Towards a Universal Physical Property Data Index

Peter J. Linstrom

National Institute of Standards and Technology

Summary

- Problem description
- What can potentially be indexed
- Previous and current efforts
- What is needed
- Obstacles to creating an index
- Possible (technical) solutions
- Future actions

Problem Description

- Most common question posed by users of the NIST Chemistry WebBook:

  Why don't you have property X for species Y?

- Second most common question:

  Where can I find this data?

Problem Description

- Generic answer:

  Look in the chemical literature (library, SciFinder, Internet sites, etc….)

- Drawbacks:
  - Potentially high cost (time and difficulty of access).
  - Does not leverage experience of archivists in commercial and public enterprises.

Problem Description

- From the data vendors perspective, the problem is reversed:

  How can our organization let potential customers know we have the data they need?

- Advertising, business directories, and search engines address general, not specific needs.

What Can Be Indexed

- Data of defined quantities for well-defined chemical species
- Need reliable, consistent identification chemical species.
  - Names, registry numbers, structure based IDs
- Need agreed upon definition of physical property
  - IUPAC definitions

What Can Be Indexed

- Data in various media types with cooperation and assistance of owner.

Media Type	Need Permission?	Need Assistance?
Public online sites		maybe
Private on-line systems	yes	yes
Private data products	yes	yes
Hardcopy publications	maybe	probably

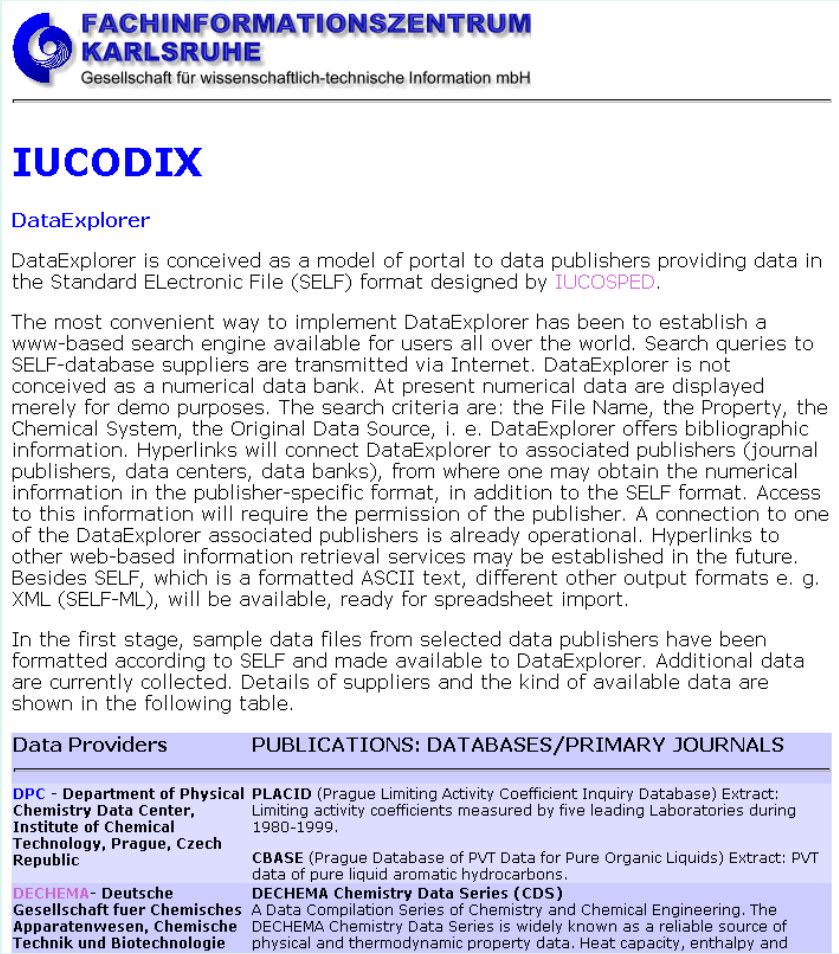
Previous and Current Efforts

- Internet search engines – Google, ChemIndustry, etc.
  - Text based only – can't identify chemical species and properties.
  - Can't index data in searchable databases
- Self contained systems – only support affiliated sources

Previous and Current Efforts

Data Explorer

- Henry V. Kehiaian
- IUCOSPED / ICSU Fiz-Karlshrue
- Physical property database with limited meta-data
- Gateway to literature
- Site currently closed



FACHINFORMATIONSZENTRUM KARLSRUHE
Gesellschaft für wissenschaftlich-technische Information mbH

IUCODIX

DataExplorer

DataExplorer is conceived as a model of portal to data publishers providing data in the Standard ELectronic File (SELF) format designed by IUCOSPED.

The most convenient way to implement DataExplorer has been to establish a www-based search engine available for users all over the world. Search queries to SELF-database suppliers are transmitted via Internet. DataExplorer is not conceived as a numerical data bank. At present numerical data are displayed merely for demo purposes. The search criteria are: the File Name, the Property, the Chemical System, the Original Data Source, i. e. DataExplorer offers bibliographic information. Hyperlinks will connect DataExplorer to associated publishers (journal publishers, data centers, data banks), from where one may obtain the numerical information in the publisher-specific format, in addition to the SELF format. Access to this information will require the permission of the publisher. A connection to one of the DataExplorer associated publishers is already operational. Hyperlinks to other web-based information retrieval services may be established in the future. Besides SELF, which is a formatted ASCII text, different other output formats e. g. XML (SELF-ML), will be available, ready for spreadsheet import.

In the first stage, sample data files from selected data publishers have been formatted according to SELF and made available to DataExplorer. Additional data are currently collected. Details of suppliers and the kind of available data are shown in the following table.

Data Providers	PUBLICATIONS: DATABASES/PRIMARY JOURNALS
DPC - **Department of Physical Chemistry Data Center, Institute of Chemical Technology, Prague, Czech Republic**	**PLACID** (Prague Limiting Activity Coefficient Inquiry Database) Extract: Limiting activity coefficients measured by five leading Laboratories during 1980-1999. **CBASE** (Prague Database of PVT Data for Pure Organic Liquids) Extract: PVT data of pure liquid aromatic hydrocarbons.
DECHEMA- **Deutsche Gesellschaft fuer Chemisches Apparatenwesen, Chemische Technik und Biotechnologie**	**DECHEMA Chemistry Data Series (CDS)** A Data Compilation Series of Chemistry and Chemical Engineering. The DECHEMA Chemistry Data Series is widely known as a reliable source of physical and thermodynamic property data. Heat capacity, enthalpy and

Previous and Current Efforts

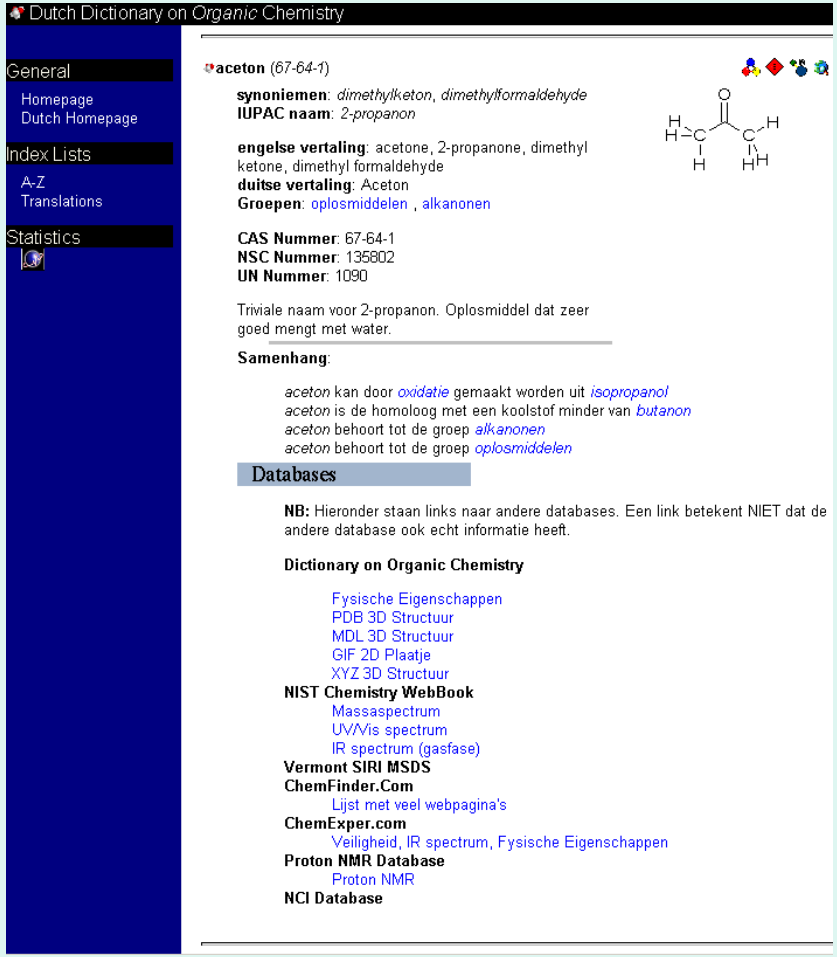
- Contributions from
  - Institute of Chemical Technology, Prague
  - DECHEMA
  - Glushko Thermocenter (RAS)
  - TRC (NIST)
  - Chemical Industry Press (Beijing)
  - ELDATA Journal
  - Interperiodica Publishing

Data Providers	PUBLICATIONS: DATABASES/PRIMARY JOURNALS
DPC - **Department of Physical Chemistry Data Center, Institute of Chemical Technology, Prague, Czech Republic**	**PLACID** (Prague Limiting Activity Coefficient Inquiry Database) Extract: Limiting activity coefficients measured by five leading Laboratories during 1980-1999. **CBASE** (Prague Database of PVT Data for Pure Organic Liquids) Extract: PVT data of pure liquid aromatic hydrocarbons.
DECHEMA- **Deutsche Gesellschaft fuer Chemisches Apparatenwesen, Chemische Technik und Biotechnologie e.V. (German Society for Chemical Apparatus, Chemical Engineering, and Biotechnology)**	**DECHEMA Chemistry Data Series (CDS)** A Data Compilation Series of Chemistry and Chemical Engineering. The DECHEMA Chemistry Data Series is widely known as a reliable source of physical and thermodynamic property data. Heat capacity, enthalpy and entropy, phase equlibria, PVT, interfacial tension and transport property data have been emphasized. Traditionally the series treats systems of relatively small, commercially important organic molecules in the fluid phase. The main purpose of the CDS is to provide chemists and engineers with those data needed for process design and optimization. Today the series exceeds fifty volumes with twenty-five thousand pages and several thousand compounds. The data compiled in CDS is part of the DETHERM database. Sample parts of the following selected volumes will be made available for the Data Explorer. **Critical Data of Pure Substances (Volume II).** Extract: The critical data for 20 selected compounds. **Vapor-Liquid Equilibria for Mixtures of Low Boiling Substances (Volume VI).** Vapor-liquid equilibrium data for mixtures consisting of two or more low-boiling components. Extract: Selected data sets for 20 selected compounds or mixtures. **Electrolyte Data Collection (Volume XII).** This is a collection of properties of electrolyte solutions. Included are values of solvent properties and distance parameters, conductivity's of dilute and concentrated solutions, transference numbers, limiting ionic conductivities, dielectric properties and viscosities. The studies published are in conformity with the recommendations of the IUPAC. Extract: Selected data sets for 20 selected compounds or mixtures.
GTCRAS - **The Glushko Thermocenter of the Russian Academy of Sciences**	**THERMEL** Thermodynamic Properties of Chemical Elements and Simple Mono-, Di-, and Polyatomic Substances in Solid, Liquid, and Gaseous States (including gaseous ions) up to 20000 K. Extract: Heat capacities, Gibbs energies, entropies, enthalpy changes, and heats of formation, sublimation, and phase transistion for 20 selected elements. **HTEDATA (HTEQUIDATA)** Primary Experimental Data on High Temperature Equilibrium Constants for Chemical Reactions (involving mainly inorganic substances). Extract: Experimental data on equilibrium constants of 100 chemical reactions for iron compounds. **UODB** Thermodynamic data for 23 U and U-O solid, liquid, and gaseous (neutral or ionic) species. Extract: Heat capacities, Gibbs energies, entropies, enthalpy changes, and equilibrium constants up to 12.000 K for 10 selected compounds. **DTCS** Database on Thermal Constants of Substances. Critically selected enthalpies and Gibbs energies of formation, enthalpy increments, entropies and heat capacities at 298.15 K, enthalpies of formation and dissociation energies at 0 K, temperatures and enthalpies of phase transition, crystallographic and critical parameters, ionization potentials and electron

Previous and Current Efforts

DADML

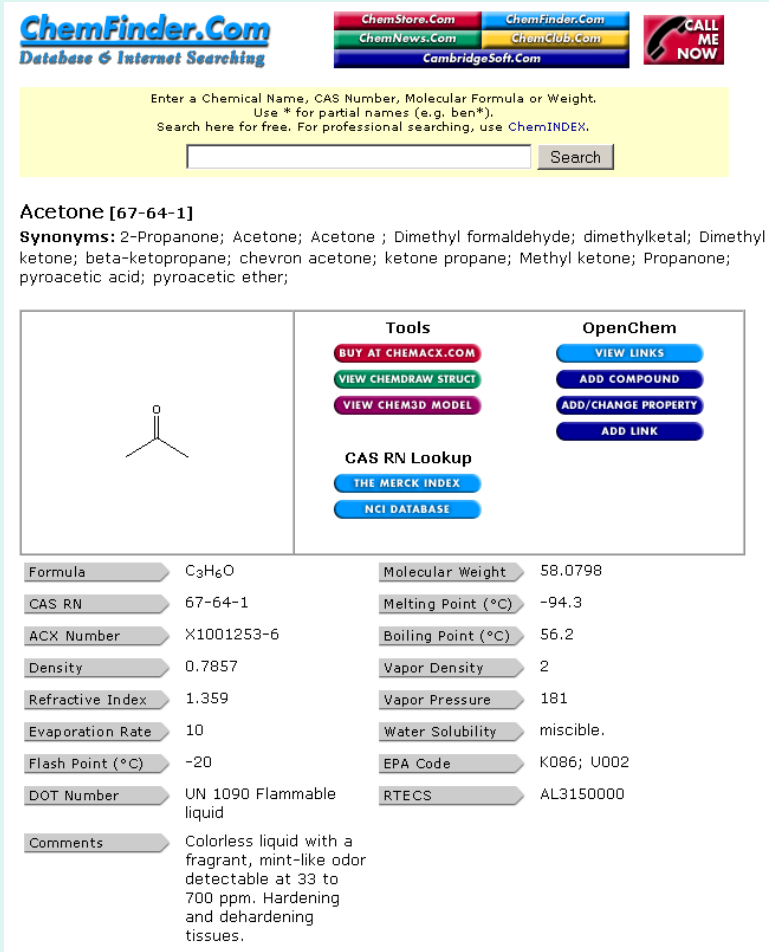
- E.L. Willighagen
- “Definition of how to Access a Database” Markup Language
- Tool for WWW molecule data access
- Dutch Dictionary on Organic Chemistry



Previous and Current Efforts

Chemfinder

- http://chemfinder.cambridgesoft.com/
- Cambridgesoft Corporation
- Links to various sites and some physical property data



Previous and Current Efforts

- Free and subscription services
- Access to various databases (subscription model)

More information about the chemical is available in these categories:

Biochemistry | Chemical Online Order | Health | Misc
MSDS | Pesticides/Herbicides | Physical Properties | Regulations
Structures | Usage

Biochemistry (2)

Biocatalysis/Biodegradation Database
Information about this particular compound
Ligand Chemical Database for Enzyme Reactions
Information about this particular compound

Chemical Online Order (1)

Available Chemicals Exchange
Information about this particular compound

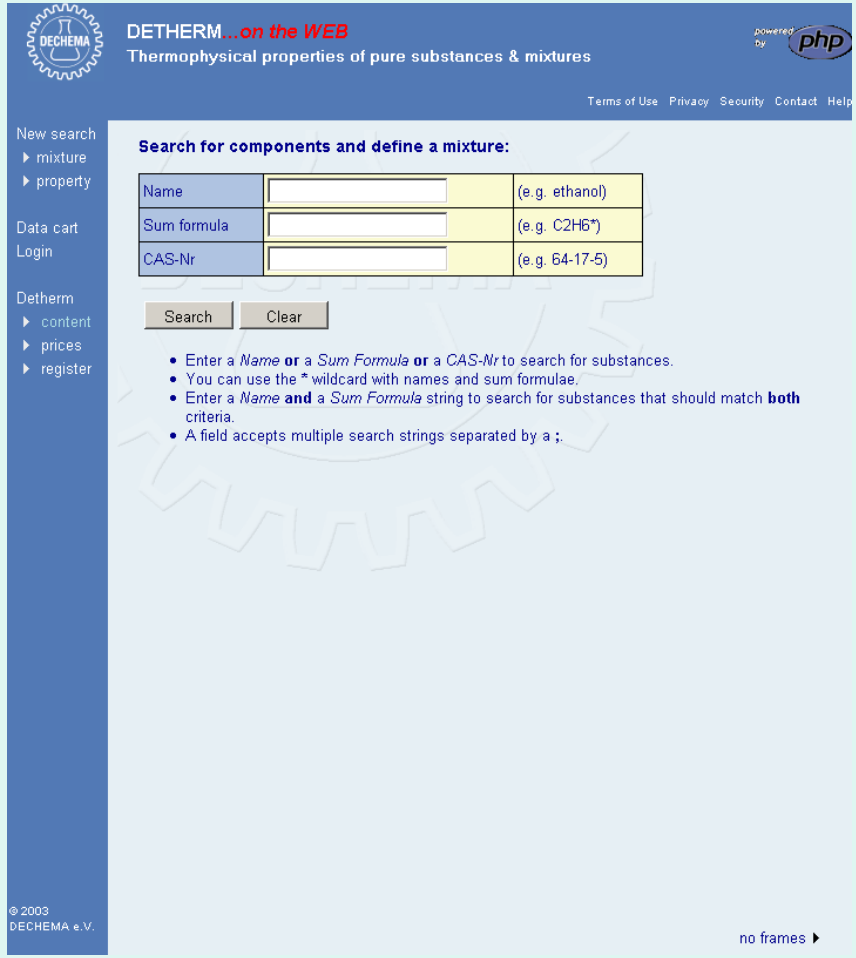
Health (17)

8(e) TRIAGE Chemical Studies Database
ATSDR Internet HazDat Site Contaminant Query
Information about this particular compound
ATSDR ToxFAQs
Information about this particular compound
Australian Atmospheric Exposure Standards
Information about this particular compound
CHEMICALS STUDIED through NIEHS's Reproductive Toxicology Group
Drugs and Toxic Chemicals in Breast Milk
Hazardous Chemicals Database at the University of Akron
Information about this particular compound
Idaho Toxic and Hazardous Substances
International Chemical Safety Cards
Information about this particular compound
International Toxicity Estimates for Risk
Information about this particular compound
Lab Chem Safety Summaries
Information about this particular compound
Minimal Risk Levels for Hazardous Substances
North American Emergency Response Guidebook 1996 (NAERG96)

Previous and Current Efforts

DTHERM

- http://i-systems.dechema.de/detherm/
- DECHEMA
- Index and access to data from a number of sources



Previous and Current Efforts

- Data from
  - DDBST GmbH
  - University of Regensburg
  - FIZ Chemie
  - Institut for Chemical Technic
  - Dow
  - DECHMA

DETHERM...on the WEB

Thermophysical properties of pure substances & mixtures

powered by php

Terms of Use Privacy Security Contact Help

New search
- mixture
- property

Data cart
Login

Detherm
- content
- prices
- register

Search results: 63 hits

page 1 /5

Lines		USD	property description	substance/mixture	year	temperature	pressure
☐	8	58.32	Molar heat capacity (cP)	2-propanone	1998	298 K- 333 K	-
☐	5	29.16	heat capacity	2-propanone	1996	372 K- 515 K	101325 Pa
☐	1	14.00	heat capacity	2-propanone	1995	515 K	101300 Pa
☐	150	139.96	heat capacity	2-propanone	1995	13 K- 178 K	-
☐	51	139.96	heat capacity	2-propanone	1995	179 K- 300 K	-
☐	1	14.00	Polynomial coeffients for the heat capacity. T= 273 ...1200 K	2-propanone	1995	-	-
☐	11	58.32	heat capacity	2-propanone	1995	213 K- 413 K	-
☐	50	139.96	Molar heat capacity (cP)	2-propanone	1995	181 K- 300 K	-
☐	1	14.00	heat capacity	2-propanone	1993	298 K	-
☐	1	14.00	Molar heat capacity (cP)	2-propanone	1993	298 K	-
☐	50	139.96	compressibility,cubic expansion coefficient,heat capacity	2-propanone	1991	278 K- 323 K	1 bar- 3750 bar
☐	1	14.00	density,heat capacity	2-propanone	1991	298 K	-
☐	1	14.00	Molar heat capacity (cP)	2-propanone	1991	298 K	-
☐	5	29.16	Molar heat capacity (cP)	2-propanone	1991	278 K- 323 K	100000 Pa
☐	45	139.96	Molar heat capacity (cP)	2-propanone	1991	278 K- 323 K	200 bar- 3750 bar

Go to page: 1 2 3 4 5

< Back | Add selection | Remove selection | Next >

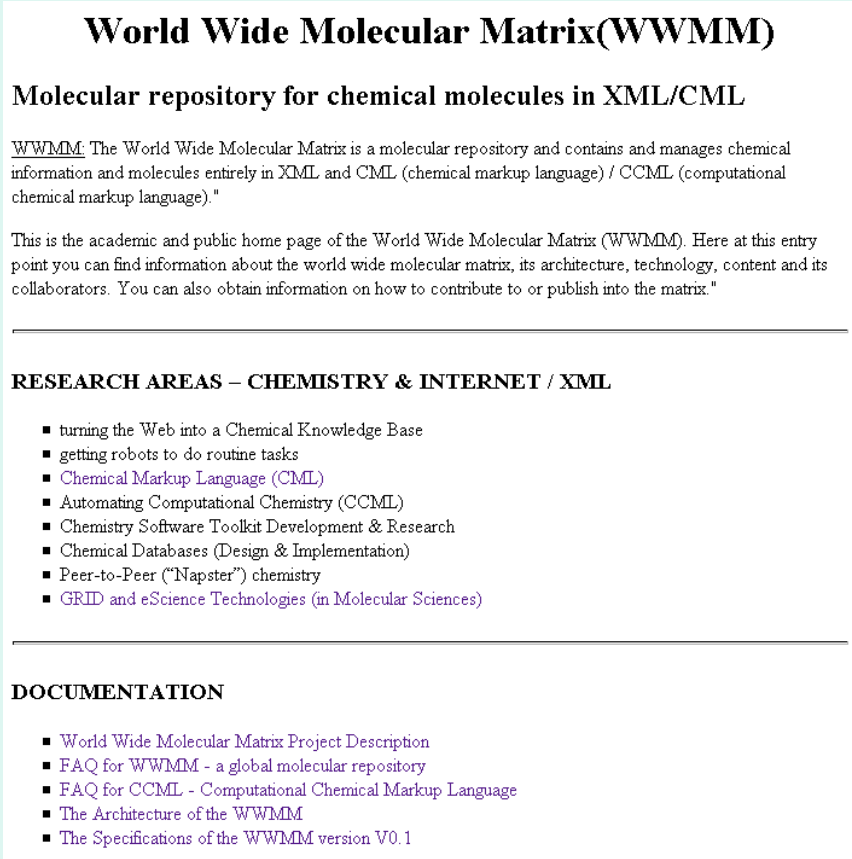
Select data sets and press *Add/Remove selection* on each result page to define your selection. Then press *Next* to add all selected items to your data cart

© 2003 DECHEMA e.V.

Previous and Current Efforts

World-Wide Molecular Matrix

- http://wwmm.ch.cam.ac.uk/
- Peter Murray-Rust, Robert C. Glen, Jürgen Harter, Yong Zhang



World Wide Molecular Matrix(WWMM)

Molecular repository for chemical molecules in XML/CML

WWMM: The World Wide Molecular Matrix is a molecular repository and contains and manages chemical information and molecules entirely in XML and CML (chemical markup language) / CCML (computational chemical markup language)."

This is the academic and public home page of the World Wide Molecular Matrix (WWMM). Here at this entry point you can find information about the world wide molecular matrix, its architecture, technology, content and its collaborators. You can also obtain information on how to contribute to or publish into the matrix."

RESEARCH AREAS – CHEMISTRY & INTERNET / XML

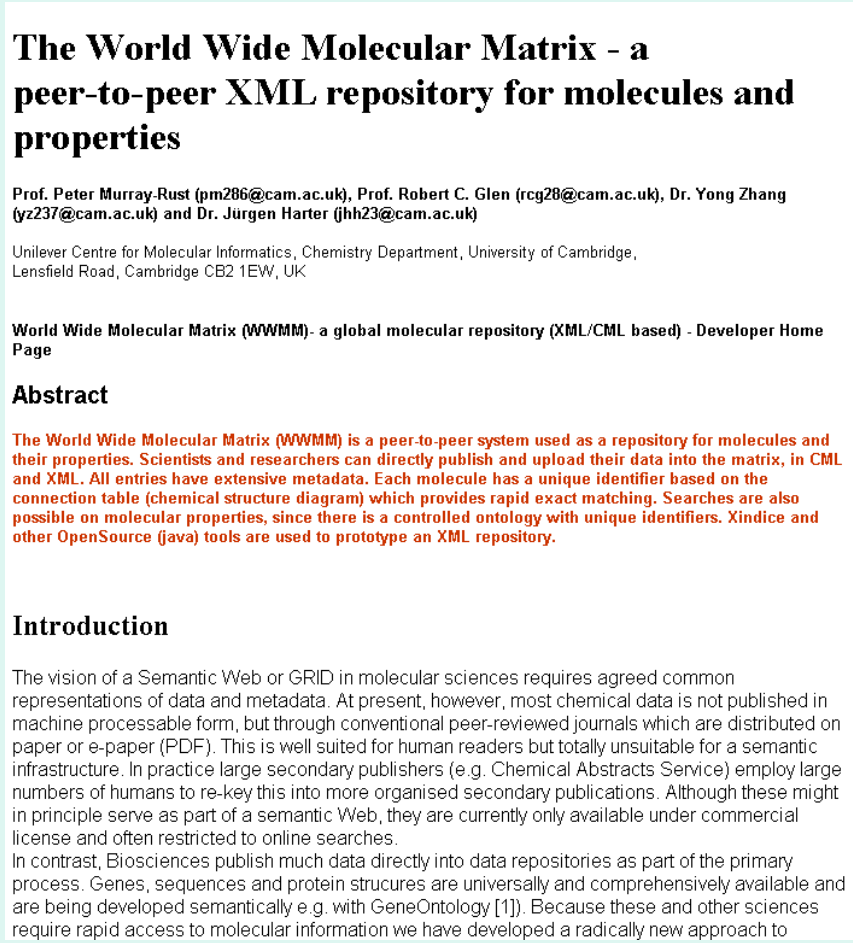
- turning the Web into a Chemical Knowledge Base
- getting robots to do routine tasks
- Chemical Markup Language (CML)
- Automating Computational Chemistry (CCML)
- Chemistry Software Toolkit Development & Research
- Chemical Databases (Design & Implementation)
- Peer-to-Peer ("Napster") chemistry
- GRID and eScience Technologies (in Molecular Sciences)

DOCUMENTATION

- World Wide Molecular Matrix Project Description
- FAQ for WWMM - a global molecular repository
- FAQ for CCML - Computational Chemical Markup Language
- The Architecture of the WWMM
- The Specifications of the WWMM version V0.1

Previous and Current Efforts

- University of Cambridge
- Peer to peer system
- Uses CML and related languages
- Repository for molecules and their properties.



The World Wide Molecular Matrix - a peer-to-peer XML repository for molecules and properties

Prof. Peter Murray-Rust (pm286@cam.ac.uk), Prof. Robert C. Glen (rcg28@cam.ac.uk), Dr. Yong Zhang (yz237@cam.ac.uk) and Dr. Jürgen Harter (jhh23@cam.ac.uk)

Unilever Centre for Molecular Informatics, Chemistry Department, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

World Wide Molecular Matrix (WWMM)- a global molecular repository (XML/CML based) - Developer Home Page

Abstract

The World Wide Molecular Matrix (WWMM) is a peer-to-peer system used as a repository for molecules and their properties. Scientists and researchers can directly publish and upload their data into the matrix, in CML and XML. All entries have extensive metadata. Each molecule has a unique identifier based on the connection table (chemical structure diagram) which provides rapid exact matching. Searches are also possible on molecular properties, since there is a controlled ontology with unique identifiers. Xindice and other OpenSource (java) tools are used to prototype an XML repository.

Introduction

The vision of a Semantic Web or GRID in molecular sciences requires agreed common representations of data and metadata. At present, however, most chemical data is not published in machine processable form, but through conventional peer-reviewed journals which are distributed on paper or e-paper (PDF). This is well suited for human readers but totally unsuitable for a semantic infrastructure. In practice large secondary publishers (e.g. Chemical Abstracts Service) employ large numbers of humans to re-key this into more organised secondary publications. Although these might in principle serve as part of a semantic Web, they are currently only available under commercial license and often restricted to online searches.
In contrast, Biosciences publish much data directly into data repositories as part of the primary process. Genes, sequences and protein strucures are universally and comprehensively available and are being developed semantically e.g. with GeneOntology [1]). Because these and other sciences require rapid access to molecular information we have developed a radically new approach to

What is Needed

- Agreement that a centralized index benefits both users and distributors.
- Central resource shared by many
  - Lists both public and private sector resources
  - Provides results in a non-biased manner
- Protocols for identifying chemical species and data types

Obstacles to Creating an Index

- Database structures are not consistent
  - Different ways of expressing the identity of a chemical species
    - Different naming conventions
    - Registry numbers not available or inconsistent
  - Different conventions for recording data types.
    - Inconsistent state information – e.g., boiling point as a function of P vs. normal boiling point
    - Granularity issues – records of individual properties may not be available

Obstacles to Creating an Index

- Some database structures are difficult to index
  - Hardcopy publications
  - Databases / documents which are not organized by chemical species or property type.
- It is difficult to maintain an updated index
  - In the on-line era resources can rapidly change content

Possible (Technical) Solutions

Centralized solution

- Develop agreed upon protocols for providing index data and updates to a central site
- Provide open access to index on the site
- Central site maintains chemical identification database (a composite of all the chemical identification data provided)

Possible (Technical) Solutions

Distributed solution

- Develop agreed upon protocols for querying resources for data
- Publishers / distributors implement protocols on their servers
- Multiple indices can be constructed through meta-searches (one site queries several other sites)

Possible (Technical) Solutions

- Chemical Identification
  - Need a database linking names, structures, registry numbers
  - Will organizations release this information if it provides them with a competitive advantage?
- Property Identification
  - IUPAC Gold book
  - IUPAC project on data definitions in XML

Future Actions

- NIST is interested in working with others to create an index of physical property data
  - Our “customers” have been asking for this information.
  - We are restricted in our interactions with commercial ventures – we can’t favor any specific organization.
  - If you or your organization is interested, please contact us.

peter.linstrom@nist.gov