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”

D1.4 - Data Management Plan

WP1

Deliverable status	Final
Type	Report
Dissemination level (according to the proposal)	Public
Work Package	WP1 - Management with Communication & Outreach
Lead Beneficiary (deliverable)	2 - UKAEA
Due Date	29/2/2020
Date of submission	27/2/2020

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

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

Document Information

AUTHOR

Author	Organisation	Contact (e-mail, phone)
Shaun de Witt	**UKAEA**	**Shaun.de-witt@ukaea.uk** **+44 1235 464585**

DOCUMENT CONTROL

Document version	Date	Author/Reviewer – Organisation	Change
V1	**17/02/20**	**Shaun de Witt – UKAEA**	First version
V2	**24/02/20**	**Irakalis Angelos Klampanos – NCSRD** **David Coster – MPG** Shaun de Witt - UKAEA	Second version
V3 – Final			Third version

DOCUMENT DATA

Keywords	Data Management, Experimental Data, Documentation
Point of contact	Name: Shaun de Witt Partner: UKAEA Address: Culham Science Park, Abingon, Oxfordshire, OX14 3EB, UK Phone: +44 (0)1235 464585 E-mail shaun.de-witt@ukaea.uk
Delivery date	February 27, 2020

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

Executive Summary

This document defines the data management plan for digital outputs generated from the FAIR4Fusion project. We identify the different types of digital output from the project, separate from the experimental data, and detail the lifecycle they will go through during the course of this project. Digital deliverables include data, software and documentation.

1. Introduction

Research data is defined as information, in particular facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings, and images[1]. The focus of the Open Research Data Pilot in Horizon 2020 is on research data that is available in digital form[2].

The Open Research Data Pilot applies to two types of data:

1. the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;
2. other data (e.g. curated data not directly attributable to a publication, or raw data), including associated metadata.

The obligations arising from the Grant Agreement of the projects are (see article 29.3):
Regarding the digital research data generated in the action (‘data’), the beneficiaries must:

1. deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:
    a. the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;
    b. other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan';
2. provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).

As an exception, the beneficiaries do not have to ensure open access to specific parts of their research data if the achievement of the action's main objective, as described in Annex 1, would be jeopardised by making those specific parts of the research data openly accessible. In this case, the data management plan must contain the reasons for not giving access.

[1] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

[2] Guidelines on Data Management in Horizon 2020
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Within the context of this project, new experimental data generated by the tokamak devices during the lifetime of this project are not considered to be a part of this plan and their management will continue to follow the data management strategies at the respective institutes. This call will make specific recommendations for future open data policies, with some open data being used as exemplars.

Thus the primary digital artefacts of this project are software and documentation in the form of deliverables, training materials and publications (both internal and external).

The rest of this document follows the format prescribed in the Horizon 2020 FAIR Data Management Plan (DMP) Template[3].

[3] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

2. Data Summary

In this section we describe the types of data generated or collected by the project specifically. We describe the types of data generally in quite high level terms (paper, software, etc). Experimental, modelling and engineering data are outside the scope of this project; for these, well established processes exist within the participating centres and it is not proposed to change these procedures. However, recommendations on modifications or additions to existing processes and procedures may be a final outcome of this project and these will be captured in papers or presentations generated within this project.

2.1 Types of Data

2.1.1 Summary Experimental Data (Used)

Purpose	Allows display of search parameters on portal either textually for scalar data or graphically for higher dimensional data. Usable as a discriminator for selection of full resolution data, or possibly downloaded as CSV dependent on policies.
Format	IDS[4] or CSV
Data Re-Use	This data will be gathered from existing experimental data which has been converted into IDS format for testing purposes and will only be ephemerally stored for a user session
Origin	Existing experimental data held at participating sites
Volume	<100GB
Data Utility	Summary data is useful for fusion researchers as it represents a quick look at the data without having to perform a full analysis. It is also useful to the general public who may be interested in fusion data

2.1.2 Metadata and Mapping Tools (Generated)

Purpose	Tools to provide mapping service between locally held metadata elements and ITER Data Dictionary where no such tools already exist.

[4] F. Imbeaux et al 2015 Nucl. Fusion 55 123006

Format	Software
Data Re-Use	-
Origin	Developed - based on existing community tools where possible
Volume	<100kB
Data Utility	Service to allow local metadata to be translated into IDS data dictionary elements

2.1.3 Metadata Annotations (Generated - optional)

Purpose	To enrich existing metadata and provide additional unstructured information related to data
Format	Metadata annotations
Data Re-Use	-
Origin	User input
Volume	<100MB
Data Utility	Allows for additional user supplied metadata in free text to be associated with existing experimental metadata. This could include elements such as the summary from the Responsible Officer or additional information from an authenticated user.

2.1.4 Metadata Schema from Existing Devices (Collected)

Purpose	Used for making queries against local databases or harvesting of information
Format	text/xml
Data Re-Use	
Origin	Originating site
Volume	<100kB
Data Utility	Used for querying on searchable parameters. In general sites will only

	provide searchable physical parameters. In no cases are these currently extended to include common schema's such as Dublin Core or DataCite. One recommendation of this could be a request to extend the metadata to include one of these common schemas, or elements thereof.

2.1.5 ITER Data Dictionary (Collected)

Purpose	Provide a consistent naming convention and clear definition.
Format	text/xml
Data Re-Use	-
Origin	ITER Organisation
Volume	<100kB
Data Utility	Used to provide community search terms and mapping to local metadata standards.

2.1.6 Software (Generated and Collected)

Purpose	This action will generate, reuse and distribute software. This software will support a number of requirements
Format	code
Data Re-Use	Where possible existing tools will be re-used. Software specifically developed within this project will be made open source, available on gitlab and released under the Apache-2.0 license
Origin	Various
Volume	Various
Data Utility	Provide support in WP4 (FAIR Building blocks) and WP5 (Demonstrator)

2.1.7 Presentations and Posters

Purpose	Disseminate information about the project both to fusion researchers and other communities.
Format	Powerpoint/PDF
Data Re-Use	Depends - mostly novel material with some re-use of images
Origin	Novel
Volume	4-8 presentations/poster over the course of the project
Data Utility	Useful for disseminating progress and engagement researchers. May also be useful input to cross cutting projects such as EOSC and EuroHPC.

2.1.8 Papers

Purpose	Papers will be produced both to convey any novel developments within this project but also to help promote adoption and FAIR and Open data principles amongst fusion researchers.
Format	PDFs to be published online
Data Re-Use	N/A
Origin	New
Volume	2-4 papers over the course of the project
Data Utility	Useful to fusion researchers and policy makers. Possibly also useful to other communities.

3 FAIR Data

One of the main outcomes of this action is to demonstrate to the fusion community across Europe the benefits associated with making data FAIR and, where permitted, open.

3.1 Making data findable, including provisions for metadata

At current tokamak devices, raw data is obtained from a number of diagnostic devices and from the device configuration. The number and types of diagnostic devices varies between experiments and within a single experiment can vary with time as devices are either replaced with newer versions or with a completely new device.

At each site data is discoverable by searching metadata related to individual ‘shots’ (experimental runs of a tokamak device lasting from less than a second to tens of seconds). However, this metadata is not consistent across all devices and the depth to which metadata can be searched varies from device to device. A part of this work is specifically to look at providing metadata in a consistent manner without changing the underlying data and metadata models already used at each site. In order to achieve this we will look at the current search interfaces from each device to extract common logical and physical parameters and, where necessary, add additional items to comply with generic query tools such as B2FIND, Invenio, etc. This latter will be more applicable to public access. Thus we consider two forms of metadata within the searchable parameter space; ‘public metadata’ based on well defined existing schemas and searchable physical parameters. Within this project, we anticipate providing the former as additional parameters which experimental sites will be recommended provide, but it is likely that for demonstration purposes. In defining this metadata we will make use of existing work taking place across the fusion community by adopting the ITER summary data (Summary IDS) as the source of searchable physical parameters. This has already been evolved with the participation of the community, but extra work needs to be performed to ensure all terms are searchable. Where there is information missing from the summary IDS we will seek to include have that added. It should be noted the Summary IDS has been developed by ITER and its definition is currently proprietary to the ITER organisation, but is licensed to all fusion sites.

Currently sites do not make use of common standard identifiers such as DOIs, but the community as a whole does have some level of standardisation for identification of data based on the instrument, shot number, field and version. For example, to obtain the total radiated power from the JET tokamak for shot number 96551, this has the unique identifier

```
        /JET/pulse/96551/ppf/signal/chain1/b5nn/topo
```

On the MAST device there is no direct access to the data but it goes through an API which allows unique identification of the data. The same signal on MAST is named *abm_prad_pol* (which implicitly includes metadata; *abm* is the multi-chrd bolometer detectors, *prad* means total radiated power and *pol* means measured at the poloidal array). To obtain this signal for say shot 30047 a user would make the following calls:

```
	client = pyuda.Client()
	result = client.get(‘30047’, ‘abm_prad_pol’)
	data = result.data
```

So it can be seen that for two devices at the same location different naming conventions are used for the same parameter. Our intention is to harmonise these through a query interface and using the naming conventions defined in the ITER Data Dictionary, which is an evolving standard being adopted by the community. Note this will only be for querying; the underlying names adopted by each device will remain unchanged and the tools provided will provide a mapping layer between the common name and the site specific name.

3.2 Making data openly accessible

As stated earlier, no data will be generated or modified by this project. At all times, the sites currently hosting the experiment will maintain full responsibility for the management of their own experimental data. This project provides a means of making this data more closely align to the FAIR principles by improving discoverability, interoperability and reusability.

The project aims at providing search capability and common access to data, primarily to the fusion users, without having to understand device specific terminology. Currently for most sites access to the data is restricted based on various different policies dependent on the site; data generated by one device may not even be accessible across the whole fusion community. This is due to the nature of the research; building a commercially viable fusion device clearly has potentially enormous economic benefits and the goal of most institutes is to support national programs for the commercialisation of fusion energy. In many cases they are both competitors to, and collaborators with, industrial partners. For the devices included within this project, they are only partly funded through the Commission via EUROfusion, with the rest coming from national governments.

That said, one of the political goals of the project is to instill the value of making data ‘as open as possible, as closed as necessary’ in order to achieve better FAIR compliance than is currently possible. However, it is understood there are some restrictions. We will work with the governing bodies both individually and through the EUROfusion General Assembly to promote this concept. Currently only one device provides full and open access to data after a 36 month embargo period and at least we will make this data accessible.

Currently none of the data is assigned a globally recognised persistent identifier (DOI’s, ARK’s pURLS, etc). However, the community does have community specific means of identifying the data (though even this is not homogeneous). Experimental data is accessed through one of three commons APIs:

- MDSplus[5] has long been a community standard for accessing experimental data which allows access to and browsing of data in a tree like manner
- Simple Access Layer (SAL)[6] has been developed to support similar functionality to MDSplus but with a lower overhead, although to date it has only been implemented for the JET experiment
- The Unified Data Access layer (UDA)[7] has been adopted at the standard for data access by ITER and is also used by MAST (note it also has a plugin to allow access to MDSplus data structures)

In all cases data is accessed in a tree like structure, but each API is distinct as shown in the following code snippets:

```
        # Get the plasma current (IPLA) from shot number 84600
        # from JET using MDSplus
        import MDSplus
        connection =
        MDSplus.Connection('mdsplus.jet.efda.org:8000')
        result = connection.get('_sig=jet("PPF/MAGN/IPLA", 84600)')

        # Get the total power (topo) for shot 96551 from JET
        # using SAL
        from jet.data import sal
        result =
    sal.get('/pulse/96551/ppf/signal/chain1/b5nn/topo')

        # Get the plasma current from MAST for shot 30047 using UDA
        # python API
        import pyuda
        client = pyuda.Client()
        result = client.get('amc_plasma_current', '30047')
```

The SAL example best exposes the hierarchical data access where the elements are broken down as follows:

- /pulse indicates this is pulse (shot) data
- /96551 indicates the specific pulse(shot)
- /ppf is the type of data - in this case a pulsed physics file
- /signal indicates the data is derived from a diagnostic signal

[5] http://mdsplus.org/index.php/Introduction
[6] https://github.com/simple-access-layer/source
[7] https://nucleus.iaea.org/sites/fusionportal/Shared%20Documents/Data%20Acquisition/10-05/Castro.pdf

- /chain1 indicates the quantity is derived from the chain 1 processing workflow
- /b5nn indicates the data is derived from the bolometric detector
- /topo is an abbreviation of total power.

In some cases sites already provide remote access methods (e.g. JET), but in other cases users are required to ssh onto a node at the hosting site before they can access data. Both SAL and MDSplus are open source protocols which support remote access. UDA is currently ITER licensed to the fusion community. In both UDA and MDSplus access requires authentication (which is based on local policies), and information on who is accessing data is recorded in access logs.

It should also be noted that there is a proposal in progress which will provide a EUROfusion wide AAI system, which will simplify the authentication and authorization process reducing the need for integration of local systems.

3.3 Making data interoperable

Data generated by fusion experiments exists in a wide number of formats, some of which are site specific (e.g. IPX files on MAST), while in other cases the data is already in well defined formats. The table below indicates the formats currently used at the different partner sites for experimental data.

Table 1: Data Formats Currently in Use at Partner Sites

Site	Format	Description
MAST[8]	IPX	Used for storing video information - essentially a series of uncompressed JPEG200 frames with timing metadata
	netCDF/HD5	Self describing structured file
	IDA3	Locally defined file format
MAST-U	netCDF/HD5	Self describing structured file
ASDEX Upgrade	MDSplus	Self describing structured file
WEST	IMAS	ITER style hierarchical data structures

[8] Note that detailed contents and formats have changed between MAST and MAST-Upgrade

TCV	MDSplus	Self describing structured file
JET	PPF	Pulsed Physics File (locally defined file format)
	CPF	Central Physics File (SQL database)
	JPF	Jet Physics File (locally defined file format)

In most cases metadata is associated with specific signals, but as mentioned earlier there is currently no consistency between sites. We will provide mapping information between site vocabularies and the ITER data dictionary and also try to aid users from other communities by describing these at a very coarse level.

3.4 Increase data re-use (through clarifying licences)

Public documents will be placed in a public repository which provides persistent identifiers, such as Zenodo, after suitable quality checking and approval. This approval will include approval by the project reviewers. Private documents generated as a part of this work will remain the responsibility of the Project co-ordinator who will be responsible for ensuring such documents are accessible to members of the consortia and funders/reviewers, but not more widely distributed. These documents will be maintained using an institutional repository available at Chalmers University of Technology (CTH).

Training material and presentations will be handled in the same way as public documents, with the exception that release will only need approval from the project steering committee, and will be released within one month of presentation (or immediately following approval in the case of training materials). Additionally, these will be placed on the project website for download.

New software (or other digital assets except as outlined previously) developed wholly within this project will be placed on the public gitlab under an Apache-2 license. Suitable documentation and test cases will be provided including user documentation and any other guides necessary. As software will be hosted on a public gitlab repository, user and other documentation associated with the software will also be held on this site.

Currently different experiments have different policies with regards to licensing. In general, there is no specific license associated with the data but is made freely available to members of the EUROfusion consortium. The exception is MAST data, which is openly licensed under CC-BY-NC-SA for public data. This project will recommend licensing be assigned to data, even when not open, to clarify rights, restrictions and responsibilities with reference to the use of the data. This work is also being undertaken in parallel with the EUROfusion Open Data Working Group (several members of the current consortium are also active within this group).

As only MAST data is open access, it is the only example with an official embargo period; currently this is 36 months or coincident with publication of papers based on the experimental data, whichever is earlier. The EUROfusion Working Group are currently recommending a EUROfusion wide embargo period of 24 months for open access to data. Across the community, generally experimental data is made available immediately with no embargo period for Principal Investigators.

Currently retention policies and quality assessments, where they exist, are carried out at each site. Since each experiment has unique

4 Allocation of Resources

Data produced by this project as identified in section 2 will be made public and, wherever possible, will make use of existing open repositories such a Zenodo[9] or schema.org[10] in order to reduce costs. Data Management for the project will be the responsibility of the Project Steering Group and CTH as the co-ordinator. Ultimately, the project co-ordinator will be responsible for the management of the outcomes of this project, but parts may be delegated to the relevant partners.

The costs associated with the provision of persistent identifiers will depend on the technology chosen and what is currently available. Our current plan is to use persistent identifiers provided by ePIC, which costs between 20 and 200€ p.a. Which will be paid from the project budget. As an alternative, we will also look at the use of DOIs provided through DataCite; however these are significantly more expensive (~6000€ p.a.). One complication is hosting cite for the prefix; since in both cases an agreement must be signed by a legal entity, the prefix owner can not be EUROfusion or the current project. Options to be investigated in this respect are for one site to hold the prefix and allow other sites to make use of it, or for each site to obtain it’s own prefix. The latter would be difficult in the case of JET, where the machine itself is owned by the commission, but the data is jointly owned by the legal entities participating in the EUROfusion project.

[9] https://zenodo.org/
[10] http://schema.org/

Since all software products will be open source, the use of gitlab will not require additional funding since open source projects can be freely hosted. Similarly, using Zenodo for documents and training materials does not require additional costs.

5 Data Security

At all times experimental data, including access to summary data, will be the responsibility of the site generating the data and will be outside the control of this project. We will work with the sites to ensure any sensitive data is not made publicly accessible. Sensitive information in this case is personally identifiable data such as names and e-mail addresses. This information will still be available to authenticated users, but will not be available publicly.

Zenodo, while not an officially certified trusted digital repository (i.e. has not obtained DSA or CTS certification) it has been recognised as a repository be re3data and openAIRE. Gitlab is an open repository with the backing of several large commercial ventures and supports over half a million projects.

6 Ethical Aspects

There will be no personal data generated as a part of this project. In some cases, databases at each site do contain personal identifiable information such as the Responsible Officer, lead Scientist or Diagnosticians. In this work we will ensure that this data is not harvested from local databases and not made public. However, the identification of these people are important for fusion researchers to ensure they fully understand the experimental results and shall be made available to the community on the same basis to that which currently exists.

There are no other ethical aspects to this work.

7 Other Issues

Data from existing experimental facilities will be handled according to existing data management procedures at each site. Together with the EUROFusion Open Access Working Group, this project will discuss with sites the harmonisation of these policies with respect to FAIR principles and Open Data, while ensuring sites retain autonomy for their own procedures.