Data Catalog Project - A Browsable, Searchable, Metadata System

Joshua Stillerman, Thomas Fredian, Martin Greenwald, Gabriele Manduchi

Consorzio RFX, Euratom-ENEA Association

June, 2015

Plasma Science and Fusion Center
Massachusetts Institute of Technology
Cambridge MA 02139 USA

This work was supported by the U.S. Department of Energy, Grant No. DE-SC0012470. Reproduction, translation, publication, use and disposal, in whole or in part, by or for the United States government is permitted.
Modern experiments are typically conducted by large, extended, where researchers rely on other team members to produce much of the data they use. The experiments record very large numbers of measurements which 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 used 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.

Keywords: Data Acquisition, Data Management

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 15GB / shot representing over 40,000 different measurements. Finding and understanding relevant data from this set can be challenging even for experienced local users.

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 or analyzing 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 his 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…), metadata 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 groups 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. Off-site 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 homogenize 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.

1. 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 pieces 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).

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 ‘T0’, 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.

3. 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 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 that interest them, 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 fro 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 addition, each change to the catalog is recorded in a history table. This allows users to trace the history of the values stored.

The database and the initial user interfaces will be implemented using django[2], and its database agnostic object relational mapper (ORM). This platform makes it very easy to provide basic Create Update Delete (CRUD) operations, as well as fairly polished searching and browsing through the records. Django-rest-framework[3] will be used to generate a RESTful API[4] for use by user applications in any language, as well as purpose built web applications.

So we have django providing the underlying database abstraction, and initial record input and modification. We have django-rest-framework providing the restful API for user applications in languages such as python, matlab, javascript, etc…

4. Data Store

The ‘traces’ in the data catalog have URIs associated with them. These are character strings, which specify where and how to retrieve the data. A protocol specific library is required, which parses the URI and calls into a data system to read the data. Using these libraries, applications that search and browse the database, can then access the data referred to, and display or analyze it.

Many fusion experiments are using MDSplus to manage their data. Several others are using the MDSplus provided network API (mdsip) to provide remote access to local proprietary data stores. All of these can share the same data store protocol, and URI syntax. This will be the first one implemented.

Access to the full complement of standardized metadata associated with individual traces is needed by data display and analysis applications. Values such as ‘title’, ‘x label’, ‘y label’, ‘units’ and usage tags drive defaults for display. It has not been decided if there will be copies of these in both the data store, and the data catalog, or the data store will access the values stored in the catalog. The case in mind is an application which looking directly in the data store, with out utilizing the catalog, but needs access to the meta data to understand or render the data. The likely solution is to make a function that given a data catalog entry’s name returns a structure containing its associated meta data.

The general URI scheme allows the catalog to describe measurements regardless of the underlying data system used to store the data. This is particularly useful for groups that work collaboratively on multiple experiments. In the best case, each experiment would provide this functionality to their users and collaborators. In the cases where they do not, off-site groups with access to an experiment’s data products could construct this index independently. The URIs in these local catalogs could either refer to remote data access APIs provided by the experiments, or to a local data repository. In turn, a local data repository could refer to the remote data in its terminal nodes, or contain copies of those data extracted from the experiment and stored locally. Regardless, these off-site users could then augment the stored quantities with locally stored data analysis results.

In addition to the use of this document template to prepare your manuscript please refer to the Guide for Authors for FED.

5. Applications

Almost every fusion experiment has developed high level applications that display interesting and useful data. These applications utilize site-specific, measurement-specific knowledge to render data whose complex metadata are often not stored, or not stored completely. These applications tend to be very useful, since they generate the specific displays that local users rely on. This is however, in conflict with their generality and usefulness for other experiments. The more details about the local experiment that these applications rely on and the more tailored to the local requirements, the more immediately useful they will be, but the less applicable to other experiments they will be. This system provides a set of building blocks that can be used to build these applications in a data driven way, and make them applicable to other experiments.

At MIT we have a local application called VIDEO_DISPLAY that displays captured video from the Alcator C-Mod experiment. While it is data driven, populating its menus from metadata stored with the video streams it relies on the specific structure and location of these metadata to operate so would not be applicable to other experiments. Using the catalog and associated data store, it should be relatively easy to rewrite this more generally, making it applicable to other experiments, or to other video data from Alcator C-Mod.

The ReviewPlus application[5] integrates many high level data display functions. While not driven by stored metadata, it is shared with and applicable to a variety of experiments. This would also be relatively easy to generalize given the data catalog. The choices of things to display could be driven by the catalog’s browsing and searching functions, and the rendering of the displays could be driven by the stored metadata values.

jScope[6] is a java data display application developed by the RFX group at IGI in Padova. It provides general displays of data stored in MDSplus as well as other data stores. Users specify the data to display by referring to stored data, or expressions of stored data. This is very general, but, since it can plot any of 40,000 signals, it would benefit from the catalog’s browsing and searching functions. Particularly data consumers that are not intimately familiar with the data will feel these benefits.

The Metadata Provenance Ontology project (MPO[7,8]) has a lot of synergy with the data catalog. It is a set of tools to document the provenance of data.
products. One of its central tenets is that its data objects are references to a persistent data store and that they are described with URIs that document how to retrieve them. It may be possible for MPO to use data catalog entries for some of its data objects. This would obviate the need for MPO workflows to create new data objects for each piece of input data they use.

6. Conclusions

The data catalog project when completed will allow diagnosticians to document their data products. This will be very useful to a variety of data consumers including the diagnosticians themselves, other onsite users, new users, visitors and off-site collaborators. The barrier to entry will be low, and payoff significant. The catalog will also facilitate the development of general data driven applications that will have utility across multiple experiments. In addition, users who do research on multiple experiments will be able to use the catalogs to lower the learning curves for data access at each site, and potentially homogenize the data APIs across sites.

Acknowledgments

This work is being performed under US Department of Energy award number DE-SC0012470. The authors would also like to thank several sources, which provided many of the underlying ideas. These are: Matt Reinke (York University), The ITER commissioned data archive format assessments by Tessella, and Hideya Nakanishi (LHD).

References