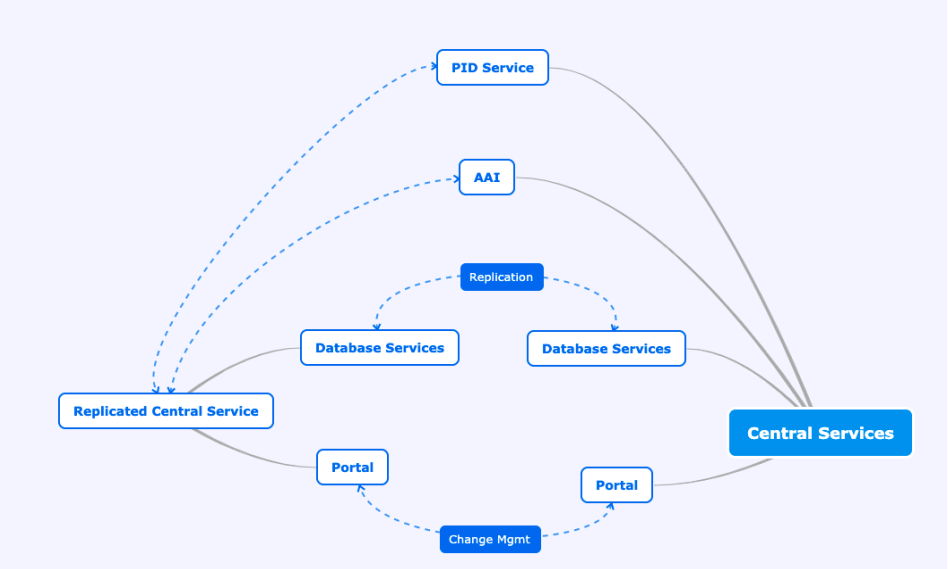


Figure 22: Diagram describing the proposed High Availability Services

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

Table 4. Costs for High Availability Service

High availability option	EF partner (FTE)	1	1,1	0,8	0,8	0,8	
	k€ [including infrastructure)	99,75	103,125	75	75	75	427,875
	Cloudprovider (FTE)	1	0,4	0,4	0,4	0,4	
	k€ [including infrastructure]	96,75	40,5	40,5	40,5	40,5	258,75

10. Recommendations

Item	Recommendation
Policy	Establish a central metadata catalogue, accessible and searchable (through a Web Portal), gathering data from multiple experiments available immediately. Make the actual data also publicly accessible, after an embargo period. Use IMAS Data Dictionary to publish metadata and data in an interoperable way, thus progressively develop mappings of data of EU experiments to IMAS. Where suitable IMAS data formats are not available or appropriate, extensions to IMAS should be promoted or other standards be adapted. FAIRness of the research results/DATA SETS should be taken into consideration while evaluating progress on project and tasks.
Licenses	We recommend using Creative Common licenses for opening fusion data and metadata, after an embargo period of a few years for data and immediately for physics metadata. The precise combination of CC flavour to be chosen is left to each experiment, but we recommend using CC-BY-NC-SA.
Architectural fundamentals	The system should integrate metadata coming from many sources, which may be natively using different formats, and present it in an unified format to the users. In order to enable easy usage of the system, the integration of metadata and its provisioning should be managed centrally, with help of specialised services accessible with well-defined remote interfaces.
Metadata handling	The metadata structures should be based on existing standards, in particular on IMAS Data Dictionary and its internal data structures like Summary IDS and "dataset_fair" IDS. If some metadata is not supported by IDSes, as a first step it should be analysed if the IDSes could be extended to support this metadata.
Integration of metadata from experimental sites	Experiments internally use different formats for their metadata. Assuming IMAS as the underlying format for metadata within the Fusion Open Data Frameworkenvironment, there is a need to provide a set of translation services to enable translation of custom formats to IMAS. The *push* and *pull* styles of data retrievement can be considered. In any case, if it is not denied by certain policies or/and technical restrictions on sites, it is recommended to base the integration on the REST protocol. While the *pull* style can be seen as more reliable, it may be disallowed by certain experiments. Therefore support for both methods shall be provided.
Physical data access	Access to physical data not stored within the central Fair4Fusion services should be possible for Fusion Open Data Framework clients: based on search results it should be possible to retrieve experimental data. EUROFusion encourages the sharing of data generated by EUROFusion funded projects through the IMAS/IDS infrastructure. For the initial implementation of the system, the access to experimental data on resources may be realised independently, based e.g. on PIDs and third-party services. For the final solution development of custom service, which will play a role of a proxy to IMAS compatible resources and a role of translation broker to non-IMAS resources, should be considered.
Client components integration	The central services of the system should be accessible for both the client tools being the integral part of the Fusion Open Data Framework system as well as for possible external clients. In order to provide an easy and standardised way of client access, it is recommended to provide a well-defined REST API as the main entry-point to the system.

Core services	The role of core services is to store aggregated metadata coming from different experiments as well as other kinds of data and metadata essential for the system that is provided by users or inferred from existing data. The data structures employed by the core services should be based on IMAS and possibly other widely-accepted standards (e.g. W3C PROV for provenance). Whenever possible and justified the usage of existing services, such as CatalogueQT, is recommended. Note that this service has been already extended as part of the Demonstrator work carried out during WP5. The central services should be provided with High Availability principles (e.g. using Kubernetes cluster).
Supplementary services	In order to support generic functionality an integration of few additional services should be considered. An example could be a PubSub service to enable asynchronous communication within the system or proxy service enabling access to experimental data. Additional services may be also needed to enable remote configuration and accounting. It is recommended to use existing software to provide these services.
User-level components	The main entry point to the system needs to be a Web Portal offering an extensive set of features for searching, filtering, displaying, comparison and management of metadata. It should also enable access to experimental data. It is recommended to implement the portal in a modern way, i.e. divide it into frontend (responsible for the presentation and user interactions) and backend (aimed to provide a flexible way of integration with services). In addition to Web Portal, there should be a command-line client interface provided for all use cases that need some sort of automation (possibly based on REST API).]In order to improve the performance of the system, Web Portal can utilise search engines (e.g. Elasticsearch, Solr), server side caches (e.g. Memcached, Redis) and finally it may need to be deployed in an environment ensuring High Availability (e.g. in a Kubernetes cluster).
AAI	In order to ensure good-level of security and enable easy integration of components, the system should use federated AAI. AAI should be implemented in accordance with the AARC blueprint architecture. Once EUROfusion AAI Proxy is made operational its usage is recommended.
Persistent Identifiers (PIDs)	All data exposed by the system, including metadata and experimental data that can be referenced, should be marked with Persistent Identifiers (eg. DOI or ePIC) so that they can be uniquely recognized. The architecture also uses PIDs (or other unique identifiers) to provide lined version histories for data sets

11. Roadmap

Roadmap of the Blueprint implementation

Fair4Fusion recommends data policy and actions that can be supported by the developed prototypes. From these recommendations the roadmap below underlines the main actions to be taken:

EUROfusion will have to:

- Decide on policy for data publication, licenses (and make its DMP consistent with the policy)
- Define an implementation project : scope, organisation, resources, timeline
    - It should include the production quality of central metadata services and metadata ingestion part, and the portal component
- Organise dissemination of the services and user feedback, evolutions of the project

Central Services are required to:

- Host the Services - (the recommendation is to start with the prototypes developed by Fair4Fusion)
- Provide AAI
- Provide PIDs
- Set up a helpdesk system with first line support
- Set up a monitoring service

- Guarantee data preservation and archiving

Experimental sites will have to:

- Provide Summary physics metadata
- Implement mechanisms to submit the data to Central Services (including remote data access)
- Find and implement a mechanism to ensure previously submitted data can be updated, following for instance data improved (re)processing.
- Progressively map more physics data to IMAS
- Select the data to be exposed by each of the experiments, and progressively increasing the amount of data
- Progressively add provenance information in the output of experiments-related workflows
- Maintain FAIR information related to the submitted data (dataset validity, publications, …)
- Extend the data submission to simulations related to experiments
- Can analyse the option of integration with the pinboard services

The implementation of the roadmap can be divided into following phases:

- Agreement between stakeholders and definition of the implementation project
- Implementation of the production quality services
- Deployment of the infrastructure and services
    - If needed organise a call for providers
- Operational maintenance and support of the services

Necessary building blocks needed for roadmap implementation:

- AAI Rollout in EUROfusion
- Persistent Identifiers (service to buy)
- Existence of EUROfusion Data Management Plan

12. Summary

Final messages:

- Native access to experimental data is still somewhat limited, there is a lot to gain if the community can harmonise on tools and methodologies for bringing data to end users
- The Blueprint is a joint effort of the experiments, modelling community, technological providers, and is taking into account feedback of the European fusion community
- Achieving FAIRness for fusion data is feasible and it is a big opportunity for the whole fusion research community by:
    - Help promote broad internal and new/novel collaborations and increase the secondary use of data
    - Facilitate cross device research initiatives and reduce thresholds for large scale data mining and AI/ML approaches
    - Integration of the horizontal eTASC (theory, simulation, V&V) activities with experiments
    - Consolidate knowledge and tools towards a consistent “data and software” ecology for exploitation of ITER and DEMO.
- Making data open and FAIR requires: policy changes, development and deployment of production quality toolsets
- Fair4Fusion project proposed an architecture complemented with the technical prototypes that could be further developed and reused to deliver necessary services for the community. Two integrated,

container-based, demonstrator systems were developed using open-source software components, while the in-house tools and technologies employed address requirements set out by the fusion research community

- One the first steps towards achieving this goal is to open the metadata (using Creative Common licences), i.e. the physical information summarising a plasma experiment and allowing identifying datasets of interest
- A roadmap towards FAIRness and Open Data Access is proposed. Its implementation needs financial and organisational support, as well as the commitment of the experiments.

References

[1] https://www.go-fair.org/fair-principles/
[2] Imbeaux, F., Pinches, S. D., Lister, J. B., Buravand, Y., Casper, T., Duval, B., ... & Strand, P. (2015). Design and first applications of the ITER integrated modelling & analysis suite. *Nuclear Fusion*, *55*(12), 123006. https://doi.org/10.1088/0029-5515/55/12/123006
[3] https://en.wikipedia.org/wiki/Open_data
[4] https://www.sciencedirect.com/science/article/pii/S0010465510000214?via%3Dihub
[5] https://www.euro-fusion.org/
[6] https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex
[7] https://www.slideshare.net/Etalab/g8-plan-daction-open-data-pour-la-france
[8] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433&uri=CELEX:32019L1024
[9] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9. https://doi.org/10.1038/sdata.2016.18
[10] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg
[11] Genova, F., Aronsen, J. M., Beyan, O., Harrower, N., Holl, A., Hooft, R. W., ... & Jones, S. (2021). *Recommendations on FAIR metrics for EOSC*. Publications Office of the European Union. https://doi.org/10.2777/70791
[12] Frédéric Imbeaux, David Coster, Pär Strand, Joan Decker, & Shaun de Witt. (2020). Data Inventories and Policy Landscape (1.5). Zenodo. https://doi.org/10.5281/zenodo.4336791 (Deliverable D2.1)
[13] http://www.mdsplus.org/index.php/Introduction
[14] https://www.iter.org/org/team/fst/itpa
[15] Koukourikos, Antonis, Ikonomopoulos, Andreas, Klampanos, Iraklis Angelos, Sissy Themeli, Karkaletsis, Vangelis, Bosak, Bartosz, Płociennik, Marcin, Palma, Raul, Imbeaux, Frédéric, & de Witt, Shaun. (2020). Report on Technology Survey and Demonstrator Requirements (Version 16). Zenodo. https://doi.org/10.5281/zenodo.4338059 (Deliverable D3.1)
[16] David Coster, Pär Strand, Frédéric Imbeaux, Shaun de Witt, Marcin Płociennik, Andreas Ikonomopoulos, Irakalis Angelos Klampanos, & Joan Decker. (2020). Final Report on Open Science Use Cases for Fusion Information (1.0). Zenodo. https://doi.org/10.5281/zenodo.4337222 (Deliverable D2.3)
[17] https://pypi.org/project/fusionprov/
[18] https://www.w3.org/TR/prov-overview/
[19] https://prace-ri.eu/
[20] https://www.eocoe.eu/
[21] https://aarc-project.eu/architecture/
[22] https://eduteams.org/
[23] https://github.com/mkopsnc/catalogue_qt_docker
[24] https://github.com/mkopsnc/catalogue_qt_docker/blob/master/docker-compose/README.md
[25] Iannone, F., Bracco, G., Cavazzoni, C., Coelho, R., Coster, D., Hoenen, O., … Voitsekhovitch, I. (2018). MARCONI-FUSION: The new high performance computing facility for European nuclear fusion modelling. *Fusion Engineering and Design*, *129*, 354–358. https://doi.org/10.1016/j.fusengdes.2017.11.004
[26] https://kubernetes.io/docs/concepts/extend-kubernetes/
[27] https://www.dcc.ac.uk/guidance/how-guides/license-research-data
[28] Jyh-An Lee, *Licensing Open Government Data*, 13 Hastings Bus. L.J. 207 (2017). Available at: https://repository.uchastings.edu/hastings_business_law_journal/vol13/iss2/2
[29] https://creativecommons.org/2010/07/15/cern-supports-creative-commons/
[30] https://creativecommons.org/2016/11/02/atlas-cern/
[31] https://creativecommons.org/licenses
[32] https://creativecommons.org/licenses/by-nc-sa/4.0/
[33] https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode
[34] https://wiki.creativecommons.org/wiki/NonCommercial_interpretation

This project has received funding from the European Community's Horizon 2020 Framework Programme under grant agreement 847612

Annex A - Common Licensing Options

Open Data Licensing

There is a strong recommendation from the European Commission and projects such as the European Open Science Cloud and OpenAIRE that open data be licensed with one of the Creative Commons licenses. This type of license has been widely adopted across the members of EIROforum, as seen in the table below, and across the wider international science community. It should be noted that these permissive licenses typically are applied after some embargo period, to allow early exploitation of results by researchers or other commercial usage. We recommend using the same approach for fusion data (although metadata may be published without any embargo).

While details of the Creative Commons licenses can be found elsewhere, these are summarised below:

- CC-BY: Requires the data creator to be cited. In principle any secondary data created from derived data should also cite the originating data and its creator.
- CC-BY-SA: As CC-BY plus requires any derived data to be shared under the same license (effectively a copy-left license). Can discourage commercial exploitation.
- CC-BY-ND: This license lets others reuse the data for any purpose, including commercially; however, it cannot be shared with others in adapted form, and credit must be provided to you. This can prevent publication of results, since any data derived from an experimental source cannot be shared.
- CC BY-NC: This license lets others remix, adapt, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don't have to license their derivative works on the same terms.
- CC0: Strictly this is not a license, but a public domain dedication. Releasing data under CC0 means creators and owners of copyright- or database-protected content waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

It should be noted that while the Creative Commons are the most widely adopted 'open data' licenses in Europe, others do exist such as the meta-share and the UK Open Government Licence. A reasonably comprehensive set of open licenses can be found at the LINUX Foundation SPDX project website[1], although this list covers data, software and documentation so care must be taken in selecting an appropriate license. Finally, open licenses are not generally rescindable; once data is licensed with an open data license, this license cannot be replaced later by a more restrictive license.

Member	Number of Sites	Open Metadata?	Open Data?	License	Embargo Period (yrs)
CERN	4[2]	Varies	Varies	CC0	3-10
EMBL	6	No	No	Specific	N/A
ESA	>20	Varies	Varies	CC-BY-SA-3.0	Varies
ESO	6	Yes	Yes	CC-BY-4.0	1

[1] https://spdx.org/licenses/

[2] While CERN is an individual device, each major experiment has its own data policies, although they are somewhat harmonised.

ESRF	1	Yes	Yes	CC-BY-4.0	3
E-XFEL	1	Yes	Yes	CC-BY-4.0	3
ILL	1	Yes[3]	Yes	CC-BY-4.0	3
EUROfusion	8	No	1 site	Custom	3

Table A1: Comparison of Data Licenses and Embargo Periods between EIROforum Members

Databases introduce their own complexities. Under European Law, database contents are protected under *sui generis* rights, while the schema's associated with them are protected under standard copyright protections. A good example of sui generis rights is explained by the EU in https://europa.eu/youreurope/business/running-business/intellectual-property/database-protection/index_en.htm. To further complicate issues, these rights are not applicable universally. For instance the US and Australia do not recognise these rights. The Creative Commons Licenses should not be applied to databases; these would be typically made open under the Open Data Commons licenses[4],[5].

Implications of the Non-Commercial Rider in Creative Commons Licenses

The following is reproduced from Creative Commons Interpretaion of the Non-Commercial rider in CC licenses[6].

The NonCommercial limitation applies to licensed uses only and does not restrict use by the licensor.

As with all CC licenses, the NC licenses only restrict what a reuser may do under the license and not what the licensor (rights holder) can do. Licensors that make their works available under an NC license are always free to monetize their works.

NonCommercial turns on the use, not the identity of the reuser.

The definition of NonCommercial **depends on the primary purpose** for which the work is used, **not on the category or class of reuser**. Specifically, a reuser need not be in education, in government, an individual, or a recognized charity/nonprofit in the relevant jurisdiction in order to use an NC-licensed work. A reuser that is not obviously noncommercial in nature may use NC-licensed content if its use is NonCommercial in accordance with the definition. The context and purpose of the use is relevant when making the determination, but no class of reuser is per se permitted or excluded from using an NC-licensed work.

Reusers may make NonCommercial uses only, even when reusing NC material with other works.

The NC licenses limit reusers to NonCommercial uses of the work only, which includes when the work is used in a collection or when it is adapted. For example, an NC essay may not be included as part of a collection in a commercially distributed book of essays, even if it is only a small portion of the book. For an example of an adaptation, an NC song may be used as the basis for a video where the visual elements are under a different license such as the BY license. When the music video is distributed as a whole, it may not be used commercially because of the NC license of the song.

[3] This only applies to publicly funded research
[4] https://opendatacommons.org/
[5] https://wiki.creativecommons.org/wiki/data#What_is_the_difference_between_the_Open_Data_Commons_licenses_and_the_CC_4.0_licenses.3F
[6] https://wiki.creativecommons.org/wiki/NonCommercial_interpretation

The NonCommercial term does not limit uses otherwise allowed by limitations and exceptions to copyright.

Nothing in the NC licenses (or any CC license) controls or conditions uses—even commercial uses—covered by an exception or limitation to copyright or similar rights, or otherwise controls any activity for which no permission under such rights is required. For example, a person may commercially use an NC-licensed work for purposes of criticism in jurisdictions where this is a fair use or otherwise covered by an exception to copyright. Similarly, because posting a link to a work does not require permission under copyright, a for-profit university may still include a link to NC-licensed courseware in a syllabus or on its paywalled website. In such cases, the CC license never comes into play and the NC restriction (and other limitations or conditions contained in the license) may be disregarded.

Explanations of NC do not modify the CC license.

Some licensors or website providers state expectations or interpretations about what NC means. Those explanations never form part of the CC license, even if included in terms of service or another resource designed to contractually bind reusers. CC strongly discourages the practice when such statements carve back (rather than expand) on reuses allowed by the NC definition or contradict the plain meaning of the licenses. When those statements are intended to bind reusers or to modify the CC license, no CC trademarks may be associated with either the work or the terms under which it is offered. For more information about CC's license modification policy, see the Creative Commons web page discussing this[7].

NonCommercial licenses are non-exclusive.

Like all CC licenses, the NC licenses are non-exclusive. This means that an NC licensor is free to offer the material under other terms, including on commercial terms. A frequently discussed use case for the NC licenses is a creator who wishes to allow NonCommercial use but also authorizes commercial uses in exchange for payment. (Additional permissions such as this may always be offered; licensors may also use CC+ protocol to offer these in a standardized manner[8]). Also, licensees are always free to contact licensors to ask permission to use the work for commercial purposes.

For a given work, permitted NC uses may still be restricted due to non-copyright rights.

Even if a use is NonCommercial for purposes of the CC license, it may still not be permitted because of other rights that prevent that particular use of the work. For example, a use that is otherwise NonCommercial could violate the publicity or personality rights of an individual featured in the work.

[7] https://wiki.creativecommons.org/wiki/Modifying_the_CC_licenses
[8] https://wiki.creativecommons.org/wiki/CCPlus