Reliable Solubility Data in the Age of Computerized Chemistry: Why, How, and When?

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

IUPAC's *Solubility Data Series* (*SDS*), begun in the mid-1970s, is an exhaustive compilation and critical evaluation of all the world's published results of experimental determinations of solubility. Since 1979, over 70 *SDS* volumes have been published, including 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. Since 1998, and following an agreement with the National Institute of Standards and Technology (NIST), the *SDS* compilations have been published four times a year in the *Journal of Physical and Chemical Reference Data.* Since January 2002, a Subcommittee on Solubility and Equilibrium Data of the IUPAC Analytical Chemistry Division has continued the coordination of the *SDS*-related projects.

Although IUPAC has worked to maintain high scientific standards and a uniform approach for data compilation and evaluation throughout its existence, the evolution of personal computing and the rise of the Internet have substantially changed the ways that research is conducted and results are made available. In 1980, most communication between compilers, evaluators, and editors occurred through the mail, and most critically evaluated research data were published commercially in printed volumes. Only a few hundred libraries—virtually all in developed countries—maintained standing orders for the *SDS*. Today, collaboration is primarily via electronic means, with drafts passed around the world as e-mail attachments. Again with the help of NIST, and as described in this article, the computerization of the entire collection is being considered as a Web-accessible database.

Introduction

Every aspect of chemistry is being affected by the growth of chemical informatics and the Internet/Web explosion. The once tedious task of building databases and disseminating them widely has become 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 profoundly changing the nature of 21st-century chemistry research. Already, industrial development and innovation flow 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 makes 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 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 researchers and others may use measurement 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.

Modeling, regardless of the discipline, has one common feature: Reliable data are an essential element.

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 in which 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

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 perform 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 Web.[3] It also sells about 45 individual use databases, usually installable on PCs.

For many years, NIST and the American Institute of Physics (AIP) have published the *Journal of Physical and Chemical Reference Data*. NIST and AIP are now committed to creating an electronic journal and, since 1 January 2000, an online, full-text version of the *Journal* has been available to subscribers. NIST is building a complementary database, which will contain important data from the tables and graphs of various articles. 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 will be available through the *Journal* database.

NIST and IUPAC Solubility Data

In 1998, NIST and IUPAC signed an agreement to publish four volumes per year of the *IUPAC SDS* in the *Journal of Physical and Chemical Reference Data.* NIST is providing some help with respect to manuscript preparation, but the bulk of the work is still performed 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 *SDS* available online. In 1999, NIST and IUPAC concluded an agreement to achieve this. Over the next five years, it is hoped that all data still valid will be made available