SAND2006-4616P

Collaboratory for Multi-scale Chemical Science (CMCS): Project Final Report

Tom Allison,[9] Jared Chase,[1] Michael Chen,[2] Brett Didier,[1] Zoran M. Djurisic,[10] Wendy Doyle,[2] Todd Elsethagen,[1] Michael Frenklach,[10] William H. Green,[6] Vidhya Gurumoorthi,[1] Deepti Kodeboyina,[3] David Leahy,[4] Jun Li, [1] David Montoya,[8] James D. Myers,[5] Jan Nobel,[2] Oluwayemisi O. Oluwole,[5] Carmen Pancerella,[2] William Pitz,[7] Reinhardt Pinzon,[3] Larry A. Rahn,[2*] Branko Ruscic,[3] Karen Schuchardt,[1] Jing Song,[6] Lisong Sun,[1] Gregor von Laszewski,[3] Al Wagner,[3] Christine Yang,[2] Lili Xu[8]

[1]Pacific Northwest National Laboratory, Richland WA 99352; [2]Sandia National Laboratories, Livermore, CA 94551-0969; [3]Argonne National Laboratory, Argonne, IL 60439-4844; [4]Stanford University, Stanford, CA 94305; [5]National Center for Supercomputing Applications, Urbana, IL 61801; [6]MIT, Cambridge, MA 02139; [7]Lawrence Livermore National Laboratory, Livermore, CA 9455; [8]Los Alamos National Laboratory, Los Alamos, NM; [9]National Institute of Standards and Technology, Gaithersburg, MD 20899; [10]University of California, Berkeley, CA 94720

*Point of Contact: Larry A. Rahn, Sandia National Laboratories, PO Box 969 MS9291, Livermore, CA 94551-0969; Phone: 925-294-2091, Fax: 925-294-3282; Email: rahn@sandia.gov

1.0 Introduction

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 vision was stated as "CMCS will enhance chemical science research by breaking down the barriers to rapid sharing of validated information and by opening new paradigms for multi-scale science."

The resulting integrated capability was envisioned as a 'CMCS Knowledge Grid' as graphically depicted in Figure 1. 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. These chemical reactions occur in an environment that is defined by both thermal conduction and radiation. 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

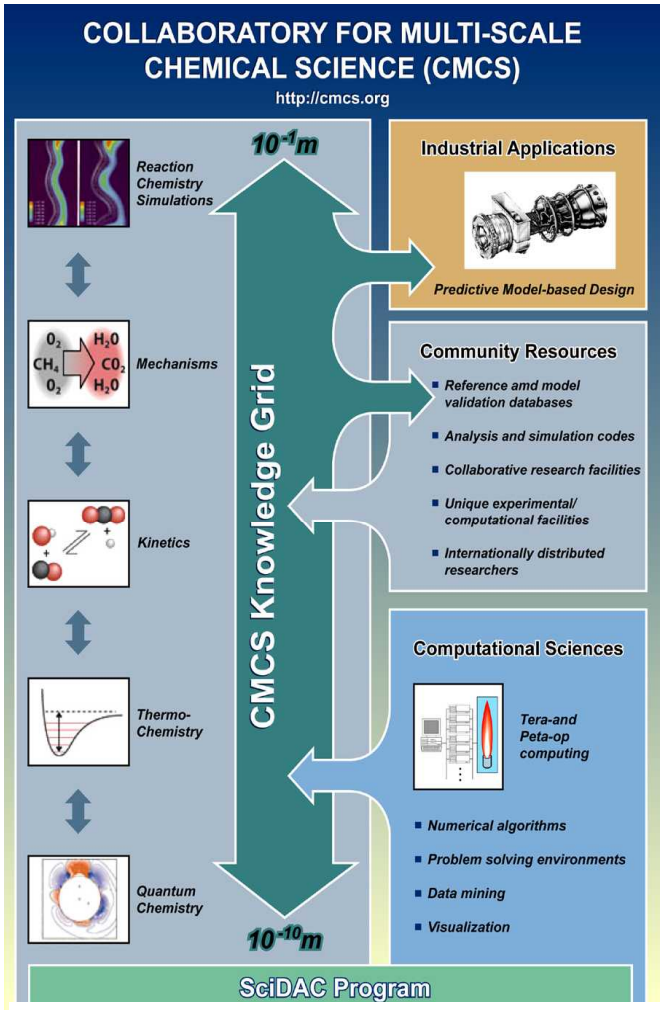

Figure 1. Predictive combustion modeling requires the integration of scientific knowledge over a large range of scales and an effective coupling among disciplines and community resources.

reduced forms of them) are used in detailed simulations that investigate the coupling of reaction chemistry to fluid dynamical processes. These interactions are then further modeled in codes that seek to provide predictive model-based design for combustion devices or systems.

Using such a knowledge grid, we envisioned diverse combustion research communities sharing data and analysis tools as they create verified, documented data sets and reference data. The grid would provide advanced pedigree tracking, data interoperability, and multi-level application support and facilitate collaboration among scientists and unique facilities. Thus the knowledge grid would couple scientists working across varying physical scales and disciplines in combustion research. Our vision was also that such a knowledge grid would not only facilitate interdisciplinary collaboration, but also enable scientists to move toward tackling data-intensive and multi-scale problems with systems science approaches. These approaches can enhance the discovery of knowledge gaps, clarify research priorities, and potentially accelerate scientific impact on industrial development and societal needs.

To enable this vision, the multi-disciplinary project team worked to develop a multi-scale informatics portal toolkit. The team also worked to integrate key chemistry resources, and to develop chemistry-specific informatics applications through an iterative development and deployment process. This process was driven by guiding use-cases and feedback from pilot user groups. This work has been communicated in numerous refereed publications [R1-R18], other publications [O1-O9], and conference presentations [P1-P18]. We found that the portal and underlying informatics toolkit developed in the CMCS project are general-purpose and useful to other science communities that have a need to develop and publish data, deploy science applications, and work together in both public and private groups. By separating the chemistry-specific applications from the infrastructure and making this collaborative science environment available to others, we have also sought to enable science projects in other disciplines. Thus, the resulting open source informatics toolkit, the Knowledge Environment for Collaborative Science (KnECS), has been separated from CMCS and licensed as open source software [A1]. KnECS has already been used by at least one other project [A2] in the development of a knowledge grid for their community. Our vision is that developers from an array of follow-on projects like CMCS will contribute enhancements to KnECS.

2.0 Significant accomplishments and scientific impact

Early in the CMCS project, the team focused on requirements definition, identifying detailed technical solutions, and on an iterative process of building prototypes for infrastructure and a few combustion applications. This process allowed iterative refinement of our vision and, most importantly, refinement of our scope. Necessarily, much of the emphasis was on getting an enabling infrastructure into place on which the team could explore actual science use cases and build application capabilities. The latter two years of project effort has thus been focused much more developing and integrating a range of applications and data into a combustion knowledge grid while refining the infrastructure, interfaces, and enabling tools. Significant progress has been made in direct collaboration with application scientists to modify and integrate legacy codes, develop schema, transform data, develop new applications, and to understand how such new approaches can benefit science. With the breadth of disciplines, applications, and data required, it is clear that different styles of application integration and community interaction must be supported. In the remainder of this section we summarize the project team's accomplishments, first in construction of the KnECS infrastructure, then highlighting a number of science capabilities and results. The latter accomplishments are ordered from the smaller scales to the larger scales (see Fig. 1), with the final one focused on the application of KnECS in other disciplines. We also learned how difficult some barriers are to overcome; these are assessed briefly in Section 3. The appendices contain publications associated with this project as well as other ancillary information.

2.1 Infrastructure and Tools

The Knowledge Environment for Collaborative Science (KnECS) KnECS defines a multilayer architecture as depicted in Figure 2. The primary user environment is a web portal based on the CompreHensive collaborativE Framework (CHEF) [A3] software which enhances standard portal technology with support for teams and community interaction tools such as chat and announcements. KnECS also leverages several middleware components. Of most significance is the Scientific Annotation Middleware (SAM) [R1] software for data and metadata services. SAM 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 is based on Apache Slide [A4] and implements the WebDAV protocol [A5].

As illustrated, KnECS also includes middleware components that implement publish/subscribe messaging, authentication, and authorization. Authorization is managed via JAAS [A6] interfaces to support pluggable security. Mechanisms for integrating Fortran, C, or other computational codes are provided through synchronous and asynchronous web services. Grid services have been demonstrated via the Commodity Grid (CoG) toolkit [A7].

KnECS provides APIs for notification subscriptions, tasks, and other core portal objects such as users and teams. The Data Storage Interface (DSI) provides an easy to use API to the SAM data repository that hides the details of the data access protocol. Additional classes support the construction of structured metadata such as multi-valued container properties, as specified by RDF [A8], and XLinks [A9] for relationships.

KnECS maintains a separation of business logic from portal infrastructure so that non-portal applications that access KnECS APIs can be written. One such application, the Notification Email Daemon (NED), listens for events matching user or team notification subscriptions and sends immediate, daily or weekly digest email messages summarizing the events. This is typically used by teams as a light-weight method of keeping up to date on recent team activities.

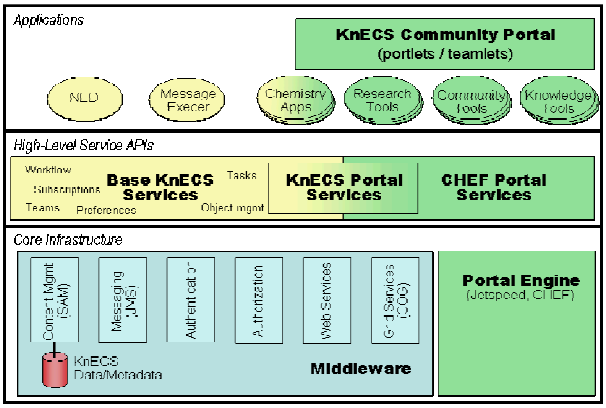

Figure 2. A schematic view of the KnECS multi-tier approach integrating data, applications, and informatics tools.

KnECS includes several knowledge tools, the DataBrowser portlet being the primary example. The DataBrowser supports standard file and directory operations including single file and bulk upload, control over permissions, mouse-over preview of metadata, full metadata viewing and editing; data translation, and visualization. A lightweight file chooser was created from the DataBrowser code base to improve file selection interaction from other portlets. Furthermore, four additional tools are integrated to provide general knowledge management capabilities. A provenance graph tool displays data relationships. A search tool utilizing Lucene [A10] indexing of metadata properties, in conjunction with DASL [A11] queries, allows for data discovery. An annotation tool allows users to annotate data with text, sound, images, equations, or whiteboard drawings. A subscription tool allows users to set up individual and team notifications.

KnECS extends the CHEF team management portlet to automatically create both public and private workspaces on the data server upon team creation. Private workspaces are used to share and perhaps develop data within a community while the public area can be used to publish verified reference data. Support for threaded, archived team email lists has been recently added as email remains a very popular way for teams to communicate.

2.2 Quantum Chemistry

CMCS accomplishments related to the quantum chemistry scale focused on the integration of Ecce [A12] and implementation of the Basis Set Exchange.

2.2.1 Ecce Integration

Ecce provides a suite of graphically rich tools that support setup, execution, analysis and management of quantum chemistry calculations using NWChem [A13], Gaussian [A14], and other codes which compute, from first principles, important properties such as minimum energy and vibrational frequencies. Ecce captures and saves, in their native formats, all the data (inputs, outputs, and setup parameters) as well as metadata to describe each calculation. Given that Ecce already uses WebDAV for its data and metadata management, we elected a lightweight integration where Ecce and SAM were made interoperable at the WebDAV protocol level. Ecce adopted the KnECS provenance strategy for fully describing relationships using XLink. Users can now directly use the KnECS SAM server for a calculation repository or export data to KnECS where portal tools can be used to share, discover, and visualize data or participate in community processing. Together, these are significant advancements towards our goal of multi-scale integration.

2.2.2 Basis Set Exchange

The Basis Set Exchange (BSE) was developed to help computational chemists use and share the basis sets that are used to expand the wave function. It was modeled after the EMSL tools, but leveraging KnECS infrastructure. The BSE is defining XML standards for various types of basis sets, adopting KnECS pedigree standards, and developing XSLT translators. A primary benefit of the KnECS infrastructure is the support for a community of basis set developers who can collectively contribute to and curate the data. Contributing data is a multi-step process and a navigation bar capability was developed to show the process visually as well as to support navigation. Curators can be notified when new data is contributed and access control lists (ACLs) control which community members can curate data. Finally, an XML-based generic logging capability was developed to track usage statistics with plans to develop summary statistics through translators. Statistic summaries can be readily tailored to sponsors, contributors, and users.

2.3 Thermochemical Science

2.3.1 Active Thermochemical Tables (ATcT) Community Service

ATcT represent a synergism of breakthrough developments in the chemistry domain with those in collaborative computer and computational science in association with the CMCS. ATcT implements a Thermochemical Network (TN) concept that explicitly exposes the manifold of inherent interdependencies ignored in traditional approaches. This effort, led by Dr. Branko Ruscic of Argonne National Laboratory, is focused on improving the accuracy, reliability, and internal consistency of thermochemical values that are fundamental to many areas of chemical science and industrial application. This work involves a growing number of collaborations, and an IUPAC community focused on the thermochemistry of radicals.

Many of the advantages of the ATcT approach stem from its use of the Thermochemical Network (TN) that explicitly exposes the manifold of inherent interdependencies among thermochemical properties and the experiments and computations from which they are derived. Thus, ATcT provides quick and painless propagation of new knowledge to update all affected thermochemical values. We have designed and implemented an ATcT community service that can expose the ATcT functionality as part of the CMCS portal. The service is reusing design principles and technologies pioneered by the Java CoG Kit. Crucial collaborative capabilities are derived from the CMCS portal, which makes the ATcT application and associated data broadly available to collaborators and supports

data evaluation and other community processes. The ATcT interface in the CMCS Portal with a visualization of a TN is shown in the figure.

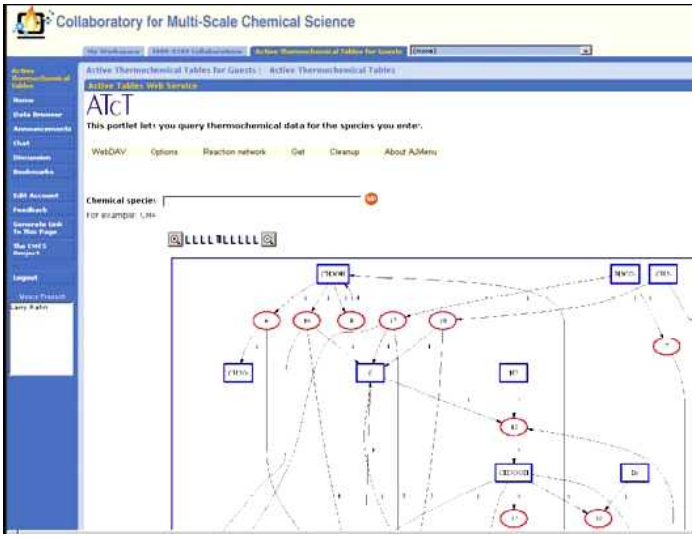

Figure 3. An ATcT thermochemcial network displayed in the CMCS Portal.

An important capability deriving from the systems approach of ATcT allows researchers to discover 'weak links' in the TN. Such 'weak links' point scientists to new experimental or theoretical determinations will significantly improve the TN.

Utilizing the Active Thermochemical Tables (ATcT) approach, new thermochemistry has been developed numerous molecular and some atomic species that have often motivated further work or new collaborations. A direct consequence of new results for the HO_2 radical, for example, was the implication that the accepted reverse kinetic rate constant of the famous Howard reaction ($HO_2 + NO \rightarrow OH + NO_2$) must be wrong by a factor of nearly 2. In order to verify this explicitly, new kinetic measurements were undertaken in collaboration with J. Michael (ANL), which proved exactly the ATcT prediction. Two papers on this subject have just appeared. [R15, R16] The HO_2 development also leveraged from the concomitant ATcT improvements in the NOx thermochemistry (also very relevant both in combustion and atmospheric chemistry), which in turn was made possible by the ATcT discovery of a 'weak link' in the network of values in related NOx compounds. The "weak link" discovery triggered new ATcT-related photoionization measurements of the threshold for formation of N+ from N2, conducted recently at the Advanced Light Source in Berkeley in collaboration with C.-Y. Ng (UC Davis). The direct measurements have been recently published,[R8] and a description of the resulting improvements to NOx thermochemistry is in preparation. The HO2 development is also part of a larger project that is developing definitive thermochemistry of H_NO_M species, including OH. A paper on this topic is also currently in preparation. ATcT also played a crucial role in developing a new electronic structure method (W4, which aims at consistent sub-kJ/mol accuracy) by the group of J. M. L. Martin (Weizmann), where ATcT provided the needed benchmarks and feedback as to which corrections are necessary in the method's formulation.[R17] Also, the first paper in the series of papers addressing the ATcT development on the topic of thermochemistry of carbon, which was another ATcT "weak link" discovery, addressed by a collaborative effort with A. G. Csaszar (Eotvos, Budapest) and J. F. Stanton (UT Austin), has been recently submitted for publication.[R18] In addition, ATcT has been used to update - to the extent possible - the largest thermochemical database of polynomials currently in existence (A. Bucat, Technion), resulting in a joint report.[O9]

2.3.2 Group Additivity Thermodynamic Properties (GATP)

Chemists often need thermodynamic properties for chemical species when no experimental or theoretical determinations exist. GATP [P4] is a relatively new effort within CMCS that provides these estimates, particularly for larger molecules where more accurate approaches are not available. GATP uses an internal additivity database and the connectivity of the species to compute thermodynamic properties. A CMCS portlet provides an interface to the java-based GATP code and provides access to computed properties, via XSLT translations, in a tabular or polynomial fit form commonly used by legacy combustion, atmospheric and chemical-engineering codes. Ultimately, the GATP additivity

database will be available for community contribution and curation processes. Easy access to GATP, shared GATP results, and its additivity database allows scientists to address problems which involve chemical species with unknown thermochemical properties.

2.4 Kinetics and Kinetic Mechanisms

Chemical kinetics mechanisms are used to simulate the chemistry in combustion, the atmosphere, and chemical engineering processes. Scientists and engineers need to discover chemical kinetic mechanisms for a particular fuel or reactant, understand the range of their applicability, and access the data in a format suitable for their application. Accomplishing this requires the ability to search, view provenance and examine the sources for the detailed rate constants and thermodynamic parameters; to record comments on the performance and applicability of the mechanisms under various use cases; and to access conversion software that translate chemical kinetic mechanisms among various combustion and atmospheric modeling formats. The Large Mechanisms group has used the CMCS portal to develop schemas and translators to facilitate the accurate use of currently available chemical mechanisms and is providing documentation and validation data that will speed the development of new or improved mechanisms. A related capability is the Range Identification and Optimization Tool described below. Also included below is a summary of accomplishments from the PrIMe development group.

2.4.1 Range Identification and Optimization Tool (RIOT)

For many years researchers have tried to predict how fuel changes will affect the performance of new engines on the drawing boards, but the many complicated details of combustion chemistry have made this difficult-to-impossible. An automated method has been developed for extracting only the essential details from a complicated fuel chemistry mechanism, so they can be incorporated into efficient computer programs for the design of novel high-efficiency engines. The Range Identification and Optimization Tool (RIOT) has been developed to provide this reduction [R11]. It has been made available in the CMCS portal to facilitate setup, execution, and results analysis for scientists collaborating from other institutions and disciplines [O7]. The new mechanism-reduction method, developed by the CMCS, allows the user to control the error vs. speed-up trade-off, and in certain steady flame simulations the user can even rigorously bound the error that will be introduced by neglecting the minor chemistry details. This sort of "coarse-graining with error control" makes it feasible to accurately predict the behavior of macroscopic devices based on first-principles models of the microscopic molecules reacting inside the engine, i.e. reliable multiscale simulation. The CMCS portlet interface assists the user with specifying inputs to the RIOT code, generates the input file, ships the request to a web service interface to the code, and uploads and presents the results when the reduction completes. Preliminary use of the RIOT/CMCS web service to build a reduced chemical reaction mechanism to model iso-octane HCCI yielded greater than a 20-fold improvement in computational efficiency while preserving the accuracy of the HCCI predictions.

2.4.2 Process Informatics Model (PrIMe):

PrIMe—Process Informatics Model (http://primekinetics.org)—is an international initiative fostering a new approach for developing predictive models of chemical reaction systems that is based on the scientific collaboratory paradigm and takes full advantage of existing and developing cyber infrastructure. The primary goals of PrIMe are collecting and storing data, validating the data and quantifying uncertainties, and assembling the data into predictive models with quantified uncertainties to meet specific user requirements. The principal components of PrIMe include: a data Depository, which is a repository of data provided by the community, a data Library for storage of evaluated data, and a set of computer-based tools to process data and to assemble data into predictive models. Two guiding

principles of PrIMe are: open membership—any qualified individual can register to participate in the project; and open source—all submitted data, tools and models will be in the public domain. The PrIMe Initiative was officially launched on April 21-22, 2006. At the time of this writing, there are 60 signed-in members.

Having members of PrIMe on the CMCS team provided numerous benefits to the CMCS project. PrIMe scientists participated in the development of the CMCS User Agreements, provided the initial CMCS demonstration projects of XML documents, participated in the design of various features of the CMCS portal, and developed the initial realistic WebDAV data collection for testing the CMCS software. Being in the vanguard of the project, PrIMe has developed XML standards for chemical elements, chemical species, chemical reactions, and experimental records, led in the development of WebDAV search capabilities, and introduced the CMCS team to the concept of data submission forms.

The PrIMe Warehouse is hosted by a Dell PowerEdge 800 computer, with 2.8 GHz Intel P4 CPU and 1 GB memory and 36 GB Ultra 320 SCSI hard disk rotating at 15000 rpm. The computer is running Microsoft Windows 2003 Server Standard Edition operating system. The web access to the PrIMe warehouse is facilitated using operating system's built-in Internet Information Services 6 web server software with enabled WebDAV extensions compliant with RFC2518 Standard. To facilitate search, files are indexed upon depositing into the Warehouse, or when re-deposited after editing. Indexing is automatic, performed by operating system's built-in Indexing Services software. The search for XML files with criteria related to a particular node (for example, for all the species composed of four carbon atoms and less than seven hydrogen atoms), standard full-text indexing by Indexing Services is extended with an XML-specialized plug-in filter QLXFilter by QuiLogic Inc. Filtering criteria are designed by the PrIMe development team members, to make most of the XML nodes available for node-specific search requests. The search queries must adhere to MSSQL grammar. The XML interface to the search engine is accessible through XML message compliant with WebDAV DASL standard (RFC proposal draft-reddy-dasl-protocol-04). Several client-side interfaces are commercially available for such a WebDAV search. The PrIMe development team created two custom interfaces for the PrIMe community, ASP-based form web page (available via publicly-accessed http://primekinetics.org) and Matlab-based graphical interface. For disaster-recovery and historical purposes, the warehouse contents are regularly recorded on CD and DVD discs, using Roxio Creator Classic software. The backup CD file system is ISO. The data are files are stored uncompressed, without anything specific to backup software used.

The current collection is mostly based on data imported from NIST Kinetics Database, GRI-Mech 3.0 project data, and the experimental data from the Stanford University (the latter work co-funded by other sources, GCEP and NSF Chemistry Division). The Warehouse currently contains 87200 files in 18500 folders, consuming 342 MB of disk space.

With the current hardware and data collection, the performance is as follows. Newly deposited or edited XML files are indexed and available for search within 5 seconds of depositing. Searching the data, apparently regardless of search query complexity, generates response with a list of files satisfying search criteria within typically 20 ms. Realistic response time to search query requested by a remote computer, thus depends mostly on time for transfer of search results through the network between the PrIMe server and the remote user's computer. The time for WebDAV retrieval of a file from the Warehouse depends exclusively on the network traffic; retrieval time without the network overhead was too short for reliable measurement, typically shorter than 0.5 milliseconds. The disaster recovery or the collection rollback to one of the archived states would involve copying the files from the backup medium (typically one hour for the current hardware and the collection size). Such a restore requires complete re-build of index catalog, which is automatic, and currently takes between two and three hours.

The work will continue for another 12 months as a no-cost extension of the present grant to the University of California at Berkeley. This work will be devoted to completing the data management

software and associate manuals. The data management software developed under this grant will open-sourced and made available at souceforge.net at the time of completion. A complete final report will be prepared and submitted at the end of the extension period.

A team of PrIMe members (UC Berkeley, Stanford, MIT) received an NSF Chemistry Division award on Cyberinfrastructure. PrIMe became affiliated with the Center for Information Technology Research in the Interest of Society (CITRIS; http://citris-uc.org), and has already benefited from the support provided on web-site and hierarchical-group portal technologies.

2.5 Reacting Flow Simulation and Experiment

2.5.1 Feature Tracking and FD Tools

A common and useful data analysis method, especially for time dependent simulation results, is to identify regions of interest and study their behavior over time. The process of finding coherent structures that persist over time, such as flames in combustion simulations and hurricanes in climate simulations, is called Feature Identification and Tracking. We have developed the Feature Detection & Tracking Library (FDTools), a serial framework to support the easy assembly of an extensible set of feature identification and tracking algorithms into a feature analysis pipeline. We used FastBit searching technology from the SciDAC Scientific Data Management Center to rapidly identify points satisfying user-specified conditions. We then group the points into regions, and compute the overlap among regions among consecutive time steps to identify the evolution of regions. One of the goals was to use features extracted from large-scale DNS (Direct Numerical Simulation) combustion data as searchable metadata at the CMCS Portal. FDTools was used to help discover new scientific knowledge in the combustion domain. In the study of autoignition, preignition kernels are generally identified as areas of high radical concentration. All autoignition sites start in preignition kernels, but not all preignition kernels ignite. FDTools was used to find, track, and examine the preignition kernels, and revealed that the kernels did not all ignite at the same time. By examining statistics of the features, the scientists were able determine a correlation between the dissipation rate and the fate of a preignition kernel.

During the course of implementation, it became obvious that a completely separate tool is not a good way to proceed. Any complicated analysis requires interpolation, derivatives and related functions and variables that are developed within the simulation code itself. Thus, while the concept was proven valuable, further development was deferred and ultimately submitted as part of proposal to the recent SciDAC 2 call.

2.5.2 Metadata Library for Premixed Turbulent Flame Experiments

The Premixed Turbulent Flame Working Group (http://purl.oclc.org/NET/preturb/) teamed up with CMCS to document and share information about their available data sets. These combustion scientists collect large data sets on measured velocities, temperatures and species concentrations in premixed, turbulent, laboratory-scale flames. Scientists need to store their important data, retrieve it, and share it with others. They also need to keep track of key information about the experimental data and computational results. The Premixed Turbulent Flame Working Group meets every two years to discuss their needs at the International Workshop on Premixed Turbulent Flames.

The Group meets in August of 2006 in Mainz, Germany to learn about and begin using new capabilities developed in collaboration with the CMCS team that enable the working-group scientists to upload MS Excel® data sheets that document their large datasets. During the upload, the CMCS portlet automatically converts the Excel® datasheet file into an XML file and an XSLT translator is invoked to extract metadata from the XML. This process is all done in the background. The critical metadata is then available which enables implementation of advanced search capabilities. This metadata includes

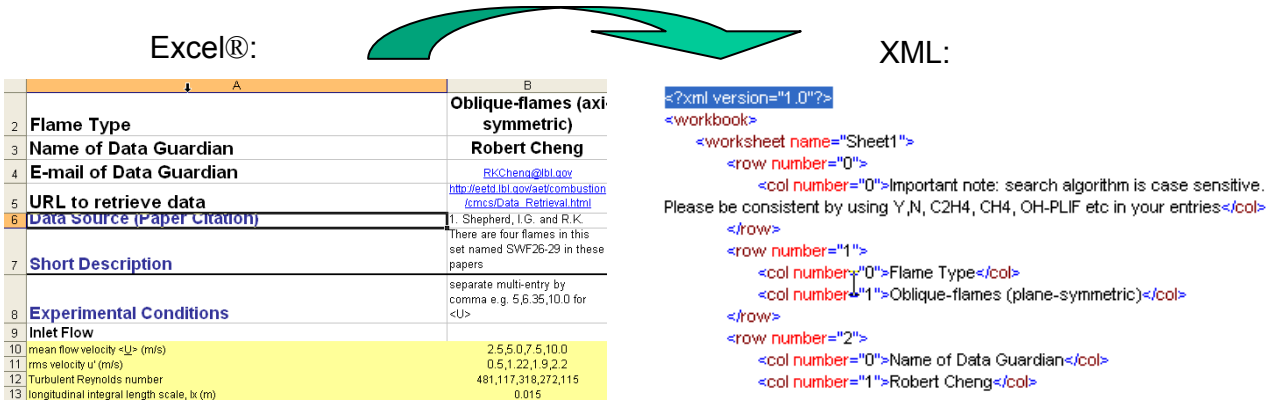

Figure 4. Illustration of the automatic extraction of metadata from and Excel® file.

experimental flow parameters such as equivalence ratio, inlet temperature, Karlovitz number, and fuel type. An advanced search form is made available that can easily be modified by user groups to meet their needs. Together, these capabilities and data enable scientists and engineers around the world to search on this metadata with a tailored search form in the portlet, allowing easy discovery and access to experimental datasheets of interest. The leaders of the Working Group anticipate that these capabilities will help resolve longstanding barriers to data sharing in this research community.

Figure 5. Premixed Turbulent group customizable search form:

2.5.3 Publication of Experimental Combustion Data Sets

A new approach for publication of combustion data that offers an inter-operable format available from an easy to access, automatically updated web page has been prototyped for two international combustion science teams in collaboration with the CMCS. These teams, an international consortium for Synchrotron Photoionization Mass Spectrometry (PIMS) for investigation of combustion chemistry [A15], and the International Workshop on the Measurement and Computation of Turbulent Nonpremixed Flames (http://www.ca.sandia.gov/TNF/) are meeting in conjunction with the 31st International Symposium on Combustion during August, 2006 to present and discuss the new approach for data publication in their communities.

The publicly accessible publication web page features the ability to not only view and download the data in a translation back to the current text based, white-space delimited format, but also to view it in a tabular web view and/or in graphical plots. The web archive provides the option of browsing the data in a hierarchical folder structure as well as the ability to search and display the data based on key properties or metadata. Also, it is possible to simultaneously provide public access to multiple folders of data that are curated by different institutions. The new capabilities are built upon a data-sharing infrastructure developed by the CMCS and are designed to meet the specific requirements for web-based publication by these communities. The implementation involves updating current data using a java application designed to convert current text files into a semantic, self-descriptive XML format which contains added metadata. When such a formatted XML file is deposited into the archive, its contained

metadata is automatically read and stored separately to enable rapid searches. From the archive, the XML file can then be viewed or downloaded in a variety of different forms, each associated with a (XSLT) translator offering a different viewable or downloadable format of the data. For example, one might translate the file into a graphical scatterplot view while another offers a downloadable text file in a newly defined format.

The demonstration of this prototype enables, and challenges, these scientific communities to invest more in the curation and publication their data, and to work on new agreements for broader participation in data publication. The vision is for an archive where experimental, computation, and modeling data all exist in a form that is easily searched and displayed by a computer while being available in formats that are easily read by other software applications.

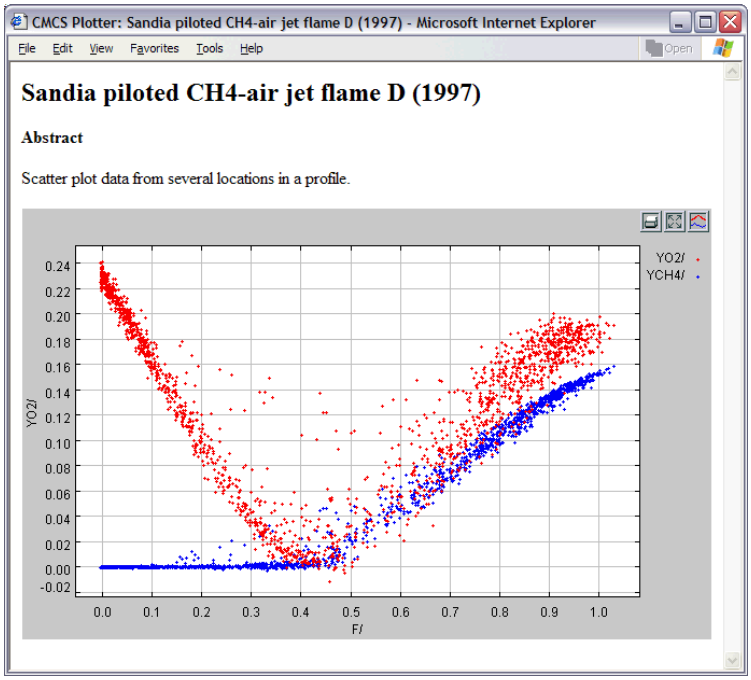

Figure 6. Example Scatterplot of Sandia Piloted CH_4-Air Jet Flame D.

2.6 KnECS Implementations In Other Science Communities

CMCS is a unique example of how cyberinfrastructure can actively support research processes rather than just being a means to provide access to finished, 'text-book' information and services. Similar requirements for such a 'cyberenvironment' that provides end-to-end support for the research lifecycle exist in many fields. A current example of the implementation of KnECS for a different discipline is a project in which a group of biomedical scientists leading the development of new technique termed 'MS3D' are collaborating with researchers in the CMCS to build the Collaboratory for MS3D (C-MS3D) [A2]. This is a multi-institution project with partners funded by the National Institutes of Health and the National Science Foundation.

This collaboration has greatly facilitated the new project, enabling deployment of a C-MS3D production portal and a development portal within the first 6 months. MS3D uses chemical crosslinking and mass spectrometry to probe the structure and dynamics of proteins, RNA, and macro-molecular complexes. The objective of this 5-year project is to broadly enable an emerging MS3D collaborative community as it develops new tools, analysis approaches, and data schema to integrate constraint data from chemical crosslinking and mass

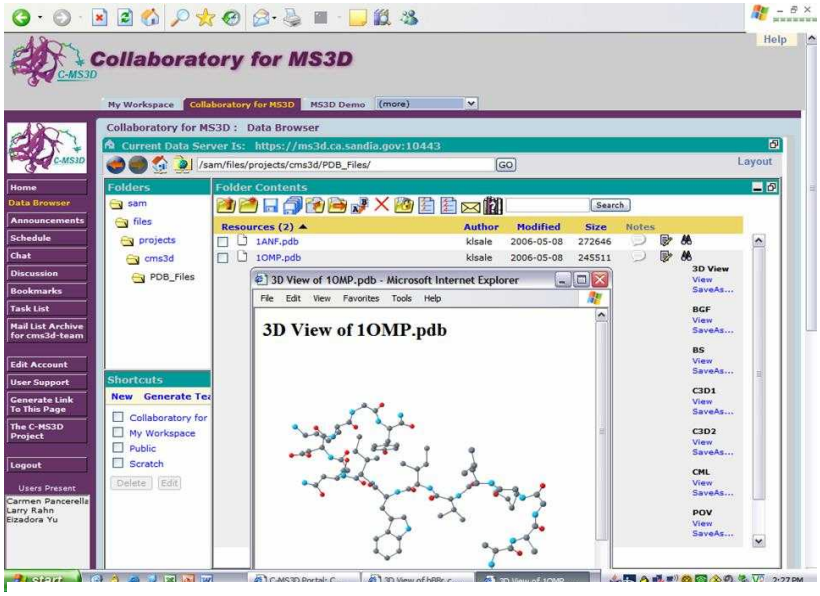

Figure 7. A screen shot of the C-MS3D Portal illustrating one of the many translations available for a Protein Data Base file.

spectrometry with other information (including that obtained from computation, NMR, EPR, FRET, X-Ray crystallography) to determine otherwise inaccessible macro-molecular structures.

The more general capabilities developed by CMCS have been made available to others as the Knowledge Environment for Collaborative Science (KnECS). KnECS is an open source portal and data management environment that can be customized with discipline-specific tools, data translators, and data viewers. The KnECS deployment process was revised by the CMCS team to support configuration of the portal for other projects such as C-MS3D. The C-MS3D team branded the portal to reflect the C-MS3D project identity. Using the development library for creating asynchronous web services in KnECS, and the C-MS3D team integrated existing MS3D codes into the portal, making the latest tools accessible across the group. They are using some translators, metadata extractors, and visualization tools developed by CMCS while developing others that are more specific to biology.

3.0 Overall Assessment

3.1 Challenges

The CMCS project team has identified critical challenges facing developers of science applications using KnECS. Some are significant, and must be addressed if portal-based knowledge environments are to become ubiquitous in science. The technology must offer modularity that empowers application scientists, the flexibility and adaptability to integrate rapidly changing technologies, applications, and other data stores. We discuss these issues by theme.

3.1.1 Data Management Challenges

Performance: As data repositories grow in size, and multiple data repositories are used, data federation will be necessary. Performance issues associated with very large data sets must be addressed. Large data sets are also an issue for real time translation and XML description. The latter is addressed by emerging standards such as the Data Format Definition Language [A16] which has been recently implemented with open source from the Defuddle project [A17].

Data Curation: Several applications require support for curation of reference data sets. However the model(s) for what curation means is not yet well defined. Individual communities will define curation mechanisms that meet their needs. We expect some common elements to span many data sets/communities. These include: facilitating data entry (e.g. using schema driven forms), annotating data with curation pedigree and status information and curator notes (though the schemas will vary), versioning of data sets, setting up automated validation steps (though the processes themselves will vary), and manual editing/correcting of data which could use generated forms for editing XML content. Curated data shares a common trait of an associated publication reference. Standards, tools and on-line access to a reference repository may be beneficial.

This area is also subject to cultural and policy issues, in that the time and skills required to curate scientific data sets are not yet highly valued in much of the scientific culture. For example, many sponsors of science do not value curated data enough provide a way for it to be included as a metric in scientific performance. Of course, this is related to one of the central issue the CMCS team has tackled, the fact that there are few ways to publish curated data where it can be referenced and reliably discovered, explored, and downloaded.

3.1.2 Integration Challenges

Legacy Integration: As discussed in Section 3, many communities require integration of legacy codes. There are many possible models for integration including: direct process invocation, simple synchronous or asynchronous web services with or without load balancing or QOS support, grid services, and ultimately multiple processes connected via workflows. The application with which

CMCS worked use several models: synchronous web services, asynchronous web service jobs and asynchronous grid tasks via the COG toolkit. However, more work is needed to create a simple API that can exploit different underlying implementations.

Portlet Integration: In general, developing complex applications in a portal environment is difficult. JSR168, the Java portlet API [A18], addresses a standard portlet specification; however, it does not make portlet development easier and does not address collaboration-style portlets, thus resulting in non-standard extensions.

Support for Different Granularities of Application Integration: From the diverse application areas discussed above in Section 2, it is clear that different styles of integration are required, so we have implemented several models. These range from simple file sharing, to loose integration of external tools, to portlet wrappers for computational codes, to applets or java web start enabled applications. The model proposed by Web Services for Remote Portlets (WSRP) [A19] is a desirable alternative that addresses important intellectual property issues but robust implementations are needed. The choice of which to use depends on many factors including ownership, whether the software exists yet or not, the resources required, and the availability of skilled portlet developers. A community portal toolkit should support all of these models.

3.1.3 Reuse Challenges

Customization: When a new application is integrated into a portal, data input and conversion tools typically need to be customized to support this application. The development and deployment of data input forms and data converters should be easy. Furthermore, as new groups or communities form, each project and/or institution wants a unique workspace, so portal branding for projects and/or organizations must be easy and provide the appropriate level of identity for the community. While individual groups can brand their CMCS Portal view with a graphic, the reuse of the KnECS infrastructure by other disciplines (e.g., the biomedical community discussed in Sec/ 2.6) requires a careful identification and separation of all elements in the code base that contribute to branding. This was undertaken by the CMCS team as a part of our effort to make KnECS more usable (see Sec. 3.1.5 below.).

Even with customization, it is difficult for one infrastructure implementation to meet enough needs of some groups to be satisfactory. The reasons for this appear to be both technical (*e.g.*, tradeoffs of features vs. performance) and sometimes cultural (*e.g.*, disagreement on common features, issues over who is in control or hosting infrastructure, etc.). The technical issues support arguments for modularity, for the adoption of standards for interoperation of modules, and for the use of open source code bases. The cultural issues are discussed more in the following section.

3.1.4 Usage Challenges

Social and Intellectual Property Issues For Scientists: Though science collaboratories have been in existence for over fifteen years, there are still social and science issues impeding adoption. Preserving the ownership and controlling the use of data and tools is necessary in a scientific community. Once KnECS is adopted and deployed to a science community, that community must also adopt guidelines for data publication and address intellectual property issues surrounding contributed data and software tools. Many of threes issues will be more easily resolved by adoption of policies and practices of example communities who have successfully established and use a common cyberenvironment. This seems like a significant impact that pilot efforts such as CMCS team can offer the scientific community.

Need for an End-to-End Solution: Until the infrastructure, data, and applications that support collaborative science fits as an integral part of scientist's day-to-day work and a part of a 'full solution' for research, collaboration tools will always be in the background and used only occasionally. This was

evidenced, for example, in our experience presenting computational tools via web portlets. In this case issues such as unique molecule identifiers, definition of the uncertainty in the results, documentation of the pedigree of the results, and the capture all the data for inputs and outputs in user friendly ways were found to be barriers to adoption across disciplinary boundaries. It seems that this requires considerable flexibility (*e.g.,* modularity, modes of access, etc.) and more standards to facilitate interoperability.

3.1.5 Open Source Challenges

Active Open Source Development Community: While the KnECS infrastructure has been licensed as open source, much work remains to organize and document it so that other communities can easily deploy and extend it to create their own knowledge grids. KnECS leverages many third party capabilities, such as SAM and CHEF, creating a challenging maintenance environment. An active community with long-term sustainability is required to convince scientists to build upon these capabilities. The CMCS project team is currently in the final stages of reorganizing the open source code base for release in a more usable form on SourceForge.net[A1].

3.1.6 Collaboration Challenges

Security and Group Management: As collaborative groups and communities form, there is a need for hierarchical groups, with flexible approaches to access control and privileges. Also, different communities have varying requirements on how teams should be formed and managed. The security access controls are intimately related to the group or file structure. From a scientific collaboration point of view, however, this mapping does not work well. Most scientists participate and share data in more than one collaborative group. Thus, different data may be contributed to more than one collaboration, each having a different control hierarchy. Fully enabling such mixed-mode collaborations requires a much more general way to manage security, membership, and hierarchy for overlapping collaborative groups.

3.2 Realization of the CMCS Vision

Many of the elements of the original CMCS vision were largely met. For example, pedigree browsing, data interoperability via XML technologies and automatic translations among schema, integration of collaboration tools with data management and application sharing capabilities, and the stimulation of new approaches for science. Other aspects of our vision are clearly still more futuristic than accomplished; such as the formation of active, sustained interdisciplinary collaborations across all of the scales in combustion (see Fig. 1) or the use of CMCS as a route to 'publish' community curated data. Note that we made progress against these goals, but it will clearly take efforts that are sustained beyond the typical project life time to change the culture and practice of science. Perhaps it is just too early to make such judgments since there are continuing activities from related projects, and the fact that Sandia's Combustion Research Facility will continue to operate the CMCS production servers. It is clear, however, that continued support for research, development, and pilot activities is still a critical need if cyberenviornments are to developed and adopted for the benefit of DOE's science mission.

4.0 Acknowledgements

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. Livermore National

Laboratory is operated by University of California for the U.S. Department of Energy under contract No. W-7405-Eng-48, Argonne National Laboratory is operated by the University of Chicago for the U.S. Department of Energy under Contract No. W-31-109-ENG-38. The development of ATcT was supported by U.S. Department of Energy, Division of Chemical Sciences, Geosciences and Biosciences of the Office of Basic Energy Sciences (chemistry aspects), and by the Mathematical, Information, and Computational Science Division of the Office of Advanced Scientific Computing Research (computer science aspects).

The authors thank program manager Mary Anne Scott for her vision, encouragement, and budgetary support of this work. The authors wish to acknowledge members of the CompreHensive collaborativE Framework (CHEF) and Scientific Annotation Middleware (SAM) projects, members of the IUPAC Task Group on Thermochemistry of Radicals (IUPAC Project 2003-024-1-100), the members of the Java CoG Kit project, and all the past and present researchers involved in CMCS science applications and development.

5.0 Publications, Presentations, and Ancillary References

5.1 Refereed Publications

[R1] Myers JD, Chappell A, Elder M, Geist A, Schwidder J. 2003. Re-integrating the research record. *IEEE Computing in Science and Engineering* 2003; **5**(3):44-50. Also available at http://www.scidac.org/SAM/.

[R2] Binita Bhattacharjee, Douglas A. Schwer, Paul I. Barton, & William H. Green, Jr., "Optimally-Reduced Kinetic Models: Reaction Elimination in Large-Scale Kinetic Mechanisms", Combustion & Flame 135, 191-208 (2003).

[R3] 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: p. 25-38.

[R4] Myers, J. D., Allison T. C., Bittner S, Didier B, Frenklach M, Green WH Jr., Ho Y-L, Hewson J, Koegler W, Lansing C, Leahy D, Lee M, McCoy R, Minkoff M, Nijsure S, von Laszewski G, Montoya D, Oluwole L, Pancerella C, Pinzon R, Pitz W, Rahn LA, Ruscic B, Schuchardt K, Stephan E, Wagner A, Windus T, Yang C., "A Collaborative Informatics Infrastructure for Multi-scale Science," Proceedings of CLADE 04, Honolulu, Hawaii, 2004, p.24. http://csdl.computer.org/comp/proceedings/clade/2004/2115/00/2115toc.htm.

[R5] Ruscic B, Pinzon RE, Morton ML, von Laszewski G, Bittner S, Nijsure SG, Amin KA, Minkoff M, Wagner AF. 2004. Introduction to active thermochemical tables: Several "key" enthalpies of formation revisited. *Journal of Physical Chemistry A* 2004; **108**:9979-9997.

[R6] Ruscic B, Pinzon RE, Von Laszewski G, Kodeboyina D, Burcat A, Leahy D, Montoya D, Wagner AF. 2005. Active thermochemical tables: Thermochemistry for the 21st Century. *Journal of Physics: Conference Series.* 2005; **16**:561-570.

[R7] Ruscic B, Boggs JE, Burcat A, Csaszar AG, Demaison J, Janoschek R, Martin JML, Morton ML, Rossi MJ, Stanton JF, Szalay PG, Westmoreland PR, Zabel F, Berces T. 2005. IUPAC critical evaluation of thermochemical properties of selected radicals: Part I. *Journal of Physical and Chemical Reference Data* 2005; **34**(2):573-656.

[R8] X. Tang, Y. Hou, C. Y. Ng, and B. Ruscic, Pulsed Field-Ionization Photoelectron-Photoion Coincidence Study of the Process $N_2 + h\nu \rightarrow N^+ + N + e^-$: Bond Dissociation Energies of N_2 and N_2^+, *J. Chem. Phys.* **123**, 074330/1-7 (2005).

[R9] Karen Schuchardt, Carmen Pancerella, Larry A. Rahn, Brett Didier, Deepti Kodeboyina, David Leahy, James D. Myers, Oluwayemisi O. Oluwole, William Pitz, Branko Ruscic, Jing Song, Gregor von Laszewski, Christine Yang, "Portal-based Knowledge Environment for Collaborative

Science," accepted, *Concurrency and Computation: Practice and Experience*, Wiley, December 2005.
[R10] Binita Bhattacharjee, Panayiotis Lemonidis, William H. Green and Paul I. Barton "Global Solution of Semi-infinite Programs", *Mathematical Programming (Series B)* 103(2), 283-307 (2005).
[R11] O.O. Oluwole, B. Bhattacharjee, J.E. Tolsma, P.I. Barton, and W.H. Green, "Rigorous Valid Ranges for Optimally-Reduced Kinetic Models", *Combustion and Flame* (submitted Jan. 2005).
[R12] Binita Bhattacharjee, William H. Green, Jr., and Paul I. Barton, "Interval Methods for Semi-Infinite Programs", Computational Optimization and Applications (2004) accepted.
[R13] Oluwayemisi O. Oluwole, Binita Bhattacharjee, John E. Tolsma, Paul I. Barton, William H. Green, "Rigorous Valid Ranges for Optimally-Reduced Kinetic Models", *Combustion and Flame* **146**, 348-365 (2006).
[R14] Oluwayemisi O. Oluwole, Paul I. Barton, and William H. Green, "Obtaining Accurate Solutions using Reduced Chemical Kinetic Models: A new Model Reduction method for models rigorously validated over ranges", *Combustion Theory and Modelling* (submitted).
[R15] B. Ruscic, R. E. Pinzon, M. L. Morton, N. K. Srinivasan, M.-C. Su, J. W. Sutherland, and J. V. Michael, Active Thermochemical Tables: Accurate Enthalpy of Formation of Hydroperoxyl Radical, HO_2, *J. Phys. Chem.* **110**, 6592-6601 (2006).
[R16] N. K. Srinivasan, M.-C. Su, J. W. Sutherland, J. V. Michael, and B. Ruscic, Reflected Shock Tube Studies of High-Temperature Rate Constants for $OH + NO_2 \rightarrow HO_2 + NO$ and $OH + HO_2 \rightarrow H_2O + O_2$, *J. Phys. Chem.* **110**, 6602-6607 (2006).
[R17] A. Karton, E. Rabinovich, J. M. L. Martin, and B. Ruscic, W4 theory for computational thermochemistry: in pursuit of confident sub-kJ/mol predictions, J. Chem. Phys., submitted.
[R18] G. Tasi, R. Izsak, G. Matisz, A. G. Csaszar, M. Kallay, B. Ruscic, and J. F. Stanton, The Origin of Systematic Error in the Standard Enthalpies of Formation of Hydrocarbons Computed via Atomization Schemes, *Phys. Chem. Chem. Phys.* **8**, xxxx-xxxx (2006) *accepted.*

5.2 Other Publications

[O1] Carmen Pancerella, James D. Myers, Thomas C. Allison, Kaizar Amin, Sandra Bittner, Brett Didier, Michael Frenklach, William H. Green, Jr., Yen-Ling Ho, John Hewson, Wendy Koegler, Carina Lansing, David Leahy, Michael Lee, Renata McCoy, Michael Minkoff, Sandeep Nijsure, Gregor von Laszewski, David Montoya, Reinhardt Pinzon, William Pitz, Larry Rahn, Branko Ruscic, Karen Schuchardt, Eric Stephan, Al Wagner, Baoshan Wang, Theresa Windus, Lili Xu, and Christine Yang, *Metadata in the Collaboratory for Multi-Scale Chemical Science,* Proceedings of the 2003 Dublin Core Conference (DC-2003), September 28 - October 2, 2003, Seattle, Washington, pp. 121-130.
[O2] James D. Myers, Carmen Pancerella, Carina Lansing, Karen L. Schuchardt, and Brett 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, October 20, 2003, Sanibel Island, Florida.
[O3] Pisi Lu, Binita Bhattacharjee, Paul I. Barton and William H. Green, "Reduced Models for Adaptive Chemistry Simulation of Reacting Flows", Computational Fluid and Solid Mechanics: Proceedings of the Second MIT Conference on Computational Fluid and Solid Mechanics, ed. by K.J. Bathe (Elsevier, 2003).
[O4] Z. M. Djurisic, D. Amusin, T. Bereknyei, T. C. Allison, M. Frenklach, "Reporting of experimental data for development and validation of chemical kinetic models," Fall Meeting of the Western States Section of the Combustion Institute, Stanford, CA, October 17-18, 2005.

[O5] Schuchardt, K., O. Oluwole, W. Pitz, L.A. Rahn, J. William H. Green, D. Leahy, C. Pancerella, M. Sjöberg, and J. Dec, "New Approaches for Collaborative Sharing of Chemical Model Data and Analysis Tools," 2005 Joint Meeting of the US Sections of the Combustion Institute. 2005.

[O6] O.O. Oluwole and W.H. Green, "Rigorous Error Control in Reacting Flow Simulations Using Reduced Chemistry Models", *Computational Fluid and Solid Mechanics: Proceedings of the Third MIT Conference on Computational Fluid and Solid Mechanics*, ed. by K.J. Bathe (Elsevier, 2005, in press).

[O7] Schuchardt K, Oluwole O, Pitz W, Rahn LA, William J, Green H, Leahy D, Pancerella C, Sjvberg M, Dec J. 2005. Development of the RIOT Web Service and Information Technologies to enable mechanism reduction for HCCI simulations. *Journal of Physics: Conference Series* 2005; **16**:107-112.

[O8] Ruscic B. 2005. "Active Thermochemical Tables" in: *2005 Yearbook of Science and Technology*, an annual update to the McGraw-Hill Encyclopedia of Science and Technology, McGraw-Hill, New York, 2004 pp.

[O9] A. Burcat and B. Ruscic, Third Millennium Ideal Gas and Condensed Phase Thermochemical Database for Combustion with Updates from Active Thermochemical Tables, Joint Report ANL-05/20, Argonne National Laboratory, Argonne, IL, USA, and TAE 960, Technion – Israel Institute of Technology, Haifa, Israel (2005).

5.3 Conference Presentations

[P1] L. Rahn, C. Yang, C. Pancerella, J. Hewson, W. Koegler, D. Leahy, M. Lee, J. Nichols, B. Didier, T. Windus, J. Myers, K. Schuchardt, E. Stephan, A. Wagner, B. Ruscic, M. Minkoff, L. Liming, G. von Laszewski, S. Bitner, B. Moran, W. Pitz, D. Montoya, T. Allison, W. Green, M. Frenklach,"Collaboratory for Multi-scale Chemical Science," Poster at the 29th International Symposium on Combustion, Sapporo, Japan, July 21-26, 2002.

[P2] von Laszewski, G., Ruscic, B., Wagstrom, P., et. al., "A Grid Service Based Active Thermochemical Table Framework" in Third International Workshop on Grid Computing, Lecture Notes in Computer Science, Baltimore, MD, 18 November 2002.

[P3] B. Ruscic, R.E. Pinzon, M.L. Morton, B. Wang, A.F. Wagner, G. von Laszewski, S.G. Nijsure, K.A. Amin, S.J. Bittner, and M. Minkoff. "Further Refinements of the Bond Dissociation Energy in Water and Hydroxyl Radical Using the Active Thermochemical Tables Approach", Proceedings of the 58th International Symposium on Molecular Spectroscopy. 2003. Columbus, Ohio.

[P4] Song Jing, Sumathi R, Yu Joanna, Green William H Jr. 2004. Next generation model construction software and new approaches to estimating rates and thermochemistry for combustion. *The 228th ACS National Meeting,* Philadelphia, Pennsylvania.

[P5] James D. Myers, Larry Rahn, David Leahy, Carmen M. Pancerella, Gregor von Laszewski, Branko Ruscic, and William H. Green, *Adaptive informatics infrastructure for multi-scale chemical science Division of Chemical Information*, 227th American Chemical Society (ACM) National Meeting, March 28-April 1, 2004, Anaheim, Calif.

[P6] L. Rahn, T.C. Allison, S. Bittner, B. Didier, M. Frenklach, W. H. Green, Jr., D. Hale, M.F. Hategan-Marandiuc, C. Lansing, G. Von Laszewski, D. Leahy, J.D. Myer, M. Minkoff, D. Montoya, L. Oluwole, C. Pancerella, R. Pinzon, W. Pitz, J. Riese, B. Ruscic, K. Schuchardt, A.F. Wagner, T. Windus, C. Yang, and G. Young, "A Web Portal for Multi-scale Chemcial Science Data and Applications," Poster at the 7th International Workshop on the Measurement and Computation of Turbulent Nonpremixed Flames, Chicago, July 22-24, 2004.

[P7] W.H. Green, O. Oluwole, and B. Bhattacharjee, "Predicting reacting flows with complex chemistry: Rigorous error control in adaptive chemistry calculations." American Chemical Society National Meeting, Philadelphia, August 2004.

[P8] Myers, J.D., Allison, T.C., Bittner, S., Didier, B., Frenklach, M., Green, W.H. Jr., Ho, Y.-L., Hewson, J., Koegler, W., Lansing, C., Leahy, D., Lee, M., McCoy, R., Minkoff, M., Nijsure, S., von Laszewski, G., Montoya, D., Oluwole, L., Pancerella, C., Pinzon, R., Pitz, W., Rahn, L.A., Ruscic, B., Schuchardt, K., Stephan, E., Wagner, A., Windus, T., Yang, C., *"A Collaborative Informatics Infrastructure for Multi-scale Science",* Sept. 2004, Cluster Computing.

[P9] Karen Schuchardt, Oluwayemisi Oluwole, William Pitz, Larry A. Rahn, William H. Green, Jr., David Leahy, Carmen Pancerella, Magnus Sjöberg, John Dec, "New Approaches for Collaborative Sharing of Chemical Model Data and Analysis Tools", *2005 Joint Meeting of the U.S. Sections of the Combustion Institute*, March 2005, Philadelphia, Pennsylvania.

[P10] W.H. Green, "Predicting Hydrocarbon Chemistry: A Path Forward", Olah Award Symposium, American Chemical Society, San Diego, March 2005.

[P11] W.H. Green, "New Data Model and Advanced Algorithms for Predicting Chemical Kinetics", *Proceedings of the World Congress of Chemical Engineering*, Glasgow (2005).

[P12] O.O. Oluwole and W.H. Green, "Obtaining Accurate Solutions Using Reduced Chemical Kinetic Models", AIChE National Meeting, November 2005

[P13] W.H. Green, "Combustion Models and Data: A New Paradigm for the 21st Century", Plenary Lecture, Eastern States' Section of the Combustion Insitute, Orlando, FL, Nov. 2005.

[P14] Karen Schuchardt, Oluwayemisi Oluwole, William Pitz, Larry A. Rahn, William H. Green, Jr., David Leahy, Carmen Pancerella, Magnus Sjöberg, John Dec, "New Approaches for Collaborative Sharing of Chemical Model Data and Analysis Tools", Poster, *Sixth International Conference on Chemical Kinetics*, July 25-29, 2005, NIST, Gaithersburg, Maryland.

[P15] Karen Schuchardt, Carmen Pancerella, Larry A. Rahn, Brett Didier, Deepti Kodeboyina, David Leahy, James D. Myers, Olumwole Oluwole, William Pitz, Branko Ruscic, Jing Song, Gregor von Laszewski, Christine Yang, "*Portal-based Knowledge Environment for Collaborative Science* ", Proceedings of the GCE 2005: Workshop on Grid Computing Portals, Seattle, WA, Nov. 18, 2005.

[P16] W. J. Pitz, R. K. Cheng, J. Chase, B. Didier, T. Elsethagen, V. Gurumoorthi, C. Pancerella, K. Schuchardt, L. A. Rahn and F. C. Gouldin, "A Data Sharing Environment for Premixed Turbulent Flames," Work-In-Progress Poster, 31st International Combustion Symposium, Heidelberg, Germany, August 7-11, 2006.

[P17] N. Hansen, P. R. Westmoreland, and L.A. Rahn, "A Web-based Library of Low Pressure Flame Data," Work-In-Progress Poster, 31st International Combustion Symposium, Heidelberg, Germany, August 7-11, 2006.

[P18] Erwin Dunbar, Larry A. Rahn, Michael Chen, Robert Barlow, "A New Web Archive for TNF Workshop Data," Poster presentation at 8th International Workshop on the Measurement and Computation of Turbulent Nonpremixed Flames, Hiedelberg, Germany, August 4-5, 2006.

5.4 Ancillary References

[A1] KnECS open source software. Available at http://sourceforge.net/projects/knecs/, 2005.

[A2] Collaboratory for MS3D Website. Available at http://ms3d.org, 2005.

[A3] CHEF Collaborative Portal Framework Website. Available at http://www.chefproject.org/. University of Michigan, 2005.

[A4] Jakarta Slide Java Content Management System website. Available at http://jakarta.apache.org/slide/. Apache Jakarta Project, 2005.

[A5] Stein G. Web Digital Authoring and Versioning (WebDAV) Resources Community website. Available at http://www.webdav.org/, 2005.

[A6] Java Authentication and Authorization Service (JAAS). Available at http://java.sun.com/products/jaas/, 2005

[A7] Commodity Grid Kits. Available at http://www-unix.globus.org/cog/. University of Chicago, 2005.
[A8] Resource Description Framework (RDF). Available at http://www.w3c.org/RDF/.
[A9] XML Linking Language (XLink) Version 1.0. Available at http://www.w3.org/TR/xlink/. World Wide Web Consortium (W3C), 2001.
[A10] Lucene, Open Source search software, using Java. Available at http://lucene.apache.org/. Apache Project, 2005.
[A11] DAV Searching and Locating (DASL) website. Available at http://www.webdav.org/dasl, 2005.
[A12] Black G, Gracio D, Schuchardt K, Palmer B. 2003. The extensible computational chemistry environment: A problem solving environment for high performance theoretical chemistry. In: *Proceedings of Computational Science - ICCS 2003, International Conference*, Eds. PMA Sloot, D Abramson, A Bogdanov, JJ Dongarra, A Zomaya, Y Gorbachev. Vol. 2660, Lecture Notes in Computer Science. Springer-Verlag, Berlin.
[A13] Kendall RA, Aprà E; Bernholdt DE, Bylaska EJ, Dupuis M, Fann GI, Harrison RJ, Ju J, Nichols JA, Nieplocha J, Straatsma TP, Windus TL, Wong AT. High performance computational chemistry: An overview of NWChem, a distributed parallel application. *Computer Physics Communication* 2000; **128**:260-283.
[A14] Gaussian website. Available at http://www.gaussian.com/, 2005.
[A15] Taatjes CA, Hansen N, Mclloy A, Miller JA, Senosiain JP, Klippenstein, SJ, Qi F, Sheng L, Zhang Y, Cool TA, Wang J, Westmoreland PR, Law ME, Kasper T, Kohse-Hoinghaus K. 2005. Enols are common intermediates in hydrocarbon oxidation. *Science* 2005; **24**:1887-1889.
[A16] Data Format Description Language (DFDL) website. Available at http://forge.gridforum.org/projects/dfdl-wg, 2005.
[A17] Defuddle website, A Parser Implementation using the DFDL. Available at http://defuddle.pnl.gov/, 2005.
[A18] Java Portlet Specification, Version 1.0, JSR 168, October 2003.
[A19] Web Services for Remote Portlets Specification. Available at http://www.oasis-open.org/committees/WSRP, 2005.