

Numerical Data Retrieval in the U.S. and Abroad[†]

G. C. CARTER

National Academy of Sciences, Washington, D.C. 20418

Received February 12, 1980

A short overview is given of the Numerical Data Advisory Board, NDAB, of the National Academy of Sciences-National Research Council, and some historical developments leading toward today's picture of science reference data. Activities of the Committee on Data for Science and Technology, CODATA, of the International Council of Scientific Unions, will be described. Other national and international data programs, including the National Standard Reference Data System, will be briefly reviewed. A number of problems arising in numerical data handling (such as storage and retrieval, compilation, evaluation, dissemination, publication) will be reviewed and some CODATA and NDAB activities addressing these issues listed.

The function of the Numerical Data Advisory Board (NDAB) is to assess the adequacy of, and stimulate improvement of the quality, reliability, availability, accessibility, dissemination, utilization, and management of numerical data. Subject coverage broadly includes numerical data of the physical, chemical, biological, and geological sciences, and engineering and technology. The information sciences are necessarily an integral part of the subject matter. The Numerical Data Advisory Board itself provides a direct path to a major international data program in that it houses the U.S. National Committee for CODATA (the Committee on Data for Science and Technology) of the International Council of Scientific Unions (ICSU). From this vantage point, this paper will give an overview of chemical data activities, both government-sponsored and industrial, national and international. A comprehensive listing and review of all such activities is outside the scope of this paper.

SOME HISTORICAL DEVELOPMENTS

The National Research Council (NRC) of the National Academy of Sciences (NAS) has a long history of being an international focus for numerical data activities. In 1919, the International Union of Pure and Applied Chemistry established the "International Critical Tables". The executive, financial, and editorial responsibilities were given to the NRC. The International Research Council (which later became the ICSU) gave its endorsement to this project.

The "International Critical Tables" were contained in seven volumes,¹ published between 1926 and 1930. It was considered at the time one of the most comprehensive compilations to cover the broad range of physics, chemistry, and technology. This international effort was done in close collaboration with the National Bureau of Standards (NBS). Indeed, the editor-in-chief, Edward W. Washburn, was from that organization.

At the time this major work was done, it was thought that this single source book would satisfy most of the data needs in physics, chemistry, and technology. As science and technology progressed, new orders of magnitude of precision and new disciplines emerged. It became less and less practicable to produce a single data source. As a result, individuals in scientific subdisciplines set up specialized data projects. Some were in universities, others in national laboratories, or in industry. For example, a critical compilation project on Chemical Thermodynamical Properties² was taken up at NBS in 1940, representing essentially an outgrowth of work done for the International Critical Tables. This compilation is still being kept up to date by this group today.³

The NRC also continued to be concerned about numerical data, focussing on a few specific needs. In 1924 it had established within its Division of Physical Sciences, a Committee on Line Spectra of the Elements, in order to encourage and contribute to the structural analysis of atomic spectra, and eventually to publish the results in a series of monographs. However, this effort as well as other literature gave emphasis to analysis, paying little attention to numerical data. In 1946, the same NRC Committee established and sponsored at NBS an atomic energy levels data⁴ project.

The Subcommittee on Fundamental Constants of NRC's Committee on Physical Chemistry is another example. In 1952, it issued a set of recommended values of the fundamental constants for chemistry. In 1963, an international agreement was reached based on an internationally accepted temperature scale and a unified scale of atomic masses. The new Committee on Fundamental Constants of NDAB works together with ICSU's international bodies (CODATA and the International Union of Pure and Applied Physics, IUPAP), to continue to refine the best values of the internationally accepted fundamental constants, and their reliability.

The Numerical Data Advisory Board. Many more numerical data compilation activities of a wide variety of character evolved than can be enumerated. As these projects evolved independently, duplication of efforts increased, while leaving gaps elsewhere. This piecemeal approach to data activities plus the inherent problem of lack of continuity in such an approach, led the NRC to establish the Office of Critical Tables in 1957. This later became the Numerical Data Advisory Board.

CODATA, the Committee on Data for Science and Technology. CODATA was established as a committee of ICSU in 1966, largely as a result of efforts on the NRC's Office of Critical Tables. The establishment of CODATA formally expanded the collecting, collating, evaluating, and dissemination of data to the international arena. Subjects covered today by CODATA are physical, chemical, technological, biological, and geological data. The U.S. National Committee for CODATA is approved by the Chairman of the NRC and resides within the NDAB. It provides the main link between the international CODATA body and the U.S. scientific community, providing input from the U.S. to the CODATA program, on the one hand, and assisting in communicating CODATA accomplishments to the U.S. scientists, on the other hand. Many of the data concerns mentioned in the next section, including data retrieval, are shared on an international level, and CODATA has been successful in addressing a number of these issues. Some of the CODATA Task Groups relevant to chemical data and their dissemination are listed in Figure 1. These Task Groups address topics such as improvement of data dissemination, of data reporting in the primary literature, of data treatment and error assessment by

[†]Presented before the Division of Chemical Information, Symposium on "Techniques and Problems in Retrieval of Numerical Data", 178th National Meeting of the American Chemical Society, Washington, D.C., Sept 12, 1979.

SOME CODATA TASK GROUPS

ACCESSIBILITY AND DISSEMINATION OF DATA

CHEMICAL KINETICS

COMPUTER USE

(PRESENTATION IN PRIMARY LITERATURE)

DATA FOR CHEMICAL INDUSTRY

FUNDAMENTAL CONSTANTS

THERMODYNAMICS - KEY VALUES

- INTERNATIONALIZATION AND STANDARDIZATION OF DATA

TREATMENT AND EVALUATION - EDUCATION

Figure 1. Some CODATA Task Groups relevant to chemical data compilation and their dissemination. The fourth entry is in brackets due to a reassignment from a separate Task Group to the individual technical groups in the specific disciplines where needs exist.

authors of research, of documentation of this research, and of interchange of data.

The National Standard Reference Data System. In 1963, the U.S. Government officially recognized the problem of scientific and technological data by establishing the National Standard Reference Data System at NBS, administered by the Office of Standard Reference Data (OSRD). This was later strengthened by signing into law the Standard Reference Data Act. Thus, a new national focus for chemical and physical numerical data was established that could coordinate and stimulate data activities in technical areas where needs may arise. A description of the National Standard Reference Data System is given in ref 5.

The present relation between the NRC and the OSRD (NBS) is similar to its previous one in the days of the International Critical Tables. However, today it does not handle the funds to be expended for a single, comprehensive data compilation. Rather it advises the OSRD and other data programs in matters related to policy in data evaluation programs, while OSRD administers the funding of certain data compilation projects of broad applicability to the scientific community.

Other sources also support data activities, but often these tend to be more specific data programs oriented toward more specialized R&D such as, for example, in the development of coal gasification plants, energy storage systems, or nuclear applications.

OTHER NATIONAL AND INTERNATIONAL DATA PROGRAMS

This section briefly describes a few of the many existing data programs to illustrate the diversity of types of programs that have developed, in addition to those mentioned in the previous section. Other papers in this Symposium have described yet other data activities.

The Nuclear Data Network. Coordination of critical compilation of nuclear data, as carried out in various organizations, is done at Brookhaven's National Nuclear Data Center. In the case of A-chain data, the effort is international, bringing evaluations together, in a uniform format, as prepared in some eight or so countries. Also, a powerful evaluated nuclear structure numerical data file is being built up at Oak Ridge National Laboratory for online performance of numerical data manipulations, such as needed for correlations and graphics. Figure 2 shows two graphical displays produced using this system.

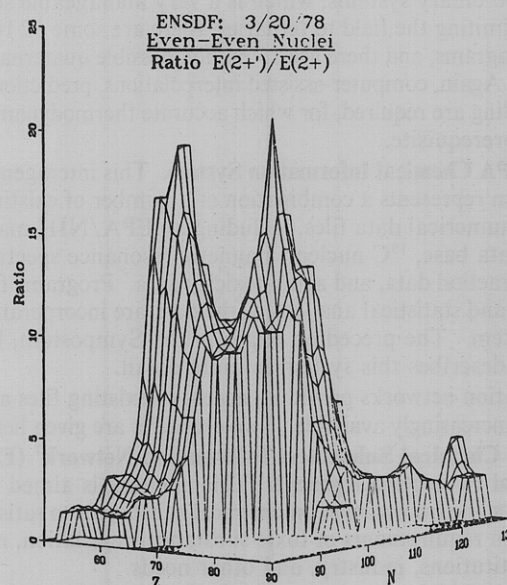
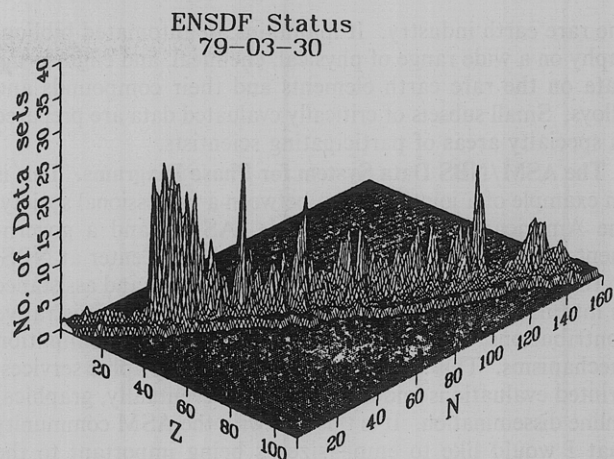


Figure 2. Graphical displays created with an automated numerical data file. Top: Status of data entered in the Evaluated Nuclear Structure Data File (ENSDF). Bottom: Graphical display after mathematical calculation using the same file.

A separate NRC Panel has been in existence for some time to help advise DOE and its precursors of inadequacies and needs in the nuclear area—the nuclear data picture is in far better shape today than many other areas of science and technology (for various reasons in addition to continuous scrutiny by leading nuclear experts).

Design Institute for Physical Property Data. This is a large industrial effort being established by the American Institute of Chemical Engineers. It will develop, organize, maintain, and make available a compilation of numerical data for about 1000 industrial compounds important in chemical process and equipment design. Chemical companies can join on an annual fee basis. Proposed evaluation projects are voted on by the companies using a weighting scheme based on annual sales of chemicals. The motivation for the establishment of this data base is that data are needed for better accuracy to be used in computer simulation of pieces of equipment and process flow sheets, for more productive and accurate designs. New federal regulations and product liability are in part responsible for an increased industrial need for reliable and accessible data banks.

Rare Earth Information Center. This is a relatively small data center, established in 1966, that is supported entirely by

the rare earth industry. It maintains an automated bibliography on a wide range of physical, chemical, and engineering data on the rare earth elements and their compounds and alloys. Small subsets of critically evaluated data are prepared in specialty areas of participating scientists.

The ASM/NBS Data System for Phase Diagrams. This is an example of a joint program between a professional society, the American Society for Metals (ASM), and a government-sponsored data center, the Alloy Data Center at NBS. In this program, NBS provides technical input and assistance in international coordination. ASM supplies administrative contributions, and most importantly, data dissemination mechanisms. This will include online bibliographic services, printed evaluations and drawings, and, eventually, graphical online dissemination. It is this link with the ASM community that I would like to emphasize as being important to the dissemination of these data for the wide user community reached by that society.⁶ This data program intends to coordinate phase diagram evaluation projects throughout the world for binary and multicomponent systems. There are under 3000 binary systems, which is a very manageable set. But even limiting the field to ternaries, there are some 32 160 possible diagrams, and there are 15 million possible quaternary diagrams. Again, computer-assisted interpolations, predictions, and modelling are required, for which accurate thermodynamic data are prerequisite.

NIH/EPA Chemical Information System. This interagency data system represents a combination of a number of existing chemical numerical data files, including an EPA/NIH mass spectral data base, ¹³C nuclear magnetic resonance spectra, X-ray diffraction data, and acute toxicity data. Programs for numerical and statistical analyses of the data are incorporated in the system. The preceding paper in the Symposium, by Hawkins, describes this system in more detail.

Information networks providing access to existing files are becoming increasingly available. Two examples are given here.

(1) The Chemical Substances Information Network⁷ (Environmental Protection Agency). This network is aimed at providing online information on chemical substances to satisfy information requirements of toxic substances legislation, research institutions, industry, and other needs.

(2) EURONET. An online information network, is being established for access from European countries. Through this network, a wide variety of data and information files, including several of U.S. origin, can be queried. The network will provide information or data for scientists, engineers, managers, documentalists, information scientists, environmentalists, and legal and socioeconomic data.

The following are some information and data projects carried out under international auspices other than CODATA.

(1) Solubility Data Project. This is a project of the International Union of Pure and Applied Chemistry's Analytical Chemistry Division, coordinated at the Hebrew University in Jerusalem, Israel. This critical compilation program receives contributions in kind from groups throughout the world. A standard format has been agreed upon so that the diversified sources of input will produce a uniform set of data. An example of a page is shown in Figure 3. Subsets are published by Pergamon Press. The final product is projected to amount to some 80 volumes of 300 pages each in a time span of from 10 to 15 years.

(2) The International Atomic Energy Agency. One of their data projects is preparing eleven volumes of evaluated chemical thermodynamic data for the actinide elements, compounds, and alloys.

(3) The General Information Program (PGI) of UNESCO. This is an intergovernmental program which has five major areas of concern: (1) promotion of information policies and

Solubility of Chlorine in Carbon Tetrachloride

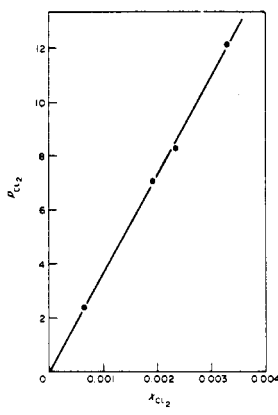
COMPONENTS: (1) Chlorine; Cl ₂ ; 7782-50-5 (2) Methane, tetrachloro; CCl ₄ ; 56-23-5		ORIGINAL MEASUREMENTS: Jones, W.J. <i>J. Chem. Soc.</i> 1911 , 99, 392.																					
VARIABLES: Concentration ?		PREPARED BY: W. Gerrard, Jan. 1976																					
EXPERIMENTAL VALUES:																							
<table><thead><tr><th>T/K</th><th>s</th><th>p_{Cl₂} /mm Hg</th><th>x_{Cl₂}</th></tr></thead><tbody><tr><td>288.16</td><td>50.7</td><td>2.34</td><td>0.000634</td></tr><tr><td>:</td><td>51.4</td><td>7.03</td><td>0.00193</td></tr><tr><td>:</td><td>53.7</td><td>8.29</td><td>0.00237</td></tr><tr><td>:</td><td>51.0</td><td>12.14</td><td>0.00331</td></tr></tbody></table> <p>(Extrapolation to p_{Cl₂} = 1 atm is not valid.)</p> <p>760 mm Hg = 1 atm 1 atm = 1.013 × 10⁵ Pascal</p>		T/K	s	p _{Cl₂} /mm Hg	x _{Cl₂}	288.16	50.7	2.34	0.000634	:	51.4	7.03	0.00193	:	53.7	8.29	0.00237	:	51.0	12.14	0.00331		
T/K	s	p _{Cl₂} /mm Hg	x _{Cl₂}																				
288.16	50.7	2.34	0.000634																				
:	51.4	7.03	0.00193																				
:	53.7	8.29	0.00237																				
:	51.0	12.14	0.00331																				
AUXILIARY INFORMATION																							
METHOD: Chlorine, carried away from the solution in a stream of air, was determined by iodometric titration. The results were given as a partition coefficient, $s = \frac{\text{concn. Cl}_2 \text{ in moles/dm}^3 \text{ solution}}{\text{concn. Cl}_2 \text{ in moles/dm}^3 \text{ gas phase}}$ s was deemed to be constant and independent of pressure, p _{Cl₂} .		SOURCE AND PURITY OF MATERIALS: Cl ₂ , not specified. CCl ₄ , redistilled.																					
APPARATUS/PROCEDURE: Bubbler and wash bottles. Concentration of the original solution determined iodometrically.		DATA CLASS:																					
NOTES: Compiler has given the approximate p _{Cl₂} and x _{Cl₂} values on the assumption that the volume of solution is equal to that of the original liquid for these low concentrations.		ESTIMATED ERROR:																					
		REFERENCES:																					

Figure 3. Sample page illustrating structured format of the evaluated data sheets for an internationally cooperative critical evaluation task undertaken by a IUPAC group, the "Solubility Data Project".

plans on a regional, national, and international level; (2) development of norms and standards of information handling; (3) development of information infrastructures, primarily in developing countries; (4) development of specialized information systems in science education and culture; (5) education and training of information specialists and users. The greater part of this large effort is devoted to infrastructure development and education activities mostly in developing countries. PGI-UNESCO and CODATA have collaborated in a number of programs. Examples are a series of education and training courses in data handling that have been, and will continue to be, held in various countries, and the preparation of a Sourcebook⁸ on data handling for science and technology.

The U.S. interacts with PGI in two ways: (1) an NAS advisory committee on science in UNESCO, advisory to the State Department; (2) a newly established general advisory committee on general information programs. The latter committee provides general advice on the State Department on the PGI programs, identifies U.S. participants for PGI, and seeks advice from groups such as USNC/CODATA.

ISSUES AND PROBLEMS IN CRITICAL COMPILATION

This section describes some of the prominent ingredients

RETRIEVAL PROBLEMS - USER'S VIEW

CAN DATA BE TRUSTED

- CHARACTERIZATION?
- MEASURED, INTERPOLATED/EXTRAPOLATED?
- PREDICTED?
- SOURCE OF DATA?

NON-UNIFORMITY OF DATA

- DIFFERENT UNITS
- DIFFERENT REFERENCE STANDARDS
- DIFFERENT EVALUATION CRITERIA

IDENTIFY DATA SOURCE

- CONFLICTING EVALUATIONS
- LACK OF DATA
- UNAVAILABILITY (PROPRIETARY)

HOW TO QUERY SYSTEM

- INDEXES
- AUTOMATED
 - BIBLIOGRAPHIC
 - NUMERICAL TABLES, GRAPHS

Figure 4. Some problems encountered in data retrieval by users of data compilations.

in compilation and critical evaluation that affect, or are affected by, data retrieval. There are those encountered by the user of the compilations, and those encountered by the provider.

Problems in Data Retrieval—Users View. Figure 4 outlines some of the types of questions asked by the conscientious user.

- Is sufficient characterization information given to make the data useful? What was the source of the material, the impurity level of certain contaminants, thermal history, etc.?

- Are the data evaluated from experimental measurements, or are they interpolated or extrapolated—or are they a blend of experiment mixed in with least squares or other statistical treatment to an expected (theoretical) expression?

- If so, what models were used; how would the data have looked if another model would have been used—what is the variation between the two treatments?

- What literature references were used—is it work from early years—at which laboratory(ies) was the work done? Are any references cited at all?

- Which system of units is being used? (The internationally accepted system of units (SI) is not adhered to by everybody.) Often good critical data are ignored because they either are or are not using SI units depending on whether you were educated in one or the other system (e.g., scientists vs. engineers). These questions are far from trivial in fields such as electromagnetism. At times, compilers respond to this problem by giving the data in both units (e.g., providing a phase diagram with both an atomic percent and a weight percent scale; this is not a linear transformation.)

- Is the reference standard stated? Have the data been adjusted correctly to a common standard? How was the conversion made? What was the value of the correction?

- Was the evaluation made for ball park figure use? Is it strictly an evaluation of experimental data so that a new theory can be tested against it, or does the evaluation consciously or unconsciously include previous theoretical considerations? Does the listed error represent statistical variations only, or also include sample-to-sample variations and other experimental errors?

- Sometimes more than one evaluation is available. How does one choose? Is the more recent one better? Usage of a handbook "already on the shelf" is common and can deprive

RETRIEVAL PROBLEMS - "STRATEGIC"

EDUCATION

- USER
- EVALUATOR
- DATA ORIGINATOR

FILE MAINTENANCE

- UPDATING
- QUALITY CONTROL
- STATE-OF-THE ART ADP USE

ADP TECHNOLOGY

- DEVELOP SYSTEMS TO SUIT NEEDS
- NETWORKING
- INTERFACING

DIVERSITY/COORDINATION

- DATA ACTIVITIES
- SPONSORS
- USER NEEDS
- FORMATS
- TECHNOLOGIES

DETERMINE END USE

- BROAD/NARROW?
- RELIABILITY ADEQUATE FOR APPLICATION?
- FORMAT USEFUL FOR APPLICATION?
- MISUSE OF DATA LIKELY?

DISSEMINATION

- MARKETING; COST-EFFECTIVENESS
- PUBLISHING
 - FORMATS
 - COPYRIGHT
 - LINKING TO POTENTIAL USERS
- HUMAN ASSISTANCE
 - LIBRARIAN/INFORMATION SPECIALIST
 - DATA CENTER SPECIALIST
 - REFERRAL SPECIALIST

Figure 5. Some important elements in the design and maintenance of a data system.

scientists of better reference data, but sometimes earlier data remain valid while more recent data can be incorrect.

- How does one verify that no data exist at all for a particular property of a particular substance?

- Sometimes data have been measured or evaluated for internal company use or other classified access. Or, they are in the wrong format and therefore effectively not available—this becomes especially true for large data bases that are on card files rather than machine readable.

- Are there indexes, or indexes of indexes available to direct users to the right place in the information center or library?

- In the case of automated files, how does one sign on? How does one identify key words? The user's jargon may be different from the system's. In the case of numerical tables, how does one ask for the needed number? Can the tables be printed in numerical increments? Can they be retrieved graphically?

Problems in Data Retrieval—Considerations for the Data Provider. Figure 5 summarizes some important considerations to be factored into an effective data system, referred to as "strategic" considerations, to facilitate data retrieval. The next paper in this Symposium expands further on some of these points.

- Better education of the user is certainly one contributing ingredient in improving the flow and proper use of information. A survey through questionnaire is summarized by Sarkisian:⁹ "...the responses overwhelmingly supported altering curricula

on both the undergraduate and graduate level to better prepare the future generation of chemists in information retrieval." Not too many years ago, scientists could get by with much less effort. Today, with so much more information available, and the development of automated retrieval methods, better education in data retrieval is needed.

- Methods of data evaluation should also be taught. This includes treatment of data, statistical and otherwise. Both the originator of the primary literature and the evaluator need a thorough background in this. The CODATA/UNESCO training courses mentioned under PGI, above, address this need. Additionally, the evaluator should have ample experimental training in the subject matter being evaluated, and must keep current with theoretical developments and changes in experimental methods. It may take several years to train an already competent scientist to become a data evaluator—to learn how to track down errors, which are the common errors, who are the most reliable authors, how not to panic at a desk full of conflicting papers which, typically, provide insufficient detail to be able to unambiguously arrive at a "best value".

- The author of research papers often could do better in the presentation of data, and in data treatment and massaging before publication. There are entirely too many papers being published that are virtually useless to the evaluator because
 - the experimental setup is not described
 - the sample treatment is not described
 - the sample analysis is not stated
 - the reference standard against which the measurement was made is not given
 - the numerical data upon which the conclusions are based are at times not even given.

Often an evaluator must simply discard published data when presented inadequately. This effectively represents a loss of data. Both researcher and referee of the journal should become more sensitized to such basic criteria for scientific reporting.

- File maintenance is highly important. Usefulness of a data file often drops sharply with age, especially in new and developing fields. Budgetary requirements for this important function are often underestimated. Similarly, updating of ADP techniques used in data centers should be periodically reviewed. At a certain point, it becomes cost effective to update the data system.

- ADP technology is advancing at a fast pace, and the equipment and computer time are becoming more and more within budgetary reach. Considerable detail has been devoted to the description of one automated system developed at the Office of Standard Reference Data (NBS) in the paper by Molino in this Symposium.

- As is illustrated in the previous section of this paper, there is a wide diversity in data projects and programs, data uses, needs, applications, ADP technologies, data center/user interfaces, and so on. Increased coordination of data activities is generally desirable to reduce overlap of efforts and costs, and to increase interchange and compatibility of data and data systems. An increase in cross-checking of data that are interrelated by well-established theories, and are evaluated by independent groups, would further enhance accuracy of the data.

- Determination of end-use of the data is important. What are the user needs? Will the user span a broad range of interests or disciplines? How foolproof should the system be? Or how elementary should the steps be in which the data files are to be queried? Is the user going to know what the tolerance or error bar means? There is a tendency for data users to take what is flashed on a screen as gospel, even more so at times than from the printed page. Will the user be too naive to read the footnotes, disclaimers, and warnings written all over the tabulated numbers? Will he incorporate these data in more

condensed tables that will ignore the footnotes?

Questions of this sort must be asked before starting a data evaluation project. Users generally assume handbook data to be 100% right, often minimizing the importance of the stated error of the numerical value. And, consequences of using data incorrectly or ignoring the estimated error, reliability factor, or tolerance can be serious. They can be costly and in some instances even life-threatening.

Three prominent dissemination issues are noted in Figure 5. *Marketing* has been discussed in the first paper in this Symposium, by Murdock.

Publishing of data brings many problems with it. There are page charges, which may limit an author's desire for preparing detailed documentation. There is the push to publish it soon! This causes documentation of data treatment to suffer even more because then "letters", "communications", "comments", and "abstracts" are published, which tend to give very little experimental detail. Then there are certain editorial rules ("don't bore the reader with lengthy tables—give interpretations instead"). This type of paper has a very short half-life as a rule, because the interpretation may not even survive the next publication on the subject. And since the data themselves are not properly documented, the paper cannot be used for compilation/evaluation purposes either. Authors sometimes resort to giving data in graphical form so that the "lengthy-tables" objection may not hinder speedy publication of the work. But often these graphs get squeezed to the size of a stamp by the publisher, straining their function of data dissemination.

Sometimes though, the author is at fault. He or she sometimes will choose to graph the data as a function of a parameter other than one that would be expedient to pinpoint the numerical data. Examples are a plot of resistivity vs. thermal conductivity, or magnetic susceptibility vs. nuclear magnetic resonance shift, both with temperature as the intrinsic parameter. Sometimes the literature source of the second parameter is not even given, or neither parameter is given as a function of the common variable, temperature, or worse yet, the data are adjusted before incorporation in the plot without letting the reader know.

There are data repositories to which the author of a paper can transfer the numerical data upon which a publication is based in any detail and length. The fact that these data are available in the repository can then be noted in the publication. However, this system remains mostly unused at this time. The new copyright law has put dissemination in a spot other than it has been, although we have no clear picture of all the ramifications at this time. A different problem in copyrighting is in propriety of the numbers disseminated by online methods. Often, particular subsets of data are drawn from different sources. Some large online files draw from automated files held by different authors.

With the *topic of human interactions* is meant the whole multidimensional spectrum of users. At one extreme of this spectrum are those researchers who shy away from asking the "librarian" and prefer to do the job themselves, even though the results may be poor. They are inhibited for one reason or another to ask nontechnical personnel a technical question, and prefer to leave only the reshelfing of books to them. At another extreme of that spectrum there are those who make heavy use of their organization's information center, where many technical inquiries are directed to the information specialist. In my experience in a data center, frequently an information specialist has called me and relayed the request to me. If it turned out to be in my data center's jurisdiction, I could at times simply answer the information specialist, or otherwise obtain the scientist's telephone number to talk to him directly. The important point is that he was led to the

NDAB & CODATA - DISSEMINATION & HANDLINGSOME CODATA PROJECTS - NUMERICAL

SOURCE BOOK ON HANDLING SCI & TECH DATA

STUDY COMPUTER PREPARATION OF SCI DATA

STUDY OF STANDARD INTERCHANGE FORMAT FOR COMPUTERIZED NUMERICAL DATA

GRAPHICS - STATE-OF-ART REVIEW

EDUCATION COURSES

GUIDES FOR DATA PRESENTATION IN THE LITERATURE

DIRECTORY OF DIRECTORIES

LISTS OF DATA HANDBOOKS, DATA CENTERS, REFERRAL CENTERS

RATE CONSTANTS & RELATED DATA

, CA 150 ATMOSPHERIC REACTIONS

FUNDAMENTAL CONSTANTS

KEY VALUES FOR THERMODYNAMICS

INTERNATIONAL AGREEMENT BELGIUM, CANADA, SWEDEN,
UK, USA, USSR

SYSTEMIZATION & INTERNATIONALIZATION - THERMODYNAMICS

DATA FOR INDUSTRY: SURVEY OF NEEDS

OF ESTIMATION PROCEDURES

Figure 6. NDAB and CODATA tasks addressing data dissemination issues.

contact via the human assistance path.

SEEKING SOLUTIONS

NDAB and CODATA maintain an awareness of these problems both in the broad sense and as seen for particular disciplines. They then form groups on the national or international level, respectively, to set out to provide assistance in alleviating them. Figure 6 lists some projects aimed at facilitating data dissemination. The education courses were mentioned earlier under UNESCO's General Information Program. Note also the guides for presentation in the literature. These are in an effort to improve the quality of data reporting in the literature, and there are some indications that they have a positive effect in areas for which they have been produced.

Master directories and lists of sources form another method of linking users to sources, especially for interdisciplinary purposes. They also form an important link in data dissemination for less developed countries.

The Sourcebook⁸ provides an introductory survey of the basic aspects of handling scientific and technical data, and indicates to the reader selected sources from which more details can be obtained. The text is addressed to a varied body of users, including those who generate, publish, abstract, collect, evaluate, repack, disseminate, and apply data, as well as those who provide training courses in the handling of data, and those who administer the funding for all these activities.

Figure 7 lists other task groups engaged in data reliability aspects. One that may be of particular interest to this Symposium is the "Key Values for Thermodynamics".¹⁰ A set of thermodynamic values has been agreed upon by representatives from the countries indicated in the figure. Once these key values become used as reference for other measurements, these later measurements, based on a consistent set of data, will be more readily used by scientists throughout the world. The next step in this work is to provide wide dissemination of these data and promote their use throughout that international scientific community.

CONCLUSIONS

The conclusion of this paper is that much more is needed in all the areas covered: more evaluation, more compilation, more education, more improvement of data reporting, coordination, dissemination, usage, graphics development, and so on. This is not only my own conclusion. An NRC committee was convened specifically to study the problem of "National Needs for Critically Evaluated Physical and Chemical Ref-

Figure 7. Some CODATA tasks aimed at numerical data evaluation, systemization, and standardization.

erence Data".¹¹ This group was able to quantify the needs for data programs. In its report it notes that even to catch up with our current backlog in existing data centers, the total data effort should be nearly tripled. And this assessment of needs is independent of the needs for increased dissemination, which further aggravates the problem.

Some excellent data publication channels have already been created to provide better linkage. For example, there is a joint publication by the American Chemical Society, the American Institute of Physics, and NBS, "The Journal of Physical and Chemical Reference Data", which receives publicity through each organization. There are other effective data journals mostly published by commercial publishers, and there are various other valuable channels, but not all compilations find their way to channels with ready access, or that are scanned by abstracting services. Also, the published page is far from the only method of data dissemination, and the other methods (computer access, networking, human interaction) should also be given attention.¹² Continued efforts should be made on all fronts. Yet, I am personally especially concerned that those data already measured and those already evaluated, that is, those in which large sums of money and technical scrutiny have already been invested, by and large, do not yet find their way readily and timely to many of those who would profit from them. This Symposium has brought out many facets of the problem. Much work lies ahead.

Note Added in Proof. Figures with scientific data are intended only to illustrate data formats and handling techniques. *Numerical values should not be used*; data publications or centers should be consulted for quantitative data.

REFERENCES AND NOTES

- (1) "International Critical Tables", Washburn, E. W., Ed. Published for the National Research Council by McGraw-Hill: New York: Vol. I, 1926; Vol. II, 1927; Vol. III, 1928; Vol. IV, 1928; Vol. V, 1929; Vol. VI, 1929; Vol. VII, 1930; Index, 1933.
- (2) Rossini, F. D., Wagman, D. D., Evans, W. H., Levine, S., Jaffe, I. "Selected Values of Chemical Thermodynamic Properties", NBS Circular 500; Government Printing Office: Washington, D.C., 1952.
- (3) Values reported in ref 2 above are being kept up to date by Wagman, D. D., et al., in a NBS Technical Note 270 series starting in 1968. Available from the National Technical Information Service, Springfield, Va. 22151.
- (4) Moore, C. E. "Atomic Energy Levels", NBS Circular 467; Government Printing Office: Washington, D.C. 20402: Vol. I, 1949; Vol. II, 1952; Vol. III, 1958. This project continued to prepare revised and other evaluated data, serving data needs in astrophysics as well.
- (5) NBS Technical Note 947, "Critical Evaluation of Data in the Physical Sciences—a Status Report on the National Standard Reference Data System", January 1977, Rossmassler, S. A., Ed.; U.S. Government Printing Office: Washington, D.C., May 1977.
- (6) The "Alloy Phase Diagram Bulletin" is planned as distribution medium to be published by the American Society for Metals. This Bulletin will circulate subsets of critically evaluated phase diagrams.
- (7) "The Chemical Substances Information Network by the Public Liaison Subcommittee of the Interagency Toxic Substances Data Committee";

- U.S. Environmental Protection Agency: Washington, D.C., April 1979.
- (8) "Data Handling for Science and Technology—An Overview and Sourcebook", Watson, D. G., and Rossmassler, S. A., Eds.; North Holland: Amsterdam, 1980.
 - (9) Sarkisian, J. E. "The Status of the Teaching and Use of Chemical Information in Academe", *Chem. Info. Bull.* 1979, 31, No. 2, 11.
 - (10) CODATA "Recommended Key Values for Thermodynamics", CODATA Bulletin No. 28, April 1978. This and other CODATA Bulletins can be obtained from: CODATA Secretariat, 51 Boulevard de Montmorency, 75016, Paris, France.
 - (11) "National Needs for Critically Evaluated Physical and Chemical Data", Committee on Data Needs, Numerical Data Advisory Board, National Academy of Sciences, 1978.
 - (12) Problems in data dissemination of "nonpublications" take on a staggering scale in the geosciences, and solutions which must be found in this area may facilitate progress in the chemical sciences. See, for example, DeGraffenreid, J. A. "Changing Patterns in Geoscience Communication, and the Proliferation of 'Non-Publications'", Annual Meeting of the Association of Earth Science Editors, Tulsa, OK, Oct 14-17, 1979.

More Questions from a Data Compiler†

DONALD M. KIRSCHENBAUM*

Department of Biochemistry, College of Medicine, Downstate Medical Center, State University of New York, Brooklyn, New York 11203

Received February 12, 1980

What are the duties, obligation, and responsibilities of a numerical data compiler? This and other questions are examined but not answered.

For the past eight years I have been compiling data about proteins. Why do I compile, what do I compile, and how do I compile? I compile data because I had a need for the kinds of data I compile and no data compilations were available. I also compile because of what I once read: this sentence by Herbert Spencer—"Science is organized knowledge".¹ I liked the "organized" part. I believe in the organization of numbers we need but can't find when we need them.

What I compile are amino acid analyses of protein² and ultraviolet and visible absorption spectra data.^{3,4}

How do I compile these data? I examine some 250-300 separate issue numbers of journals (this is about 20-30 journals) every year, page by page, looking for these data.⁵ When I find the data I need, there are two paths I can take: (1) I make a note of the volume of the journal, page or pages in that number on which the data appear. When I collect a large amount of such information, I have the pages containing the data photocopied. The first page of the article, which provides the title, the names of the authors, the volume number, inclusive pages, and the name of the protein and its source, are also copied as this is the information I need in order to publish the data compilations. (2) When I find simple numerical data, $A_{\text{cm}}^{1\%}$ or molar absorption values, then I prepare a 4 × 6 in. card containing all the necessary information, i.e., protein name, source, title of the article, author(s), journal name, volume number and inclusive pages, and numerical data.

Occasionally I must send for a reprint of an article which I believe contains the data I need and occasionally I get back a totally useless response (Figure 1). If I am fortunate there is an illegible signature which I may be able to decode and so be able to trace the article desired. My last resort is to use the excellent interlibrary loan service available to me.

I file the photocopied pages alphabetically in notebooks and the file cards in boxes. The data now sit on shelves waiting to be used. To which journals can I send compilations of data? In the past I was able to send my numerical data compilations to a journal which would publish them. This is no longer feasible and I now must seek another journal to accept the data compilations, and while I seek such a journal the data ages. How long can data be kept? Are old data useful? Even if

replaced by newer and more exact data? I don't believe that data have a life span. Old data are useful but not for the same reasons that they were useful when they were new data. Old data are useful because of what they tell about the methods and equipment used to obtain the data at that time in the past. For example, when I compare the amino acid analysis of a protein obtained by column chromatography with the more recently obtained analysis obtained from sequence analysis of the protein, I can learn something about the stability of certain amino acids to the hydrolytic conditions used, about the hydrolytic stability of the bond between certain amino acids, about the ability of the technique used to separate and quantitate the amino acids, and something about the accuracy of the total technique. I find that a retrospective examination of this kind is a useful teaching tool.⁶

Some time after I started my compiling activities I realized I could use some support funds. I applied for a grant and was not awarded one. One of the reasons given was that "my" data were neither reliable nor valid. Figure 2 gives the definitions of reliable and valid.

It is necessary to remind the reader that "my" data were not mine but were taken from articles published in reputable journals after review. In my naivete I assumed that the data contained in such articles were reliable and valid because the reviewer had checked the data.

I was thus led to wonder about the numerical data I was compiling from reviewed articles published in reputable journals. Who decides if the numerical data in the article are valid and reliable? How is this decision made? I don't have answers to these questions. I have been checking the numerical data I compile for what I like to call "Internal Consistency" (see Figure 3). I define internal consistency as the agreement between an experimentally determined numerical datum describing a property of the protein, in this case its absorption at 280 nm, and a value for the same property calculated from other available information, the amino acid analysis of the protein.⁷ If these two values, one experimentally determined and one calculated from another set of experimentally determined numbers, agree within the assigned limits,⁷ then there is internal consistency. If there is a lack of agreement, then there is an error in one or both of the experimentally determined numerical values.

What do I do now that I can check published numerical data for internal consistency? Do I, as a compiler, correct incorrect

†Presented before the Division of Chemical Information, Symposium on "Techniques and Problems in Retrieval of Numerical Data", 178th National Meeting of the American Chemical Society, Washington, D.C., Sept 12, 1979.

* Faculty Exchange Scholar, State University of New York.