Int J Thermophys (2012) 33:22–33
DOI 10.1007/s10765-011-1107-7

NIST/TRC SOURCE Data Archival System: The Next-Generation Data Model for Storage of Thermophysical Properties

A. Kazakov · C. D. Muzny · K. Kroenlein · V. Diky · R. D. Chirico · J. W. Magee · I. M. Abdulagatov · M. Frenkel

Received: 20 September 2011 / Accepted: 7 October 2011 / Published online: 28 October 2011

Abstract A new data model for storage of experimental thermophysical and thermochemical property data was developed and implemented for the NIST/TRC SOURCE data archival system. Substantial improvements in data quality, as well as system usability and extendability, are achieved. Substance identification based on chemical structures was implemented. Availability of stored chemical structures will facilitate the use of property estimation methods to supplement the experimental information.

Keywords Database · Data model · Thermochemical properties · Thermophysical properties

1 Introduction

The NIST/TRC SOURCE data archival system (SOURCE) is a large, general-purpose archive of experimental data covering thermophysical and thermochemical properties for pure compounds and mixtures of well-defined composition, as well as for chemical reactions. It has been developed and maintained for nearly 30 years [1–3], and at present contains over 4 million experimental data points (see Table 1 for detailed

A. Kazakov (✉) · C. D. Muzny · K. Kroenlein · V. Diky · R. D. Chirico · J. W. Magee · I. M. Abdulagatov · M. Frenkel
Thermophysical Properties Division, National Institute of Standards and Technology, Boulder, CO 80305-3337, USA
e-mail: andrei.kazakov@nist.gov

Table 1 SOURCE statistics on September 7, 2011

	Count
Compounds with data	24 711
References with data	43 145
Pure compounds data points	1 270 408
Binary mixture data points	2 402 892
Ternary mixture data points	637 529
Reaction data points	14 437
Total data points	4 325 266

statistics). Every stored data point is associated with an experimental uncertainty [4], and only original experimental data from the literature are captured (i.e., no derived or predicted data). SOURCE is an essential element of the dynamic data evaluation system, ThermoData Engine (TDE) [5–9], developed at NIST. TDE is a combination of SOURCE with expert-system software, designed to automatically generate recommended data based on available experimental data and prediction methods, thus providing the ability to produce critically evaluated data dynamically. Dynamic data evaluation, in turn, is one of the central elements of emerging global communication systems in thermodynamics [10–12] and is gaining acceptance in facilitating advances in scientific and industrial research [13,14], as well as in the global data validation process, currently involving five major journals in the field [15,16].

Over the years [1–3], SOURCE and its underlying data model have evolved from a flat-file system to a non-relational, and, finally, a relational database. Recent rapid advances in information technology have generated the need for further evolution of the data model to address the demand for new features and to eliminate limitations hindering further development. This article describes our recent efforts on redesigning the SOURCE schema.

2 New Data Model

2.1 Need for a New Data Model

The general relational data model concept for storage of property data [1–3] is shown in Fig. 1 and contains four basic building blocks: "Substances," "References," "Samples," and "Data." The "Substances" block provides unambiguous definition of substances

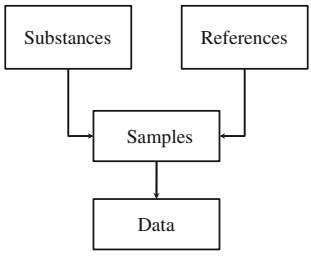

Fig. 1 Conceptual data model for storage of property data

(objects of property measurements), the "References" block defines bibliographic sources of information. The substance and reference are brought (linked) together with the "Samples" block. An accurate definition of experimental samples (source, purity, purification methods, etc.) is essential for defining experimental measurements, especially considering the significant effect of sample purity on experimental uncertainties. Finally, samples are associated with data in the "Data" block. While this global data structure is the same as in a previous version of SOURCE [3], the specific implementation of individual blocks requires special consideration. The previous SOURCE data model [3] was designed almost 10 years ago, guided by hardware and software options available at the time. In particular, storage was viewed as a major factor controlling design decisions. While SOURCE is one of the world's largest collections of thermophysical and thermochemical properties (Table 1), by today's standards, it is a relatively small database, and storage considerations are of minor concern. On the other hand, storage of all relevant metadata and multimedia information, provision for automation of most of maintenance tasks, and extendability (e.g., possibility to store new types of information without major revisions of underlying schema) are essential. In addition, our extensive experience with the collection of diverse data has revealed several drawbacks in the previous design that cause significant problems for archive maintenance and operation, and cannot be addressed without major changes to the data model. For example, the individual data blocks shown in Fig. 1 were not properly encapsulated (i.e., they are, in fact, connected via multiple logical relations); consequently, even small design changes within a given block could require major changes in the database structure. The remainder of this section will describe in more detail the new design of the schema blocks, "Substances," "References," and "Data," that addresses these problems.

2.2 "Substances" Block

A critical issue in designing a modern archival system for thermophysical properties is the identification of chemical substances. Identification implies the ability to unambiguously define a substance for storage, associating it with a specific identifier (key), and the ability to unambiguously find a substance when database searches (queries) are performed. The latter is also an important step for loading new data that need to be associated with the correct substance.

As the objective of this study is data-driven, the term "substance" in this context refers to the object of a reported thermophysical property measurement that is rigorously defined (i.e., the definition should be sufficient to independently reproduce the experiment). The majority of these objects of thermodynamic studies are chemical compounds and their mixtures. However, in some cases (e.g., hydrates in a crystal phase), these substances can be represented by sets of chemical compounds that exist as a single bound chemical system only at certain conditions.

Identification of chemical substances in the previous schema [3] was performed by chemical name and, if available, by the Chemical Abstract Service (CAS) registry number (internal use only). Both choices posed significant problems. Uniqueness of a chemical name is difficult to enforce. Modern International Union of Pure and Applied

Chemistry (IUPAC) [17] and CAS [18] specifications allow for systematic and unambiguous naming of compounds. However, these standards evolve over time, which makes their use as primary identifiers problematic for long-term projects. In addition, the standards often fall behind as new classes of compounds are introduced. Finally, some systematic names tend to be rather long and complex, and require a high level of expertise for their generation and interpretation. As such, they are prone to human errors. The use of CAS registry numbers is hindered by their rather restrictive licensing. They also possess the same problems as the systematic names: they change as errors in the CAS registry are discovered (mostly due to deletion of "deprecated," duplicated entries) and often lag behind in assigning numbers to new substances as they appear in the literature.

The obvious choice of an identifier that is "permanent" and can be generated independently in-house (i.e., is not controlled by an external entity) is the chemical structure. Chemical structure implies the combination of composition (inclusive of specific isotopes and charges, if present) and bonding (connectivity) information (inclusive of bond stereo, if applicable). A two-dimensional layout for depiction is also desirable. Several chemical structure formats are in wide use [19–21]; all of them include the basic functionality required. The MDL (now Symyx) MOL format [19] is one of the most popular historically. It is also supported by virtually all software used for molecular drawing and analysis, and was adopted in this study. Having a substance defined via a MOL file, however, does not solve the search/query problem: one cannot perform matching of two structures defined by MOL files directly. Efficient matching of chemical structures is one of the central problems in chemoinformatics and is the subject of numerous studies [22]. In practice, it is accomplished by associating complete chemical structure information with a short (usually string) notation; these string representations can be easily compared and matched. Two popular string representations of molecular structures are the Simplified Molecular Input Line Entry System (SMILES) [23] and IUPAC International Chemical Identifier (InChI) [24]. In its original formulation, SMILES notation was not unique; this problem was addressed by introducing *canonical* SMILES. Additional SMILES extensions were also developed to represent bond stereo, radical centers, etc. Presently, several commercial SMILES versions exist that can potentially represent chemical structures in a unique manner and with most of the structural features necessary for our purposes (e.g., [25]), yet an accepted standard implementation is still lacking (although some initial efforts are underway [26]). InChI [24] is a recent development, and, being a well-documented IUPAC standard, it quickly gained widespread acceptance in the chemoinformatics community, and was also adopted here. It should be pointed out that InChI string generation depends on a number of options controlling structure perception. A specially selected set of options (absolute stereo, no distinction between tautomers, etc.; see [24] for a detailed list) produces a *standard* InChI string intended for use in information exchange. Finally, because the InChI string can be rather long for complicated structures, a hashed version of the InChI string, the InChIKey, was introduced. InChIKey is a condensed digital signature of the InChI string of fixed length (27 characters) and was developed to accommodate search and indexing applications.

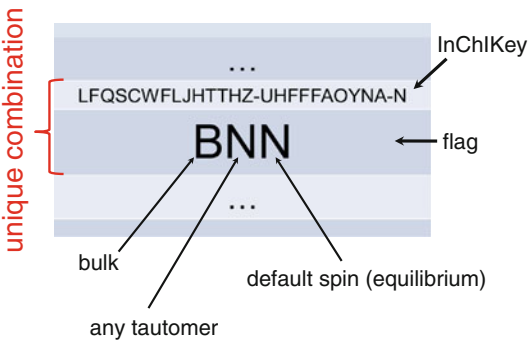

Fig. 2 Illustration of compound identification

The standard InChI technology addresses the majority of our present needs for substance identification. However, based on the contents of the current SOURCE collection, there are three known exceptions:

(1) Standard InChI makes no distinction between tautomers. This is consistent with the overwhelming majority of stored data; however, there are some cases where this distinction is necessary;
(2) InChI contains no information on spin multiplicity; again, distinction between compounds with the same structures but different spin quantum numbers is needed in a limited number of cases;
(3) Finally, a distinction between the *bulk* and the *species* states is needed. This is also a rare situation when a compound has a tendency either to associate or to dissociate, and the data are reported separately for the overall mixture (bulk) as well as for the individual components that include a compound in its non-associated or non-dissociated state (species). Distinction between the bulk and the species cannot be made with standard InChI.

Considering the above, compound identification (enforced by the uniqueness constraint in the database) was formulated as follows (see Fig. 2). The first part of the identifier is a *non-standard* InChIkey generated by activation of the distinction between the tautomeric forms in the set of options for standard InChI. The second part of the identifier is a three-register flag: the first register has settings for "bulk" and "species," the second has settings for "specific" or "any" tautomer, and the third defines the value of the spin multiplicity. A combination of the two (non-standard InChIkey and a flag) provides unique identification for all substances presently available in SOURCE. Addition of registers to the flag is possible if new exceptions are discovered.

Having resolved the problem of compound identification, the "Substances" block of the database is developed as shown schematically in Fig. 3. The table "Substance Header" contains a unique internal numerical substance identifier and a flag indicating the type of substance. The table "Names" stores substance chemical names, and the table "Registry Numbers" stores the CAS Registry numbers if available. Four types of substances are presently supported: "Compound," "Complex," "Compound and Complex," and "Material." "Compound" refers to a substance that can be represented by a single MOL file and, consequently, by a single InChI/InChIKey. This represents the majority of substances in SOURCE. The table "Compound" contains unique

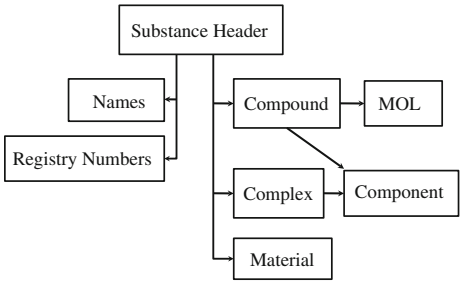

Fig. 3 Schematic representation of the "Substances" data block

compound identification (non-standard InChIKey and the flag of registers, as described above) and a number of other compound characteristics that can be used for searches (standard InChIKey, molecular formula, molecular weight, etc.). The table "MOL" contains "larger items" such as the MOL file, full InChI string, structure depictions in several graphical formats, etc. These items are stored outside of the "Compound" table for database efficiency. The "Compound" table is heavily used for searches, and the columns containing large items would slow the retrieval process. The table "Complex" contains information for substances composed of disconnected compounds (such as ionic structures, hydrates, etc.). Individual components of the entries of the "Complex" table are listed in the "Component" table along with their stoichiometric coefficients or mole/mass fractions if applicable. Equilibrium or "undefined" mixtures of components can also be assigned explicitly. This feature is reserved for mixtures of isomers that can coexist in samples used in experimental measurements. If the collection of disconnected components can be defined with a single MOL file, it is defined as "Compound and Complex" type and stored in *both* the "Compound" and "Complex" tables. This is done to take advantage of indexing with InChIKeys and to accommodate searches for the overall structure and the individual components. The type "Complex" (cases stored only in the "Complex" table, without duplicated entry in the "Compound" table) includes, for example, special mixtures that cannot be represented by a MOL file, but commonly are defined as unique chemical systems (such as air), and substances that can be described only in terms of relative stereochemistry and have to be defined as equilibrium mixtures of two or more steroisomers. Finally, the table "Material" is reserved for substances that cannot be described with the above formalism, but may be considered for future extensions of the database scope (e.g., polymers, biomaterials, etc.).

It should also be emphasized that the "Substances" block in the present formulation (Fig. 1) is encapsulated; the relational link to the rest of the database is done via a single internal numerical identifier defined in the table "Substance Header." This simplifies any future modifications or extensions within this block.

2.3 "References" Block

The new structure of the "References" block is schematically shown in Fig. 4. Conceptually, the structure remains similar to that in the previous design [3]. Efforts were focused on converting lumped bibliographic descriptions from the previous schema

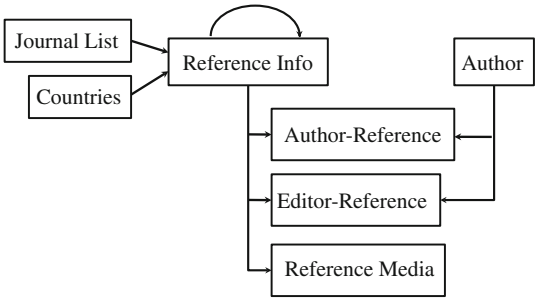

Fig. 4 Schematic representation of the "References" data block

into collections of structured fields, following the documented ideas from bibliography management software (see, for example, [27]). The table "Reference Info" contains an internal numerical reference identifier and fields to accommodate the diverse types of literature documents supported by SOURCE (journal article, report, thesis, conference proceedings, patent, book/book chapter, etc.). The arrow linking this table to itself (Fig. 4) indicates the introduced ability to link the original publication with its follow-up erratum, or the publication in native language with its English translation. Parsing and formal field standardization for the large collections of diverse types of documents is extremely challenging and will proceed gradually over time. Because journal articles constitute the vast majority of bibliographic information stored, initial efforts were focused on this type of document, with emphasis on enforcing standard journal names and abbreviations compiled in the table "Journal List." This will help prevent duplication of reference entries and eliminate journal name misspellings.

As in the old schema [3], the table "Author" stores names of authors. The reference and its authors are linked via the table "Author-Reference." A similar link table, "Editor-Reference," was added to link the reference and the editors (for books). A new table "Reference Media" provides storage for the original sources and supplementary information in electronic form. (Multiple formats are supported.) Finally, the auxiliary table "Countries" provides a standard list of countries used for patent definition.

As for the "Substances" block, "References" is also encapsulated and is connected to the rest of the database via a single internal numerical reference identifier.

2.4 "Data" Block

Developed previously [2,3], the data model for storage of pure substance and binary and ternary mixture data based on the Gibbs phase rule has proven to work well over the years and was adopted in the new data model. The pure or mixture data are organized in similarly structured blocks, as illustrated in Fig. 5. The separate tables contain single-valued ("0VAR"), one-variable ("1VAR"), two-variable ("2VAR"), and three-variable ("3VAR") data. (Note that the pure substance system does not have the "3VAR" table.) The "variables" imply variation in one or more state variables, while the remaining degrees of freedom are defined via numerical or non-numerical constraints. When variables are defined ("1VAR," "2VAR," or "3VAR"), the table contains the header information describing the dataset and the constraints; the variable and property values

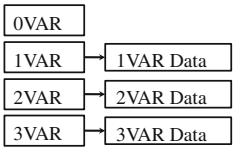

Fig. 5 Schematic representation of data block for storage of pure compound or mixture data (3VAR and 3VAR Data are not present for pure compound data)

are stored in "1VAR Data," "2VAR Data," or "3VAR Data," respectively. Among the significant changes as compared to the old schema is the introduction of uncertainty fields for all variables and numerical constraints. Previously, only the combined uncertainty for the property value was stored. The new schema provides a more complete representation of the data. In addition, explicit non-state variables and their uncertainty fields were added. Description of non-state variables (e.g., wavelength, frequency, etc.) is important for rigorous definition of some properties but had not been included in the old schema.

The data blocks describing reaction data are shown in Figs. 6 and 7. Conceptually, they remain similar to the data structures in the old schema. Changes primarily eliminate unnecessary limitations and make storage more flexible. For example, "ReactChanged" data (Fig. 6), associated with a change of state during a chemical reaction, can now accommodate an arbitrary number of reaction participants (given separately in the table "ReactChanged Components"). Specific ordering of participants mandated in the old schema is no longer required. The "ReactEquil" table (Fig. 7) that stores reaction equilibrium data now combines data stored previously in two separate tables due to the more flexible storage schema. Components, variables, and constants for the system are stored in separate tables ("ReactEquil Components," "ReactEquil Constants," and "ReactEquil Variables," respectively), and may have any necessary counts associated with the system under experimental study. As with the other tables of the "Data" block, the uncertainty fields were introduced for all variables and numerical constraints.

Fig. 6 Schematic representation of data block for storage of reaction data involving change of state

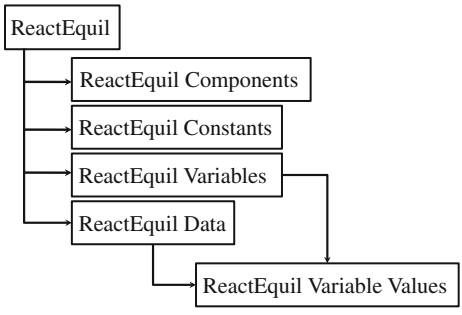

Fig. 7 Schematic representation of data block for storage of reaction data involving chemical equilibrium

2.5 Data Transfer

The crucial and most time-consuming part of data transfer from the old to the new schema was generation of molecular structure (MOL) files for over 20 000 substances present in SOURCE, as this information was not present in the old schema. This work required human judgment and expertise, and was performed manually. A consistent protocol for generation and validation of the structure files was established. The procedure relied on the use of the chemical names from the old schema and two commercial tools providing conversion of "name to structure" [28,29]. The capabilities of earlier versions of these tools were tested [30], and it was demonstrated that the software produced more accurate results than those of an average human expert.

Failure to achieve consistency among the chemical structures generated from the stored chemical names resulted in further analysis. If erroneous or ambiguous names could be identified with certainty by a human compiler, they were corrected or eliminated. More complicated cases required investigation of the original literature sources to identify the correct chemical system for which the experimental data were reported. Finally, the most difficult cases were resolved by nomenclature experts. An additional benefit of this effort was that numerous substance-related database errors were corrected. A similar, but smaller-scale, effort was conducted to manually associate journal names from unstructured reference fields of the old schema with the standard journal abbreviations. Encountered reference errors were corrected.

The data load to the new schema was accomplished with a framework of Perl scripts specifically developed for this purpose. Web-based access tools needed for database maintenance are being developed with the same framework.

3 SOURCE and NIST/TRC Data Processing

SOURCE is the central element of the NIST/TRC data processing system (Fig. 8). Large-scale data capture from the literature, both in-house and by external contractors,

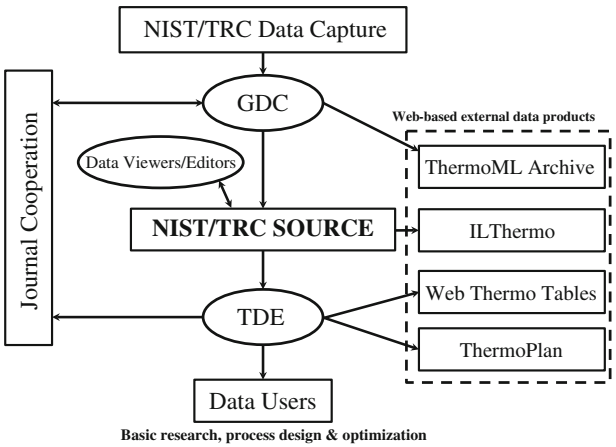

Fig. 8 NIST/TRC data processing flow

is performed with the Guided Data Capture (GDC) software [31]. Captured data are verified and, if needed, corrected by the NIST/TRC staff and subsequently loaded into SOURCE. A system of data viewers and editors is used for low-level visualization and correction of SOURCE data. The TDE software uses SOURCE information to provide data compilations, generate recommendations, identify erroneous (inconsistent) data sets, give guidance in experimental planning, etc. (see [5–9] for a complete description of the TDE capabilities). TDE is used as a stand-alone application with a user interface [32], as part of a process simulation system [14], or as a batch processing system customized in-house for a specific task.

The NIST/TRC processing system also provides the basis for quality control of published data, a process endorsed by five scientific journals (*Journal of Chemical and Engineering Data*, *The Journal of Chemical Thermodynamics*, *Fluid Phase Equilibria*, *Thermochimica Acta*, and *International Journal of Thermophysics*) [16]. The data from submitted articles are captured with GDC, thus providing an initial level of verification enforced by the software (e.g., completeness of system definition, absence of obvious outliers, etc.). The next level of data verification is carried out by comparing the captured data with the TDE-generated recommendations based on the current body of knowledge (experimental data and predictions). If problems are found at any level, they are communicated back to the authors and journal editors. As a service to the authors, the list of references containing experimental data for the system studied in the submitted article is also provided.

Finally, SOURCE provides the basis for several web-based data products offered by NIST/TRC. As a part of the cooperation with scientific journals, the data published in these journals and loaded into SOURCE are also made available to the general public in IUPAC-standard ThermoML format [33] via the ThermoML Archive [34]. The IUPAC Ionic Liquids database (ILThermo) [35,36] represents a subset of data extracted from SOURCE devoted to ionic liquids. Results of TDE evaluations of SOURCE data are presented by two more data products, Web Thermo Tables (WTT) [37–39] and ThermoPlan [40]. WTT is a subscription database that provides thermophysical properties critically evaluated with TDE for over 27 000 compounds. The open-access web application ThermoPlan offers recommendations for the relative merit of a given measurement via assessment of the existing body of knowledge, including availability of experimental thermophysical property data, variable ranges studied, associated uncertainties, state of prediction methods, and parameters for deployment of prediction methods and how these parameters can be obtained using targeted measurements.

4 Summary: Advantages of the New Data Model

A new data model for storage of thermophysical property data was developed and implemented in the NIST/TRC SOURCE data archival system. While building upon prior experience [1–3], it presents significant improvements over the earlier schema. Numerous design deficiencies were eliminated. As a result, more rigorous representation of experimental data and improvements in overall data quality were achieved. The new data model is designed to be readily extendable in anticipation of future developments. Provisions for storage of additional information (e.g., supporting information,

annotated tables used in data capture, etc.) further improve data traceability and the usability of the system.

One of the most important aspects of the new data model is substance identification and indexing based on chemical structures. This significantly simplifies substance entry, management, and maintenance by making it independent of external registry numbers and complicated and potentially ambiguous chemical names. This is also expected to result in improvements in overall data quality, as many fewer errors associated with substance identification will be made. Finally, the availability of chemical structures (MOL files) provides opportunity for large-scale use of prediction methods to supplement experimental data during dynamic data evaluation with TDE [5–9]. Empirical methods based on group contributions can use group decompositions obtained by parsing stored two-dimensional structures. These structures can also be used for auto-generation of optimized three-dimensional structures and subsequent estimation of properties either directly from quantum-chemical calculations or indirectly, via empirical correlations based on quantitative structure–property relationships [41].

References

1. R.C. Wilhoit, K.N. Marsh, J. Chem. Inf. Comput. Sci. **29**, 17 (1989)
2. R.C. Wilhoit, K.N. Marsh, Int. J. Thermophys. **20**, 247 (1999)
3. M. Frenkel, Q. Dong, R.C. Wilhoit, K.R. Hall, Int. J. Thermophys. **22**, 215 (2001)
4. R.D. Chirico, M. Frenkel, V.V. Diky, K.N. Marsh, R.C. Wilhoit, J. Chem. Eng. Data **48**, 1344 (2003)
5. M. Frenkel, R.D. Chirico, V. Diky, X. Yan, Q. Dong, C.D. Muzny, J. Chem. Inf. Model. **45**, 816 (2005)
6. V. Diky, C.D. Muzny, E.W. Lemmon, R.D. Chirico, M. Frenkel, J. Chem. Inf. Model. **47**, 1713 (2007)
7. V. Diky, R.D. Chirico, A.F. Kazakov, C.D. Muzny, M. Frenkel, J. Chem. Inf. Model. **49**, 503 (2009)
8. V. Diky, R.D. Chirico, A.F. Kazakov, C.D. Muzny, M. Frenkel, J. Chem. Inf. Model. **49**, 2883 (2009)
9. V. Diky, R.D. Chirico, A.F. Kazakov, C.D. Muzny, J.W. Magee, I. Abdulagatov, K. Kroenlein, M. Frenkel, J. Chem. Inf. Model. **51**, 181 (2011)
10. M. Frenkel, Pure Appl. Chem. **77**, 1349 (2005)
11. M. Frenkel, J. Chem. Eng. Data **54**, 2411 (2009)
12. M. Frenkel, Comput. Chem. Eng. **3**, 393 (2011)
13. M.L. Huber, T.J. Bruno, R.D. Chirico, V. Diky, A.F. Kazakov, E.W. Lemmon, C.D. Muzny, M. Frenkel, Int. J. Thermophys. **32**, 596 (2011)
14. S. Watanasiri, Pure Appl. Chem. **83**, 1255 (2011)
15. M. Frenkel, R.D. Chirico, V. Diky, C. Muzny, Q. Dong, K.N. Marsh, J.H. Dymond, W.A. Wakeham, S.E. lStein, E. Königsberger, A.R.H. Goodwin, J.W. Magee, M. Thijssen, W.M. Haynes, S. Watanasiri, M. Satyro, M. Schmidt, A.I. Johns, G.R. Hardin, J. Chem. Inf. Model. **46**, 2487 (2006)
16. P.T. Cummings, T. de Loos, J.P. O'Connell, W.M. Haynes, D.G. Friend, A. Mandelis, K.N. Marsh, P.L. Brown, R.D. Chirico, A.R.H. Goodwin, J. Wu, R.D. Weir, J.P.M. Trusler, A. Padua, V. Rives, C. Schick, S. Vyazovkin, L.D. Hansen, Int. J. Thermophys. **30**, 371 (2009)
17. IUPAC, Commission on Nomenclature of Organic Chemistry. A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993) (Blackwell Scientific, Oxford, 1993)
18. Naming and Indexing of Chemical Substances for Chemical Abstracts, 2007 edn. (Chemical Abstracts Service, American Chemical Society, Columbus, OH, 2008)
19. CTfile Formats, Symyx Solutions, Inc., Sunnyvale, CA (June 2010)
20. Tripos MOL2 format, http://www.tripos.com/data/support/mol2.pdf. Accessed 21 Oct 2011
21. P. Murray-Rust, H.S. Rzepa, J. Chem. Inf. Comput. Sci. **39**, 928 (1999), http://cml.sourceforge.net. Accessed 21 Oct 2011
22. J. Gasteiger, T. Engel (eds.), *Chemoinformatics* (Wiley, Weinheim, 2003)

23. SMILES—A Simplified Chemical Language, Daylight Chemical Information Systems, Inc., Laguna Niguel, CA (2008), http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html. Accessed 21 Oct 2011
24. IUPAC International Chemical Identifier (InChI) Programs, InChI version 1, Software Version 1.03 (2010), User's Guide, http://www.inchi-trust.org. Accessed 21 Oct 2011
25. Extended SMILES, SMARTS format. ChemAxon, http://www.chemaxon.com/marvin/help/formats/cxsmiles-doc.html. Accessed 21 Oct 2011
26. OpenSMILES project, http://www.opensmiles.org. Accessed 21 Oct 2011
27. P. Lehman, The BibLatex Package, Version 1.16, July 29, 2011, http://mirror.ctan.org/macros/latex/contrib/biblatex/doc/biblatex.pdf. Accessed 21 Oct 2011
28. ChemDraw Ultra 12.0, CambridgeSoft Corporation, Cambridge, MA (2011)
29. ACD/Name to Structure. Version 12.0 for Microsoft Windows, Advanced Chemistry Development, Inc., Toronto, ON (2010)
30. G.A. Eller, Molecules **11**, 915 (2006)
31. V.V. Diky, R.D. Chirico, R.C. Wilhoit, Q. Dong, M. Frenkel, J. Chem. Inf. Comput. Sci. **43**, 15 (2003)
32. M. Frenkel, R.D. Chirico, V. Diky, C.D. Muzny, A.F. Kazakov, J.W. Magee, I. Abdulagatov, J.W. Kang, NIST ThermoData Engine, Version 5.0—Pure Compounds, Binary Mixtures, and Chemical Reactions, NIST Standard Reference Database #103b (Standard Reference Data Program, National Institute of Standards and Technology, Gaithersburg, MD, 2010)
33. M. Frenkel, R.D. Chirico, V. Diky, P.L. Brown, J.H. Dymond, R.N. Goldberg, A.R.H. Goodwin, H. Heerklotz, E. Königsberger, J.E. Ladbury, K.N. Marsh, D.P. Remeta, S.E. Stein, W.A. Wakeham, P.A. Williams, Pure Appl. Chem. **83**, 1935 (2011)
34. http://trc.nist.gov/ThermoML.html. Accessed 21 Oct 2011
35. Q. Dong, C.D. Muzny, A. Kazakov, V. Diky, J.W. Magee, J.A. Widegren, R.D. Chirico, K.N. Marsh, M. Frenkel, J. Chem. Eng. Data **52**, 1151 (2007)
36. NIST Ionic Liquids Database, ILThermo, NIST Standard Reference Database 147 (Standard Reference Data Program, National Institute of Standards and Technology, Gaithersburg, MD, 2006), http://ilthermo.boulder.nist.gov/ILThermo/mainmenu.uix. Accessed 21 Oct 2011
37. K. Kroenlein, C.D. Muzny, V. Diky, A.F. Kazakov, R.D. Chirico, J.W. Magee, I. Abdulagatov, M. Frenkel, J. Chem. Inf. Model. **51**, 1506 (2011)
38. NIST/TRC Web Thermo Tables (WTT), NIST Standard Reference Subscription Database 2-Lite Edition, Version 2-2011-3-Lite (Standard Reference Data Program, National Institute of Standards and Technology, Gaithersburg, MD, 2011), http://wtt-lite.nist.gov. Accessed 21 Oct 2011
39. NIST/TRC Web Thermo Tables (WTT), NIST Standard Reference Subscription Database 3-Professional Edition, Version 2-2011-3-Pro (Standard Reference Data Program, National Institute of Standards and Technology, Gaithersburg, MD, 2011), http://wtt-pro.nist.gov. Accessed 21 Oct 2011
40. K. Kroenlein, V. Diky, C.D. Muzny, R.D. Chirico, J.W. Magee, M. Frenkel, ThermoPlan: Experimental Planning and Coverage Evaluation Aid for Thermophysical Property Measurement, NIST Standard Reference Database #167 (Standard Reference Data Program, National Institute of Standards and Technology, Gaithersburg, MD, 2011)
41. A. Kazakov, C.D. Muzny, V. Diky, R.D. Chirico, M. Frenkel, Fluid Phase Equilib. **298**, 131 (2010)