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”

D2.3 - Final 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	31/5/2020
Date of submission	29/5/2020

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

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

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

DOCUMENT CONTROL

Document version	Date	Author/Reviewer (Organisation)	Change
V0.1	**2020-05-17**	**David Coster (MPIPP)**	First version (based on the separately maintained living document)
V0.1_review_SdW	**2020-05-20**	**Shaun de Witt (UKAEA)**	Review
V0.1_review_MP	**2020-05-20**	**Marcin Plociennik (PSNC)**	Review
V0.2	**2020-05-25**	**David Coster (MPIPP)**	Incorporate reviewers comments
V0.3	**2020-05-28**	**David Coster (MPIPP)**	Incorporate additional comments
V0.3_NCSRD	**2020-05-29**	**Andreas Ikonomopoulos (NCSRD)**	Additional comments
V0.4	**2020-05-29**	**David Coster (MPIPP)**	Incorporate additional comments
V1.0	**2020-05-31**	**Pär Strand (CTH)**	Final Review PO for release

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

DOCUMENT DATA

Keywords	Use cases, user stories
Point of contact	Name: David Coster Partner: MPG Address: Boltzmannstr 2, Garching bei Muenchen, D85748 Phone: +49.89.3299-1965 E-mail: David.Coster@ipp.mpg.de
Delivery date	May 29, 2020


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

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
FAIR	Findable, Accessible, Interoperable, and Reusable
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 produced by fusion devices more widely and homogeneously available to the fusion community, other science communities, funding bodies, and the public at large in order to maximize the impact of the data and demonstrate the importance of the work done at relevant sites.[1]

In this deliverable[2], "Final 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. This extends on the submitted deliverable "Interim Report on Open Science Use Cases for Fusion Information" by adding a few new use cases, and more significantly, by providing an analysis of each use case in terms of any blocking issues.

For the use cases, a range 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 (one or more scalars as a function of time) as well as profiles (one or more scalars as a function of a radial coordinate, typically at a user specified time).

The open science cases extend beyond what is currently available for researchers 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 a unified manner.

A set of user stories from the perspective of FAIR[3] principles has also been added, with some overlap of the previous stories but also adding a new perspective.

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

[1] Based on Fair4Fusion Proposal
[2] D2.3 - Final Report on Open Science use Cases for Fusion Information (M9) extends D2.2 to cover additional use cases which we consider may be of interest to wider audiences including other research organisations and funders, enabling them to quantify their investment in fusion research. (MPIPP)
[3] Findability, Accessibility, Interoperability, and Reusability

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".[4] The data mentioned here is interpreted to be the data produced by fusion devices within the EU, with an initial emphasis on the devices funded by EUROfusion. In order to make the data more readily usable, it is intended to increase the homogeneity with which the data can be accessed.

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) (outlined in D4.1 "Architecture for a Metadata Interoperability Service and Proposed Metadata Model"), much of the work in this project will become easier[5]. Since this also aligns with work that is being done within EUROfusion and ITER, 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 (summarized in D4.1 "Architecture for a Metadata Interoperability Service and Proposed Metadata Model"), 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. Possibly help the device teams fill the fields in the summary IDS by providing tools/expertise

[4] Fair4Fusion Proposal
[5] 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

4. Look at mechanisms that could provide solutions for data and/or metadata that cannot fit into the existing IDS's[6]

When starting to provide use cases, the fusion physicists provided them in the form shown in Appendix A. As a result of feedback[7] from the technical work packages, the use cases were recast as user stories in section 2. Section 3 provides a summary and outlines some of the next steps.

The original use cases have been preserved in Appendix A as they might still be useful for providing additional information and Appendix B provides some background technical information (on the experimental data formats).

1.1 Access levels

In developing the use cases, it became clear that the project would need to distinguish between different categories of users and their associated access levels. The exact level of granularity of access rights will have to be developed with the experiments and the community[8].

Five levels are easily identified:

1. The general public,
2. Funding agencies,
3. External collaborators (defined as researchers not covered by EUROfusion agreements),
4. General EUROfusion collaborators,
5. Internal (to the experiment) scientists[9],

and the need for more might develop. The data manager mentioned in the FAIR section would --- in most cases --- be in either category 4 or 5. Non-fusion researchers would --- for the moment --- sit in category 1, 3, 4 or 5 depending on their relationship with the experiments or EUROfusion.

These different levels might have access to different subsets of the data. As examples,

- category 5 might have access to all of the data associated with their experiment, but only a subset of the data available on other experiments
- category 4 would have access to all data whose collection was funded by EUROfusion
- category 3 might have access to very detailed data, but only after any embargo period has expired

[6] Examples of such data might be free text annotations about the data, provenance data as well as data related to citations
[7] Perhaps we could consider using the "user stories" paradigm (more info *e.g.* http://www.agilemodeling.com/artifacts/userStory.htm)
[8] 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 and, being a joint co-fund action under the H2020 programme, it also has the collaborative processes and procedures including IP and access rights already developed.
[9] This would include any external scientists that have established an agreement of data access to that experiment.

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

- category 1 might have access to less detailed data after the expiry of any embargo period

The exact access rights, and any limitations as to what level of data is to be provided, is likely to evolve as a result of interactions within this project, with the experiments, with the funding agencies and the development of attitudes to open-data.

This project should consider the mechanisms to implement any such restrictions, even if, at the beginning, there might not be any special data associated with a particular category. EUROfusion is working on an AAI implementation but the use of this is not a requirement of any demonstrator built as part of this project.

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

2 User stories

Here we have collected a number of user stories about searching for and accessing data and/or metadata, as well as some of the wishes from the data providers. Those use cases are presenting the different perspective of the members of the general public, researchers and data providers that are the main target users of analyzed scenarios. The breakdown into categories is historical and is maintained as other parts of the project have retained links to these stories in the order that was used at the beginning. A certain amount of repetition is also present since the user stories were developed independently, and this is reflected in the added commentary. What is also apparent is a wide variability in the specificity of the questions --- the intention is to use the specific form of the question to drive a more general specification.

After each user story, a commentary has been added to identify any blocking issues and, in some cases, to provide a generalization of the specific story. It is expected that this analysis will be extended by other Work Packages as they work to deliver any prototypes and a Blueprint Architecture.

2.1 Member of the general public

At least some of the following questions could come from funders and policy makers even though the question is phrased in terms of “a member of the general public”.

2.1.1 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.

No policy issue would block this; however it is not clear that the experiments have this data in a form that could be easily mined for presentation. It is possible that this data could be derived from the search interface.

2.1.2 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, …

No policy issue would block this and at least some of the data should be available in the summary IDS. Extensions to the summary IDS might be needed to address any missing data.

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

No policy issue would block this, but the data is not readily available. When more complete metadata tracking is performed by the experiments, this could become possible.

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

2.1.4 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, …)

No policy issue would block this and the data should be available in the summary IDS.

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

There might be policy issues related to how the data could be calculated and whether the results for different devices would be comparable. Longer term discussion on this could take place, but this query might be dropped for this project, though the project could provide a mechanism for supplying an answer if the experiments have made any such data available.

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

No blocking policy is seen --- but rather the technical issue of whether the discharge plans (which would document the goals expected) are stored by the experiments in a way that could be mined and compared to the experimental results.

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

There might be policy concerns related to this and also uncertainty about how best to generate a metric that is comparable across devices.

2.1.8 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.

No policy issues should block this but the identification of disruptions across all devices might be challenging.

2.2 Researcher finding suitable shots

Mostly this subsection returns a list of shots that meet various criteria and that could then be used to drill down to get the actual data, as described in later subsections.

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

No policy issue should block such a request. To what extent the descriptions are standardized across all of the devices is questionable, and there might also be language issues. A controlled vocabulary

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

would be useful --- the challenge is to reach agreement on such a controlled vocabulary and then to map the data from the experiments onto this.

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

No policy issue should block such a request. It must be ensured that the summary IDS contains the date and time of each discharge.

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

No policy issue should block this request. It should be possible to return the product for each device and then allow the user to sort the data. The generalization of this query would be to allow for simple mathematical operations for data that are on the same temporal grid; a further generalization might be to look at mapping data on different temporal grids to the same grid.

2.2.4 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.

No policy issue should block this request. From the data in the summary IDS it must be possible to identify shots with an H-mode; identifying the length of flat-top phases might require additional data mining effort and possible extensions to the summary IDS.

2.2.5 As a researcher, I want to obtain a list of typical discharges (shot numbers) satisfying a set of criteria, for instance a scan of one parameter or even more. As an example, on TCV, one typical shot with elongation~1.0, one at 1.1, … one at 2.8.

No policy issue should block this request. For the data available in the summary IDS, searching for shots which have a chosen parameter in a range should be possible. For the scenario described in this user story, the user could either have to repeat the search for each value (in the example, elongation) and then make a choice of his/her own of the matching shots; or get a list of shots with elongations and then refine his/her choice from that list.

2.2.6 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.

No policy issue should block this request. An issue that might need to be addressed is if the list becomes excessively long.

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

2.2.7 As a researcher, I want to obtain the parameters, and their values, used to prepare a particular discharge.

No policy issue should block this request. It is not clear that this data will be available in the initial sets of summary IDS's --- but if the experiments provide this data, providing it should not be a problem. Some way of flagging parameters "as planned" or "as performed" will be needed.

2.2.8 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.

No policy issue should block this request. This would require that the system to store data on behalf of its users and the exact mechanics of this would need to be examined in detail. Possible data size issues, period to keep the data and protection of the user's data would all need to be considered.

2.2.9 Search the EUROfusion databases for device/shots matching criteria determined by the particular database[10].

No policy issue should block this request. If all of these databases are storing their data in the summary IDS, no additional work should be required; if data is stored outside the summary IDS, work might be needed to implement this in the system

2.2.10 As a researcher, I would like to know which shots have resulted in publications and to how many publications and access to their references. And vice versa, I would like to know data from which shots have been used for a particular reference (and codes?).

No policy issue should block this request. Unless this information is already available from the experiments, it would be impossible for this project to provide this information.

2.2.11 As a researcher, I want to retrieve the information and data used for a given figure in a given paper.

No policy issue should block this request. This, however, would require a level of provenance that is, in most cases, not yet present in the information available from the experiments. The data might also be made available by the publishers in the future and some way of working with that might be necessary. Most of the laboratories (and EUROfusion) have internal pinboards where publications are tracked from the proposal to the published phase --- it might be possible to link these into the future open data system so that the use of experimental data in publications can be better tracked.

[10] 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

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

2.3 Researcher obtaining time traces

The data sought here tends to be a scalar which is a function of time.

2.3.1 As a researcher, I want to plot the plasma current[11] for a particular shot on a particular machine.

No policy issue should block this request. If the data is not in the summary IDS, then a mapping from the quantity to where it might be found would need to be maintained. It may not be necessary to distinguish between plotting and downloading the data --- this is a presentation issue which can be iteratively refined on the software providing the data to the user --- while multiple formats for data output might be required.

2.3.2 As a researcher, I would like to be able to download the current versus time data that I have just plotted, as a simple ASCII file.

No policy issue should block this request. This is a refinement of 2.3.1.

2.3.3 As a researcher, I would like to be able to download the current versus time data that I have just plotted, as a NetCDF file.

No policy issue should block this request. This is a refinement of 2.3.1.

2.3.4 As a researcher, I would like to be able to download the current versus time data that I have just plotted, as a HDF5 file.

No policy issue should block this request. This is a refinement of 2.3.1.

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

No policy issue should block this request. This is an extension of the basic scenario in that multiple signals would now be present. The user interface would need to be able to specify that more than one basic data source should be selected.

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

No policy issue should block this request. The user interface described in 2.3.5 would need to allow for selecting data from different devices.

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

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

2.3.7 As a researcher, I want to compare the time traces of H-98[12] for different shots from different machines.

No policy issue should block this request. The only issue would be the availability of the data (as a function of time) in some IDS. The derived generalization is the ability to plot any time trace data that is available from the different devices.

2.3.8 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*.).

No policy issue should block this request. Part of this is a presentation issue --- just how fancy a user interface should be supported? The easiest option might be to push anything other than simple plots to the user, allowing him/her to download the data and then use whatever tools he/she wants to use.

2.3.9 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.

No policy issue should block this request. As already mentioned, this is a presentation issue and a staged, modular approach might be foreseen --- start with a minimal set of options but allow for modules that could transform the data before supplying them to the user.

2.3.10 As a researcher, I want to have access to – and be made aware of – essential provenance metadata, including (automatic or manual) data processing code parameters and version (svn version or git hash).

A possible policy issue might be that the requested provenance data exposes protected personal data. It is also not yet guaranteed that this data will be available for all devices.

2.3.11 Researchers should have the capacity – and be encouraged – to verify that they are comparing times/shots/devices that are comparable.

No policy issue should block this request. It is not clear how the system can help the user in doing this.

2.3.12 As a reviewer or a reader, I want to have access to all the provenance metadata associated with the data used in a manuscript or publication.

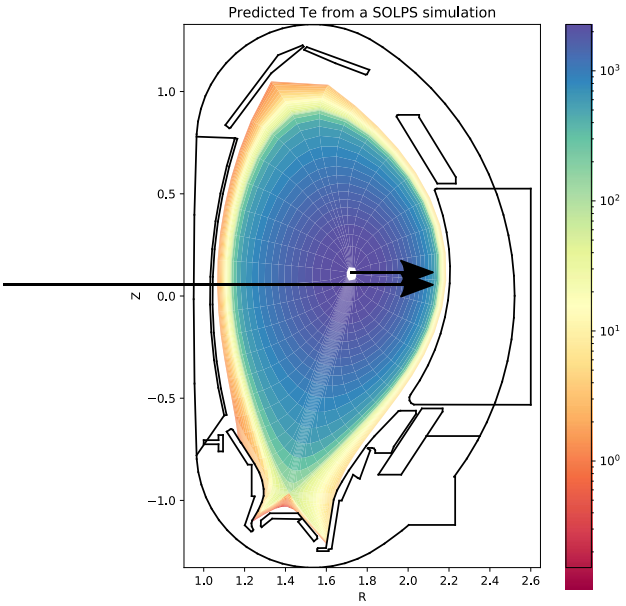
A possible policy issue might be that the requested provenance data exposes protected personal data. It is also clear that the provenance data might not be available, but this project could look into ways of providing the data if it is available.

[12] 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

2.4 Researcher obtaining profile data

This section adds the complication of the profile (radial) coordinate. In principle a number of coordinates are possible (rho_toroidal, normalised rho_toroidal, rho_poloidal, normalised rho_poloidal, rho_volume, major radius, distance between the flux surface to the magnetic axis taken at the outer mid-plane …) and different diagnostics and different experiments tend to use different coordinates. The IDS's do standardize these, but the favoured (by the user) coordinate might not be available. On the experiments this need is often met by introducing a general mapping library – but whether that can or should be considered by this project is open.



The issue with the radial coordinate can be explained by looking at the figure on the left. Some diagnostics return information as a function of (R,Z). Others are mapped, assuming symmetry on flux surfaces, to a radial coordinate that can have its origin at Z=0 (the longer arrow, corresponding to the major radius), or the magnetic axis (the shorter arrow, corresponding to the minor axis). The length of the shorter arrow can be given in meters, or it can be related to one of two magnetic fluxes (toroidal or poloidal, giving Φ or Ψ), and then converted into a distance by appropriate normalisation with the toroidal field. These two distances, and the two fluxes, can also be normalised to 0 on axis and 1 at the last closed flux surface. Another possible radial coordinate is by the normalised volume contained within the flux surface.

2.4.1 As a researcher, I want to plot density profiles[13] for a particular shot on a particular machine.

No policy issue should block this request. The basics of this are the same as for 1D data with the complication that the data might not be on a regular grid --- how to make this available and (whether to) plot the data will need to be addressed. Whether to support mappings of the radial coordinate will also need to be considered.

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

No policy issue should block this request. This is a slightly more complicated version of 2.4.1 but the only additional complication seems to be the same one mentioned in 2.3.5.

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

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

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

No policy issue should block this request. Same comment as for 2.4.2 but with the addition that radial coordinate mappings become more important.

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

No policy issue should block this request. This would require maintaining some information about where such data might be stored (different run numbers, different IDS's, different occurrences, possibly different parts of the same IDS) --- the maintenance of this information might become a problem. It might be that all this project could do is to specify that such functionality will need to be supported.

2.4.5 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.

No policy issue should block this request. This is a presentation issue --- see the comment for 2.3.9.

2.5 Researcher obtaining data needed for simulations

This section introduces the possibility of multi-dimensional data.

2.5.1 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[14].

No policy issue should block this request. If the presentation layer is built in a modular form this could be supported, but not all formats need be supported during the lifetime of this project.

2.5.2 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.[15]

No policy issue should block this request. Again this is a presentation problem --- see the comment for 2.5.1.

2.5.3 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.

No policy issue should block this request. The limitations would be the availability of the data from the experiments and the complexity of the presentation layer.

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

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

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

2.5.4 As a researcher, I would like to select cases from multiple experiments, based on the catalogued 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).

No policy issue should block this request. The complexity comes in implementing this in the presentation layer and a staged approach in building this layer will probably be necessary. While more complex cases may not be addressed in the near future, they may be taken into consideration.

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

No policy issue should block this request. It is not clear when/how the experiments will be able to provide information about data updates. While there is no need for developing all of the functionality, the issue could be taken into consideration. It is also noted that the issue of granularity might be different for different devices as to which data for a particular shot was changed versus that something for this shot might have changed.

2.5.6 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.

No policy issue should block this request. The challenge here will be whether the experiments can support this in a transparent fashion and how to store this information in the database.

2.5.7 As a researcher, I would like to know if another set of data representing the same physics quantity is available and select them (for example various I_p methods, T_e from Thomson scattering and ECE[16] (and simulation, and real-time)).

No policy issue should block this request. Here the challenge would be maintaining mappings of what data sources exist that might provide the same information. If the IDS's are built in a consistent manner (same data paths) it might be possible to automate this, otherwise the mappings would need to be maintained manually.

2.5.8 As a researcher, I want to know whether given data has been officially validated by a data processing expert.

No policy issue should block this request. This would only be possible if the experiments have a way of exposing this data and the data is made available to the system in some form.

[16] Thomson scattering and Electron Cyclotron Emission (ECE) are two ways of measuring the electron temperature in a plasma

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

2.6 Researcher providing additional meta-data or data

This section introduces the issue of user supplied annotations.

2.6.1 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;

No policy issue from the experiments should block this request. If the comments are maintained as private, no significant problems are foreseen. If the comments are public some mechanism of policing them might become necessary (the same problem any system allowing public comments has). [17]For some of the comments, a standard set of "tags" could be useful as this would make this information more interoperable. For information about publications, this information should be standardized so that it can be used to establish a provenance web.

2.6.2 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.

No policy issue should block this request. However this seems to be outside the scope of the implementation currently foreseen for this project. What might be possible would be to establish a cross-reference to data stored elsewhere.

2.6.3 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).

No policy issue should block this request. The specification of the publications should be standardized to make it most inter-operable.

2.7 Data provider

While the following are not all strictly user stories requesting something, they do provide additional constraints that need to be kept in mind.

[17] It might be necessary that public comments should only be allowed from authenticated users and the system makes clear that these are the responsibility of the person posting them.

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

2.7.1 [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.

No policy issue arises from this. It is not clear that this information is relevant for the implementation of this project but it is useful to know.

2.7.2 [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.

No policy issue arises from this. It is not clear that this information is relevant for the implementation of this project but it is useful to know.

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

No policy issue arises from this. It is not clear that this information is relevant for the implementation of this project but it is useful to know.

2.7.4 [JET] As a data provider, I currently provide external access to a subset of my local data via SAL[18].

No policy issue arises from this. It is not clear that this information is relevant for the implementation of this project but it is useful to know.

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

The project will need to ensure (help ensure) that this concern is addressed.

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

The project will need to be able to respond to push requests where data has been changed.

[18] The Simple Access Layer API.

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

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

This might impact how some operations are performed and whether caching of data might be necessary. Some form of throttling of access requests might be required and how this is managed would need to be considered.

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

The project will need to consider costs at various levels.

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

The project will need to ensure that any tools that are implemented meet these requirements as best possible.

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

This is a policy requirement that will need to be enforced. Some mechanism for the devices to propagate information about embargo periods will need to be provided, also allowing for these to be changed.

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

This is a policy issue that relates to authentication and authorization. The project will need to consider it, even if not implemented.

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

The level of logging will need to be considered by the project, and how to make this information available to stakeholders.

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

The level of logging will need to be considered by the project, and how to make this information available to stakeholders.

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

2.7.14 More generally, as a community of data providers, we would like to build and access statistical analytics about portal usage.

The level of logging will need to be considered by the project, and how to make this information available to stakeholders.

2.8 FAIR User Stories

This section takes a different perspective in that it specifically does not cover the aspect related to requirements from researchers, but instead takes the approach of what elements of FAIR are needed from the perspective of a data manager. At the end of each user story, the elements of FAIR addressed are shown in brackets. Note that there is some overlap with earlier user stories, but this is preserved as these cases arose separately and we would like to preserve this.

2.8.1 Data Manager Stories

1. As a local data manager responsible for compliance with policy, I want to make sure that I can make my data available in a manner which complies with my local, national, regional or global responsibilities. This includes making data 'open' after a defined embargo period (or sooner if policy requires), restricting access to sensitive data (primarily this will be commercially sensitive data) to authorized users both within my local organization and with collaborating individuals, groups or organizations, and ensuring that any data released is compliant with current restrictions on the release of personal information. (**FA**)
2. As a local data manager I want to know who has accessed the data and to understand how many users are researchers from the fusion community, how many from other research areas, and how many are from the general public so that I can report this to the institute manager and funder. (**A**)
3. As a data manager, I want to be to able to report what data has been used in publications, whether those publications are available through open access or not to demonstrate to funders the importance of the work I am doing, as well as being able to ensure that correct use has been made of the data by the author(s). (**AIR**)
4. As a data manager I want to ensure that people who are using data which I am responsible for always have access to the most up-to-date version, but that users of older data from the same data set (e.g. data that has undergone some form of reprocessing), are still able to access this data.(**FAR**)
5. As a data manager I want to ensure that data under embargo does not leave my local site and access is only granted through well defined APIs which I can control. The same holds for any non open data as some of the work at my institute is funded nationally and we are not required to make this data open to the public. (**A**)

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

2.8.2 Funder Stories

1. As a funder I want to know how the data produced by an experiment I am funding is used for either scientific, engineering or industrial purposes in order to assess whether continued funding is justified. In order to assess this I would like to know how much data has been used in publications, how often data sets are accessed, how many shots are deemed of scientific value and how many were aborted, and understand the purposes of each experimental campaign to ensure money is being used as efficiently and possible. (**IR**)
2. As a funder I want to know how much of the data is accessed by researchers in my country and how much is accessed by external researchers since I need to ensure money supporting fusion research is being well spent in university departments and research organizations associated with fusion research departments, and to demonstrate that we are collaborating on an international basis to both continue leading aspects of this work and are working towards commercialization of fusion energy. (**IR**)
3. As a funder I want to ensure my site is compliant with local, national and international obligations in respect of providing data appropriately. This may include releasing data for educational purposes, to support industry or ensuring that only identified data is released. (**FAIR**)
4. As a funder I want to ensure that data is appropriately licensed regardless of whether it is publicly accessible or not. (**R**)

2.8.3 Non-Fusion researcher stories

In these stories the term researcher refers specifically to non-fusion researchers who might want to make use of the data. I have tried to use appropriate use cases where I believe there is at least a moderate level of overlap with the fusion community. This list is definitely non exhaustive and highly speculative.

1. As an astronomer interested in modelling H I regions, or one interested in modelling stellar atmospheres which share similar properties to magnetically confined plasmas, I am interested in comparing results I obtain with those obtained in tokamaks and understanding what codes are used for calculating transport and radiative transfer. While the data itself might be of little benefit for my research, information on which codes were is of interest to identify potential partners to develop codes which are applicable to both fields; I would therefore like to see the provenance of data derived from different tokamak devices. (**FR**)
2. As an instrument scientist on a synchrotron light source (e.g. ILL) who has to perform tomographic reconstruction, I am interested in the code used in fusion research which also performs this type of analysis. I would like to be able to use the provenance information to find related publications and see whether similar reconstruction techniques could be used within my own institute. (**FR**)
3. As an astronomer working in the field or coronal mass ejections from twisted magnetic fields, I am interested in investigating the causes of these events and whether disruption or ELM events in plasma physics can be used as a model for these larger scale events. To do this I

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

need to have a good understanding of the different diagnostics and be able to find these events easily using an interface which is aimed at non-fusion users, and I would like the data to either be in a self describing format or with a well documented format that I can use to interpret the results. I would also like to understand what modelling codes are used within the fusion community for modelling these events.

4. As an educator I am interested in making use of data, either from modelling or experiments, to educate students about basic nuclear fusion and support their learning with data. I am primarily interested in audio-visual data, but also in obtaining some interesting parameters such as plasma current or electron density in a simple format suitable for minor project work. However, as I am not familiar with many of the terms used in plasma and fusion physics in general, I would need a simple guide to these terms. As most of this work will use tools such as excel, being able to access AV data and contextual data in common formats (e.g. JPEG, MPEG4, CSV, XLSD) would make the data more accessible. (**FAR**)
5. As a citizen scientist I am looking at the history of fusion research to form my own conclusions about how we are progressing towards commercialization. I understand there is a value Q which is the ratio of output power to input power and also the length of a ‘pulse’ or ‘shot’ varies between experiments. I would like to be able to get these values for all of the different tokamaks as a table with columns (Experiment, Date, Pulse Number, Q Value, Shot Duration). Based on initial analysis I might want to do the same again but strip out shots from each experiment less than a certain duration which I interpret as not being of interest. If the Q value is not available I would like to know how to estimate it based on the parameters which I can find and download. (**FA**)
6. As a potential supplier I am interested in knowing the volumes of various materials consumed during plasma operation, specifically hydrogen and deuterium or other gasses/liquids consumed, so that I can assess whether I am in a position to supply these to an experimental facility in the future. In particular I am interested in finding out how much of each resource is consumed in a typical experimental ‘pulse’ and how many ‘pluses’ are produced each day (**FA**)

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

3 Summary and Outlook

This document provides the final report on a number of use cases that could be considered for implementation by the Fair4Fusion project. A later deliverable[19] 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

[19] D2.4 Characterization of non-experimental data (M18) Report on what non-experimental data (e.g. from modelling) is being produced, and what steps would be necessary to bring it into the developed open data framework. (MPIPP)

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

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

A.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

Shot	B_t [T]	I_p [MA]	n_e [10^{19} /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

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

Figure 2: Example of a search box from AUG.

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

A.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

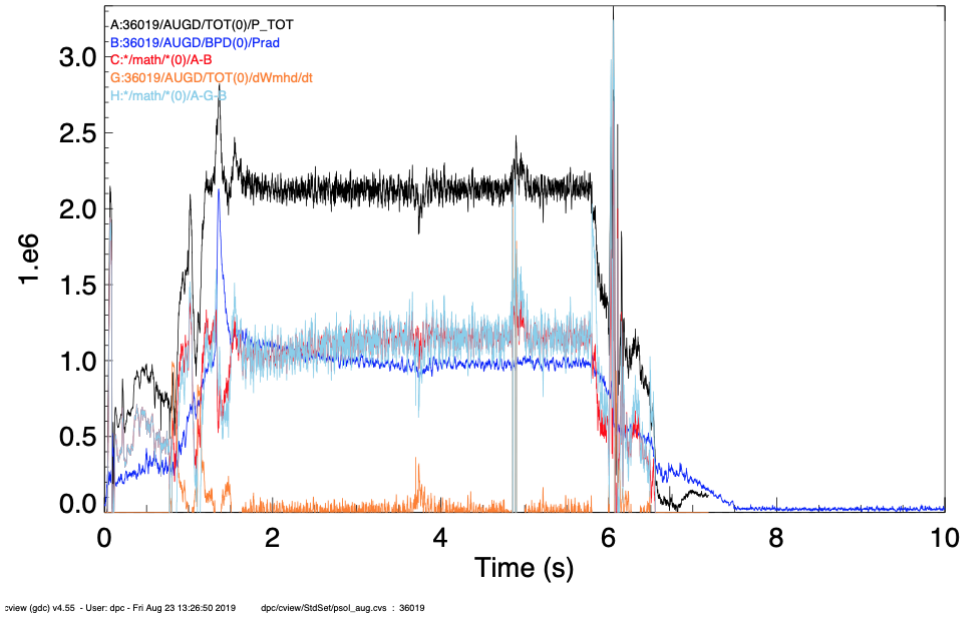


Figure 3: Example time traces plot.

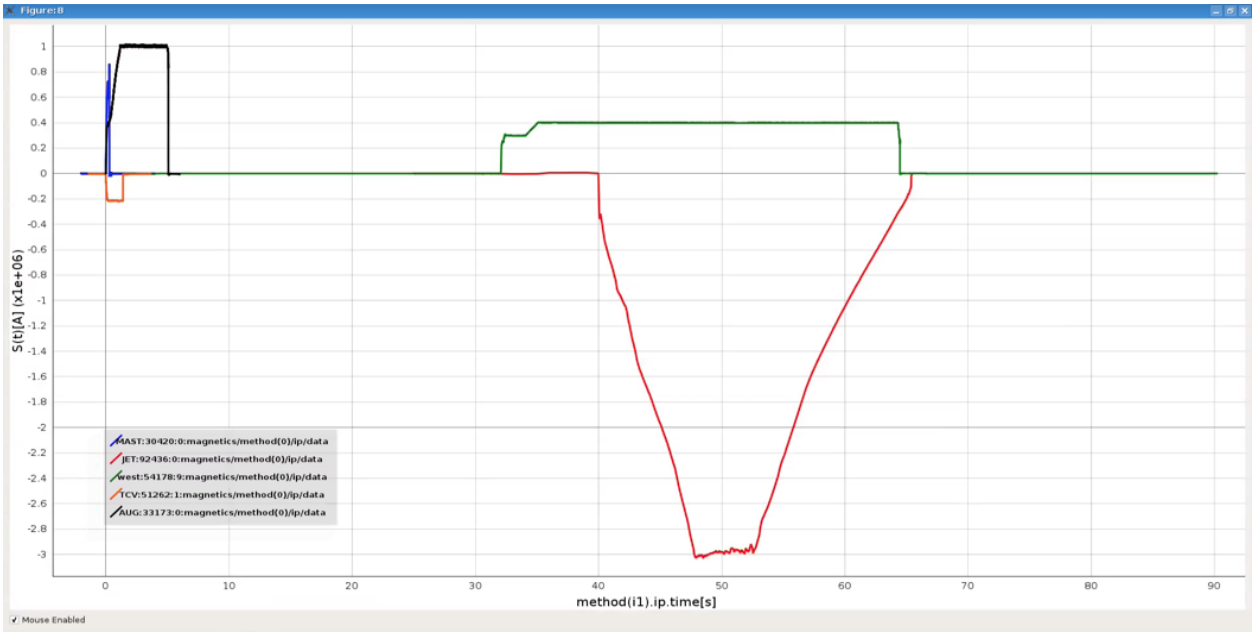


Figure 4: Another example of a time traces plot.

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

A.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

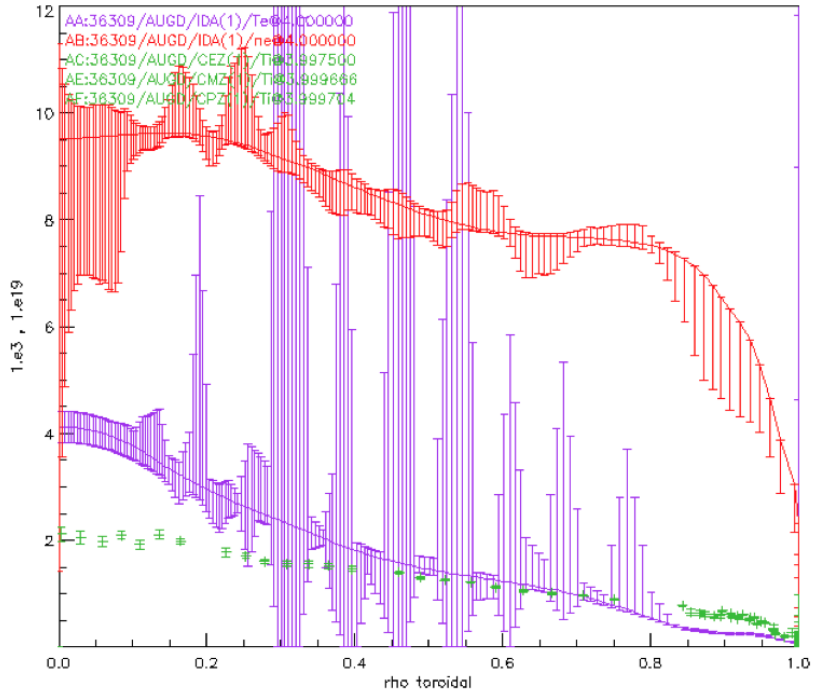


Figure 5: Example of a plot showing a number of profiles.

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

A.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:

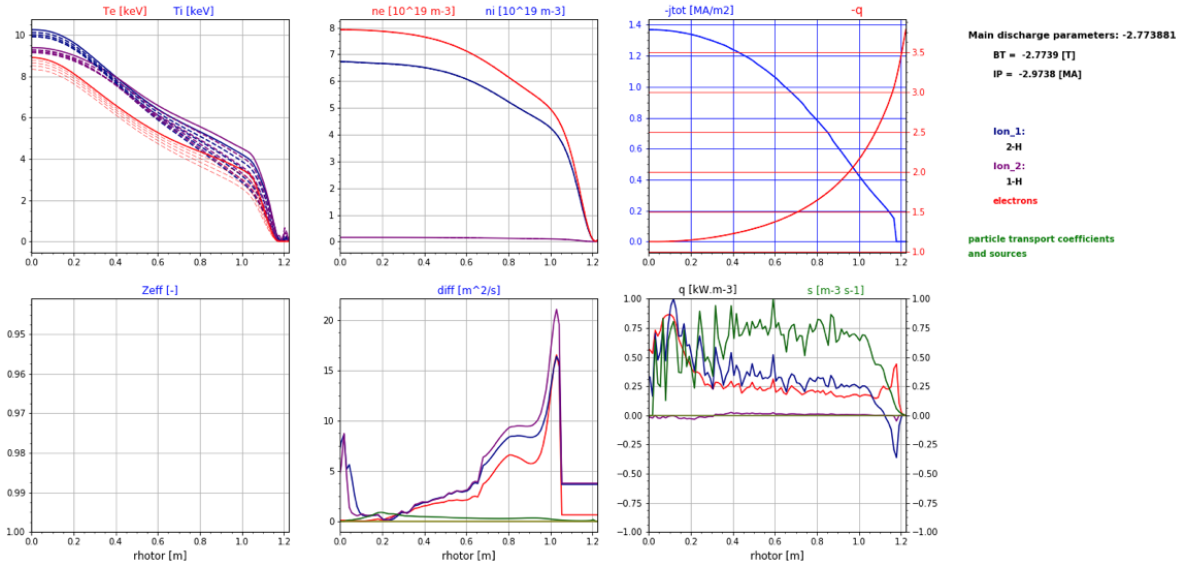


Figure 6: Example of simulation output.

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

A.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

Appendix B Background technical information

B.1 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
- Almost all (all?) of the machines currently also supports remote access to data via MDSplus[20] 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”[21]
- Almost all (all?) of the machines also have a way of searching for data (see A.1 for an example). These tend to be different for each machine in available criteria and method of implementation.
- Within EUROfusion there is an effort to provide data via UDA (Unified Data Access)[22].

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

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

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