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"

D2.2 - Interim Report on Open Science Use Cases for Fusion Information

WP2

Deliverable status	Final
Type	Report
Dissemination level (according to the proposal)	Public
Work Package	WP2 - Policy and use Case Definition
Lead Beneficiary (deliverable)	6 - MPG
Due Date	30/11/2019
Date of submission	29/11/2019

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)
David Coster	**MPIPP**	**David.Coster@ipp.mpg.de**
Pär Strand	**CTH**	par.strand@chalmers.se
Frédéric Imbeaux	**CEA**	**frederic.imbeaux@cea.fr**
Shaun de Witt	**UKAEA**	shaun.de-witt@ukaea.uk
Marcin Plociennik	**PSNC**	marcinp@man.poznan.pl
Andreas Ikonomopoulos	**NCSRD**	anikon@ipta.demokritos.gr
Iraklis Angelos Klampanos	**NCSRD**	iaklampanos@iit.demokritos.gr
Joan Decker	**EPFL**	joan.decker@epfl.ch

DOCUMENT CONTROL

Document version	Date	Author/Reviewer – Organisation	Change
V0.4	**2019-11-24**	**David Coster, MPIPP**	First version
V0.4_review	**2019-11-29**	**James Edwards, UKAEA**	Second version
V0.5	**2019-11-29**	**David Coster, MPIPP**	Third version

DOCUMENT DATA

Keywords	Use cases, user stories
Point of contact	Name: David Coster Partner: MPG

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

	Address: Boltzmannstr 2, Garching bei Muenchen, D85748 Phone: +49.89.3299-1965 E-mail: David.Coster@ipp.mpg.de
Delivery date	November 30, 2019

Contents

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

List of Figures

Terms and definitions

Acronym	Description
F4F	Fair4Fusion
IDS	Interface Data Structure
IMAS	Integrated Modelling & Analysis Suite

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

Executive Summary

The overall objective of Fair4Fusion “is to make European (or nationally when permitted) funded data more widely available to the fusion community, other science communities, funding bodies, and the public at large in order to maximise the impact of, the data and demonstrate the importance of the work done at relevant sites”.[1]

In this deliverable[2], “Interim Report on Open Science Use Cases for Fusion Information” we present a collection of user stories to capture existing and hypothetical open science use cases. The “Final Report on Open Science use Cases for Fusion Information” is due in month 9 of the project and will further elaborate and extend this collection of use cases.

For existing use cases, a range of cases starting with single site access (currently readily available) and then extending to multi-site access (currently only partly available) are provided. These take the form of plotting or obtaining the data for time-traces as well as profiles.

The open sciences cases extend beyond what is currently available for researchers and extend to the desire to download data matching certain criteria from a number of devices that could then be used for data-mining. Another extension beyond what is presently available, are the user stories wanting to search for appropriate discharges across devices in an unified manner.

These cases will provide input into the architecture and technical Work Packages and could be updated or elaborated as a result of further feedback from them.

[1] Fair4Fusion Proposal

[2] D2.2 “Interim Report on Open Science Use Cases for Fusion Information” presents existing use cases currently used by fusion researchers and hypothetical open science use cases. The use cases will cover improved collaboration between different groups within the same site, increased cross-site data analysis and machine learning across multiple data sets. It will also identify a subset of these requirements which should be addressed by the demonstrator being developed within WP5 based on maximising the impact of the demonstrator. (MPIPP)

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

1 Introduction

The overall objective of Fair4Fusion "is to make European (or nationally when permitted) funded data more widely available to the fusion community, other science communities, funding bodies, and the public at large in order to maximise the impact of, the data and demonstrate the importance of the work done at relevant sites".[3]

This document attempts to identify typical use cases that will form the basis of the work to be performed in other work packages of Fair4Fusion. It concentrates on use cases addressing data from the existing major experiments as this data is well managed and (relatively) homogeneous.

As the use cases were being prepared, it became clear that if we make the assumption that all of the data (and possibly metadata) from the devices were to be provided via IMAS (Integrated Modelling & Analysis Suite) IDS's (Interface Data Structure), much of the work in this project will become easier[4]. Since this also aligns with work that is being done within EUROfusion, it is likely that Fair4Fusion will adopt this assumption.

If we make this assumption, then much of the work in the project is dramatically simplified, but we will then need to

1. Get this principle accepted by the experiments
    a. Intermediate possibility: F4F (Fair4Fusion) implements a tool that queries the device databases and populates IDS's
2. Ensure that sufficient data/meta-data is available in IDS's from the experiments to provide input to any prototypes we build in the project
3. Identify any gaps in the present IDS's that will need to be filled
    a. If the first port of call for finding data will be based on the summary IDS, we should ensure this provides slots for all of the fields that are currently supplied by the existing devices as part of their search criteria?
    b. Can we help provide tools/expertise to the device teams to help fill slots in the summary IDS that would be useful but are not currently provided as part of their database search options?

When starting to provide use cases, the fusion physicists provided them in the form shown in Appendix A. As a result of feedback ('Perhaps we could consider using the "user stories" paradigm (more info *e.g.* http://www.agilemodeling.com/artifacts/userStory.htm)') from the technical work packages, the use cases were recast as user stories in section 2. Section 4 provides a summary and outlines some of the next steps. The original use cases have been preserved in Appendix A as they

[3] Fair4Fusion Proposal
[4] For a description of the concepts behind the IMAS IDS's, see F. Imbeaux *et al* 2015 *Nucl. Fusion* **55** 123006, https://doi.org/10.1088/0029-5515/55/12/123006; for those with access to the ITER confluence server, further information on IMAS can be found at https://imas.iter.org.

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

might still be useful for providing additional information and Appendix B provides some background technical information (on the categorization of users and access levels and experimental data formats). .

2 User stories

Here we have collected a number of user stories about accessing data and/or metadata, as well as some of the wishes from the data providers. Those use cases are presenting the different perspective of the members of the general public, researchers and data providers that are the main target users of analyzed scenarios.

2.1 Member of the general public

As a member of the general public, I would like to know how many shots per day and per year are performed by each of the experiments.

As a member of the general public, I would like to know some key data about each discharge: e.g. length of the pulse, toroidal field, plasma current, flat-top length, maximum and integrated heating power, maximum and average stored energy, maximum and average neutron production rate, …

As a member of the general public, I would like to know which shots have resulted in publications, and to how many publications.

As a member of the general public, I would like to know how the discharges of the various devices compare with respect to key figures of merit (e.g. n*tau*T, H-98, …)

As a member of the general public, I would like to know how much each discharge costs, and what the cost breakdown is.

As a member of the general public, I would like to know to what extent each discharge meets the goals expected for the discharge.

As a member of the general public, I would like to know about the reliability of each device.

As a member of the general public, I would like to know how many discharges end in disruptions for each device, and be able to dig into this data to try to detect trends.

2.2 Researcher finding suitable shots

As a researcher, I want to find shots that have the keyword “Standard H-mode” in their description, across a selection of devices.

As a researcher, I want to find shots that were performed on a particular day on a particular device.

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

As a researcher, I want to find shots that were in the top 5% for the product of Ip * Bt for a particular device.

As a researcher, I want to locate all H-mode shots that had a flat-top phase of longer than 0.5 seconds, across a selection of devices.

As a researcher, for the shots that I find matching my criteria, I would like to be able to download the list of shots together with data characterizing the shots from a list of such fields that I can choose from, across a selection of devices.

As a researcher, I would like the system to remember what choices I have made, so that I can refine (or generalize) my search criteria at a later date, or to search for new discharges that meet my criteria.

Search the EUROfusion databases for device/shots matching criteria determined by the particular database[5].

2.3 Researcher obtaining time traces

As a researcher, I want to plot the plasma current[6] for a particular shot on a particular machine.

As a researcher, I would like to be able to download the current versus time data that I have just plotted, in various formats (such as ASCII, netCDF, HDF5 or JSON).

As a researcher, I want to compare (plot) the plasma current traces for different shots on the same machine.

As a researcher, I want to compare (plot) the plasma current time traces for different machines.

As a researcher, I want to compare the time traces of H-98[7] for different shots from different machines.

As a researcher, I would like to be able to annotate any plots that I have produced, and then be able to save the plots in both vector and bitmap formats (such as postscript, PDF, CSV, PNG, JPG *etc*.

[5] https://users.euro-fusion.org/iterphysicswiki/index.php/Database identifies four databases: disruptions, confinement, pedestal and stellarator; some fields will be shared but some fields will be unique to each database

[6] magnetics/method(:)/ip/data(:), equilibrium/time_slice(:)/global_quantities/ip or core_profiles/global_quantities/ip(:)

[7] summary/global_quantities/h_98/value(:). See summary/global_quantities for other examples of what might be wanted.

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

As a researcher, for data that I have been able to plot, I would like to be able to download the data in various formats.

2.4 Researcher obtaining profile data

As a researcher, I want to plot density profiles[8] for a particular shot on a particular machine.

As a researcher, I want to compare (plot) density profiles for different experiments on the same machine.

As a researcher, I want to compare density profiles for different machines.

As a researcher, I want to compare the different data related to the density profile for a particular machine.

As a researcher, for data that I have been able to plot, I would like to be able to download the data in various formats.

2.5 Researcher obtaining data needed for simulations

As a researcher, I want to obtain the equilibrium information for a particular shot at a particular time, being able to plot and save the data[9].

As a researcher, I want to obtain a description of the core profiles for a particular shot at a particular time, being able to plot and save the data.[10]

As a researcher, I want to obtain the heating and current drive sources as a function of space and time for a particular shot, being able to plot and save the data.

As a researcher, I would like to select cases from multiple experiments, based on the cataloged metadata, and then to obtain a dataset containing data extracted from the selected cases as specified fields (from the entire data tree, not only from the metadata); these fields may be scalars but could also be of higher dimensionality. I must be able to save the returned dataset in a file format adequate to cope with large data volumes (*e.g.* HDF5).

[8] For example, core_profiles/profiles_1d(:)/electrons/density(:) as a function of core_profiles/profiles_1d(:)/grid/rho_tor_norm(:)

[9] *E.g.* as an EQDSK file, see https://w3.pppl.gov/ntcc/TORAY/G_EQDSK.pdf

[10] Fields below core_profiles/profiles_1d(:)/

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

As a researcher, I would like to be able to register for updates for data.

As a researcher, I would like to be able to identify if earlier versions of the data exist, and if so, compare the different versions of the data.

2.6 Researcher providing additional meta-data or data

As a researcher, I would like to be able to provide my own public or private metadata to augment the available metadata. The sorts of metadata that I would like to add are:

- comments on the shot itself (*e.g.* not useful for my purpose because …);
- comments that data from this shot has been used in this publication of mine;
- comments that data from this shot has been inserted into this database that I am creating;
- comments that data from this shot are being used in a particular project of mine;

As a researcher, I would like to be able to provide my own public or private data for particular shots to augment the available data (processed data). This could, for example, be that I have combined data from a number of diagnostics for a particular shot to produce my own version of an electron temperature profile which might be of interest to others, or I have combined different signals to calculate the power crossing the separatrix that will be the basis of future work.

As a researcher, I would like to add references to publications that use data for a particular shot to the metadata for that shot (or shots).

2.7 Data provider

[AUG] As a data provider, the experimental data is stored in standardised shotfiles stored in AFS accessible via libraries provided for a number of computer languages.

[JET] As a data provider, the experimental data is stored in standardised pulse files (of various types) where the data is accessible locally via library calls supported for a number of computer languages.

[AUG+JET+TCV] As a data provider, I currently provide external access to much of my local data via MDSplus.

[JET] As a data provider, I currently provide external access to a subset of my local data via SAL.

As a data provider, I want to ensure the appropriate availability of my data without breaking the law (*e.g.* GDPR).

As a data provider, I want to ensure that the provided data is up-to-date so that local changes are appropriately propagated.

As a data provider, I want to ensure that the provision of this data has no impact on plasma operations.

As a data provider, I want to provide the data with least cost to my institute.

As a data provider, any solutions to provide data/metadata must be well documented, maintainable and as future-proof as possible.

As a data provider, I want to ensure that any embargo periods are enforced.

As a data provider, for data not yet out of the embargo period, I want to ensure that the appropriate paperwork has been completed.

As a data provider, I would like to know who (by category) has accessed my data.

As a data provider, I would like to know how much data is accessed via the remote access portal.

3 Summary and Outlook

This document provides an interim report on a number of use cases that should be addressed for implementation by the Fair4Fusion project. This will be followed by a deliverable in six months providing the final report on use cases which will also identify any policy issues (raised by D2.1, "Data Inventories and Policy Landscape", M6) that might block implementation and suggest means of overcoming them. A later deliverable will characterize the non-experimental fusion data.

The selection and prioritization of the cases for implementation will be done on the basis of

- coverage over the various categories provided in section 2.
- feasibility of implementation within the constraints of a two year project
- the availability of the necessary data from the experiments
- demonstrating the usefulness of the implemented tools to go beyond what is presently available

4 Appendix A: Additional background in support of the use cases

The following sections provide some additional information in support of the use cases described in the user stories. Some of the use cases are repeated and supplemented with examples of the sort of output that is currently available for a particular device, mostly from AUG but similar output is available from other devices.

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

4.1 Finding suitable discharges

The scientist is trying to find shots on one or more tokamaks which meet his/her requirements.

The following examples are used to demonstrate possible search cases:

1. Find all AUG shots which have more than 1.05 MA of plasma current
2. Find all JET shots with more than 3.5 MA of current
3. Find the top 10 discharges on each machine with regard to injected energy
4. Find all phases in all discharges where key quantities (toroidal field, plasma current, line averaged density, stored energy) are constant for more than two energy confinement times reporting key values for these discharges
5. Find all shots where good pedestal and SOL profiles of Te and ne are present
6. Find all shots where the H-98 factor exceeds 1.2 and return key parameters

Help will be needed to identify the names of fields that can be used to identify the cases, and to identify what data can be returned to characterize the discharges.

An example (for AUG) of the output for such a search is presented below:

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

ASDEX Upgrade Journal: Shotlist 30.07.19, 17:31

ASDEX Upgrade Journal

Shot	B_t [T]	I_p [MA]	n_e [1019 /m2]	q_{95}	P_{aux} [MW]	Remarks Experiment / Shot req. Program	Beha-viour	Gas	Conf δ_u	Useful Type
03383 01.07.93 09:39	-1.3	0.73	4.1			**** Standard-H-Mode-#** 4 min geglimmt in He Ip - Messung falsch Program: *189*	STABLE HMOD	D	SNu	yes plasma
03540 16.07.93 10:01	-1.3	0.79	4.1			**** Standard H-Mode Test #** ! Vor Schuss 4 min in He geglimmt ! CoI-Rampe noch unsymmetrischer Program: *189*		D	SNu	yes plasma
11492 04.02.99 15:08	-2.1	1.01	8.1	3.4	NI 5.3 IC 0.1	**** Standard H-Mode #** Quellen 2,1 statt 3,2 kein Nachlauf Q3 -> DL Disr i. Ramp-down ne-soll-max = 6E19 1 Program: *1010*	STABLE DISE HMOD VDEu	D He	SNu	yes plasma
11493 04.02.99 15:33	-2.1	1.01	8.1	3.4	NI 5.4 IC 0.1	**** Standard H-Mode #** ne-soll-max = 8E19 kein Nachlauf Q3 -> DL Disr i. Ramp-down 1 Program: *1010*	STABLE DISE VDEu	D He	SNu	yes plasma

Figure 1: Example from the AUG Journal.

Another example of such output is as a CSV list:

Shot," Ip MA"," Bt T"," ne"," pheattot MW"," nbi4m MW"," icrh4m MW"," ecrh4m MW",

3383," 0.7284"," -1.298"," 4.08e+19",,,,,
3540," 0.7939"," -1.345"," 4.12e+19",,,,,
11492," 1.007"," -2.053"," 8.11e+19","5.444","5.333","0.1111",,
11493," 1.008"," -2.054"," 8.08e+19","5.495","5.381","0.1136",,
11494," 1.006"," -2.057"," 9.78e+19","5.395","5.283","0.1123",,
11502," 1.007"," -2.059"," 9.32e+19","5.277","5.277",,,
11511," 1.005"," -2.057"," 9.76e+19","5.278","5.278",,,
11533," 1.004"," -2.057"," 1.00e+20","5.29","5.29",,,
11554," 1.009"," -2.053"," 8.73e+19","5.347","5.347",,,
11574," 1.001"," -2.055"," 5.60e+19","5.37","5.37",,,

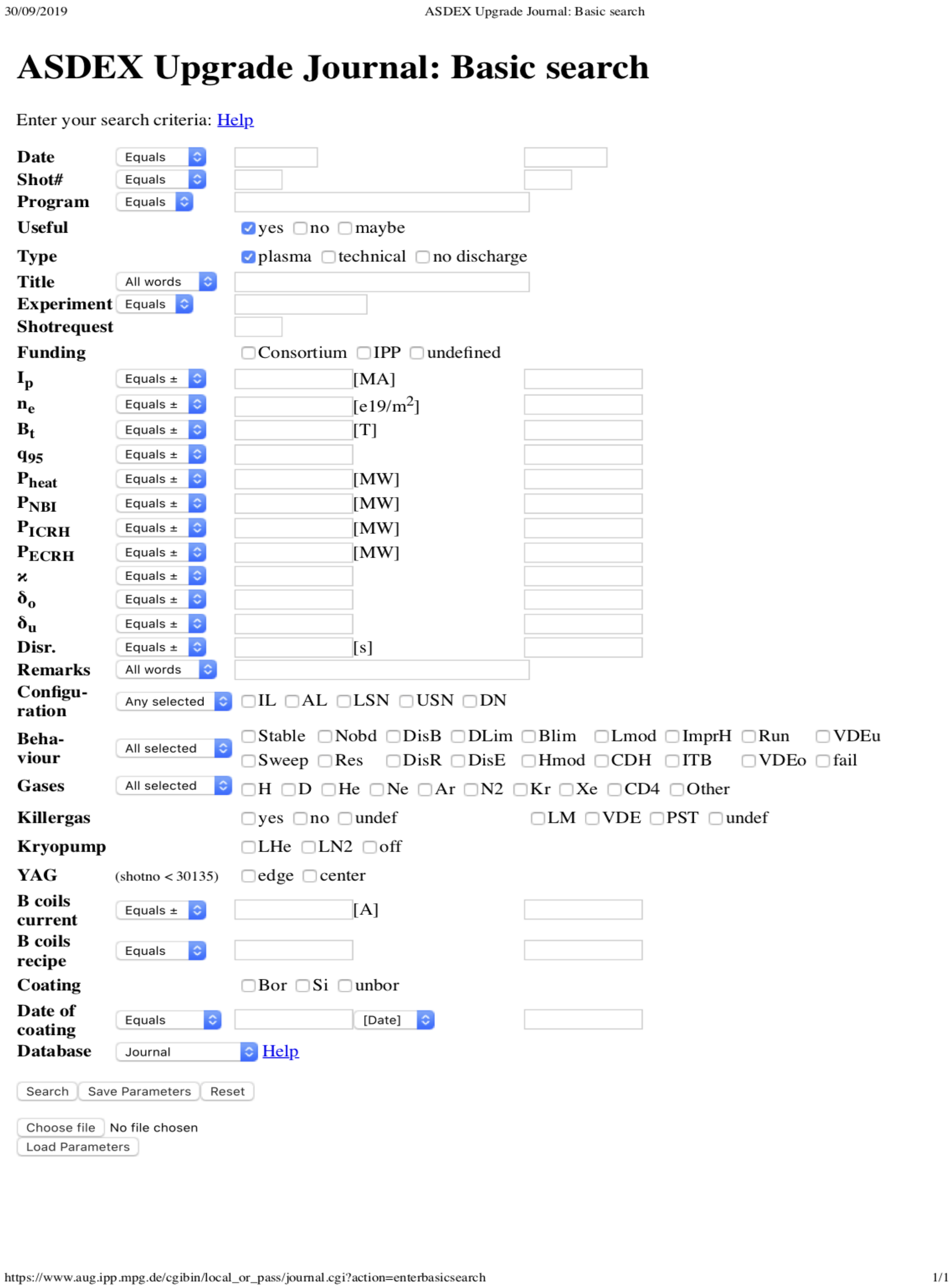
This project has received funding from the European Community's Horizon 2020
Framework Programme under grant agreement 847612

Many of the tokamaks have such a search system already, and can be divided into two general categories:

1. A database of information characteristic of each shot is created, with a combination of scalar fields and text information. This is often created at the time the shot is performed.
2. One or more databases of a more specialized nature where data from one or more time points in a shot is captured. This is usually from one or more post-processing codes that are run after the shot has been performed.

An example of the interface for the first type of query for AUG is shown below:

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



30/09/2019 ASDEX Upgrade Journal: Basic search

ASDEX Upgrade Journal: Basic search

Enter your search criteria: Help

Date	Equals			
Shot#	Equals			
Program	Equals			
Useful		☑ yes ☐ no ☐ maybe		
Type		☑ plasma ☐ technical ☐ no discharge		
Title	All words			
Experiment	Equals			
Shotrequest				
Funding		☐ Consortium ☐ IPP ☐ undefined		
I_p	Equals ±		[MA]	
n_e	Equals ±		[e19/m^2]	
B_t	Equals ±		[T]	
q_{95}	Equals ±			
P_{heat}	Equals ±		[MW]	
P_{NBI}	Equals ±		[MW]	
P_{ICRH}	Equals ±		[MW]	
P_{ECRH}	Equals ±		[MW]	
$\varkappa$	Equals ±			
δ_o	Equals ±			
δ_u	Equals ±			
Disr.	Equals ±		[s]	
Remarks	All words			
Configu- ration	Any selected	☐ IL ☐ AL ☐ LSN ☐ USN ☐ DN		
Beha- viour	All selected	☐ Stable ☐ Nobd ☐ DisB ☐ DLim ☐ Blim ☐ Lmod ☐ ImprH ☐ Run ☐ VDEu ☐ Sweep ☐ Res ☐ DisR ☐ DisE ☐ Hmod ☐ CDH ☐ ITB ☐ VDEo ☐ fail		
Gases	All selected	☐ H ☐ D ☐ He ☐ Ne ☐ Ar ☐ N2 ☐ Kr ☐ Xe ☐ CD4 ☐ Other		
Killergas		☐ yes ☐ no ☐ undef		☐ LM ☐ VDE ☐ PST ☐ undef
Kryopump		☐ LHe ☐ LN2 ☐ off		
YAG	(shotno < 30135)	☐ edge ☐ center		
B coils current	Equals ±		[A]	
B coils recipe	Equals			
Coating		☐ Bor ☐ Si ☐ unbor		
Date of coating	Equals		[Date]	
Database	Journal	Help		

Search | Save Parameters | Reset

Choose file No file chosen
Load Parameters

https://www.aug.ipp.mpg.de/cgibin/local_or_pass/journal.cgi?action=enterbasicsearch 1/1

Figure 2: Example of a search box from AUG.

4.2 Obtaining and/or plotting scalar data as a function of time

The scientist is trying to understand/characterize a particular discharge, or to compare discharges from the same or different machines.

Three levels of access can be identified:

1. Providing an interface where the data can be plotted (this could be entirely web based, or via a program that is downloaded and queries the server)
    a. Ideally, the user should have options to change the labelling and scaling of the plots
    b. It should be possible to save the plots as bitmap (GIF, PNG, …) and vector (PS, PDF, SVG) files
2. Provide an interface where the data can be downloaded (this could be entirely web based, or via a program that is downloaded and queries the server)
    a. The output data formats to be supported are probably fairly simple (ASCII) but could also extend to netCDF and/or HDF5.
3. Provide an API where the data can be accessed by a user written program

At the start I suggest targeting 1 and 2 above via a web interface.

The following examples are used to demonstrate particular use cases:

1. Plot key parameters (*e.g.* plasma current, toroidal field, line averaged density, H-98, betan) for one particular discharge
2. Plot key parameters (*e.g.* plasma current, toroidal field, line averaged density, H-98, betan) for different discharges on the same machine
3. Plot key parameters (*e.g.* plasma current, toroidal field, line averaged density, H-98, betan) for discharges from different machines

Help will be needed for the scientist to identify the names of the signals to be plotted. The names for these signals as currently implemented is likely to be different for each device.

Some additional information about the data is given in Appendix B.

An example (again for AUG) is shown below:

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

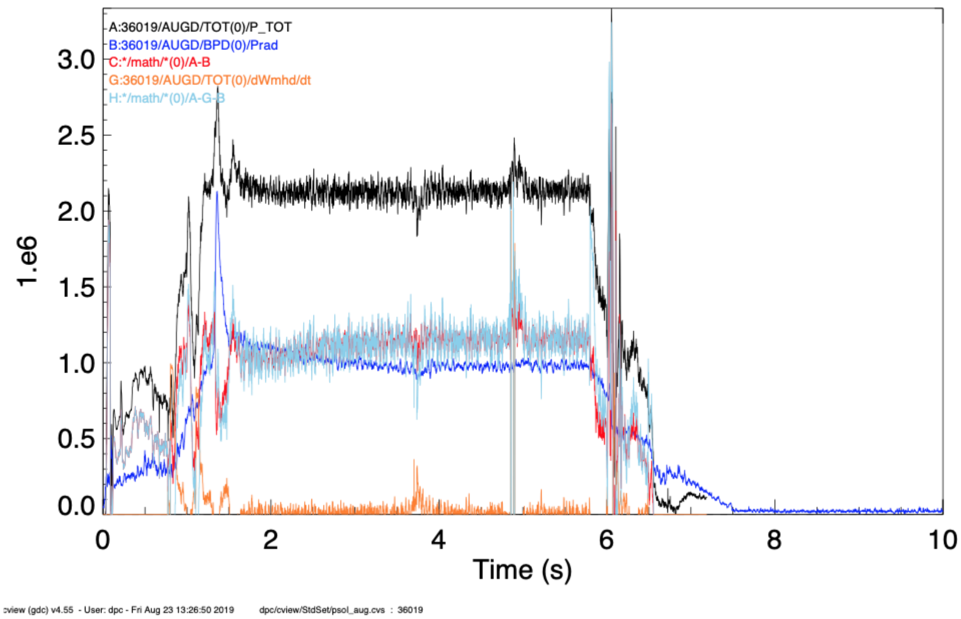


Figure 3: Example time traces plot.

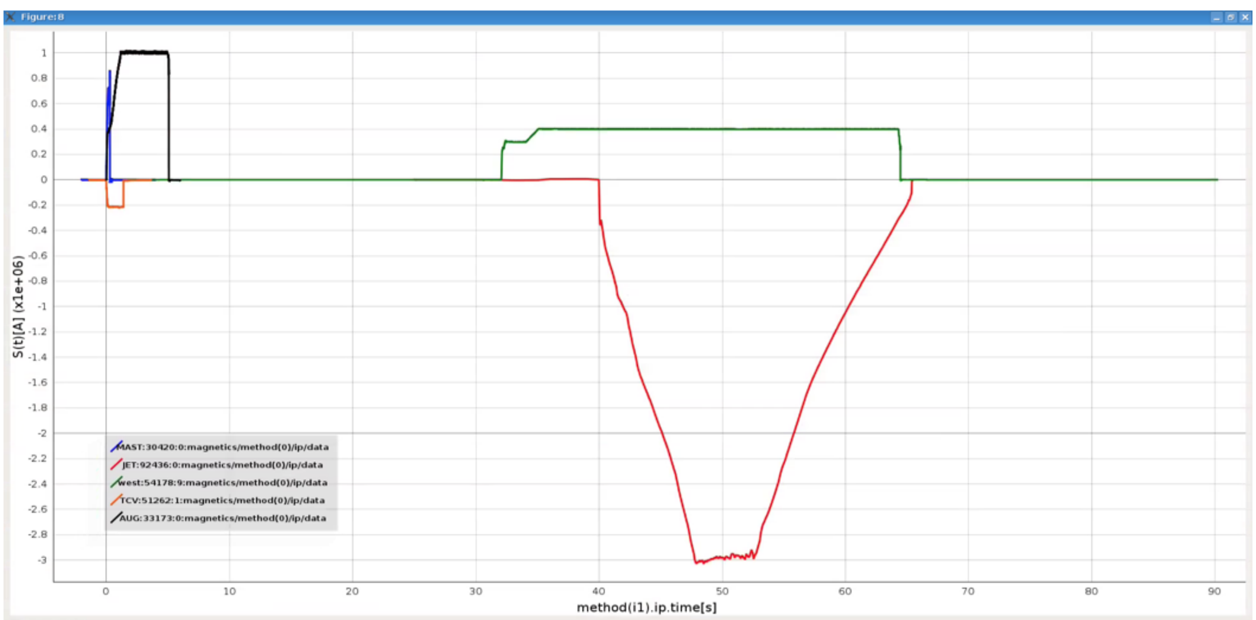


Figure 4: Another example of a time traces plot.

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

4.3 Obtaining and/or plotting profiles as a function of time

The scientist is trying to understand/characterize a particular discharge, or to compare discharges from the same or different machines

Three levels of access can be identified:

1. Providing an interface where the data can be plotted (this could be entirely web based, or via a program that is downloaded and queries the server)
    a. Ideally, the user should have options to change the labelling and scaling of the plots
    b. A new feature here is the ability to choose the radial coordinate
    c. It should be possible to save the plots as bitmap (GIF, PNG, …) and vector (PS, PDF, SVG) files
2. Provide an interface where the data can be downloaded (this could be entirely web based, or via a program that is downloaded and queries the server)
    a. The output data formats to be supported are probably fairly simple (ASCII) but could also extend to netCDF and/or HDF5.
3. Provide an API where the data can be accessed by a user written program

At the start I suggest targeting 1 and 2 above via a web interface.

The following examples are used to demonstrate particular use cases:

1. Plot profiles of key parameters (*e.g.* the electron temperature, the electron density, the ion temperature) for one particular discharge
2. Plot profiles of key parameters (*e.g.* the electron temperature, the electron density, the ion temperature) for different discharges on the same machine
3. Plot profiles of key parameters (*e.g.* the electron temperature, the electron density, the ion temperature) for discharges from different machines

Help will be needed for the scientist to identify the names of the signals to be plotted. Again, each machine currently has different names and conventions for the signals; it is possible that multiple different versions of the same physics quantities might be available from different diagnostics.

An example (again from AUG) is shown below:

Here, three different diagnostics are used to provide the ion temperature.

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

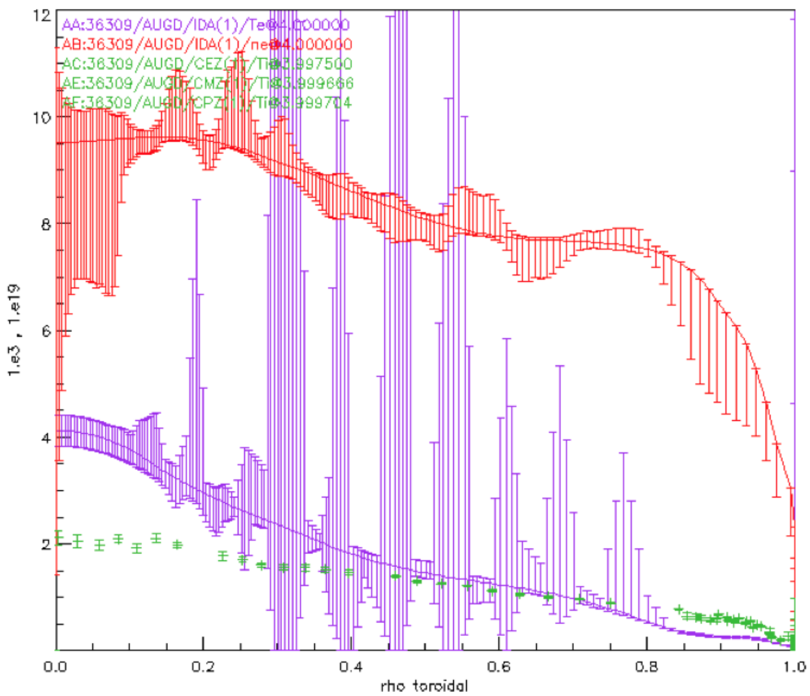


Figure 5: Example of a profiles plot.

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

4.4 Provide the data needed for simulations

The use case here is to provide the data needed to provide the inputs to simulations of the experimental scenarios. In the past this might be done by extracting scalar fields or profiles and storing them in ASCII data files or Excel Spreadsheets.

Within EUROfusion there is an effort to convert codes to expect their input in IMAS IDS's. It would be good if the infrastructure could supply these IDS's.

An example simulation is shown below:

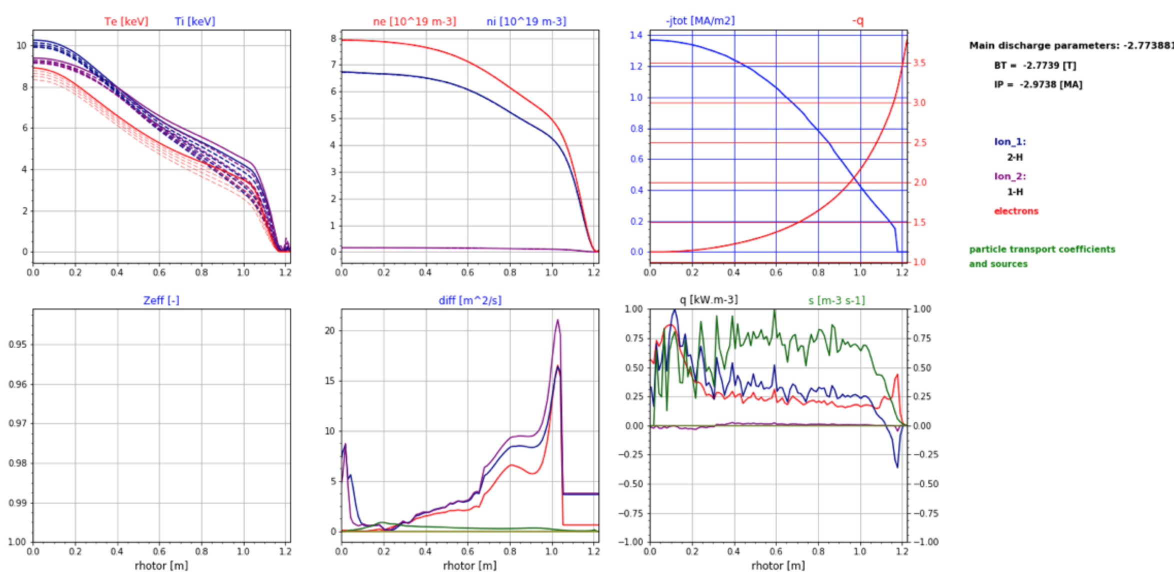


Figure 6: Example of simulation output.

At the moment, programs have been written to provide the data with IMAS IDS's for a number of the devices, and these are then used to provide initial and boundary conditions for the simulations.

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

4.5 Gather large datasets to perform dedicated analysis

Start from Use Case 1 to operate a selection of a dataset from multiple experiments, based on the cataloged metadata.

Ask to return a data file containing data extracts from that selected dataset: specify what fields to return (from the entire data tree, not only from the metadata). These fields may extend beyond simple scalars to higher dimensionality. The data file must contain the returned fields concatenated in a file format adequate to cope with large data volumes (*e.g.* HDF5).

Then the user will perform his/her own analysis using the returned data file.

This use case is typically the initial step to perform further data mining, or neural network construction from a dataset, *etc*.

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

5 Appendix B: Background technical information

5.1 Categorisation of users and access levels

The level of granularity of access rights will have to be developed with the experiments and the community. The “community” here is to be taken as the EUROfusion supported researchers and institutions. The reason for this is that this is by far the largest congregation of fusion researchers in Europe. Also, as a joint co-fund action under the H2020 programme it also has the collaborative processes and procedures including IP and access rights already developed.

Three levels are easily identified

1. The general public
2. External collaborators (defined as fusion researchers not covered by EUROfusion agreements)
3. EUROfusion

There will be different level of processes and procedures for the different categories. A complication is that EUROfusion contain facilities that are only partially embedded within the EUROfusion project and in principle falls in between category 2 and 3.

Hence for the EUROfusion there are three levels

a) General access to EUROfusion members
b) EUROFusion collaborators (possibly with enhanced access and no embargos)
c) Internal device access (validation of data, national projects *etc.*)

Fair4Fusion will need to work with EUROfusion and the experiments to be able to identify users, ascertain what access rights they have and then to enforce those rights.

5.2 Experimental data formats

- At the moment each of the machines has its own way of storing the data internally, as well as names for the data fields that are specific to each machine. The units of the quantity, as well as the time-base might also be different from machine to machine.
    - The data is often accessed via a hierarchy consisting of a
        - Sub-system name
        - Signal within that sub-system
        - And some local embellishments
- Each machine provides libraries for accessing the data locally. Typically supported are
    - Fortran and C
    - IDL and/or Matlab
    - Python

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

- Almost all (all?) of the machines currently also supports remote access to data via MDSplus[11] but the user would still need to know the names of the signals on each machine, as well as (possibly) having to manage ssh tunnels.
  - JET also supports “Simple Access Layer”[12]
- Within EUROfusion there is an effort to provide data via UDA (Unified Data Access)[13].

[11] http://www.mdsplus.org/index.php/Introduction

[12] https://data.jet.uk/guides/sal/

[13] https://nucleus.iaea.org/sites/fusionportal/Shared%20Documents/Data%20Acquisition/10-05/Castro.pdf