Fusion Engineering and Design xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Fusion Engineering and Design

journal homepage: www.elsevier.com/locate/fusengdes

Data catalog project—A browsable, searchable, metadata system

Joshua Stillerman[a,*], Thomas Fredian[a], Martin Greenwald[a], Gabriele Manduchi[b]

[a] MIT Plasma Science and Fusion Center, Cambridge, MA, USA
[b] Consorzio RFX, Euratom-ENEA Association, Corso Stati Uniti 4, Padova 35127, Italy

ARTICLE INFO

Article history:
Received 23 May 2015
Received in revised form 29 April 2016
Accepted 2 May 2016
Available online xxx

Keywords:
Data acquisition
Data management

ABSTRACT

Modern experiments are typically conducted by large, extended groups, where researchers rely on other team members to produce much of the data they use. The experiments record very large numbers of measurements that can be difficult for users to find, access and understand. We are developing a system for users to annotate their data products with structured metadata, providing data consumers with a discoverable, browsable data index. Machine understandable metadata captures the underlying semantics of the recorded data, which can then be consumed by both programs, and interactively by users. Collaborators can use these metadata to select and understand recorded measurements.

The data catalog project is a data dictionary and index which enables users to record general descriptive metadata, use cases and rendering information as well as providing them a transparent data access mechanism (URI). Users describe their diagnostic including references, text descriptions, units, labels, example data instances, author contact information and data access URIs. The list of possible attribute labels is extensible, but limiting the vocabulary of names increases the utility of the system. The data catalog is focused on the data products and complements process-based systems like the Metadata Ontology Provenance project [Greenwald, 2012; Schissel, 2015].

This system can be coupled with MDSplus to provide a simple platform for data driven display and analysis programs. Sites which use MDSplus can describe tree branches, and if desired create 'processed data trees' with homogeneous node structures for measurements. Sites not currently using MDSplus can either use the database to reference local data stores, or construct an MDSplus tree whose leaves reference the local data store.

A data catalog system can provide a useful roadmap of data acquired from experiments or simulations making it easier for researchers to find and access important data and understand the meaning of the data and how it was obtained. This is particularly useful in research facilities that study the results of many different experiments or simulations and may not know the intricacies of the data organization in use where the data was generated. It is also possible to store a local copy of key data items in local MDSplus trees and then add processed data to the local catalog.

1. Introduction

Modern fusion experiments generate very large data sets. These are composed of many heterogeneous elements, ranging from scalars to 100 MB time-series signals and high-speed video sequences. For example, during a recent campaign, the Alcator C-Mod experiment acquired 15 GB/shot representing over 40,000 different measurements. Finding and understanding relevant data from this set can be challenging even for experienced local users, this is problem is compounded by the passage of time, and the changes in personnel. A coherent metadata system provides a framework so that the recorded data can be understood by a wider group of people over a long period of time.

2. Motivations

Diagnostic owners know where to find their results, and how to interpret them. Frequent consumers of data also know how to find and understand them. These users who already know how to document and index the stored quantities will not be the primary consumers of this information since they already know it. This project is a platform for creating and consuming this documentation. The tools will make it easy to produce compelling applications for searching and browsing this information, and then displaying

* Corresponding author.
E-mail addresses: jas@psfc.mit.edu, jas@mit.edu (J. Stillerman).

http://dx.doi.org/10.1016/j.fusengdes.2016.05.004
0920-3796/

or analyzing the measurements found. The hope is that once an experiment has documented an interesting subset of their measurements, users will be motivated to provide further information about their data, and descriptions of more of their data. To facilitate this behavior, emphasis will be placed on making the process of entering this information as simple as possible.

The MDSplus [1] data system allows diagnosticians to record metadata associated with their measurements. These metadata document the recorded data so that it retains its value over time. The metadata fall into three categories, standardized metadata that are associated with all data items (on/off-status, date/time written, data-type, usage, etc...), metatdata that MDSplus provides specialized data-types for (with_units, with_error, signal independent parameters, ...) and user defined metadata stored in associated nodes. However, it does not provide for searching and discoverability nor does it provide any standardization for these user defined metadata. Making these data searchable and browsable further increases the likelihood that future consumers be able to utilize these data.

Many different sets of users will benefit from this system. The owners of a diagnostic can use it to write down all of the information that they thought was obvious and that they would never forget, only to discover that they have forgotten as time passes. Local users other than the owner of the measurements that want to utilize the data as part of their data analysis will be able to look at these descriptions to find and understand the data, off-loading these queries from the original diagnosticians. Visiting scientists have an even greater need for this, since they may not know the personnel associated with any of the diagnostics. Offsite collaborators, in addition to these difficulties, encounter additional problems associated with remote collaboration, ranging from time zones, language, culture, and lack of 'presence' information.

These tools are particularly interesting for scientists working on multiple off–site experiments or comparing data from several machines. In addition to benefits described above, this system coupled with a local data store can provide uniform access to a whole set of experiments, regardless of the data systems that those experiments use.

Annotating measurements with homogeneous metadata will facilitate the development of data driven high-level data display applications that can be shared between multiple experiments. These programs have tended to be written in less general ways, encoding site knowledge and configuration and metadata storage in the application code. Unifying and regularizing these metadata makes generalizing and sharing these tools possible. A single application can browse or search for measurements, and then by inspection know how to render them.

3. Data catalog schema

The quantities described in the data catalog are heterogeneous. They will range from scalars to multi-dimensional arrays. These items and their constituent parts will have both required and optional metadata associated with them. These metadata will be labeled with terms described in ontologies (i.e. with fixed unambiguous vocabularies). This ontology will be populated scientific and computing staff in collaboration. Once the general entries are in place, users can expand the vocabulary as needed. Decisions about how carefully these expansions should be curated are important, and will depend on sites scientific administration.

The main entries in the catalog are measurements. These correspond to the high level products of the diagnostics. For example at Alcator C-Mod the HIREX-SR diagnostic measures ion temperature, toroidal rotation and impurity density, which are all functions of both space and time. Each of these would have general metadata about the diagnostic and measurement and one trace describing a two dimensional array of the data. The Grating Polychrometer (GPC), which is one of the Electron Cyclotron Emission (ECE) diagnostics, records nine temperatures versus time and each of these has a corresponding position versus time. In addition the channel that is looking closest to the axis is designated 'Te(0)', so that less sophisticated users can locate that channel. Again, general information about the diagnostic would be stored with the GPC measurement, while specific metadata would be stored with the individual associated traces.

Traces are also labeled with one or more tags, which provide hints about how applications can render the data. For example the HIREX-SR traces would be marked as '2D Profiles', alerting applications that the data can be rendered as a surface or sliced by time or position. The GPC traces would be labeled as 'Time Series' and 'Temperature', indicating that they are functions of time. If appropriate, a tag denoting time series at variable positions could be created for these. This last case points at the need for a controlled vocabulary for these tags. Once a name for this tag is created, it should be used for all appropriate cases, rather than inventing different names for the same thing. To facilitate this, the system will require that the tags be chosen from an extensible ontology of tag names.

4. Database implementation

The entries in the data catalog are organized around measurements. These are high-level abstractions of quantities being measured by diagnostics. Each measurement is in turn made up of one or more traces, which describe quantities that can be retrieved from the experiment's data store and rendered or analyzed.

Each measurement has a set of required metadata that describe their general properties. These include, Name, Description, Owners, Diagnostic, WWW References, Canonical or Example Shot, . . . These metadata can be used to drive browsing and searching applications. Users can find out what measurements a given diagnostic produces, or search for measurements that have particular strings in their descriptions. They can filter by owner or vintage of the canonical example shot. Once they have located a set of records of interest, they can view their details, including who to contact, or papers to read about the measurements.

Measurements are in turn made up of traces that refer to particular retrievable records in the experiment's data store. These items, which can be any shape or size, are annotated with a similar set of metadata to the measurements with a few critical additions. Traces have a uniform resource identifier (URI), units, labels, geometric information. . .and a set of tags indicating likely uses of the data.

A URI is a character string that specifies where the trace can be retrieved from, and what mechanism to use to retrieve it. In the case of MDSplus the URIs take the form:

mdsplus://server-ip-name/tree/shot?path=mdsplus-tree-path.

A sample URI for the plasma current from Alcator C-Mod:

mdsplus://alcdata.psfc.mit.edu/cmod?path=/MAGNETICS:IP.

There is no shot specified in this case, since the shot the user is interested will be added to this specification before the data is retrieved. Given this string and a shot number, applications have enough information to retrieve the data.

Traces also have a set of usage tags that describe the ways in which applications are likely to use and render the data. For example the frames from a video camera might be tagged as 'video' and those from an infrared camera also tagged with 'temperature'. Traces that are profiles would be tagged as 'profile' and time evolving profiles would have tags denoting that. These tags come from an extensible fixed ontology or vocabulary.

Every item stored in the catalog is tagged with time inserted or modified, and the user-id of the user that inserted or modified it. In

Please cite this article in press as: J. Stillerman, et al., Data catalog project—A browsable, searchable, metadata system, Fusion Eng. Des. (2016), http://dx.doi.org/10.1016/j.fusengdes.2016.05.004