Ref. Ares(2020)3059127 - 12/06/2020

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"

D4.1 - Architecture for a Metadata Interoperability Service and Proposed Metadata Model

WP4

Deliverable status	Final
Type	Report
Dissemination level (according to the proposal)	Public
Work Package	WP4 - Data Foundations for Open Access of Fusion Data
Lead Beneficiary (deliverable)	2 - UKAEA
Due Date	29/05/2020
Date of submission	12/06/2020

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)
Shaun de Witt	**UKAEA**	**Shaun.de-witt@ukaea.uk** **+44 1235 464585**
George Gibbons	**UKAEA**	George.gibbons@ukaea.uk
Frederic Imbeaux	**CEA**	frederic.imbeaux@cea.fr +33 4 42 25 63 26
Antonis Koukourikos	**NCSR-D**	**kukurik@iit.demokritos.gr**
Iraklis Klampanos	**NCSR-D**	**aklampanos@iit.demokritos.gr**
Andreas Ikonomopoulos	**NCSR-D**	**anikon@ipta.demokritos.gr**

DOCUMENT CONTROL

Document version	Date	Author/Reviewer – Organisation	Change
V1	**2020-05-22**	**Shaun de Witt – UKAEA** **George Gibbons - UKAEA**	First version
V2	**2020-05-29**	**Frederic Imbeaux – CEA** **Irakalis Klampanos – NCSR-D**	Added additional input
V3	**2020-06-02**	**Shaun de Witt - UKAEA**	Additional input incorporated into word version
V4	**2020-06-09**	**Shaun de Witt - UKAEA** **Marcin Plociennik (PSNC)** **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	Metadata, Search, Interoperability, Nuclear Fusion
Point of contact	Name: Shaun de Witt Partner: UKAEA Address: Culham Science Centre, Abingdon, OXON, UK Phone: +44 (0)1235 464585 E-mailshaun.de-witt@ukaea.uk
Delivery date	2020-06-11

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
STEM	Science, Technology, Engineering and Mathematics
OAI-PMH	Open Access Initiative Protocol for Metadata Harvesting
REST	Representation State Transfer
MAST	Mega-Amp Spherical Tokamak
JET	Joint European Torus
COCOS	Coordinate Conventions
EFDA	European Fusion Development Agreement
IDS	Interface Data Structure
CERIF	Common European Research Information Format
XML	eXtensible Markup Language
EOSC	European Open Science Cloud
EPFL	Ecole polytechnique fédérale de Lausanne
PID	Persistent Identifier
EDMI	EOSC Dataset Minimum Information
METS	Metadata Encoding and Transmission Format
MODS	Metadata Object Description Schema
RDA	Research Data Alliance
FGDC	Federal Geographic Data Center
FAIR	Findable, Accessible, Interoperable and Reusable

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

Executive Summary

Metadata is one of the key elements in making data FAIR (Findable, Accessible, Interoperable and Reusable) but currently across the fusion community there is not a widely adopted standard which would allow easy cross site discovery of data of interest, largely due to the presence of long standing existing practices. This document presents a proposed metadata model to overcome this issue. This model has been built based on existing schemas previously designed to support the fusion community with the involvement of a wide range of stakeholders from across the European Fusion community. To support 'open data' and ease discovery for non community users we have also adopted some elements of Dublin Core Elements, which also allow us to some very basic provenance tracking of versions. A full provenance model will be incorporated in a later deliverable.

While as a project we are not recommending sites change their existing practices of even adopt this schema, we propose to aggregate metadata from sites and convert it to this schema to allow for simpler cross site discovery and access, something which is currently impossible to do.

In conclusion we find that the use of the Summary IDS which has been co-developed with the fusion community is suitable as a metadata schema for the community, providing a rich set of metadata, a few additions are required to better support the FAIR principles such as provision for a globally unique, actionable persistent identifier, suitable licensing information and provenance.

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

1 Introduction

This document details the work carried out into the definition of a common metadata standard which is both FAIR compliant and acceptable to the fusion community. FAIR (Findable, Accessible, Interoperable and Reusable)[1,2] and 'open' data have become the standard in many scientific disciplines driven by a desire to maximise the economic benefit of publicly funded data, promote better science by allowing additional scrutiny of data, enhance the opportunities for cross disciplinary research and help encourage careers in STEM subjects globally. An additional benefit of ensuring data the FAIR is that it can make research more efficient and ease the burden of training on the next generation of researchers. These fundamentals are as true for nuclear fusion as for any other discipline generating data.
A significant part of ensuring data is FAIR is the definition of an agreed common metadata standard; indeed eleven of the 14 FAIR principles defined by Force 11 relate to the content, management and accessibility of metadata. As a key element much of the first part of this project has been devoted to agreeing on a standard and working with the experimental sites to ensure that this standard is acceptable to them, without requiring them to adopt it in full (q.v. Section 3.2). This work is still ongoing and subject to some revisions as will be outlined later.

1.1 Background to Metadata in Fusion Research

Across Europe, and globally, there are relatively few fusion experiments, but all of them have the same long term goal – to create the science and engineering base to realise commercial fusion energy. And, like many such experiments where instrumentation is scarce such as particle physics and astronomy, different sites work together in a collaborative/competitive way. Across Europe, fusion research has been governed overall by the European Atomic Energy Community (EURATOM) since the Treaty of Rome in 1957. Since then a number of fusion experiments have been built across different sites in Europe, ranging from the EC funded Joint European Torus to nationally funded devices such as the Wendelstein-7X stellarator, the spherical tokamak MAST in the UK and the tungsten divertor device WEST in France. In addition, a large number of smaller devices using different technologies are used at other sites around Europe. In most cases, these device are either wholly funded by national or institutional bodies or as part funded between national bodies and the EC through projects such as EFDA or EUROfusion.
In addition to these devices being developed somewhat autonomously, various parameters relating to the plasma and the reactor vessel are monitored using a number of diagnostic techniques. While many of these are common to all experimental devices, there are a number that are unique and even where these parameters and diagnostics may be recorded differently due to difference in details of construction, convention or monitoring. While COCOS[3] has provided conventions on coordinate

[1] https://www.force11.org/group/fairgroup/fairprinciples
[2] https://repository.eoscsecretariat.eu/index.php/s/C3a5WkpsFHL6GD3
[3] O.Sauter and S. Yu. Medvedev, "Tokamak Coordinate Conventions: COCOS", Computer Physics Conventions, Volume 184, Issue 2, February 2013, Pages 293-302

systems which have been adopted widely, and permit simple mapping between these conventions, no similar standard exists for metadata. Even with essential parametric outputs for parameters such as the plasma current Ip, these are presented with different sign conventions by different devices.

Recent work on standardisation has been driven by ITER, the next generation of tokamak device. With the support of EUROfusion and in the frame of the ITER Integrated Modelling and Analysis Suite (IMAS), a device-neutral ontology known as the IMAS Data Dictionary has been developed. While still not widely adopted as a native format, work has been ongoing into allowing access to data using IMAS Data Dictionary naming conventions and providing mappings between local naming conventions and the Interface Data Structures (IDS), which are high level structured objects defined in the IMAS Data Dictionary. Currently the IMAS Data Dictionary is licensed by the ITER consortium and is not openly available, but this group and EUROfusion are discussing making it open to adhere to FAIR principles.

1.2 Introduction to FAIR

The acronym FAIR is now widely known across Europe and wider by many science communities and provides a framework of policies for making data Findable, Accessible, Interoperable and Reusable. This has largely been driven by the G8 declaration on open data, tempered with some realisation that both making data 'open' is insufficient to provide the expected scientific and economic benefits, and that not all data can be made 'open'. FAIR provides a framework for easing discovery of data, encourage suitable licensing and ensure that data (or information about the data) can persist over time spans of ten years or more as well as ensuring suitable Authentication and Authorization processes are in place. For data to be FAIR there are 15 policies which should be adhere to, and most of these relate in some way to either metadata, persistent identifiers and licensing. However, there have been many nuances and interpretation of these, notably from the Research Data Alliance Working Group on Fair Data Maturity Model[4] and the ESOC Secretariat FAIR Working Group recommendations on FAIR metrics for EOSC[5] which add a level of complexity and clarity.

1.1.1 Common Metadata Standards

As previously stated, metadata is central to making data FAIR and much work has been published on metadata standards some of which has become widely adopted and some of which has had a distinct lack of uptake. Arguably the first serious attempt to define an overall metadata standard is the well-known Dublin Core Standard[6], which was primarily developed by librarians for publications. While this has been extended it still remains a core to publication metadata despite some clear drawbacks (such as all the metadata elements being optional and free test). However, for many science disciplines Dublin Core and Dublin Core Elements do not provide sufficient richness to describe scientific data. There was work funded by the commission on trying to define an all-encompassing high level standard CERIF[7]; unlike Dublin Cores 15 simple elements, CERIF has 293 metadata entries and 1814 metadata

[4] doi:// 10.15497/rda00045
[5] https://repository.eoscsecretariat.eu/index.php/s/C3a5WkpsFHL6GD3
[6] http://www.dlib.org/dlib/July95/07weibel.html
[7] https://www.eurocris.org/cerif/main-features-cerif

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

attributes. However, it's complexity and deliberately abstract nature means it has not received such widespread adoption. Even before Dublin core, the importance of metadata was realised by the geospatial community in the United States which led to the 300 element FGDC metadata standard in 1995, which demonstrates that even for a relatively isolated parameter space, rich descriptions are essential. The Digital Curation Centre has published a list of almost 40 metadata schemas in use today[8], and this is far from exhaustive.

In recent years, in part driven for the need for FAIR compliance and also the realisation that metadata is primarily now read by services rather than humans, metadata standards are no longer just published in papers to be followed but are defined in machine readable schemas in commonly used formats such as XML Schema Documents, JSON schemas and RDF. This has the significant advantage of minimising errors in the metadata by being able to validate both its format and values against such a schema. The metadata itself references the web accessible schema which allows multiple versions of a schema to be concurrently supported.

1.1.2 Persistent Identifiers

The other core pillar of ensuring FAIR compliance is the provision of persistent identifiers to both the data and metadata. A persistent identifier is an globally unique, actionable identifier which resolves to a dataset (or a landing page with further information regarding the dataset). In terms of FAIR the use of persistent identifiers provides three main features – to make (meta)data available regardless of location using a resolver, support reproducible science by ensuring data used in a publication can be used to repeat an analysis and also as a citation mechanism to demonstrate to funders that the research is valuable and is achieving impact. A more detailed discussion of PID technologies will be provided in a later deliverable.

Having said that PID's are also a key element of metadata since the RDA recommendation call for persistent identifiers on both data and metadata, so in the design of a proposed schema we need to ensure this can be accommodated. As we will see later, this does potentially have an impact on both the architecture of a final FAIR portal and potentially on operations at sites. We have based significant parts of the proposed schema on part of the IDS structure known as the Summary IDS. However, these summary IDS's are immutable and putting PID's within them may accidentally enforce changes in data management policies at sites. We discuss this later in more details and potential ways of overcoming this.

1.1.3 Provenance

While provenance is an important part of a metadata schema, it is deliberately excluded from this document as it is the subject of a later deliverable. However, work has already commenced in extending the proposed metadata model to include provenance information already captured by experiments and to promote the capture of provenance information at least within automated processing chains and extending to sensor or machine specific provenance information. However, to develop and populate a full provenance model will require more implementation at sites than will be possible during the lifetime of this work and will take some time to adopt any recommendations. Nonetheless within the deliverable 4.3 to come we will propose a full provenance model based on

[8] https://www.dcc.ac.uk/guidance/standards/metadata/list

PROV and show how this can be populated with existing data, which will allow us to determine what gaps exist.

1.2 Fusion Specific Restrictions

Fusion is not unique in considering much of its data 'sensitive' although unlike communities such as linguistics or medicine, this is more due to commercial interested than protecting personal privacy. Aside from the JET experiment, the other major European Tokamak devices have been and continue to be funded in part through national funding bodies. This national funding is clearly through a level of self-interest; the first country to create a commercial fusion device would clearly have a significant lead over others given the potential of this technology for sustainable, clean, predictable energy for the foreseeable future. This combined with the legacy of being part of the nuclear industry means that most countries currently do not have a policy of making fusion data available publicly, even where there is an open data mandate for publicly funded data. There are some exceptions to this – for instance the MAST experiment in the UK has a policy of releasing data after a 3 year embargo period and EPFL are working towards an open data regime after a suitable embargo period.

1.3 Structure of This Document

The remainder of this document is structured as follows. Section 2 introduces the proposed metadata model, indicating how the schema has evolved and looking at how this will help empower the fusion community and address FAIR issues in a wider context. Section 3 addresses how we have attempted to minimise the impact of existing site practices and how we can use existing tools to integrate this model into a searchable dashboard. We also look at some of the outstanding issues which need to be resolved and discuss how this model impacts existing tools which will feed into the demonstrator and the overall architecture.

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

2 Proposed Metadata Model

2.1 The Interface Data Structure

Within the IMAS Data Dictionary, some structures are marked as Interface Data Structure (IDS), a very important notion. An IDS is an entry point of the Data Dictionary that can be used as a single entity to be used by a user. Examples are the full description of a tokamak subsystem (diagnostic, heating system, …) or an abstract physical concept (equilibrium, set of core plasma profiles, wave propagation, …). This concept allows tracing of data provenance and allows a simple transfer of large numbers of variables between loosely or tightly coupled applications. The IDS thereby define standardized interface points between IMAS physics components.

2.1.1 IDS Summary Metadata

Within the IMAS Data Dictionary, the Summary IDS is the placeholder for physical metadata summarizing an experiment or a simulation. It contains time traces of several global, local or space-averaged physical quantities that physicists typically use to search plasma experiments of interest. In addition to the value of each quantity, there are also placeholders for error bars and provenance information (a simple string so far). Being defined in a machine-generic way and usable for both experiments and simulations, we propose to use this ontology as the standard for metadata for making European fusion experiments data open.
The full structure of the Summary IDS is given in a separate project milestone, but some information is available in an early publication from Frederic Imbeaux[9].

2.2 Extending the Summary IDS to Support FAIR principles

The Summary IDS provides a large coverage of the physics quantities that can be captured in fusion experiments but needs improvement when it comes to more generic documentation that will help make the data more findable and accessible to non-fusion users, including funders, other researchers and the general public. Facilitating this will require additional non-physical terms to be added either to the summary IDS or as additional searchable parameters linked to the data.

Based on some the requirements identified in Work Package 2, and following investigation of some non-discipline specific schemas including Dublin Core, Datacite, the EOSC-EDMI, MODS and METS. Full comparisons of these are widely available and will not be discussed in this document. We also considered various standards from the energy sector including the Energy Industry Profile (rejected since this is too petrochemical focussed) and European Statistical Data and Metadata Exchange Metadata Structure (which is too focussed on energy generators), in the hope of increasing interoperability of metadata between these sectors.

Based on the requirements we have selected a number of Dublin Core Elements to extend the existing IDS Summary Schema. Dublin Core have curated a list of generic metadata terms known as DCMI

[9] F. Imbeaux et al 2015 Nucl. Fusion 55 123006 https://doi.org/10.1088/0029-5515/55/12/123006

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

Metadata Terms (superseded qualified Dublin Core in 2008) based on the smaller Dublin Core Metadata Element Set (DCMES). Whilst, DCMI only has two compulsory terms it is understood that by using the generic terms provided by DCMI we will improve the interoperability of the metadata schema with other schemas. As a generic schema not all DCMI terms apply to fusion but by comparing the DCMI terms and the Summary IDS a subset of DCMI can be selected to improve the FAIRness of the proposed fusion metadata schema. Only the terms from DCMI that have been selected for implementation will be discussed. This is a similar approach taken in the Wind Energy sector which extended Dublin Core with five additional parameters[10].

The table below shows the list of the Dublin Core Elements we propose to make use of, whether they map to existing fields in the Summary IDS and a link to the associated requirements in WP2.

Dublin Core Term	IDS Structural Element	Notes
Description	ids_properties/comment	Free text
Source	ids_properties/source	Source of the data (any comment describing the origin of the data : code, path to diagnostic signals, processing method, ...)
Creator	ids_properties/provider	Name of the person in charge of producing this data; for experimental data this could be the name of the device of the legal entity hosting the device
Created	ids_properties/creation_date	IDS does not specify format for the date. For search purpose, these will be converted to ISO Date format YYYY-MM-DD
Identifier		A valid persistent identifier. This field will be added to the Summary IDS
Replaces		A valid persistent identifier referencing a previous version of this shot
isReplacedBy		A valid persistent identifier indicating that this data has a later version
accessRights		Who is entitled to access the data; examples could be PUBLIC, COMMUNITY, COLLABORATOR or RESTRICTED

[10] "Taxonomy and metadata for wind energy Research &Development", 10.5281/zenodo.1199489

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

Valid		The dates between which data is valid; used in conjunction with the replaces/isReplacedBy tuple
rightsholder		The legal entity owning of the data
License		What license the data is issued under
available		The latest date after which data will be made publicly accessible
isReferencedBy		Publications (whether external or internal) referencing the data. This can be obtained from site pinboards where publications are put up for review
rightsHolder		The organisation owning or managing rights over a resource (the data in this case)

In addition, a few terms which do not appear in the Summary IDS are important in data identification are included; *experiment* and *shot.* The former represents an the experimental device on which the data was obtained, while the latter represents a distinct run of an experimental device (also known as a pulse on JET). It should be noted that the definition of 'shot' will be, and already is to some extent, anachronistic. As pulses become longer and eventually operations become continuous shots will become more like time slices based either on wallclock time or delineated based on event. Already on JET and some other devices some information is collected continuously and used in level 0 processing of raw data to physics products.

The identifier (PID) can be expected to remain constant throughout the lifetime of the dataset therefore it can be added directly into the Summary IDS. However, due to the immutability of the physics metadata once initially stored, a new way of storing the dynamic metadata demanded to satisfy the FAIR principles is required. Dynamic metadata are defined as metadata that can evolve (e.g. Replaces or isReplacedBy) or be appended later in time (e.g. tags, isReferencedBy).

There is also a requirement to be able to add annotations to the metadata. Sometimes after performing complex calculations with one of the datasets it is useful to let other users know about the results/non-results of the computation. Annotations are another dynamic quantity that can change with time and so would need to be included outside the summary IDS.

It should be made clear that where we have identified metadata elements which are not are part of the Summary IDS (or a planned addition), these elements are used primarily for searching. With extensions, it would be possible to add these to a Summary IDS returned to a user by creating 'placeholders' which are not filled by the experiments but could be filled is as part of a data retrieval service. However, this is an architectural discussion outside the scope of this document.

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

2.3 Defining Searchable Metadata

A study was carried out to see how the individual experiments allowed users to search through their metadata. A total of four experiments were surveyed (WEST, JET, MAST-U and ASDEX-U) and each term mapped onto the Summary IDS.

Each experiment's searchable metadata mainly focused on the physics summary parameters such as the average plasma current for a shot and there was little focus on more generic metadata. This meant the study soon morphed into a comparison of these physical parameters. A common set of these terms (which were made searchable by each experiment) was then formulated although there was no guarantee that the values were measured in the same way. Continuing the plasma current example this can be taken when the shot is in the flat top phase but it is likely that each experiment has subtly different definitions of this. In fact the method of measurement may not even be the same. This is not an issue though, since information on how the data was obtained can be added in the "source" node attached to each "value" node in the Summary IDS. Examples of commonly searchable physics parameters can be found below along with their Summary IDS mappings.

Commonly Searchable Physics Parameters	Mapping to Summary IDS	Data Dictionary Definition - 3.26.0
Plasma Current	global_quantities/ip	Total plasma current [A] - FLT_1D
Toroidal Field	global_quantities/b0	Vacuum toroidal field at R0. Positive sign means anti-clockwise when viewed from above. The product R0B0 must be consistent with the b_tor_vacuum_r field of the tf IDS. [T] - FLT_1D
Toroidal Beta	global_quantities/beta_tor	Toroidal beta, defined as the volume-averaged total perpendicular pressure divided by (B0^2/(2*mu0)), i.e. beta_toroidal = 2 mu0 int(p dV) / V / B0^2 [-] - FLT_1D
NBI Power	heating_current_drive/power_nbi	Total NBI power coupled to the plasma [W] - FLT_1D

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

Elongation	boundary/elongation	Elongation of the plasma boundary [-] - FLT_1D
Electron Density	local/magnetic_axis/n_e	Electron density at the magnetic axis [m^-3] - FLT_1D

Presently, experiments tend to provide only a fraction of the physical metadata available in the Summary IDS. This fraction will need to be augmented to allow for wider and more flexible multi-machine queries. However, providing all quantities present in the Summary IDS is not a requirement, but rather a long term goal in view of increasing Interoperability across EU experiments.

All the terms in the Summary IDS are searchable by the CatalogQT tool in which we will store the experimental metadata. However, using this method gives an indication about which fields of the Summary IDS would be most relevant for users of the metadata to search by and gave a good starting point for coming up with a minimum set of physics parameters that the sites would have to be able to provide.

2.4 Versioning of Data

Versioning is essential as data is reprocessed at irregular intervals due to revised calibration information or due to incorrect transcriptions or use of improved algorithms with revised physics. While this is not an uncommon problem, what makes this difficult in the case of fusion is it is very rare that a whole dataset will change in response to such a reprocessing. Normally it will only impact a single diagnostic or a single physical parameter. Each site also handles versioning differently due to different models. In our proposed model we would use the metadata elements *replaces* and *isReplacedBy* to record these changes. These would not be provided by sites in the Summary IDS but would be generated within the architecture when new metadata arrives with the same experiment-shot tuple, but with a different persistent identifier. Internally we would then store generate the version information (including the *validity* field). This is shown diagrammatically in Figure 1 below.

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

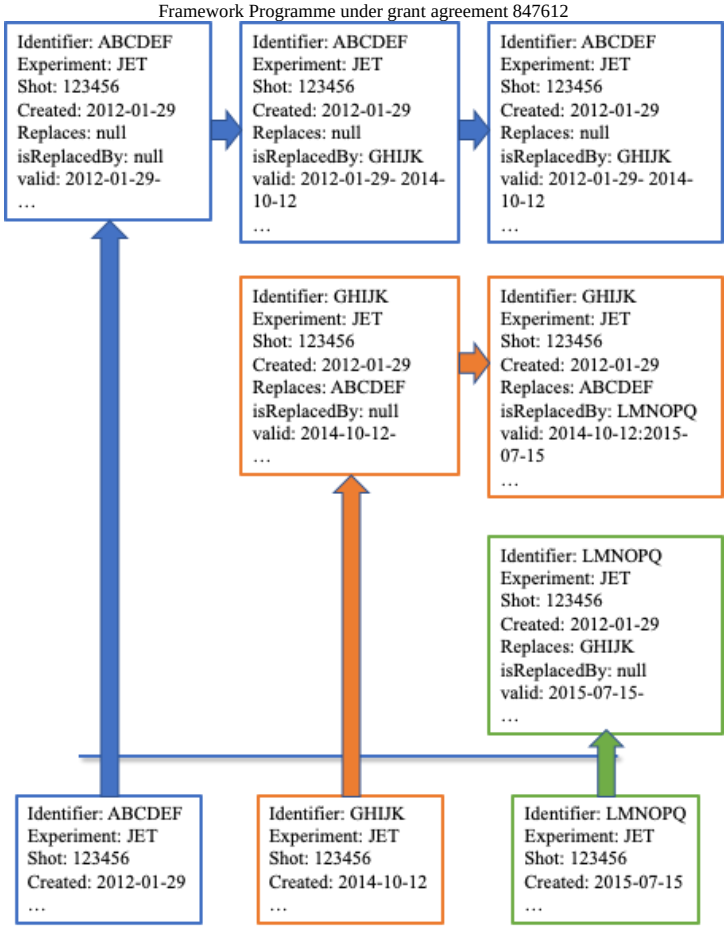

Figure 1: Proposed Handling of Versioning

In this figure the blue text boxes represent the first version of the data, the orange the second version and the green the third version. The key points to note are that the *created* date does not change through the version but the *valid* field is used to track the dates at which new versions are generated and the use of *replaces* and *isReplacedBy* to allow tracing to previous versions. A date search will use the *valid* element to get the latest version, rather than the *created* element.

3 Interoperability Considerations

3.1 Retrieving Metadata from Sites

During the course of this project we have discussed with the experiments the responsibilities between the site and the tools being developed as a part of this project, ensuring that any changes required by a site are minimised since the additional costs and risks associated with changing existing well defined practices is deemed unacceptable. It is clear that the sites will maintain responsibility for data storage and metadata generation. As a part of the work of this project we need to be able to aggregate the metadata from these sites.

Two considerations have been made to retrieving metadata, namely how metadata is sent to the central aggregator (i.e. push/pull models) and an investigation into different metadata harvesting techniques.

3.1.1 Push vs Pull Models

There are careful considerations as to whether metadata should be pushed from a site to a central aggregator or pulled by an aggregator from the repository site. The pull model, where the aggregator pulls information from the site hosting the data can make for a more reliable service since transient events can be better dealt with and accidental DoS events between the aggregator and site can be controlled. However, it would potentially mean sites having to modify their existing metadata infrastructures in the case where data is a mix of commercially sensitive and more open data which it is unlikely sites would accept. The alternative, where sites push data to a central aggregator is also not without cost to the sites since this push service would become an additional production service which would need monitoring. However, it does give sites more freedom as to when metadata can be pushed to the central aggregator, doing this during the evening so as not to interfere with ongoing operations.

In reality the project cannot dictate to sites which method to choose but can only make recommendations. If the Universal Data Access layer of IMAS is used to gather data from sites before conversion to Summary IDS this may be more easily done by making pull requests, while for data sets already adhering to the IMAS standards, either push or pull would be quite possible.

3.1.2 Harvesting Techniques

By far the most standard for metadata harvesting is the OAI-PMH standard[11] developed by the Open Access Initiative. While widely adopted it is not believed to be suitable since it is very much tied to a pull model of metadata harvesting (Figure 2). While this presents a well defined standard both in terms of requests and responses, it I not clear if it is suitable for the current situation since it assumes that the repository is open (or at least externally accessible) which may not be the case at fusion sites. In addition, it is somewhat of a legacy protocol, not being based on REST calls.

ResourceSync[12] is a NISO standard which is based on more modern technologies and allows sites a greater level of control over what and when they publish information to a central aggregator based on REST frameworks. It seems like this can accommodate the use of UDA to provide metadata translation between site specific format and the summary IDS.

[11] https://www.openarchives.org/pmh/
[12] http://www.openarchives.org/rs/toc

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

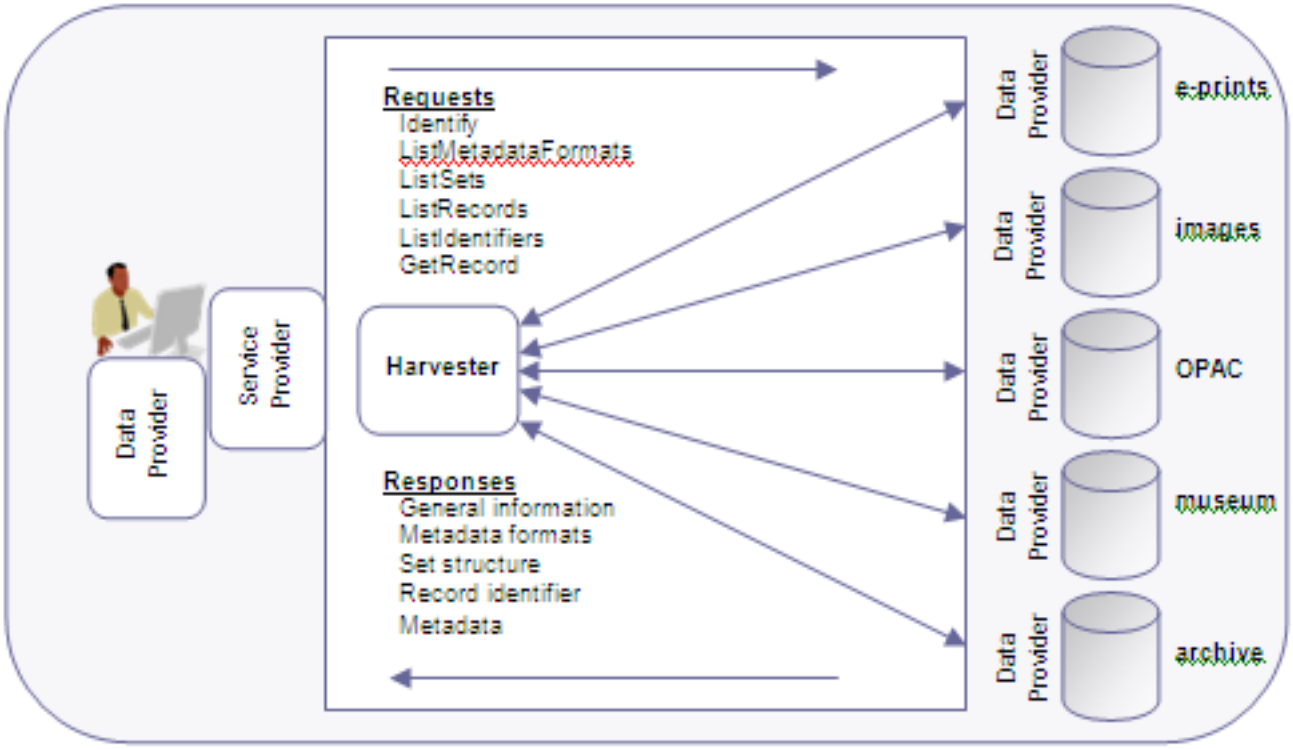

Figure 2: OAI-PMH Structure ((c) Herriot-Watt University, CC-BT-SA 3.0)

3.2 Metadata Conversion Techniques

At the moment only WEST directly outputs its data in the IMAS format. Any metadata we get from the other experiments will have to be converted to IDS. Work has already been done in conjunction with the data providers to start mapping their data into different IDSs. Furthermore, work on this will be combined with the outcomes of T4.2 for potentially adopting (semi-)automatic tools for facilitating the mapping of different standards to IDS.

UDA (Universal Data Access) was originally the data access tool developed for MAST and it has now been adopted for use with IMAS for data access. It is a tool designed to abstract away the problem accessing data in different formats and databases and provide a simple API for users. An instance of UDA is currently installed on the Eurofusion gateway and this can be used to access the data of experiments that have set up UDA. UDA has a plugin concept and there is essentially a different plug in for each experiment.

MAST example

A list of signals which map UDA plugin call to IDS address paths is stored on the MAST UDA server. A user can request individual IDS elements via UDA server and import the entire IDS using IMAS. When requests are made to the UDA server for an IDS element it uses the mapping to find the appropriate UDA plugin. It then calls the appropriate MAST endpoints and returns the data of interest.

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

3.3 Granularity of Metadata and Persistent Identifiers

One issue which is currently under discussion, with a document in preparation, is the appropriate level of granularity for both metadata and persistent identifiers. There are a number of possible implementations largely dependent on the site data models. For instance, for a specific experimental shot, some devices store data in a number of files related to either diagnostic information or physical parameters, each file containing data for the whole shot. In other sites, the same information is presented as time slices to minimise the data transfer just to the area of interest (e.g. the flat top of the plasma current). This diversity is data models makes defining a single granularity for data access difficult. In the worst case, a user could use data from a single diagnostic device with a particular time span to derive results. A simple PID at the shot level would clearly not aid reproducibility and extracting the required information in a machine-oriented manner would be difficult. Applying a unique persistent identifier to each time slice for each file would also be very difficult if using a handle system such as DataCite or EPIC die to the sheer number of PID's that would need to be minted for every single shot.

3.4 Architectural Considerations

While it is clear that the sites will be the ultimate owners and maintainers of the metadata and data, how the PID's are created is an architectural discussion. For instance, one legal entity (e.g. IPP or Euratom) could form an agreement with a provider and then sites could make use of this central service. While this minimises costs since only one prefix is required it is unclear whether this can be done (under investigation) or indeed whether it is desirable since this then represents a single point of failure in the architecture. The alternative would be each site getting its own prefix but this is more costly and requires every site to support an additional service. In addition, with each site having its own data model, each site could choose a different level of granularity to suit this model, which could cause problems for the proposed aggregator. This is currently being investigated in Task 4.2 and will be reported in more depth in a later deliverable.

It is also clear that as well as the standard types of metadata (structural, descriptive and administrative) dealing with a mix of static and dynamic metadata elements will introduce a level of complexity and necessitate policy decisions in some places. For instance, if a site moves from a protective to an open licensing model, this will need to be updated appropriately. Metadata annotations also present a similar issue. Unlike many other aspects related to metadata, the use of annotations, while identified as a clear requirement, has no central implementation and is only used at some sites. While we are currently investigating different technologies such as B2NOTE[13], Annotare[14] and Annotea[15], it is not clear whether any of these can easily store this information within the CatalogQT database, and if this cannot

[13] https://www.eosc-hub.eu/services/B2NOTE
[14] https://code.google.com/archive/p/annotare/
[15] https://www.w3.org/2001/Annotea/

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

be done then an additional database may need to be instantiated. Note Catalog QT is a Eurofusion database specifically developed to allow the storing and querying of data is IDS structures.

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

4 Conclusion

Working with the already existing schema's which have been developed within the EUROfusion consortium, we have proposed a metadata schema which has been proven as acceptable to the experimental sites with only minor additions to the existing schema. In addition, we have integrated elements of common schemas (namely Qualified Dublin Core) which provides additional metadata and have mapped some of the existing fusion schema elements to Dublin Core terms. These have been reviewed and approved by project members representing sites hosting experimental tokamak devicesand we are working on integrating the required additional elements into the Summary IDS. With this we believe we have satisfied the requirements of both the fusion community and helped support the evolution towards a more open data regime when this is deemed acceptable by sites.

It is important to note that this project cannot mandate how these elements are filled

While the schema has been agreed, it is important to recall that there are still some essential discussions ongoing and which will need further clarification in the future. In summary these are:

- provenance metadata beyond version tracing has not yet been incorporated but will be dealt with later in the project,
- the granularity of the data and metadata needs to either be standardised or, within the architecture, needs to be addressed to ensure consistency between the search results obtained from different experiments,
- thus far we have only considered experimental data; while the IDS structure is neutral on this, further analysis should be performed to ensure this metadata meets the needs of the modelling and analysis groups within the community,
- further investigation into relating related metadata standards (e.g. materials database schemas such as those proposed by the RDA),
- in principle all elements of the Summary IDS should be searchable by experienced users and this should be incorporated into the blueprint architecture; however the searchable elements as identified previously should be accessible to any user wishing to learn more about the work of the fusion community.

We also believe with the modifications made we have gone some ways to improving the FAIRness of fusion data. One further consideration is that both the IMAS Data Dictionary and Summary IDS which is derived from it are both currently owned by the ITER organisation. While the schema is openly available to members of the fusion community including funding bodies, it is not a truly open schema. As a project and through EUROfusion we are in discussions with ITER about making this schema open.