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

Interim Report

Version	Version 1.1 (Interim Report)
Type	Report
Dissemination level	Public
Lead Beneficiary	PSNC
Date	30.08.2020

Authors:

PSNC: Marcin Płóciennik, Bartosz Bosak, Raul Palma, Michal Owsiak
UKAEA: Shaun de Witt, George Gibbons, Nathan Cummings
NCSRD: Iraklis Klampanos, Iris Xenaki, Andreas Ikonomopoulos
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 https://www.go-fair.org/fair-principles/ for a detailed description of these principles.
IMAS	ITER Integrated Modelling and Analysis Suite. This suite of interoperable analysis code, sponsored by ITER Organization, is based on a machine-generic ontology, the Data Dictionary. A useful reference explaining the underlying principles of the Data Dictionary is [F. Imbeaux et al, Design and first applications of the ITER integrated modelling & analysis suite. Nuclear Fusion, 2015, 55, pp.123006. DOI : 10.1088/0029-5515/55/12/123006 https://hal-cea.archives-ouvertes.fr/cea-01576460/document]
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 https://en.wikipedia.org/wiki/Open_data
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.
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 Fair4Fusion project is to make European funded fusion data more widely available in order to maximise the impact of the research. 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 long term architecture for Fusion Open Data Framework implementation.

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 cover also 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.

We present this Blueprint to 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 for implementation on joint experiments and modelling activities. As a significant fraction of the European fusion research is done in joint collaboration with domestic programmes the support and commitment from the 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 the data providers. Those use cases are presenting the different perspectives of the members of the general public, EUROfusion researchers and data providers that are the main target users of analyzed 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 a 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 a 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:

- It is designed as a machine-generic ontology, capable of covering all experiment subsystems and plasma physics, and is extensible
- It's the only ontology standard that has been elaborated in the fusion community (with the exception of the "CPO" data model, which can be considered as the ancestor 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 access easily 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 organization. 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 ITER Organization (IO), EU labs have access to it and EUROfusion has already a formal collaboration with IO on development and usage of IMAS

The resulting architecture of the system consists of the 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 produced by experiments. In the proposed design, Metadata Ingests stay in an administrative control of particular experiments, thus the experiments 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 with 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, the key importance is given to Web

Portal that is expected to offer an extensive set of functionalities for searching, filtering or displaying metadata and data managed within the system.
The final version of this document should be a complete source of information for the stakeholders on how to develop a production-ready system. Thus, based on experiences and lessons learned during the course of the project, we are planning to supplement the ultimate version of the blueprint with the discussion aimed to advocate particular solutions and point out possible risks in the proposed design as well the recommendations.

2. Introduction

2.1 Objectives

The overall objective of Fair4Fusion project is to make 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 adhering 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 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.

This **blueprint architecture** aims for long term architecture for **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, high level architecture, how different local policies will be handled, recommendations on standards, achieving interoperability, type and granularity of metadata and persistent identifiers to expose and investigate the use of metadata annotations to allow enrichment and enhance 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 operate the WEST tokamak, MPIPP operate ASDEX-Upgrade, EPFL operate TCV and UKAEA operate both the MAST tokamak and JET on behalf of EUROfusion.

2.2 Scope

This document is the first early draft of the Blueprint Architecture, that will be distributed for the comments, and consultations. The first early version will be published at Month 12 of the project (end of the August 2020). Its final version is foreseen at the end of the project at month 24.

The scope of this document is a description of the target architecture of the Fair4Fusion system for the aggregation and management of metadata coming from distributed Fusion experimental 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 wouldn't be addressed in this document.

The current version of the document provides a high-level overview on the architecture of the target Fair4Fusion system, but does not display all detailed aspects of the foreseen architecture solutions. More verbose description of possible ways of handling provenance, annotations, simulation data, and likely other functionalities that would be recognized as needing particular attention, will be provided in the next version of the 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 art - so the starting point as well as existing obstacles in terms of policies, data access, 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. 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.

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

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 all 28 European member states are members of the Euratom treaty, with Switzerland as an additional associate member. Most of the community in Europe is now gathered mainly under the EUROfusion project[1] umbrella 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 within Europe and beyond. 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. Indeed, even security is currently delegated to sites, with different experiments operating different policies for accessing the data, no standard data format, no consistent metadata schema or naming operations.

The European fusion community has become increasingly collaborative over the last few decades with more experimental devices becoming available for broader groups of researchers. 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

[1] https://www.euro-fusion.org/

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

has been a move to make more 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[2,3] and across national borders[4].An effort towards an Open Data environment for European fusion research can spearhead also efforts to be made in ITER.

3.2 Fusion experiments

3.2.1 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 resource, including the highest Q value recorded (that ratio of power in to paper out), the highest peak energy and the highest plasma current. In support of ITER operations it is starting its second ever experimental campaign using Deuterium-Tritium (DT) as a fuel source, a fusion reaction producing more than 5 times the energy of Hydrogen-Hydrogen reactions powering the sun. JET is currently the only tokamak capable of running DT plasmas.

3.2.2 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.

[2] https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex
[3] https://www.slideshare.net/Etalab/g8-plan-daction-open-data-pour-la-france
[4] Directive (EU) 2019/1024

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

3.2.3 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. Over the nearly 30 years of operation it has performed nearly 40000 plasma discharges.

3.2.4 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.

3.2.5 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 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.

3.2.5 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 heatload by a factor of 10, overcoming one of the issues

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

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 diagnostic purposes from a wide range of sensors, and 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 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 will be brought into the FAIR process at a later stage.

3.4 FAIR

The FAIR principles[5] 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 and the ESOC Secretariat FAIR Working Group recommendations on FAIR metrics for EOSC, 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

[5] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

increase the trust in the data, and a data dictionary or ontology service to support cross disciplinary usage of the data.

4. Current state of the art

The detailed analysis of the current state of the art has been performed at the beginning of the project, and is detailed in the deliverable D2.1[6], in this section we present highlights of this 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 double password 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 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 from exploiting results of the EU fusion experiments at their full potential.

Formal Data Management Plans (DMPs) have not been established by any EU 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 and whatever degree of formalisation they request from the researchers, all experiments have established similar rules for using data in a publication, based on a formal publication clearance procedure. This clearance procedure constitutes the main feature and also the common ground of the data policies in all European experiments.

[6] Deliverable D2.1: Data Inventories and Policy Landscape

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

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 are 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, a heating & current drive system. The IMAS ontology provides the possibility to store and access easily 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 [http://www.mdsplus.org/index.php/Introduction].
- 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.

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

- 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 AFS for remote data access. The 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 enable 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, https://www.iter.org/org/team/fst/itpa]
- Accessible (via authentication, so not open), for fusion researchers having an official link to an experiment, using access methods specific to that experiment
- 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.
  - 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.

In summary, when considering a single experiment, its data has already today 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.

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.

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

5.1 Users and access levels

The target audience of the system 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 computer users that will take benefit of machine readable data. It is hardly possible to precisely define all categories of the users. This all leads to quite an extensive set of requirements and interface design decisions that need to be incorporated into the system. Therefore our goal is to make the system generic and extendable.

As a starting point, we have identified six basic user categories that are the main target of the system:

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 the 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 will 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 are presenting the different perspectives of the 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 F4F project deliverable D2.3[7], 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.

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.

[7] Deliverable D2.3: Final Report on Open Science Use Cases for Fusion Information

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

5.3 Identified Client Interactions

Figure 1 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.

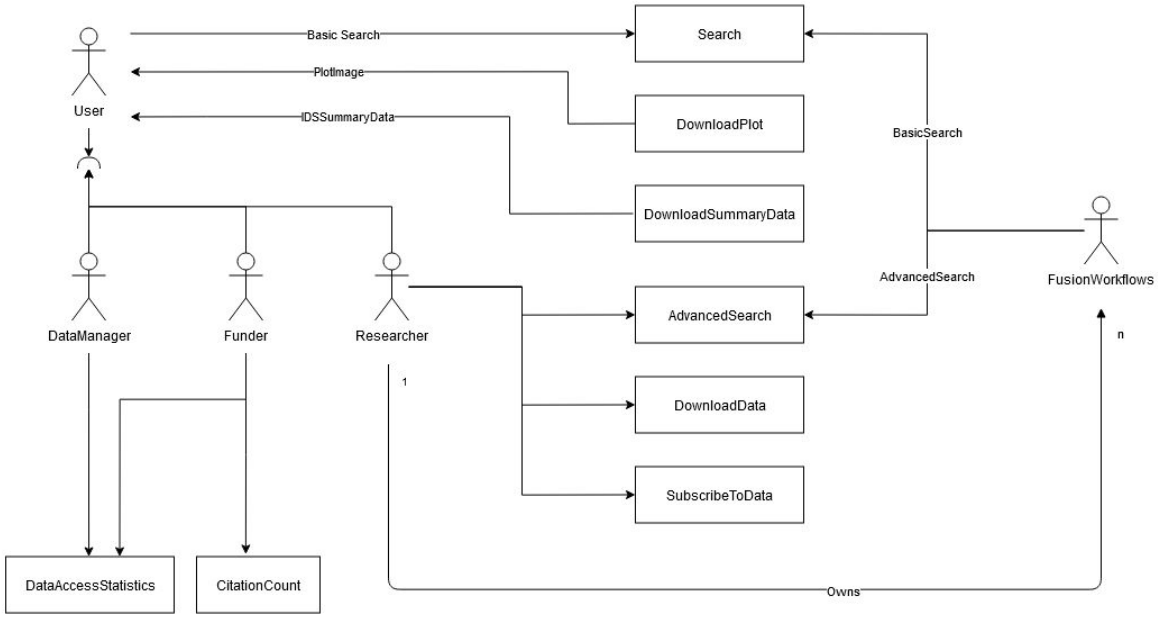

Figure 1. Detailed Client Interactions

5.3 Required Functionalities

In order to develop the Fair4Fusion system's 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
    - Conversion of 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 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 F4F 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 out the data downloads, maintain web or provenance, include initial provenance, to go back to origin)
  - 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[8], here we only focus on highlights). Towards a higher compliance with the FAIR and Open data principles following policies and practices are assumed and recommended:

[8] Deliverable D2.1: Data Inventories and Policy Landscape

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

- 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 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

In the most simplistic form, the idea of Fair4Fusion system can be depicted in a way presented in Figure 2. As it can be seen, the Experiments publish Metadata and Citation Data to the system, which collects them and exposes to the Clients. Client can Search over this data, request for Experimental Data as well as add its own Annotations to the system.

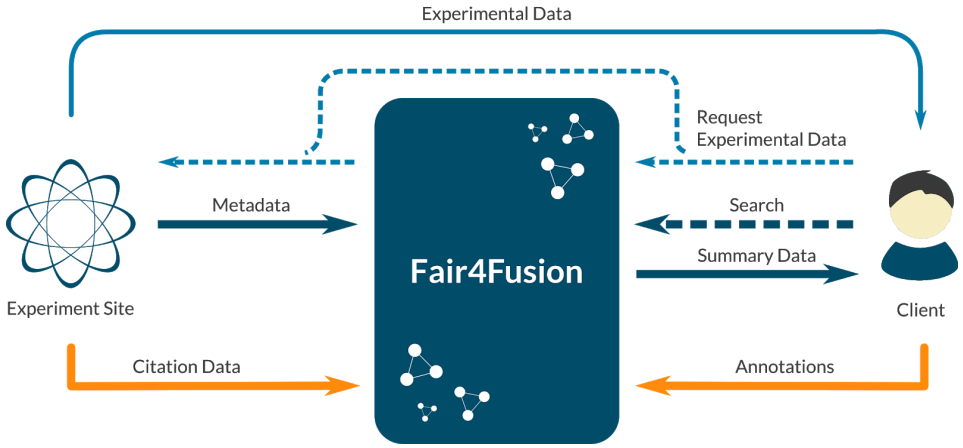

Figure 2. The generic idea of the Fair4Fusion system

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

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 Fair4Fusion architecture, 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 extended view on F4F services, user level tools and on the integration scheme between both of them. 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.

6.1 Baseline architecture

The high-level diagram of the Fair4Fusion system is presented in Figure 3.

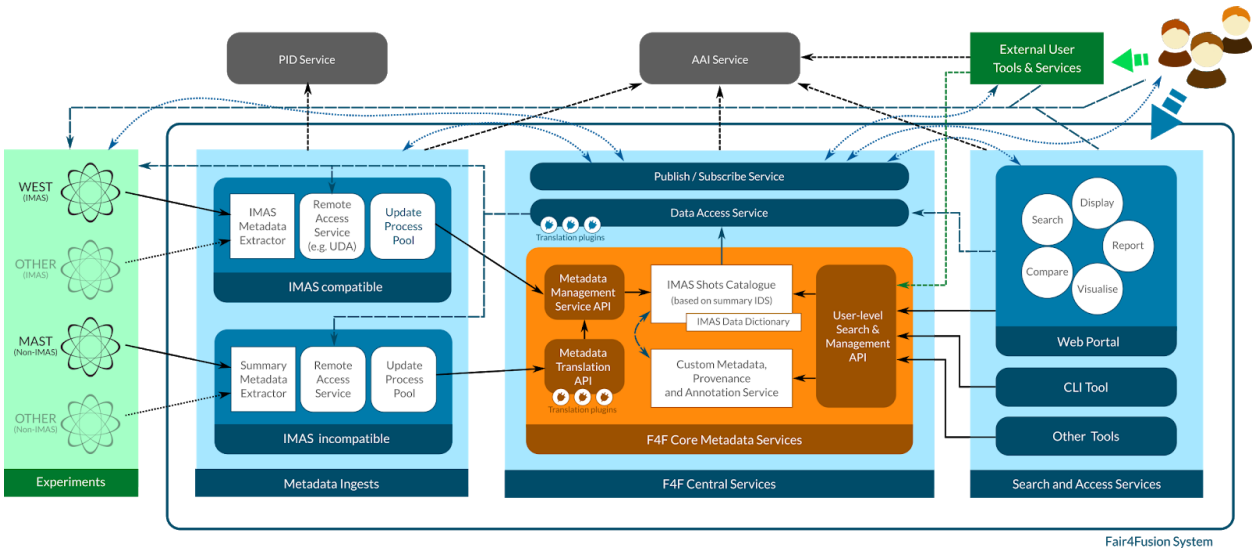

Figure 3: High-level architectural overview of Fair4Fusion system

On the most general architectural level, the whole Fair4Fusion environment can be divided into four main parts, i.e. Fusion Experiments, Metadata Ingests, Fair4Fusion Central Services and various User-level tools and services. While the last three parts constitute the integral content of the Fair4Fusion system 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.

Experiments

The experiments are EUROfusion devices 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 that is not interoperable and can't be simply reused at scale. Furthermore, 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 experiments was to treat them as black-boxes and integrate them on a higher conceptual level.

Metadata Ingests

The metadata from experiments are fed into the Fair4Fusion system through the Metadata Ingests. The role of ingests is to transfer metadata to the form which can be published to the F4F Metadata Services and its users. What is crucial and should be stressed, technically the ingests are still placed in administrative domains of specific experiments, 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 removed at this stage. Depending on a type of source data coming to the system, we can distinguish ingests operating on IMAS data and ingests operating on non-IMAS data.

Central F4F Services

Central F4F Services state a basic service layer of the proposed system. A key role is played here by a set of software components marked as **F4F 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 **IMAS Shots Catalogue**. All experiments natively supporting IMAS will be able to directly use **Metadata Management Service API** for pushing metadata. Other experiments that do not support IMAS, will need to use **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 **Custom Metadata and Annotation Service** as an additional unit. This service will be employed for the management of data pieces external to the IMAS, such as references to publications, provenance, and user annotations.

With the focus on usability, all the data managed by the F4F Metadata Services is going to be exposed to the external world via a single endpoint which will implement **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 dealing with the problem. 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 **Data Access Service** as a moderator in accessing the physical data when a user wants to get it from the experiment. The precise functionality and restrictions of the service will be described in the future updates of this document, however it is already foreseen that the service should expose data in a homogenised form and it will use a set of translation plugins to do so.

The Central F4F Services will be complemented by **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 F4F Metadata Services as well as users about updates made within the observed data collections.

Search and Access Services

The user's access to the Fair4Fusion system will be enabled primarily through a set of dedicated software components grouped in Search and Access Services. It is expected that **F4F 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 they will be permitted to 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

The open character of the Fair4Fusion system will be materialised via the opening User-level Search and Management API to the external collaborators as well as to general-public. We anticipate that the information accessible with this API could be of interest for further processing, e.g. by Data Analytics Frameworks, in order to produce more sophisticated outcomes.

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 uniquees of the data globally.

AAI

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

The system will be complemented by common Federated Authorization Authentication Infrastructure based on latest technologies, enabling easy and safe integration between components.

6.2 Architectural components

The architecture overview presented in the previous section can be already 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 Fair4Fusion system. This section will be a subject of particular extensions during the further development of the project. It is expected that a number of components can be added to the architecture and thus described in this section, but also many of existing generic components can be materialised into a form of concrete software and therefore their description will need to be amended as well.

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 4.

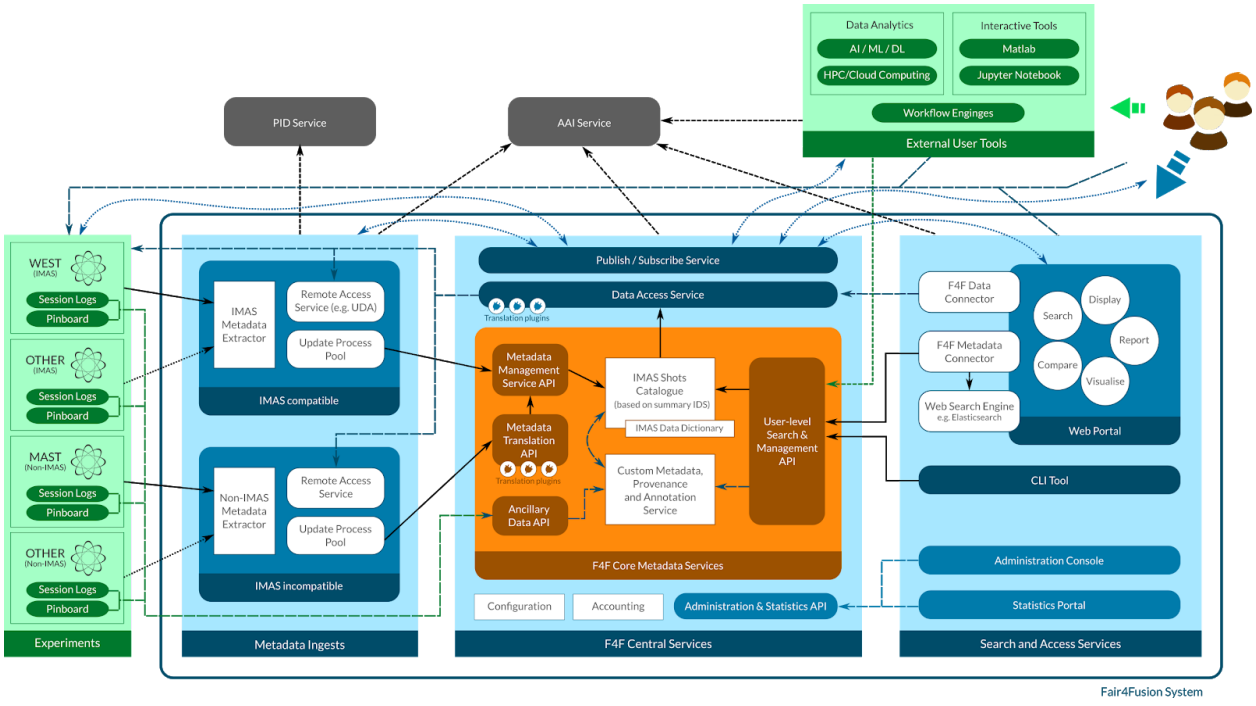

Figure 4: Architecture of the Fair4Fusion system with the detailed specification of individual components.

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

6.2.2 Experiment side components

Experiments already using IMAS format

The metadata related to shots produced by these experiments can be directly, without conversion, transferred to Central Metadata Services in the IMAS format. In this case the experiment 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 MySQL database based on the CatalogQT tool/structure). 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.

Experiments not using IMAS format

The metadata produced by these experiments needs to be converted to IMAS format in order to be processed by Central Metadata Services.

6.2.3 Metadata Ingests

The Fair4Fusion data needs to ingest metadata from the experimental sites. While the final definition of this mechanism is still to be decided there are a number scenarios which are currently being investigated for this part of the architecture. The first is whether data is *pulled* from a site as recommended in the OAI-PMH model, or whether the data is *pushed* from the site to the Fair4Fusion system. Adoption of OAI-PMH has the advantage that the APIs needed are well defined and allow interoperability with other external harvesters, but may require additional work on the part of the sites to both support the end point, and a means of separating any information which should not be made accessible through the portal. Allowing sites to push metadata will need the Fair4Fusion system to design an API which allows for site specific metadata to be pushed to a specific end point.

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 F4F 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 with Fair4Fusion, this has not been finalised yet.

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 Shots Catalogue

The central service that integrates IMAS metadata (Summary IDS) coming from different experiments. 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 Shots 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 Shots 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 Shots Catalogue and Portal.

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 can directly use this API to publish new metadata to the system as well as to update existing one. 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.

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 API.

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

This API will make use of one, or more, translator modules which will be provided 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.

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 either 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 which contains information about which datasets have been used for publication. Accessing this information is required to meet the requirement of the portal, and we will need to discuss with experiments whether APIs already exist to allow access to this information or whether one will need to be developed.

6.2.5 Fair4Fusion Central Services

Data Access Service

This service will enable automated client access to experimental data stored at particular sites. Natively it will be a proxy for the data stored in the IMAS-compatible resources, but with an extendable set of plugins it will be also ready to translate non-IMAS data to the IMAS format on-the-fly. The concrete functionality and scope of accessible data with this service will be a matter of further analysis within the project.

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 privileged users 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.

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. Web Portal will be implemented in modern software development technologies with diversification on front-end and back-end. The front-end will be responsible for the graphical presentation and user-interactions, while the back-end side will perform background tasks related to interactions with services. Conceptually, the latter will consist of the three main elements: F4F Metadata Connector, F4F Data Connector and F4F Search Engine.

F4F Metadata Connector

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

F4F Data Connector

This backend component will integrate with F4F Data services in order to enable access to experimental data stored on individual sites.

F4F Web Search Engine

In order to improve the performance of the system Web Portal will use some state-of-the-art Web Search technologies such as elasticsearch or memcached. With this component, the data accessed frequently via F4F Metadata Connector will be indexed and / or cached for further usage.

CLI Tool

Command-line interfaces remain a prefered way of accessing software systems for many scenarios and users. To support this way of interaction with the Fair4Fusion system, we are going to develop a dedicated tool and offer it to the Fusion community as an alternative to the basic part of the functionality of 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 informations 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 F4F platform, they are important F4F Central Services clients, 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 supports the scientists in their research are another example of the clients for F4F Central Services, searching and retrieving the data. Those tools are not part of the F4F platform.

Data Analysing Frameworks

Feature Extraction and Data Mining

F4F 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

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

Fusion community is making usage of many available computational infrastructures, such as HPC systems in particular dedicated system for fusion - currently Marconi@CINECA, those provided by PRACE, EoCoE, 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 of 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[9], such as eduTeams[10] (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 F4F 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 F4F components. The presented mapping is a result of both, the ongoing state of the art analysis aimed at juxtaposing existing solutions with the general F4F assumptions and the survey performed by the project to point out possible technologies for fulfilling defined F4F 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, but this is still work in progress and will be presented in detail in further reports. Based on a current findings, the specificity of the Fusion community, i.e. decentralised experimental devices having their own data, procedures and software, brings the F4F 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

[9] https://aarc-project.eu/architecture/
[10] https://eduteams.org/

only from the technological choices, but also from the political aspect of agreeing 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 components of the blueprint architecture is presented in Table 1. It should be noted that some of the components have been already incorporated into the F4F Demonstrators [D3.1]. It is expected that further analysis will allow to fine-tune this set and evaluate applicability of the components for the ultimate system.

F4F component	Technology candidates
PID Service	DOI, ePID
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 Shots Catalogue	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
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+ EOSC ecosystem: OneData, EUDAT/B2SHARE 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, BeeGFS, 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

Table 1. Mapping between F4Fcomponents and technologies

6.4 Relationship between components and services

6.4.1 Metadata Conversion

At the moment only WEST directly outputs its data 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.

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 Deny 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.

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.

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), a very important notion. 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.

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 data open.

A 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-U) and each term mapped onto the Summary IDS.

Each experiment's searchable metadata mainly focused on the physics summary parameters such as the average 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.

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. Another IDS, the “dataset_description” IDS, is used for the generic description of the dataset (outside of its physical quantities) and is a natural placeholder for additional FAIR information.

Based on the requirements we have selected a number of Dublin Core Elements to extend the existing IDS Dataset_description Schema. 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_description IDS a subset of DCMI can be selected to improve the FAIRness of the proposed fusion metadata schema.

7. Summary and next steps

This document is the first version of the blueprint architecture of the Fair4Fusion system. In the current form it already provides a comprehensive view on the design of the core part of the system, from the high-level scheme to detailed description of components, standard and protocols. The information presented in this paper is a result of in-depth analysis of the project use cases and is grounded in the vast set of collected functional requirements, current state of art as well as discovered restrictions and policies associated with the Fusion experiments.

In the future, based on further analysis and evaluation, this document will be variously extended and revised. The architecture itself will be fine tuned and expanded with a set of more specialised components, such as those for provenance in particular. The generic components of the current design will be mapped to more precise technological solutions and finally concrete software products. Not without significance for the final shape of the architecture will have the outcome from the realisation and tests of the demonstrator. It is expected that both discovered issues and successes in the realisation of the demonstrator will help to importantly advance the quality of the target architecture.

In the end, we expect that the final version of the document should be a complete source of information for the stakeholders on how to develop a production-ready system. Thus, based on experiences and lessons learned during the course of the project, we are planning to supplement the ultimate version of the blueprint with the discussion aimed to advocate particular solutions and point out possible risks in the proposed design.