Cluster Computing 8, 243–253, 2005
© 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.

A Collaborative Informatics Infrastructure for Multi-Scale Science

JAMES D. MYERS
National Center for Supercomputing Applications, Champaign-Urbana, IL 61820

THOMAS C. ALLISON
NIST, Gaithersburg, MD 20899-8381

SANDRA BITTNER
Argonne National Laboratory, Argonne, IL 60439-4844

BRETT DIDIER
Pacific Northwest National Laboratory, Richland, WA 99352

MICHAEL FRENKLACH
University of California, Berkeley, CA 94720-1740

WILLIAM H. GREEN JR.
NIST, Gaithersburg, MD 20899-8381

YEN-LING HO
Los Alamos National Laboratory, Los Alamos, NM 87545

JOHN HEWSON and WENDY KOEGLER
Sandia National Laboratories, Livermore, CA 94551-0969

CARINA LANSING
Argonne National Laboratory, Argonne, IL 60439-4844

DAVID LEAHY and MICHAEL LEE
Sandia National Laboratories, Livermore, CA 94551-0969

RENATA MCCOY
Pacific Northwest National Laboratory, Richland, WA 99352

MICHAEL MINKOFF, SANDEEP NIJSURE and GREGOR VON LASZEWSKI
Argonne National Laboratory, Argonne, IL 60439-4844

DAVID MONTOYA
Los Alamos National Laboratory, Los Alamos, NM 87545

LUWI OLUWOLE
MIT, Cambridge, MA 02139

CARMEN PANCERELLA
Sandia National Laboratories, Livermore, CA 94551-0969

REINHARDT PINZON
Argonne National Laboratory, Argonne, IL 60439-4844

WILLIAM PITZ
Lawrence Livermore National Laboratory, Livermore, CA 94551

LARRY A. RAHN
Sandia National Laboratories, Livermore, CA 94551-0969

BRANKO RUSCIC
Argonne National Laboratory, Argonne, IL 60439-4844

KAREN SCHUCHARDT and ERIC STEPHAN
Pacific Northwest National Laboratory, Richland, WA 99352

A. WAGNER
Argonne National Laboratory, Argonne, IL 60439-4844

THERESA WINDUS
Pacific Northwest National Laboratory, Richland, WA 99352

CHRISTINE YANG
Sandia National Laboratories, Livermore, CA 94551-0969

Abstract. The Collaboratory for Multi-scale Chemical Science (CMCS) is developing a powerful informatics-based approach to synthesizing multi-scale information in support of systems-based research and is applying it within combustion science. An open source multi-scale informatics toolkit is being developed that addresses a number of issues core to the emerging concept of knowledge grids including provenance tracking and lightweight federation of data and application resources into cross-scale information flows. The CMCS portal is currently in use by a number of high-profile pilot groups and is playing a significant role in enabling their efforts to improve and extend community maintained chemical reference information.

Keywords: collaboratory, knowledge grid, provenance, multi-scale data, system science, community data, cyberenvironment

1. Introduction

Combustion research is, in some ways, representative of many areas that address complex multi-scale phenomena, such as earth system studies, fusion research, and high-energy physics. In such fields, an understanding of environment, device, and/or system scale phenomena requires more than simply applying one type of computation, with increased computing power, across scales. Different physical phenomena dominate system dynamics at these different scales, leading to a variety of conceptual models and associated experiments and computations relevant in the different regimes. However, researchers working in areas of chemical science relevant to combustion do not consider themselves part of a large project team to the degree seen in other fields. Rather they see combustion as one of myriad application areas that rely on more fundamental community reference data.

One of the major bottlenecks in multi-scale research today is in the passing of information from one level to the next in a consistent, validated, and timely manner. Traditionally, this information flow has been accomplished at the community level through the research literature. Increasingly, the literature is supplemented by community databases. In more project oriented disciplines, collaboratory and grid models for virtual organizations are starting to play a role. However, these approaches do not yet scale well with respect to community, process, and data/metadata heterogeneity [30,37,38].

Bridging these two paradigms (project and community-style interaction) involves numerous challenges but is essential to the full realization of systems-science approaches. The Collaboratory for Multi-scale Chemical Science (CMCS) [36] is designed to address many of these challenges and provide rich group-level collaboration capabilities, facile bi-directional data flow between groups and larger communities and between communities, and community-level review and curation mechanisms.

Once communication across scales and between projects and communities is sufficiently transparent, the character of research can change. Researchers can use information from multiple scales/communities to actively guide their research and interdisciplinary efforts can form to address related data and model requirements simultaneously across a range of scales. The ready availability of genome and protein data and analysis tools has had such an effect at the molecular scale in bioscience. The already revolutionary impacts there are motivating biologists to extend, and the broader scientific community to investigate, informatics-based approaches across multiple scales and disciplines [5,14,15,35].

2. Combustion, and the collaboratory for multi-scale chemical science

The CMCS project was initiated in 2001 with a long-term vision of multi-scale science enabled by modern informatics, and a commitment to realizing this vision in support of combustion research. The team proposed a general collaborative informatics infrastructure, reusable across disciplines, that would then be customized within the project to support combustion research. This science focus is motivated by the importance of the combustion problem, its classic multi-scale nature, and the opportunities for near-term impact by such a systems approach.

Combustion research relies on chemical information spanning more than nine orders of magnitude, in terms of both length and timescale. For realistic fuels, the chemistry of combustion involves hundreds to thousands of chemical species participating in thousands of reactions. The structural and thermochemical properties of these species are determined from spectroscopic experiments and, increasingly, from computational quantum chemistry. Reaction rates as a function of temperature and pressure are determined experimentally and by a number of computational methods using detailed data from quantum chemistry computations. Collections of these properties and rates are assembled into chemical mechanisms to model chemical transformation associated with a whole suite of reactions. These models (and often reduced forms of them) are used in detailed simulations and experiments that investigate the coupling of reaction chemistry to fluid dynamical processes in real-world device geometries. These interactions are then further modeled in codes that seek to provide predictive model-based design for combustion devices or systems.

Combustion scientists have developed numerous proven techniques, models, and reference data that span multiple physical scales. Multiple experimental and theoretical techniques are available to provide complementary constraints on the values of fundamental chemical properties. Advances in laser-based measurement methods and rapidly advancing computational algorithms and computer technology are offering new approaches for tackling the complexities of real-fuel chemistry, turbulence-chemistry interactions, and multi-phase processes [2,3,19,33]. While the combustion community is broadly distributed today with many stakeholders, often with conflicting priorities, researchers in the field are recognizing that a new 'systems approach' is required. This approach would simultaneously span disciplines, physical scales, and geography, making data, tools, and new approaches broadly available to significantly enhance progress towards the ultimate goal of predictive model-based device design. Further, given the extensive base of chemical knowledge available, such an informatics-based systems approach is seen as having the potential for immediate impacts.

The expanded role of informatics envisioned requires a new research infrastructure and improved tools and approaches. The heterogeneity inherent in multi-scale science requires strong support for converting information between conceptual models and across formats. In addition, scalable and maintainable solutions require increased abstraction of services and resources, while researchers will need better mechanisms to coordinate their activities within dynamically evolving groups and communities. The emerging vision for meeting these requirements is the 'knowledge grid,' which incorporates advances being made in semantic web, informatics, collaboratory, and grid communities. Effective use of knowledge grids will also require cultural changes and iterative approaches involving long-term collaborations of domain and information researchers to fully realize the potential of cross-scale, systems-oriented informatics.

3. Developing a collaboratory for multi-scale chemical science

To enable such a 'knowledge grid' approach, the CMCS project's approach couples a multi-disciplinary team with efforts to develop a multi-scale informatics portal toolkit, to integrate key chemistry resources, and to develop chemistry-specific informatics applications. This is accomplished through an iterative development and deployment process that is driven by guiding use-cases and feedback from pilot user groups.

As depicted in figure 1, the CMCS portal provides its capabilities through a multi-tiered architecture and layered services which integrate a wide range of chemistry community data, application resources and state-of-the-art computing technologies. In many ways, CMCS leverages current web and grid distributed computing architectures. However, to support the heterogeneity and rapid change inherent in multi-scale

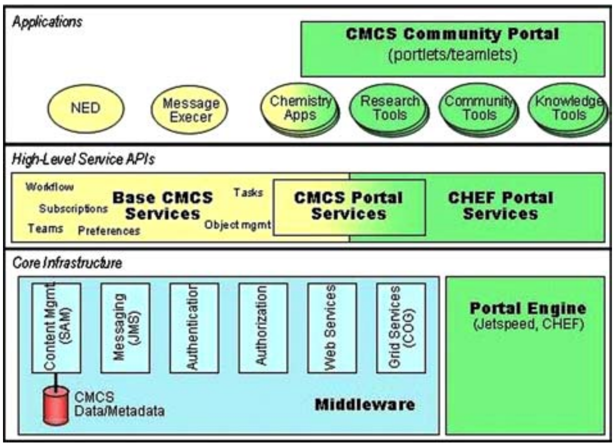

Figure 1. A schematic view of the CMCS multi-tier approach integrating chemistry data, applications, and informatics tools to enable multiscale community research.

Community Interaction
Knowledge Management
Research Support

Figure 2. CMCS provides a coherent suite of tools for developing and using knowledge repositories, interacting within scientific communities, and performing research at the level of individual projects.

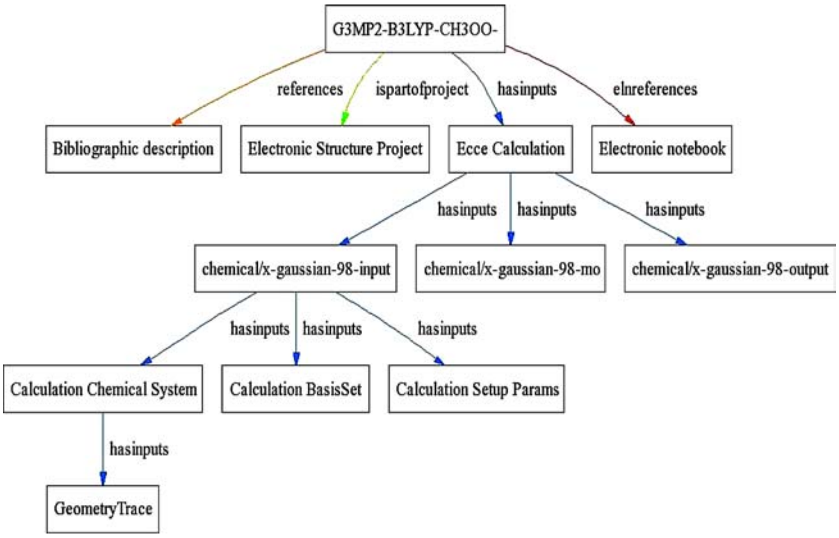

Figure 3. The Provenance Tool displays a labeled graph showing data relationships. This example shows the chain of data sets used to derive a result, and that this result is cited in a specific paper, is part of a larger project and is discussed in an electronic notebook. In the actual display, relationships are color-coded, making it easy to quickly identify relationships and patterns of interest.

science, CMCS places emphasis on lightweight integration supported by aspect-oriented design [20].

Key to the CMCS design is the principle that application-oriented services, middleware, and grid-level services should interact based on aspects of information in the system, e.g. a data set's name or type, rather than a common object model. Another key to this approach is the elevation of data and metadata transformation and semantic interpretation to first class operations that can be invoked by applications and services as needed. By making the management of scientific data and process models explicit, this design eliminates the need for up-front, global agreement on these models and decouples the evolution of the science being researched from the lifecycle of the software infrastructure.

To realize this design, CMCS leverages middleware standards and open source toolkits related to the concepts of portals, content management, and publish/subscribe messaging that embody these principles. The project team, along with a number of collaborators, is extending and evolving this base and approach to build the CMCS combustion research portal and to further develop the concept of lightweight, powerful knowledge grids.

In the subsections below, we first discuss the core infrastructure and application-level services, followed by the portal itself. We group the primary portal and data exchange interfaces into three categories for discussion—knowledge management, community interaction, and research support, though in use, researchers can access capabilities across these categories as needed.

3.1. Core infrastructure

The concept of portals and portlets is primarily one of integration at the user interface layer. CMCS reviewed a number of technologies before selecting the Apache Jetspeed [18] based CompreHensive collaborativE Framework (CHEF) [8] as a portal engine. CHEF extends the portlet concept to that of collaborative 'teamlets' that can communicate between instances of themselves, e.g. to provide real-time chat capabilities.

For data and metadata management, CMCS employs the scientific content management services developed within the DOE-funded Scientific Annotation Middleware (SAM) project [22]. SAM software provides a range of capabilities for storing and retrieving data and metadata, searching, versioning, locking, and providing access control, as well as extensible mechanisms for extracting metadata from files, performing translations, and managing provenance and other data relationships. SAM, based on the open source Jakarta Slide [17] content repository, presents an XML-centric Web-based Distributed Authoring and Versioning (WebDAV) protocol [34] view of underlying data and metadata repositories. WebDAV is a widely adopted Internet Engineering Task Force (IETF) standard set of extensions to the HTTP/1.1 protocol supporting basic content management over the web. Both WebDAV and SAM support the aspect-oriented design integral to CMCS.

To support messaging, the team selected the standard Java Messaging Service (JMS) publish/subscribe protocol and the open source OpenJMS implementation [28]. The publish/subscribe model maintains the independence of the producers and consumers of notifications, allowing dynamic changes in message routing and workflows.

For authentication and authorization, CMCS chose to initially leverage the basic username/password and role-based access control mechanisms available from underlying tools coupled with wire-level encryption provided via the Secure Socket Layer (SSL). As a longer term strategy, the team has designed the architecture to allow authentication and authorization to be handled by separable services and to support an incremental security model which allows anonymous access to public CMCS resources while supporting certificate-based credentials for access to grid resources. This plan minimizes barriers to adopting CMCS while meeting the current and anticipated needs of users. To support single-sign-on across the portal environment, CMCS initially modified the CHEF user management service to synchronize user information with the SAM repository but this is currently being updated to allow the SAM repository to directly share the CMCS user database and authentication service by replacing its default Java Authentication and Authorization Service (JAAS) based security module with a CMCS specific one.

CMCS has implemented a number of web services based on Apache's Axis [1] toolkit. CMCS has used Axis to wrap several tools as noted below and has simplified the tasks of launching and monitoring web services through the development of a higher-level programming interface. The Commodity Grid (CoG) toolkit [10] has also been used to demonstrate the ability to run applications as web or grid services as desired.

3.2. Application-level service interfaces

Application level services are high-level abstractions that combine multiple low-level services to more directly support the requirements of applications. A simple example is account creation, which, in CMCS, involves not only creation of an identity within an authorization service but association of that identity with portal preferences, the creation of directories in the repository for the user, etc. Toolkits such as CHEF and SAM provide a number of high-level service abstractions for, for example, user/team management and collaboration, and annotation and provenance tracking, respectively. CMCS is modifying, extending, and integrating these services and has developed additional ones related to messaging. In general, CMCS is aiming for a clean separation of application-level services from both the portal engine and lower-level services, to enable, for example, the real-time interaction of desktop applications with data analysis and visualization portlets. We believe this level of separation is critical to interoperability and evolution of knowledge grids over the longer term. While much of the development work in CMCS to date has been focused on immediate needs for practical integration, defining appropriate abstractions for such services is a key longer term goal of the project that is being pursued with collaborating projects. The discussion here focuses on two areas rather than an exhaustive list in an attempt to provide the general flavor of the changes while avoiding details that are changing rapidly as CMCS, CHEF, SAM, and other tools evolve.

A significant effort was made to develop a general Data Storage Interface (DSI) providing a high-level abstraction for managing metadata and data access. DSI encapsulates the WebDAV-centric SAM client library, providing higher level functionality and simplifying error handling. One simple example is that DSI supports the submission of data and associated metadata in a single call, hiding the fact that this requires two separate WebDAV operations. DSI is used within many CMCS developed tools and is also available to third party developers as an integration point. Another service abstraction that has been created is an annotation mechanism generalized from SAM's electronic laboratory notebook (ELN) and now available as a service that can be invoked from within a portlet.

CMCS has also developed new services that monitor JMS events and use these events to trigger email notifications and simple workflows to produce derived data. Currently, the events published in CMCS are limited to those from the repository indicating changes to data and metadata. We anticipate additional message sources including other services and portlet-based tools. The Notification Email Daemon (NED) service provides portlet interfaces allowing users to register interest in events based on data's location, type, and metadata values and to request individual or digested email notification of data creation, modification, and/or access. NED monitors the events from the repository, applies appropriate filters, generates digests, and sends email. A simple workflow engine uses similar methodology to trigger third party applications to analyze data and produce derived data sets.

3.3. The CMCS portal/community data exchange

The most visible parts of the CMCS infrastructure development effort are the specific portlets providing the knowledge management, community interaction, and research support capabilities. These portlets appear within the overall CMCS portal interface of the user's web browser and interact with portal engine, high-level and base services, and each other as needed. These portlets appear in the overall portal view coordinated by CHEF and Jetspeed. When a user logs in to the CMCS portal, the teams to which he belongs appear as tabs across the top of the portal workspace. Within each team space, multiple pages showing customized aggregations of portlets are available.

3.3.1. Knowledge management

Due to the very flexible and powerful data and metadata capabilities being developed in CMCS and given work being conducted within CHEF on group collaboration tools and within other Grid and collaboratory projects on research support tools (e.g. grid job launching), CMCS placed an early emphasis on developing data-centric tools supporting knowledge management activities. Knowledge management is defined here as the general ability to discover and interpret data in semantically meaningful ways. Capabilities in this category include mechanisms for searching and browsing data, organizing and finding information based on metadata, submitting new content, browsing provenance and other relationships, and discovering and managing vocabularies, schema, and ontologies.

Given the capabilities of the CMCS infrastructure, it was not possible to directly leverage data interfaces developed in other portal projects. The Explorer portlet was designed as the default CMCS data and metadata browsing tool providing a full set of basic content management operations including file upload, copy, move, and delete, and the creation of hierarchical directory structures. The Explorer supports traversing directories, viewing files sorted by different criteria (e.g. creation date, author), creating bookmarks, and setting access controls. It also allows users to create, view, and edit arbitrary key/value metadata (i.e. WebDAV properties) associated with files and directories. Metadata can describe a variety of aspects of data including its type and format, its owner and creators, the type of measurement represented by the data, its categorization, judgments of its quality, and its relationships to other data. Different metadata are relevant in different scientific tasks. The CMCS Explorer makes all of the metadata available through a common interface and provides filtering mechanisms to limit the metadata shown in the current view. For example, the CMCS has defined a relatively small set of common metadata [3] based in part on the Dublin Core [12] that is shown by default, but groups may define their own standards and customize their Explorer view to highlight their metadata.

The Explorer interprets some metadata, versus simply displaying it, to provide live links to related data including 'virtual' derived/translated formats, and to support viewing data within the portal. The Explorer creates links for any metadata including an Xlink [42] to other data within the repository that will refocus the Explorer on the link target, providing an alternative to browsing via directory hierarchies. The Explorer provides a similar mechanism to trigger the creation of derived/translated data. This capability relies on a dynamically generated "hastranslations" property generated by SAM based on the capabilities of registered translators. Translated data can be downloaded or directly copied to user-specified locations in the repository. When translations are available that produce output viewable within a browser (e.g. HTML or JPEG images), Explorer creates a "View As" list that can be used to trigger popup windows with the specified view(s) of the data. Since views generated via translation can include interactive applets, the Explorer can be used not only as a way to quickly discover and evaluate data, but as a means to then explore the data interactively.

Several additional tools are associated with the Explorer and can be launched from its interface. These tools share state, and changes in one tool can trigger changes in the others. This capability is being developed as a prototype for a more general mechanism to support inter-portlet communication. Specific tools include:

Provenance Graph Tool—provenance (or pedigree) [7] is a broad term describing the history of a piece of data. The CMCS Provenance Graph tool [24] displays a node-and-arc graph of data relationships, providing a powerful alternative to the tabular metadata view in the Explorer. The Provenance Graph is fully configurable and can be used to show any set of relationships. This is critical in allowing researchers to focus on workflow provenance (data processing history) or scientific provenance (e.g. association with refereed papers), or to highlight or hide annotations and project relationships.

Advanced Search Tool—search is an essential tool for data discovery. Search involves finding data based on comparisons with metadata patterns specified by the user. Due to CMCS's rich metadata capabilities, very complex searches are possible. To support a range of user needs, the CMCS Search tool was designed with multiple interfaces. For the simplest cases, a Google-like search box allows users to specify words to be matched across all metadata properties. A second configurable interface allows search based on specified values for multiple metadata properties. Extension of these interfaces to support, for example, specification of the search scope in semantic terms (e.g. search through data used as inputs for data in the current project), are anticipated. These interfaces are built upon a search programming interface which is also accessible to integrate search capabilities directly into portlets and applications.

Annotation Tool—annotation is the attachment of free-form information to existing data. The Annotation tool supplements the simple metadata creation capability of the Explorer with a much richer model derived in part from the SAM-based electronic laboratory notebook (ELN) [23]. The Annotation tool allows data owners and third parties to associate text, sound, images, equations, or whiteboard drawings with existing data. Annotations are stored as independent files linked via metadata to the specified data and can have their own metadata and additional relationships. Thus annotations can be protected with access controls and threaded to organize related topics of discussion.

The Explorer and associated knowledge management tools continue to evolve in response to user feedback and as other parts of the infrastructure mature and make new capabilities possible.

3.3.2. Community interaction

The CMCS portal provides a variety of portlets and tools to support community interactions. In general, these tools provide research groups and communities the ability to form, organize and scope their activities, and to disseminate their findings. Capabilities in this category include mechanisms for defining groups, managing membership, creating shared group data repositories, supporting real-time and asynchronous discussion, customizing the layout and mix of tools within the group's portal view(s), managing group processes, assigning tasks, maintaining public data collections, and managing broader community processes such as the review and annotation of data. In some sense, they bridge between the public-oriented knowledge management capabilities and the small-group oriented research/project-oriented capabilities.

Many CMCS capabilities in this area are provided by tools, or minor modifications of tools within CHEF. Examples include Chat, Discussion, and Calendar portlets. CMCS has developed a number of additional tools, some in response to specific user needs and others to begin to explore longer term issues for knowledge grids. Examples of the former include an email subscription tool for managing the event-driven notification service discussed above and a sophisticated task management portlet. The subscription portlet can be used to set up notifications for teams, sub-teams, or individuals triggered by a range of conditions from new files appearing in a given directory to matches to persistent queries based on the CMCS search capability. The task management portlet allows tasks to be categorized, prioritized, and filtered by priority, status, category, assignee, and milestone/version.

The prototype capabilities include the annotation tool described above, which is an initial step towards support for scientific peer review. The team also created a mechanism for "expert discovery" that allows users to publish requests for additional data (i.e. at other scales) that are compared with advertised expertise and interest profiles. When a match is found, the system forwards the request by email providing an introductory contact. Additional capabilities, including application and schema registries and associated management tools are in development.

CMCS has also modified and extended the basic user and team management capabilities provided by CHEF. In addition to automating the creation of access-restricted and publicly readable website/data areas in the repository for users and groups, modifications were also made to expose user and team definitions outside the portal and to allow users more control over individual and group profiles. Groups can request a URL of the form http://cmcs.org/<groupname> to provide direct public access to their CMCS website/data areas. A New User portlet allows an individual to create an account on CMCS and provide information about her combustion-related interests and expertise. The Teams portlet allows any CMCS user to create a team, thereby becoming its administrator. The administrator can then add and remove members, promote other members to the administrator role, set descriptive metadata about the team, and specify an (internal or external) team web page. Further modifications including, for example, an email invitation mechanism allowing team administrators to invite/include people who have not yet registered with CMCS, are planned.

3.3.3. Research support

The tools discussed under the topics of knowledge management and community interaction clearly provide some support for research, but they address only part of the requirements. To perform detailed experiments, calculations, and modeling runs, researchers need access to domain applications and ways to transfer and transform data, automate repetitive tasks, visualize and compare data and metadata, and document their progress. Thus, research support in the CMCS infrastructure is primarily in the form of interfaces supporting inclusion of domain resources. These are described here, with examples of resources that have been integrated to support multi-scale chemistry given in the following section.

CMCS's use of WebDAV provides a number of possibilities for data integration. WebDAV-based file system drivers exist for most platforms allowing applications to view the CMCS repository as a file system and to directly upload and download information without any modification to the application itself [11,40]. WebDAV libraries exist in a number of languages [9,25,34], and the CMCS DSI library provides a higher-level interface in Java. To support integration of external data sources, CMCS relies on capabilities within SAM to map local or geographically distributed relational databases, GridFTP stores, and other sources into the WebDAV namespace using Java or XML-based interfaces.

WebDAV also plays a role in CMCS support for metadata integration. Supplementing the manual metadata entry available within the Explorer, CMCS supports direct metadata submission and retrieval via WebDAV clients and DSI. Thus, WebDAV-aware applications have full access to CMCS metadata. To support applications that interact through file system drivers, CMCS provides a SAM-based metadata extractor mechanism that can pull information from within arbitrary (e.g. binary) files and populate WebDAV properties, converting to any schema desired.

As noted previously, CMCS has developed mechanisms to enact simple workflows based on JMS triggers. However, to date, CMCS has emphasized mechanisms to document and view workflow-based relationships over traditional workflow enactment services. The metadata extraction mechanism discussed above and WebDAV/DSI interfaces allow applications to contribute workflow documentation that is then available within the Explorer and Provenance Graph Tool. CMCS also has internal mechanisms to track provenance, adding metadata to specify the source of copies and translations requested via the Explorer or the data interfaces. The team continues to track grid workflow developments [41] and is relying on feedback from the pilot communities to prioritize inclusion of a general computational gateway in the system.

The standard portal/portlet mechanism is the primary display integration mechanism available in CMCS. CHEF provides a richer model that supports interacting portlets including server-mediated communication between portlets and a mechanism for portlets to invoke other portlets. As noted, CMCS also has a lightweight mechanism to allow portlets to share state information, making it possible, for example, for an application to control independently developed data viewer portlets to display its results. Combined with the mechanism to register translations capable of producing browser-viewable renderings of data, these mechanisms provide sufficient capabilities to allow portal environments to approach the richness of more traditional problem solving environments.

CMCS's design allows very powerful integration of all information related to an experiment. The repository can store data and associated notes, images, documents, and other material in any format and store all relationships between them. In addition to the tools for managing data and metadata, and for adding annotations that were mentioned in previous sections, CMCS provides an electronic laboratory notebook application that can be launched from the portal and stores its output directly in the CMCS repository [23]. As with other CMCS tools, the data and metadata created with the ELN, including chapter/page/note relationships, are directly available via WebDAV and the DSI interface. The notebook can be extended with additional ‘editors’ supporting entry of new types of annotation and can be configured to use any translators registered with CMCS to render data on notebook pages.

3.4. Combustion resources

While clearly relevant to combustion research, the infrastructure and tools discussed in previous sections are essentially generic. To tailor CMCS for the combustion research community, a wide range of existing combustion data collections and applications have been integrated with CMCS using these mechanisms. A number of leading file-based collections ranging from Quantum Chemistry calculation results [13] to highly annotated information on kinetic rates with related validation experiments [29], chemical reaction mechanisms for complex fuels, and feature annotations of direct numerical simulations of hydrogen-air autoignition data [21] have been made available directly within with CMCS repository. These sources have been uploaded manually via the portal or directly via applications supporting WebDAV. Other sources, of which the NIST Kinetics Database [26] is the first, will be maintained at their home institutions and mapped into the CMCS repository namespace via SAM.

Applications have been integrated with CMCS using WebDAV and via wrapping as data translators or as part of event-triggered workflows. For example, the Extensible Computational Chemistry Environment (Ecce)[6,32], a state-of-the-art problem solving environment supporting a growing number of quantum chemistry codes, has been modified to export molecular scale data to CMCS including full provenance information. Additional modifications include a new mechanism to combine multiple quantum mechanical calculations to directly produce thermodynamic information used as input to higher-scale computations. As another example, Fitdat, which converts between different community standards for parameterization of the temperature dependence of thermodynamic information, has been wrapped and can be triggered as part of a simple workflow to automatically produce alternate parameterizations of new data that are then stored in CMCS with links to the original data

The CMCS metadata and translator/viewer capabilities provide some of the most striking demonstrations that rich, lightweight integration is possible. Extractors have been created that populate the core CMCS metadata types (the default metadata displayed in the Explorer) based on information within numerous file types. Translators have been created across the range of combustion scales to convert between popular community file formats. For example, OpenBabel [27], a tool for converting between molecular geometry file formats, has been integrated via a web service wrapper registered with the CMCS repository to provide on-demand translations of molecular geometry files. Sophisticated viewers generate interactive 3D displays of molecular structures and selected molecular properties. Similarly, an XY-graph viewer supports the visual analysis of a variety of tabular data sets such as special profiles of temperature and species mass fractions or thermodynamic properties vs. temperature. Overall, the CMCS production server currently has 48 metadata extractors and 99 translations/views configured.

3.5. Chemical informatics applications

Three new informatics based applications have been developed which take advantage of the CMCS structure and capabilities. They, and the pilot communities they are enabling, represent a new level of sophistication in combustion research that is directly enabled by the ability to integrate large amounts of heterogeneous data and coordinate group analysis of the entire corpus provided by CMCS.

Active Thermochemical Tables (ATcT)—ATcT [39] is a new approach to evaluating thermodynamic data in which fundamental data representing a network of related measurements are statistically analyzed to calculate the most accurate estimate of species enthalpy given current information and to explore the impact of additional measurements. CMCS developed an Active Tables portlet and associated web service to allow users to import and export data to ATcT from their CMCS workspace and to interactively run analyses and perform queries against the ATcT chemical network.

The IUPAC (International Union of Pure and Applied Chemistry, http://www.iupac.org) Task Group ”Selected Free Radicals and Critical Intermediates: Thermodynamic Properties from Theory and Experiment” [16] is using ATcT and related tools, and CMCS’ team coordination capabilities, to critically evaluate data for a number of radicals important in combustion and atmospheric chemistry and produce recommended thermochemical values. This data will be subsequently used by the chemical community at large to create accurate models of relevant chemical reactions for complex systems such as flames or the atmosphere. The 13 members of this Task Group have been selected by IUPAC from among the most prominent scientists in the field. They will be working across as many as ten time zones to analyze data related to 30+ ephemeral chemical species.

ReactionLab—ReactionLab is a Matlab-based open source software package for modeling reaction kinetics. ReactionLab addresses key issues related to the efficient evaluation and validation of kinetics data and models and provides a richer suite of tools for exploring and comparing experimental data and model results than is available through the CMCS portal directly. ReactionLab has been extended using DSI to directly store and retrieve information from CMCS workspaces.

Range Identification and Optimization Toolkit (RIOT)—RIOT [4] generates reduced kinetic mechanisms which, when used in place of comprehensive mechanisms, significantly reduces the computational cost of solving the equations describing a reacting system, without significant loss of accuracy. RIOT can rapidly and automatically determine the smallest subset of a given mechanism that sufficiently mimics the full mechanism at specified conditions and optionally provide a rigorous estimate of how much the reaction conditions can be varied before the model becomes invalid. CMCS developed a portlet interface and associated web service that enables users to interactively view mechanisms, set tolerances and reaction conditions, interactively run RIOT, and view the results.

3.6. Community-driven development

Earlier this year, CMCS released a 1.0 version of its infrastructure and began operating a 'production' portal supporting a number of groups across a wide range of chemistry disciplines with goals ranging from simply sharing and comparing results between related research projects to gathering and analyzing data on a global scale. While many of these groups are still learning about CMCS and making fairly limited use of its capabilities, others are being quite aggressive in their adoption of CMCS and are guiding new developments. As an example of the latter the PrIMe (Process Informatics Model) group, a research team of about 40 international scientists, including kineticists, thermodynamicists, quantum chemists, experimenters, and modelers, is actively loading CMCS with highly annotated data sets, defining processes for coordinating their analysis work, directly engaging CMCS developers in refining capabilities and adding new features. PrIMe has formed to assemble thermodynamic, chemical kinetic, and transport data for use in developing a curated public multi-scale library and associated tools that can mine the library to produce, for example, customized optimal kinetics models. The PrIMe group plans to use ReactionLab, ATcT, RIOT, and numerous other CMCS capabilities to coordinate activities, evaluate data, and interact with the larger community. PrIMe is a primary driver behind current CMCS efforts to provide form-based data/metadata input, a means to create hierarchical sub-groups within a community, and mechanisms to support community-defined review processes.

4. Conclusion

The CMCS team, guided by feedback from pilot users and in collaboration with other software development projects, has developed a powerful multi-scale combustion portal and underlying informatics toolkit. The toolkit is itself will soon be available under an open source license and is expected to be used in projects developing knowledge grids for other science communities. A variety of chemistry data resources have been integrated and innovative informatics applications have been developed and made available within the portal. International teams of scientists have formed more than half a dozen pilot groups tackling important open issues in combustion chemistry.

The portal allows researchers to work naturally, accessing heterogeneous data, working within community groups, and performing research that directly integrates community data, tools, and processes. The infrastructure provides the multiple information pathways necessary to integrate tools and allow researchers to maintain a focus on their science rather than the mechanics of data transfer. Community data exchange (including translation and extraction mechanisms) and standard programming interfaces are designed into the portal, underlying services, and the data/metadata repository to make it simple to integrate third-party data and application resources without requiring modifications that would prohibit continued independent development of these resources.

These are all issues that will present challenges to future knowledge grids spanning virtual organizations. Although not discussed directly in this paper, CMCS is also helping to elucidate other knowledge grid related issues such as licensing of community data from myriad sources, procedures for direct community data review, and mechanisms for assessing community contributions independent of the scientific literature. As knowledge grids lower barriers to discovering, analyzing, and generating chemical information, technologies and research processes will need to co-evolve. Researchers will need to easily participate in multiple communities, and sub-communities will need to be able to independently develop and evolve their domain resources while contributing to multi-scale goals. We believe the approaches taken within CMCS are advancing our ability to meet these goals.

Acknowledgments

This work was supported as part of the Collaboratory for Multi-Scale Chemical Science (CMCS) project within the National Collaboratories Program sponsored by the U.S. Department of Energy's Office of Mathematical, Information, and Computational Sciences. 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 U.S. Department of Energy under Contract No. DE-AC06-76RLO 1830.

The authors wish to acknowledge members of the CompreHensive collaborativE Framework (CHEF) and Scientific Annotation Middleware (SAM) projects, and the researchers involved in CMCS pilot groups, for helpful discussions related to CMCS' design and contributions to hardening the CMCS software. Baoshan Wang and Kaizar Amin are also acknowledged for contributions to the CMCS project.

References

[1] Apache Axis Website, http://ws.apache.org/axis/ (2004) Apache Jakarta Project.

[2] A Science-Based Case for Large-Scale Simulation, Volume 1, http://www.pnl.gov/scales/docs/volume1_300dpi.pdf. Proceedings of the Workshop on the Science Case for Large-scale Simulation, (2003. Arlington, Virginia: Office of Science, Department of Energy).

[3] R.S. Barlow, Turbulent Nonpremixed Flame (TNF) Workshop, http://www.ca.sandia.gov/TNF/. This site provides links to good examples of data that the international community is posting to validate combustion models.

[4] Binita Bhattacharjee, Douglas A. Schwer, Paul I. Barton and William H. Green, Jr., Optimally-reduced kinetic models: reaction elimination in large-scale kinetic mechanisms, Combustion and Flame (2003).

[5] BioSPICE Website, https://biospice.org/ (2004) DARPA.

[6] G. Black, D. Gracio, K. Schuchardt and B Palmer, 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] P. Buneman, S. Khanna and W.C. Tan, Why and where: A characterization of data provenance, in: *Proceedings of the International Conference on Database Theory (ICDT)*, 2001.

[8] CHEF Collaborative Portal Framework Website, CHEF Collaborative Portal Framework Website, http://www.chefproject.org/, University of Michigan (2004).

[9] Patrick Collins, PerlDAV: A WebDAV client library for Perl5, http://www.webdav.org/perldav/. (2001).

[10] Commodity Grid Kits, http://www-unix.globus.org/cog/ (2004) University of Chicago.

[11] Davfs WebDAV Linux File System, http://dav.sourceforge.net/.

[12] Dublin Core Website, http://www.dublincore.org/ (2004) Dublin Core Metadata Initiative.

[13] EMSL Computational Results Database (CRDB), http://www.emsl.pnl.gov/proj/crdb/. 2004, Pacific Northwest National Laboratory.

[14] A. Gupta, B. Ludäscher and M.E. Martone, Knowledge-based integration of neuroscience data sources, in: *12th Intl. Conference on Scientific and Statistical Database Management (SSDBM)*, Berlin, Germany, IEEE Computer Society, (July 2000).

[15] Information and Communications: Challenges for the Chemical Sciences in the 21st Century. 2003, Washington, D.C.: National Academy Press.

[16] IUPAC Project 2000-013-1-100, http://iupac.chemsoc.org/projects/2000/2000-013-1-100.html, continued as 2003-024-1-100, http://www.iupac.org/projects/2003/2003-024-1-100.html. 2003, International Union of Pure and Applied Chemistry.

[17] Jakarta Slide Java Content Management System Website, http://jakarta.apache.org/slide/ (2004) Apache Jakarta Project.

[18] Jetspeed, an Open Source implementation of an Enterprise Information Portal, using Java and XML, http://jakarta.apache.org/jetspeed/site/, Apache Jakarta Project (2003).

[19] A.N. Karpetis and R.S. Barlow, Measurements of scalar dissipation in a turbulent piloted methane/air jet flame. Proc. Combust. Inst., **29** (2002) 1929–1936.

[20] Gregor Kiczales, John Lamping, Anurag Mendhekar, Chris Maeda, Cristina Videira Lopes, Jean-Marc Loingtier and John Irwin, Aspect oriented programming, in: *Proceedings of the European Conference on Object-Oriented Programming (ECOOP)*, Finland. Springer-Verlag LNCS 1241. 1997).

[21] W. Koegler, Case study: Application of feature tracking to analysis of autoignition simulation data. 12th IEEE Visualization 2001 Conference (VIS 2001), San Diego, CA 2001.

[22] J.D. Myers, A. Chappell, M. Elder, A. Geist and J. Schwidder, Re-integrating the research record. IEEE Computing in Science and Engineering, Available at http://www.scidac.org/SAM/ **5**(3) (2003) 44–50.

[23] J. Myers, E. Mendoza and B. Hoopes, A collaborative electronic notebook, in: *Proceedings of the IASTED International Conference on Internet and Multimedia Systems and Applications (IMSA 2001)*, Honolulu, Hawaii (2001).

[24] J.D. Myers, C. Pancerella, C. Lansing, K.L. Schuchardt and B. Didier, Multi-scale science: Supporting emerging practice with semantically-derived provenance, *Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, held at the 2nd International Semantic Web Conference, Sanibel Island, Florida (2003).

[25] Neon: an HTTP and WebDAV client library, with a C interface, http://www.webdav.org/neon/.

[26] NIST Chemical Kinetics Database, Standard Reference Database 17, http://kinetics.nist.gov/index.php. 2000, National Institute of Standards and Technology.

[27] Open Babel, http://openbabel.sourceforge.net/

[28] OpenJMS Website, http://openjms.sourceforge.net/ (2004) The OpenJMS Group.

[29] Overview of GRI-Mech, http://www.me.berkeley.edu/gri_mech/overview.html.

[30] Particle Physics Data Grid (PPDG) Website, http://www.ppdg.net/ (2004) DOE.

[31] C.M. Pancerella, J.D. Myers, et. al., Metadata in the collaboratory for multi-scale chemical science, in: *Proceedings of the 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice-Metadata Research and Applications (DC 2003)*, Seattle, Washington (2003).

[32] K.L. Schuchardt, J.D. Myers and E.G. Stephan, A web-based data architecture for problem solving environments: Application of distributed authoring and versioning to the extensible computational chemistry environment, Cluster Computing 5 (2002) 287–296.

[33] Scientific Discovery through Advanced Computing (SciDAC), http://www.osti.gov/scidac/SciDAC.pdf (2000) Office of Science, Department of Energy.

[34] G. Stein, Web Digital Authoring and Versioning (WebDAV) Resources Community Website, http://www.webdav.org/ (2004).

[35] R.D. Stevens, A.J. Robinson and C.A. Goble, myGrid: personalised bioinformatics on the information grid, Bioinformatics, Eleventh International Conference on Intelligent Systems for Molecular Biology, 19 (Suppl. 1) (2003).

[36] The Collaboratory for Multi-scale Chemical Science Website, http://cmcs.org/ (2004) Sandia National Laboratories.

[37] The Earth System Grid (ESG) Website, https://www.earthsystemgrid.org/ (2004) DOE.

[38] The National Fusion Grid Website, http://www.fusiongrid.org/projects/ (2004) DOE.

[39] G. von Laszewski, B. Ruscic, P. Wagstrom, S. Krishnan, K. Amin, S. Nijsure, S. Bittner, R. Pinzon, J.C. Hewson, M.L. Morton, M. Minkoff and A.F. Wagner, A grid service-based active thermochemical table framework. Lecture Notes in Computer Science (2003) 2536: 25–38.

[40] WebDrive, http://www.webdrive.com/ (2004) South River Technologies.

[41] Workflow in Grid Systems, Workshop, Global Grid Forum 10, Berlin, Germany, 9, (2004).

[42] XML Linking Language (XLink) Version 1.0, http://www.w3.org/TR/xlink/ (2001) World Wide Web Consortium (W3C).

James D. Myers received his B.A. in Physics from Cornell University in 1985 and his Ph.D. in Chemistry from the University of California at Berkeley in 1993. He is currently the Associate Director for Collaborative Technologies at the National Center for Supercomputing Applications (NCSA) at the University of Illinois, Urbana Champaign. Dr. Myers is the lead investigator on the U.S. Department of Energy (DOE) sponsored Scientific Annotation Middleware project (http://www.scidac.org/SAM/) (scientific content

management, semantic annotation, and records functionality) and is serving as the Chief Technical Officer for the DOE-sponsored Collaboratory for Multiscale Chemical Science (CMCS) project. His is also the lead architect for the Mid-America Earthquake Center's MAEViz hazard risk management collaboratory and co-lead of NCSA's Collaborative Large-scale Engineering Analysis Network for Environmental Research (CLEANER) related cybercollaboratory effort. Open source software developed by Dr. Myers and his colleagues including the electronic laboratory notebook (ELN) and the Collaborative Research Environment (CORE) real-time collaboration environment have been downloaded from the Pacific Northwest National Laboratory (PNNL) Collaboratory website (http://collaboratory.pnl.gov) by thousands of researchers and educators.
E-mail: jimmyers@ncsa.uiuc.edu

Due to space limitations, individual bios for all 28 authors are not shown. The CMCS project is led by Dr. Larry Rahn (rahn@sandia.gov) at Sandia National Laboratories. The team includes combustion researchers and computer science researchers and developers at five DOE National Laboratories (Argonne, Lawrence Livermore, Los Alamos, Pacific Northwest, and Sandia National Laboratories), the National Institute of Standards and Technology, Massachusetts Institute of Technology, and the University of California, Berkeley. Current contact information and biographic information for team members is available at http://cmcs.org/team.php.