

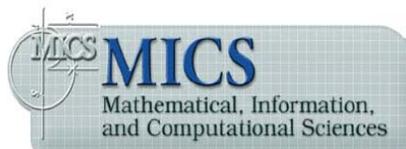


Chemical Informatics and You: Data Annotation, Translation and Visualization

David Leahy

djleahy@sandia.gov

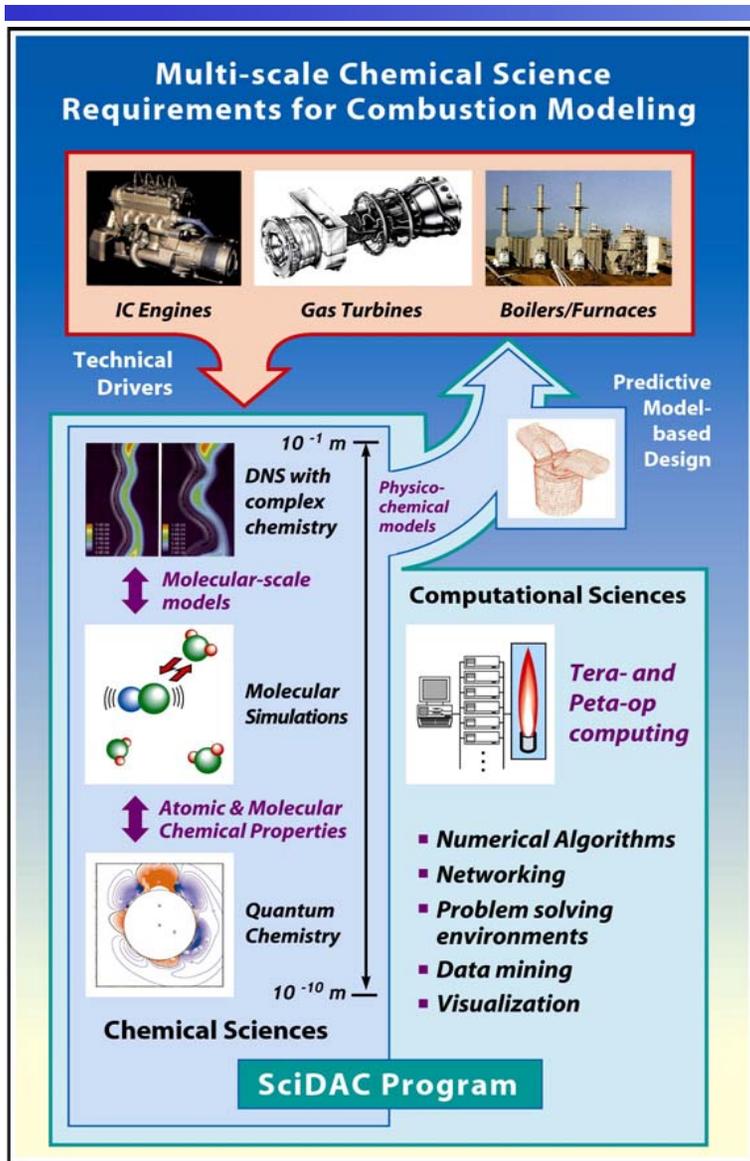
Funding provided by



National Collaboratory Program

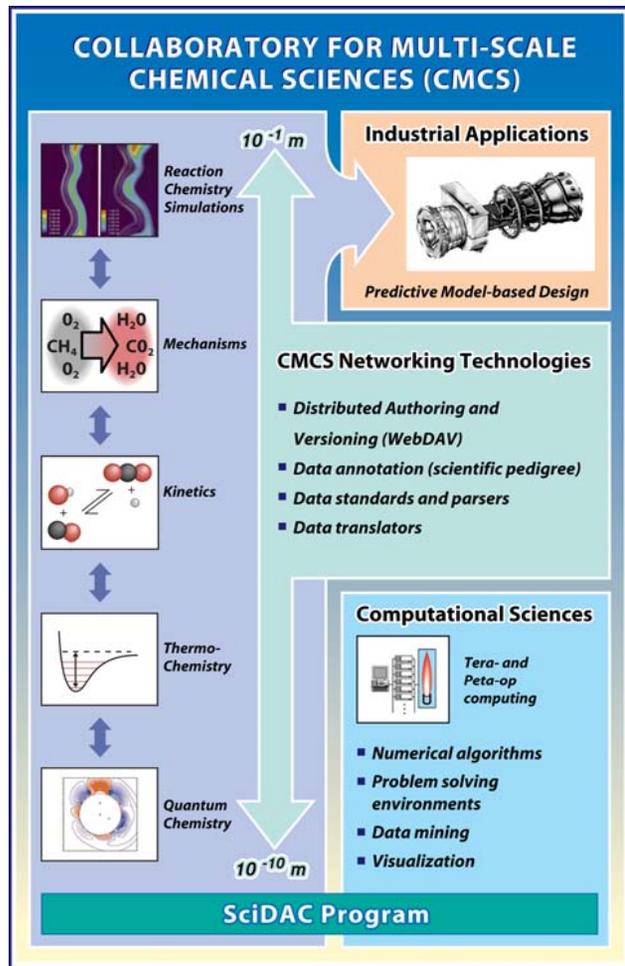
Dept. 8358 Seminar - January 29, 2003

Overview



- CMCS project
 - ▶ Who we are
- Problem statement
 - ▶ Need for shared knowledge bases: “Adaptive Informatics Infrastructure”
- Technology DEMO
- New Standards, Technologies
- Scientific Data Annotation
- Translation/Visualization
- Long term vision

Collaboratory for Multi-scale Chemical Sciences (CMCS)



- A collaboration of eight national labs and universities
 - ▶ Chemical scientists spanning the scales from electronic structure of molecules to simulations of reacting flow
 - ▶ Computer and information scientists expert in emerging web-based technologies
- Funded by DOE/SC MICS office
 - ▶ Part of the National Collaboratory Program
 - ▶ Supports BES SciDAC and Chemical Sciences Community
- Pilot project within combustion community with much broader goals in the longer term



The Data Problem

- **Data-centric perspective of scientific content**
 - ▶ Scientific content, at its core, is digital data and annotation
 - ▶ Scientific publications represent a great deal of distillation of data
- **Data growth is exponential**
 - ▶ Experiments coughing up more and more complex data
 - ▶ Computation methods coming to the fore; huge datasets
- **Human brain growth is not exponential**
 - ▶ Data analysis capacity rising linearly at best

→ Data Sharing: There is a Growing Need for Capabilities in Data/Knowledge Sharing



Thermochemist



Thermo-
Chemistry
Application

Thermo-
Chemistry
data

Parsers
Annotators

Shared Data Repository

Distributed Authoring and Versioning (WebDAV) protocol

DAV prop
Pedigree
Metadata

XML
Quantum
chemistry
data

Link

DAV properties
Pedigree
Metadata

XML
Thermo
data

Link

DAV prop
Pedigree
Metadata

XML
Thermo
data

Thermochemist



Thermo-
Chemistry
Application

Thermo-
Chemistry
data

Parsers
Annotators

Kineticist



Kinetics
Application

Thermo-
Chemistry
data

Kinetics
data

Parsers
Annotators
Translators

Shared Data Repository

Distributed Authoring and Versioning (WebDAV) protocol

Annotation

XML

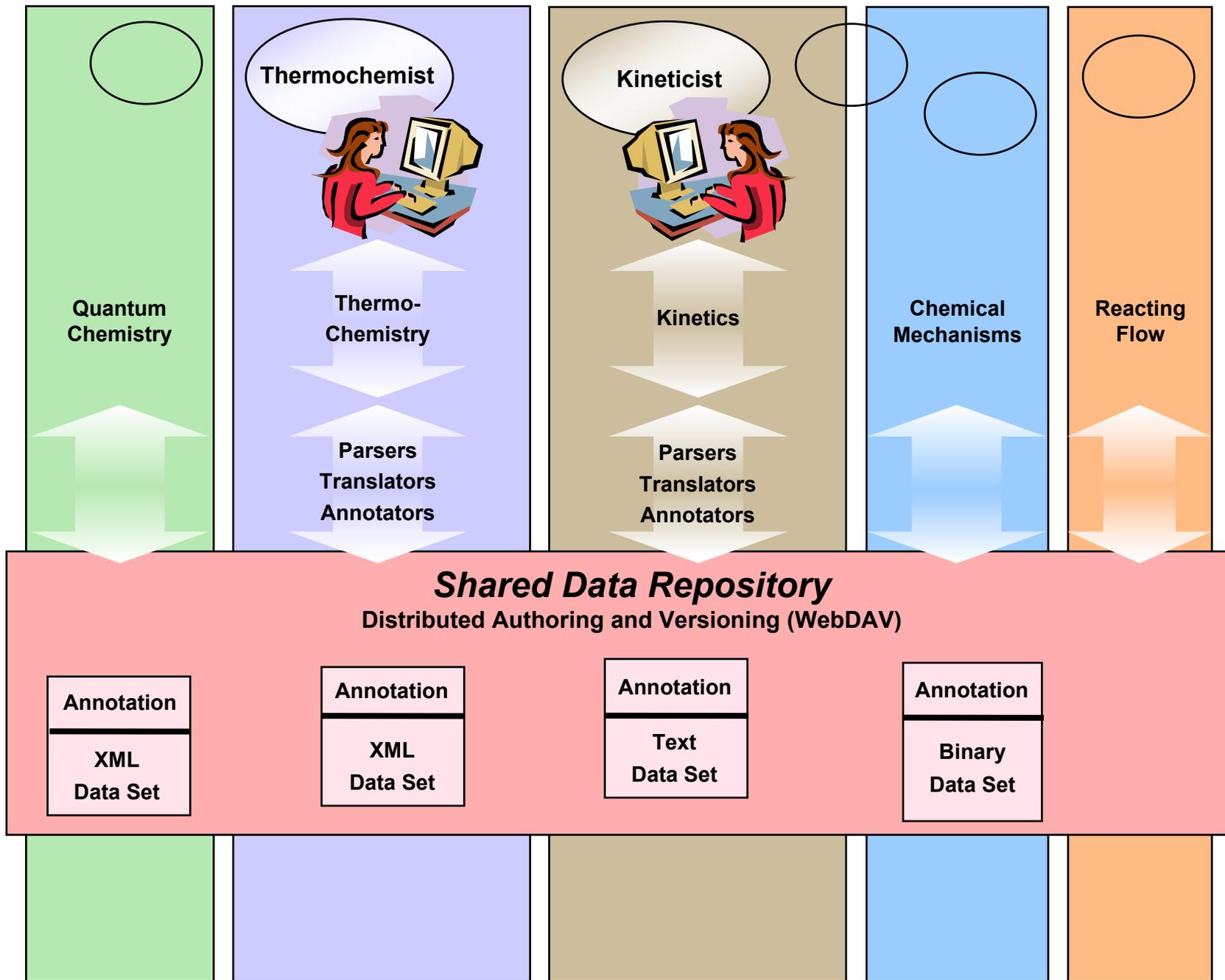
Quantum
Chemistry
Data Set

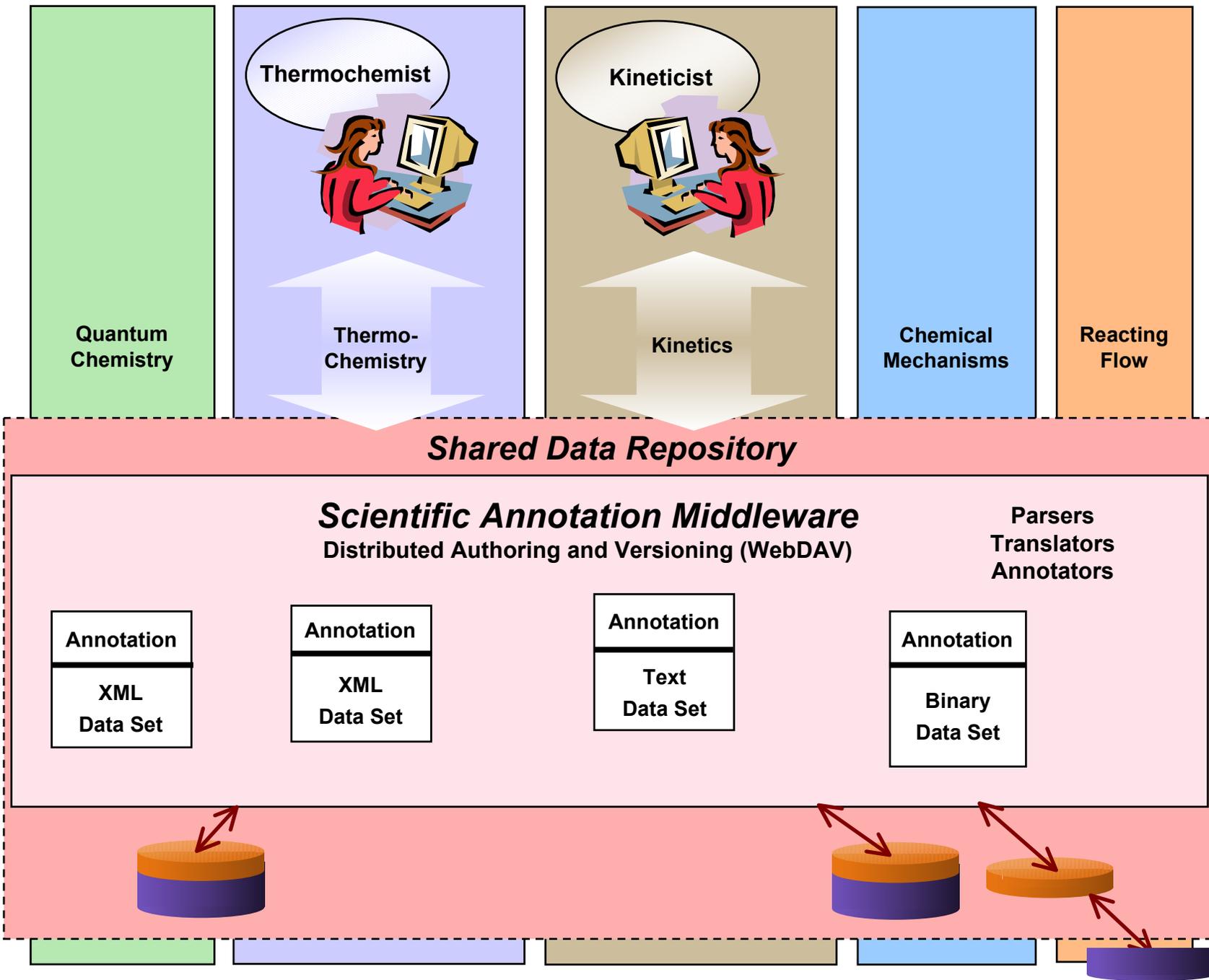
Annotation

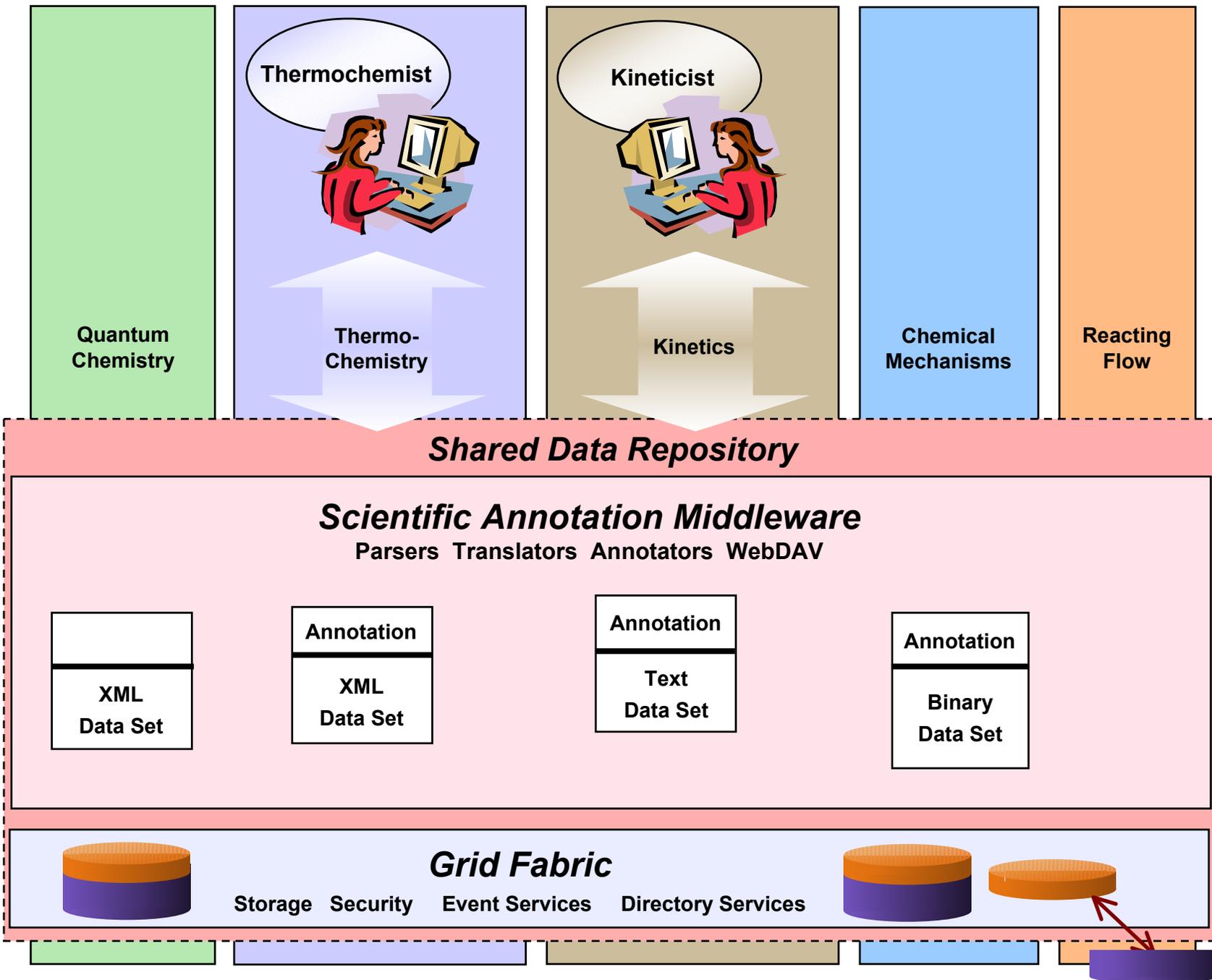
XML Thermo
Data Set

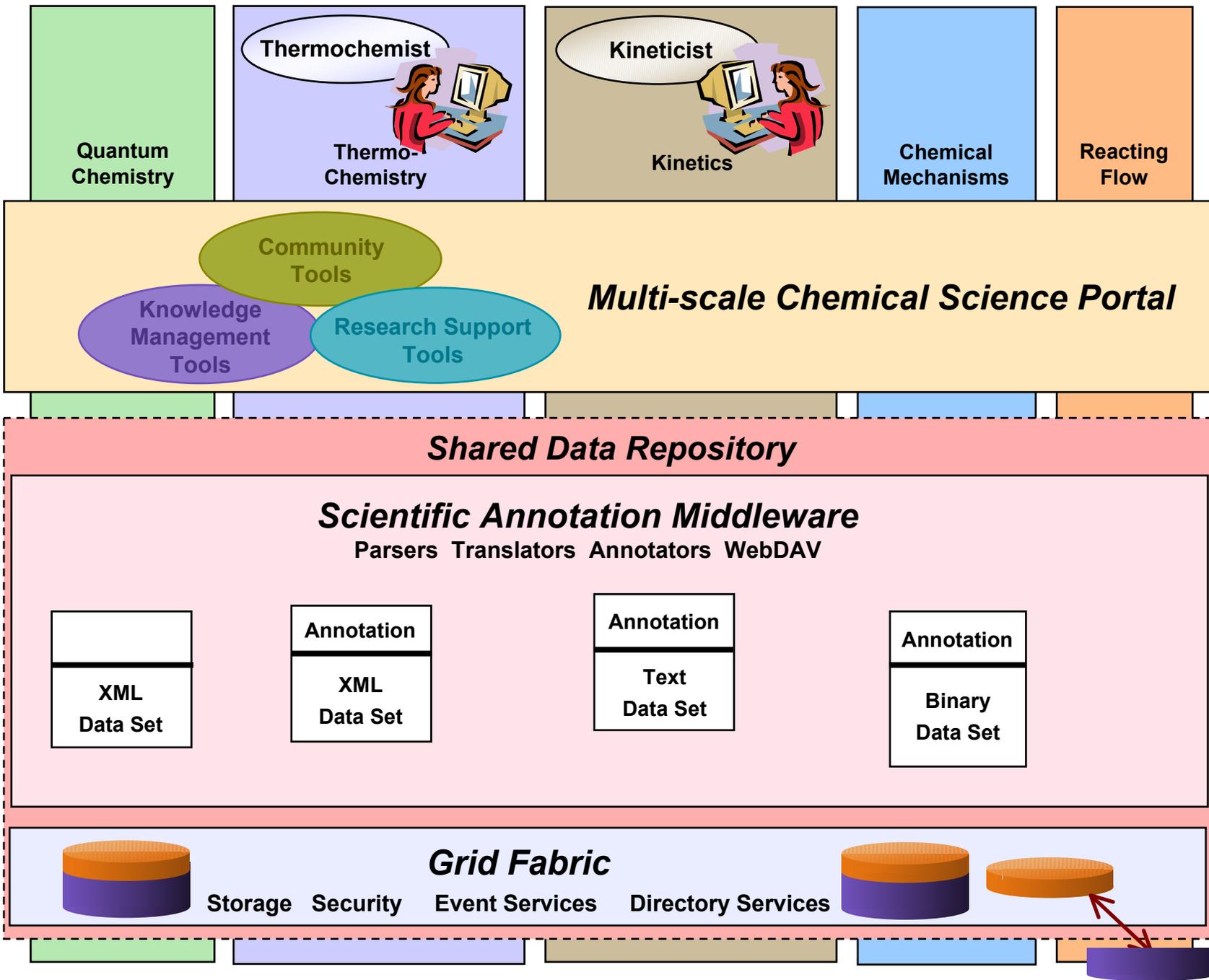
Annotation

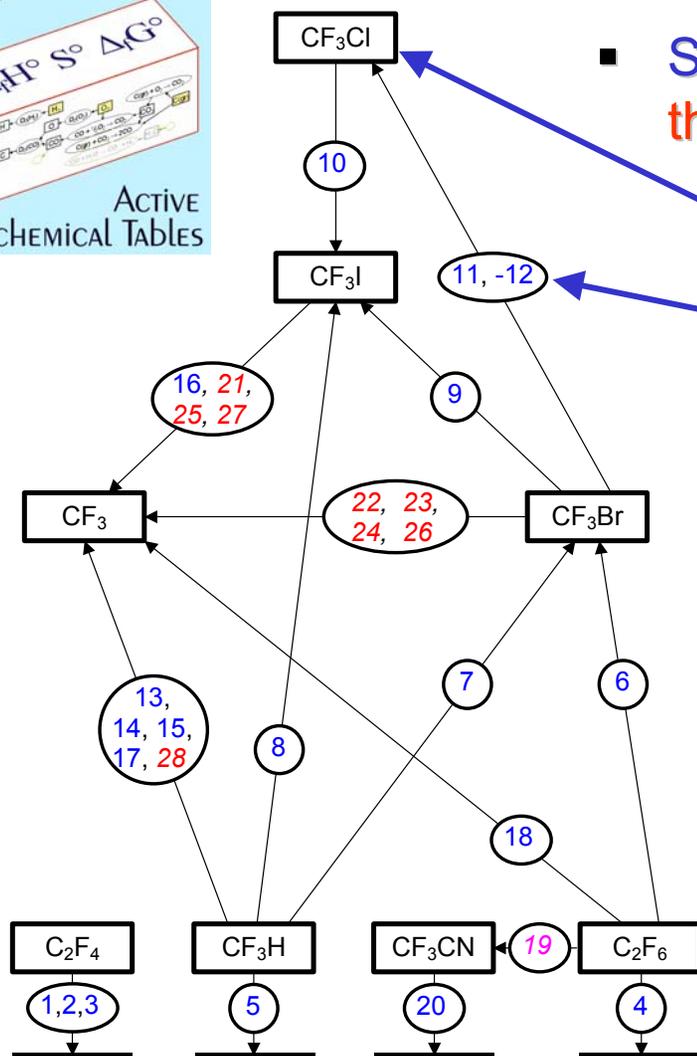
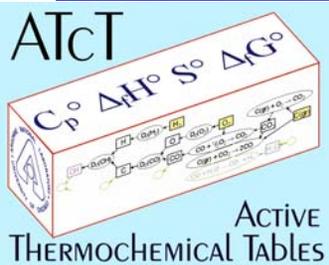
XML
Kinetics
Data Set







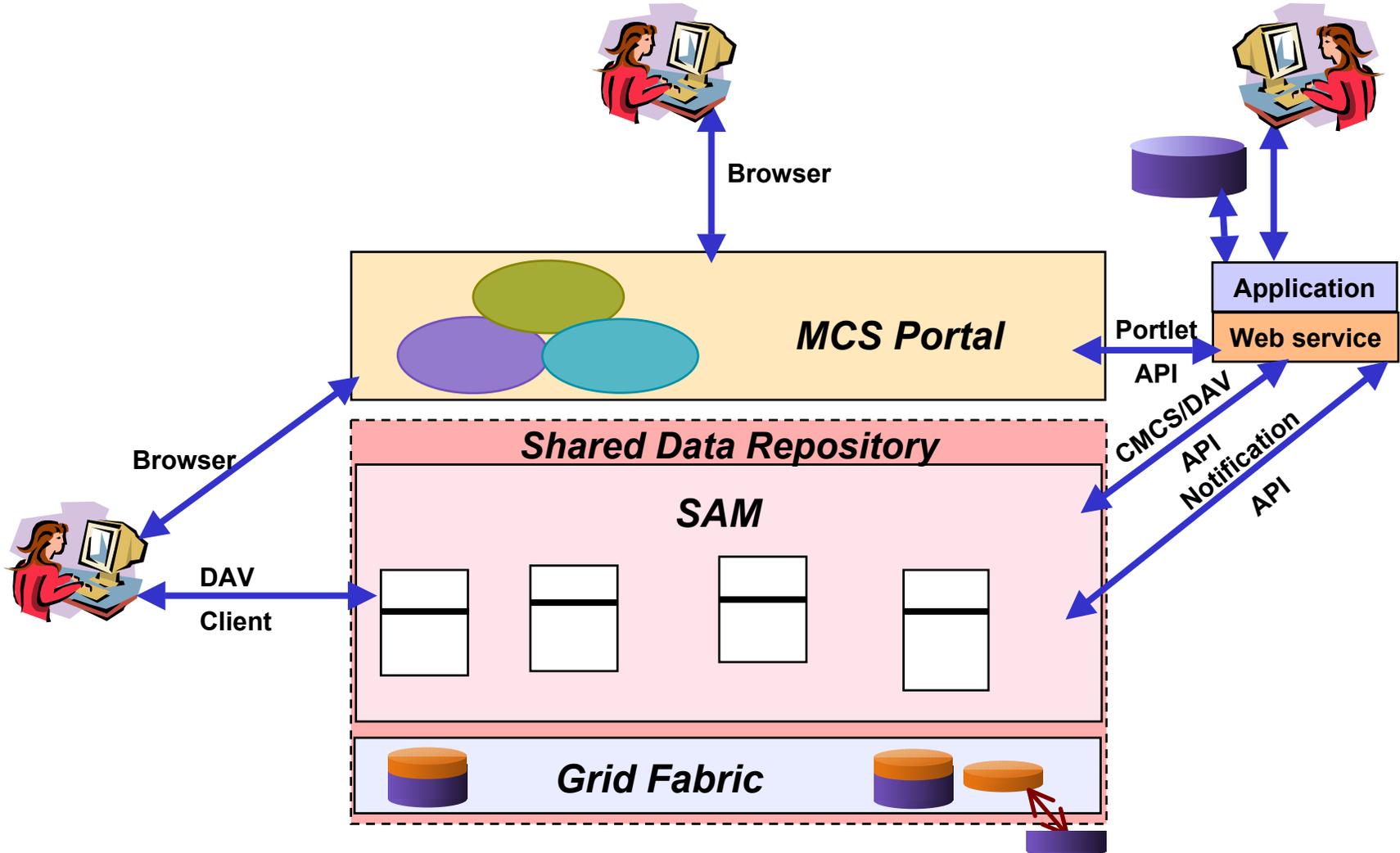




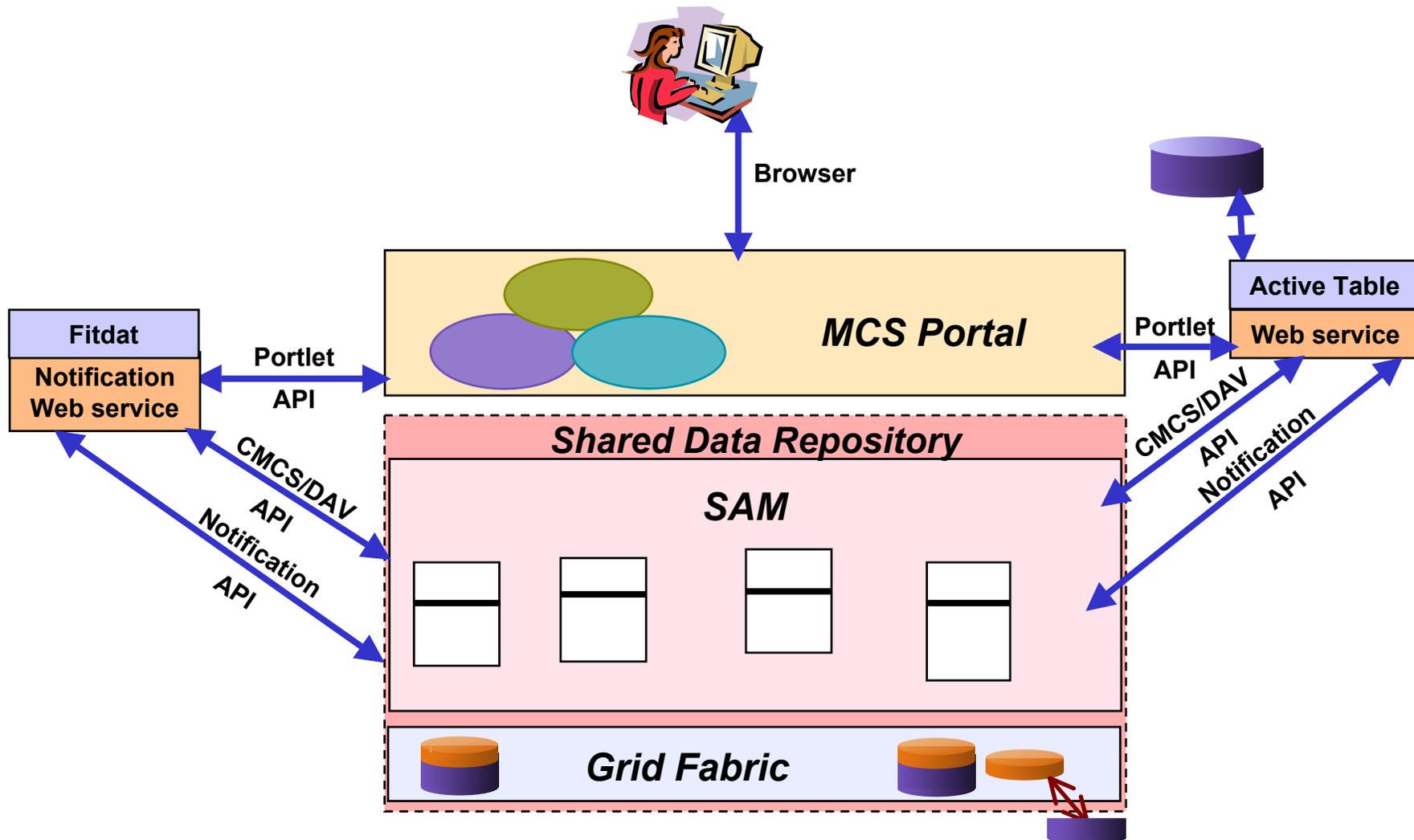
Species-interconnecting data form a thermochemical network

- Vertices (nodes): sought-for enthalpies
- Edges (links): species-interconnecting measurements (reaction enthalpies, equilibria,...)
- Generally there are:
 - ▶ multiple links between nodes (competing measurements)
 - ▶ alternative paths to the same node
- The best set of enthalpies describing the underlying knowledge is obtained not by choosing a particular sequential path through particular links, but by simultaneous solution of the thermochemical network

Adaptive Infrastructure Enables Applications



Adaptive Infrastructure Enables Applications



Challenges in Data/Knowledge Sharing



- **Infrastructural challenges**

- ▶ **Un- or under-documented data**
 - ▶ Scientific pedigree is supremely important
- ▶ **One-off data formats**
- ▶ **Legacy media**
 - ▶ Limited access to potentially valuable data stores
- ➔ **Limited ability to find, assess, analyze data**

- **Behavioral challenges**

- ▶ **Big science vs. little science**
 - ▶ Modern science problems grow ever larger and more complex
 - ▶ Scientists are accustomed to working in isolation on their own problem spaces
- ▶ **The glare of the public spotlight**
 - ▶ Concern for how one's data will be (mis-)interpreted

New Standards for Data Management



- **Widespread adoption of new (and some not-so new) standards**
 - ▶▶ **TCP (the Net) and HTTP (the Web)**
 - ▶▶ **WebDAV – Digital Authoring and Versioning**
 - ▶ HTTP-based digital content management protocol
 - ▶ Resource annotation via “properties”
 - ▶ Search interfaces (property searching)
 - ▶▶ **XML (Extensible Markup Language)**
 - ▶ HTML
 - ▶ Stronger mapping of data objects/models into data files
 - ▶ XML Schema – self-describing data file formats
 - ▶ XSLT (XML Stylesheet Language—Transformation)
 - Translation from XML to HTML
 - Translation from XML to application input formats
 - Progressive condensation of diverse XML formats into more widely adopted formats
 - ▶▶ **Dublin Core**
 - ▶ Digital Library Standard, using XML and WebDAV

New Tools for Data Management



- **Standards-driven technologies**

- ▶ **Apache Jakarta Project**

- ▶ **Java-based Open-Source web services and technologies**

- ▶ **Jakarta Tomcat – Servlet Container**

- ▶ **Jakarta Slide – WebDAV Implementation**

- ▶ **Digital data content management**

- ▶ **Jakarta Jetspeed – Web Portal Environment**

- ▶ **XML Technologies**

- ▶ **Jakarta Xerces, Crimson XML Parsers**

- **DOM (Document Object Model)**

- ▶ **Jakarta Xalan (XSLT Engine)**

- ▶ **SOAP (XML-based web application protocol)**

- ▶ **Jakarta Axis (SOAP implementation tools)**

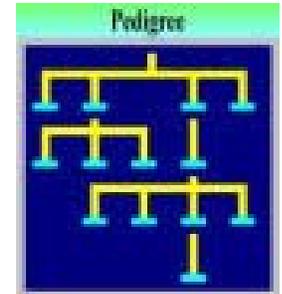
- ▶ **Castor (XML – Data Object “Binding”)**

- ▶ **...**

Scientific Pedigree: A Special Kind of Annotation



- Data pedigree or data provenance is a relationship which provides a “line of ancestors”.
- Pedigree allows for the categorization and tracing of the scientific data, and for the identification of the data’s ultimate origin.
- Pedigree metadata is associated with CMCS resources, and is browsable and searchable from the CMCS portal.
- Data is linked to projects, references, inputs, and outputs.

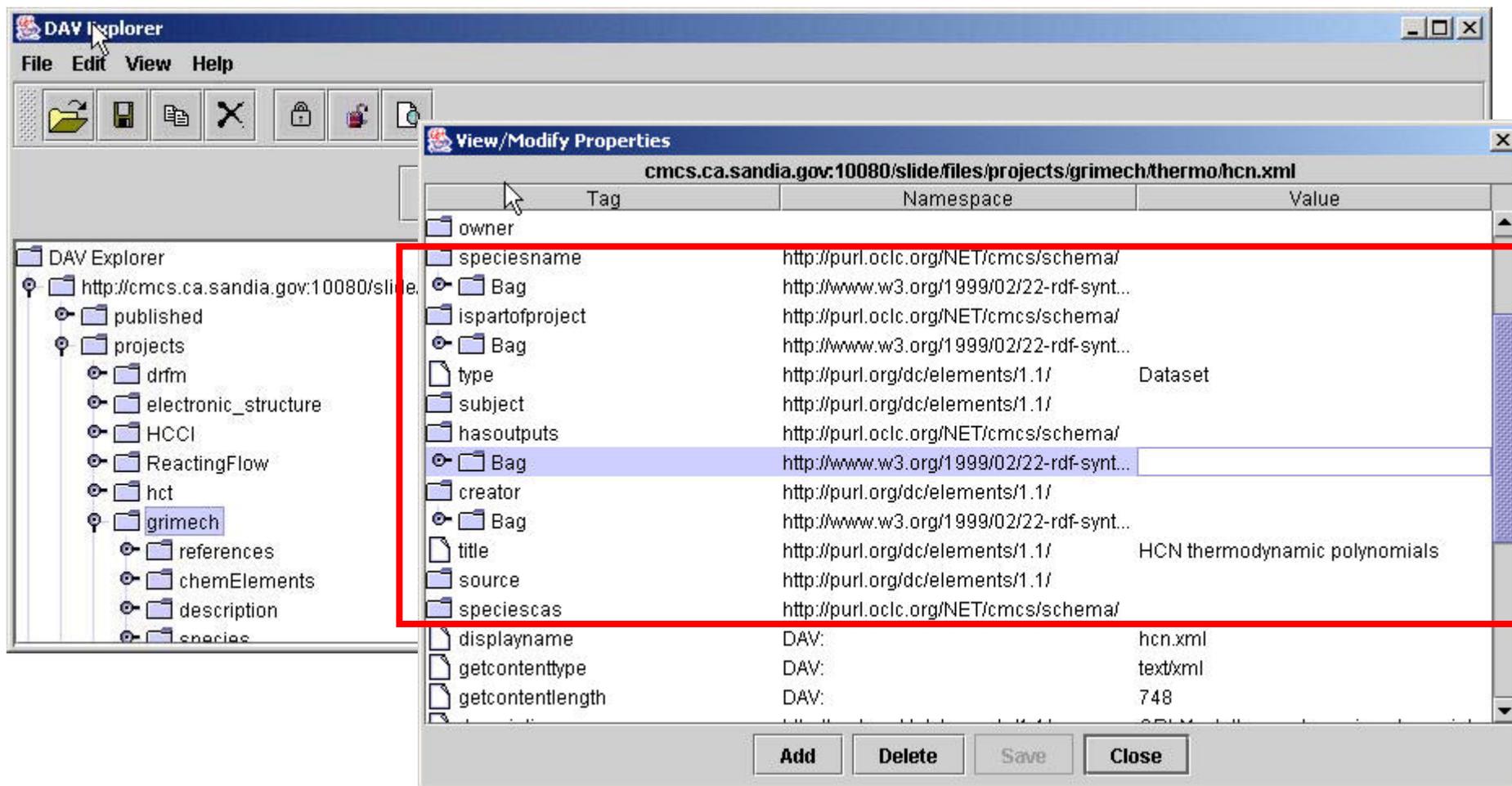


Scientific Pedigree is Critical to CMCS



- Can identify and trace data.
- Can trace scientific data to its ultimate origin, possibly across scales.
- Can track data to program versions, and hence, to program bugs reported for that version.
- Can retrieve information about the input parameters and configuration files.
- Can retrieve literature references which describe or reference this data.

Annotation Stored as WebDAV Properties

The screenshot shows the DAV Explorer window with a tree view of a WebDAV collection. The 'View/Modify Properties' dialog is open, displaying a table of properties for the file 'hcn.xml'. A red box highlights the XML annotations, which are keyword/value pairs in the form of namespace:tag.

Tag	Namespace	Value
owner		
speciesname	http://purl.oclc.org/NET/cmcs/schema/	
Bag	http://www.w3.org/1999/02/22-rdf-synt...	
ispartofproject	http://purl.oclc.org/NET/cmcs/schema/	
Bag	http://www.w3.org/1999/02/22-rdf-synt...	
type	http://purl.org/dc/elements/1.1/	Dataset
subject	http://purl.org/dc/elements/1.1/	
hasoutputs	http://purl.oclc.org/NET/cmcs/schema/	
Bag	http://www.w3.org/1999/02/22-rdf-synt...	
creator	http://purl.org/dc/elements/1.1/	
Bag	http://www.w3.org/1999/02/22-rdf-synt...	
title	http://purl.org/dc/elements/1.1/	HCN thermodynamic polynomials
source	http://purl.org/dc/elements/1.1/	
speciescas	http://purl.oclc.org/NET/cmcs/schema/	
displayname	DAV:	hcn.xml
getcontenttype	DAV:	text/xml
getcontentlength	DAV:	748

DAV property is a keyword/value pair: namespace:tag, and a well-formed XML value.

How Annotation is Populated in CMCS



- SAM Metadata Services Layer
 - ▶ When data is put into WebDAV, SAM causes XSLTs to be executed to extract metadata from XML files, based on MIME type.
 - ▶ Similarly, Binary File Descriptor (BFD) provides an interface to extract metadata from binary files.
- CMCS data management/pedigree API to facilitate insertion and modification of metadata, in the proper XML format.
 - ▶ Java code which allows software developers and scientists to easily write programs to add/edit metadata.
 - ▶ Scientists can use these APIs to integrate with existing or new chemical science applications.
 - ▶ Uses open source DAV and XML libraries.
- Any WebDAV client application
 - ▶ DAVExplorer: Java application
 - ▶ CMCSExplorer: Integrated in our CMCS portal



Nov 14, 2002 06:33 pm



Collaboratory for Multi-Scale Chemical Science

My Workspace

CMCS team

CMCS Dev

[Home](#)

[News](#)

CMCS Explorer

[Calendar](#)

[Resources](#)

[Team
Management](#)

[Edit Account](#)

[Logout](#)

Users Present

Carmen Pancarella

Address:

Folders

Search

Notify

Pedigree

Search

- [Basic](#)
- [Molecular](#)
- [Thermochemistry](#)
- [Kinetics](#)
- [Combustion](#)
- [Feature Analysis](#)
- [Species Dictionary](#)
- [Last Search Results](#)

Keywords

Electronic Structure
Feature Tracking

Properties

Enthalpy of Formation
Entropy
Specific Heat

CAS #



SEARCH IN PROGRESS - PLEASE WAIT...

Users can search CMCS data
repositories on a number of
different criteria.

Data Translation: XML and XSLT



- XML Stylesheet Language: Transformation (XSLT) is an easy-to-use, powerful tool for translating XML documents
 - ▶▶ HTML pages
 - ▶ Simple text and table view
 - ▶ Applets for interactive views
 - ▶▶ Application input formats
 - ▶▶ Automatic extraction of annotation
- Scientific Annotation Middleware (SAM) provides a way to provide XSLT translations to certain data formats in the CMCS portal

Translations: XSLT Data Viewer Registered With SAM



Collaboratory for Multi-Scale Chemical Science

Nov 14, 2002 12:50 pm

My Workspace
CMCS team
CMCS Dev

Address:

Folders
Search
Notify
Pedigree

Search

- [Basic](#)
- [Molecular](#)
- [Thermochemistry](#)
- [Kinetics](#)
- [Combustion](#)
- [Feature Analysis](#)
- [Species Dictionary](#)
- [Last Search Results](#)

Keywords

Electronic Structure

Feature Tracking

Properties

Entropy

Specific Heat

Heat Capacity

Edit:

Resource

[hen.xml](#)

Author: GRI-Mech team

Size: 748

This resource has a browser/viewer associated with it.

SAM Allows Data Viewer to be Launched



GRI-Mech Thermochemistry Viewer - Microsoft Internet Explorer

GRI-Mech Thermochemistry Viewer

T	Cp	S	delta-H_f
K	cal/mol K		kcal/mol
300	8.59	48.29	31.26
400	9.36	50.87	31.22
500	9.97	53.02	31.17
600	10.48	54.89	31.11
700	10.91	56.54	31.04
800	11.31	58.02	30.97
900	11.67	59.37	30.9
1000	12.01	60.62	30.83
1100	12.29	61.78	30.76

User is able to manipulate HCN data.

Begin by selecting a species from the drop-down menu on the left.

HCN names mol weight show xml file

Temperature 300:100:3000 K calculate units cal

(e.g. 1200 or 1200:100:2000)

[View DAV grimech directory](#) [Explanations](#)

Vision: Towards an Adaptive Informatics Infrastructure



- **Forward-looking approach to technology adoption**
 - ▶ **Long-term solution for data storage (WebDAV protocol)**
 - ▶ **Long-term solution for data annotation (WebDAV protocol)**
 - ▶ **Long-term solution for data accessibility (XML, XML Schema)**
 - ▶ **Application interfaces**
 - ▶ **Long-term solution for facile data transformation (XSLT)**
 - ▶ **Application interfaces**
 - ▶ **Visualization tools**
 - ▶ **Open-source applications and tool**

→ **Social contract of openness with the users**

Vision: Towards an Adaptive Informatics Infrastructure



- **Evolution, not revolution**

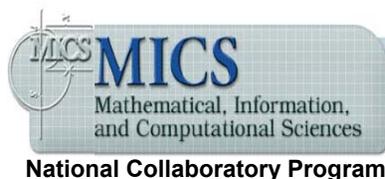
- ▶ Provide services that show clear value to data owners
- ▶ Easy adoption
 - ▶ Examples
 - ▶ Tutorials
- ▶ Visible rewards
 - ▶ Data discovery
 - ▶ Citations!

- **Scientific Publication of Data**

- ▶ Complementary to paper publication
- ▶ . . .



L. Rahn, C. Yang, C. Pancerella, J. Hewson, W. Koegler, D. Leahy, M. Lee, E. Walsh, *Sandia National Laboratories*; T. Windus, B. Didier, J. Myers, K. Schuchardt, E. Stephan, C. Lansing, E. Mendoza, *Pacific Northwest National Laboratory*; A. Wagner, B. Ruscic, M. Minkoff, G. von Laszewski, S. Bittner, R. Pinzon, S. Nijsure, K. Amin, B. Wang, *Argonne National Laboratory*; W. Pitz, *Lawrence Livermore National Laboratory*; D. Montoya, L. Xu, Y.-L. Ho, *Los Alamos National Laboratory*; T. Allison, *National Institute of Standards and Technology*; W. Green, *Massachusetts Institute of Technology*; M. Frenklach, *University of California at Berkeley*



SAM



Contact

email: rahn@sandia.gov

<http://cmcs.ca.sandia.gov>