Pure & Appl. Chem., Vol. 67, No. 6, pp. 1027–1030, 1995.
Printed in Great Britain.

The realities of developing computer readable numeric databases

Stephen R. Heller

Agricultural Research Service, US Department of Agriculture,
Beltsville, MD 20705-2350 USA

Abstract

With more data being made available in electronic form, the issue of the technical, economic, and political realities in developing such databases is presented. This paper emphasizes the technical and economic problems related to the development of scientific numeric databases. Examples from a number of groups in both the scientific community and IUPAC sponsored database are highlighted.

Introduction

Over the past two decades a great deal of scientific information has been made available in computer readable form to the community. The chemical community has been one of the leaders in this field, primarily in the USA and Germany. In the USA, Chemical Abstracts has been the world leader in developing chemical bibliographic database, while ISI has also developed a number of related useful products for chemists. Germany, which was the leading country in Europe for chemistry up to the mid 1940's has recently undergone a renaissance in this area. Today it is one of the two leading countries with centers of chemical information, particular numeric and factual chemical data. As this paper will stress the development of numeric databases, the excellent main bibliographic works of Chemical Abstracts and ISI will not be mentioned further.

Numeric data in chemistry have been compiled for many years, with the Beilstein Handbook being one of the first compilations. Since the early 1970's, there has been a considerable amount of activity throughout the world in the creation and maintenance of numeric databases in chemistry. Most of the initial activities came from the USA, particularly the NIST Office of Standard Reference Data (1), with some efforts being undertaken in Europe and even a few from Japan.

Economic Issues

The title of this paper is "The Realities of Developing Computer Readable Numeric Databases", perhaps should have been the economic future of numeric databases because economics is the main issue. The issue of the economics of numeric data in chemistry is not new (2-4). The related issues of the technical quality and the quantity of data, while both quite important, plays a secondary role compared to the major issue of economics. The reason is that the cost of data quality, which involves mostly highly trained and educated manual labor is increasing much faster than the current usage of this information. When this is coupled with a relative lack of quantity of data, in comparison to bibliographic data, the result is that income derived from numeric databases is quite low (3). As Weiske has pointed out "It is therefore understandable that various national institutions and international institutions have been participating in creating such datafiles"(4) . Thus it seems clear that the costs relative to the income of numeric databases is such that profit making companies have virtually stayed away from this area of databases.

International Scientific Societies

There are two major international scientific organizations which are heavily involved in scientific data. The first is CODATA (5), the Committee on Data for Science and Technology. CODATA was established in 1966 and is concerned with all types of quantitative data resulting from experimental measurements or observations in the physical, geographic, meteorologic, biological, geological, and astronomical sciences, and so on.

The second is IUPAC, the International Union of Pure and Applied Chemistry. Besides being involved in printed data compilations, such as the Solubility Data Series publications, in the mid 1980's IUPAC initiated projects to develop computer readable databases from internal IUPAC projects. In addition IUPAC established a Committee on Chemical Databases (6). To date this group has produced a database on enthalpies of vaporization (7) and a database of stability constants (8) and others are in various stages of discussion and preparation for dissemination to the scientific community. With the size of slightly more than 600 entries for the database of enthalpies of vaporization is not surprising that the sales to date over about 3 years are very small in number. The second database project, a database of stability constants, which contains more than 20,000 entries has produced many more sales in just the first few months than the previous database has produced in years. It still remains a question as to when IUPAC will actually be able to recover its investment in the stability constants database.

Database Size Issues

One of the problems with the economics of numeric data is what may be thought of at first as a contradictory statement. There is both too little and too much data. The problem of the small volume of data makes it difficult to attract users to search or use the database, while the cost of storing and searching the large numeric databases leads to very costly systems and access fees.

For the matter of large volumes of data two examples will are presented, the first of which is in biology. While the nucleotide sequence database may contain over 100 million base pairs, the total human or maize (corn) genome have over 3 billion base pairs each! Add to that the number of bases in a few of the other more important genes (soybean, wheat, yeast, e-coli, mouse, cow, barley, rice, and so on), so one sees that is a lot of data to be stored, and even with the ongoing reduction in the costs of disk storage fees these costs are not trivial. In chemistry the largest collection of organic numeric scientific data is Beilstein Online. However as one looks carefully at the database one soon discovers that, while there are some 5-6 million compounds in the database, few have much data. That is, of the almost 400 data fields, only a fraction contain data. For example, in the Beilstein database of the 5-6 million compounds, there are about 600,000 compounds with boiling points and just over 3,000 with enthalpy of formation values. Thus a user might not be pleased to find so little data for the particular information they want. Furthermore, even if the same user found the one (or a few) data values of interest, it is not likely the user would go back to the database for more of the same information for two reasons. The first is that all the existing numeric data in the database has been found from the first search. The second is that the database is not updated very often (certainly not weekly or daily as bibliographic databases are generally updated). Thus one has a situation where large volumes of high quality data needs to be stored and is likely to be accessed relatively infrequently. And it has taken 150 years of the scientific chemical literature and years of careful evaluation by the chemists at the Beilstein Institute to get even this much data which has been published!

In the area of thermodynamics and material properties data, the subject of a detailed

presentation in this symposium, the same situation exists. The amount of data is so small as to preclude the possibility of massive use of the database. The same can be said for the data contained in over 500 volumes of the inorganic Gmelin Handbook database.

Successful Database Example

One of the few successful numeric databases is the NIH/EPA mass spectral database, which is now being maintained and distributed by the US Government agency NIST (9). This database of a over 60,000 spectra of organic compounds is of a very good quality, but remains small for a number of reasons. One is because so little good published mass spectral data can be extracted from the literature for such a database. The second reason is that to run mass spectra from scratch, the cost to obtain a sample and run it on mass spectrometer exceeds $ 250 per sample. Why then has it been so successful (with revenues of almost $ 1 million per year)? One reason is that the US Environmental Protection Agency (EPA) has demanded that this database be used in all contract and regulatory chemical analysis. Hence due to government regulations, this database is widely used. Similar efforts by NIH, EPA, and NIST in the fields of infrared (IR) and nuclear magnetic resonance (NMR) spectroscopy have met with a much lower level of acceptance and use. I believe the reason for this is the lack of a large database coupled with the absence of a regulatory requirement.

Low Database Usage

Additionally the obvious fact the these databases don't contain as much information as desired, are there other reasons for the low usage? I would offer a few suggestions which address this question. First, much of the numeric data does not appear the scientific literature. Journal publishers are quite cost conscious. Thus they, and the journal editors, want as many papers as possible, using as little space and paper as possible. Authors are interested in publishing to enhance and advance their careers. Who then is there to look after the larger and longer range question of data as a foundation for future scientific work? Even the recent policy changes which allow for additional data to be submitted to journals, with such information in supplementary materials, there has been little overall improvement in the situation. Also one must remember that, at present, virtually no one gets credit for publishing supplementary materials.

Second, journal publishers don't pay for what they publish. For the most part, scientists quite willingly submit their research results for nothing, and the journals pay essentially nothing to scientifically process the papers being published. The cost of the editors, advisors, and reviewers is rather minimal. Physically publishing, marketing, and selling the journal is where the costs are. Once the printed journal is published there is essentially no ongoing cost for maintenance, updating, or corrections.

Third, all publishers sell, at a single price, everything they publish, be they scientific publishers, or publishers of a daily newspaper. This includes materials readers don't want. In any given journal, exactly how many articles do you read, let alone want? The publisher is able to sell pages and pages of articles which the reader will never look at! Thus the reader (or in most cases the library) buys a product with all the accessories, bells and whistles included, and all at a single price (even if this price be a subsidized one as in the case of individuals or, in some cases, non-profit organizations). When you go into a computer readable online (or even PC based) database you are able to quickly (and cheaply) find out if what you want is there, and if so, get it directly and quickly. If it is not there, then you quickly leave the system. That is hardly the sort of economic incentive to convince companies to invest in numeric databases for their future well being and economic survival.

Summary

In summary the outlook for commercially viable numeric databases remains poor and there is little reason to believe it will improve in the near future. Some governments, domestic, and international organizations realize they must subsidize such activities. Thus the economic problems have been counter-balanced by the longer term policies, politics, and foresight of such groups. Overall it would seem that things are in better shape than one would expect at this time. Hopefully over time the recognized value of these activities will swell, usage will increase, and it will become more apparent to the scientific community as to the value of this information. Thus as the subsidies begin to dwindle, the likelihood for these databases becoming economically viable will improve.

References

1. D. R. Lide, "Critical Data for Critical Needs", Science, 212, 1343-1349 (1081).

2. S. R. Heller, "The Economics of Online Data Dissemination", Proceedings of the 7th International CODATA Conference, pages 578-585, Ed. P. S. Glaeser, Pergamon Press (1981).

3. Harry Collier, "Strategies in the Electronic Information Industry - A Guide for the 1990s", by Harry Collier (1991). Published by Infonortics Ltd., 9A High Street, Calne, Wiltshire, SN11 OBS, UK. ISBN#: 1 873699 00 X.

4. C. Weiske, "Chemical Information in a Changing Europe", Kemia-Kemi, 18, 23-25 (1991).

5. For details about CODATA, please contact the CODATA Executive Secretary: Mrs. Phyllis Glaeser, CODATA, 51 Blvd. de Montmorency
75016 Paris, France.

6. For details on the IUPAC Committee on Chemical Databases (CCDB), please contact the CCDB secretary: Dr. Rudolph Potenzone, Jr., CAS - New Product Development, 2540 Olentangy River Road Columbus, OH 43210 USA. Phone: +1-614-447-3600; FAX: +1-614-447-3813; Internet: RXP07@CAS.ORG.

7. ENTVAPOR, a retrieval and computation system is available from the IUPAC publisher, Blackwell Scientific Publications Ltd., PO Box 88, Oxford, UK.

8. The IUPAC Stability Constants Database is available from Academic Software, Sourby Old Fram, Timble, Otley, Yorks, LS21 2PW, UK. Phone: +44-943-880-628

9. NBS Mass Spectral Database, PC Version. Program by Dr. Stephen E. Stein, NIST (formerly the National Bureau of Standards), Office of Standard Reference Data, Building 221, Room A-325, Gaithersburg, MD 20899 USA.