The Impact of Electronic Publishing on the Academic Community

Session 6: Access to scientific data repositories

From private data to public knowledge

Olga Kennard

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1BZ, U.K.

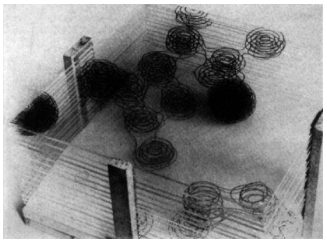
Introduction

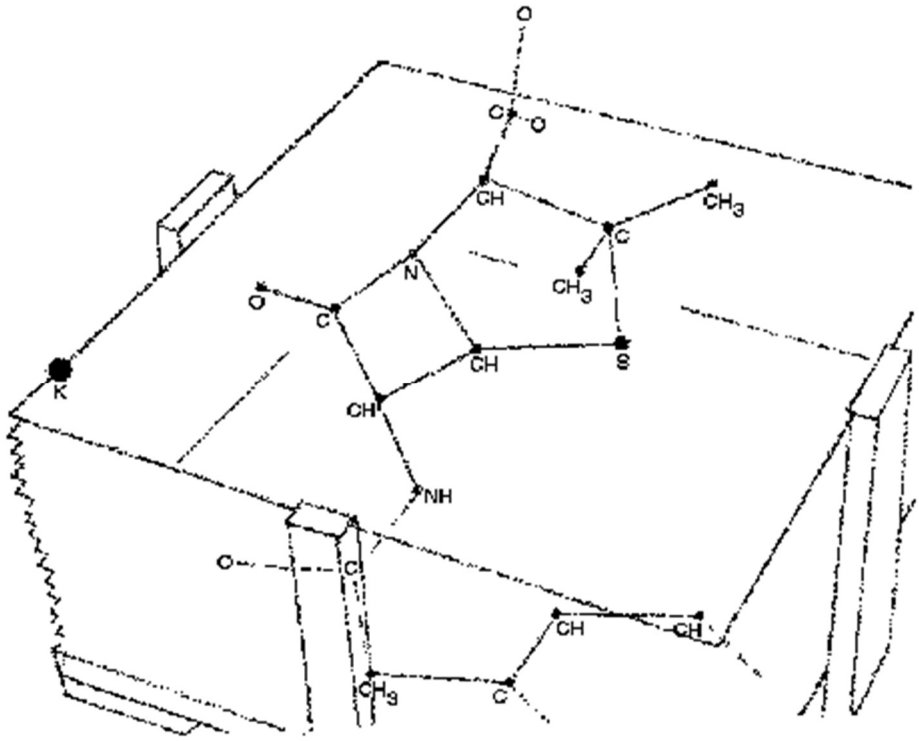
The Cambridge Structural Database (CSD) [1] is one of the oldest numeric scientific databases and thus provides a good example of the gradual transition from the traditional printed dissemination of data to the present electronic age. The database was established in 1965 to fulfil a dream of myself and a great scientist, the polymath J.D. Bernal. We had a passionate belief that the collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments.

What is crystallography?

In individual experiments we use X-rays to perform the structural analysis of crystals. Such analyses aim at defining, in numeric terms, the three-dimensional structure of molecules, and also indicate the forces between molecules which lead to the beautiful regularity of crystals. X-rays can be used to 'see' the individual molecules and arrays of molecules, not directly as in a light microscope, but indirectly by measuring the intensities of the X-rays scattered by the regular array of molecules in the crystal and recombining them mathematically as electron density maps. Figure 1 shows such a recombination for penicillin, carried out by Dorothy Hodgkin [2].

Figure 1. Electron density map of potassium benzyl penicillin and the skeletal representation of the penicillin molecule

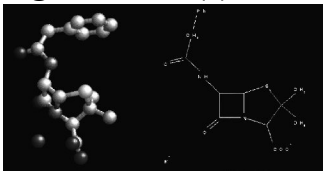




Derived by Dorothy Hodgkin from the X-ray diffraction pattern of penicillin crystals. Reproduced with permission from Princeton University Press, at http:\\pup.princeton.ed\class-_use\

The size and shape of the molecules is defined by the co-ordinates of the electron density peaks and the molecule can be constructed from these co-ordinates as a three-dimensional object (see Figure 2).

Figure 2. The (a) chemical and the (b) three-dimensional structure of penicillin



This figure also shows the chemical structure of the molecule. In the case of penicillin the chemical structure itself was deduced from the X-ray analysis, since a four-membered ring of carbons and nitrogen was previously unknown. Identification of the chemical structure is of course essential if the molecule is to be synthesized chemically and not merely isolated from natural products. The key numeric information is the co-ordinates of the atoms forming the molecule which can be expressed in the form shown in Figure 3.

Figure 3. Numeric entry on penicillin stored in the CSD

Potassium benzylpenicillin
C16 H17 N2 O4 S1 1-,K1 1+
D.Crowfoot,C.W.Bunn,B.W.Rogers-Low,A.Turner-Jones
The Chemistry of Penicillin, , 310,1949

Authors' cell dimensions

a	b	c	alpha	beta	gamma
9.360	6.370	30.350	90.0	90.0	90.0

K1	.352	.530	-.002
O1	.379	-.123	.056
O2	.359	.202	.058
C1	.436	.046	.061

.
.
.

This is the kind of data which was incorporated in the Cambridge Database with the eventual aim of using the numbers in a collective way. Other information stored was text, the usual bibliographic data (author's name, journal citation, etc.), as well as information about the symmetry relations between molecules forming the regular crystal.

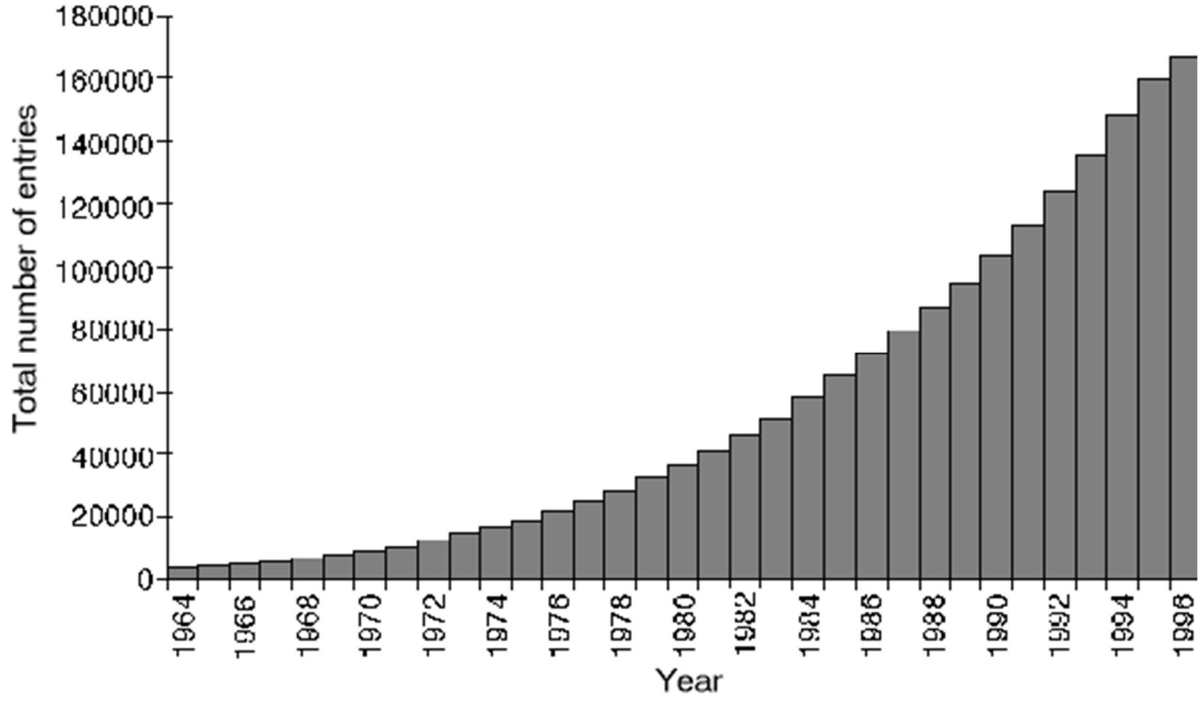
Validation of data

It was fortunate that there was a well-defined aim of using the numbers for specific purposes, for this made us realize, right from the outset, that it was essential to check each data set to ensure that future users would be able to rely on the numbers, and also for some collective use. This

may seem obvious, but it was not generally of serious concern to other database compilers at the time.

When the CSD was started numeric data was only available for relatively small molecules with less than 30 atoms, and all the co-ordinate numbers were printed as tables in the usual learned journals. There were about 1\500 such publications and it was feasible to devise methods to validate the entries both manually and by developing computer tests. About 20% of the published structures contained errors and inconsistencies which had to be corrected, often by correspondence with the authors. The annual increase at the time was slow, but the situation changed quite rapidly as shown in Figure 4.

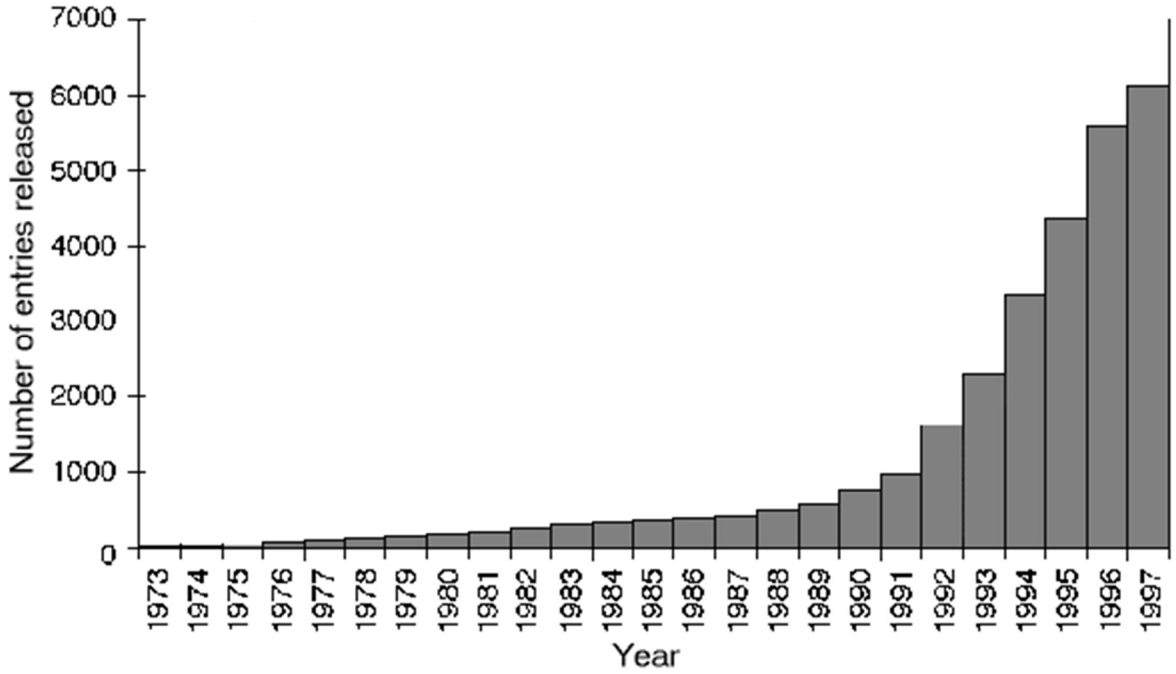
Figure 4. Growth in the number of entries in the CSD in the period 1965--1996



However, our checking procedures have remained at the same level or indeed have improved and the Data Centre has become a vital party in the refereeing process. Our check programs were made available to individual scientists and this helped to ensure that data entering the literature, whether through printed publications or electronically, was at least to some degree validated. Recently there has been much discussion about the loss of quality control if results are placed directly on the Web, but the experience of the Cambridge Crystallographic Data Centre (CCDC) shows that data centres can play a significant role in the refereeing process.

The growth of the related protein database, the Protein Databank (PDB), maintained by the Brookhaven National Laboratories in the U.S.A. [3], was much slower, as shown in Figure 5.

Figure 5. Growth in the number of entries in the Protein Databank in the period 1975--1996



Proteins contain many thousand of atoms, in contrast to the 'small' molecules of the CSD, and the analysis of protein structures only became possible with major improvements in methodology, using large computers and powerful X-ray sources. The PDB also started, as shown in Figure 5, with relatively few structures, but the extensive numeric data generated by the analyses was not included in printed publications. Instead it was deposited by individual scientists directly in the Databank. There was no special quality control and the 'curators' of data relied on the authors to take care of the accuracy of the entries. Unfortunately, this approach was often too optimistic and the PDB has only recently begun to develop validation tools to ensure the integrity of the data.

Enhancement of the data

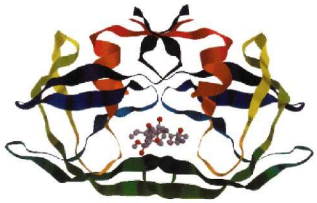
The second effect of building a database with a clear scientific aim was that we realized the need for enhancing the data with additional information. The enhancement involved correlating the three-dimensional structures of the molecules in the Database with their chemical structures. Such correlation made it possible to search for a combination of chemical and structural attributes. An example is the derivation of the length of chemical bonds between specific atoms in various chemical environments. A search for distances between say a carbon and oxygen within molecules stored in the Database would yield the value of the length of the carbon--oxygen bond, determined in hundreds (or thousands) of experiments. These values could then be analyzed statistically and the 'best' values derived. This is one example of how 'private' data can be brought together to build up a 'public' knowledge base. We have indeed done this and have produced extensive tables of bond lengths for different types of bonds in different environments [4]. This knowledge was then utilized by others to write computer programs to construct the accurate three-dimensional structure of molecules directly from their chemical structures without

the need for further experiments. Such programs are now used to generate libraries of three-dimensional co-ordinates for many millions of compounds.

Three-dimensional versus two-dimensional

The shape of molecules is as important as their dimensions and this is an aspect which cannot as yet be readily computed. Figure 2 illustrates the leap of imagination or experiment needed to translate the stylized two-dimensional chemical information into a real three-dimensional object. A special shorthand, a string of 'words', each word representing one of the chemical units (amino acids) of a protein is used to describe the chemical structure of a protein. This is even more difficult to visualize as a three-dimensional object than in the relatively small molecules so far considered. In Figure 6 the three-dimensional structure of a moderately sized protein, HIV protease, is shown with a small inhibitor molecule bound to it [5].

Figure 6. The three dimensional structure of the protein HIV protease and a small inhibitor molecule bound to the active site



A knowledge of the three-dimensional structure of protein--inhibitor or protein--drug complexes permits a systematic approach to the search for and development of new drugs. The CSD system is an important tool for this, since numeric information is available for many hundreds of molecules. The Database is now widely used for rational drug design by pharmaceutical companies.

Dissemination

This brings me to the methods of dissemination and the question of electronic access to the data. Here too our experience might be helpful.

Free access to validated and enhanced data worldwide is a beautiful dream. The reality, however, is more complex. Firstly, on the technical side, data input must be standardized. We have been heavily involved in specifications for a crystallographic input format (CIF) but this raises too many problems to consider in detail here. The crystallographic community has found that our experiences as old fashioned editors and data validators are relevant to the specifications of CIF files, and the programs we devised at the CCDC are used to validate data submitted directly from authors. Priority problems, however, rear their ugly heads, since some journals are formulating rules that deposition on the Web (or with our Database) will count as prior publication and the work will not then be accepted by the journals. We also receive data deposited with journal

editors as part of a printed paper, which again has its own problems, although the Data Centre can offer considerable help with the refereeing process.

Our experience with electronic access is that users who wish to access individual data items, particularly journal references and citations, are well served by electronic networks. However, when it comes to the use of numeric data for in-depth research, networks are not as suitable as compact discs which can be installed in individual laboratories. We issue the Database and associated programs on CD twice a year and distribute these through a *human* network of Affiliated Centres in some 40 countries. Local electronic networks can be set up under this scheme, thus providing for both the casual and the in-depth user. This network has two advantages. First, that each national centre knows the needs of its local community, and can initiate workshops and training sessions to ensure the skilled use of the information in the Database. The second advantage is economic. Each country pays an annual subvention to the CCDC to help with our activities. This usually comes in the form of a government grant and can be adjusted to the size of the country's scientific population. We introduced this policy in the 1970s when there was no concept of actually paying for information, in order to ensure that young scientists, in particular, had access to the Database. Added to this is income from industrial companies, which has put the CCDC on a very sound financial footing. We do not have to rely on special grants from government sources unlike some data centres, whose data is accessed, free of charge, electronically, but whose existence could well be threatened if the special grant is withdrawn. Indeed this is what happened to SwissPro, an essential database of protein sequences. In future it might well be possible to offer subscriptions for access of data on the Web, but the CCDC is a good model for the present situation.

Compression of data

I would like to return to science, to the provision of compressed data, in the form of knowledge bases. I have already mentioned one, the knowledge base of molecular dimensions. We are now engaged in constructing another knowledge base, of intermolecular interactions, which extracts from the individual data items general rules for the way molecules interact with each other. Each individual user could of course construct the rules for such interaction themselves, utilizing hundreds of data sets in the Database. However, providing these rules in an easily searchable and visual way frees users from this task and gives them the opportunity of using the knowledge base in a variety of ways from basic science to practical applications.

Summary

It has been the experience of the CCDC that ensuring access to a highly specialized and rapidly evolving database has not been technologically limited since technological developments were well able to keep pace with the exponential growth of information stored in the database. It has, however, been important for the curators of the data to strive and anticipate likely problems, such as the potential errors in the data and hence the development of validation tools; the enhancement of data so that it can be used optimally; and the development of knowledge bases to reduce the labour of having to access large data sets for fundamental information. The cost of maintaining the activities of the data centres is a major factor and the way this is achieved by the CCDC is one possible model. Other models are evolving and we need to pool our experience, as

at this conference, to ensure that the best use can be made of unprecedented opportunities arising from new technological developments in electronic communications.

References

1. Allen F.H. and Kennard O. (1993) Chemical Design Automation News **8**, 31--37

2. Crowfoot, D., Bunn, C.W., Rogers-Low, B.W. and Turner-Jones, A. (1949) The Chemistry of Penicillin (Clarke, H.J., Johnson, J.R. and Robinson, R., eds.), p. 310, Princeton University Press, Princeton, NJ

3. Bernstein, C., Koetzle, T.F., Williams, G.J.B., Meyer, Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. **112,** 535--542

4. Allen, F.H., Kennard, O., Watson, D.G., Brammer, L., Orpen, G. and Taylor, R. (1987) J. Chem. Soc. Perkin Trans. **II**, S1--S19

5. Rose, R.B., Craik, C.S., Douglas, N.L. and Stroud, R.M. (1996) Biochemistry **35**, 12933--12944; Abdel-Meguid, S.S., Metcalf, B.W., Carr, T.J., Demarsh, P., Desjarlais, R.L., Fisher, S., Green, D.W., Ivanoff, L., Lambert D.M. and Murthy, K.H. (1994) Biochemistry **33**, 11671--11677
