A general purpose tool-set for representing data relationships: converting data into Knowledge

Joshua Stillerman, Thomas Fredian, Martin Greenwald, John Wright
MIT Plasma Science and Fusion Center
Cambridge, MA, USA
jas@psfc.mit.edu

Abstract—**Rich metadata is required to find and understand the recorded measurements from modern experiments with their immense and complex data stores. Systems to store and manage these metadata have improved over time, but in most cases are ad-hoc collections of data relationships, often represented in domain or site specific application code. We are developing a general set of tools to store, manage, and retrieve data-relationship metadata. These tools will be agnostic to the underlying data storage mechanisms, and to the data stored in them, making the system applicable across a wide range of science domains.**

Data management tools typically represent at least one relationship paradigm through implicit or explicit metadata. The addition of these metadata allows the data to be searched and understood by larger groups of users over longer periods of time. Using these systems, researchers are less dependent on one on one communication with the scientists involved in running the experiments, nor to rely on their ability to remember the details of their data. In the magnetic fusion research community, the MDSplus system is widely used to record raw and processed data from experiments. Users create a hierarchical relationship tree for each instance of their experiment, allowing them to record the meanings of what is recorded. Most users of this system, add to this a set of ad-hoc tools to help users locate specific experiment runs, which they can then access via this hierarchical organization. However, the MDSplus tree is only one possible organization of the records, and these additional applications that relate the experiment 'shots' into run days, experimental proposals, logbook entries, run summaries, analysis work flow, publications, etc. have up until now, been implemented on an experiment by experiment basis.

The Metadata Provenance Ontology project, MPO, is a system built to record data provenance information about computed results. It allows users to record the inputs and outputs from each step of their computational workflows, in particular, what raw and processed data were used as inputs, what codes were run and what results were produced. The resulting collections of provenance graphs can be annotated, grouped, searched, filtered and browsed. This provides a powerful tool to record, understand, and locate computed results. However, this can be understood as one more specific data relationship, which can be construed as an instance of something more general.

Building on concepts developed in these projects, we are developing a general system that could be used to represent all of these kinds of data relationships as mathematical graphs. Just as MDSplus and MPO were generalizations of data management needs for a collection of users, this new system will generalize the storage, location, and retrieval of the relationships between data. The system will store data relationships as data, not encoded in a set of application specific programs or ad hoc data structures. Stored data, would be referred to by URIs allowing the system to be agnostic to the underlying data representations. Users can then traverse these graphs. The system will allow users to construct a collection of graphs describing ANY OR ALL OF the relationships between data items, locate interesting data, see what other graphs these data are members of and navigate into and through them.

Keywords—metadata, provenance, data relationships

I. Background

A. Evolution of data handling systems.

Over time the tools for data management are becoming more general at higher and higher levels of abstraction. Handwritten observations were supplanted by automated recording devices. These eventually gave way to computer based data acquisition programs, purpose built to record specific measurements or diagnostics. In the 1980s general purpose data collections systems abstracted data acquisition and data storage. This was followed by systems such as MDSplus[1] which added higher level abstractions and significant metadata to create a system where data could retain their meaning over long period of time. Information that previously had to be gathered by speaking with the original owner of the data, or by reading their lab notebooks was now tightly bound to the recorded data. Electronic lab notebooks[2][3], document management systems, and experiment planning systems provide searchable metadata that can be used for exploration and discovery. These representations of the connections between data give the results meaning and context. The tools for storing and exploiting these metadata tend to be purpose built, analogous to the purpose build data acquisition programs of the past. Over time we have generalized successively higher level functions. The next logical step is a generalization of the metadata system, supporting data discovery not only through search, but enabling more powerful, multi-faceted browsing.

II. Introduction

A. Data is easy, usable knowledge is hard.

Data collection has evolved from hand written observations, to automatically recorded instrument readings, and now includes a wide variety of high resolution, high speed recording devices. It is relatively easy to acquire gigabyte or even petabyte data sets. At the same time the cost of storage for these data has been steadily decreasing. This makes it feasible

This work supported by DOE Contracts DE-SC0008736 and DE-SC0012470

978-1-4673-9051-4/16/$31.00 ©2016 IEEE

to record extremely large data sets. Increases in available computing power along with decreases in hardware and software costs have largely kept pace, so that we can not only store but also retrieve and process these data.

For example the Alcator C-Mod experiment, which has been operating for 25 years, began collecting about 1 megabyte of data per shot, totaling 30 megabytes per day, or 1.5 gigabytes per year. At that time this was achievable using one or two mini-computers and one rack of magnetic disks. Twenty-five years later the experiment acquires 15 gigabytes per shot — totaling 45 Terabytes per year, and this can still be done by one current generation server, and one drawer of disks.

Over this period the data increase was due to both the size of the average data item, and in the number of items. The current 15 gigabyte Alcator C-Mod pulses have over 240,000 records, recording the measurements from approximately 100 diagnostics. The ease with which these data can be collected makes it imperative that the scientists involved put significant effort into data organization.

Ease of collection and storage can lull users into thinking they don't need to work at organization. Scientists used to carefully file or paste into lab notebooks their hand drawn graphs, and later polaroid pictures of oscilloscope or computer storage displays. These could reasonably be found, understood, and used for further analysis, assuming that chronology was a sufficient organizing principle, simply by finding the notebooks in question or speaking with the people responsible for the data.

The growth in size and complexity of scientific data sets exacerbates the problems associated with data discovery and understanding. It is no longer possible to view even a small fraction of the total data set. Larger projects are products of collaborations with large often geographically distributed staffs and students who may not be involved over the whole life of the experiment. These issues are domain independent. Researchers in physics, earth science, life sciences, genetics, social sciences all share these concerns and recent community workshops have highlighted these challenges [4][5].

III. Approach

A. Generalize the representations of data relationships

A general metadata scheme needs to be able to refer to heterogeneous stored data. This is the case even within a single project, as there are almost always a variety of kinds of data that the system must represent. For the system to be applicable across varied projects in different research domains, this is even more imperative, since they are unlikely to share underlying data storage mechanisms. Uniform Resource Identifiers (URIs) provide a mechanism to homogenize heterogeneous resources. A URI is a string of characters used to identify a resource. Such identification enables interaction with representations of the resource over a network, typically the World Wide Web, using specific protocols [6].

The system needs to be cognizant of the complexity and granularity of the data objects it refers to and the needs of usable, user-facing interfaces. While the URIs must be very specific in order for the graphs that contain them to resolve to a

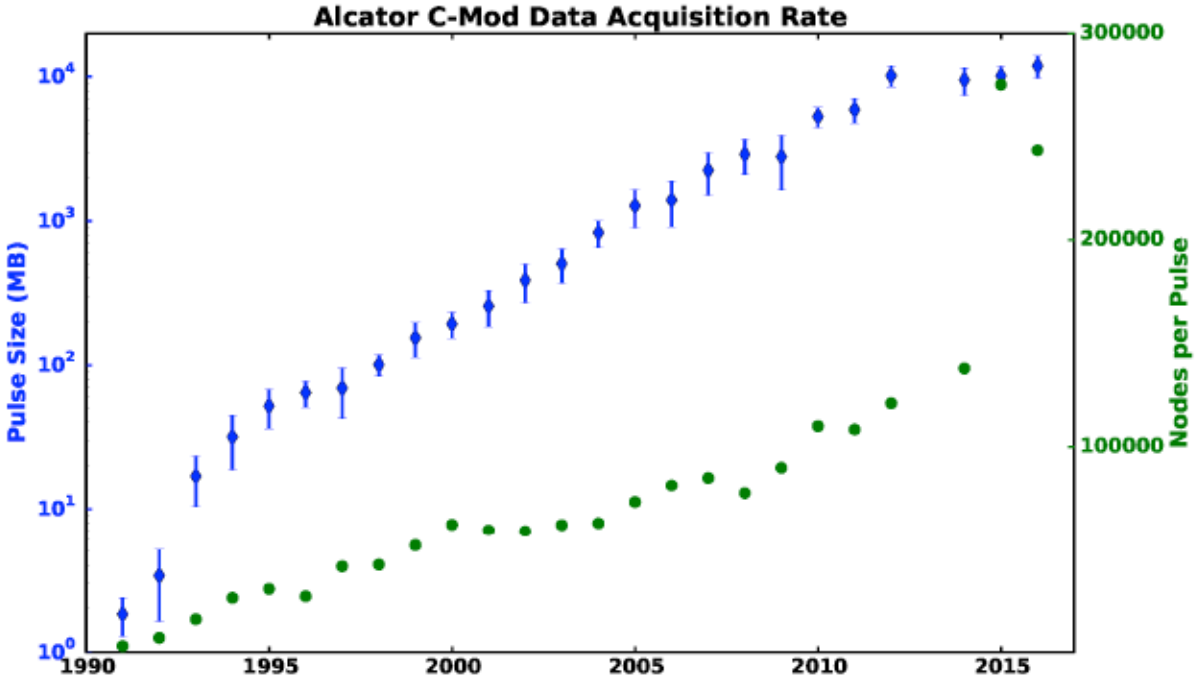

Figure 1—Yearly average data size and number of items for each pulse of the Alcator C-Mod Tokamak (~2500 pulses per year)

particular data object, it is often desirable to have an easier to understand shorthand to display in some applications. For example, given an analysis that refers to a specific viewing cord or profile from a diagnostic, this detailed information is required to know the provenance of the quantities computed. At the same time, it may be advantageous to display this by naming the diagnostic, or experiment pulse, or experiment as a whole. To be useful the URIs must refer to specific immutable things, this in turn requires that storage mechanisms support versioning.

B. *Relationships are graphs*

Relationships between stored data are represented as graphs. That is, collections of vertices (nodes) and edges (connections). There are relatively few classes of graphs that the system will need to represent. Collections of these metaschemas form the schemas for the relationship data for each application domain. Tools can be developed to operate on each graph type. These tools can then be applied to all of the graphs of that type in the domain specific schema. At this time we plan to support hierarchical trees, ordered and unordered sets, timelines, directed acyclic graphs, and threaded conversations. This list will expand as domain specific needs arise.

C. *Schema as data*

The schemas, that is the collection of graph types, for a given domain will be represented as uniformly accessible data, rather than in application code or disjoint data structures. This allows administrators, or even users, to easily add new relationships which can then be exploited by the user community. We will provide application programming

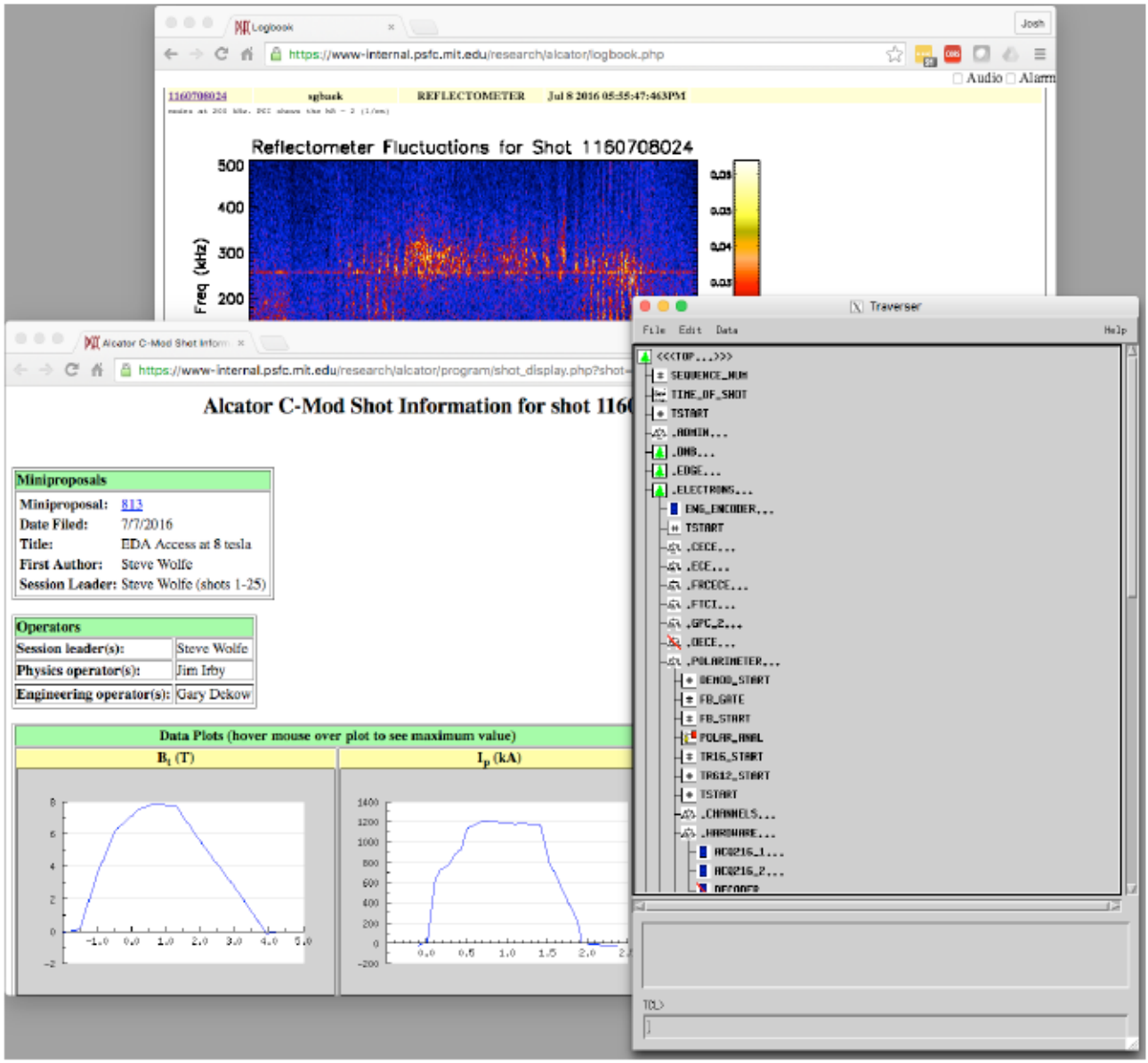

Figure 2 —MDSplus tree hierarchy, logbook and shot display, Alcator C-Mod tokamak.

interfaces to construct and explore the defined relationships. We also plan to implement a web based graphical user interface to interact with these schemas.

D. Instances as data

The relationship data can be divided into three broad categories based on their rate of evolution, though they may share representation and implementation. Each use of the system will have collections of static descriptive metadata that describe the project as a whole. The nodes of these graphs may contain URIs to retrieve detailed records about the project. For example, the mechanical drawings of the experiment, the material properties of its components or the major parameters of its design.

The next class of metadata are applied to slowly changing quantities. Again, they may or may not contain URIs for retrievable objects. For example, in MDSplus the model is the basis or schema for new tree instances that hold the data from subsequent runs of the experiment. This structure changes slowly, and its coherence allows familiar users to navigate the trees which contain the data for each experiment pulse.

The final class of metadata describe the stored data records from the experiment. The nodes of these graphs contain URIs for retrievable data instances. This will, by far, contain the largest number of records. Examples of these include the MDSplus tree structure, logbook entries about specific shots or stored quantities and so forth.

As with the schemas, we will provide application programming interfaces and web based graphical user interfaces to populate and explore the descriptive metadata of the experiment. After navigating to any data reference, users can ask: what other graphs is this object a member of? That is, what other data is related to this item? What are its neighbors in these graphs. This provides a rich environment to document and exploit stored results.

IV. Examples

A. Library

A library provides a simple analogy which illustrates the main features of the system. A library holds a collection of resources, books, journals, databases, etc. These items have a primary organization — the order of the books on the shelf or the order of the cards in the card catalog. This was generalized when the catalogs were computerized, so that customers could effectively re-sort and search the catalog entries by any of the metadata that the library has for each item. However, there is something *special* about the order of the books on the shelf — it is the only relationship that can be browsed physically. Using the catalog users can track down a specific resource, then go to the stacks, find it and browse through the adjacent items. What if they could easily browse items with different adjacency criteria? What if these criteria were extensible and annotatable? What if the new criteria and annotations were accessible to all of the library's customers?

B. Magnetic Fusion Research

Much of the world wide magnetic fusion energy research community uses the MDSplus system to store and organize their acquired and computed results. The data have a primary organization into shots or pulses, each of which has a hierarchically organized record store, with a URI-like access mechanism (though with rather different syntax). The hierarchy, or tree, allows users to create associations between stored data items. For example, a temperature diagnostic might have a tree structure with textual descriptions and final computed results at the top, and the details of the acquired data, and the computation at lower levels of the tree.

Users familiar with the overall layout of the tree can browse and search to find measurements they are interested in. Once found, the associated nodes near them, provide the context that gives them meaning.

Additional site-local data structures and programs store additional metadata that can be used to locate shots and records of interest. Textual logbook entries describe shots and their contained measurements. Numerical tables summarize and index the shots. Drawings and specifications are stored in document management systems. Experiment planning and scheduling documents are indexed, and related to the run days assigned to them. Figure 2 shows the heirarchy and sof the the loosely coupled tools for the Alcator C-Mod experiment.

The navigation between these domains tends to be unidirectional and only between specific classes of records. Applications need to be programmed using specific knowledge of each underlying data and metadata structure. This new system will allow the data to be members of as many relationship graphs as needed. It will enable very general applications which allow users to navigate within and between these graphs at will.

C. Metadata Provenance Ontology Project

The metadata provenance ontology (MPO)[7] system constructs provenance graphs of computational workflows. These directed acyclic graphs can be used to locate and understand the origins of computed results, showing which acquired and computed quantities were used in the computational pipeline. The graphs can be grouped into collections, annotated with comments and augmented with structured metadata. Note that some of the nodes of these provenance graphs refer to data stored in MDSplus. Navigating these graphs provides context needed to understand the meaning and provenance of stored results. However, the system is limited to the prescribed data relationships. Data associations can be seen from the MPO tools (i.e. what raw data was used for a computation) cannot be seen symmetrically from the data storage tools referred to. This is another asymmetric, purpose built relationship graph, that can be generalized providing additional functionality to the users. Figure 3 shows web based data provenanace displays and navigation.

D. Other Research Fields

Since the same issues come up in life sciences, earth sciences, and even social sciences, similar examples can be found in other areas of research. The hyperstudio group at MIT has several projects involving exploring, sharing, and annotating data from the humanities, an application area seemingly as far from magnetic fusion as possible and providing a measure of the generality of the problem and solution. From their web site[8]:

> *HyperStudio explores the potential of new media technologies for the enhancement of education and research in the humanities. Our work focuses on questions about the integration of technology into humanities curricula within the broader context of scholarly inquiry and pedagogical practice.*
>
> *Our central goal is to provide individual digital humanities project participants flexibility in modeling, analyzing, and presenting their materials as they choose — while also allowing researchers to combine features from other projects in innovative ways.*

This mission and goal align well with the work we are proposing and their *annotation studio*[9] has many of the features we are describing. The data from Comédie-Française Registers Project[10] is an example of a large set of data that can be explored interactively by looking at the connections and relationships between the stored values. In this project indexes the box office receipts and other records from over 100 years of theater performances in Paris. This data set, could be represented in this new system, this would allow users to extend the relationships between items, and with historical events.

V. Conclusion

The relationships between pieces of data are what gives them meaning. Preserving and documenting these relationships allows these data to be used by wider groups of users over longer periods of time. Over time data collection, organization, and retrieval has undergone successive generalization. As this has taken place it has become easier for users to exploit their data. We are creating a system to represent relationships between stored data in a general, extensible, data driven

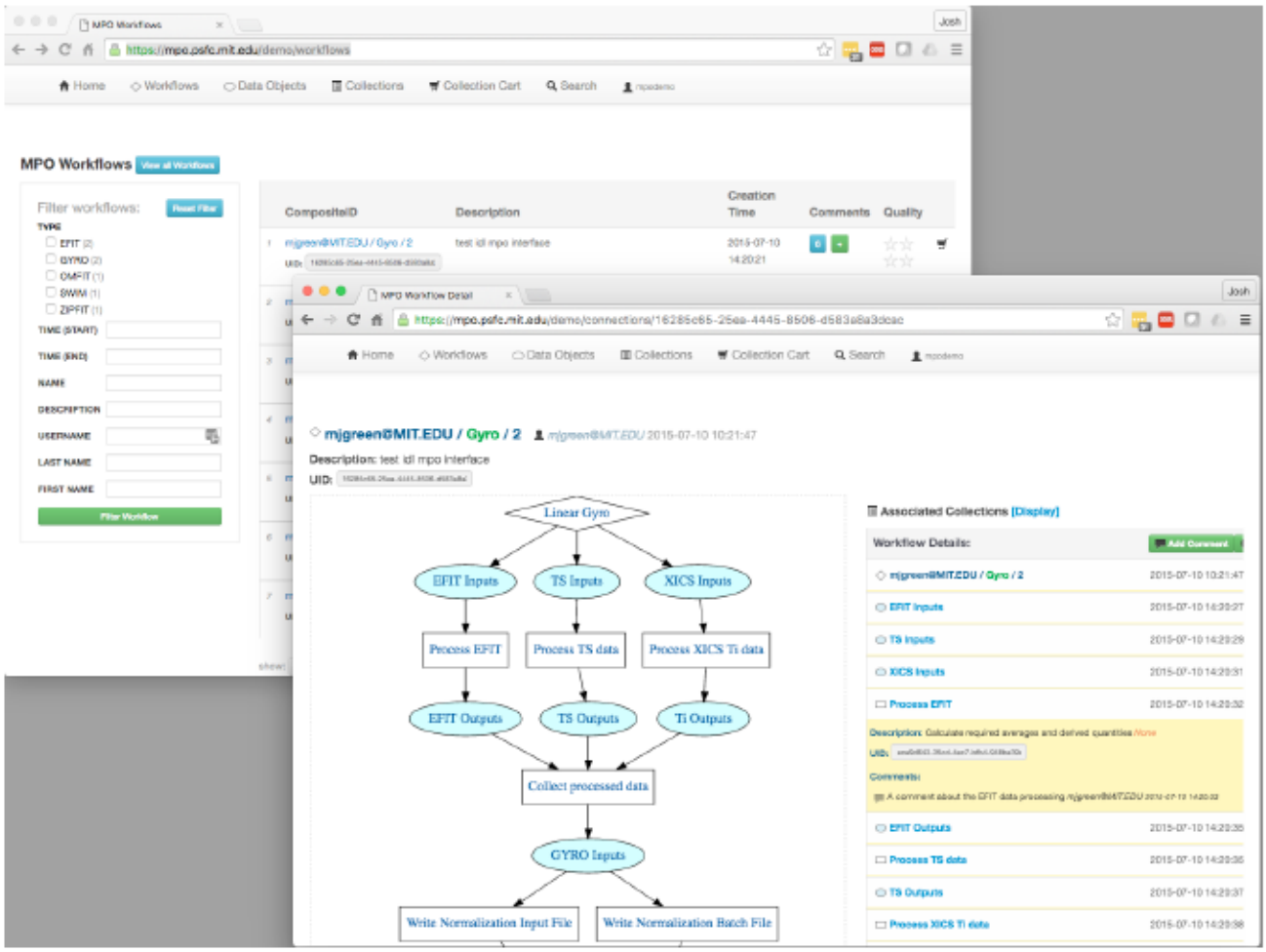

Figure 3 —Metadata Ontology Provenance (MPO) displays. http://mpo.psfc.mit.edu/ .

manner. This system will facilitate exploration and discovery activities on store data from research endeavors.

Data will be represented by URI, regardless of their source or storage mechanism, this enables the system to operate on a wide varied of stored data, and across research domains. The URIs will be very specific as to the data they refer to, but will can be referred to or displayed by the identifiers of enclosing or related items.

The system will make data findable through searchable, browsable, filterable metadata and comments. Since provenance is one of the data relationships we will support, it will facilitate the verification of computed results, helping to preserve the traceability and integrity of research products.

Using these tools data will be able to retain its meaning over time and through personnel changes. This problem is general, and its solution will be applicable across a wide variety of research domains.

ACKNOWLEDGMENT

The authors would like to thank the Alcator C-Mod scientific staff, the MDSplus developers and users, the MPO project team.

REFERENCES

[1] MDSplus data acquisition system, Stillerman, J. A. and Fredian, T. W. and Klare, K.A. and Manduchi, G., Re- view of Scientific Instruments, 68, 939-942 (1997).

[2] Butler, Declan. "Electronic notebooks: A new leaf." Nature 436.7047 (2005): 20-21.

[3] Fredian, T. W., and J. A. Stillerman. "Web based electronic logbook and experiment run database viewer for Alcator C-Mod." Fusion engineering and design 81.15 (2006): 1963-1967.

[4] Report of the Workshop on Integrated Simulations for Magnetic Fusion Energy Sciences, June 24, 2015, U.S. Department of Energy-Office of Science, http://science.energy.gov/~/media/fes/pdf/workshop-reports/2016/ISFusionWorkshopReport_11-12-2015.pdf, accessed June 23, 2016.

[5] Report of the workshop on Management, Analysis and Visualization of Experimental and Observational Data, Sept 29-Oct 1, 2015, U.S. Department of Energy-Office of Science,. http://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/ascr-eod-workshop-2015-report_160524.pdf

[6] https://en.wikipedia.org/wiki/Uniform_Resource_Identifier, accessed June 23, 2016.

[7] M. Greenwald, T. Fredian, D. Schissel, J. Stillerman, A metadata catalog for organization and systemization of fusion simulation data, Fusion Engineering and Design, Volume 87, Issue 12, December 2012, Pages 2205- 2208, ISSN 0920-3796, http://dx.doi.org/10.1016/j.fusengdes.2012.02.128.

[8] http://hyperstudio.mit.edu/ Identifier, accessed June 23, 2016.

[9] http://hyperstudio.mit.edu/projects/annotation-studio/ Identifier, accessed June 23, 2016.

[10] http://hyperstudio.mit.edu/projects/comedie-francaise-registers-project/ Identifier, accessed June 23, 2016.