Extraction and Application of Environmentally Relevant Chemical Information from the ThermoML Archive

Ekstrakcja i Użycie chemicznych Informacji odnoszacych się do Środowiska z Archiwum ThermoML

Axel Drefahl
axeleratio@yahoo.com

Presentation at the ENVIROINFO 2007 in Warsaw, Poland, on September 12, 2007

Overview

- **ThermoML quick tour**
- **Chemical identification**
- **Chemical Property Viewer (CPV)**
- **ThermoML compounds and properties of environmental interest**
- **Property estimation methods: Modeling with ThermoML data**
- **Future developments and applications**

ThermoML is an XML application

XML = eXtensible Markup Language

ThermoML = Thermodynamic Markup Language to capture and exchange thermodynamic data

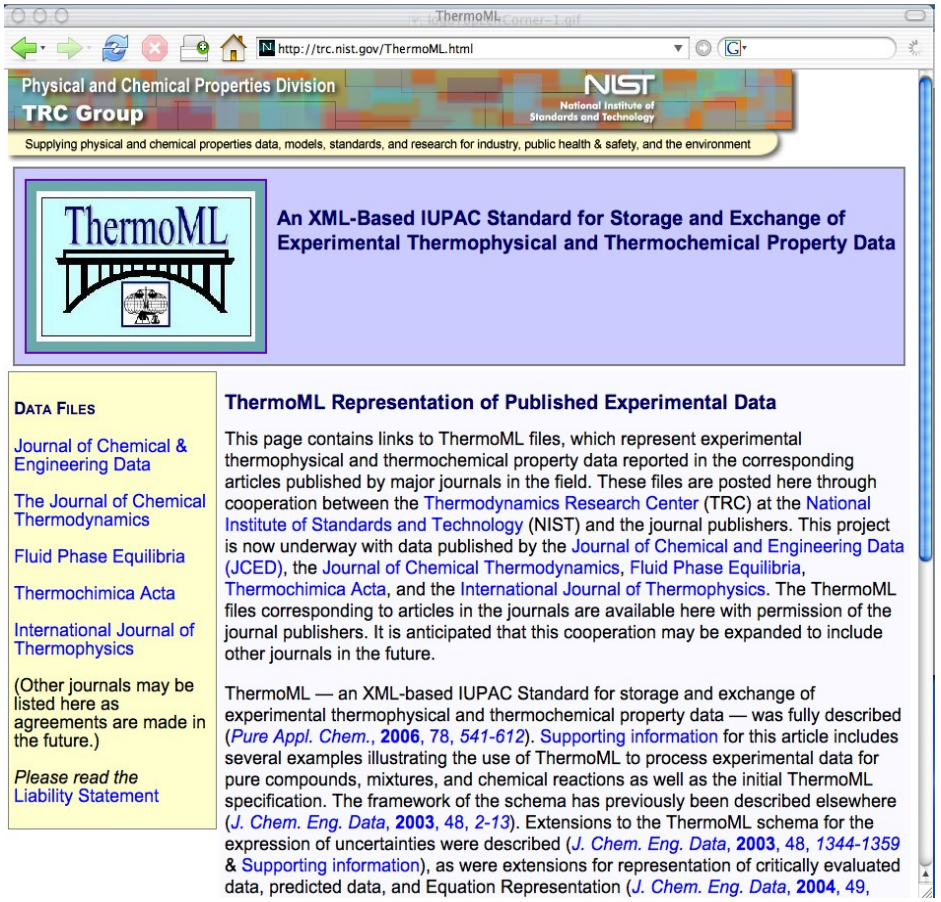
Other XML applications of interest in science and environmental chemistry:

- **MathML to represent and apply equations, functions, etc.**
- **CML to encode molecular structure**
- **CDX for Central Data Exchange of environmental information at US-EPA**

To explore XML applications and initiatives go to:
http://xml.coverpages.org/xmlApplications.html

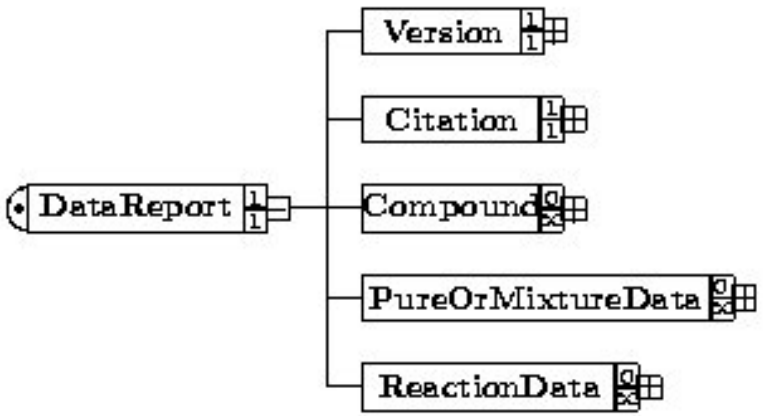
ThermoML Archive Portal

http://trc.nist.gov/ThermoML.html



- General Information
- Links to publications about ThermoML
- Links to ThermoML files with chemical property data of articles from five journals
- Schema: trc.nist.gov/ThermoML.xsd

ThermoML root and first layer nodes



- Exactly one **<Version>** and one **<Citation>** subtree
- None to many **<Compound>**, **<PureOrMixtureData>** and **<ReactionData>** subtrees

Programming approaches
using the Document Object Model (DOM)

Off-line scripting

Python, XML access via `xml.dom.minidom` module

Python scripts implemented for

- **Inspection of ThermoML files**
- **Extraction of data**
- **XML-to-XML conversions**

(chemical dictionary generation)

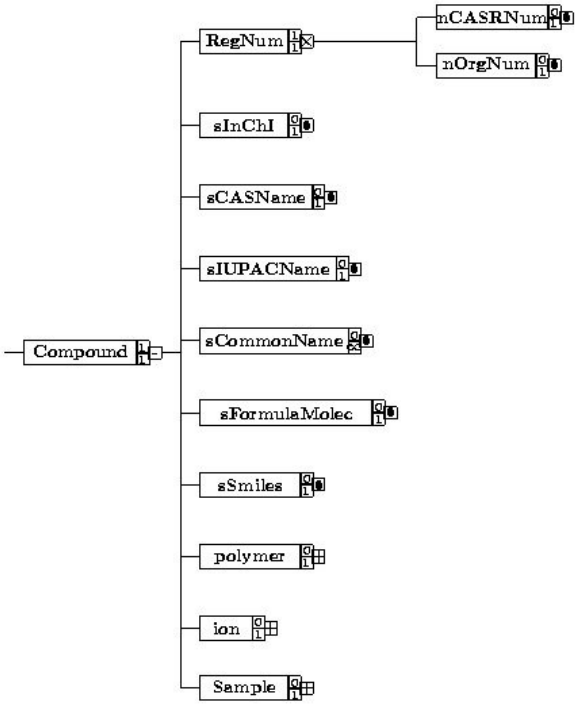
Web design

JavaScript for browser-side tasks, DOM functions slow for huge XML files

PHP for server-side tasks including dictionary browsing and generation of result pages

(`XMLReader` extension for parsing huge XML documents)

Compound Block
for chemical identification



- **Cross-referencing: <nOrgNum>, <nCASRNum>**
- **Name(s): one or more <sCommonName>**
- **Chemical composition: <sFormulaMolec>**
- **Molecular structure: <sInChI>, <sSmiles>**
- **Others: <polymer>, <ion>, <Sample>**

Inspection of currently available ThermoML files shows:

- **Cross-referencing within a file mostly done through <nCASRNum>**
- **Typical nodes used for compound identification: <sCommonName> and <sFormulaMolec>**
- **Structural information not (yet) available from within ThermoML files**

Scope of ThermoML Archive

Total number of ThermoML Files:

1,568 (Feb'07)

1,737 (July'07)

1,016 (with pure compound data for over 40 different properties)

Counting property data nodes:

17,226 (total, Feb'07)

7,764 (for pure compounds, Feb'07)

8,277 (for pure compounds, July'07)

Counting compounds (July'07):

1,113 (organics by name)

58 (inorganics by name)

1,154 (distinct CASRNs)

716 (distinct molecular formulae)

Most frequent properties:

Vapor or sublimation pressure

Mass density

Refractive index (Na-D-line)

Viscosity

Molar heat capacity at constant *P*

Conversion of ThermoML files into customized XML files

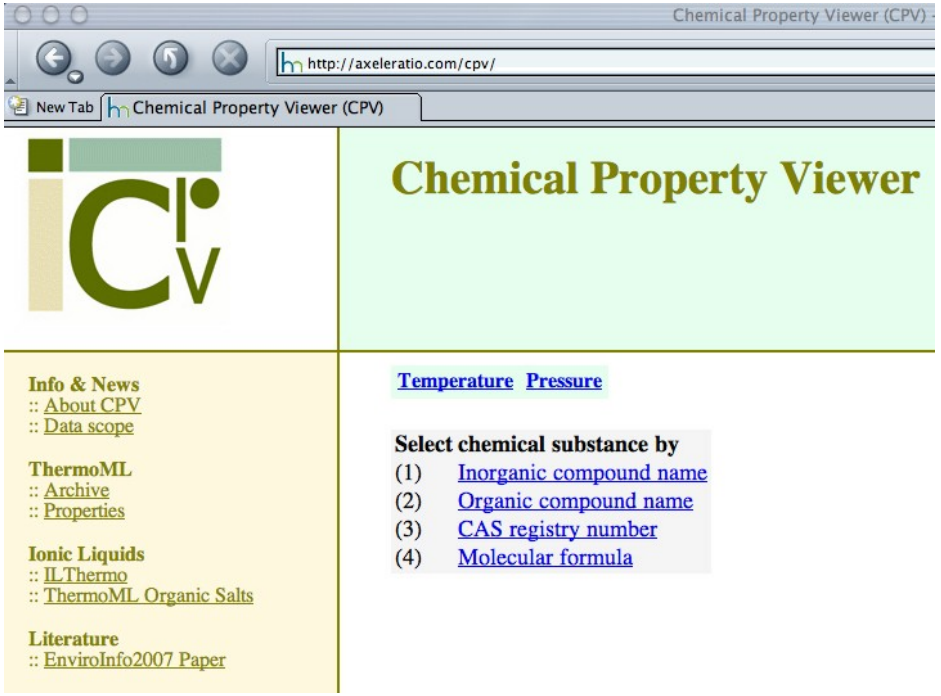
The ThemoML Archive is organized by article. Location of chemicals and properties requires looping over all archive file.

Mark-up provision for numerical accuracy, chemical purity, and exact physical state gives strength to ThermoML, but such info not needed for every task.

- **Generation of chemical dictionaries for look-up by name, formula, and CASRN**
- **Generation of lean versions of ThermoML Archive to efficiently retrieve chemical systems (pure, binary, ternary) and properties of interest**

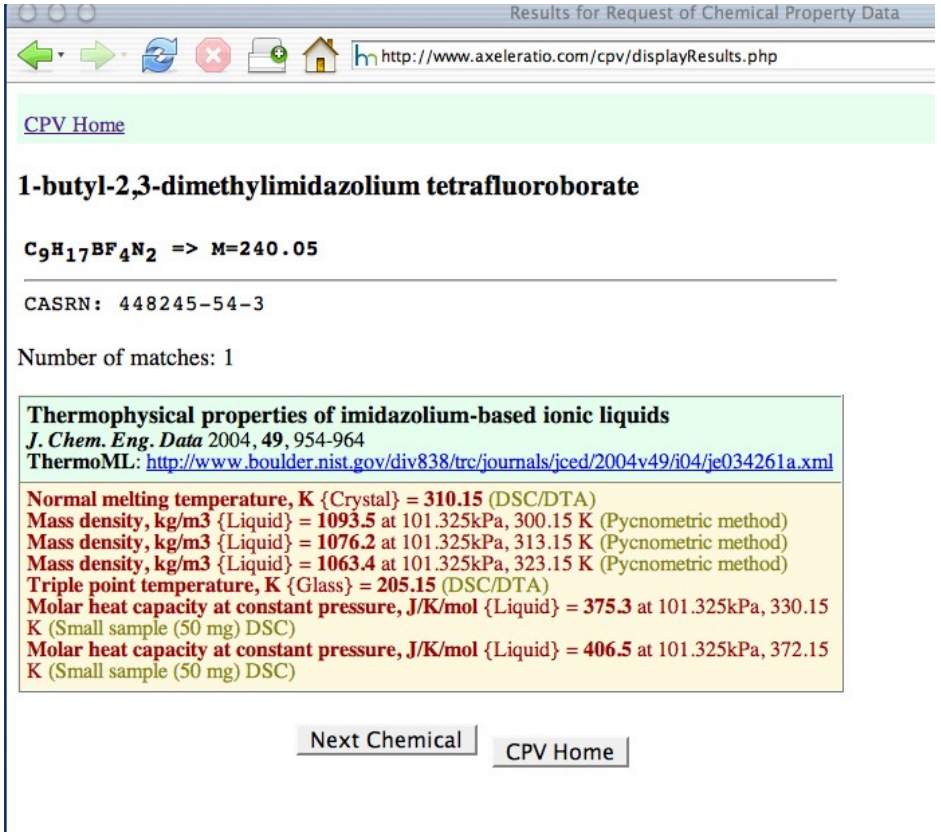
Chemical Property Viewer (CPV)

www.axeleratio.com/cpv



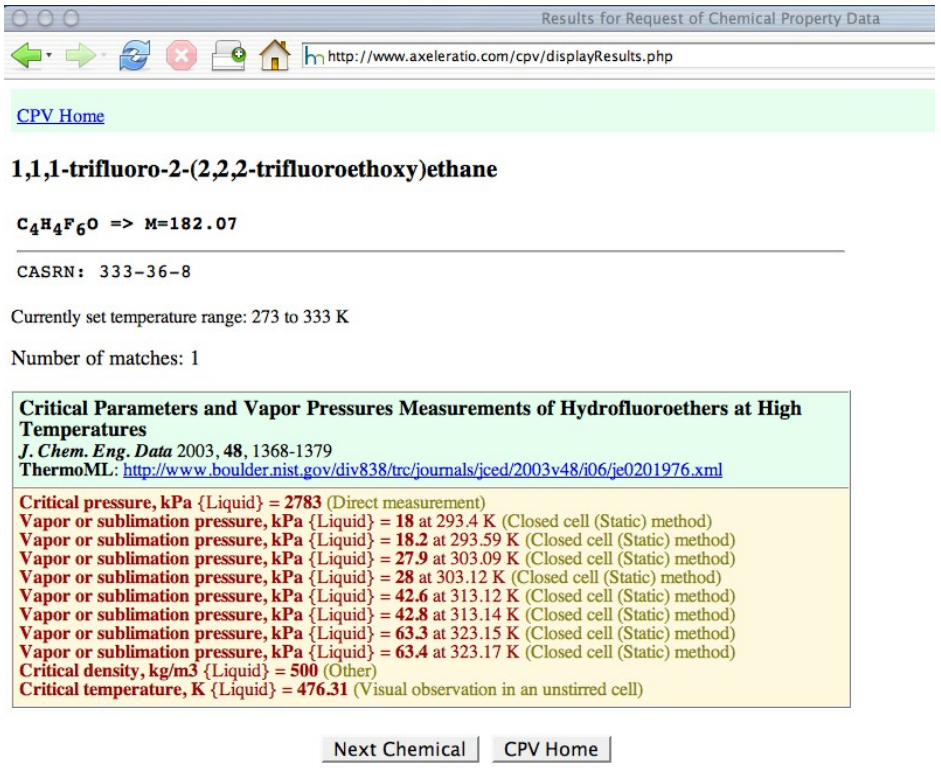
- **Define temperature and pressure range**
- **Select by name for inorganic (non-carbon) compound**
- **Select by name for organic (carbon-containing) compound**
- **Select by CASRN**
- **Select by molecular formula**

Display of CPV results



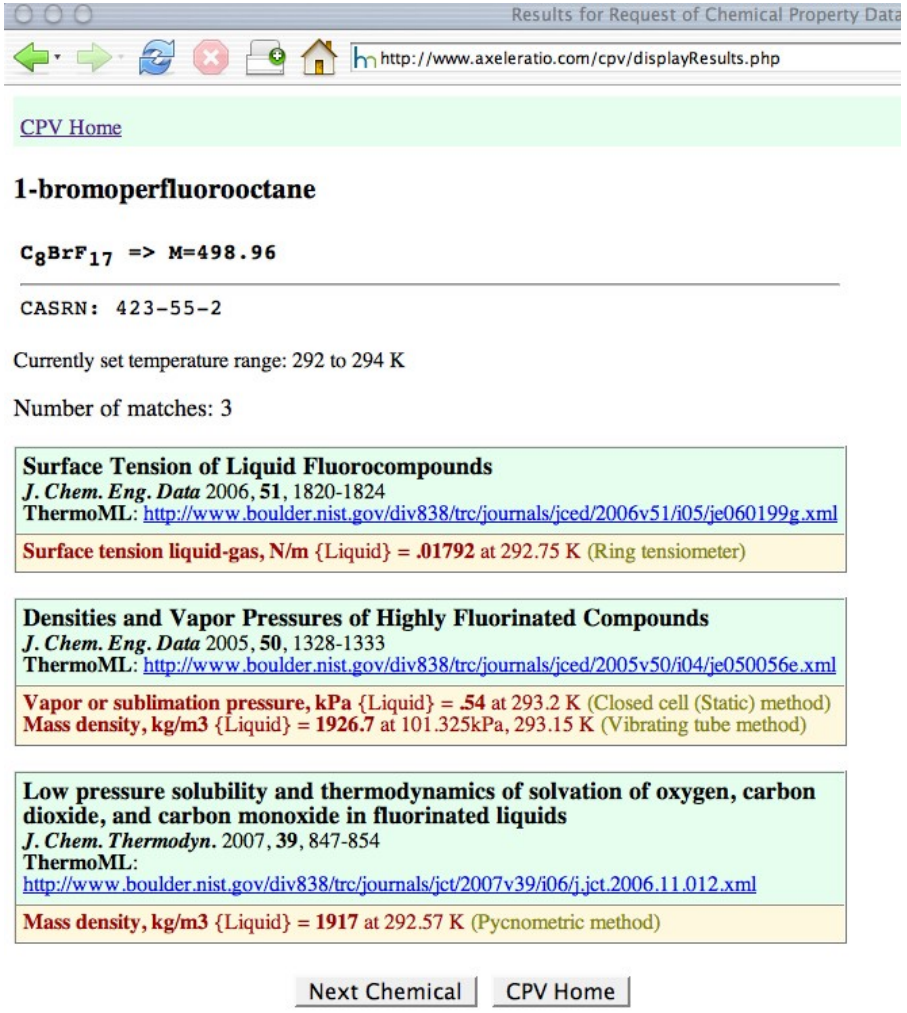
- 1 Match, referring to 1 article
- Link to ThermoML file
- Property data given line-by-line
- Some properties at different temperatures

CPV results with user-defined temperature range



- Default setting: data at any temperature (*T*) and pressure (*P*)
- User option: to define lower and upper limits for *T* and *P*

CPV results including multiple matches



- 3 Matches
- Narrow temperature range
- Data comparison: mass density occurs in 2 matches at similar temperatures

Water H_2O 7732-18-5

Current number of matches: 61 articles

Almost all articles report pure water properties in context with properties of aqueous solutions and (water + chemical) systems.

Typical (and exotic) *T*, *P* Ranges

Temperature range: 273 to 400 K
(hexagonal ice: 0.5 to 38 K)
Pressure range: 100 to 3,500,00 kPa
Many properties at 101,325 kPa

- Mass density
- Vapor pressure
- Viscosity
- Surface tension
- Molar heat capacity
- Thermal conductivity

(Water + Chemical) Systems
for over 400 chemicals

- **Mass density, viscosity, surface tension**
- **Molar enthalpy of solution**
- **Activity and diffusion coefficients**
- **Henry's Law constants**

A list of all chemicals and available properties with ThermoML links can be found at

www.axeleratio.com/EnviroInfo2007/AquBinSys.html

Properties of Ionic Liquids (ILs)

IUPAC Ionic Liquids Database (ILThermo)

provides forms to look up data and literature. ilthermo.boulder.nist.gov/ILThermo/mainmenu.uix

ILThermo supports search by

- Literature
- Property
- Ions
- Ionic Liquids

but no XML access.

ThermoML Archive

currently contains over 50 files with data on organic salts including pure ILs and mixtures. www.axeleratio.com/EnviroInfo2007/OrganicSalts.html

Most frequent properties:

- triple, melting, boiling temp.
- vapor or sublimation pressure (!)
- density, viscosity, surf. tension
- molar heat capacity
- thermal, electrical conductivity

Design and Testing of Chemical Property Estimation Models

Broad range (*T*, *P*, and molecular-structure-wise) of ThermoML data available for

- **theoretical modeling (e.g., corresponding states principle using T_c , P_c , V_c)**
- **(semi)empirical modeling (e.g., QPPR, QSPR, GCM, ANN, molecular similarity)**
- **molecular descriptor calculation**
- **generation of training and test sets**

ThermoML provides a clear, well-defined interface to select and evaluate data within the request context.

Example: **Polarizability**

www.axeleratio.com/EnviroInfo2007/CompareAlphas.pdf

- **Experimental data from ThermoML Archive: Mass Density, Refractive Index (Na-D line) at T/K = 293.2, 298.2**
- **Atom Additivity (AA) approach** (Bosque and Sales: *J. Chem. Inf. Comput. Sci.* **2003**, *42*, 1154-1163)

Experimental polarizability:

$$\alpha_{\mathrm{LL}} = 0.3964 \frac{n_{\mathrm{D}}^2 - 1}{n_{\mathrm{D}}^2 + 2} \frac{M}{\rho}$$

based on Lorentz − Lorenz (LL) equation.

Estimated polarizability:

$$\alpha_{\mathrm{AA}} = 0.32 + \Sigma\, c_{\mathrm{A}}\, N_{\mathrm{A}}$$

where

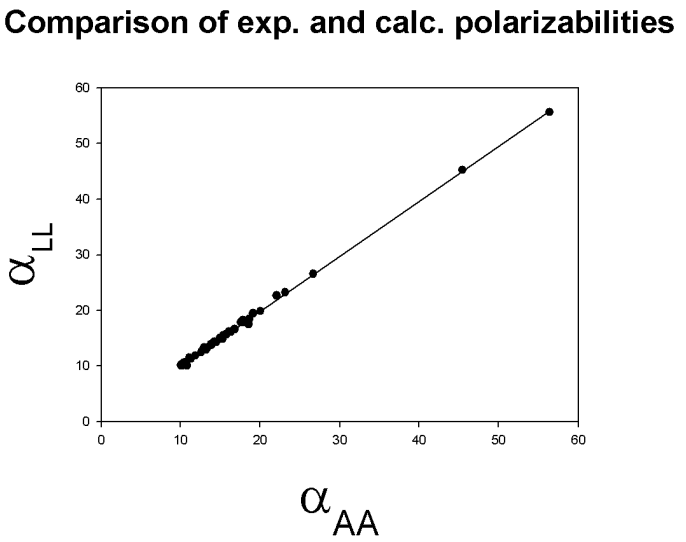
c_A = contribution for atom A

N_{A} = number of atoms A per molecule

Results: **Polarizability**

www.axeleratio.com/EnviroInfo2007/CompareAlphas.pdf

- **64 compounds with data that were not part of the original work by Bosque and Sales could be extracted from ThermML Archive**
- **Excellent correlation between exp. and est. polarizabilities at 298.2K: *R* = 0.9996**



BioaccuML? EcotoxML? FirehazML? NanomatML?

The success of ThermoML encourages XML presentation of other chemical information.

Are publishers of environmental journals/literature ready?
What is the current status?

Of interest:

Parr (2007): Open Sourcing Ecological Data.
BioScience, 57 (No. 4), pp. 309-310.

Swan(2007): Open Access and the Progress in Science.
Am. Sci. 95 (No.3), pp. 197-199.

Customization of Chemical Property Viewer

- **Chemical identification based on molecular structure and substructure**
- **Data interpolation at given *T* and *P***
- **Interface for binary and ternary chemical systems**
- **Data fitting**
- **Design of property estimation methods (correlations, molecular similarity, ...)**

Conclusions

- **ThermoML supports open access screening, filtering, and comparing of chemical information.**
- **The Chemical Property Viewer (CPV) provides quick “first-glance” access to chemical property data and associated files/publications.**
- **Chemical data critical to environmental modeling is abstracted with ThermoML and extractable as context demands.**

Future Developments

may include

- **Integration of ThermoML data with environmental modeling tools, chemical life-cycle assessment, and alternative materials (re)search.**
- **Probing ThermoML property + reactivity data in predictive models for biodegradation, synergistic or antagonistic environmental behavior and solar detoxification.**

Ongoing ThermoML activities:

- **Updating the Chemical Property Viewer with data from the latest publications**
- **Adding functionality to the Property Viewer in concert with advancing research goals**

This slide show can be revisited at

www.axeleratio.com/EnviroInfo2007/slides.pdf