Data Provence in the CMCS

Carmen Pancerella, Jim Myers, Larry Rahn
carmen@ca.sandia.gov, jim.myers@pnl.gov, rahn@sandia.gov

CMCS Overview

The Collaboratory for the Multi-scale Chemical Sciences (CMCS) is a research project involving chemical science researchers and computer scientists from multiple DOE laboratories, other government laboratories and academic institutions in the development of an informatics-based approach to synthesizing multi-scale information to create knowledge in the chemical sciences.

The CMCS will use advanced collaboration and metadata-based data management technologies to develop a Multi-scale Chemical Sciences portal providing community communications mechanisms and data search and annotation capabilities. The portal will also provide capabilities for defining and browsing cross-scale dependencies, for example, those associated with data produced at one scale that is used as input for computations at the next. Notification mechanisms will make both researchers and their applications aware of updated values of relevant information such as reaction rates. The CMCS and its portal will provide mechanisms to enhance the coordination of research efforts across related sub-disciplines in the chemical sciences, focusing research at one scale on obtaining or refining values critical in the next, reducing work performed using limited or outdated values, and enhancing the ability of the community to meet the national research challenges of the DOE. We believe that the approaches and technology that CMCS is piloting will not only increase collaboration and coordination across disciplines and chemistry scales, but will also enable new, possibly revolutionary, approaches in chemical science.

The current state of the art in chemical science collaboration is through literature references. However, there is additional information about data that is not always contained in these references; key examples are sensitivity to error and relationships to other data. We intend to capture this as part of CMCS pedigree.

The CMCS community does not have metadata standards, and it is important for each scientist to establish his/her own metadata properties. This particular community will not be standardizing on metadata anytime in the near future. In the CMCS, pedigree and annotations will enhance the potential for collaboration and at the same time the scientists will not have to agree on community metadata standards.

Data Provenance and Pedigree

Data provenance is at the heart of the CMCS project and is a key enabling technology for our data-centric collaboratory. It enables researchers to categorize and trace the scientific data across disciplines and scales and to identify the ultimate origin of scientific data. We use the term *data pedigree* to describe the metadata which uniquely defines data and provides a traceable path to its origin.

A normal source of scientific data is scientific journal articles. The scientific article is a part of the data provenance, and it also further describes the data's pedigree by indicating, for example, the methods by which that data was obtained. In the CMCS, the pedigree relationships can also be output and input data from related scientific processes that are stored in our data repository, as well as self-describing links. Traditionally, a researcher has kept a detailed, *private* record of pedigree in notebooks, parameter files, etc. and then publishes results. The publication becomes the source of the data's *public* pedigree. CMCS blurs this distinction, allowing more of the detailed, pre-publication pedigree to be made public to CMCS users. Furthermore, we are capturing the data provenance, the information about electronic data (when created, by whom, where did this come from, etc.).

The CMCS has two types of data: *published data*, available to all CMCS users, and *private data*, available to an individual or collaborating group. While pedigree can and should be attached to any data, the CMCS only requires pedigree information is attached to all published data. However, as data sets and results are referenced without a formal publication, the CMCS will seek approaches that provide access to data and results prior to publication.

In addition to pedigree, CMCS plans to provide data review and annotation. Chemical science data is often reviewed by external sources and further analyzed after publication and made available to researchers through web accessible databases. CMCS plans to capture this annotation information so that it is available to researchers. All of this information will be included in the CMCS core pedigree and annotations.

Pedigree Implementation in CMCS

Pedigree information can be contained within data files or documents, and/or it can be included in *metadata* properties associated with the data file (in our case, metatdata properties are stored as Web-Based Distributed Authoring and Versioning (WebDAV or DAV) [1] properties). When a data file is published on the CMCS server, we can extract much of the pedigree within an XML data file and place it into DAV properties using SAM (Semantic Annotated Middleware) [2]. The XML file does not have to be in a standard format. Instead, an XSLT translation can be provided which extracts the corresponding pedigree information from the XML data file.

In DAV, a URL corresponds to a document and a set of properties. One can think about the set of properties plus the document as being the 'content' of the URL in the way that a file is the content of a filename. What is different is that properties can be retrieved and set independent of the document. That separation does provide some efficiency; for example, an analysis program pulls down the data file, while a pedigree browser can retrieve literature references and other properties without actually retrieving the data itself. Perhaps more importantly, the split between document and properties can guide the idea of how coupled information is. We can now separate the pedigree information (literature references, specific comments about the data, etc.) from the data itself, and so new pedigree information and annotations (forward links, reviewer comments, data descriptions, sensitivity analysis, etc.) can be added by either the owner or others without "touching" the original data set. Also, modeling and analysis programs can use the data (likely in some standard format) without having to sift through the pedigree information. The caveat is that once a data file is downloaded, its pedigree may be lost unless the DAV properties are also downloaded.

Our DAV-aware pedigree browser can easily find the pedigree data (as well as annotations) and allow users to search, browse, and retrieve a data set's pedigree. When data is used or referenced by other scientists, the associated metadata can be retrieved as well.

Pedigree Properties

Examples of CMCS pedigree include creator, contributors, keywords (used for searching in the future), dates (modified, created, etc.), input files (distinguishing between parameter files, input data, etc.), program or technique used to create data (versions, platforms, specific comments, etc), relationship to other data (this data set "replaces" another, this data "is part of" a larger set, etc.) and literature references.

In this first phase of the CMCS, we are primarily using XML as a way of describing some of the pedigree properties. During the next phase, we will move to RDF as a way of describing more complex pedigree relationships (i.e., "subject verb object" relationships).

The CMCS has adopted the Dublin Core [2] as a way of representing some of the pedigree information in the CMCS. Our reason for adopting this is to conform to a "standard" that digital libraries are adopting. Furthermore, the CMCS is publishing chemistry data electronically, and Dublin Core is designed to capture publication. The name "Core" indicates an assumption that Dublin Core will coexist with other metadata sets. In the CMCS, we have also defined a schema that allows us to define a project space, reviewer comments, and other chemistry-specific metadata. Furthermore, individuals or groups can extend the CMCS pedigree set and define additional pedigree information in order to meet specific project needs.

Summary

We are interested in capturing pedigree information across scales without having to define and enforce standards in the chemical science community.

References

[1] Web-based Distributed Authoring and Versioning (Web DAV), http://www.webdav.org
[2] Dublin Core Metadata Initiative, http://www.dublincore.org/
[3] Scientific Annotation Middleware project, http://www.scidac.org/SAM/