FOREWORD

If the knowledge is
undigested or simply wrong,
more is not better

How to communicate and disseminate numerical data effectively in chemical science and technology has been a problem of serious and growing concern to IUPAC, the International Union of Pure and Applied Chemistry, for the last two decades. The steadily expanding volume of numerical information, the formulation of new interdisciplinary areas in which chemistry is a partner, and the links between these and existing traditional subdisciplines in chemistry, along with an increasing number of users, have been considered as urgent aspects of the information problem in general, and of the numerical data problem in particular.

Among the several numerical data projects initiated and operated by various IUPAC commissions, the *Solubility Data Project* is probably one of the most ambitious ones. It is concerned with preparing a comprehensive critical compilation of data on solubilities in all physical systems, of gases, liquids and solids. Both the basic and applied branches of almost all scientific disciplines require a knowledge of solubilities as a function of solvent, temperature and pressure. Solubility data are basic to the fundamental understanding of processes relevant to agronomy, biology, chemistry, geology and oceanography, medicine and pharmacology, and metallurgy and materials science. Knowledge of solubility is very frequently of great importance to such diverse practical applications as drug dosage and drug solubility in biological fluids, anesthesiology, corrosion by dissolution of metals, properties of glasses, ceramics, concretes and coatings, phase relations in the formation of minerals and alloys, the deposits of minerals and radioactive fission products from ocean waters, the composition of ground waters, and the requirements of oxygen and other gases in life support systems.

The widespread relevance of solubility data to many branches and disciplines of science, medicine, technology and engineering, and the difficulty of recovering solubility data from the literature, lead to the proliferation of published data in an ever increasing number of scientific and technical primary sources. The sheer volume of data has overcome the capacity of the classical secondary and tertiary services to respond effectively.

While the proportion of secondary services of the review article type is generally increasing due to the rapid growth of all forms of primary literature, the review articles become more limited in scope, more specialized. The disturbing phenomenon is that in some disciplines, certainly in chemistry, authors are reluctant to treat even those limited-in-scope reviews exhaustively. There is a trend to preselect the literature, sometimes under the pretext of reducing it to manageable size. The crucial problem with such preselection - as far as numerical data are concerned - is that there is no indication as to whether the material was excluded by design or by a less than thorough literature search. We are equally concerned that most current secondary sources, critical in character as they may be, give scant attention to numerical data.

On the other hand, tertiary sources - handbooks, reference books and other tabulated and graphical compilations - as they exist today are comprehensive but, as a rule, uncritical. They usually attempt to cover whole disciplines, and thus obviously are superficial in treatment. Since they command a wide market, we believe that their service to the advancement of science is at least questionable. Additionally, the change which is taking place in the generation of new and diversified numerical data, and the rate at which this is done, is not reflected in an increased third-level service. The emergence of new tertiary literature sources does not parallel the shift that has occurred in the primary literature.

With the status of current secondary and tertiary services being as briefly stated above, the innovative approach of the *Solubility Data Project* is that its compilation and critical evaluation work involve consolidation and reprocessing services when both activities are based on intellectual and scholarly reworking of information from primary sources. It comprises compact compilation, rationalization and simplification, and the fitting of isolated numerical data into a critically evaluated general framework.

The *Solubility Data Project* has developed a mechanism which involves a number of innovations in exploiting the literature fully, and which contains new elements of a more imaginative approach for transfer of reliable information from primary to secondary/tertiary sources. *The fundamental trend of the Solubility Data Project is toward integration of secondary and tertiary services with the objective of producing in-depth critical analysis and evaluation which are characteristic to secondary services, in a scope as broad as conventional tertiary services.*

Fundamental to the philosophy of the project is the recognition that the basic element of strength is the active participation of career scientists in it. Consolidating primary data, producing a truly critically-evaluated set of numerical data, and synthesizing data in a meaningful relationship are demands considered worthy of the efforts of top scientists. Career scientists, who themselves contribute to science by their involvement in active scientific research, are the backbone of the project. The scholarly work is commissioned to recognized authorities, involving a process of careful selection in the best tradition of IUPAC. This selection in turn is the key to the quality of the output. These top experts are expected to view their specific topics dispassionately, paying equal attention to their own contributions and to those of their peers. They digest literature data into a coherent story by weeding out what is wrong from what is believed to be right. To fulfill this task, the evaluator must cover *all* relevant open literature. No reference is excluded by design and every effort is made to detect every bit of relevant primary source. Poor quality or wrong data are mentioned and explicitly disqualified as such. In fact, it is only when the reliable data are presented alongside the unreliable data that proper justice can be done. The user is bound to have incomparably more confidence in a succinct evaluative commentary and a comprehensive review with a complete bibliography to both good and poor data.

It is the standard practice that the treatment of any given solute-solvent system consists of two essential parts: I. Critical Evaluation and Recommended Values, and II. Compiled Data Sheets.

The Critical Evaluation part gives the following information:

(i)   a verbal text of evaluation which discusses the numerical solubility information appearing in the primary sources located in the literature. The evaluation text concerns primarily the quality of data after consideration of the purity of the materials and their characterization, the experimental method employed and the uncertainties in control of physical parameters, the reproducibility of the data, the agreement of the worker's results on accepted test systems with standard values, and finally, the fitting of data, with suitable statistical tests, to mathematical functions;

(ii) a set of recommended numerical data. Whenever possible, the set of recommended data includes weighted average and standard deviations, and a set of smoothing equations derived from the experimental data endorsed by the evaluator;

(iii) a graphical plot of recommended data.

The Compilation part consists of data sheets of the best experimental data in the primary literature. Generally speaking, such independent data sheets are given only to the best and endorsed data covering the known range of experimental parameters. Data sheets based on primary sources where the data are of a lower precision are given only when no better data are available. Experimental data with a precision poorer than considered acceptable are reproduced in the form of data sheets when they are the only known data for a particular system. Such data are considered to be still suitable for some applications, and their presence in the compilation should alert researchers to areas that need more work.

The typical data sheet carries the following information:

(i) components - definition of the system - their names, formulas and Chemical Abstracts registry numbers;

(ii) reference to the primary source where the numerical information is reported. In cases when the primary source is a less common periodical or a report document, published though of limited availability, abstract references are also given;

(iii) experimental variables;

(iv) identification of the compiler;

(v) experimental values as they appear in the primary source. Whenever available, the data may be given both in tabular and graphical form. If auxiliary information is available, the experimental data are converted also to SI units by the compiler.

Under the general heading of Auxiliary Information, the essential experimental details are summarized:

(vi) experimental method used for the generation of data;

(vii) type of apparatus and procedure employed;

(viii) source and purity of materials;

(ix) estimated error;

(x) references relevant to the generation of experimental data as cited in the primary source.

This new approach to numerical data presentation, formulated at the initiation of the project and perfected as experience has accumulated, has been strongly influenced by the diversity of background of those whom we are supposed to serve. We thus deemed it right to preface the evaluation/compilation sheets in each volume with a detailed discussion of the principles of the accurate determination of relevant solubility data and related thermodynamic information.

Finally, the role of education is more than corollary to the efforts we are seeking. The scientific standards advocated here are necessary to strengthen science and technology, and should be regarded as a major effort in the training and formation of the next generation of scientists and engineers. Specifically, we believe that there is going to be an impact of our project on scientific-communication practices. The quality of consolidation adopted by this program offers down-to-earth guidelines, concrete examples which are bound to make primary publication services more responsive than ever before to the needs of users. The self-regulatory message to scientists of the early 1970s to refrain from unnecessary publication has not achieved much. A good fraction of the literature is still cluttered with poor-quality articles. The Weinberg report (in 'Reader in Science Information', ed. J. Sherrod and A. Hodina, Microcard Editions Books, Indian Head, Inc., 1973, p. 292) states that 'admonition to authors to restrain themselves from premature, unnecessary publication can have little effect unless the climate of the entire technical and scholarly community encourages restraint...' We think that projects of this kind translate the climate into operational terms by exerting pressure on authors to avoid submitting low-grade material. The type of our output, we hope, will encourage attention to quality as authors will increasingly realize that their work will not be suited for permanent retrievability unless it meets the standards adopted in this project. It should help to dispel confusion in the minds of many authors of what represents a permanently useful bit of information of an archival value, and what does not.

If we succeed in that aim, even partially, we have then done our share in protecting the scientific community from unwanted and irrelevant, wrong numerical information.

A. S. Kertes