

Collaboratory for Multi-Scale Chemical Science
Status as of September 2003 / Quarterly Report for Q3-Q4 of FY 2003
Larry A. Rahn, rahn@sandia.gov

Project Staff

SNL-Larry Rahn*, Christine Yang, Carmen Pancerella, David Leahy, Michael Lee, Darrian Hale, PNL-Theresa Windus*, James D. Myers, Karen Schuchardt, Brett Didier, Eric Stephan, Carina Lansing, ANL-Al Wagner*, Branko Ruscic, Michael Minkoff, Sandra Bittner, Gregor von Laszewski, Reinhardt Pinzon, Kaizar Amin, Shasshank Shank, Baoshan Wang, LLNL-William Pitz*, LANL-David R. Montoya*, Bill Barber, NIST-Thomas C. Allison*, MIT-William H. Green, Jr.*, Luwi Oluwole, UCB-Michael Frenklach*

* denotes Institutional Point of Contact

Summary

This performance period began with the successful Peer Review of the pilot Collaboratory for Multi-Scale Chemical Sciences (CMCS). The review panel offered constructive feedback and numerous insights into the impacts that CMS can make in the scientific community. Significant development during this period focused on the implementation of Version 1.0 of our software on a production server (in May), and its evaluation and improvement for the science pilot activities. This work culminated in updating the production server to Version 1.1 with many improvements at the end of the period. In addition, chemical applications were updated, including a new Version 1.1 of Active Tables. A process to develop and gain consensus approval of Web Data-sharing Agreements was initiated, and much progress made towards approved documents for pilot activity. Finally, the progress of the CMCS project was reported or submitted to six public scientific forums by project team members during this period.

Progress

This document summarizes the work done over the third and fourth quarter of FY03 by the CMCS project. First, a summary listing of project activities is presented, then more a detailed discussion of accomplishments is presented, highlighting an improved pedigree user interface.

- **Summary of CMCS Project Activities – May 2003 through September 2003**
 - CMCS project Peer Review (4/30-5/1/2003)
 - DOE Open Data Policy Team Conference Call (5/7/2003)
 - CMCS project team meeting (5/15/2003)
 - CMCS Release Candidate 1.0 software up on productions portal server. (5/23/2003)
 - DOE Open Data Policy Team Conference Call (5/29/2003)
 - DOE Networking Workshop in Wash DC-L. Rahn attending (6/2-6/5/2003)
 - Portal ‘Walk Through’ for application scientists (6/4/2003)
 - Portal ‘Walk Through’ for application scientists (6/5/2003)

- Portal 'Walk Through' for application scientists (6/6/2003)
- Implemented www.cmcs.org as URL for CMCS Project and Portal (6/14/2003)
- CMCS project team meeting (6/19/2003)
- CMCS project team meeting (7/17/2003)
- Infrastructure Overview presentation to MIT graduate students (7/29/2003)
- Version 1.0 of Open Data Policy and User Agreements approved by SNL, submitted to CMCS Team (8/15/2003)
- CMCS project team meeting (8/21/2003)
- Poster presentation at Gordon Conference on Laser Diagnostics for Combustion, Oxford, UK, August 17-22, 2003
- CMCS project team meeting (9/18/2003)
- Presentation and Poster at the NIST Workshop on Combustion Simulation Databases for Real Transportation Fuel, September 4-5, 2003.
- Presentation at 2003 Dublin Core Conference (DC-2003), September 28 – October 2, 2003, Seattle, Washington
- Version 1.0 of Open Data Policy and User Agreements approved by CMCS Team contingent upon DOE Open Data Policy Team approval (10/1/2003)

- **Project Management, Structure and Planning**

CMCS priorities focused on deriving feedback on Version 1.0 of the infrastructure software from pilot scientific leaders and groups, and on preparing for presentations at SC03 in November. Changes to the management structure of CMCS are being developed to further enhance our ability to work well with current and future science pilot groups. We started by teaming each lead chemical scientist with a CMCS teammate working on the infrastructure. A task to further formalize enhanced coupling to scientific pilot activities is in progress, with anticipated changes proposed to the Management Plan.

Another important task initiated during this performance period is the development of a Data Policy and related User and Content Provider Agreements. The products of this task will be agreements that visitors and users of the CMCS service must agree to in order to protect the rights of data providers and limit the liabilities of the host of the service (currently SNL). We developed drafts for Content Provider Agreement, User Agreement (for those who don't plan to contribute digital content), and a Security and Privacy Policy and vetted these agreements with Paul Gottlieb (DOE HQ), Dickson Kehl (DOE Albuquerque), Gary Drew (DOE Oakland), Larry Adcock (NNSA), Kurt Olsen and Craig Smith (SNL), Steve May (PNNL), and Helen Cordell (ANL). These drafts were accepted by SNL counsel and the CMCS Pilot Users for their consideration, but still have a final iteration with the broader legal counsel team led by Paul Gottlieb.

The resource redistribution issue resolved with the sponsor during the last quarter was implemented and is complete.

- **The CMCS Portal and Infrastructure**

Version 1.1 of the CMCS infrastructure software was installed on the production server on September 26. This release remained in release candidate status until October 3 to

provide sufficient time for users to report significant problems with the new software that would be addressed in an immediate fashion by the development team. The CMCS cvs software tree was tagged with version 1.1 on October 3. This version is responsive to the feedback received from chemical scientists that were pilot users of version 1.0. New capabilities in this version include: anonymous browsing of public data, controlling access on a per user basis to data within the portal, and incorporating tutorials and help items within the portal to describe its capabilities. This version incorporates new versions of the CompreHensive collaborative Framework (CHEF v1.1.01) and the Scientific Annotation Middleware (SAM V1.1). The upgrade to SAM provides a more robust data repository and the CHEF upgrade provides a new look and feel to the portal. Performing the CHEF upgrade was a significant undertaking, because of changes they performed to their base software. Furthermore, literally dozens of improvements were made to the portal user interfaces in response to feedback, evaluation, and testing activities. The new look and feel and the new feature of fine-grained access control in a group are illustrated in Figure 1 below.

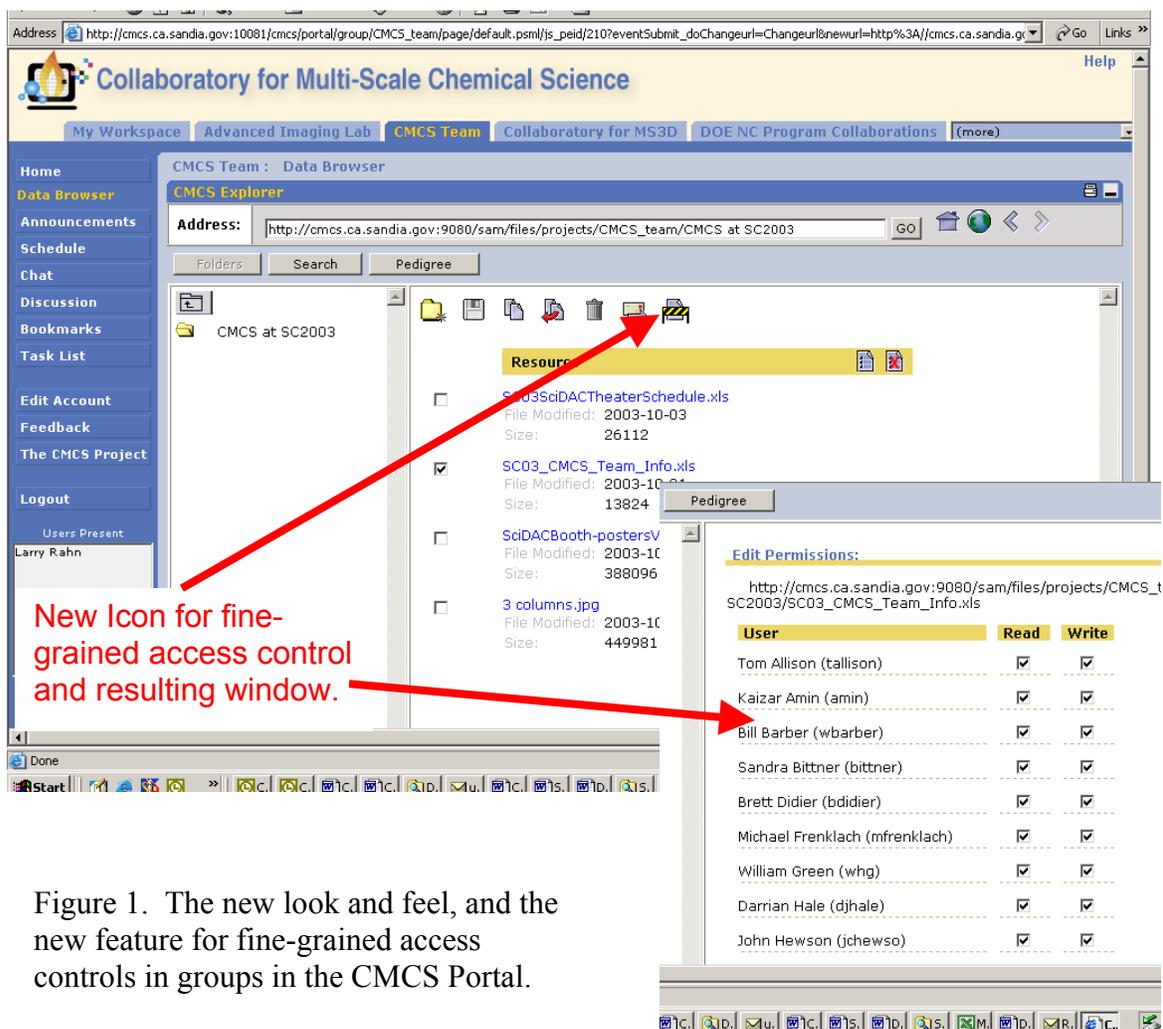


Figure 1. The new look and feel, and the new feature for fine-grained access controls in groups in the CMCS Portal.

Working with the SAM project, CMCS developers have designed a mechanism to automatically invoke external web services to generate metadata and provide data translations from within the CMCS portal. This mechanism builds on the current ability to invoke Binary Format Description (BFD) and XSLT scripts to generate metadata during data upload and to dynamically generate data translations. As a first use of this new capability, CMCS has wrapped the OpenBabel (openbabel.sourceforge.net) library as a web service. Open Babel is a cross-platform program and library designed to interconvert between many file formats used in molecular modeling and computational chemistry. The current version provides services for over 40 file types.

- **Application Science and Pilot Activities**

The existing CMCS 1.0 functionality was reviewed with Michael Frenklach and helped to further identify and prioritize requirements for PrIME. CMCS-friendly XML formats for initial PrIME library data were developed along with software to produce such XML. An updated set of PrIME library data including valuable new annotations was then created. Automatic annotation tools for the staging of these data on CMCS were developed along with basic viewing tools. Considerable useful feedback into Version 1.1 improvements as well as the Data Policy and related User and Content Provider Agreements were provided, derived from the PrIME vision and use cases for advanced approaches to collaborative chemical science.

Further requirements and capabilities enabling the LLNL chemical science have been pursued. In collaboration with Tom Allison the need for multiple requests for CAS numbers from the species dictionary was evaluated as a way to enable deriving CAS numbers from a species list that then could be added to the database. In collaboration with Michael Frenklach, ReactionLab software is being evaluated to convert the LLNL species thermodynamics and reaction mechanisms to XML files that can then be stored in CMCS data storage. One of the large iso-octane species thermodynamic files has been converted using ReactionLab and placed in the HCCI working group area on the production CMCS portal. The large iso-octane data set could be part of a public data set available for Supercomputing.

Formal support for the IUPAC Task Group on Thermochemistry of Radicals was begun during this period. The IUPAC Task Group was introduced to CMCS capabilities by making a call to all members to encourage them to visit the CMCS portal and download the just submitted manuscript for a joint publication. Accounts were opened for all members and we started organizing the file repository structure and depositing some of the initial files. The initial response was quite enthusiastic. However, progress after that was sluggish, partly arising from the fact that the portal does not seem to be entirely intuitive/self explanatory. Some of these non-intuitive aspects have been subsequently corrected in Version 1.1 of the infrastructure and the user agreements. Work on what will become the final versions of CODATA Library, JANAF Library and Gurvich Library has started. While many chemical species from these libraries were included in the initial version. The aim is now to include in these libraries all species (gas and non-

gas) appearing in the CODATA Key Values, JANAF and in Vols. 1 & 2 of the Russian compilation.

A new version (1.100) of the Active Thermochemical Tables kernel that introduces the capability to treat and solve aqueous thermochemistry. This considerable complication was not planned initially at all, but turned out to be extremely important, since gas-phase and aqueous thermochemistry are intimately connected at the very core (e.g. doing computational thermochemistry for halo-compounds and non-halo compounds that are networked to halo-compounds is impossible without aqueous thermochemistry). Also, considerable progress in developing the core Thermochemical Network (TN) for the Main Library of ATcT: ~100 new species and ~ 600 measurements were added during the summer. The core TN now has ~200 species. This is still work in progress, and we want to introduce many more compounds, particularly radicals. However, the present Network is already close to producing new publishable results; work on three separate thermochemical topics (and hence three papers in chemical journals) is in progress. -- During the development of the core Thermochemical Network a number of "weak links" have been discovered. Some of these are now being pursued experimentally in our lab, in collaboration with the Advanced Light Source in Berkeley, and in collaboration with theorists that can conduct state-of-the-art quantum mechanical calculations (at least one order of magnitude more accurate than G3 or its variants).

CMCS has designed a portlet and web service framework for the integration of MIT's Range Identification and Optimization Tool (RIOT). RIOT is a numerical package for computer-aided kinetic model reduction with valid range analysis. Reliable kinetic model reduction allows accurate chemistry models to be employed in computational fluid dynamics (CFD) simulations, i.e. it allows one to bridge from molecular chemistry scales up to the centimeter scale. The program takes as input a comprehensive kinetic mechanism (in CHEMKIN format) and returns the smallest possible mechanism that satisfies user-specified species and energy flux error tolerances at given points (sets of temperature and species concentrations). This is achieved by eliminating negligible reactions. If requested, the valid range of the resulting reduced model (range of temperature and species concentrations in which model satisfies user tolerances) is also determined. The portlet design consists of a series of screens for user input of options that are driven by existing CHEMKIN full model files. Depending on the options selected, the RIOT program may take considerable time to execute. Thus, the web service infrastructure will support both synchronous and asynchronous requests. When the request is completed, the user will be notified via email and the files placed in the users working directory on the CMCS server where outputs can be viewed using existing CMCS translation and presentation services. The web service architecture is based on Apache Axis with JMS messaging for asynchronous web services.

- **Conference Presentations, workshops, and Publications**

“Enabling Collaborative Combustion Data and Metadata Sharing,” Larry A. Rahn and David Leahy, Poster presentation at the Gordon Conference on Laser Diagnostics for Combustion, Oxford, UK, August 17-22, 2003.

“Metadata in the Collaboratory for Multi-Scale Chemical Science,” Carmen Pancerella,¹ James D. Myers,² Thomas C. Allison,⁶ Kaizar Amin,³ Sandra Bittner,³ Brett Didier,² Michael Frenklach,⁸ William H. Green, Jr.,⁶ Yen-Ling Ho,⁵ John Hewson,¹ Wendy Koegler,¹ Carina Lansing,³ David Leahy,¹ Michael Lee,¹ Renata McCoy,² Michael Minkoff,³ Sandeep Nijsure,³ Gregor von Laszewski,³ David Montoya,⁵ Reinhardt Pinzon,³ William Pitz,⁴ Larry Rahn,¹ Branko Ruscic,³ Karen Schuchardt,² Eric Stephan,² Al Wagner,³ Baoshan Wang,³ Theresa Windus,² Lili Xu,⁵ Christine Yang¹, Presented by Brett Didier at the 2003 Dublin Core Conference (DC-2003), September 28 – October 2, 2003, Seattle, Washington. Paper is published online at: http://www.siderean.com/dc2003/401_Paper67.pdf

“Multi-scale Science: Supporting Emerging Practice with Semantically-Derived Provenance,” James D. Myers,¹ Carmen Pancerella,² Carina Lansing,¹ Karen L. Schuchardt,¹ and Brett Didier¹, Presented at 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. This paper has been published Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data. Paper is published online at <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-83/>.

“Enabling Collaborative Science for Real Fuels Combustion,” Presentation and Poster by Larry A. Rahn and David Leahy at the NIST Workshop on Combustion Simulation Databases for Real Transportation Fuel, National Institute of Standards and Technology, September 4-5, 2003.

"Collaboratory for Multi-scale Chemical Science." Presented by Theresa Windus at American Chemical Society (ACS), New York City, NY on September 8, 2003.

“Optimally-Reduced Kinetic Models: Reaction Elimination in Large-Scale Kinetic Mechanisms”, Binita Bhattacharjee, Douglas A. Schwer, Paul I. Barton, & William H. Green, Jr., *Combustion & Flame* (2003) in press.

- **Abstracts of conference Presentations, and Publications**

Enabling Collaborative Combustion Data and Metadata Sharing

Larry A. Rahn and David Leahy

Sandia National Laboratories, Livermore, CA 94550

In this poster we present a new informatics infrastructure designed to facilitate collaborative combustion science. Web-based data sharing for collaborative combustion research activities has become important in a growing number of projects. Two well-known examples in this community are the past work on GRI-Mech (http://www.me.berkeley.edu/gri_mech/) and the Turbulent Nonpremixed Flame workshop series (<http://www.ca.sandia.gov/TNF/>). The new informatics infrastructure

reported here is the first release from the Collaboratory for Multi-scale Chemical Science (CMCS, <http://cmcs.ca.sandia.gov/>). CMCS is implementing novel informatic data-sharing concepts, and is piloting this infrastructure among a multi-disciplinary team of chemical scientists working to advance combustion science.

The data infrastructure takes advantage of a variety of standards and open-source information technologies to provide an unprecedented ability to share data, data pedigree, and project information within groups and across communities. A shared data service provides configurable capabilities for automating the generation of metadata, translating data between standard formats, and federating multiple data stores. A portal serves as the web interface for the adaptable informatics infrastructure being developed by the CMCS team. The portal provides an array of functionality to support group and community processes, with an emphasis on simplifying the discovery and use of data. A pedigree browser can easily display pedigree data (as well as annotations) and allows users to search, browse, and retrieve a data set's pedigree. Pedigree data may also be an active link to a different, but associated, data resource.

To support the chemistry community, the CMCS team has integrated a variety of powerful chemistry applications, data viewers, and data translators. Examples of the implementation of the CMCS infrastructure in support of concepts and data in the context of experimental data and validations of combustion models are discussed. The capabilities that would support a collaborative data store for quantitative combustion diagnostics is also discussed.

Supported by the U. S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research; Mathematical, Information and Computational Science.

Metadata in the Collaboratory for Multi-Scale Chemical Science

Carmen Pancerella,¹ James D. Myers,² Thomas C. Allison,⁶ Kaizar Amin,³ Sandra Bittner,³ Brett Didier,² Michael Frenklach,⁸ William H. Green, Jr.,⁶ Yen-Ling Ho,² John Hewson,¹ Wendy Koegler,¹ Carina Lansing,³ David Leahy,¹ Michael Lee,¹ Renata McCoy,² Michael Minkoff,³ Sandeep Nijsure,³ Gregor von Laszewski,³ David Montoya,⁵ Reinhardt Pinzon,³ William Pitz,⁴ Larry Rahn,¹ Branko Ruscic,³ Karen Schuchardt,² Eric Stephan,² Al Wagner,³ Baoshan Wang,³ Theresa Windus,² Lili Xu,⁵ Christine Yang¹

¹Sandia National Laboratories, Livermore, CA 94551-0969

²Pacific Northwest National Laboratory, Richland, WA 99352

³Argonne National Laboratory, Argonne, IL 60439-4844

⁴Lawrence Livermore National Laboratory, Livermore, CA 94551

⁵Los Alamos National Laboratory, Los Alamos, NM 87545

⁶NIST, Gaithersburg, MD 20899-8381

⁷MIT, Cambridge, MA 02139

⁸University of California, Berkeley, CA 94720-1740

carmen@sandia.gov

Abstract

The goal of the Collaboratory for the Multi-scale Chemical Sciences (CMCS) [1] is to develop an informatics-based approach to synthesizing multi-scale chemistry information to create knowledge in the chemical sciences. CMCS is using a portal and metadata-aware content store as a base for building a system to support inter-domain knowledge exchange in chemical science. Key aspects of the system include configurable metadata extraction and translation, a core schema for scientific pedigree, and a suite of tools for managing data and metadata and visualizing pedigree relationships between data entries. CMCS metadata is represented using Dublin Core with metadata extensions that are useful to both the chemical science community and the science community in general.

CMCS is working with several chemistry groups who are using the system to collaboratively assemble and analyze existing data to derive new chemical knowledge. In this paper we discuss the project's metadata-related requirements, the relevant software infrastructure, core metadata schema, and tools that use the metadata to enhance science.

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

Multi-scale Science: Supporting Emerging Practice with Semantically-Derived Provenance

James D. Myers,¹ Carmen Pancerella,² Carina Lansing,¹
Karen L. Schuchardt,¹ and Brett Didier¹

¹Pacific Northwest National Laboratory, Richland, WA 99352

²Sandia National Laboratories, Livermore, CA 94551-0969

jim.myers@pnl.gov, carmen@sandia.gov, carina.lansing@pnl.gov,
karen.schuchardt@pnl.gov, brett.didier@pnl.gov

Abstract

Scientific progress is becoming increasingly dependent of our ability to study phenomena at multiple scales and from multiple perspectives. The ability to recontextualize third party data within the semantic and syntactic framework of a given research project is increasingly seen as a primary barrier in multi-scale science. Within the Collaboratory for Multi-scale Chemical Science (CMCS) project, we are developing a general-purpose informatics-based approach that emphasizes “on-demand” metadata creation, configurable data translations, and semantic mapping to support the rapidly increasing and continually evolving requirements for managing data, metadata, and data relationships in such projects. A concrete example of this approach is the design of the CMCS provenance subsystem. The concept of provenance varies across communities, and multiple independent applications contribute to and use provenance. In CMCS, we have developed generic tools for viewing provenance relationships and for using them to, for example, scope notifications and searches. These tools rely on a configurable concept of provenance defined in terms of other relationships. The result is a very flexible mechanism capable of tracking data provenance across many disciplines and supporting multiple uses of provenance information.

Collaboratory for Multi-scale Chemical Science (CMCS)

Presented by Larry A. Rahn and David Leahy
Sandia National Laboratories, Livermore, CA

Abstract

The goal of the CMCS is to enhance chemical science research by breaking down the barriers to rapid sharing of validated information and by opening new paradigms for multi-scale science. To accomplish this CMCS is developing an adaptive informatics infrastructure and demonstrating proof-of-concept by publicly deploying an integrated set of key collaboration tools and chemistry-specific applications, data resources, and services. We have implemented a prototype of the Version 1 infrastructure software, integrated initial application data and tools, produced use-cases illuminating the central aspects of the project, and are now testing these capabilities in collaboration with initial pilot scientific activities.

Collaboratory for Multi-scale Chemical Science

Theresa L. Windus, Molecular Science Software Group, Pacific Northwest National Laboratory, 902 Battelle Boulevard, P.O. Box 999, MSIN: K1-96, Richland, WA 99352, Fax: 509-375-6631, theresa.windus@pnl.gov

The Collaboratory for Multi-scale Chemical Science (CMCS) is a DOE sponsored environment for enabling chemical information to be communicated, translated and annotated across several chemical scales. Enabling a dynamic environment in which to perform new informatics based manipulations is the ultimate goal of this project. The initial scales are the molecular (computational, *ab initio* data), thermochemical, kinetic, kinetic mechanism, and the numerical simulation scales (including computational and experimental data). This talk will present the data involved, the formats used to describe this data, the pedigree information associated with the data, and the collaboratory infrastructure and portal that enable researchers to access, annotate and manipulate the data. The chemistry communities piloting use of the CMCS will also be discussed.

Optimally-Reduced Kinetic Models: Reaction Elimination in Large-Scale Kinetic Mechanisms

Binita Bhattacharjee, Douglas A. Schwer, Paul I. Barton, and William H. Green
Dept. of Chemical Engineering, Massachusetts Institute of Technology

Abstract

A new optimization-based approach to kinetic model reduction is presented. The reaction-elimination problem is formulated as a linear integer program which can be solved to guaranteed global optimality. This formulation ensures that the solution to the integer program is the smallest possible reduced model consistent with the user-set tolerances. The method is applied to generate optimally-reduced models for isobaric, adiabatic homogeneous combustion. The computational cost and accuracy of the reduced models are compared to those of the full mechanism. Results are shown for GRI mech 3.0 and the Lawrence Livermore n-heptane combustion mechanism. The accuracy of the integer programming approach is compared to existing reaction elimination methods. The method is also applied to generate a library of reduced kinetic models for an adaptive chemistry simulation of a 2-D laminar, partially-premixed methane burner flame. Preliminary results are presented comparing the computational cost of the full GRI mech 3.0 chemistry to that of the reduced model library.