Criteria of Quality Assessment for Scientific Databases[†]

Peter Kuhn,* René Deplanque,[‡] and Ekkehard Fluck

Gmelin-Institut für Anorganische Chemie der Max-Planck-Gesellschaft, Varrentrappstrasse 40-42,
D-60486 Frankfurt am Main, Germany

Received September 14, 1993[⊗]

Quality assessment of scientific databases by well-defined standards is becoming increasingly important for users of scientific information. For a database producer it is then necessary to develop rules by which input to the database is performed and to explicitly state these rules. The conceptual difficulties of this approach are discussed. It is shown that quality standards cannot be defined in a formal manner. It is proposed to define the quality of scientific databases by the criteria of relevance, comprehensiveness, and reliability. These criteria should be applicable both at the user's and at the producer's level and thus may provide an objective basis. An outline is given on how data input to a scientific database is to be carried out along these criteria. The Gmelin Factual Database is taken as an example to illustrate procedures and problems encountered.

1. INTRODUCTION

In the long tradition of science, quality is thought of as being defined in itself. The measure of the quality of a scientific idea, method, or result is its impact to promote scientific progress. This retrospective procedure of quality assessment has efficiently worked, and there is no reason to change it.

Nevertheless, within the recent past the situation has changed to some extent. Stimulated by computer technology, new industries and services have evolved to distribute scientific information using database systems. Producers and customers agree that commercial products or services of this kind of information have to satisfy the same high quality standards as the original publication from where the information is taken. Apparently, however, no common notion exists of how to measure quality.

To justify the tremendous efforts and costs involved in providing this type of scientific information, clearly defined and accepted criteria for assessing the quality of scientific databases are needed. A purely heuristic approach is being used here with the following objectives:

(1) to define criteria for quality assessment of scientific databases applicable to both customer and producer;

(2) to let the customer realize the complexity of data compilation and management to reach high quality standards and, moreover, to boost the awareness of the customer concerning the costs in money terms of high-quality scientific information.

The Gmelin Factual Database, which the Gmelin Institute offers besides the *Gmelin Handbook of Inorganic and Organometallic Chemistry*, is used here to illustrate the problems associated with quality control during data input.

2. WHAT IS QUALITY OF SCIENTIFIC DATABASES?

The goal is to find criteria for the quality of databases which a user can apply to compare the output of different databases and which a producer has to apply to offer a database conforming to the user's needs. Databases referred to in this paper consist of numeric and factual data that are collected from the primary literature. This requires a functional approach for defining the purpose scientific databases ought to serve. However, there presently appears to be no agreement within the scientific community which requirements should be met by scientific databases. Even the role of scientific data has changed considerably due to their ever-growing amount and concurrently the ever-growing facilities to store and retrieve them.[1]

Quality features of numeric/factual databases needed to serve the user have recently been outlined by Barrett in a paper presented at an ACS/CODATA symposium.[2] A guide for formatting and use of material property and chemical property data and database quality indicators have been proposed as a new standard by the ASTM.[3,4] Also recently, Zass and co-workers,[5] on behalf of the scientific advisory board of the German *Fachinformationszentrum Chemie* (FIZ), have proposed a classification scheme for database properties that are important for a user. These papers show that database quality may have quite different aspects for a user.

The present paper focuses on a few aspects, namely, the quality of information provided by the processes of data collection from the literature and input to a database. The main goal is to find quality attributes that are controllable both at the input level and at the user level, with the intent to arrive at a consensus between database users and producers about quality attributes and to make quality criteria objective.

The approach therefore is purely heuristic; it is proposed to assess scientific databases along the following criteria: relevance, comprehensiveness, and reliability. The first criterion, relevance, is a measure of *what* data relate to the user's question and the second, comprehensiveness, of *how many* he will find. The third criterion, reliability, is an indicator of *how* feasibly or reproducibly the data can be accessed.

In the following section an attempt is made to formulate the criteria above in definite terms. Several examples are to demonstrate how the criteria are checked at the stage of data input to the Gmelin database and what problems have been encountered at the quality control stage. Data at the input level that are not checked for relevance, comprehensiveness, and reliability are meaningless.

3. HOW TO FULFILL THE CRITERIA

3.1. Relevance. A chemical database like that of Gmelin[6] contains chemical and physical data of compounds (Gmelin: inorganic and organometallic) in a substance-oriented data structure. The data are taken from the current and past

[†] We kindly acknowledge financial support by the German Ministry of Research and Technology (BMFT).

[‡] Present address: Fachinformationszentrum Chemie, Franklinstrasse 11, 10587 Berlin, Germany.

[⊗] Abstract published in *Advance ACS Abstracts*, March 15, 1994.

0095-2338/94/1634-0517$04.50/0

scientific literature. Pictorially speaking, the data structure of the database is spread over the scientific literature to see what falls into the grid. The screening is performed by asking questions such as "how large is ..?", "how is compound XY synthesized?", or "to what degree is compound XY soluble in ...?".

The problem at this stage is that the genuine problems of chemistry or physics reside at the level of concepts, that is, at a level of abstraction beyond a purely material- or data-oriented notion. From this follows that most studies communicated in the scientific literature cannot be addressed by simple questions regarding, for example, material properties, nor can results be expressed in the form of single data points.

As a consequence, data related to individual compounds, even though they may be essential for the phenomena or processes and the laws that govern them, are not the object of primary interest. For example, spectra may be recorded to determine chemical structures, reactions may be carried out to elucidate reaction pathways or kinetics, and electrical properties may be studied to get insight into the mechanism of charge carrier transport.

When data of individual compounds from the literature are to be entered into a database, they get segregated from the problems which initiated their determination. Therefore, the question of how relevant they are must be raised each time. There is no way to define relevance in any formal manner. It always must be discerned from the context within which data are used in a publication and the motivation of the author to report them.

An example may help to clarify this: Let us take six different papers in which ceramic superconductors of the general formula $YBa_2Cu_3O_{7-x}$ are described. All papers contain data on the critical temperature T_c which, depending on oxygen stoichiometry, may attain values between 38 and 92 K. In one paper the dependence of T_c on oxygen stoichiometry is the main issue and T_c without any doubt is relevant. In the remaining five papers the determination of T_c is not of primary interest. In two of them, T_c was redetermined, since it is a prerequisite to understand the subjects of main interest. In three papers, T_c from different samples and their variation with oxygen stoichiometry to all intent and purposes are presented only to indicate homogeneity and composition of the sample. The two former T_c values should reasonably be taken to be relevant, the latter three to be nonrelevant.

This example shows that a scientist doing quality control on data acquisition, to evaluate data as correct and meaningful, must have the knowledge and insight at a level commensurate with those of the bench scientist.

3.2. Comprehensiveness. Comprehensiveness is a measure of the degree of completeness of a database. To be useful for scientists, this criterion must be satisfied so that the information adequately represents the state of science, rather than be archivally self-contained. Therefore, the term comprehensiveness should be preferred instead of completeness.

Representing the state of science in databases is only possible to the extent by which science can be represented by data. Therefore, the database producer has to ensure that all relevant data are contained in the database. This requires that the database has to be up to date. Comprehensiveness is thus tightly associated with relevance. The user of a database will judge comprehensiveness by examining its literature coverage, i.e., how many primary sources can be accessed via the database. Thus, the producer of a database should open up as many as possible sources of information, i.e., scientific papers covering topics that pertain to the data structure.

3.3. Reliability. When analyzing data retrieved from a database, users may find that certain data are missing, although they know or may reasonably assume that the data are contained in the database. They also may find data that have been indexed or classified along guidelines they find difficult to conceive. Users may further experience that several modifications of a query, which are judged not to alter the meaning of the question, may be necessary to obtain a sufficient answer set.

The outcome is that under conditions which users cannot foresee, data happen not to satisfy their expectations, which in the users' judgment renders the database from which they were retrieved unreliable.

Apart from communication language problems which cannot be solved at the input level, some of the shortcomings may result from a lack of relevance or comprehensiveness. This is hardly recognizable by the user. The question therefore is, what else besides relevance and comprehensiveness must be controlled at the stage of data input to warrant reliability?

At this point it is important to note that data before being transferred from the scientific literature to a database very frequently need to be correctly interpreted, since the meaning of the data is not always evident to a degree of exactness required by the rigid data structure of the database. The extent to which the interpretation is correct is, however, limited by the capability of the individual involved in data processing, and a possible source of error can be due to insufficient knowledge or an inherent arbitrariness.

Semantic ambiguities always present in scientific terminology are one source of error. In the example above we used the term "critical temperature" in connection with superconductivity. It may, however, be associated with critical phenomena of whatever kind, be it conductivity, magnetization, or liquefaction of a gas.

For the Gmelin database, most problems encountered with a reliable data input arise from the structure of organometallic compounds and coordination compounds within the scope of the database. An atoms-and-bonds description of structures is appropriate for them. This information is transformed in the database to connection tables.[7] However, structures in the chemical literature quite frequently are not described in sufficient detail.

The problems can be demonstrated by three selected examples:

1. By-name-only designations of organometallic or coordination compounds with acronyms for organic ligands and without structure drawings are commonly used in the chemical literature. Given, for example, the name or half-formula "*Pd(II)(phbpy)Cl*",[8] it is not difficult to deduce the structure of *6-phenyl-2,2′-bipyridine* (*phbpy*) from textbook knowledge. However, an error can occur unintentionally if one is not aware of the coordination chemistry of this particular ligand, which is not bidentate (*2,2′-bipyridine*), but has a third coordination site at the 2″-carbon atom of the phenyl ring.

2. A great variety of cluster compounds are known in organometallic chemistry which contain a cage of several metal atoms enclosing an interstitial atom. This may be a metal atom but can also be a carbon or hydrogen atom. A brief look into the chemical literature shows that the unique bonding situation of the interstitial atom cannot be described in common terms. Sometimes, structures are presented with bonds between the interstitial atom and the metal cage atoms. Where the bonds have not been drawn, the authors might prefer not to specify the bonding situation of the interstitial any further, or the bonds may have been omitted in order not to clutter

the drawing. However, representations for structures in chemical databases have to be chosen independent of perceptional ambiguities and of the accidental nature of two-dimensional structure drawings used in the literature.

3. It is a matter of perception how significant intramolecular hydrogen bonds, for instance, in ligands such as *dimethylglyoxime*, are for describing the structure. As a rule, intramolecular hydrogen bonds are not taken into account for representing the structure in the Gmelin database. It is not possible, however, to generalize this rule. It would be highly unreasonable to apply it to B–H–B bonds in boranes. Furthermore, hydrogen bonds may become significant to the extent to which they determine substance properties relevant to the Gmelin data structure. This may hold for so-called agostic interactions between a metal and a hydrogen atom in organometallic compounds, for which increasing evidence is found in the literature.

4. CONCLUSIONS

From a database producer's view the following conclusions can be drawn:

1. There is no way to apply criteria for quality assessment in a strict and formal manner. It is of course possible to include everything from the literature in the database dependent on the data structure. In this case there would be no clear relevance, and the reliability and the comprehensiveness of the data would not mirror the state of science adequately. It follows that none of the above criteria should be taken alone but always in context and together. The only way to judge relevance, comprehensiveness, and reliability in a meaningful way is to look at "what comes in" at the input level and "what comes out" at the user's level.

2. Data collection has to follow very strict and formal rules, but unless the data are critically evaluated and checked by experts before being entered into the database, the above criteria cannot be fulfilled. This step of quality control must be carried out along a truly scientific concept of "looking beyond the facts" or "asking what the real meaning is", to provide a sensible access to the wealth of information distributed in the scientific literature.

3. The high quality standards of a database necessary to render it valuable for scientists can hardly be implemented by only a single step of quality control at the data input. However, all subsequent steps of quality improvement require facilities and procedures by which data already processed and stored in the database can be modified and edited.

In pursuing the last conclusion, Gmelin has developed software tools that are increasingly used together with the database as a source for the production of Gmelin *Handbook* volumes. The volume editors compile and critically evaluate selected data related to certain classes of compounds. The software tools allow any user-defined view to the data, that is, on selected subjects related to a certain class of compounds from a selected set of papers. This is a benchmark of the Gmelin Integrated Information System.[9] It offers the opportunity to apply the high scientific standards of the Gmelin *Handbook* to the Gmelin database.

REFERENCES AND NOTES

(1) Dubois, J. E. Data are not as simple as they were. *CODATA Bull.* **1990**, *22* (4), 51–61.

(2) Barrett, A. J. Socioeconomic aspects of materials data: Serving the user. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 22–26.

(3) ASTM. *Standard Guide for Formatting and Use of Material and Chemical Property Data and Database Quality Indicators*; Philadelphia, 1992, Standard E1484-92.

(4) Reference 3 came to our knowledge by the paper of Barrett (ref 2). The original document of ASTM has not yet become available to us.

(5) Zass, E.; Donner, W.; Leuther, P.; Lockhoff, A.; Römelt, J.; Spanagel, H. D. *Mitteilungsbl. Ges. Dtsch. Chem., Fachgruppe Chem.-Inf. Comput.* **1993**, *26*, 31–40.

(6) Deplanque, R.; Boehmer, U.; Kunz, M.; Fluck, E. *Mitteilungsbl. Ges. Dtsch. Chem., Fachgruppe Chem.-Inf. Comput.* **1992**, *22*, 11–36.

(7) Roth, B.; Boehmer, H.-U.; Deplanque, R. Registration of substances in the Gmelin Factual Database. *Anal. Chim. Acta* **1992**, *265*, 301–304.

(8) Karlen, T.; Ludi, A.; Güdel, H. U.; Riesen, H. One-dimensional migration of ^{3}MLCT excitation energy in Pd^{II}(phbpy)Cl (phbpy = 6-phenyl-2,2′-bipyridine). *Inorg. Chem.* **1991**, *30*, 2249–2251.

(9) Nebel, A.; Toelle, U.; Maass, R.; Olbrich, G.; Deplanque, R.; Lister, P. The Integrated Gmelin Information System. *Anal. Chim. Acta* **1992**, *265*, 305–312.