

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

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

**Axel Drefahl**

[axeleratio@yahoo.com](mailto: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** = e**X**tensible **M**arkup **L**anguage

**ThermoML** = **T**hermo**d**ynamic **M**arkup **L**anguage 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>

Physical and Chemical Properties Division  
TRC Group  
NIST  
National Institute of Standards and Technology  
Supplying physical and chemical properties data, models, standards, and research for industry, public health & safety, and the environment

**ThermoML**  
An XML-Based IUPAC Standard for Storage and Exchange of Experimental Thermophysical and Thermochemical Property Data

**DATA FILES**

- [Journal of Chemical & Engineering Data](#)
- [The Journal of Chemical Thermodynamics](#)
- [Fluid Phase Equilibria](#)
- [Thermochimica Acta](#)
- [International Journal of Thermophysics](#)

(Other journals may be listed here as agreements are made in the future.)

[Please read the Liability Statement](#)

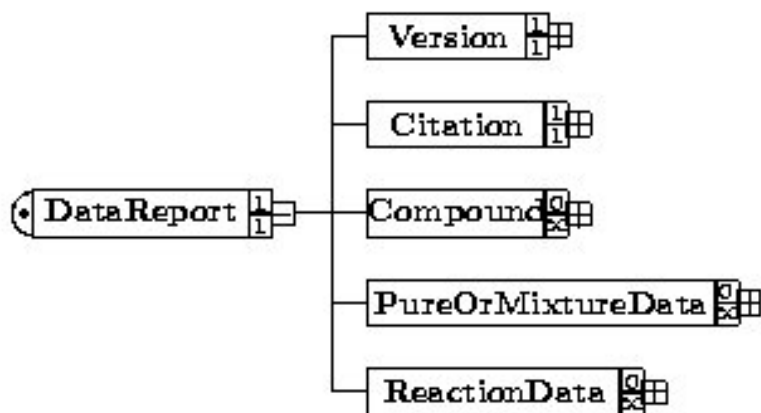
**ThermoML Representation of Published Experimental Data**

This page contains links to ThermoML files, which represent experimental thermophysical and thermochemical property data reported in the corresponding articles published by major journals in the field. These files are posted here through cooperation between the [Thermodynamics Research Center \(TRC\)](#) at the [National Institute of Standards and Technology \(NIST\)](#) and the journal publishers. This project is now underway with data published by the [Journal of Chemical and Engineering Data \(JCED\)](#), the [Journal of Chemical Thermodynamics](#), [Fluid Phase Equilibria](#), [Thermochimica Acta](#), and the [International Journal of Thermophysics](#). The ThermoML files corresponding to articles in the journals are available here with permission of the journal publishers. It is anticipated that this cooperation may be expanded to include other journals in the future.

ThermoML — an XML-based IUPAC Standard for storage and exchange of experimental thermophysical and thermochemical property data — was fully described ([Pure Appl. Chem., 2006, 78, 541-612](#)). [Supporting information](#) for this article includes several examples illustrating the use of ThermoML to process experimental data for pure compounds, mixtures, and chemical reactions as well as the initial ThermoML specification. The framework of the schema has previously been described elsewhere ([J. Chem. Eng. Data, 2003, 48, 2-13](#)). Extensions to the ThermoML schema for the expression of uncertainties were described ([J. Chem. Eng. Data, 2003, 48, 1344-1359 & Supporting information](#)), as were extensions for representation of critically evaluated data, predicted data, and Equation Representation ([J. Chem. Eng. Data, 2004, 49,](#)

- 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](http://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

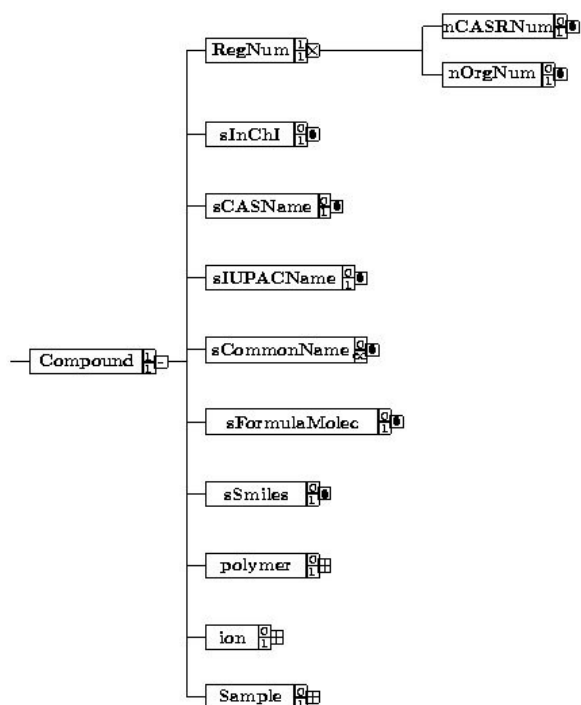
- Inspection of ThermoML files
- Extraction of data
- XML-to-XML conversions  
(chemical dictionary generation)

## 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 compounds (July'07):

1,113 (organics by name)

58 (inorganics by name)

1,154 (distinct CASRNs)

716 (distinct molecular formulae)

## Counting property data nodes:

17,226 (total, Feb'07)

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

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

## 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 ThermoML Archive is organized by article.

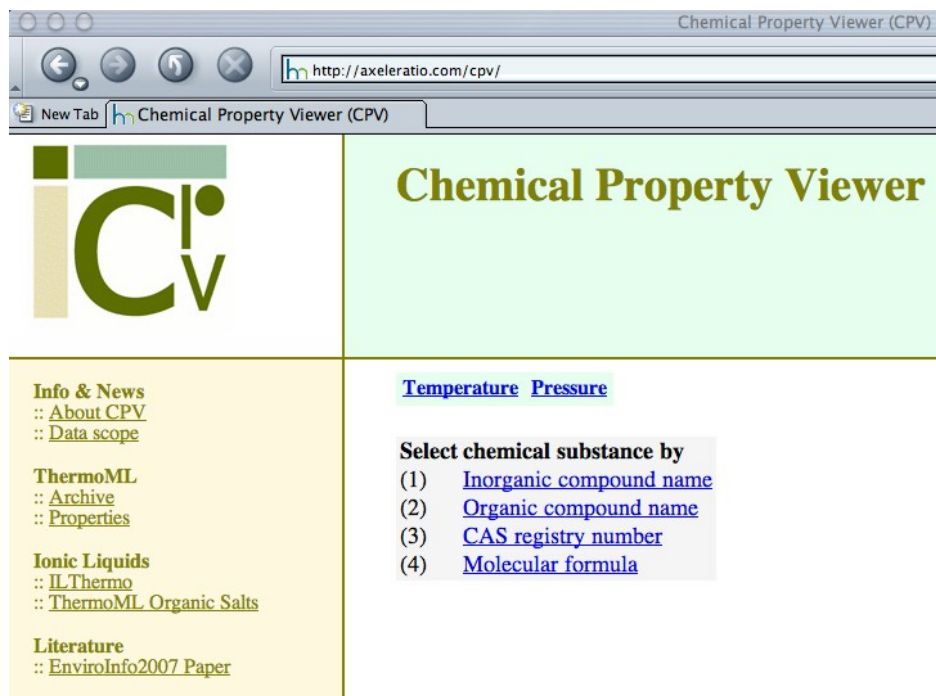
Location of chemicals and properties requires looping over all archive file.

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](http://www.axeleratio.com/cpv)



The screenshot shows a web browser window titled "Chemical Property Viewer (CPV)". The address bar contains the URL "http://axeleratio.com/cpv/". The page layout includes a logo on the left, a main header area, and a sidebar with navigation links.

**Chemical Property Viewer**

**Temperature Pressure**

Select chemical substance by

- (1) [Inorganic compound name](#)
- (2) [Organic compound name](#)
- (3) [CAS registry number](#)
- (4) [Molecular formula](#)

**Info & News**  
:: [About CPV](#)  
:: [Data scope](#)

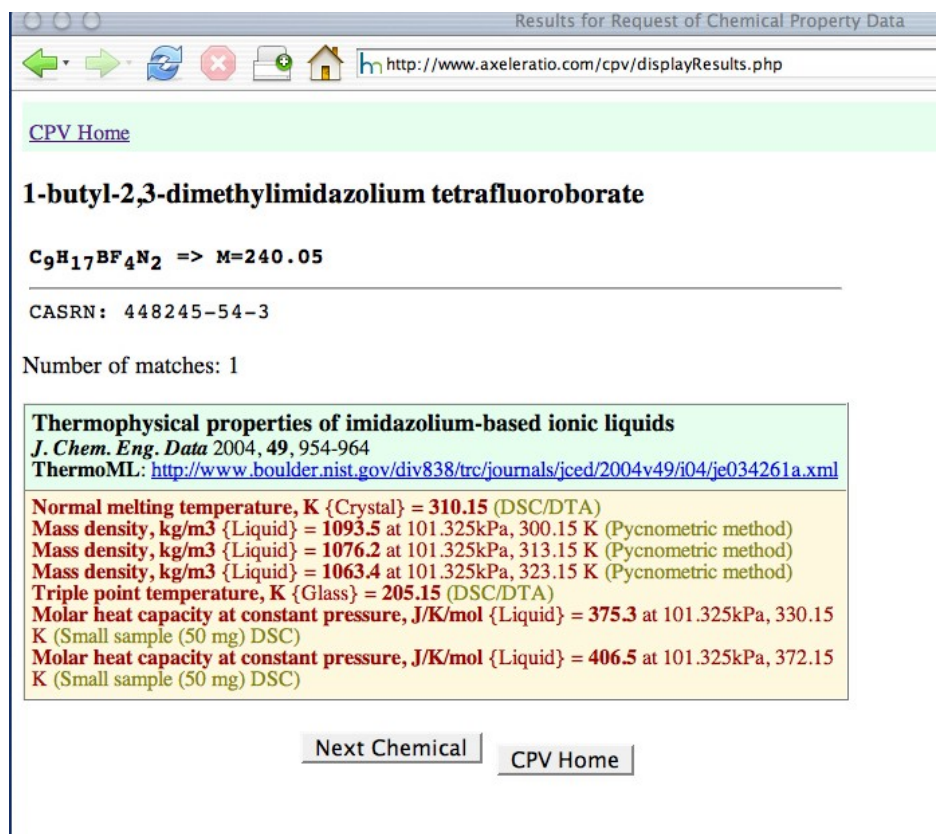
**ThermoML**  
:: [Archive](#)  
:: [Properties](#)

**Ionic Liquids**  
:: [IL Thermo](#)  
:: [ThermoML Organic Salts](#)

**Literature**  
:: [EnviroInfo2007 Paper](#)

- **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



Results for Request of Chemical Property Data

CPV Home

**1-butyl-2,3-dimethylimidazolium tetrafluoroborate**

$C_9H_{17}BF_4N_2 \Rightarrow M=240.05$

CASRN: 448245-54-3

Number of matches: 1

**Thermophysical properties of imidazolium-based ionic liquids**  
*J. Chem. Eng. Data* 2004, 49, 954-964  
ThermoML: <http://www.boulder.nist.gov/div838/trc/journals/jced/2004v49/i04/je034261a.xml>

Normal melting temperature, K {Crystal} = 310.15 (DSC/DTA)  
Mass density, kg/m<sup>3</sup> {Liquid} = 1093.5 at 101.325kPa, 300.15 K (Pycnometric method)  
Mass density, kg/m<sup>3</sup> {Liquid} = 1076.2 at 101.325kPa, 313.15 K (Pycnometric method)  
Mass density, kg/m<sup>3</sup> {Liquid} = 1063.4 at 101.325kPa, 323.15 K (Pycnometric method)  
Triple point temperature, K {Glass} = 205.15 (DSC/DTA)  
Molar heat capacity at constant pressure, J/K/mol {Liquid} = 375.3 at 101.325kPa, 330.15 K (Small sample (50 mg) DSC)  
Molar heat capacity at constant pressure, J/K/mol {Liquid} = 406.5 at 101.325kPa, 372.15 K (Small sample (50 mg) DSC)

Next Chemical CPV Home

- 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

Results for Request of Chemical Property Data

<http://www.axeleratio.com/cpv/displayResults.php>

[CPV Home](#)

**1,1,1-trifluoro-2-(2,2,2-trifluoroethoxy)ethane**

**C<sub>4</sub>H<sub>4</sub>F<sub>6</sub>O => M=182.07**

---

CASRN: 333-36-8

Currently set temperature range: 273 to 333 K

Number of matches: 1

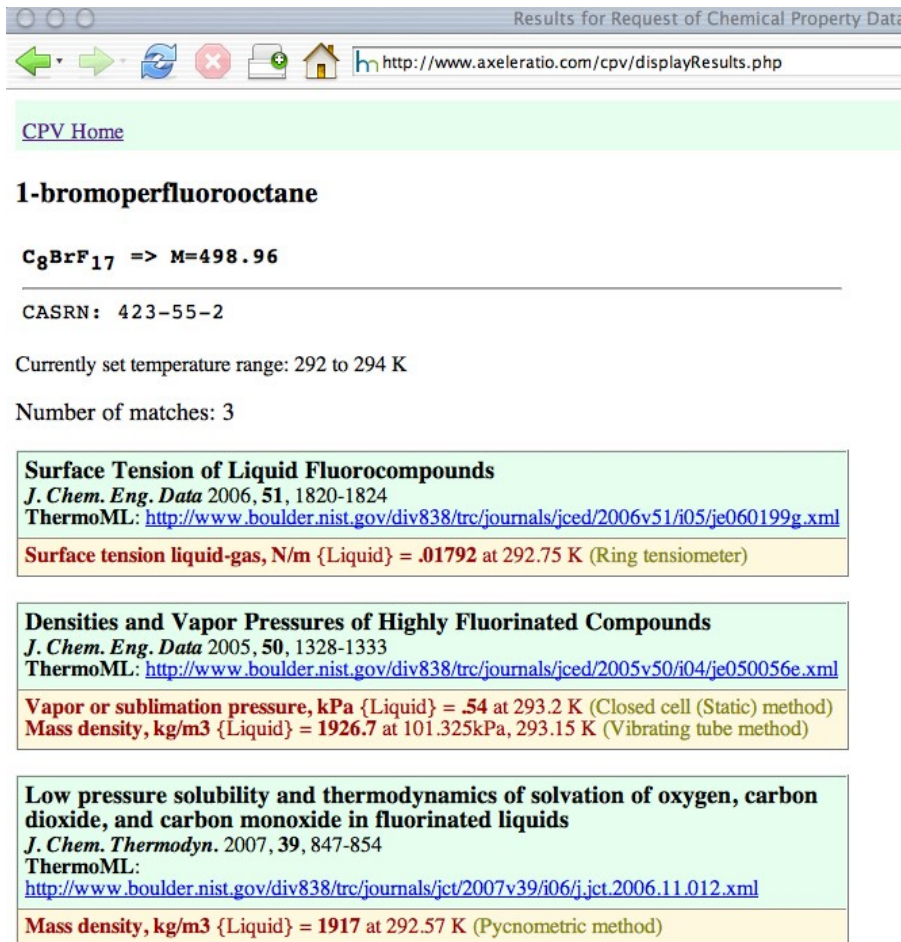
Critical Parameters and Vapor Pressures Measurements of Hydrofluoroethers at High Temperatures	
Critical pressure, kPa {Liquid}	= 2783 (Direct measurement)
Vapor or sublimation pressure, kPa {Liquid}	= 18 at 293.4 K (Closed cell (Static) method)
Vapor or sublimation pressure, kPa {Liquid}	= 18.2 at 293.59 K (Closed cell (Static) method)
Vapor or sublimation pressure, kPa {Liquid}	= 27.9 at 303.09 K (Closed cell (Static) method)
Vapor or sublimation pressure, kPa {Liquid}	= 28 at 303.12 K (Closed cell (Static) method)
Vapor or sublimation pressure, kPa {Liquid}	= 42.6 at 313.12 K (Closed cell (Static) method)
Vapor or sublimation pressure, kPa {Liquid}	= 42.8 at 313.14 K (Closed cell (Static) method)
Vapor or sublimation pressure, kPa {Liquid}	= 63.3 at 323.15 K (Closed cell (Static) method)
Vapor or sublimation pressure, kPa {Liquid}	= 63.4 at 323.17 K (Closed cell (Static) method)
Critical density, kg/m <sup>3</sup> {Liquid}	= 500 (Other)
Critical temperature, K {Liquid}	= 476.31 (Visual observation in an unstirred cell)

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

[Next Chemical](#)

[CPV Home](#)

# CPV results including multiple matches



Results for Request of Chemical Property Data

CPV Home

**1-bromoperfluorooctane**

**C<sub>8</sub>BrF<sub>17</sub> => M=498.96**

---

CASRN: 423-55-2

Currently set temperature range: 292 to 294 K

Number of matches: 3

**Surface Tension of Liquid Fluorocompounds**  
*J. Chem. Eng. Data* 2006, **51**, 1820-1824  
ThermoML: <http://www.boulder.nist.gov/div838/trc/journals/jced/2006v51/i05/je060199g.xml>  
**Surface tension liquid-gas, N/m {Liquid} = .01792 at 292.75 K (Ring tensiometer)**

**Densities and Vapor Pressures of Highly Fluorinated Compounds**  
*J. Chem. Eng. Data* 2005, **50**, 1328-1333  
ThermoML: <http://www.boulder.nist.gov/div838/trc/journals/jced/2005v50/i04/je050056e.xml>  
**Vapor or sublimation pressure, kPa {Liquid} = .54 at 293.2 K (Closed cell (Static) method)**  
**Mass density, kg/m<sup>3</sup> {Liquid} = 1926.7 at 101.325kPa, 293.15 K (Vibrating tube method)**

**Low pressure solubility and thermodynamics of solvation of oxygen, carbon dioxide, and carbon monoxide in fluorinated liquids**  
*J. Chem. Thermodyn.* 2007, **39**, 847-854  
ThermoML: <http://www.boulder.nist.gov/div838/trc/journals/jct/2007v39/i06/j.ct.2006.11.012.xml>  
**Mass density, kg/m<sup>3</sup> {Liquid} = 1917 at 292.57 K (Pycnometric method)**

Next Chemical

CPV Home

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

# Water $\text{H}_2\text{O}$

# 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](http://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](http://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](http://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](http://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**

Experimental polarizability:

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

based on Lorentz – Lorenz (LL) equation.

- **Atom Additivity (AA) approach**  
(Bosque and Sales:  
*J. Chem. Inf. Comput. Sci.*  
**2003**, 42, 1154-1163)

Estimated polarizability:

$$\alpha_{AA} = 0.32 + \sum c_A N_A$$

where

$c_A$  = contribution for atom A

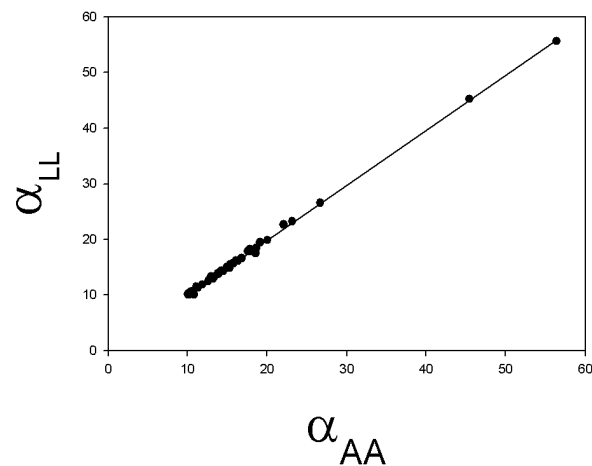
$N_A$  = number of atoms A per molecule

# Results: Polarizability

[www.axeleratio.com/EnviroInfo2007/CompareAlphas.pdf](http://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:  
 $R = 0.9996$**

Comparison of exp. and calc. polarizabilities



# **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](http://www.axeleratio.com/EnviroInfo2007/slides.pdf)**