Computer Physics Communications 33 (1984) 79–84
North-Holland, Amsterdam

DATA BASE FOR INORGANIC CRYSTAL STRUCTURES

G. BERGERHOFF

Anorganisch-Chemisches Institut der Universität Bonn, Gerhard-Domagk-Strasse 1, 53 Bonn, Fed. Rep. Germany

Crystal structure data bases form a typical example for the advantages of numerical data bases. They store the long row of primary results in such a manner that they can be checked and evaluated by computer programs. Thus giving results of structural work in an appropriate presentation to all who need them.

1. Introduction

The increase of information in all branches of science will soon overwhelm us and will lead to a completely uneconomic use of our resources. Thus, what appears to be a progress in science will in reality be a regress. We shall waste our time, our man-power and our means if we neglect things already known to other people and if we rediscover these things once again. In fact today we develop such well and fast-working instruments that some people think it would be easier to repeat the measurements than to look for them in the literature. But it seems to me that this point of view underestimates the diligent and hard work that is done daily in scientific laboratories. Instead we should use the results found by our predecessors and we should pass on our own results to our successors in such a way that they can find them and they can work with them.

Looking around for the reasons of this avalanche of information we shall find that one reason is the computer. It accelerates measurements and evaluations and I am sure it also seduces us to collect more data than we need. But now we must learn to use the computer to solve our information problem as well.

Indeed the computer can store any amount of data, it works fast and reliably, but it only does what we say that it should. Any successful use of the computer for information storage and retrieval depends on the way in which the data are handled. The first step in using computers for automatic storage and retrieval was an imitation and – I concede – an improvement of the classical registry. The technical processes of sorting, checking and printing of keywords and citations became faster and more reliable. But more important than this, computers were able to sort on logical combinations of keywords. Thus, they were able to select relevant references with high speed by direct access to data bases.

Nevertheless users are not always satisfied because they get references they do not want and they do not get references that they do want. Today information retrieval is no longer a technical problem; but it is the logical and semantic problem of preparing a scientific paper for retrieval in such a way that the user finds what he has in mind.

There are extensive and wide ranging discussions among information people about classification problems, thesauri, freetext search, defined vocabulary, keywords, role and link indicators, etc. All this shows that the philosopher's stone has still not been found.

This also can be demonstrated by searching for the same facts by different methods. There is a recent investigation by Baerns et al. from the University of Bochum. They compared searches in Chemical Abstracts by hand and by automatic retrieval via SDC (System Development Corpora-

0010-4655/84/$03.00

tion, Santa Monica, California). While there were some differences in the different branches of chemistry they recovered only about 70% of the stored information, measured with the other method, respectively. In course of setting up our Inorganic Crystal Structure Database (ICSD) we have searched for relevant papers by three methods: (1) direct scanning of primary journals, (2) via Chemical Abstracts, (3) via Bulletin Signaletique. The discrepancy between the sets of relevant papers found by these methods is frightening. Further, 80% of a retrieval using a profile search for "crystal structure of inorganic compounds" in Chemical Abstracts regularly consists of 80% non-relevant papers.

This special difficulty reflects the different meanings of the term "crystal structure" or even "structure of crystals". For chemical compounds it would be worthwhile to differentiate the terms:

Crystal structure (all atomic coordinates known),

Crystal structure type (similarity to known crystal structure type verified),

Crystal morphology (form, habit, etc.),

Crystal data (unit cell dimensions, space group, Z only),

Powder diagram (list of d-values, I(observed) only),

Microstructure (description of domains, dislocations, texture, stacking faults, shear structures, block structures, etc.).

Structure (connectivity of atoms only).

Such a more precise use of words is the precondition for any improvement of the information process.

It is interesting to note that users are looking for other ways to get their information. A traditional way is to identify a name of an author with the subject of his research. Thus, a very diffuse field has been condensed to one concept: the author's name which is easy to locate. For several years the Citation Index has been refining this method into a very efficient tool.

2. Numerical data bases

But even while we appreciate this tool, the disadvantages are obvious and our goal should be to classify papers according to clear-cut concepts as far as possible.

Of course we cannot solve the logical and semantic difficulties in one step. But we can separate the contents of papers into hard data obeying strict rules and into soft data reflecting the comments, the explanations and ideas of the author. We then select all the hard data from one paper and group them together into as many data sets as there are groups of logically connected concepts. In chemistry the central point of such data sets will be the substances whose preparation, properties and the physical conditions under which they exist have been described. In cases of higher complexity it may be appropriate to separate the data for low and high temperature, etc. into different sets. All these datasets now form the body of a new kind of data bases: numerical data bases, data bases for facts and not just for references. They can be handled with common data base management systems at least in principle. In German we say "we can kill several flies with one blow":

1) We have a precise data definition and the correct linkage between different data types. This makes it easier to find all relevant papers. We reduce the irrelevant and the increase the relevant recall.

2) The scientist has immediate access to the data he wants and – through the citation – a more reliable approach to the comments and ideas of the author.

3) The data entered in the data base can be checked for consistency giving some guarantee of their correctness.

4) The results of complex investigations, themselves often complex and unintelligible to the non-specialist, are easily interpreted by the computer that reads the data base.

5) A single result gains in value when compared with other related results, a feature easily incorporated into retrieval programs.

3. The ICSD

Crystal structure analysis is a typical example to demonstrate how scientific literature could be prepared in the future to provide for efficient and economic access. It is a field whose bulky results not only lend themselves to storage in a numerical data base but it also needs interpretation by calculations with the computer. Without this nobody can recognize from such tables of atomic coordinates what the influence of the temperature on the structure is.

We installed such a data base for inorganic crystal structures to complement the well known Cambridge data base for organic structures started by Dr. Kennard and her coworkers. To avoid duplication between the two data bases we have defined an organic compound as one that contains at least one C–C- or C–H-bond and an inorganic compound as one that contains neither. On the other hand, we include metal carbides but not metals and alloys which are included in the metals data file, produced by Dr. Calvert at Ottawa.

Our data base contains all 29000 structures from the relevant sections of "Strukturbericht" and "Structure Reports" and in addition several structures from Chemical Abstracts, Bulletin Signaletique, Crystal Data and Landolt–Börnstein. All structures can be retrieved for bibliographic data and chemical elements, but at the moment about 50% only have the whole data set necessary for a full description of a crystal structure. This means:

Unit cell dimensions – space group – occupation of positions by atomic sorts and degree – atomic coordinates – thermal parameters – parameters of state and conditions of measurement. Of course in addition we store different names and bibliographic data.

It is our aim to complete the data base. But already it is publicly available via the German "Fachinformationszentrum Energie, Physik, Mathematik" at Karlsruhe (FIZ).

The FIZ and the German Ministry for Research and Technology as well as some other institutions have supported our work. Many colleagues from several parts of Germany and other parts of the world have also contributed.

Let us now look at the ICSD and how we try to fulfil the following five conditions that define an ideal numerical database:

1) it should be correct,
2) it should be complete,
3) it should be up-to-date,
4) it should be user-friendly and
5) it should be versatile.

To be correct: Data are extracted from the primary literature and keyboarded (using a terminal, punch cards or a typewriter with an optical character recognition (OCR) font) into a well-defined machine readable "input" file which is then accumulated at Bonn in an "input" archive file. To ensure that the information in the database is reliable it is subjected to objective and systematic checks by computer programs. The first stage is to check the syntax of the data. Although the input is in free format, the various data must follow one another in a defined sequence and therefore the fields must be of the correct type: integer, real number of alphanumeric. Once this check is passed the second stage is to check the values of the different fields (fig. 1). Here we can distinguish four types of data:

1) Data which must be internally consistent (e.g. lattice constants, formula and density, atomic coordinates).
2) Data which are in principle independent but whose values typically lie within a restricted range (e.g. temperature factors).
3) Data which must be exactly defined and correctly used within the data base in order that they may be used for retrieval (e.g. compound name, minerals, author name).
4) Optional data (e.g. the conditions of measurement, the origin of minerals) whose existence in the data base depends on the thoroughness of the abstractor. These data are nearly incapable of being checked.

The types of errors detected during these tests are of three kinds. The simplest to locate and correct are errors resulting from carelessness in abstracting. More difficult are those resulting from ambiguities in the authors intention, resulting, e.g., from undefined or inconsistent space group settings, the use of undefined symbols, missing data

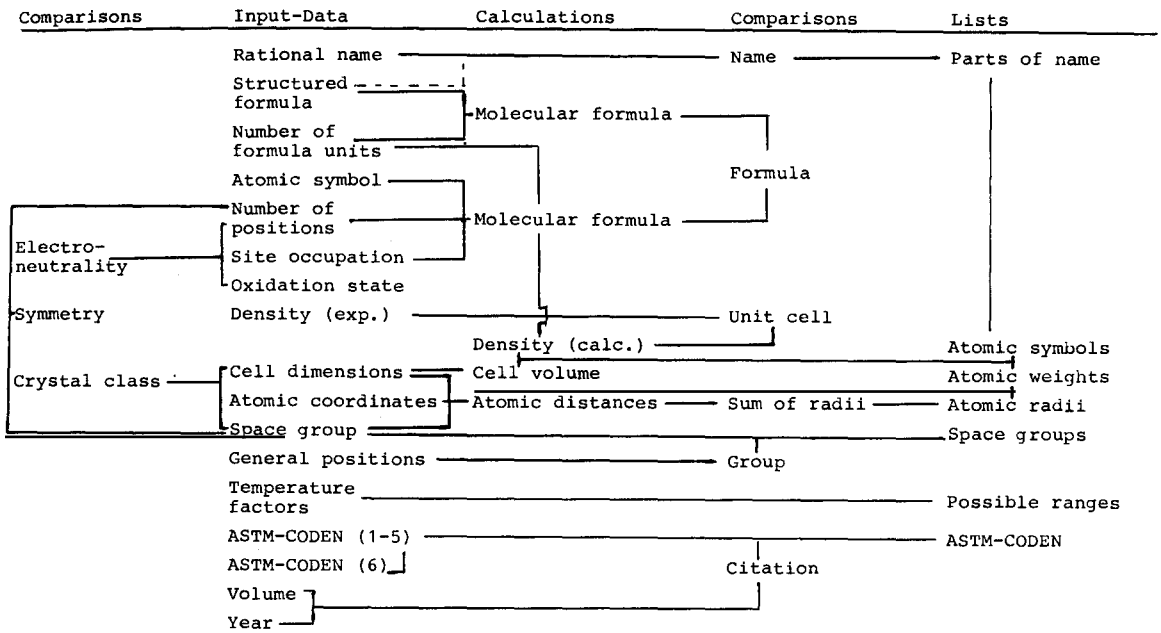

Fig. 1. Check of input-data for the inorganic crystal structure data base

and uncertain chemical composition. The third type are misprints appearing in the original publication. When possible, ambiguities and errors are corrected by correspondence with the author.

The computer check results in a variety of possible error messages and a corresponding code that is inserted into the data base itself. The checked data base entries are then compressed and stored in a memory-efficient Z-file (Z = Zugriff = access). During this compression all the blanks and leading zeros are removed and the address of each datum is defined by a directory at the beginning of the entry. A typical entry in this format occupies just over 1 kbyte.

To be complete: To ensure that the data base is complete is a difficult task. Papers with crystal structures are found in many journals. In our experience: 22% are found in 1 journal; a further 51% are found in 17 journals; the next 20% are found in 58 journals, and the final 7% are found in 254 journals. It is not possible to scan regularly all these journals and we must rely on abstract journals. Once again we encounter the problems of effective recall in bibliographic data bases. I refer once more to a foregoing paragraph showing different sets of relevant papers found by different types of literature search.

It shows how difficult it is to be complete. On the other hand, it seems to me, completeness is one of the most important goals for a data base. Not only from the general point of view mentioned in the beginning, but also from another point of view. There is still a competition between the classical printed reference journals and books and modern computer retrieval. As long as computer retrieval cannot convince all users in a laboratory, both tools must be present. This is economically unjustifiable and the modern method will have much more difficulty in succeeding in spite of its obvious advantages.

To be up-to-date: Also in science information must be up-to-date to avoid inefficient and duplicated work. Data bases can in principle fulfil such a goal. Any information entered in the data base can immediately be retrieved whether it has recently been added or it has been present for a long time.

Compared with the potential of the data base for being up-to-date, the time taken for results to pass from the author through the publisher and the reference journal to the data base producer is too long. Moreover, several data can be lost along the way. There is also a trend in scientific journals to omit important details of structural investigations when they are not directly related to the main part of the paper (e.g. the primary results of a structural investigation like the atomic coordinates).

We have to aspire to changes in operation of scientific publications. There should be much better cooperation between authors, publishers, reference journals and data base producers. Authors could transfer bulky results on magnetic tape directly to the data bases; journals could restrict the papers to soft data, relieve their expensive printed pages of large tables of data and make reference to the data bases; reference journals could sharpen their keyword definitions, hunt out obscure publications and make them available to the data base producers.

I am glad to be able to report that several journals have already started with such cooperation. Acta Crystallographica, Chemical Communications, Angewandte Chemie and some other journals give crystallographic data to the depositories collaborating directly with the relevant data centers in Cambridge, Ottawa, Brookhaven and Bonn.

To be user-friendly: High speed is one of the advantages of the computer but it depends on an appropriate program and an appropriate machine. In Germany we have special information centers, whose function is to make data bases available on-line to all users. It is easier to keep data bases up-to-date, to maintain computers and programs when they are all kept at one place. Any scientist is well advised to look for a connection to such a center through data networks like EURONET, TYMNET or others. Two requirements should be met: Firstly the costs should be payable and directly correlated to the information retrieved; we should pay for information units and not for the connection time. Secondly a uniform set of commands should be used for as many data bases as possible in order to make the use of data bases as easy as the use of catalogues in libraries.

On the other hand, even in the future not everyone will have access to data bases. We should not exclude them from the progress in science. In spite of the loss of many advantages of data bases we should publish the contents in form of books. There is no technical problem and no high costs. The data are already in the correct format and we only need to think about an appropriate order. Books can only present a one-dimensional sequence of facts but a registry with different entries may help. We plan such an output from our ICSD.

Further, crystal structure analysis is typical of fields where the data themselves are the subject of intensive research. For example the concept of isotypism, the theory of bond order or a set of ionic radii, all could be confirmed or evaluated from the stored data. Date centers cannot make programs available for such special tasks. Interested users should have the possibility of transferring the data base or parts of it to their own computer and to access these with their own programs. Here microcomputers offer interesting opportunities. To be user-friendly is a fundamental requirement of data bases because data, once carefully collected, well-defined, formatted and correctly checked can be used for many purposes. One must avoid preventing such access by administrative provisions.

To be versatile: To meet this requirement for the ICSD we have arranged two modes of access.

For general use we have developed a retrieval system specially designed to meet the demands of crystallographers. In preparing this system, inverted files are generated indicating the addresses where data are to be found. At present we have inverted files for:

authors, journals, year of publication;
chemical elements and their oxidation states, element groups, number of different elements, molecular formula (also in ranges);
mineral names, standardized remarks about the condition of measurement, reliability index;
space group symbols;
unit cell volume and crystal class (for identification);
interatomic distances.

Firstly you can explore the inverted files. In the

case of author or mineral names you may wish to know how the names are written. Secondly you can combine the inverted files by the logical operators "and, or, not" and in addition you may look for special strings of characters appearing in names, etc. By this means you can select for example nitrato-complexes of lanthanides and alkali metals using the command:

find LAN. and. N. and. O.and. ALK. and. ELC = 4
with nitrato in = name

The display will show you the number of true references. You can then ask for details. This may be the formula only for first inspection or for the whole data set. In general you will not be interested in the table of atomic coordinates but you may ask the program to calculate atomic distances or bond angles. On our university computer we can also display a stereographic picture of the structure or a calculated powder diagram.

This last item indicates the next stage of development. X-ray powder diagrams are easy to measure from unknown substances and the lines can be matched with the patterns caluclated from the data base. In many cases this will be the most convenient way to a complete analysis. If you have a single crystal, automatic diffractometers allow you to find the unit cell in a very short time. Once more you can match it with the values from the data base and determine what the compound is.

For special purposes we make available a set of programs to provide easy access to each single datum of any data set and then to combine them as the user requires. For example the user can look for crystal structures being rhombohedral with all angles 90°. Other users will be interested in finding crystal structures with copper in elongated octahedral coordination or all the distorted tetrahedra in silicates.

Such examples show the large range of possibilities data bases offer to the user. In principle the technical problems have been solved, logical problems have been recognized and represent an interesting field of research. But many data are waiting in the literature to be collected, to be updated and to be checked. We need as much information as possible to overcome the problems of our world. Our problem is not to have too much information, it is to have the right information at the right time and at the right place. We now have a good instrument to deal with this problem, we should now work up what we know in order to put this instrument into such a position that it serves us to the optimum extent possible.

Many of the problems that still remain to be solved are not scientific, but are organisational, political and psychological problems. The international character of science could help to link together all the producers and users of information who are in fact the same people. Then we would have no alternative but to live in peace because we need the information.