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

Axel Drefahl[1]

Abstract

The ThermoML Archive supports an open-source approach to science. The XML-structured plain-text files in this archive provide thermodynamic data for chemical compounds and mixtures abstracted from recently published articles. We introduce a Web-based Property Viewer that displays all currently available property data for user-selected pure compounds. Extraction and transformation of ThermoML data using Python-based scripts and JavaScript is discussed. We demonstrate the general use of the Property Viewer and in particular discuss access to data on organic salts such as ionic liquids. The Property Viewer is an easy-to-use tool assisting comparison of "conventional" solvents and those labelled as "green". The ThermoML Archive provides a rich source of data on multi-component systems. We demonstrate access to information on environmentally important (water + chemical) systems.

This paper emphasises the request-to-target precision achieved by applying the Document Object Model (DOM) guided matching while inspecting ThermoML-structured files. Such a precision cannot be obtained by text-based search-key matching since the chemical and environmental literature in general does not confine to unique, unambiguous terms, neither for chemical names nor for terms of properties and their associated units. The DOM-based granularity and annotation of the chemical information in the ThermoML Archive efficiently supports the design of software for customer specified applications such as screening and selective data extraction. As an example, we evaluate a published atom additivity method for the estimation of molecular polarisability, a critical molcular descriptor in ecotoxicological modelling, by extracting from the archive new experimental mass density and refractive index data not used in the original development of that method.

We forecast that the ThermoML Archive will not just serve as an outstanding source for chemical properties and risk assessment but will support eco-integrated chemical design and thus advance future strategies in sustainable chemistry.

1. Introduction

ThermoML is a new standard of the International Union of Pure and Applied Chemistry (IUPAC) for thermodynamic and transport property data of pure compounds, chemical mixtures and chemically reactive systems (Frenkel, 2003, 2004, 2006 and Chirico, 2003). Based on the eXtensible Markup Language (XML), ThermoML allows unambiguous and annotated encoding of chemical property data and platform-independent storage and exchange thereof.

The Physical and Chemical Properties Division of the Thermodynamic Research Center (TRC) Group at the National Institute of Standards and Technology (NIST) offers public access to the ThermoML Archive which currently includes ThermoML files with experimental data from peer-reviewed articles published in five different journals (TRC-ThermoML, 2007).

Managing information on chemical substances and their benefits-versus-hazards profiles requires the knowledgeable translation of a request into a task queue that includes search, data validation and property estimation procedures. Typically, a substantial amount of interactions by human experts is necessary to fulfill such complex tasks that involve reviewing and mining documents and databases with different for-

[1] Owens Technology, Inc., 5355 Capital Court, Suite 106, Reno, Nevada 89502, U.S.A.,
email: axel@owenstechnology.com, internet: http://www.axeleratio.com

mats and access policies. The ThermoML approach and its open-source archive brings researchers a significant step closer to a coherent representation of a vast amount of chemical property data structured for efficient, application-oriented querying.
Herein, we explore the ThermoML Archive by designing a Property Viewer for pure compounds and by screening the archive for data of interest in environmental investigations and molecular-structure/property modeling. The Property Viewer provides access to property data of current archive compounds (Drefahl, 2007a).

2. Approach

The Uniform Resource Identifier (URI) of a particular ThermoML file consists of two parts: the archive path `http://www.boulder.nist.gov/div838/trc/journals/` and the part that identifies the specific journal article from which the ThermoML data have been abstracted. For example, the URI part `jced/2006v51/i02/je050368h.xml` refers to an article published in 2006 in the second issue of volume 51 in the Journal of Chemical Engineering Data (`jced`). Throughout this paper and our internal software dictionaries we use solely the article part of the ThermoML URI (excluding the `xml`-extension) as a short-hand source identifier. The complete URI, then, is easily reconstructed where needed.
For the purpose of this work, all ThermoML files have been downloaded into a local site while preserving the directory and file structure of the on-line archive. This downloaded version includes all files that were availabe in February 2007. The following evaluations and discussions are based on this downloaded archive.
The tree-based structure of each plain-text archive file complies with the language-independent Document Object Model (DOM). Our scripting and web design uses languages that support DOM interfacing. We use the scripting language Python for off-line inspection, extraction and transformation of ThermoML data. The `xml.dom.minidom` module (Python-XML, 2007) provides the needed access and parsing functionality.
Design of the Property Viewer is based on HTML and JavaScript. We apply the Browser Object Model (BOM) in JavaScript (Zakas, 2005) and its XML DOM objects that supply methods to load, parse and traverse XML dictionary files generated by our Python-based scripts.

3. Inspection of the ThermoML Archive

Current ThermoML files use the molecular formula, the Chemical Abstract Service Registry Number (CASRN), and common names to identify chemical elements and compounds. Registry numbers of regulatory and ecotoxicological interest such as the European Inventory of Existing Commercial Chemical Substances (EINECS) number and the Registry of Toxic Effects of Chemical Substances (RTECS) number are not supported in the current ThermoML version. But the ThermoML definition provides element nodes for chemical identifiers that encode detailed structural information such as the IUPAC name, the Simplified Molecular Input Line Entry Specification (SMILES) notation and IUPAC's International Chemical Identifier (InChI). However, our inspection script could not locate any files applying these structural identifiers yet. Since molecular structure information is critical for advanced search strategies and for the design of novel and the evaluation of existing structure/property relationships with the valuable new data the archive offers, we have integrated our database of SMILES notations with our dictionary-generating scripts such that we have access to the majority of archive compounds, which are molecular substances, via structural inquiry.
Before describing the dictionaries, we briefly tour the current content of the archive. On occasion, we came across a ThermoML file for which a parsing error was reported on loading the file into the browser.

In those cases, visual inspection of the source revealed incomplete mark-up and mismatched tags. We will include these files in our future work as soon as they are available as well-formed documents. Our currently downloaded version of the archive contains 1,568 ThermoML files that successfully passed the Python `minidom.parse` call.

Each ThermoML file is organized in blocks: (1) a mandatory version block with the ThermoML version designation, (2) a mandatory citation block with bibliographic information regarding the abstracted journal article, and, depending on article content, (3) a compound block to label and identify the investigated chemicals, (4) a chemical property block with multiple property data nodes including pure-compound and mixture data, and (5) a chemical reaction block with multiple reaction data nodes. The detailed XML structure of each block has been published (Frenkel, 2006). Our inspection script found that, in addition to the article title, 897 of the 1,568 files included the article abstract and 1,417 files contained key words. We found a total of 7,520 compound nodes, many compounds with multiple occurrences. The number of compounds with distinct CASRN is 1,706 and the number of distinct molecular formulae is 1,274. Further, we counted a total of 17,226 property data nodes, of which 7,764 belonged to pure compounds. The remaining property data nodes belonged to two- and three-component systems with 7856 and 1606 nodes, respectively. Finally, 176 reaction data nodes were counted.

ThermoML specifies 10 property groups together encompassing over 120 distinct experimental properties (Frenkel, 2006). Currently, we detected 39 different properties with data for pure compounds in 945 files. The five most frequently occurring properties are vapour or sublimation pressure, mass density, refractive index (Na D-line), viscosity, and molar heat capacity at constant pressure for 473, 452, 244, 232, and 189 compounds, respectively. In many cases, property data are available for an array of temperature or pressure points, typically covering temperature ranges consistent with environmental conditions.

Since the ThermoML Archive - at the top level - is organized by published article, not by chemical structure, composition, or property, ”conversion of the archive” into convenient dictionaries is a prerequisite for automatic compound/property processing. We have implemented Python scripts that generate chemical dictionaries. Our dictionaries enable applications to efficiently look up chemicals by their name, CASRN, and molecular formula. Separate name dictionaries have been generated for inorganic compounds and carbon-containing compounds to allow a respectively organised compound selection in the Property Viewer.

4. ThermoML Property Viewer

The ThermoML Property Viewer (Drefahl, 2007a) is a Web-based application that enables a user to display the list of all currently available property data of a pure compound for which data have been abstracted into the ThermoML Archive. Figure 1 shows an example in which the selection of the organic compound 3-ethoxysalicylaldehyde resulted into one match. A match consists of a reference and a property section. The reference section provides a link to the ThermoML source file and shows the title and reference of the article from which property data were abstracted. The property section contains a list of property lines. Each property line is a concatenation of textual data and numerical values taken unaltered from the ThermoML source. A line begins with the property name including the property units. The property name is followed by a phase descriptor inside curly braces, the equal sign, the property value, the state variables (temperature and pressure, unless they are obvious for a particular property as in the example of Figure 1) and finally, inside parenthesis, the descriptive phrase for the experimental method. A property line in the Viewer contains the basic information to qualify a property value. To keep the property section easy to survey, we restrained from including further metadata such as chemical purity, experimental repeatability, numerical uncertainties, and similar descriptors of data accuracy and validity. But we like to underline that the ThermoML specification provides for a rich set of mandatory and optional annotation nodes to encode such information. Inspection of archive files shows that they have been applied exhaustively if corresponding data is available in the original article. The intent of the Proper Viewer is to get a

first glance of chemical property data and to identify sources for further examination whenever demand arises. The current uncompressed version of the Viewer including HTML, XML, and JavaScript files has a size of about 7.1 megabytes.
The example in Figure 1 was chosen for its small size. For an impressive number of compounds, more than one match will be obtained and sometimes a long list of lines with property values at various temperatures and pressures is displayed. We opted to stay with the per-article organisation since this presentation type directly shows the extent to which properties have been measured in terms of temperature and pressure variations within a particular work and therefore supports the user in deciding which original sources he may want to study in depth according to his goals.

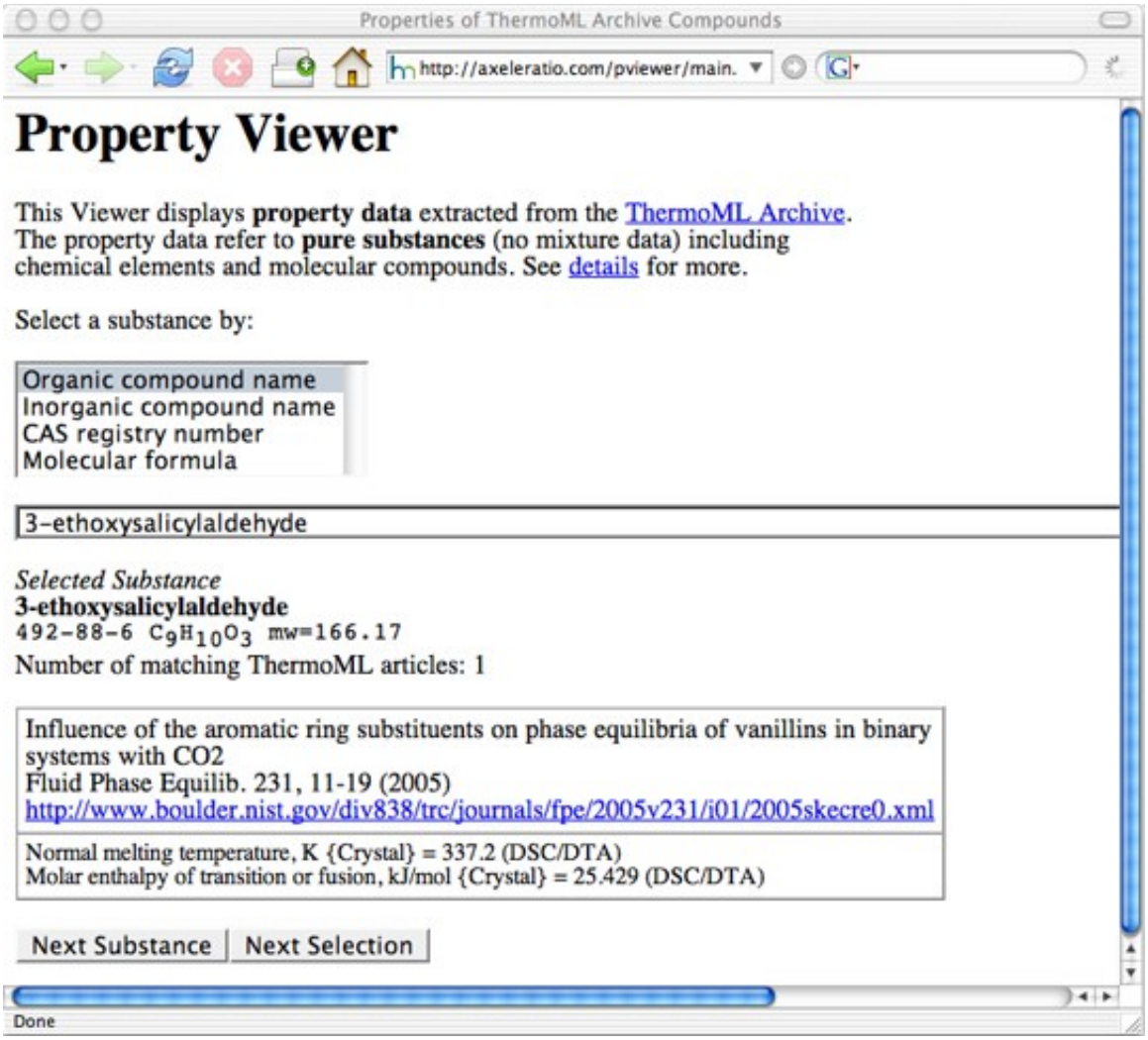

Figure 1
Screenshot of the Property Viewer in the Web Browser displaying results for 3-ethoxysaliciclyaldehyde

5. ThermoML Compounds and Properties of Environmental Interest

The ThermoML archive contains property data for compounds that are of environmental interest as both media and potential pollutants. Important media are water as an ubiquitous compound and 1-octanol to simulate the lipid phase of organisms. Water can be selected in the Property Viewer either by its inorganic compound name, water, by CASRN, 7732-18-5, or by molecular formula, H_2O. Currently, 59 matches are displayed including data for mass density, vapour pressure, surface tension, viscosity, molar heat capacity, thermal conductivity, electrical conductivity, and speed of sound. For 1-octanol (111-87-5, $C_8H_{18}O$) 18 matches are found including data for mass density, vapour pressure, surface tension, viscosity, and molar heat capacity.

By browsing the list of organic compound names in the Property Viewer, we identify a large number of liquids including environmentally problematic solvents and also solvents, such as low- and high-molecular-weight esters and ionic liquids. Many esters and ionic liquids have been acknowledged as green solvents in contrast to solvents such as volatile organic compounds (VOCs) and halogenated hydrocarbons. A typical request for a solvent to be green is to have low volatility and good thermal and hydrolytic stability. Depending on the compound selection, the Viewer provides easy access to temperature-dependent property data to assess volatility and stability behaviour and other properties desirable for a particular application, thus, allowing comparison of alternatives and critical evaluation of a green solvent claim.

Ionic liquids are usually defined as organic salts with a melting point below 100 °C. Besides their intriguing physicochemical properties that qualify them for replacement of "conventional" reaction media, the debate is still on about how green ionic liquids really are. A study with a set of 1-butyl-3-methyl-imidazolium salts has demonstrated that bioaccumaltion in aquatic organisms will not occur (Belvèze, 2004). Results of a recent study on biodegradability and ecotoxicity of selected ionic liquids (Garcia, 2005), however, cast a mixed view on a generalised green solvent claim. Clearly, ionic liquids have to be analysed individually and, ideally, in combination with an early life-cycle assessment regarding their proposed industrial applications (Jastorff, 2005). Although ecotoxicological data are outside the scope of ThermoML, physicochemical properties that are a prerequiste to model environmental fate and to design suitable bioassays are readily available in the archive for various ionic liquids. We identified 47 organic salts with ThermoML property data (Drefahl, 2007b).

In addition to properties of pure compounds, the ThermoML Archive contains property data of multi-component systems of which aqueous binary systems such as solutions of chemicals in water are of eminent interest in environmental science. We have generated a list of compounds for which such data are found in the archive (Drefahl, 2007c). Currently, there are 424 compounds. For each compound the investigated properties and links to the ThermoML sources are given. Investigated properties include mass density, surface tension, viscosity, molar enthalpy of solution, molar binary diffusion coefficient, activity coefficient, Henry's Law constant and various equilibrium-composition properties.

6. Evaluation of Estimation Methods with New ThermoML Data

Since the ThermoML delivery process is dedicated to unambiguous abstraction of data that facilitates data validation and quality assessment, chemical property data from the ThermoML Archive are ideally suited for the evaluation of existing and the development of new quantitative property/property relationships (QPPRs) and quantitative structure/property relationships (QSPRs). Such relationships with applications to environmental risk and fate assessment have been reviewed elsewhere (Lyman, 1991 and Reinhard,

1999). Here, we demonstrate application of ThermoML data by comparing experimental and estimated values for molecular polarisability.

Experimental polarisability, α_{LL}, values are calculate with equation $\alpha_{LL} = 0.3964 R_D$ (Bosque, 2002), where R_D is the molar refraction derived with the Lorentz-Lorenz (LL) equation (Nelken, 1991 and Drefahl, 1999) using experimental data for mass density and refractive index (Na D-line). Estimated polarisability, α_{AA}, values are calculated with the atom additivity (AA) method (Bosque, 2002). Calculation of α_{AA} simply requires the input of the molecular formula for a chemical compound. The atom additivity method of Bosque and Sales was derived with a working and predictive set of organic compounds with polarisability values between 20 and 25 °C. We extracted 64 compounds from the ThermoML Archive for which data pairs (mass density and refractive index) at either 20 or 25 °C were reported and which were not part of the original work of Bosque and Sales. We obtained 97 data pairs due to multiple occurrences of some compounds. For the 25 data pairs at 20 °C we calculated α_{LL} and α_{AA}. The Pearson's correlation coefficient between corresponding α_{LL} and α_{AA} values is 0.9994. Similarly, for the 49 data pairs at 25 °C, the correlation coefficient is 0.9996. This results shows excellent agreement between experimental and estimated values and confirms that a temperature effect is neglegible for the molecular polarisability within this temperature interval.

Various group contribution models (GCMs) to estimate R_D, and therefore α, are known and have been reviewed (Drefahl, 1999). R_D is an important variable in estimating boiling points, liquid viscosities and ecotoxocological properties. Since the atom additivity method of Bosque and Sales simply requires input of the molecular formula, whereas other methods require input of the detailed molecular structure, the atom additivity method is an excellent approach allowing rapid calculation of a molecular descriptor with physicochemical significance for a huge set of compound classes. The found agreement between experimental and estimated polarisabilities nominates α_{AA} as a significant candidate to be included in a molecular descriptor set in rational molecular design and environmental modelling.

7. Discussion

The ThermoML Property Viewer has a very simple interface. The user is not required to enter any search text, but merely selects the compound of interest by an identifier of her choice from a list box. This interface design informs the user immediately for which compounds property data are available and guarantees at least one match for her selection. For temperature dependent properties, some matches can contain a very long list of property lines. Hence, in future versions of the Property Viewer it might be desirable to offer customisation such as restrictions to properties at atmospheric pressure and ambient temperature ranges. Also, a feature for pre-selecting properties or property groups will be beneficial. By default, however, user interaction should be minimal, as in the current version, unless the user chooses otherwise.

The generation of XML dictionaries as intermediate look-up files proved to be an efficient strategy, because it is not known in advance in which files a particular compound is present. Since a rapid growth of the ThermoML Archive can be expected, preprocessing into a format that interfaces the targeted application will almost always be necessary. The XML dictionaries will increase with the growing archive and therefore the performance design of the Property Viewer needs to be adjusted to guarantee acceptable load and display times.

We showed that the ThermoML Archive contains over 1,500 article files covering five journals over approximately a three year period (2004 to 2006). This information may allow a conservative extrapolation of the number of files and data in the archive at any future time. But an upward correction of such an extrapolation will become necessary if articles from the ante-ThermoML era will be included and also if additional journals will join the ThermoML approach.

Our Property Viewer is a window into the domain of pure compounds in the archive. Similar viewer-like access to the data of multi-component systems including aqueous solutions and mixtures will be more involved. In fact, a multitude of angles is thinkable from which a user might want to view the data. Challenges associated with the implementation of such complex queries will be addressed in future work.
ThermoML property data are key parameters in assessing distribution and transport behaviour of chemicals in and between different environmental compartments. To understand chemical fate and long-term effects of compounds and formulations on humans and the environment, access to data on bioaccumulation, biodegradability, toxicity and ecotoxicity becomes essential. Currently, such data is available through published media, material safety data sheets (MSDSs), databases and on-line services like Google Scholar. However, the data comes in a view-or-print rather than extract-and-apply format as desired in computational environmental modelling and rational decision making. Even precisely formulated requests may result into incomplete or misleading answers due to less precise data annotation on the server site. We encourage a ThermoML-like presentation of original environmental data of chemicals. Certainly, the definition of an XML-based language for such data will involve increased metadata complexity, since, in addition to temperature, pressure, and composition, bioassay parameters such as pH value, exposure time, and tissue or species description need to be formatted.
Arguments for data sharing and an open-source web of ecological data (Parr, 2007) and scientific literature (Swan, 2007) have been put forward. The success of the ThermoML approach should inspire researchers and publishers in environmental sciences to an equally ambitious undertaking.

8. Conclusions

The ThermoML Archive is a powerful open-source database covering a vast amount of chemical systems and properties that are of wide-spread interest including environmental research and regulation. Consistent updating and dynamic growth of this archive with the progress of new publications makes its use very appealing and valuable. The ThermoML data structure supports critical evaluation of existing chemical data; evaluation, enhancement, and novel design of property estimation methods and ranking procedures for hazardous compound testing; and efficient screening for alternatives to hazardous chemicals.

Acknowledgement. The author likes to thank Gary L. Owens for supporting this work through research in sustainable chemistry and renewable resources at Owens Technology, Inc.

References and Notes

Belvèze, L.S. (2004): Modelling and Measurement of Thermodynamic Properties of Ionic Liquids, Master Thesis, University of Notre Dame, Indiana (U.S.A.).

Bosque, R., Sales, J. (2002): Polarizabilities of Solvents from the the Chemical Composition, in: J. Chem. Inf. Comput. Sci., 42, pp. 1154-1163.

Chirico, R.D. (et al.) (2003): ThermoML – An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 2. Uncertainties, in: J. Chem. Eng. Data, 48, pp. 1344-1359.

Drefahl, A. (2007a): The ThermML Property Viewer discussed in this work is found at http://www.axeleratio.com/EnviroInfo2007/pviewer.html; a version of the Viewer with frequently updated new data from the ThermoML Archive is at http://www.axeleratio.com/pviewer/main.html.

Drefahl, A. (2007b): http://www.axeleratio.com/EnviroInfo2007/OrganicSalts.html.

Drefahl, A. (2007c): http://www.axeleratio.com/EnviroInfo2007/AquBinSys.html.

Drefahl, A. (2007d): http://www.axeleratio.com/EnviroInfo2007/CompareAlphas.pdf.

Drefahl, A. (1999): Refractive Index and Molar Refraction, Chapter 4 (Reinhard, 1999).

Frenkel, M. (et al.) (2003): ThermoML – An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 1. Experimental Data, in: J. Chem. Eng. Data, 48, pp. 2-13.

Frenkel, M. (et al.) (2004): ThermoML – An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 3. Experimental Data, in: J. Chem. Eng. Data, 49, pp. 381-393.

Frenkel, M. (et al.) (2006): XML-Based IUPAC Standard for Experimental, Predicted, and Critically Evaluated Thermodynamic Property Data Storage and Capture (ThermoML) (IUPAC Recommendations 2006), in: Pure Appl. Chem., 78, pp. 541-612.

Garcia, M.T., Gathergood, N., Scammells, P.J. (2005): Biodegradable ionic liquids. Part II. Effect of the anion and toxicology, in: Green Chem., 7, pp. 9-14.

Jastorff, B. (et al.) (2005): Progress in evaluation of risk potential of ionic liquids – basis for an eco-design of sustainable products, in: Green Chem., 7, pp. 362-372.

Lyman, W.J., Reehl, W.F., Rosenblatt, D.H. (1991, third printing): Handbook of Chemical Property Estimation Methods, American Chemical Society, Washington, DC.

Nelken, L.H. (1991): Index of Refraction, Chapter 26 (Lyman, 1991).

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

Python-XML (2007): http://docs.python.org/lib/module-xml.dom.minidom.html.

Reinhard, M., Drefahl, A. (1999): Handbook for Estimating Physicochemical Properties of Organic Compounds, John Wiley & Sons, Inc., New York.

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

TRC-ThermoML (2007): http://trc.nist.gov/ThermoML.html.

Zakas, N.C. (2005): Professional JavaScript™ for Web Developers, Wiley Publishing, Inc., Indianapolis, Indiana.