



A FAIR based approach to data sharing in Europe

Downloaded from: <https://research.chalmers.se>, 2022-10-11 19:48 UTC

Citation for the original published paper (version of record):

Strand, P., Coster, D., Plociennik, M. et al (2022). A FAIR based approach to data sharing in Europe. Plasma Physics and Controlled Fusion, 64(10). <http://dx.doi.org/10.1088/1361-6587/ac8618>

N.B. When citing this work, cite the original published paper.

PAPER • OPEN ACCESS

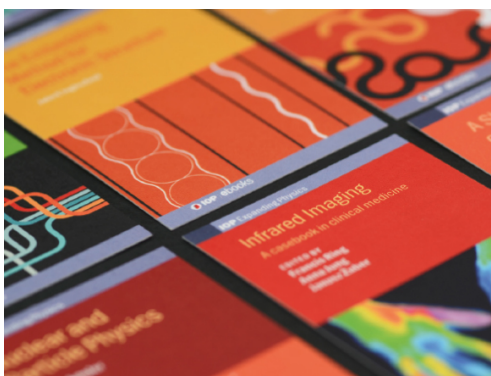
A FAIR based approach to data sharing in Europe

To cite this article: P Strand *et al* 2022 *Plasma Phys. Control. Fusion* **64** 104001

View the [article online](#) for updates and enhancements.

You may also like

- [Comparison of Doppler back-scattering and charge exchange measurements of \$E \times B\$ plasma rotation in the DIII-D tokamak under varying torque conditions](#)
Q Pratt, T Rhodes, C Chrystal *et al*.
- [DataVault: a data storage infrastructure for the Einstein Toolkit](#)
Yufeng Luo, Roland Haas, Qian Zhang *et al*.
- [Investigation of RF driver equivalent impedance in the inductively coupled SPIDER ion source](#)
Palak Jain, Mauro Recchia, Alberto Maistrello *et al*.






IOP | ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection—download the first chapter of every title for free.

A FAIR based approach to data sharing in Europe

P Strand^{1,*} , D P Coster² , M Plociennik³, S de Witt⁴, I A Klampanos⁵, J Decker⁶, F Imbeaux⁷, J F Artaud⁷, B Bosak³, N Cummings⁴, L Fleury⁷, A Ikonomopoulos⁵, S Konstantopoulos⁵, A Ludvig-Osipov¹ , P Maini⁷, J Morales⁷ and M Owsiak³

¹ Department of Space, Earth and Environment, Chalmers University of Technology, Gothenburg, Sweden

² Max Planck Institute for Plasma Physics, Garching, Germany

³ Institute of Bioorganic Chemistry PAS, Poznan Supercomputing and Networking Centre, Poznan, Poland

⁴ United Kingdom Atomic Energy Authority, Culham, United Kingdom

⁵ National Centre for Scientific Research 'Demokritos', Agia Paraskevi, Greece

⁶ Swiss Federal Institute of Technology, EPFL, Lausanne, Schweiz

⁷ CEA, IRFM, F-13108 Saint Paul lez Durance, France

E-mail: par.strand@chalmers.se

Received 24 February 2022, revised 19 July 2022

Accepted for publication 1 August 2022

Published 22 August 2022



CrossMark

Abstract

The European fusion research activities have, over recent decades, generated a vast and varied set of data. The volume and diversity of the data that need to be catalogued and annotated make the task of organising and making the data available within a broader environment very challenging. Nevertheless, there are strong scientific drivers as well as incentives and mandates from national research agencies suggesting that a more coherent approach to data referencing, dissemination and sharing would provide strong benefits to the fusion research community and beyond. Here, we discuss the technical requirements and developments needed to transition the current, and future, range of fusion research data to an open and Findable, Accessible, Interoperable, and Reusable data sharing structure guided by the principle 'as open as possible, as closed as necessary'. Here we propose a set of recommendations and technical implementations needed to form a European data sharing environment for the fusion research programmes. Consistency with the emerging IMAS (ITER Integrated Modelling and Analysis Suite) infrastructure is considered to facilitate future deployments.

Keywords: FAIR, research data, fusion

(Some figures may appear in colour only in the online journal)

1. Introduction

In the last few decades, the role and intrinsic value of research data has increased alongside data management, processing and

analysis capabilities [1]. A movement towards improved data sharing and quality management has been supported by government actors, research communities and large-scale research infrastructures and has been structured into a set of recommended practices—FAIR (Findable, Accessible, Interoperable, and Reusable) [2, 3]. Data management in most experimental facilities have largely been evolving as domestic concerns over several years based on their own national requirements, legacy practices, and technology implementation. The procedures with respect to access, provenance and quality assurance of the captured data have been largely developed to support

* Author to whom any correspondence should be addressed.



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

the internal needs in terms of formats, access protocols and internal standards. Hence, the experiments have developed a certain level of internal FAIRness for their experimental data. However, the lack of consistent definitions of physics quantities, sign conventions and data formats and the wide array of tools in use for data access as well as access and authorization of end users, mean that this internal FAIRness of data is largely lost when user access is elevated to the joint European level. For simulation and modelling data the situation is somewhat more concerning as there seem to be less consistent provenance capture and long-term data storage strategies implemented outside of the first line analysis of experimental discharges.

To address this within the scope of the fusion community the FAIR for fusion (FAIR4Fusion) project [4] was initiated. Here we discuss the technical requirements and developments needed to transition the current and future range of fusion research data to an open and FAIR data sharing structure guided by the principle ‘as open as possible, as closed as necessary’. Here, we report on the current set of recommendations and the technical developments undertaken in the FAIR4fusion project, towards providing FAIR fusion data.

1.1. Restrictions in scope

The European Fusion community has a long history of fusion research activities that mainly have developed organically via the domestic research programmes. It is only recently that it has moved towards a more project-oriented structure with a joint coordinating structure, EUROfusion [5]. Several of the EUROfusion partners maintain strong domestic research activities that remain at least partially outside of the formal EUROfusion collaboration.

In the following discussions we limit the definition of the ‘European fusion community’ to mean the activities within the EUROfusion consortium as it maintains a consistent set of agreed on rules for joint access to research results and intellectual property rights which may be more complicated with respect to data developed for the domestic projects.

On a technical level this distinction is superfluous as integration of the EUROfusion data within a facility in a beneficiary or laboratory, is trivially extended to the whole data set of that entity. In addition, focus is on modelling data and data from experimental (tokamak and stellarator) facilities where an emerging joint data description is already available through the integrated data interfaces (IDS) from the ITER (‘The Way’ in Latin) integrated modelling and analysis suite (IMAS) [6] which carries experimental data and modelling data within a single data model.

IMAS has been developed by ITER, based on original developments by EUROfusion Integrated Tokamak Modelling Task force, and consists of a ‘data dictionary’ that defines the supported variables (definitions and conventions). The data is organised in hierarchical data structures (IDSs). The IMAS infrastructure also provides for bindings of the IDSs to several programming paradigms as well as (remote) access and support for visualization and database tools.

Other types of data and research activities, in addition to modelling and experimental data, can be included provided a data ontology like the IMAS data dictionary can be furnished.

1.2. Outline

In section 2, the data and policy surveys undertaken by the Fair4Fusion consortium is analysed in view of the FAIR requirements and recommendations are provided. In section 3, a blueprint architecture is outlined that provides a proposed implementation for an FAIR based data sharing infrastructure in Europe building on existing tools and practices within current experiments and based on a minimal impact philosophy for the devices.

A key component in opening the data is to maintain and improve the provenance captured from the origins of the data collection (be it physical or digital) to the final users and usages, and an extended discussion of provenance management and implementation is provided in section 4. Section 5 discusses experiences with developing demonstrators used to encapsulate expressed user needs and the resources needed to move from the demonstrator implementation to a production level tool is indicated in section 6, ending with a closing summary section.

2. Data and policy surveys

Surveys targeting the experimental and the modelling communities have been performed exploring the current data management structures and available policies.

2.1. Experimental data

Fusion data exists in various forms, but the one that most people think of immediately is the data arising from the major fusion experiments. Each of the major fusion experiments in the EU has developed a local infrastructure to acquire and then store data gathered during a plasma discharge. They have also developed procedures for converting the raw experimental data into scientifically useful data, and to produce derived data. The challenges to make this data FAIR are: the different formats that the data are stored in, the different definitions of what is stored, the variety of tools for processing and accessing the data and the policies that the data owner imposes. A common data format, with well-established conventions, such as the IMAS/IDS paradigm, supplemented with provenance capture is needed to resolve the situation.

An assessment of the FAIRness of present experimental data management practises has been made by contacting representatives of the following European experiments: ASDEX Upgrade, COMPASS, FTU, JET, MAST, W-7X, WEST:

- **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 some quite specific

International Databases—which are not populated in a systematic way by the experiments.

- **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), partially justified by the differences in experimental diagnostic systems and data processing routines.
- **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 fusion community has no simple means to exploit it in an FAIR way. 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.

Among the European experiments, only MAST-U has presently an active Open Data policy: by default, a three year embargo period 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'. Licences for released data are not used yet by any experiment.

The following recommendation towards evolution of policies and practises to make data more FAIR are promoted:

- **Findable:** establish a central metadata catalogue, 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. We propose also to make this metadata catalogue open to the public without any embargo period, since (a) a web interface makes it easy to use for the general public so no additional effort would be needed here and (b) most scientific publications make use of both metadata and data, the latter being accessible only after an embargo period, preserving the 'publish first' capability of the experiment's team. Hence, we recommend making metadata available as soon as practicable to allow the community to discover it and, if necessary, request access.
- **Accessible:** following the querying step allowing finding data of interest, providing a single authenticated method to access data across multiple experiments secured by appropriate authorization policies at each site and, after some

embargo period, accessible even to the public (in some simplified form). The recommendation (in conjunction with the interoperability bullet below) is to use the IMAS Access Layer for this [6], although maybe through a simplified, more user-friendly interface for the public.

- **Interoperable** between various experiments (both data and metadata) by using a standard ontology (IMAS) [6]. This means mapping local ontologies to the IMAS data dictionary at some stage, before exposing it to users/public. Interoperable data must also carry information on provenance to track data processing and ontology mapping. More details about the use of IMAS for interoperability are given in section 3.2.
- **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 (see section 4).

2.2. Simulation and modelling data

Parallel to the experimental data are modelling results that are used to predict future experimental results, to interpret present experimental results, or to make predictions for future devices. Whereas the number of EU experimental fusion devices is relatively limited, the number of codes in use or under development is large, and there is little standardisation in the output formats of the codes. While it was possible to engage with each of the major experiments directly, it was decided to distribute questionnaires to code developers and to people running codes ('data producers'). Forty responses from data producers were received, and thirty-one from code developers. Some responses are shown below in figures 1–4.

Here Fair4Fusion has endorsed the recommendations of EUROfusion working groups for the creation of a Long Term Simulation Storage Facility where simulation results can be archived, and recommended that key simulation results be stored in IMAS IDSs, including the SUMMARY IDS that is the basis for making data findable. This would provide a good start to making much of the modelling data FAIR. Other fusion related data exist, and mechanisms will need to be found to make these data FAIR. This process will need to start by identifying these other categories of data, and then launching activities to increase the FAIRness of the data.

3. Blueprint architecture

The Blueprint architecture for Fusion Open Data Framework is one of the main outcomes of the Fair4Fusion project. It describes the current state of the art in terms of policies and data access and elaborates on the FAIRness status of the experiments and their repositories. Furthermore, it proposes, based on a set of user stories and requirements, architectural components along with their descriptions and technological options. It is complemented with a costs and benefits analysis, licensing options, gap analysis and an implementation roadmap. In this section we focus on community standards and

What input formats does your code use?

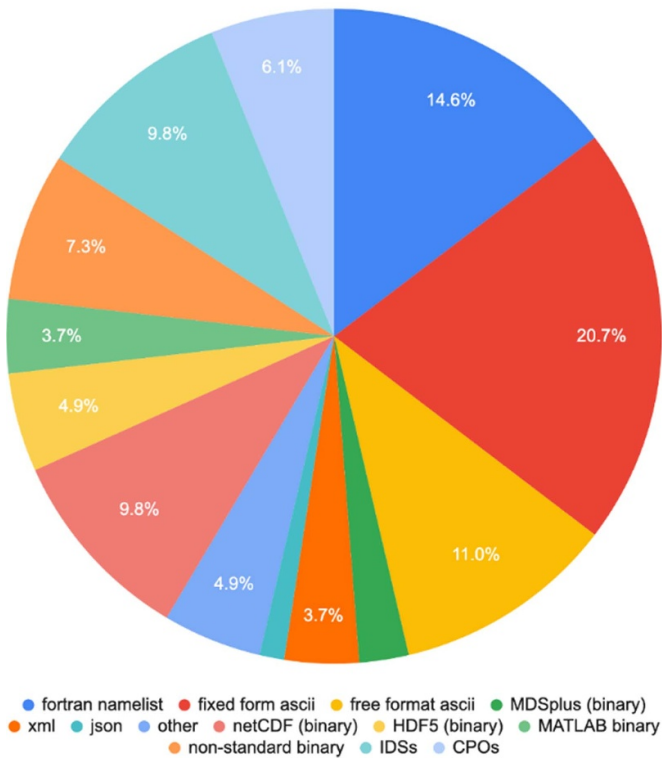


Figure 1. Characterisation of code input formats.

What outputs does your code use?

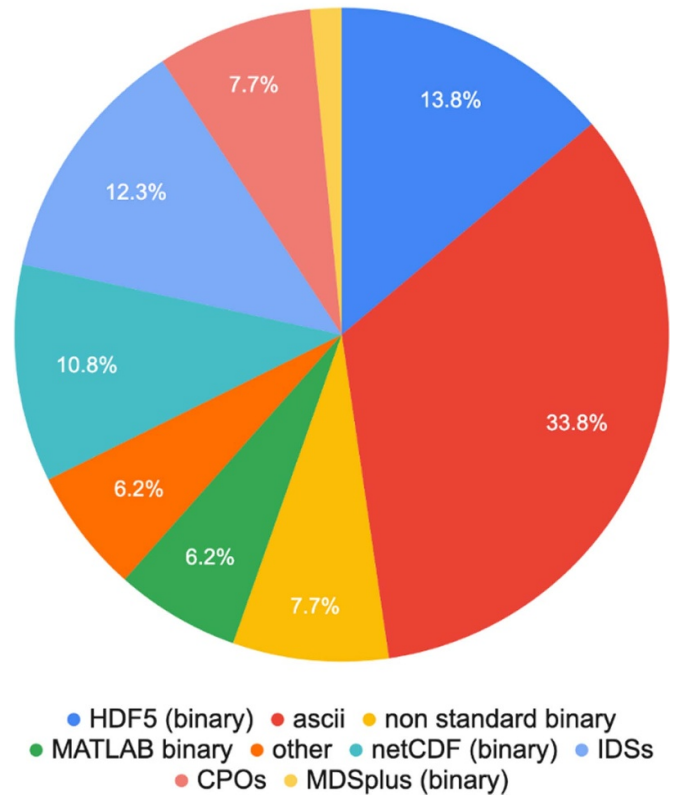


Figure 2. Characterisation of code output formats.

the proposed licensing schema, more details can be found in the Blueprint architecture report [7].

3.1. User stories and requirements

We have collected several user stories about searching for and accessing data and/or metadata, as well as user stories from the perspective of the data providers. These use cases highlight the different perspectives from the members of the public, fusion researchers and data providers. The basic requirements and user stories have been transformed into a list of desirable functionalities to be fulfilled.

These functionalities have been grouped in several categories: search, visualisation and data access, 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.

3.2. Architecture choices

In the architecture 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 an extensible machine-independent ontology, capable of covering all experiment subsystems and plasma physics
- It is the only ontology standard that has been elaborated in the fusion community
- 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 descriptions, calibration coefficients, as well as raw and processed signals), while such information may be sometimes difficult to find in present experiment databases.
- 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 useful in the context of this blueprint architecture
- It is the standard ontology for ITER scientific exploitation

Within the IMAS Data Dictionary, the Summary IDS is the place for physical metadata summarising an experimental or simulated plasma discharge. 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

Would you be interested in adding metadata to your code's output to improve the FAIRness of your code output?

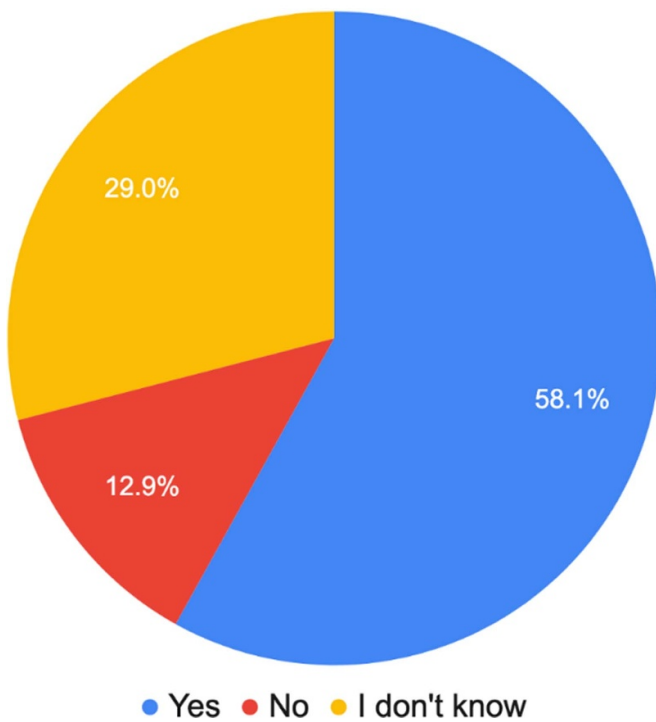


Figure 3. Attitudes to improving FAIRness of code output.

Would you be interested in making your data "open"?

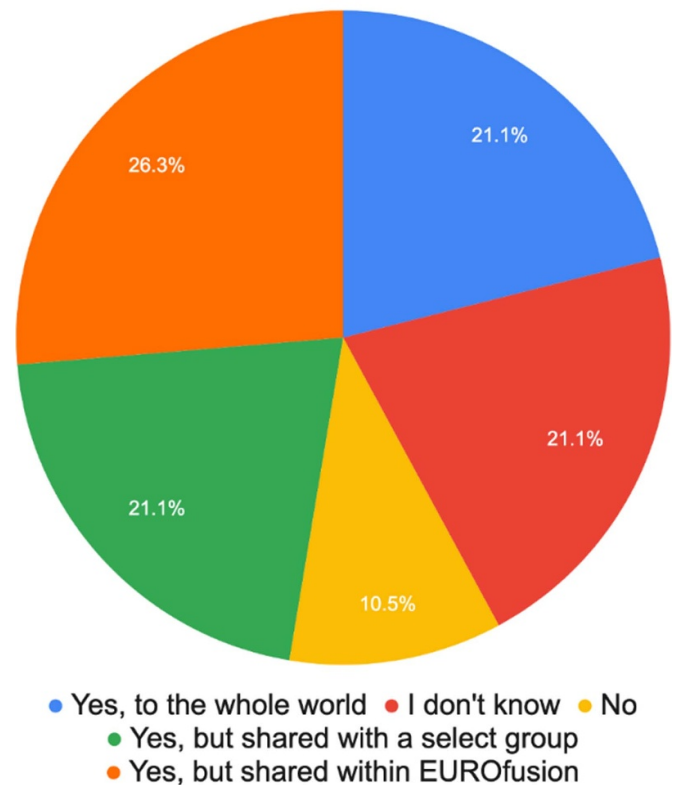


Figure 4. Attitudes to making code data 'open'.

also placeholders for error bars and basic provenance information. The Summary IDS is defined in a machine-generic way and is usable for both experiments and simulations, we propose to use this ontology as the basis for the metadata standard in order to make European fusion experiments and simulation data FAIR.

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.

3.3. Architectural overview

The proposed architecture, outlined in figure 5, consists of three 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 specific 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 can accept different formats of

metadata as input through the translation components. Central Fair4Fusion Services provide supplementary functionality for the 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 focus is given to the Web Portal that is expected to offer an extensive set of functionalities for searching, mining, filtering, or displaying metadata and data managed within the system.

Once a particular dataset has been selected as a result of queries on metadata (e.g. in the Web Portal), the Data Access Service will enable automated client access to the corresponding data. This service is the final gate for users of the Fair4Fusion services to the original data that is referenced (by metadata) in the Fair4Fusion portal. This data will not reside within the Fair4Fusion services but remain at the originating sites (e.g. experiment sites), therefore remote connections will have to be open to transfer data on the fly (upon user request).

Depending on local policy, the requested data will not necessarily be open and accessible directly via the FAIR4Fusion portal. In such a case, the minimal requirement is that the data access service will provide the instructions for accessing available datasets, 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

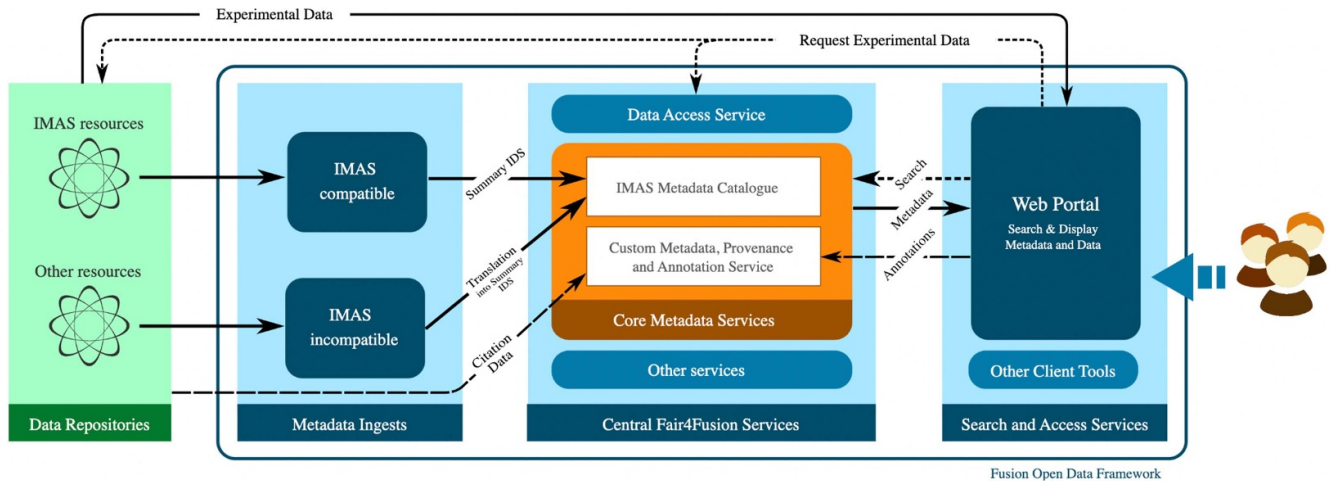


Figure 5. System diagram for the proposed FAIR fusion portal.

case of non-open data) would be to embed an authentication mechanism in the data access service, so that the user of the FAIR4Fusion service can authenticate themselves towards the data server.

The system is envisaged to be complemented by a common Federated Authorization Authentication Infrastructure based on the latest technologies, following the Authentication and Authorisation for Research and Collaboration blueprint architecture, such as eduTeams (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.

3.4. Proposed extension of community standards

The Summary IDS provides a large coverage of the physics quantities that can be captured in fusion experiments but does not in general include more generic data entities which could help make the data more findable and accessible to non-fusion users, including funders, other researchers, and the public. We have thus decided to extend the Data Dictionary with additional FAIR information. A dedicated Dataset FAIR 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 several Dublin Core Elements to include in this new IDS.

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.

3.5. The licencing schema

The use of a creative common (CC) licence for opening validated fusion data and metadata to the public and wider research community is recommended. An embargo period of a few years should be given for data to provide sufficient time for the Institutes running the experiment or collaborating with it to exploit the data first. Data used for publication should be released at the time of publication.

The precise combination of CCs flavour to be chosen is left to each data owner (experiments, modellers, etc), but the Fair4Fusion project recommends using CC-BY-NC-SA: this licence allows users to distribute, remix, adapt, and build upon the material in any medium or format for non-commercial purposes only with derived data distributed under the same licence, and it is a requirement to give attribution to the creator. It is a good trade-off in terms of openness and fair usage of the data produced by a huge effort based on public funds. Looking into the attributes:

- BY (attribution): requires citing the data creator.
- SA (share-alike): requires that, if derived data is produced and exploited, it has to be published with the same licence as the original data. This would also allow the experiments to benefit from any adaptation of their own data under the same licence.
- NC (non-commercial): prevents commercial usage of the data and of its adaptations. 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.

It should also be noted that if different data producers use different licences, it will create difficulty when trying to combine data for the benefit of the community. For strategic, commercial aspects, this licence does not preclude any site from partnering with industry and sharing data with them under a

bespoke licence or even charging for data access requested by commercial entities.

4. Provenance

4.1. Provenance in fusion

The current situation in fusion is that aspects of provenance are recorded in myriad ways across the various sites. Through a combination of digital and analogue methods, provenance data is captured in ad-hoc formats, describing aspects of a given signal's history to various degrees of detail. Additionally, this information is usually not accessible via any standard protocol, e.g. MDSplus [8]. Consequently, an objective comparison between sites as to how effectively provenance is recorded is difficult to make.

The FAIR4Fusion project looked in detail at the provenance capture of MAST/MAST-U and found that it was possible to extract some provenance information from the session log files, but these files are not in any standard format, meaning that the extraction of provenance information had to be done manually, or by parsing the logs with regular expressions, a brittle method with often unexpected results. The amount and type of provenance information available varied depending on the signal, as analysis codes are provided by the various responsible officers and diagnosticians, who each document and log their part of the signal processing chain in their own way.

Many of these issues are, to various extents, addressed by IMAS, by providing within its data dictionary places to record provenance information within IDSs. These can be filled manually or programmatically, but as free-text fields, it can still present a challenge to unify the numerous ways that they could be (and are) filled. There is also an issue, in that an IDS in isolation can be left disconnected from the other data from the same experiment/simulation, without a reference to the top-level process that generated the data, be that a particular shot, or a run of some integrated modelling workflow. Recent versions of IMAS have improved upon this, featuring additions to the data dictionary that allow data producers to describe the provenance of their IDS in more detail, with reference to the top-level data entry.

Some questions that are important for data consumers to ask themselves, as their answers can speak to the reproducibility of their analysis, are:

- Can any analysed signal that you are working with be reproduced from just the raw data, and provenance metadata?
- Do you always know what raw signals the data you are analysing come from?
- Do you have access to those raw signals?
- What do you know about the calibration of the device that recorded the raw signal?
- Is the code that processed that raw signal findable and accessible? Is it under version control and is the version stored within provenance information?

4.2. The W3C-PROV standard for provenance

There already exists a standard model for capturing and reporting on provenance chains, called W3C-PROV [9] (hereafter referred to as PROV). The standard defines an ontology capable of describing provenance for any given entity and is well suited to describing the provenance of data. The core of the model is composed of 'Entities', 'Agents' and 'Activities', with well-defined edges linking these objects.

A key strength of this model, besides its ubiquity and high level of generalisability, is that it supports serialisation in formats that are both machine *and* human readable. This means that provenance capture and propagation can be automated programmatically, and manual scrutiny is possible using tools provided by the PROV developers.

A number of other disciplines and datasets already make use of PROV either directly (e.g. DBpedia [10] and OECD Linked data [11]) or using domain specific extensions (e.g. SEIS-PROV for seismology [12]). As of 2013, over 60 datasets or frameworks implemented the PROV model. Indeed recently, the F-UJI tool [13] developed under the FAIRsFAIR [14] project assumes the use of a machine-readable version of PROV (or PAV [15] which is just a specialisation of PROV) when assessing a datasets compliance to FAIR principles, and the DARE [16] platform provides support for automating the output of PROV formatted provenance.

The adoption of this standard in the fusion community would significantly improve the trust and reliability attributed to its research, as well as fostering a culture of best practice, as the PROV model is designed such that the granularity of provenance description can be iteratively refined. A simple description linking data products as 'entities' to its ancestors can be implemented, then enriched over time, with a more richly defined attribution of responsibility emerging over time.

4.3. Fusionprov—a demonstration and exploration of applying PROV to fusion data

As part of the efforts to explore and summarise the state of provenance capture and report in fusion, the FAIR4Fusion project has developed a tool, *fusionprov* that generates provenance reports from existing fusion data. It supports both MAST/MAST-U data and IDSs.

The tool provides an interface that takes a given dataset or signal and using its own awareness of the ontology and locations of data as defined for both MAST/-U and IMAS data, retrieves the relevant information and builds a W3C-PROV compliant provenance document for the input signal or IDS, by calling the *prov* python package, provided by the PROV developers. *fusionprov* is itself a python package, registered with the Python Package Index and can be installed using python's package manager, *pip*.

A significant caveat for this tool is that for MAST/-U data and IDSs, the tool can only retrieve the data it needs within a carefully curated environment. In practice, this means that the user will need to be running the tool from one of a very limited number of locations, as the data access server utilises an IP-whitelisting security mechanism. This is one of

the most notable findings of the provenance element of the FAIR4Fusion project, that while data is in some cases ostensibly ‘open’, the tools used to retrieve the data fall short of the standards necessary to describe the data as truly ‘findable’ or ‘accessible’. If data is to be presented as open, it needs to be retrievable from a publicly facing server, with a well-documented and well-maintained suite of access APIs.

5. Demonstrator

Within the course of the FAIR4Fusion project two demonstrators addressing different aspects of the use cases that were collected during the initial phases of the project were implemented and assessed. For technical testing purposes a database with low resolution, but representative data in the IDS-summary format was set up and used for assessing the demonstrators. An example of searchable variables from the test-database can be seen in figure 6, where the output of a process identifying flattops and visualizing plasma characteristics (confinement factor H_{98} , confinement time τ_E plasma current I_p , magnetic field B_T , and heating powers (total, NBI, ICRH and ohmic power) P_{tot} , P_{NBI} , P_{ICRH} and P_{Ohm} is shown as functions on time. An identified flattop is shown by the light green area in the top subplot.

5.1. Demonstrator I

The main goal of Demonstrator I was to provide reference implementation of the Blueprint in the following areas: integration of Demonstrator I with IMAS framework, remote access to experimental data using Uniform Data Access layer (UDA) of IMAS, proof of concept for Jupyter Notebook based development and direct data access from Python scripts, integration with AAI platform, and Docker based deployment.

Because the IMAS platform defines a placeholder for metadata—the Summary IDS—it was a natural choice to base data ingestion on these structures. In addition to that, FAIR4Fusion project defined one more IDS—Dataset Fair—for storing metadata describing the source of the information in greater detail. Based on these two sources of information, Demonstrator I collect, transforms, stores and then presents data to final users. Due to the fact both IDSes—Summary and Dataset Fair—provide distilled information extracted from experimental data, it is possible to run uniform comparison of data coming from different sources. Demonstrator I was able to extract this information from both local and remote data sources. Local data access was based on MDSPlus format, while remote access was realised using UDA services. As the development has been based on IMAS framework, data access itself is very similar to the way regular developers access IDS based data. Thanks to its modular approach, it is possible to extend Demonstrator I’s data source list by implementing custom plugins. It means that IMAS is not a limitation here, it rather serves as a reference, and different data sources are possible. Remote data access was realised using UDA technology.

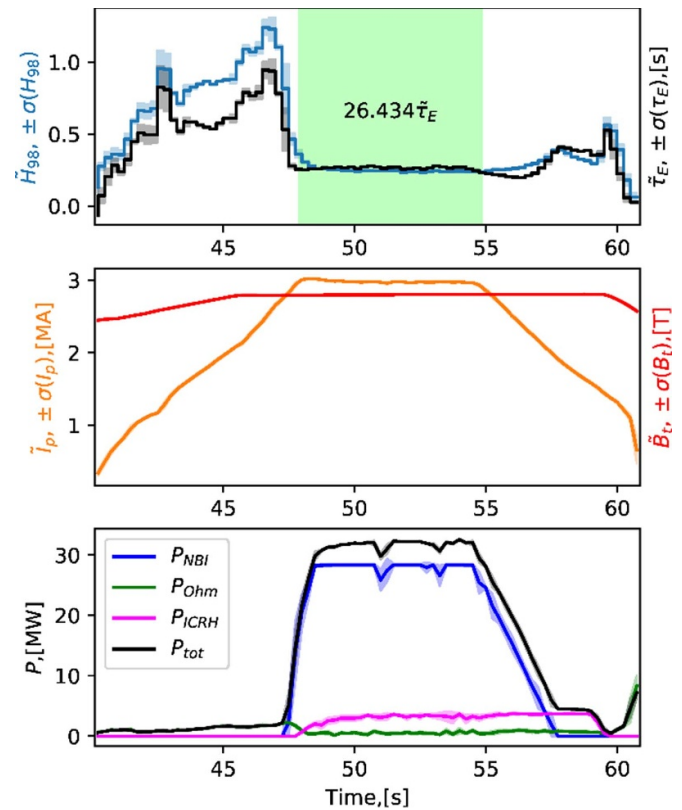


Figure 6. An example of a test-database entry.

UDA allows the use of remote data (stored in IMAS based format) directly from the code. From a developer’s perspective, there is not much difference between direct access to local and remote data. Thanks to this approach, it was possible to provide users of Demonstrator I with a Jupyter Notebook based solution where not only Summary or Dataset Fair data are available, but the whole content of the pulse file that was stored inside Catalog QT 2. Jupyter Notebook provided as part of FAIR4Fusion based solution can be either incorporated into Catalog QT 2 or can serve as a separate application. In both cases it provides a Python based template for accessing remote data via UDA.

One of the limitations, when it comes to accessing remote data, was authentication and authorization. There is no single, unified, way of accessing information regarding experimental data. Inside Demonstrator I we have applied an AAI solution for user authentication and authorization. It was successfully integrated with eduTEAMS Identity Provider. This way, it was possible to share Catalog QT 2 with any participant of eduTEAMS initiative. Authentication and authorization mechanisms were developed using a well-established software component—Keycloak. It is an Open Source identity platform that provides support for several authorization services.

Demonstrator I provides not only backend services (ingestion of data, storage, IMAS integration) it also provides frontend services (figure 7). There are different ways of accessing

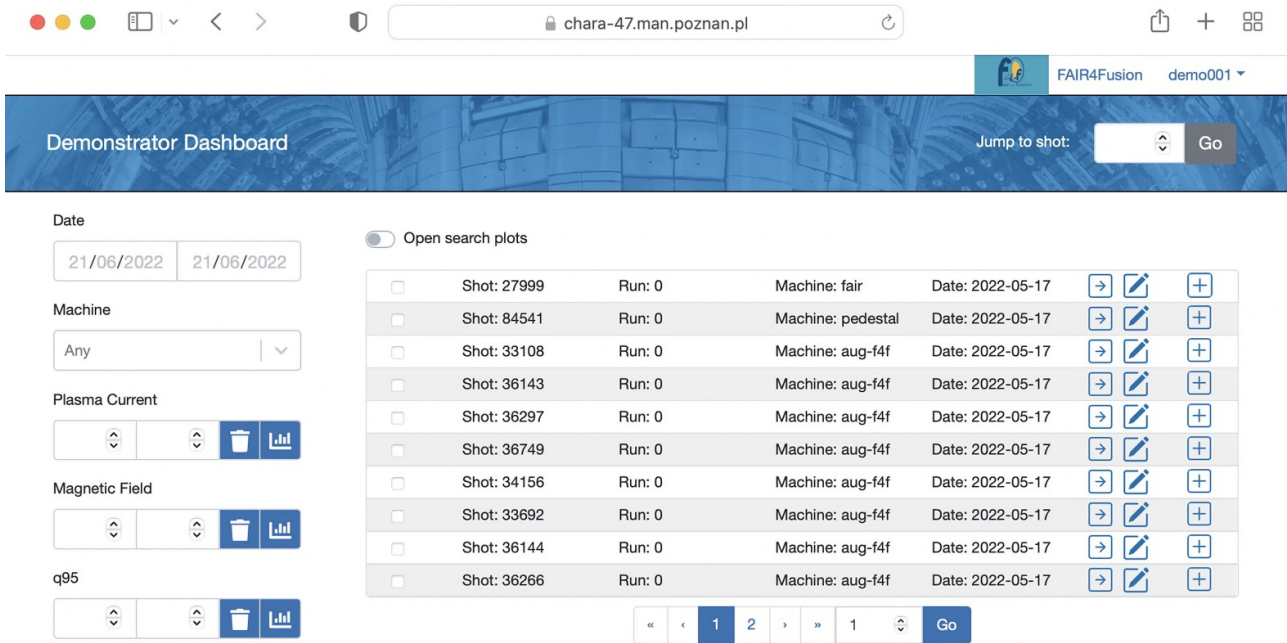


Figure 7. Interface to the fusion FAIR data portal.

data stored inside Catalog QT 2. ReactJS based frontend is the most obvious way of accessing data stored inside Catalog QT 2. It provides users with a user friendly, GUI based approach executed directly inside the browser.

However, it is not the only way data can be retrieved from Catalog QT 2. Users can also use CLI based client and Web Service client. Catalog QT 2 backend is a Web Service based solution with well-defined API documented using a tool called Swagger—de facto standard for Web Services documentation. This way, it is possible to integrate Catalog QT 2 into a custom pipeline where searching for suitable data is one of the steps of computations.

Demonstrator I was designed in a modular, extensible way, to some extent based on micro services. Each and every component can be installed separately. Communication between components is realised using Web Services. However, it is also possible to use Docker based installation that binds all components together using a docker-compose based approach. Different elements are combined into a single service. Docker based solution uses IMAS Docker—a Docker image that provides a minimal working IMAS environment. It consists of elements that are essential for running IMAS enabled codes. It is worth noting that IMAS Docker was initially developed for a different purpose, but its modular nature allowed it to be used as it was inside Demonstrator I and benefit from its ability to operate on IMAS based data.

Demonstrator I, as a reference implementation, proves that components used inside the implementation (Docker, IMAS, UDA, Jupyter Notebook) can be successfully linked together while at the same time they can serve different roles in a different environment. Thanks to applying modularity it was possible not only to extract experimental data but also present it in user friendly and appealing form.

5.2. 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 visualising shots and experiment results.

The backend explores the integration of the following elements:

- Containerization technologies, especially in the context of FAIR sharing of complete computational experiments besides experiment data.
- Cloud computing for distributed execution of computational experiments. In particular, using concepts and software from R&D work on workflows to orchestrate the execution of computational experiments defined as pipelines of containerized tools.
- Interoperability between alternative representations for metadata and summary IDS, achieved via transformers available for inclusion in pipelines.
- Following FAIR and open science principles by maintaining metadata about the provenance of the shots processed, as well as links between data, experiments that process them, and publications that describe these experiments
- Integrating the management of data, container, and computational resources with Keycloak, so that role-based access encompasses all elements of the backend.

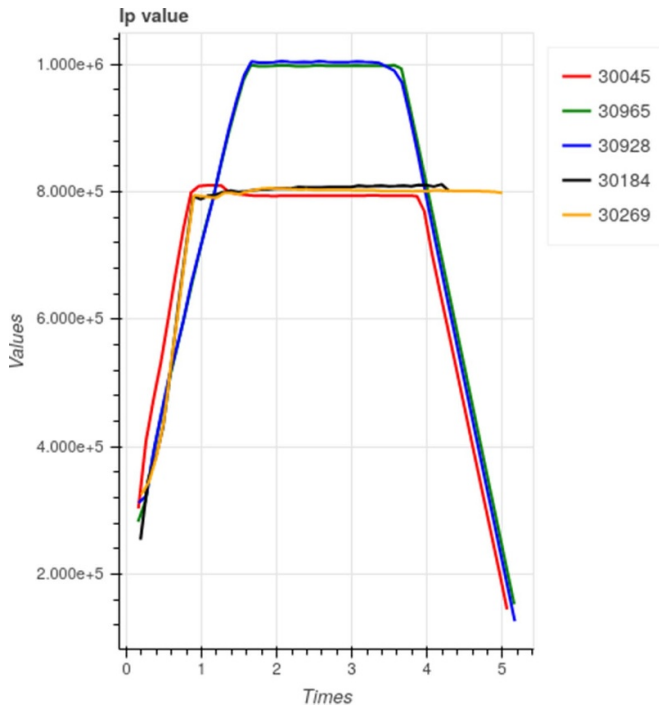


Figure 8. Example of output form demonstrator II.

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 [17], 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. 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.

The frontend is a Django application that offers new and flexible solutions for visualisation and analytics, such as Python/bokeh (see figure 8, overlaying similar shots). It also uses Elasticsearch to store metadata, allowing users to search, display and augment metadata, and to display analytical and statistical details of the pulse variables. Pulse data can be exported as raw JSON data as well as visualised charts and images.

6. From demonstrators to production

The demonstrators are implemented as building blocks towards an implementation of a production facility and not turnkey solutions. A full-scale implementation of the Fair4Fusion Blueprint Architecture requires some further resources to be made available at scale. Hardware resources for hosting of the central portal facilities require frontend/backend hardware and metadata storage facilities to be supported and

AAI and PID services to be implemented. If simulation data is to be incorporated as a facility on its own, a need for a Long Term Simulation Storage facility is needed (estimated at petabyte size).

The main investment is however in manpower: Support for central services is estimated to be of the order of 2 FTE/year (including the cost of moving from demonstrator to a hardened production environment with additional features as well as the longer-term maintenance and user support). Site services are designed to be lightweight and non-intrusive to operating resources but require 0.5 to 2FTE/year per site depending on the current level of FAIRness and level IMAS adaptation already available.

7. Summary

Extended and improved data sharing within the EU fusion community can be built by extending current data services and installations with additional practices and a set of new software technologies, together with a limited investment of hardware resources for longer term storage. The proposed implementation, detailed in the Blueprint Architecture, proposes an implementation on top of existing facilities minimizing the impact on current working practices and assumes modest extensions on top of the existing facilities

For experiments remote access between the central facilities (metadata portal and e.g. a central data dashboard) and each of the facilities need to be established together with mapping tools from the local systems to the IMAS based metadata (data) formats of the portal services. For modelling and simulation data outside of the analysis done at the experiments, there is no common storage system and a long term storage facility for simulation data need to be installed to support the modelling community. In both cases the provenance capture of the data will need to be structured and the Fair for Fusion project have been proposing and supporting extension to the IMAS data definitions to facilitate that, that are now included in the IMAS data dictionary.

The proposed infrastructure can be scaled from a system supported in a single lab, to serving a national level structure or as presented here the EU fusion community and beyond.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Acknowledgments

This project has received funding from the European Community's Horizon 2020 Framework Programme under Grant Agreement 847612.

ORCID iDsP Strand  <https://orcid.org/0000-0002-8899-2598>D P Coster  <https://orcid.org/0000-0002-2470-9706>A Ludvig-Osipov  <https://orcid.org/0000-0002-7057-6414>**References**

- [1] Editorial, Nature 2009 Data's shameful neglect *Nature* **461** 145
- [2] G20 leaders 2016 *G20 Leaders' Communique Hangzhou Summit* (available at: https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967)
- [3] Wilkinson M D *et al* 2016 The FAIR guiding principles for scientific data management and stewardship *Sci. Data* **3** 160018
- [4] Fair for Fusion 2019 Euratom research and training programme Grant Agreement No. 847612 (available at: www.fair4fusion.eu)
- [5] EUROfusion 2014 Euratom research and training programme Grant Agreement No. 101052200 (available at: www.euro-fusion.org)
- [6] Imbeaux F *et al* 2015 Design and first applications of the ITER integrated modelling & analysis suite *Nucl. Fusion* **55** 123006
- [7] Płóciennik M *et al* 2022 Blueprint architecture for a fusion open data framework (1.6) (Zenodo) (<https://doi.org/10.5281/zenodo.6759119>)
- [8] Stillerman J, Fredian T W, Klare K A and Manduchi G 1997 MDSplus data acquisition system *Rev. Sci. Instrum.* **68** 939
- [9] W3C-PROV *An Overview of the PROV Family of Documents* (available at: www.w3.org/TR/prov-overview/)
- [10] DBpedia *Global and Unified Access to Knowledge Graphs* (available at: www.dbpedia.org)
- [11] OECD linked data *OECD Data* (available at: <https://data.oecd.org>)
- [12] SEIS-PROV *SEIS-PROV: Provenance for Seismological Data* (available at: <http://seismicdata.github.io/SEIS-PROV/>)
- [13] F-UJI *Automated FAIR Data Assessment Tool* (available at: www.f-uji.net)
- [14] FAIRsfair *Fostering FAIR Data Practices in Europe* (available at: www.fairsfair.eu)
- [15] PAV *PAV - Provenance, Authoring and Versioning* (available at: <https://pav-ontology.github.io/pav/>)
- [16] Klampanos I, Themeli C, Spinuso A, Filgueira R, Atkinson M, Gemünd A and Karkaletsis V 2020 DARE platform: a developer-friendly and self-optimising workflows-as-a-service framework for e-science on the cloud *J. Open Source Softw.* **5** 2664
- [17] Iannone F *et al* 2018 Marconi-fusion: the new high performance computing facility for European nuclear fusion modelling *Fusion Eng. Des.* **129** 354–8