# The Inorganic Crystal Structure Data Base

G. BERGERHOFF,* R. HUNDT, and R. SIEVERS
Anorganisch-Chemisches Institut der Universität Bonn, Bonn, Federal Republic of Germany

I. D. BROWN
Institute for Materials Research, McMaster University, Hamilton, Ontario, Canada L8S 4M1

Received December 23, 1982

An inorganic crystal structure data base is described which will, when completed in the next year, contain details of all the 23 000 published structures of inorganic crystals. This paper describes the structure of the data base, the procedures used to check the data as they are entered, and the program used to access them. Plans for the future development of the data base system include defining search keys on the basis of bonding topologies and crystal structure types as well as plans for providing an integrated crystal structure retrieval system.

## INTRODUCTION

The growing number of publications in all branches of science has made it increasingly difficult to obtain an overview of current research in any particular area. This problem has recently been addressed by bibliographic data centers which have turned to computers to create large data bases which can be searched in a systematic fashion. In such searches there are many difficulties in matching the requirements of the scientist to the stored information in order to ensure an effective retrieval. These difficulties lie on the one hand in the need to represent the complex ideas of the original papers by concise key words and on the other in the development of a corresponding representation to express the needs of the user. Unfortunately, a poorly formulated search leads to poor recall, to many irrelevant references, and hence to frustration and disillusion on the part of the user.

Chemical publications have an advantage in that the chemical formula provides an objective key on which to search. In crystallography it is even easier to provide objective search keys. A crystal structure determination provides a precise three-dimensional description of the structure which can be searched not only for chemical constitution but also for connectivity and molecular geometry. Furthermore, a data base of crystal structures stores the complete results of an investigation, not merely a reference to where the results are to be found.

The results of crystal structure determinations lend themselves particularly well to storage in a numeric data base for a variety of reasons: (i) They are usually recorded in terms of a standard set of numbers (lattice parameters, space group, atomic coordinates) so all the entries in the data base can have the same data structure. (ii) They typically consist of tables which contain between 100 and 1000 numbers that are not themselves of immediate chemical significance. They must be converted to bond distances and angles or molecular diagrams before their significance can be appreciated. A user's program can perform this conversion directly from the data base, often providing information not given explicitly in the original publication. (iii) Because the atomic coordinates are not of direct chemical significance, some journals have refused to publish them, thereby depriving the scientific community of the primary results of a relatively expensive investigation. A crystallographic data base is an obvious place in which to archive this data. As an archive the data base is more accessible than the tradiational data depositories and more useful than printed journals.

The advantages of a computer crystal structure data base were first exploited by Kennard and her colleagues at the Cambridge Crystallographic Data Center in 1965.[1] Their data base is now available in many parts of the world and contains details of the structures of some 30 000 compounds that contain at least one C–C or C–H bond. They deliberately restricted their coverage to organic compounds because the search strategies on connectivity that were built into their data base did not lend themselves to use with inorganic compounds and metals.

Since the pioneering work of the Cambridge Data Center other data bases have been set up to cover the remaining crystal structures. These are the Protein Data Bank (Brookhaven),[2] the Metals Data File (Ottawa),[3] and the Inorganic Crystal Structure Data Base (ICSD) which is the subject of this paper. The data centers that produce these four data bases collaborate closely to ensure complete coverage of all crystal structures with minimum duplication.

## ORGANIZATION OF THE INORGANIC CRYSTAL STRUCTURE DATA BASE (ICSD)

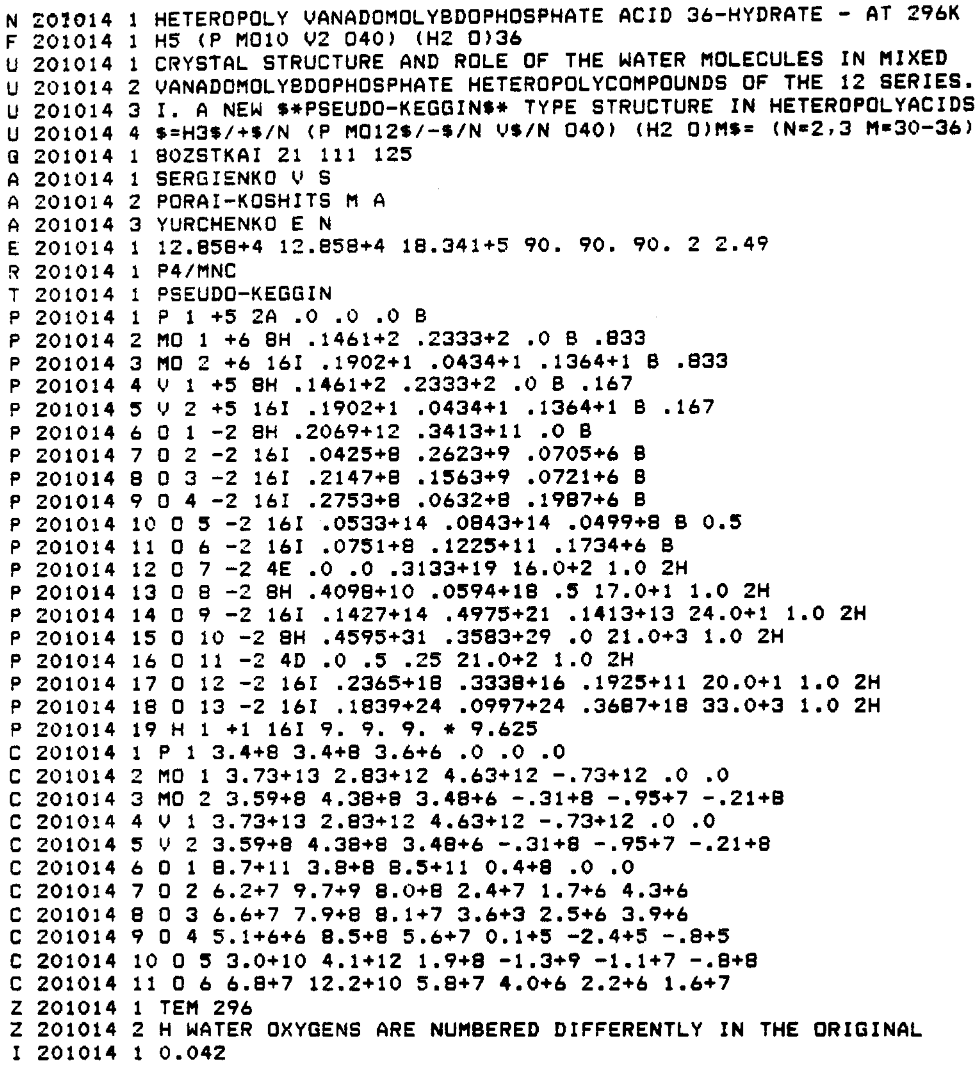
The Inorganic Crystal Structure Data Base is published by the Fachinformationszentrum Energie Physik Mathematik (FIZ) in Karlsruhe, FRG. The major part of the ICSD is produced under the direction of Dr. G. Bergerhoff at the University of Bonn where the data base is assembled and checked. A smaller contribution is prepared under the direction of Dr. I. D. Brown at McMaster University in Canada. In addition, contributions are received from correspondents in Delft (Holland), Parma (Italy), Göttingen, Clausthal, and Erlangen (FRG). Deposited atomic coordinates are routinely supplied by FIZ, The Canada Institute for Scientific and Technical Information (CISTI), the Royal Society of Chemistry (UK), and the International Union of Crystallography.

## STRUCTURE OF THE DATA BASE

The data structure of ICSD is modeled on that of the successful Cambridge Organic Crystal Structure Data Base. When complete it will contain one entry (i.e., one data set) for every complete structure determination of an inorganic compound reported in the scientific literature. Each entry contains all the data necessary for a complete description of the crystal structure together with the bibliographic and other data necessary to characterize the compound and its structure determination. Entries, each of which is assigned a six-digit collection (or entry) number, are composed of a number of logical records (each normally of less than 80 characters) whose data structures are defined by the first character of the record. The different types of records are listed in Table I and a sample entry is given in Figure 1.

Data is 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-de-

0095-2338/83/1623-0066$01.50/0



```
N 201014 1 HETEROPOLY VANADOMOLYBDOPHOSPHATE ACID 36-HYDRATE - AT 296K
F 201014 1 H5 (P MO10 V2 O40) (H2 O)36
U 201014 1 CRYSTAL STRUCTURE AND ROLE OF THE WATER MOLECULES IN MIXED
U 201014 2 VANADOMOLYBDOPHOSPHATE HETEROPOLYCOMPOUNDS OF THE 12 SERIES.
U 201014 3 I. A NEW $*PSEUDO-KEGGIN$* TYPE STRUCTURE IN HETEROPOLYACIDS
U 201014 4 $=H3$/+$/N (P MO12$/-$/N V$/N O40) (H2 O)M$= (N=2,3 M=30-36)
Q 201014 1 BOZSTKAI 21 111 125
A 201014 1 SERGIENKO V S
A 201014 2 PORAI-KOSHITS M A
A 201014 3 YURCHENKO E N
E 201014 1 12.858+4 12.858+4 18.341+5 90. 90. 90. 2 2.49
R 201014 1 P4/MNC
T 201014 1 PSEUDO-KEGGIN
P 201014 1 P 1 +5 2A .0 .0 .0 B
P 201014 2 MO 1 +6 8H .1461+2 .2333+2 .0 B .833
P 201014 3 MO 2 +6 16I .1902+1 .0434+1 .1364+1 B .833
P 201014 4 V 1 +5 8H .1461+2 .2333+2 .0 B .167
P 201014 5 V 2 +5 16I .1902+1 .0434+1 .1364+1 B .167
P 201014 6 O 1 -2 8H .2069+12 .3413+11 .0 B
P 201014 7 O 2 -2 16I .0425+8 .2623+9 .0705+6 B
P 201014 8 O 3 -2 16I .2147+8 .1563+9 .0721+6 B
P 201014 9 O 4 -2 16I .2753+8 .0632+8 .1987+6 B
P 201014 10 O 5 -2 16I .0533+14 .0843+14 .0499+8 B 0.5
P 201014 11 O 6 -2 16I .0751+8 .1225+11 .1734+6 B
P 201014 12 O 7 -2 4E .0 .0 .3133+19 16.0+2 1.0 2H
P 201014 13 O 8 -2 8H .4098+10 .0594+18 .5 17.0+1 1.0 2H
P 201014 14 O 9 -2 16I .1427+14 .4975+21 .1413+13 24.0+1 1.0 2H
P 201014 15 O 10 -2 8H .4595+31 .3583+29 .0 21.0+3 1.0 2H
P 201014 16 O 11 -2 4D .0 .5 .25 21.0+2 1.0 2H
P 201014 17 O 12 -2 16I .2365+18 .3338+16 .1925+11 20.0+1 1.0 2H
P 201014 18 O 13 -2 16I .1839+24 .0997+24 .3687+18 33.0+3 1.0 2H
P 201014 19 H 1 +1 16I 9. 9. 9. * 9.625
C 201014 1 P 1 3.4+8 3.4+8 3.6+6 .0 .0 .0
C 201014 2 MO 1 3.73+13 2.83+12 4.63+12 -.73+12 .0 .0
C 201014 3 MO 2 3.59+8 4.38+8 3.48+6 -.31+8 -.95+7 -.21+8
C 201014 4 V 1 3.73+13 2.83+12 4.63+12 -.73+12 .0 .0
C 201014 5 V 2 3.59+8 4.38+8 3.48+6 -.31+8 -.95+7 -.21+8
C 201014 6 O 1 8.7+11 3.8+8 8.5+11 0.4+8 .0 .0
C 201014 7 O 2 6.2+7 9.7+9 8.0+8 2.4+7 1.7+6 4.3+6
C 201014 8 O 3 6.6+7 7.9+8 8.1+7 3.6+3 2.5+6 3.9+6
C 201014 9 O 4 5.1+6+6 8.5+8 5.6+7 0.1+5 -2.4+5 -.8+5
C 201014 10 O 5 3.0+10 4.1+12 1.9+8 -1.3+9 -1.1+7 -.8+8
C 201014 11 O 6 6.8+7 12.2+10 5.8+7 4.0+6 2.2+6 1.6+7
Z 201014 1 TEM 296
Z 201014 2 H WATER OXYGENS ARE NUMBERED DIFFERENTLY IN THE ORIGINAL
I 201014 1 0.042
```

**Figure 1.** Entry from the ICSD in the input format.

fined machine readable "input" file which is then accummulated at Bonn in an "input" archive file. To ensure that the information in the data base is reliable, it is subjected to objective and systematic checks by computer programs in both Bonn and at McMaster University. The checking procedures are evolving at both centers, but the following description gives an idea of checks that are currently being performed. The first stage is to check the syntax of the data. Although the input is in free format, the various fields must be of the correct type: integer, real number, or alphanumeric. Once this check is passed, the second stage is to check the values of the different fields. Here we can distinguish four types of data: (i) data which must be internally consistent (e.g., lattice constants, formula, and density), (ii) data which are in principle independent but whose values typically lie within a restricted range (e.g., temperature factors), (iii) 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), (iv) 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 incapable of being checked.

The first category comprises the chemical formula, lattice parameters, symmetry, and atomic coordinates. Figure 2 shows the various tests that are routinely performed. For example, the lattice parameters, symmetry, and atomic coordinates are used to calculate bond lengths which can be compared with sums of ionic radii. Because of the great variability of bond lengths in inorganic crystals this test picks up only gross errors. A more sensitive test which we are developing is to convert the bond lengths to bond valences

**Table I.** Types of Data Stored in Each Entry in ICSD

| | |
|---|---|
| N | compound name |
| V | common or trivial name |
| M | mineral name and source |
| F | chemical formula |
| U | title of paper from which the data have been abstracted |
| Q | citation of paper |
| A | authors' names |
| E | lattice constants (with standard errors)<br>number of formula units in the unit cell<br>experimental density |
| R | space group symbol |
| S | symmetry operators |
| T | structure types |
| P | atomic parameters, viz:<br>element symbol and identifier<br>oxidation state<br>multiplicity of site<br>Wyckoff symbol<br>positional coordinates (with standard errors)<br>occupation number<br>number of attached H atoms (if H coordinates not given) |
| B, C, D | isotropic or anisotropic temperature factors (with standard errors) |
| H | *Chemical Abstracts* registry number |
| I | crystallographic agreement index (*R* factor) |
| Z | comments (standardized and free text) |
| Y | test status |

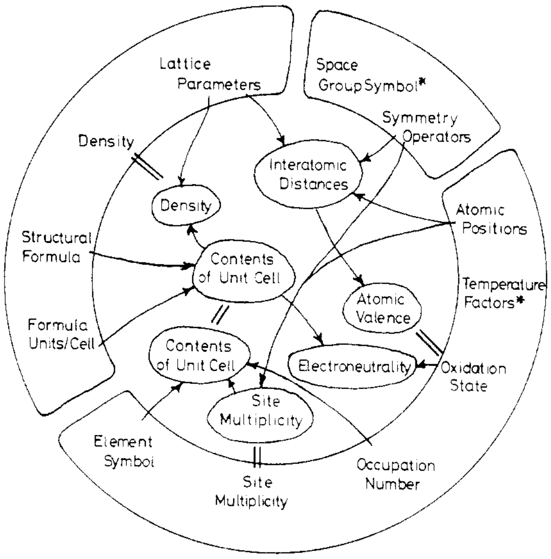


**Figure 2.** Summary of tests performed on data stored in ICSD. The stored data are shown within the ring. Quantities derived from these are shown inside the ring. Quantities connected by double lines should be equal. Quantities marked with an asterisk cannot be cross-checked.

whose sum around each atom should match the theoretical atomic oxidation states.[4]

In the second category are temperature factors and journal citations. It is always possible to check the volume number of a journal with the year of issue and to ensure that the page numbers lie within the range that occurs in that volume.

Typical problems in the third category originate from the variety of ways in which the names of authors,[5] minerals, and chemicals are written. Programs can help to standardize them by providing sorted lists in which the anomalies become apparent. Chemical nomenclature presents a special problem. The IUPAC rules[6] do not lend themselves to computer checking. We have therefore defined a list of element names and endings as well as special names for common groups (e.g., ammonium). With these it is possible to write the names of inorganic compounds according to a well-formulated set of rules.

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, and uncertain chemical composition. The third type are misprints appearing in the original publication. When necessary, ambiguities and errors are corrected by correspondence with the author. Corrections to the published data are usually noted in the comments.

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 = retrieval). 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.

## CURRENT STATUS

The intent of the data base is to include all complete reports of three-dimensional determinations of inorganic crystal structures. Collaboration with the corresponding organic and metal data bases ensures that all structure determinations are included in at least one data base and that duplication is minimized. This has resulted in working definitions of organic compounds (must contain a C–C or C–H bond) and of metals (must contain one element to the left of the Zintl line and no halogen or oxygen; the position of sulfides, suboxides, and metal hydrides is not yet clearly defined). At present the data base contains all structures from the relevant sections of ***Strukturbericht*** and ***Structure Reports***. All 23 000 can be retrieved for chemical elements and bibliographic data, but only about 70% so far contain the full crystallographic data. Completing and improving the retrospective data in the collection will be an ongoing project.

The literature since 1977 is being covered by a systematic search of the principal journals reporting inorganic crystal structures. *Acta Crystallographica* alone provides a quarter of the entries. Four journals cover the next quarter, while for the third quarter 11 journals must be searched. The remaining entries are taken from over 200 different journals, and most are located through abstract journals such as ***Chemical Abstracts*** and ***Bulletin Signaletique***. A comparison of lists of citations from these two reference works shows that they have different coverages and that neither gives complete coverage on a simple direct search. About 1200 entries are added each year from the current literature.

## MODES OF ACCESS

A system of programs has been developed for accessing the data base. At the lowest level is a set of subroutines that will extract from the Z file any given datum in a given entry. This allows the user to retrieve some or all of the data in an entry in a form suitable for listing or for further processing.

These subroutines can be used to provide the input to a user program, but they are also used in an on-line retrieval system written in FORTRAN for the IBM 370/168 computer in the regional computing center at the University of Bonn. A detailed user manual is available in German or English from the German authors. This system has also been implemented on the Siemens computer at FIZ at Karlsruhe (FRG) and for demonstration purposes on IBM computers at CISTI (Ottawa) and the University of Bern (Switzerland). It uses inverted files corresponding to frequently searched keys. Currently these are as follows: chemical element with oxidation state, group names of elements, mineral names, number of different elements present in the compound, standardized remarks, space group name, collection (entry) number, journal coden, author's name, year of publication.

The keys in the inverted files can be combined by using the logical operators .AND., .OR., and .NOT. In addition, string searches of all text fields stored in the data base are possible. The command "DISPLAY" can be used to examine the search keys used in the inverted files and the number of entries in each. "FIND" will give the number of entries having the characteristics sought, "SHOW" will list all or part of the entries on the terminal, and "COPY" will write the entries to a subfile for further processing. The retrieval routine will itself calculate bond lengths and angles, plot diagrams on a graphics terminal, and display powder patterns. The interactive command "HELP xxxx" will give instructions on the use of "xxxx" in either English or German depending on the version used. The full contents of the HELP file give a description of the use of the system.

The data base is accessible on line from FIZ in Karlsruhe, who will be responsible for marketing the data base in computer readable form. In Canada the data base will be available from the Canada Institute for Scientific and Technical Information, National Research Council of Canada, Ottawa.

## FUTURE DEVELOPMENTS

There are three thrusts for future development: (i) the completion of the data base, (ii) improvement of the retrieval system, (iii) linking with other data bases.

For the user it is particularly important to find all the relevant data in one place. The contents of *Strukturbericht* and *Structure Reports* are now included, and other sources (e.g., *Landolt-Börnstein*, *Chemical Abstracts*, *Bulletin Signalatique*, and *Crystal Data*) are being searched. Work is also continuing on improving the routines for checking the input data.

The present retrieval system allows ready access to entries on the basis of a number of important keys that are easy to program (e.g., element name). There are other types of search that involve more complex programming as well as a clarification of concepts. We plan to produce an index of chemical bonds both as an inverted file and in hard-copy form. Closely related to the concept of chemical bond is that of connectivity which would allow for the searching of fragments (e.g., $SO_4$, $SO_3$, and metal hexacarbonyl) or frameworks (e.g., in silicates). In both cases it is necessary to find an operational definition for a chemical bond, a concept which, in inorganic chemistry, has never been unambiguously defined. Another concept that is ill-defined but potentially useful for searching is that of structure type. Many inorganic compounds crystallize with structures that are the same or that are closely related. Such similarities often go unnoticed by the author, and, consequently, the "structure type" field in the data base is frequently left blank. We are developing an algorithm that will use the information in the data base itself to recognize structure types. A first step in this direction is the assignment (using the computer) of Pearson symbols, type formulas (e.g., AB, $AB_2$), and crystal class. The difficulties encountered in assignments of structure type and connectivity are not made any simpler by the existence of many structures with compositional or positional disorder or with atoms that have unassigned coordinates (typically H or interstitial $H_2O$).

The metal, inorganic, and organic crystal structure data bases, while containing the same essential crystallographic data, are organized on slightly different principles which reflect the different chemical concepts that have been found to be useful in these three areas. The boundaries between the data bases, however, are arbitrary, and many investigators will be working in areas that span two different data bases. In order to help them, we are working on a single retrieval system that would extract data from whichever data base contained the required structure. The computer, not the user, would decide which data base to access and would ensure a correctly formulated query. It is also useful to be able to access, in conjunction with the structural data bases, data bases of physicochemical information (e.g., thermodynamic data). Here it is necessary to identify unambiguously the same compound in the two different data bases. The Chemical Abstracts Registry Number works well for organic compounds but is not well designed for use with solid inorganic phases. So far no completely satisfactory registration procedure has been developed.

## CONCLUSION

In the near future details of all inorganic crystal structures will be accessible on line through the Inorganic Crystal Structure Data Base. Like the Cambridge Organic Crystal Structure Data Base, this archive will be more complete, more correct, and more accessible than the original literature. It will provide not only a useful service to chemists and crystallographers interested in reviewing previous structural work but will be a resource of incomparable value to solid-state chemists engaged in systematic studies of structure and properties.

## ACKNOWLEDGMENT

We acknowledge the encouragement given to us by the staff of the Cambridge Crystallographic Data Centre and the financial support provided by NATO, the Natural Science and Engineering Research Council of Canada, the Ministry for Research and Technology (FRG), the Canada Institute for Scientific and Technical Information, and the Commission of the European Community as well as for the support and encouragement given by our sponsor, the Fachinformationszentrum Energie, Physik, Mathematik.

## REFERENCES AND NOTES

(1) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. G.; Rogers, J. R., Watson, D. G. *Acta Crystallogr., Sect. B* **1979**, *B35*, 2331–2339.

(2) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouch, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535–542.

(3) Calvert, L. D. *Acta Crystallogr., Sect. A* **1981**, *A37*, C343–C344.

(4) Brown, I. D. *Chem. Soc. Rev.* **1978**, *7*, 359–376.

(5) Garfield, E. *Naturwissenschaften* **1981**, *68*, 519–520.

(6) IUPAC, "Nomenclature of Inorganic Chemistry", 2nd ed.; Butterworths: London, 1970.