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”

D2.1 - Data Inventories and Policy Landscape

WP2

Deliverable status	Final
Type	Report
Dissemination level (according to the proposal)	Public
Work Package	WP2 - Policy and use Case Definition
Lead Beneficiary (deliverable)	4 - CEA
Due Date	29/2/2020
Date of submission	27/2/2020

Project Name:	Fair4fusion – open access for fusion data in Europe
Grant Agreement:	847612
Project Duration:	1 September 2019 – 31 August 2021

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

Document Information

AUTHORS

Authors	Organisation	Contact (e-mail, phone)
F. Imbeaux	CEA	Frederic.imbeaux@cea.Fr
D. Coster	IPP	David.coster@ipp.mpg.de
P. Strand	CTH	Par.strand@chalmers.se
J. Decker	EPFL	Joan.decker@epfl.ch
S. De Witt	CCFE	Shaun.de-witt@ukaea.uk

DOCUMENT CONTROL

Document version	Date	Author/Reviewer – Organisation	Change
V1	10/02/2020	F. Imbeaux (CEA, author)	First version
V2	19/02/2020	Y. Martin (EPFL, reviewer) : comments+corrections, taken into account by F. Imbeaux (CEA)	Second version
V3	20/02/2020	M. Plociennik (PSNC, reviewer) : comments+corrections, taken into account by F. Imbeaux (CEA) + completed table 1	Third version
V4	21/02/2020	F. Imbeaux (CEA) : clean up	Fourth version
V5 – Final	26/02/2020	P. Strand (CTH) : addition to Introduction, F. Imbeaux (CEA) : final clean up	Final version

DOCUMENT DATA

Keywords	Open Data
Point of contact	Name: F. Imbeaux Partner: CEA Address: IRFM, CEA, 13108 Saint Paul lez Durance, France Phone: +33 4 42 25 63 26 E-mail: Frederic.imbeaux@cea.Fr
Delivery date	February 29, 2020

Contents

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
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

Executive Summary

This report presents a summary of existing experimental data holdings together with current access policies in European experiments. It also presents an assessment of their compliance with FAIR Principles and makes suggestions for how to improve the present practices/policies towards a more FAIR and more Open Data management.

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. 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. 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. In particular, Data Mining / Machine Learning activities cannot be conducted across multiple experiments, or would require the ad-hoc creation of specific databases. A few international multi-machine databases have been created in the last decades of fusion research but their perimeter is limited to specific physics topics (e.g. confinement, disruptions, …) and they are not fed on an automated/systematic basis. In addition to improving the EU fusion science community Open Science and FAIR practices, making metadata and data interoperable across EU experiments is a key target of our recommendations since it would bring unique benefits to the EU fusion research, increasing the potential for new discoveries. The IMAS Data Dictionary is recommended as the standard ontology for achieving interoperability.

1 Introduction

This report presents a summary of existing experimental data holdings together with current access policies in European experiments. It also presents an assessment of their compliance with FAIR Principles and makes suggestions for how to improve the present practices/policies towards a more FAIR and more Open Data management.

A large part of the material used to carry out our analysis was initially gathered via a questionnaire that was sent end of 2017 to several EU experiments by the EUROfusion Working Group on Open Data Access. These questionnaires have been analysed and completed by other targeted questions to produce the present overview.

EUROfusion is a joint co-fund project under the EURATOM umbrella that currently organises all 28 European member states, with Switzerland as an additional associate member. It is supporting both the Joint European Torus (JET) where it is responsible for the scientific programme and collaborative research on a range of national programmes. A strategy for open data access is being developed as part of the work programme for Horizon Europe, and the FAIR4Fusion blueprint forms an important part of the framework. A continuous exchange of information between the FAIR4Fusion group and

EUROfusion is structured through the General Assembly and associated working groups. EUROfusion is defining its work in relation to "*Fusion Electricity – A roadmap to the realisation of fusion energy".*[1]

The following European experiments are addressed in this report, which represent a fairly broad panel: ASDEX Upgrade (AUG), COMPASS, FTU, JET, MAST, TCV, WEST, W-7X. All are tokamaks, except W-7X which is a stellerator. The ISTTOK experiment has been also contacted but didn't respond to our queries.

2 Present practices in EUROfusion experiments

2.1 Data policies

All European tokamak and stellarator experiments grant access to their measured and processed data to external fusion researchers who are formally identified as members of the experiment's team, so data access is granted on an individual basis. 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. The most likely reason for this is that it would be quite difficult to implement practically in a fair way : if an experiment is funded by a specific organization (e.g. EUROfusion) but a diagnostic (or heating system) used in that experiment has been provided in-kind by a different collaboration, how to manage the access rights to the data ?

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. It essentially dictates that the content of the publication must be cleared by the experiment's Task Forces (and EUROfusion when it funds specifically a set of pulses used in the publication, or when PhD students are involved). This is often a step during which the quality of the data

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

used in the publication is further checked, leading in some cases to additional analysis and data reprocessing.

2.2 Data ontologies and access tools

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, to fulfil though very similar functionalities (data storage, data access, data model documentation, cataloguing and browsing of metadata). It clearly creates an intrinsic limit to interoperability between the European experiments. To remedy this, WEST made all its processed data and part of the measured data accessible via IMAS, the standard promoted by ITER Organization.

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 of the nature of the data, e.g. JPF (JET Pulse File) and PPF (Processed Pulse File) for respectively raw and processed data at JET. W-7X uses another system, namely a web-service based API serving data to users in JSON format. Data is uniquely addressed via a URL.

Remote data access is often provided via the MDS+ technology used as a client/server architecture on top of the native database (AUG, TCV, JET). AUG also uses 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.

Experiment	Data model and Access Method	Metadata and querying tool/portal	Remote Data Access and related authentication
WEST	IMAS Data Dictionary and Access Layer	Metadata based on IMAS Summary IDS is dumped into a SQL database. An IMAS compatible web-based querying tool allows formulating complex queries to the above SQL database	IMAS/UDA. Requires a local user account on the remote machine, opens an SSH tunnel after authentication
AUG	Custom	Custom	AFS, MDS+
JET	Custom, raw data gathered in “JPF” tree, processed data in “PPF” tree	Custom metadata, JET Dashboard	MDS+, SAL
MAST	Custom	Custom metadata, MAST Data Dashboard, Open data web page	IDAM/UDA (data download for open data)
TCV	Custom model, stored as MDS+ trees. Access via MDSIP	ALMA database accessed via TCV logbook	MDS+ (ssh+mdsip using LDAP server)
W-7X	Custom, web-service based API serving data to users in JSON format	Websites for browsing manually every signal. In addition, a separate website provides access to the archived experiment program logs	Only via VPN (Remote Computer Access)
FTU	Custom	None, data is simply documented in a Wiki page	AFS, MDS+
COMPASS	Custom, data stored in NetCDF4, HDF5, and JSON formats	Metadata are obtained from MySQL DB, there is also a web front end with some basic visualization options	Only via VPN or SSHFS+SSH tunnel/ proxy (Remote Computer Access)

Table 1 : Overview of the various data infrastructures of European experiments

2.3 Provenance capture

At each experiment site, a Plasma Reconstruction Chain (PRC) processes the raw data to obtain and store “processed data”. This step is documented and processed data contains provenance information for the sake of traceability/reusability. Typically, the name and versions of the codes used in the PRC are recorded with the output data of the PRC (WEST, MAST), which links unambiguously the output data to the source code that has produced it since the source code of the PRC components is under version control (SVN or GIT). The PRC is a typical data-driven workflow and thus it is interesting to record pointers to all data on which the processing depends which is done at AUG.

For a given plasma pulse there may exist multiple versions of the processed data. They are created as improvements in the data processing are made, and are all maintained available for the sake of reproducibility/traceability of further analysis. So even if neither DOI nor PID (persistent identifier) is formally attributed to data, each version of the processed data remains unique (within a given experiment), accessible and reproducible.

2.4 Openness, data licensing

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”. This policy is typically what could be recommended to all experiments, although the embargo period may seem a bit long.

Licenses for released data are not used yet by any experiment.

2.5 Contributions to International Databases

EU experiments contribute to international databases, for instance in the framework of the International Physics Tokamak Activities (ITPA) [https://www.iter.org/org/team/fst/itpa]. These databases often have a European counter-part organized by EUROfusion. The international databases are focusing on specific physics topics (e.g. confinement, disruptions, L-H threshold, pedestal, …) and thus their perimeter is limited to the physics quantities of relevance for their topic. The contribution to these databases is furthermore neither automated nor done on a systematic basis, instead careful data selection and validation is carried out by the experiment before any data submission. Although each of these databases is gathering data from multiple experiments using its own experiment-generic ontology, international databases are often not interoperable yet. ITER Organization plans to make them interoperable by using the IMAS Data Dictionary but this has not happened yet.

2.6 FAIRness assessment

We now assess the compliance of present practices related to experimental and processed data with FAIR Principles. Experimental and processed data 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 to find only the data of that experiment. There is no central metadata catalogue that would allow multi-machine searches, apart from the International Databases mentioned in 2.5

- 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. 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.

3 Proposals for enhancing FAIR and Open data practices among EU experiments

3.1 Recommended improvements

Towards a higher compliance with the FAIR and Open data principles, we recommend the following evolution of policies and practices to make data more:

- 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 i) a web interface makes it easy to use for the general public so no additional effort would be needed here and ii) 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 .
- Accessible: following the querying step allowing finding data of interest, 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). The recommendation (in conjunction with the interoperability bullet below) is to use the IMAS Access Layer for this, 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). 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.

It should be made clear that making data more FAIR than it presently is doesn't remove the principle of clearance before a publication: it's just another way of finding and accessing the data. It has to be made explicit in the data's license that the experiment's clearance procedure for using the data in any kind of publication (course, reports, these, conference presentations, articles, website, …), by any author, including general public, remains necessary.

These recommendations address Green data: only non-confidential, non-commercial and non-patentable data are targeted by these recommendations. CAD data is outside of the perimeter of the recommendations, mostly because there its IPR is in general owned by the experiment's operating Institute and because there is not much interest in sharing it across experiments. Note however that it appears in a simplified form as “machine description” in the data typically used for physical processing.

3.2 Using IMAS Data Dictionary as the ontology

The main arguments for recommending the use of IMAS as a standard ontology for making data and metadata interoperable across the various EU experiments are:

- 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
- 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

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

3.3 A pragmatic approach to FAIR and Open Data practices

There is clearly flexibility in the way a given experiment may implement the recommendations made in section 3.1:

- Data coverage may not be exhaustive: only a selection of data from the experiment's database can be mapped to IMAS and be made accessible to the EU fusion researcher community, although we would recommend to progressively broaden the range of accessible data to all quantities that are useful to perform research on an experiment. An even more limited data selection may be fully open to the public, although it's not in the spirit of Open Science. As a minimal initial dataset to be made accessible in IMAS, we recommend to address the quantities covered by the equilibrium and core_profiles Interface Data Structures (IDS) of the IMAS Data Dictionary, which describe key core plasma quantities, and for metadata the quantities covered by the Summary IDS.
- Making data accessible doesn't require copying data outside of the local experiment database. Data may be accessed from the local experiment database through on-the-fly remote IMAS mapping. Only metadata are required to be copied to the joint metadata catalogue for multi-experiment searches

We remind that most EU experiments have already started to map some of their experimental data to the IMAS Data Dictionary, in the frame of the EUROfusion Code Development Work Package. The knowledge presently exists in most experiments for dealing with this.

We also would like to emphasize the fact that making data accessible outside of the experiment's context doesn't imply additional data validation/clearance schemes:

- Data is already exposed to the experiment's collaborators in a state that can be opened to the EU fusion researcher community without requiring more validation
- Data related to a publication is already cleared by the experiment – again we insist on the fact that even making data open shall not change the publication clearance process
- Data for the general public may go through a special clearance procedure if this is desired by a given experiment, although we don't see this as a requirement. It shall be made clear that data is always subject to reprocessing, changes, … and is made open in a given state of the research activities

4 Conclusion

After having presented a summary of existing experimental data holdings together with current access policies in European experiments, and an assessment of their compliance with FAIR Principles, we are making recommendations for a pragmatic and acceptable evolution towards more FAIR and more Open data management. Making metadata and data interoperable across EU experiments is a key target of these recommendations since it would bring unique benefits to the EU fusion research, 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. The IMAS Data Dictionary is recommended as the standard ontology for achieving interoperability.

This report can be used to promote these recommendations towards the EUROfusion Consortium and also establishes key principles/objectives to be taken into account for the design of the Blueprint Architecture (WP3) and Demonstrator (WP5).

5 Appendices

The questionnaires gathered by the EUROfusion Working Group on Open Data Access are annexed to this Deliverable. We acknowledge the following persons for having kindly provided answers to the questionnaires : D. Borba (JET), C. Centioli and E. Giovannozzi (FTU), A. Dingklage and A. Holtz (W-7X), F. Imbeaux (WEST), A. Kirk (MAST), A. Kallenbach and C. Fuchs (AUG). D. Tskhakaya and D. Fridrich (COMPASS).