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 protein[2] 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, $A^{1\%}_{1\text{cm}}$ 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.

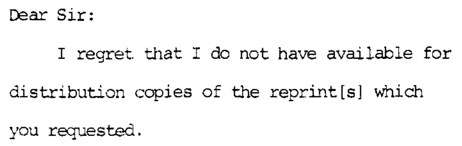
I was thus led to wonder about the numerical data I was compiling from reviewed articles published in reputable journals. Who decides if the numerical data in the article are valid and reliable? How is this decision made? I don't have answers to these questions. I have been checking the numerical data 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.

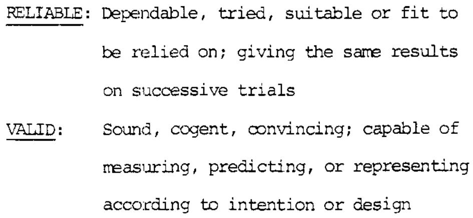
0095-2338/80/1620-0152$01.00/0



Dear Sir:

I regret that I do not have available for distribution copies of the reprint[s] which you requested.

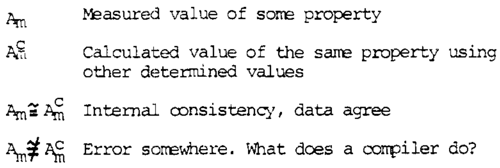
Figure 1. Unsigned response to reprint request.



RELIABLE: Dependable, tried, suitable or fit to be relied on; giving the same results on successive trials

VALID: Sound, cogent, convincing; capable of measuring, predicting, or representing according to intention or design

Figure 2. Definitions of reliable and valid.



A_m	Measured value of some property
A_m^C	Calculated value of the same property using other determined values
$A_m \cong A_m^C$	Internal consistency, data agree
$A_m \ncong A_m^C$	Error somewhere. What does a compiler do?

Figure 3. Illustration of internal consistency and related problem for a compiler of numerical data.

data? If I do correct the data and publish it in a compilation, then I am no longer publishing original data. At this point I feel that I am in a "Catch-22" situation. If I correct the data, then as a compiler I am inserting myself into the data. If I do not correct the data, then as a scientist I feel it is wrong to publish the incorrect data. Should a data compiler be a data evaluator? Should a data compiler get involved with the data compiled? What are the duties, obligations, and responsibilities of a numerical data compiler?[8]

I wish to now raise a question of a more personal nature. I want to know why compiling numerical data, organizing science, is not a good scientific endeavor for a scientist to pursue, while the use of the compiled data is a good scientific endeavor leading to promotion. A scientist who spends a good part of his time compiling numerical data would do better professionally if he used that time in some laboratory endeavor. I wish to raise a question and to provide the answer to it. The question: How do numerical data used by scientists get in the handbooks these scientists use? The answer: some capable, knowledgeable scientist sits down and compiles them and prepares them for publication. Both user and provider are doing useful scientific endeavors.

I close with two statements describing what all compilers of numerical data do in expressing the usefulness of their work. "I have only made a nose gay of culled flowers, and have brought nothing of my own but the thread that ties them together",[9] and "Some may think that the same talents and industry would be better devoted to original work; but it must be allowed that to elucidate and render accessible the labours of others may be a service as valuable as the addition of new material to the common store".[10]

ACKNOWLEDGMENT

The work reported in this paper has received no external support. I thought about the material presented while doing compilations at various libraries. I did the major portion of the library work at the Library of the Downstate Medical Center—College of Medicine. During the summer months of 1971–1975 and 1977–1979 I used the library of the Marine Biological Laboratory, Woods Hole, Mass., and during the period February 1976–July 1976, the Library of the University of Sussex, Falmer, England. I thank the librarians of each of these libaries for the kind and necessary assistance. I thank Ms. Gina Boone for typing the manuscript.

REFERENCES AND NOTES

(1) Spencer, H. "Education", cited in "The Home Book of Quotations", Stevenson, B., Ed.; Dodd, Mead and Co.: New York, 1937; p 1764.

(2) Kirschenbaum, D. "A Compilation of Amino Acid Analyses of Proteins. XIII. Residues per molecule-10", *Anal. Biochem.* **1977**, *83*, 521–550 (this is the last published paper in this series).

(3) Kirschenbaum, D. "Molar Absorptivity and $A_{1cm}^{1\%}$ Values for Proteins at Selected Wavelengths of the Ultraviolet and Visible Regions. XVII", *Int. J. Pept. Protein Res.* **1979**, *13*, 479–492.

(4) Kirschenbaum, D. "Atlas of Protein Spectra in the Ultraviolet and Visible Regions"; IFI/Plenum: New York: Vol. 1, 1972; Vol. 2, 1974.

(5) The introduction of miniprint in some journals is an added complication for data compilers to overcome. The print is too small to read rapidly, and occasionally the print is too light. If miniprint is to be used and adopted by additional journals, then data must be flagged.

(6) One use of the analytical literature is the ability to determine the kinds of equipment and the manufacturers of the equipment used in the analytical determination. This information should be of use in determining marketing strategy and buying patterns of analytical scientists.

(7) Wetlaufer, D. "Ultraviolet Spectra of Proteins and Amino Acids", *Adv. Protein Chem.* **1962**, *17*, 304–390.

(8) Perhaps journals should have professional data evaluators. See Compton, W. D. "Our Objective: Systems with Critically Evaluated Data", *Bull. Am. Soc. Info. Sci.* **1975**, *1*, 3. Innis, G. S. "Letters", *Science* **1979**, *204*, 242. Stockmayer, W. H., "Data Evaluation: A Critical Study", *ibid.* **1978**, *201*, 575. Lide, D. R., Jr. "The Standard Reference Data System", *Chem. Eng. Prog.* **1971**, *67*, 77–78.

(9) deMontaigne, M. E. "Essays", Book III, Chapter 15. Cited in "Familiar Quotations (John Bartlett)", 14th ed.; Beck, E. M., Ed.; Little, Brown and Co.: Boston, 1968: p 191. A similar quotation, "I am but a gatherer and disposer of other men's stuff", was used by Henry Wotton, "Elements of Architecture", cited in Barlett's "Familiar Quotations", 14th ed.; Beck, E. M., Ed.; Little, Brown and Co.: Boston, 1968: p 191.

(10) Lord Rayleigh, The Academy V176-77, cited in Lide, D. R., Jr.; Rossmassler, S. A. "Status Report in Critical Compilations of Physical Chemical Data", *Annu. Rev. Phys. Chem.* **1973**, *24*, 135–158.