Metadata in the Collaboratory for Multi-Scale Chemical Science

Carmen Pancerella,[1] James D. Myers,[2] Thomas C. Allison,[6] Kaizar Amin,[3] Sandra Bittner,[3] Brett Didier,[2] Michael Frenklach[8], William H. Green, Jr., [6] Yen-Ling Ho, [5] John Hewson,[1] Wendy Koegler,[1] Carina Lansing,[3] David Leahy,[1] Michael Lee,[1] Renata McCoy,[2] Michael Minkoff,[3] Sandeep Nijsure,[3] Gregor von Laszewski,[3] David Montoya,[5] Reinhardt Pinzon,[3] William Pitz,[4] Larry Rahn,[1] Branko Ruscic,[3] Karen Schuchardt,[2] Eric Stephan,[2] Al Wagner,[3] Baoshan Wang,[3] Theresa Windus,[2] Lili Xu, [5] Christine Yang[1]

[1]Sandia National Laboratories, Livermore, CA 94551-0969
[2]Pacific Northwest National Laboratory, Richland, WA 99352
[3]Argonne National Laboratory, Argonne, IL 60439-4844
[4]Lawrence Livermore National Laboratory, Livermore, CA 94551
[5]Los Alamos National Laboratory, Los Alamos, NM 87545
[6]NIST, Gaithersburg, MD 20899-8381
[7]MIT, Cambridge, MA 02139
[8]University of California, Berkeley, CA 94720-1740

carmen@sandia.gov

Abstract

The goal of the Collaboratory for the Multi-scale Chemical Sciences (CMCS) [1] is to develop an informatics-based approach to synthesizing multi-scale chemistry information to create knowledge in the chemical sciences. CMCS is using a portal and metadata-aware content store as a base for building a system to support inter-domain knowledge exchange in chemical science. Key aspects of the system include configurable metadata extraction and translation, a core schema for scientific pedigree, and a suite of tools for managing data and metadata and visualizing pedigree relationships between data entries. CMCS metadata is represented using Dublin Core with metadata extensions that are useful to both the chemical science community and the science community in general. CMCS is working with several chemistry groups who are using the system to collaboratively assemble and analyze existing data to derive new chemical knowledge. In this paper we discuss the project's metadata-related requirements, the relevant software infrastructure, core metadata schema, and tools that use the metadata to enhance science.

Keywords: chemistry, metadata, knowledge management, collaboratory, Dublin Core, WebDAV.

1. Introduction

As seen in Figure 1, chemical science research addresses complex multi-scale phenomena. Different physical phenomena dominate system dynamics at these different scales, leading to a variety of conceptual and computational models and experiments relevant in the different regimes. Information from one regime is used as input for the next, essentially "bootstrapping" from the atomistic to the device level. One of the major bottlenecks in such a multi-scale research enterprise is the passing of information from one level to the next in a consistent and validated manner.

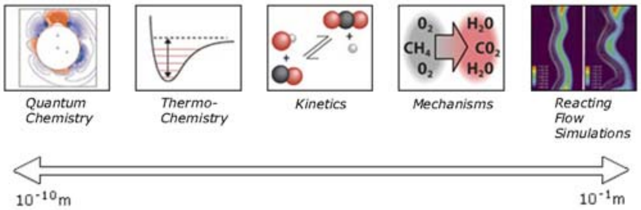


Figure 1. Multi-scale chemical science.

The scientific process described above leads to a data- and model-centric view of the communications between subdisciplines working at different scales. Data at one level is analyzed to develop a model that produces data used in turn by another, repeatedly across the range of scales and types of chemical information required. However, in this process more than just the raw data values need to be communicated. Confidence in a value's accuracy, its uncertainty, dependencies on other data, etc. must all be considered when using it in further computational and experimental research. In the direction of decreasing length and time scales, information about the sensitivity of models on particular data may place a premium on very accurate values for certain fundamental quantities. Enabling the rich bi-directional exchange of both data and metadata between scales is a critical issue in making progress.

Data provenance [2] or data pedigree -- where a piece of data came from and the process by which it arrived in the data repository – is important to the accuracy and currency of scientific data. It enables researchers to categorize and trace the scientific data across disciplines and scales and to identify the ultimate origin of scientific data. Data pedigree is the metadata that uniquely defines data and provides a traceable path to its origin. In the CMCS, we use the term scientific pedigree to capture identification of the data, the traceability of the data, as well as some of the accuracy and sensitivity values discussed previously.

Traditionally, information flow of chemical science data has been accomplished through the research literature and, more recently, through databases of chemical values. Discovery of new information in these sources is a manual process. Further, the information is fragmented. Determining whether results presented in a paper depend on obsolete values from a different regime may require searching through several papers and databases. These factors make communication difficult and time consuming and increase the likelihood of redundant and irrelevant research.

Current manual approaches to coordinating multi-scale research cannot themselves scale to the amounts of data that will be generated in future generations of chemical science research and to the level of effectiveness and efficiency required to tackle national science issues in a cost effective manner. The multi-scale communications challenges facing chemical science researchers are not discipline specific. Thus, a solution to these issues in the chemical sciences will provide a model for multi-scale science that can guide efforts in other domains.

To overcome current barriers to collaboration and knowledge transfer among researchers working at different scales, the Collaboratory for Multi-Scale Chemical Science (CMCS) [1] is developing the following information technology infrastructure for the chemical science community:

- A collaboration infrastructure to enable real-time and asynchronous collaborative development of standards for data and metadata description, inter-scale scientific communication, geographically distributed disciplinary collaboration, and project management.
- Modification of existing tools, which generate and analyze data at each scale, to enable the generation and storage of the required metadata.
- Repositories to store chemical sciences data and metadata in a way that preserves data integrity and allows web access.
- New tools to search and query metadata, and to retrieve data across all scales, disciplines, and locations.

Metadata is at the heart of this data-centric infrastructure, enabling the discovery of data across scales and preserving the data provenance or pedigree. In this paper, we discuss how Dublin Core [3] is being used in CMCS, describe our current metadata definitions for chemistry and scientific pedigree, describe our CMCS metadata infrastructure, which is built on top of DAV [4] and uses the Scientific Annotation Middleware (SAM) [5], and show how this metadata infrastructure enables data pedigree browsing, searching, and other useful science.

The main interface to CMCS is through a Multi-Scale Chemistry (MCS) portal, which provides a wide range of knowledge management, collaboration, and research productivity tools to research groups and chemistry communities. Within the portal, tools are available for exploring data collections, searching for chemical information about particular species, subscribing to receive email when new data appears, viewing the inputs used to create a data set, visualizing data, collaborating with research groups and communities, and more. Underlying the portal, CMCS has developed an advanced data repository that automates many aspects of data discovery, translation, and pedigree tracking. The overall scope of the CMCS project is described at [1]. In this paper, we ignore details about the portal architecture, and instead discuss the metadata-aware content store and the related end-user capabilities available in the portal. In section 4, we discuss metadata applications that are available through the portal.

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 and science in general.

2. Chemical Science Metadata

Metadata is commonly defined as information ‘about’ data values and data sets. However, such a definition is very dependent on one’s perspective. For example, whether a chemical formula is metadata about a molecular geometry, or whether geometry and other information such as the heat of formation are metadata about a chemical formula is a matter of perspective. Such differences of opinion, once encoded in software, are an endless source of barriers to cross-scale collaboration. Within the CMCS, we equate the term metadata with data values that have meaning across domains. In the example above, the chemical formula is metadata because it has meaning for both quantum- and thermo- chemists. Heat of formation would not be considered metadata until we expand scope and realize that both thermochemists and kineticists ascribe meaning to it.

By our definition, data is opaque and meaning-free outside a sub-discipline and, as a corollary, efforts to standardize formats and meanings between collaborators to support inter-scale search capabilities, application interoperability, etc. can be confined to metadata. Further, the system architecture can treat data as opaque as well and no restrictions need be placed on its format. In contrast, because metadata must be understood and manipulated, it must be formatted in a way that exposes its meaning in machine-comprehensible form. An important consequence of this bifurcation is that it minimizes the effort required to allow two parties to collaborate – no changes are required to any applications, and no agreements need be reached about the meaning of terms, except those directly concerning the values that will be exchanged.

Metadata is used in the CMCS in the following ways:

- To provide identification and documentation to scientific data.
- To document the context and value of the data. {For example, the theoretical atomization energy of methylhydroperoxide (and its uncertainty) computed in Ecce (Extensible Computational Chemistry Environment) [6], a problem-solving environment for computational chemistry at the molecular scale, contains information identifying

the species and the quantity, units, the theoretical method used, vibrational frequencies and geometry, reference to source file, creator, etc.}

- To facilitate cross-scale transfer of data, by showing a chain of inputs, data, and outputs across scales or by relating data to its literature references.
- To allow users to comment on the data and its quality, for example, for scientific peer review of data.
- To make collaboration of domain scientists more effective.

While individual chemistry domains have implicit and sometimes explicit schema, there is no schema that spans the scales that address CMCS requirements. Our approach is to acknowledge that such definitions have not to date been practical and to define minimal schema and provide mechanisms to map to/from them rather than attempt to force standardization ahead of being able to deliver recognized value. Enforcing metadata standards across multiple chemistry communities would not be pragmatic and would alienate scientists. Instead, we are providing guidance to users capturing metadata to simplify mapping within CMCS. We believe this model, which will allow CMCS to deliver significant capabilities related to data discovery and analysis and to group collaboration without the up-front costs of standardization and tight integration, will be key in promoting adoption. When and if metadata standards across multiple scales emerge, we can easily include these standards and map our schema to them as necessary.

3. CMCS Metadata Infrastructure and Definitions

The CMCS data repository is shown in Figure 2. This diagram shows how data sets can be annotated with metadata. We describe our implementation details as well as metadata definitions next. The CMCS metadata infrastructure, as well as the CMCS portal and SAM, are domain independent and can be and are being reused in other areas.

3.1. Data/Metadata Infrastructure

CMCS is employing the Scientific Annotation Middleware (SAM) [5] to provide metadata management capabilities. SAM presents a Web-based Distributed Authoring and Versioning (WebDAV or DAV) protocol [4] view of underlying data and metadata repositories. DAV is an Internet Engineering Task Force (IETF) standard set of extensions to the HTTP/1.1 protocol to support basic data management over the web including storage and retrieval of typed, opaque data files/objects, content locking, hierarchical collections and annotation of the data with arbitrary metadata. DAV defines the formatting of metadata in properties consisting of well-formed XML key:value pairs and provides operations for creating, removing, and querying them. Hence, all searchable and browsable metadata in CMCS is stored as DAV properties. Metadata can be associated with both files and collections. Metadata

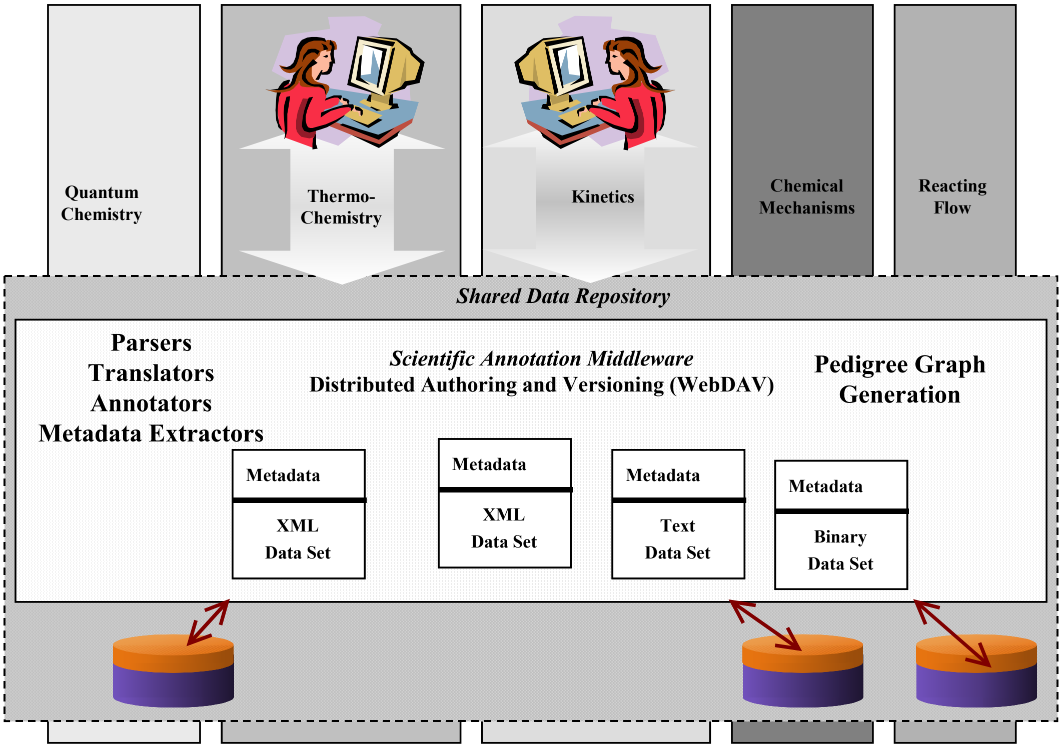


Figure 2. CMCS Infrastructure for Data/Metadata Management.

can come from multiple people - that is, not only the data creator - in multiple schema and namespaces. Metadata can also be added at any time during the lifetime of the data, that is, not only at file creation. The data itself can be any type of file.

SAM is based on Jakarta Slide [7] an open-source Java implementation of DAV. As a DAV capable server, SAM accepts arbitrary text/XML metadata, and as part of its Metadata Management Services, SAM generates user-defined metadata using metadata generation translators. Metadata generators are scripts that are executed whenever files with a given Multipurpose Internet Mail Extension (MIME) type are added to the data repository. The metadata properties to be generated are based on the MIME type (assigned based on file name extension if not assigned otherwise) and contents of the uploaded data. SAM can be configured to run a user-defined XSLT script to determine a more specific MIME type from XML files since many files have .xml extensions. When an XML file is added to the repository, SAM can run a registered Extensible Stylesheet Language Transformations (XSLT) [8] script to parse, extract, and map the metadata to the correct DAV properties. This feature is important to chemists because many new applications are using XML for data files, yet the contents of the XML files can be very diverse. When text and binary data files are added, users can specify scripts using the XML-based Binary Format Description (BFD) language to define how the data can be mapped to XML and use an additional XSLT script to populate DAV properties. The initial mapping to XML is performed by a generic BFD engine, which works analogously to an XSLT processing engine. In either case, as long as some of the metadata that is extracted and placed in DAV properties are in the CMCS schema, the data becomes searchable and browsable using CMCS tools.

SAM provides additional mechanisms for federating data repositories and for generating relationship graphs. Additional capabilities related to managing semantic relationships and implementing electronic notebook/records semantics are under development. The current state of these capabilities is discussed briefly in relation to CMCS tools in Section 4.

In addition to SAM metadata extractors, metadata can be placed into DAV properties in CMCS in three other ways. First, we have developed a portlet in our CMCS portal that allows users to browse, modify, and add arbitrary metadata in any namespace or schema to resources. Second, any DAV client application can be used to add metadata to DAV properties. Since DAV is becoming a very popular standard, the use of other DAV clients to interface with CMCS is appealing to the CMCS development team and application scientists. Third, we have developed a CMCS data management API that facilitates the insertion and modification of metadata, in the proper XML format. This Java code allows scientists to easily write programs to add/edit metadata and to integrate with existing and new chemical science applications. The CMCS data management API uses open source DAV and XML libraries. The CMCS data management API is used in CMCS behind the portlet tool and several chemistry applications, which we describe in Section 4.

3.2. Use of Dublin Core in CMCS Metadata

The Dublin Core Element Set [9] provides a good foundation for the description of electronic resources, including data. In the CMCS, we are employing the following DC elements: *Title*, *Creator*, *Subject*, *Description*, *Publisher*, *Contributor*, *Date*, *Type*, *Format*, *Source*, *Language*, *Relation*, and *Rights*. In addition, we are using the following Dublin Core Element Refinements [10]: *Abstract*, *Created*, *Valid*, *Available*, *Issued*, *Modified*, *Is Version Of*, *Has Version*, *Is Replaced By*, *Replaces*, *Is Referenced By*, and *Has References*.

In particular, some of the most useful DC elements for capturing scientific pedigree relationships are *Is Version Of*, *Has Version*, *Is Replaced By*, *Replaces*, *Is Referenced By*, and *Has References*. These elements, as well as some elements in our CMCS schema are very useful for capturing the relationship of scientific data sets to other data and for showing the traceability of data to its ultimate origin. For example, the *Replaces* and *Is Replaced* element refinements allow us to capture when data (input data, configuration files, software) has been replaced by new or updated data sets. If a scientist can easily discover that some data has been updated with newer values, it can reduce the time for producing newer results in other computations, possibly at other scales.

DC is not often used in scientific or chemistry data files; however, these files typically contain information about authors, relationships, and dates, or this information is available to the data owner. Since we can map from any data format, we use DC in our metadata properties in the data repository (note: without changing the data file itself) and have decided to build CMCS tools to use DC information while allowing researchers to use more familiar terminology or chemical science schemas within their data.

3.3. CMCS Metadata Extensions

We have defined a small core schema with the following metadata elements that enable chemistry-specific searching:

- *Species Name*: Used to identify one or more chemical species, by their common name.
- *Species CAS*: Used to identify one or more chemical species, by their Chemical Abstracts Service (CAS) number.
- *Species Formula*: Used to identify one or more chemical species, by their chemical formula.
- *Chemical Property*: Used to identify one or more chemical properties, by their common name. The current list of terms include vibrational frequency, molecular geometry, absolute energy,

enthalpy of formation, entropy, specific heat, heat capacity, free energy differences, excitation energy, bond dissociation energy, rotational barriers, potential energy surface, shock tube, premixed flame, nonpremixed flame, flow reactor, stirred reactor, static reactor, species concentration, temperature, and pressure.

We have also defined the following metadata elements, which are useful properties for defining the relationship of scientific data to projects and related inputs and outputs:

- *Has Inputs*: Used to define the input files needs to recreate the data results.
- *Has Outputs:* Used to define the output files created by the process.
- *Is Part Of Project:* Used to reference a project to a data collection or resource.

Finally, we have defined one metadata element for scientific publication and peer review annotations:

- *Is Sanctioned By*: Used to reference one or more review boards or organizations that have approved or "blessed" the data.

In addition, our CMCS schema is extensible, and new metadata can be added to this schema. Other schemas can be developed, and these metadata properties can be made available as DAV properties. We are currently exploring ways that this core pedigree schema can be refined and expanded. For example, we anticipate that some communities and some users will require finer distinctions than *cmcs:hasinputs* to differentiate parameters from data, etc. The CMCS mapping mechanisms will allow us to easily associate such new elements with the overall concept of *cmcs:hasinputs* and the wider concept of pedigree. Also, we expect to add new elements for scientific peer review. Currently, we only have defined the notion of a formal approval yet the scientific peer review process is much richer than a single stamp of approval.

In addition to the chemistry-specific metadata, we have defined some CMCS internal metadata elements which allow us to keep track of data translators and data viewers that are applicable to each resource. These values enable visualization of data files through the portal.

3.4. Metadata Examples

Since all DAV properties must be expressed in well-formed XML, we are using both XML and Resource Description Framework (RDF) to encode the metadata values. We are using the encodings for the Dublin Core Element Set and Qualified Dublin Core in XML and RDF [11, 12] and have developed a similar encoding for the CMCS core metadata set. At this time, our use of RDF is minimal but we have incorporated some RDF in anticipation of more general use in the future. We recognize the power of RDF and the subject-object-verb structure and have structured our property names as verbs in anticipation of an expanded use of RDF in CMCS in the future. We are using XLink, the XML Linking Language [13], to express values of hyperlinks. XLink provides advanced, scalable, and maintainable hyperlinking and addressing functionality for XML.

Here is an example of a value for a *cmcs:hasinputs* metadata property, which gives URIs for two different input files:

```
<cmcs:hasinputs>
  <rdf:Bag>
     <rdf:li>
       <cmcs:href xlink:href =
"http://cmcs:18081/cmcs/nwch.in"
          xlink:type="simple"
          xlink:role="nwchem-input"
          xlink:title="NWChem Input
Parameters" />
     </rdf:li>
     <rdf:li>
       <cmcs:href xlink:href =
"http://cmcs:18081/cmcs/basis.xml"
          xlink:type="simple"
          xlink:role="basis-set"
          xlink:title="Basis Set" />
      </rdf:li>
     </rdf:Bag>
</cmcs:hasinputs>
```

The use of hyperlinks as metadata within DAV properties allows users to link resources to each other. This can be very powerful to scientists when they need to find a pedigree tree for a specific resource. We show this and other applications of the CMCS metadata in the following section.

4. Tools for Utilizing and Manipulating CMCS Metadata

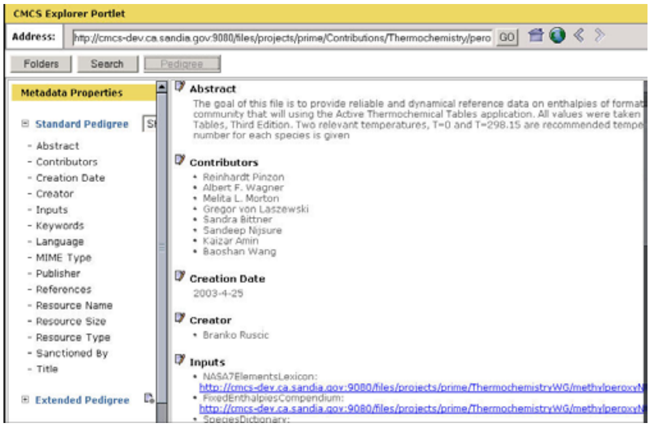
Now that we have described our metadata infrastructure, we discuss the tools we have created to create, modify, browse and search metadata in the MCS portal. In addition, we show how other applications contribute to CMCS and/or use CMCS metadata schemas to share concepts and enable collaboration of application scientists working at different chemical scales.

4.1. Portlet to Browse and Manipulate Metadata

We have developed a CMCS Explorer portlet that allows users to browse CMCS metadata in the data repository and add and edit metadata. A screenshot of the metadata viewer and editor can be seen in Figure 3. As you can see, the metadata appears in a human-readable format (text and hyperlinks), even though it is stored as XML.

On the left hand side of the portlet is the selection of metadata schemas and/or groups, and on the right hand side of the portlet are the selected property values. A user can specify metadata by known schemas and groups, including

DC, CMCS, and/or CMCS core pedigree metadata (which includes some DC and CMCS elements). Additionally, the portlet will discover and display any metadata in other namespaces and schemas. A user can edit the property values in either XML/XLink or plain text to be stored as XML and/or XLink. URIs are used to reference other data collections or resources within the CMCS data repository.



http://cmcs-dev.ca.sandia.gov:9080/files/projects/prime/Contributions/Thermochemistry/pero

Abstract

The goal of this file is to provide reliable and dynamical reference data on enthalpies of format community that will using the Active Thermochemical Tables application. All values were taken Tables, Third Edition. Two relevant temperatures, T=0 and T=298.15 are recommended tempe number for each species is given

Contributors

- Reinhardt Pinzon
- Albert F. Wagner
- Melita L. Morton
- Gregor von Laszewski
- Sandra Bittner
- Sandeep Nijsure
- Kaizar Amin
- Baoshan Wang

Creation Date

2003-4-25

Creator

- Branko Ruscic

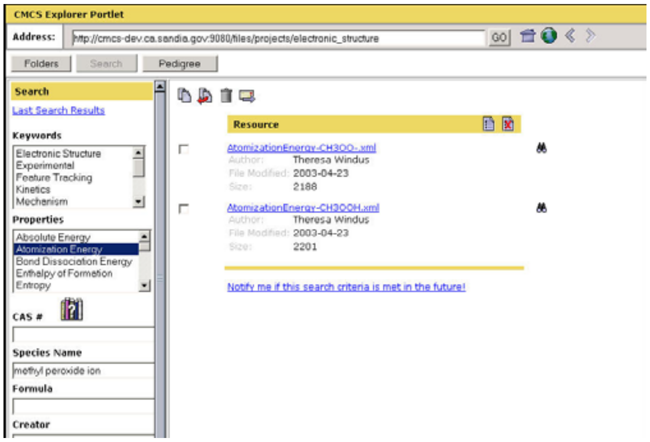
Inputs

- NASA7ElementsLexicon: http://cmcs-dev.ca.sandia.gov:9080/files/projects/prime/ThermochemistryWG/methylperoxyN
- FixedEnthalpiesCompendium: http://cmcs-dev.ca.sandia.gov:9080/files/projects/prime/ThermochemistryWG/methylperoxyN
- SpeciesDictionary:

Figure 3. CMCS Explorer portlet to view, add and edit metadata.

4.2. Search Portlet

As seen in Figure 4, we have developed a search portlet which can search the CMCS data repository based on the following metadata: keywords (*dc:subject*), creator (*dc:creator*), chemical species (search by *cmcs:speciesname*, *cmcs:speciescas*, or *cmcs:speciesformula*), or chemical property (*cmcs:chemicalproperty*). On the left hand side of the portlet the user can specify the search criteria, and the results are returned on the right hand side of the portlet. At this time, all searches begin with a DAV URL, seen in the portlet, and propagate through the DAV directory hierarchy.



http://cmcs-dev.ca.sandia.gov:9080/files/projects/electronic_structure

Species Name: methyl peroxide ion

AtomizationEnergy-CH3OO-.xml	
Author:	Theresa Windus
File Modified:	2003-04-23
Size:	2188
AtomizationEnergy-CH3OOH.xml	
Author:	Theresa Windus
File Modified:	2003-04-23
Size:	2201

Figure 4. Searching metadata in CMCS.

At present, most new data is expected to be stored in CMCS's DAV aware data repository. However, CMCS does have the ability to federate external databases to support search and annotation. For example, we have defined a mapping (again an XML document registered with SAM) for a NIST chemical kinetics database that defines how various elements of that database's internal schema should be exposed as DAV content and associated properties in the CMCS core schema. When another database is mapped in DAV, users can then add additional metadata, making a third party database annotatable by the entire chemical science community.

4.3. Pedigree Browser Portlet

As seen in Figure 3, a pedigree browser (the same tool as the metadata viewer/editor) allows users to see the scientific pedigree for a particular resource. If the user clicks on a link in pedigree browser, the user can traverse the pedigree tree through the CMCS DAV store, i.e., links are live. If a new CMCS DAV link is encountered in the pedigree browser, its pedigree is brought into the pedigree browser. If a non-CMCS-DAV link is encountered, it is treated as any other URL, and a new web browser window is brought up, loaded with that URL.

As one can imagine, there are multiple relationships (*dc:references*, *cmcs:hasinputs*, *cmcs:hasoutputs*, to name a few) and the sub-tree for each reference can get complicated. For example, a resource may have multiple links listed for each relationship value and the scientific pedigree may be traced across multiple chemical scales, making the length of each pedigree branch greater than one. By simply traversing links, a user has a difficult time seeing how the data is related to other resources. Hence, we have developed a pedigree graphing portlet, shown in Figure 5. This portlet provides scientists with a two-dimensional visualization of a data collection or file and all of its scientific pedigree relationships. The user can now easily see the relationships without following all pedigree links.

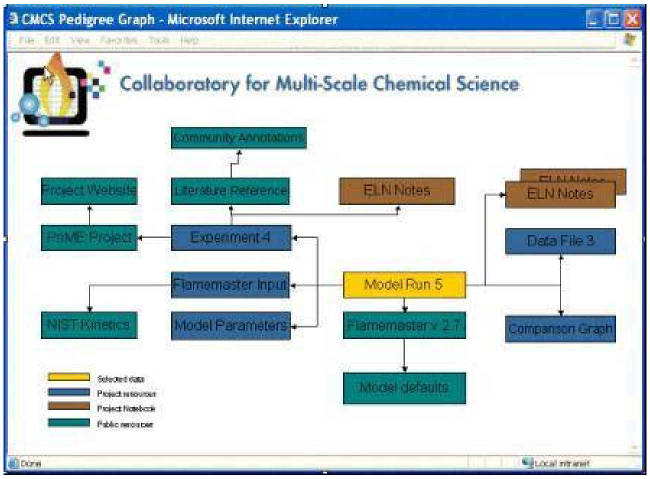


Figure 5. Pedigree graph portlet.

The CMCS pedigree graph is generated dynamically by SAM. SAM returns a value for a live DAV property called *cmcs:pedigreegraph* in Graphical eXchange Language (GXL). The SAM server is configured with all possible pedigree relationship property names. The GXL is then converted to Scalable Vector Graphics (SVG) format, viewable by many browsers. The CMCS pedigree graph is color-coded with a different color for each relationship. If a non-CMCS-DAV link is encountered while traversing the tree, it is printed as a leaf and not followed further. In the future, we plan to support the generation of pedigree graphs across DAV servers, assuming that the servers are configured to be federated.

Currently, the pedigree graph is a static picture. We are exploring the ability to create graphs with live links, and we expect to have this feature in place by the next release. Furthermore, we have plans to use the pedigree tree as an optional scoping mechanism for searching (instead of scoping based on the default DAV directory structure).

4.4. CMCS Tool Registry

The CMCS tool registry is a special metadata repository in the CMCS DAV store that stores metadata about common CMCS translators, software, science applications, and tools. This allows chemical scientist to create a pedigree link from metadata on a data set to program versions, and possibly bugs for that version. Furthermore, other scientists can comment on the quality of the software or tool in the registry, and other collaborators can then see this information.

4.5. Other Applications

Several additional applications contribute metadata and data to the system. Some of this metadata is in or is translated to the core CMCS schema and is interpreted by the CMCS portlets. The Extensible Computational Chemistry Environment (Ecce) [6], a DAV aware problem solving environment used at the molecular chemical scale, directly generates properties in the CMCS schema related to the creator, chemical species, and pedigree relationships among the large number of files produced during a computational chemistry calculation, etc. Then, using the data management API, this metadata is explicitly written to DAV properties in the CMCS data repository when data is exported to CMCS.

The electronic laboratory notebook (ELN) [15], which can be launched from a CMCS portlet, is also DAV aware and can directly store notes in the CMCS repository. It also generates DC and pedigree relationship properties (chapter-page-note relationships as well as those between data in the general CMCS repository and notes about that data in the notebook) that can be viewed using the CMCS Explorer and pedigree browser.

Another chemistry application, Active Thermochemical Tables (ATcT) [15,16,17] has been integrated with CMCS as a web service. The ATcT web service has been configured to read and write required data files to/from the CMCS repository, but it does not directly generate DAV properties. Instead, several related metadata extraction scripts are automatically applied to the various types of uploaded ATcT data to create CMCS-related properties. While all three of these applications have been able to maintain a high degree of independence, they none-the-less appear integrated with CMCS pedigree and search tools.

5. Conclusions

CMCS is developing concepts and tools for the management of scientific metadata and pedigree. We have developed a basic set of capabilities that are enabling several international groups of chemists to assemble data libraries, assess the entries for overall quality and applicability to the group's efforts via metadata and scientific pedigree, and record their group processes and final conclusions for dissemination to the public. These groups are using the CMCS metadata infrastructure and tools to collaborate and affect science. The specific details of these chemistry groups are beyond the scope of this paper; however, the data, metadata and pedigree information available within CMCS will clearly get beyond a "toy problem".

The CMCS portal, the SAM/DAV data repository, and the CMCS Explorer portlet, as well as some other tools and APIs, are independent of the chemical science community and have broad applicability to other scientists – e.g., the physics or bioinformatics communities. We expect to release a general science portal and metadata and scientific pedigree infrastructure as open source software in the near future. Other science communities can create an instance of our metadata-aware collaboratory with minimal effort.

6. Future Work

We have begun to explore both third party annotation and scientific peer review, enabled in the CMCS data/metadata repository. There are a number of social and legal issues that must be addressed to allow third party annotators to add metadata to data sets in addition to the development of a formal process and supporting tools.

There is a potential for mismatch of pedigree granularity across scales, and we are investigating methods to resolve this issue so that the scientific pedigree is unambiguous across scales. Further planned enhancements to CMCS include visually generating mappings, extracting conceptual models from the registered mappings, etc. to provide even more power to researchers.

CMCS is developing technology that will be needed for next-generation science efforts and is exploring the use of semantic technologies for managing scientific data. The result will be an enduring metadata infrastructure with an extensive, richly interconnected set of chemistry data and metadata. We believe our concepts will be particularly valuable in scientific research where knowledge is incomplete and changes rapidly in response to new data.

Acknowledgements

The Collaboratory for Multiscale Chemical Science (CMCS) is a National Collaboratories Program project sponsored by the U.S. Department of Energy's Office of Mathematical, Information, and Computational Sciences. Participating organizations include Sandia National Laboratories, Pacific Northwest National Laboratory, Argonne National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, National Institute of Standards and Technology, Massachusetts Institute of Technology, and University of California, Berkeley. Parts of this work were also supported by the U.S. Department of Energy, Division of Chemical Sciences, Geosciences and Biosciences of the Office of Basic Energy Sciences under Contract No. W-31-109-ENG-38 (Argonne National Laboratory).

Sandia National Laboratories is operated by the Sandia Corporation under Contract No. DE-AC04-94-AL85000 with the U.S. Department of Energy. Pacific North Pacific Northwest National Laboratory is operated by Battelle for the US Department of Energy under Contract No. DE-AC06-76RLO 1830. Argonne National Laboratory is operated by the University of Chicago under Contract No. W-31-109-ENG-38 for the U.S. Department of Energy. Los Alamos National Laboratory is operated by the University of California for the National Nuclear Security Administration (NNSA) of the U.S. Department of Energy under Contract No. W-7405-ENG-36. Lawrence Livermore National Laboratory is operated by the University of California under Contract W-7405-ENG-48 with the U.S. Department of Energy.

References

[1] http://www.cmcs.org

[2] Buneman, P., Khanna, S., and Tan, W.C., “Why and Where: A Characterization of Data Provenance”, Proceedings of the *International Conference on Database Theory (ICDT),* 2001.

[3] http//:www.dublincore.org

[4] http://www.webdav.org

[5] Myers, J.D., Chappell, A., Elder M., Geist A., and Schwidder, J., “Re-Integrating the Research Record”, *IEEE Computing in Science and Engineering*, pp. 44-50, May/June 2003 (Vol. 5, No. 3); Available at: http://www.scidac.org/SAM/

[6] Black, G., Gracio, D., Schuchardt, K., and Palmer, B, “The Extensible Computational Chemistry Environment: A Problem Solving Environment for High Performance Theoretical Chemistry”, in *Proceedings of Computational Science - ICCS 2003*, International Conference, Eds. P.M.A. Sloot, D. Abramson, A. Bogdanov, J.J. Dongarra, A. Zomaya, and Y. Gorbachev, vol. 2660, *Lecture Notes in Computer Science* (Springer-Verlag, Berlin, 2003).

[7] http://jakarta.apache.org/slide/index.html

[8] XSL Transformations (XSLT) Version 1.0, 1999, from World Wide Web Consortium web site: http://www.w3.org/TR/xslt

[9] Dublin Core Metadata Element Set, Version 1.1: Reference Description, 2003, from Dublin Core Metadata Initiative web site: http://www.dublincore.org/documents/dces/

[10] DCMI Metadata Terms, 2003, from Dublin Core Metadata Initiative web site: http://dublincore.org/documents/2003/03/04/dcmi-terms/

[11] Expressing Simple Dublin Core in RDF/XML, 2002, from Dublin Core Metadata Initiative web site: http://dublincore.org/documents/2002/07/31/dcmes-xml/

[12] Expressing Qualified Dublin Core in RDF / XML2002, from Dublin Core Metadata Initiative web site: http://dublincore.org/documents/dcq-rdf-xml/

[13] XML Linking Language (XLink) Version 1.0, June 2001, from World Wide Web Consortium web site: http://www.w3.org/TR/xlink/

[14] Myers, J., Mendoza, E., and Hoopes, B., “A Collaborative Electronic Notebook”, *Proceedings of the IASTED International Conference on Internet and Multimedia Systems and Applications (IMSA 2001),* August 13-16, 2001 Honolulu, Hawaii.

[15] Ruscic, B., Pinzon, R.E., Morton, M.L., Wang, B., Wagner, A.F., von Laszevski, G., Bittner, S., Amin, K.A., Nijsure, S.G., and Minkoff, M., "Active Thermochemical Tables", Version 1.00 Beta, Argonne National Laboratory, Argonne, IL, 2003.

[16] Ruscic, B., Michael, J.V., Redfern, P.C., Curtiss, L.A., and Raghavachari, K., “Simultaneous Adjustment of Experimentally Based Enthalpies of Formation of CF3X, X = nil, H, Cl, Br, I, CF3, CN, and a Probe of G3 Theory,” *J. Phys. Chem. A.* 102, pp. 10889-10899, 1998.

[17] Ruscic, B., Litorja, M., and Asher, R.L., “Ionization Energy of Methylene Revisited: Improved Values for the Enthalpy of Formation of CH2 and the Bond Energy of CH3 via Simultaneous Solution of the Local Thermochemical Network”, *J. Phys. Chem. A.* 103, pp. 8625-8633, 1999.