Pure Appl. Chem., Vol. 73, No. 5, pp. 825–829, 2001.

Reliable solubility data in the age of computerized chemistry. Why, how, and when?*

John Rumble, Jr.†, Angela Y. Lee, Dorothy Blakeslee, and Shari Young

National Institute of Standards and Technology, 100 Bureau Drive MS 2310,Gaithersburg, MD 20899-2310, USA

Abstract: Since 1979, the International Union of Pure and Applied Chemistry (IUPAC) Commission V.8 on Solubility Data has published over 70 compilations of evaluated data on the solubility of gases in liquids, liquids in liquids, and solids in liquids. These volumes represent one of the largest collections of chemical property data ever produced and are the result of work of scientists throughout the world. In 1998, IUPAC signed an agreement with the National Institute of Standards and Technology (NIST) to continue the series by replacing the monographs by articles in the *Journal of Physical and Chemical Reference Data*. Five data compilations have already been published in the Journal, and many more are under way. Recently, IUPAC and NIST have concluded another agreement about computerizing all previously published IUPAC solubility data. In this paper, we describe in detail the computerization of IUPAC solubility data, with some emphasis on harmonizing data published over a long time period. We describe the anticipated query paths that will be supported. We also discuss some of the driving forces for making these and other data resources available over the World Wide Web.

INTRODUCTION

Every aspect of chemistry is being affected by the growth of chemical informatics and the Internet/World Wide Web explosion. The once tedious task of building databases and disseminating them widely has become very much easier. Today, some data gateways point to hundreds of web sites that provide some type of chemical information. The accessibility of these data is part of a larger effort both to improve the quality of scientific data and to make them as widely available as possible. Before examining the details of computerizing IUPAC solubility data, it is useful to examine some of the broader aspects of scientific data.

Modern computers are changing the nature of 21st century research just as profoundly. Already, industrial development and innovation flows primarily from computer-aided design, model-based processing and manufacturing and virtual testing. The confluence of increased computer power, advances in applied mathematics, and a new generation of highly computer-proficient scientists and engineers make the move to model-based research inevitable.

DATA EVALUATION AND RELIABILITY

Modeling, regardless of the discipline, has one common feature: *Reliable* data are an essential element. Model-based science and engineering cannot function properly without a large data collection of known

*Lecture presented at the 9th IUPAC International Symposium on Solubility Phenomena (9th ISSP), Hammamet, Tunisia, 25–28 July 2000. Other presentations are published in this issue, pp. 761–844.

†Corresponding author

quality. The expression "garbage in, garbage out" applies in every instance. The generation and dissemination of reliable data is a complex process. Most scientific and technical data are generated in the course of research not specifically focused on data measurement and quality. In fact, most data are scattered throughout the technical literature and are poorly documented. Data users are not usually experts in how data were generated. Consequently, even if they find needed data, they cannot easily determine the quality of those data.

Several organizations collect and evaluate data so that users may use measurements results more confidently. The process of critically evaluating data involves four key steps:

- collecting the data from the published literature;
- reviewing and evaluating data by experts;
- designing databases and publications to meet user needs; and
- disseminating those data collections widely.

The evaluation of scientific data proceeds from three viewpoints. First, the data are evaluated with respect to how well their generation is documented. Have all independent variables affecting the measurement been identified? Have they all been controlled during the measurement? And how have these facts been demonstrated and documented? The second viewpoint is how do the data follow the known laws of nature. The third viewpoint is how do the data compare with other measurements that purport to look at the same phenomena.

The mixture of these viewpoints depends on the maturity of the discipline and the existence of previous data evaluation efforts. In areas such as chemical thermodynamics and atomic spectroscopy, in which knowledge of the measurement technology is quite developed, the independent variables understood, and previous evaluations exist, the emphasis in new evaluations is on the latter two viewpoints. In areas where measurements are fairly new, or the phenomena are quite complex and not totally understood, the emphasis must be on the first viewpoint.

THE NIST DATA PROGRAMS

The National Institute of Standards and Technology (NIST) has long been interested in data evaluation. Beginning with the International Critical Tables [1] in the 1920s, the National Bureau of Standards, which was renamed as the National Institute of Standards and Technology in 1987, operated a large number of data evaluation activities [2]. Why is NIST interested in data evaluation? As the U.S. national laboratory concerned with advancing measurement science and technology, NIST considers data to be a fundamental result of measurements, both experimental and calculational. Data collections summarize previous measurement experience, and data evaluation therefore assesses the quality of current measurement technology.

NIST has unique broad expertise in measurement technology, and the knowledge and experience necessary to do data evaluation. NIST measurement experts are neutral, i.e., they do not favor any particular method except on merit. Data projects often involve partnerships on a national and international scale, and NIST has much experience in such partnerships in terms of sharing responsibility, costs, and outputs.

Today, NIST operates the Standard Reference Data Program, a network of data centers and projects covering about 40 scientific and technical disciplines. NIST operates 15 online data systems, available at no charge over the World Wide Web [3]. It also sells about 45 individual use databases, usually installable on PCs with graphical user interfaces (GUIs). Table 1 summarizes some areas in which strong data activities are maintained.

For many years, NIST has published the *Journal of Physical and Chemical Reference Data* with the American Institute of Physics (AIP). (Until 1999, the American Chemical Society was also a publishing partner.) NIST and AIP are now committed to creating an electronic journal and, as of 1 January 2000, an online, full text version of the *Journal of Physical and Chemical Reference Data* has been

Table 1 NIST standard reference data activities.

Chemistry	
Mass spectrometry	Protein structure
Chemical kinetics	Thermodynamics of enzyme-catalyzed reactions
Gas-phase infrared spectroscopy	Fluid properties
Alternative refrigerants	Critical stability constants
Chemical thermodynamics	X-ray photoelectron spectroscopy
Biotechnology	Solubility
Biomacromolecular crystallization	
Physics	
Fundamental physical constants	X-ray form factors and scattering
X-ray and gamma-ray attenuation	Atomic spectroscopy
Electron and positron stopping powers	Molecular spectroscopy
Materials Properties	
Crystallographic structure – bulk and surface	High-temperature superconductors
Phase equilibrium diagrams	Corrosion
Composite reinforcement permeability	Advanced ceramics performance
Construction and Fire Science	
Insulation materials	Refrigeration engineering
Fire test data	Spectral UV-B data
Electronics	
Plasma modeling for semiconductor manufacturing	Alternative dielectric gate materials
Information Technology	
Statistical reference datasets	Gray-scale and binary images
Fingerprint images	

available to subscribers. NIST is building a complementary database that contains important data from the tables and graphs of various articles, and it is planned that this will be available sometime in 2001. Eventually, we anticipate that the printed and online full text version of the Journal will be greatly reduced in size, and the majority of data themselves will be available through the Journal database.

NIST AND IUPAC SOLUBILITY DATA

In 1998, NIST and IUPAC signed an agreement to publish future volumes of the *IUPAC Solubility Data Series* in the *Journal of Physical and Chemical Reference Data*. As of the summer of 2000, five volumes (numbers 66 through 70) have been published (Table 2), and about four volumes per year are

Table 2 IUPAC Solubility Series volumes published in the *Journal of Physical and Chemical Reference Data*.

IUPAC-NIST Solubility Data Series 66. Ammonium Phosphates, J. Eysseltova and T. P. Dirkse, *J. Phys. Chem. Ref. Data* **27**, 1289–1470 (1999)

IUPAC-NIST Solubility Data Series 67. Halogenated Ethanes and Ethenes with Water, A. L. Horvath, F. W. Getzen, and Z. Maczynska, *J. Phys. Chem. Ref. Data* **28**, 395–627 (1999)

IUPAC-NIST Solubility Data Series 68. Halogenated Aliphatic Compounds C_3-C_{14} with Water, A.L. Horvath and F. Getzen, *J. Phys. Chem. Ref. Data* **28**, 649–777 (1999)

IUPAC-NIST Solubility Data Series 69. Ternary Alcohol-Hydrocarbon-Water Systems, A. Skrzecz, D. Shaw, and A. Maczynski, *J. Phys. Chem. Ref. Data* **28**, 983–1235 (1999)

IUPAC-NIST Solubility Data Series 70. Solubility of Gases in Glassy Polymers, R. Paterson, Y. Yampol'skii, P. G. T. Fogg, A. Bokarev, V. Bondar, O. Ilinich, and S. Shishatskii, *J. Phys. Chem. Ref. Data* **28**, 1255–1450 (1999)

planned under the five-year agreement. NIST is providing some help with respect to manuscript preparation, but the bulk of the work is still done by the individual volume editors with funding raised from their own sources.

With the explosion of web-based chemical information resources, IUPAC and NIST began discussions about how best to make the contents of the entire IUPAC Solubility Data Series available online. In 1999, NIST and IUPAC concluded an agreement that is intended to achieve this. Over the next five years, it is hoped that all data still valid will be made available via the web at no charge. The remainder of this paper discusses these plans and the planned system. Because over 70 printed volumes have been printed, many of which are not available in computerized formats of any type, a subset has been selected to determine the best approach to important issues. The first subset deals with the solubility of halogenated hydrocarbons in water and covers four volumes (Table 3).

Building the NIST–IUPAC solubility data systems involves many activities, including

- data entry (for volumes not computerized)
- data uniformity, including translation from old formats
- data verification, that the numbers are entered or translated correctly
- database design
- GUI (graphical user interface) design
- search strategies
- display formats

During the next few years, these issues will be explored using the subset identified above as a test bed. A prototype database design has already been developed, and data entry is proceeding for Vol. 20. While NIST is performing this work, IUPAC V.8 will play an important role in giving advice, reviewing the proposed design and checking the data.

Table 3 Solubility Data Series used in initial online data system.

IUPAC-NIST Solubility Data Series 20. Halogenated Benzenes, Toluenes and Phenols with Water, A. L. Horvath *et al*., Pergamon Press, Oxford (1985)
IUPAC-NIST Solubility Data Series 60. Halogenated Methanes with Water, A. L. Horvath *et al*., Oxford University Press, Oxford (1995)
IUPAC-NIST Solubility Data Series 67. Halogenated Ethanes and Ethenes with Water, A. L. Horvath, F. W. Getzen, and Z. Maczynska, *J. Phys. Chem. Ref. Data* **28**, 395–627 (1999)
IUPAC-NIST Solubility Data Series 68. Halogenated Aliphatic Compounds C_3–C_{14} with Water, A. L. Horvath, and F. Getzen, *J. Phys. Chem. Ref. Data* **28**, 649–777 (1999)

The proposed system will contain compiled experimental data as well as the critically evaluated recommendations. The printed volumes do have a limited number of expressions and formats for the data. It is NIST's experience, however, that printed text often contains subtle inconsistencies and ambiguities that are identified only upon computerization [4]. Additional problems exist in making computerized data uniform, especially if the definitive publication of the data is via printed page. For example, it is common practice for small changes to be made directly onto page proofs without alteration of the computer files that were input into an automated typesetting system. Documentation of such changes is usually nonexistent, which means that every number must be compared with the printed page, and discrepancies investigated.

In spite of the diverse data expressions, the search strategies are remarkably simple. As is widely recognized, the solubility of substance A in substance B can also be viewed as the solubility of substance B in substance A. Therefore, there is no need to designate uniquely the solute and solvent in

designing the underlying database that stores the data, nor to constrain search strategies unreasonably. The three search strategies that will definitely be supported are:

- find data on the solubility of substance A in substance B;
- find data on the solubility of A; and
- find data on the solubility of substances in substance B.

At first, searching on the solubility data themselves will not be supported, but depending on user needs, such as for liquid–liquid separations, this capability could be added later. To support these searches, methods for handling uncertainties need to be worked out. However, display of data in different units will be supported. In the future, additional features may be added, as requested by users, including performing calculations and generating plots.

For this project to be successful, NIST and IUPAC will have to work closely together on several aspects, including system priorities, additional search strategies, interface design, data quality, system functionality, and display of results. This partnership has already been successfully started as evidenced by the smooth transition to publication of new volumes of the Solubility Data Series. Both parties look forward to a long partnership to make solubility data readily available via the web.

ACKNOWLEDGMENTS

The authors would like to thank David Shaw and Mark Salomon of IUPAC Commission V.8 on Solubility Data for their dedication in making the NIST–IUPAC solubility project a reality.

REFERENCES

1. E. W. Washburn (Ed.). *International critical tables of numerical data, physics, chemistry, and technology*, published for the National Research Council by McGraw-Hill, New York (1926–30).
2. T. E. Gills *et al*., "NIST mechanisms for disseminating measurements," submitted for inclusion in *NIST Journal of Research*, **102**, no. 1 (2001).
3. http://www.nist.gov/srd
4. J. H. Westbrook. *J. Chem. Info. Comput. Sci*. **33**, 6–17 (1993).