Operation of an International Data Center: Canadian Scientific Numeric Database Service†

Gordon H. Wood,* John R. Rodgers, and S. Roger Gough

National Research Council of Canada, Montreal Road, Ottawa, Ontario K1A 0S2, Canada

Received June 4, 1992

The Canadian Scientific Numeric Database Service (CAN/SND) is a focal point for Canadian activity in numeric database production, dissemination, and exploitation. Production efforts are concentrated on CRYSTMET, the NRC Metals Crystallographic Database, which contains structural, chemical, and physical data on over 40 000 metallic phases. Dissemination is principally via an international online network, featuring the complete set of crystallographic databases supported by an online document ordering service. A secondary method of distribution is via magnetic media. Enlightened exploitation of the vast reservoir of information contained in these databases is encouraged through research into novel ways of classifying, interpreting, and viewing data, especially those pertinent to materials science. CAN/SND is a component of the Canada Institute for Scientific and Technical Information of the National Research Council of Canada.

I. INTRODUCTION

As an international data center, the Canadian Scientific Numeric Database Service (CAN/SND) produces a numeric database, operates an international online numeric database service, and carries out research on the informed use and exploitation of scientific numeric data.

CAN/SND is a service provided by the Canada Institute for Scientific and Technical Information (CISTI), which is part of the National Research Council of Canada (NRCC), a corporate agent of the Government of Canada. The guiding philosophy of CAN/SND, from its inauguration in 1980, has been the importance of the scientist or engineer as end-user. Decisions concerning enhancements or changes to the services are therefore user-driven rather than computer science-driven, and user feedback is actively solicited. Modifications must make the services more useful scientifically not simply more elegant or sophisticated.

Not being a large enterprise, CAN/SND has chosen to focus its resources on a limited number of disciplines with the current emphasis being given to crystallography and molecular biology. Leverage on limited resources is obtained by acquiring databases via exchange agreements and leases and by collaborating with others in production and research endeavors.

In what follows, Sections II and III outline the database production and database dissemination activities, respectively. Section IV sketches the research and development being undertaken.

II. DATABASE PRODUCTION: CRYSTMET

Since 1984, CAN/SND has been responsible for all aspects of the production of CRYSTMET, the National Research Council of Canada Metals Crystallographic Data File. CRYSTMET is a computer-readable database containing critically evaluated crystallographic, chemical, physical, and related bibliographic data for metals, alloys, and intermetallic compounds studied by diffraction methods.[1] Prior to 1984, entries for the database were collected first by D. T. Cromer at Los Alamos Scientific Laboratories, then by W. B. Pearson at the University of Waterloo, and then by L. D. Calvert at NRCC. By virtue of a formal agreement with the NRCC in 1988, Dr. Pierre Villars of Villars Intermetallic Phases Databank in Switzerland became involved in the production of CRYSTMET and has since assumed the role of a contributing editor.

A. Description of CRYSTMET. CRYSTMET may be characterized as a numeric/factual database—*numeric* in the sense that it contains crystallographic data such as atomic coordinates and lattice parameters; *factual* in the sense that it contains structural, chemical, and physical information such as the structure type, space group designation, chemical formula, and density. In addition, each entry contains sufficient bibliographic information [author(s) and citation] to facilitate consultation of the original works. A typical entry is shown in Figure 1.

Thanks to understandings among the producers of the structural databases, CRYSTMET complements the coverage of the Cambridge Structural Database (CSD, organics and organometallics) and the Inorganic Crystal Structure Database (ICSD, inorganics). The relationships among these three databases and the U.S. National Institute of Standards and Technology's Crystal Data File (CRYSTDAT) are best understood by reference to Figure 2.

Currently at 40 000 entries, CRYSTMET exhaustively covers the literature from 1913 to 1990; annual updates of 1500–2000 entries serve to keep the database complete. For all entries, it is required that the composition of the compound be clearly defined. Candidate compounds are taken to be metallic if they are composed of elements to the left of the Zintl line in the periodic table (see Figure 3). Cross-compounds, phases formed with elements immediately to the right of the line, are included along with some dioxides. All other oxides and all compounds with halides or noble gases are excluded.

B. Abstraction of Data and Metadata. Figure 4 shows the steps involved in the production of CRYSTMET. Steps one and two are performed by Dr. Villars; steps three and four take place under the supervision of one of the authors (J.R.R.) at the NRCC in Ottawa.

Relevant entries are found primarily via cover-to-cover manual scanning of journals known to be rich in metals and intermetallics and secondarily by meticulous scrutiny of output from abstracting services. Photocopies of each pertinent article are prepared and suitably marked for the recording of all germane information.

C. Scientific Evaluation. For a database like CRYSTMET to be useful, it is vital that the data be critically evaluated and

† NRCC Publication No. 34250.

0095-2338/93/1633-0031$04.00/0 Published 1993 by the American Chemical Society

```
ID:  16291
FO:  Nb3 Re2
TY:  W  sc=cI2
AU:  Knapton AG
RE:  J. Less-Common Metals(JCOMAH),1,480,1959
AC:  a=3.203 spgr=Im-3m spno=229 sys=cubic z=.4
SL:  Nb1-xRex, x= 0-0.43 at 2073K, a= 0.3300-0.3198nm,
     (information taken from figure)
RM:  unit cell dimension taken from figure; M= Nb, Re
AT:  M 2a m-3m 0 0 0 1
```

Figure 1. Typical CRYSTMET entry. Explanations of abbreviations are as follows: ID, local identification number of the entry; FO, chemical formula; TY, structure type; sc, structure code; AU, author(s) name(s); RE, reference; AC, published crystallographic data, a, b, and c axes, α, β, and γ interaxial angles; spgr, space group symbol; spno, space group number; sys, crystal system; dx, calculated density; z, number of formula units per unit cell; SL, solute (formula range and associated cell parameters for solid solutions); RM, remarks; and AT, atomic coordinates.

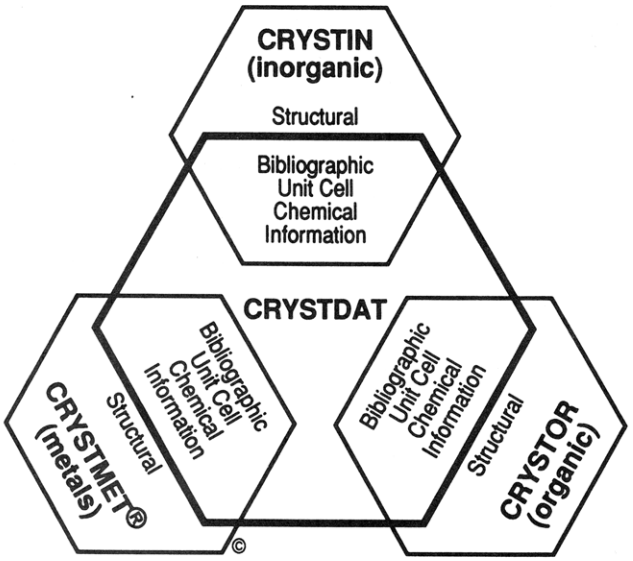

Figure 2. Relationships of the crystallographic databases. (Note that CRYSTIN, CRYSTOR, and CRYSTDAT represent the ICSD, and CSD, and Crystal Data, respectively, as implemented on CAN/SND.)

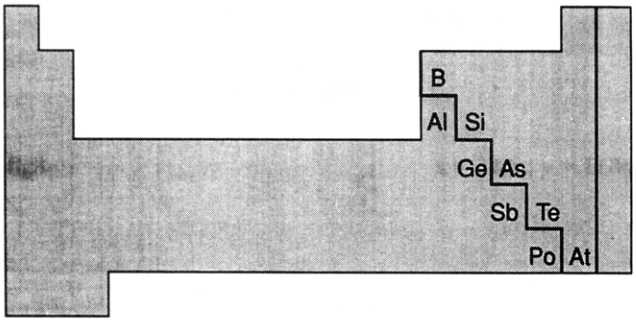

Figure 3. Division between inorganics and metallics.

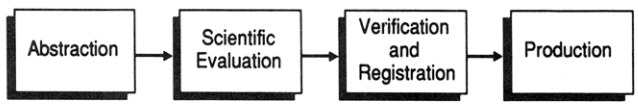

Figure 4. CRYSTMET production.

error-free to a very high degree. In this specialized, labor-intensive, time-consuming, and expensive step, the data are compared to previous work on related compounds and appraised on their crystallographic and structural merits. Concurrently, checks are made for correctness and quality. Depending on the error condition, appropriate adjustments may be made or a note of caution concerning the data may be added to the entry.

D. Registration in Archive Database. In this relatively automated stage, each entry is subjected to additional verification tests. A comprehensive cross-check of crystallographic specifications and calculations of other dependent parameters serves to identify other related entries (e.g., same elements but different stoichiometry) and to detect any internal inconsistencies. Possible duplicate studies of the same phase are sought by comparing the reduced cell and the calculated powder patterns of candidate entries with those of existing database entries.

Another elementary, yet mandatory, check is the enforcement of the standardization of authors' names necessitated by the inconsistent manner in which these names often appear in the literature, whether by variants in presentation, in spelling, or in transliteration. Related to that is the need to verify the literature reference for accuracy and possible duplication. A more detailed description of the evaluation and registration processes is given in ref 1.

E. Physical Production. This final step results in the production of both an archive and an online form of the database. The archive version, the format in which Licensees receive the database, is a flat file in which new entries are simply concatenated to the existing ones.

For the online version, techniques are employed which exploit data properties and access patterns to provide efficient data storage and access.[1] Briefly, primary indexes are created for the main components of each entry along with secondary indexes for certain fields. The records, each containing all the pertinent information for a single entry, have variable length with a variable number of fields, many of them indexed for searching. A database table gives the following additional information about each field in the record:

Index: indicates the presence of an index for a field; certain fields (e.g., atomic parameter) are not indexed.

Data types: exhibits the data types for each field (e.g., integer, decimal, or character) and the associated format (e.g., fixed or floating). This information is used during searching to convert user input strings and also during analysis to inhibit certain types of operations such as attempts to carry out a numeric application on a nonnumeric field.

Compression: indicates whether compression techniques have been employed for a field. Data in certain fields, e.g., the coordinate field, are compressed to reduce storage and buffer requirements while increasing the speed of data access and transfer. These advantages generally offset the extra processing time required for the compression and decompression operations.

This version is made available online internationally as one of the physical sciences databases as described in the next section.

III. DATABASE DISSEMINATION

The principal means by which the databases offered by CAN/SND are made available to the international scientific community is via interactive online networks with the host computers located in Ottawa. Secondary distribution is effected by the licensing of magnetic tape copies of some databases for private or other limited use. A tertiary means of access is provided by 'valet' or customized searches.

A. Online. Figure 5, a schematic of the current online system operated by CAN/SND, shows the databases grouped into three basic disciplines: molecular biology, molecular structure, and analysis. (The full names of the databases and

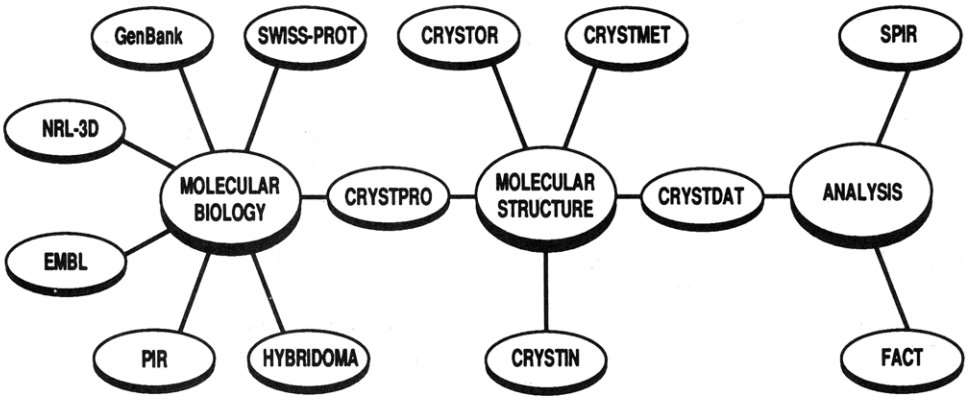

Figure 5. Databases available on CAN/SND.

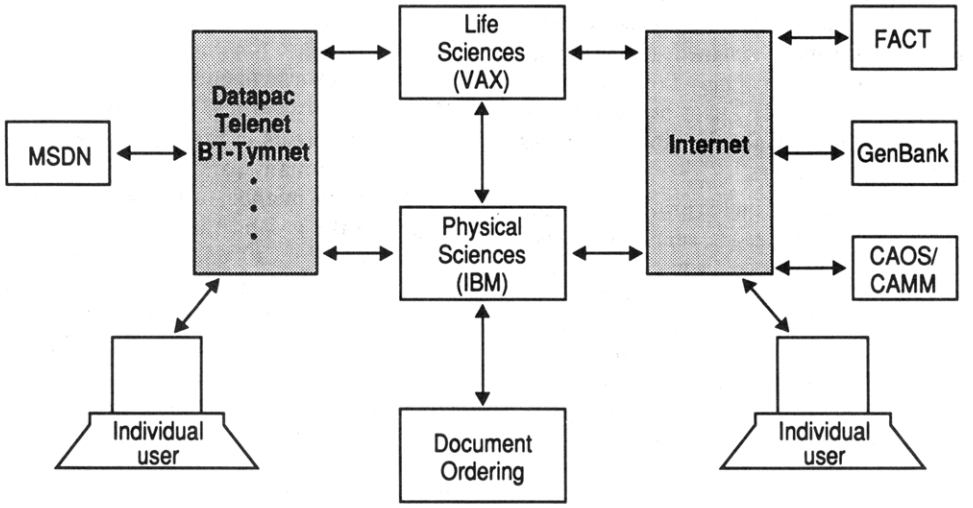

Figure 6. Means of online access.

some additional information about each are given in Appendix I.) CAN/SND is particularly fortunate to be able to offer its clients the full spectrum of structural databases along with an integrated search system.

The rights to offer these databases online, apart from CRYSTMET of course, are obtained through agreements with the database producers. One of the ongoing tasks is to ensure that the updates—most of the databases are updated at least annually—are available as soon as possible.

As illustrated in Figure 6, the databases and their associated searching and analysis software packages reside on VAX and IBM platforms. Primarily because of software considerations, the Molecular Biology ('Life Sciences') group shown in Figure 5, with the exception of 'Hybridoma' and 'CRYSTPRO', is implemented on a small VAX (Model 3400) and the Molecular Structure and Analysis groups ('Physical Sciences') are implemented on an IBM 3090 mainframe. The two platforms function essentially independently of each other in that each machine performs the required search and analysis routines; processing is not shared. Files relating to system administration are exchanged automatically, and there are full TCP/IP communication linkages between each machine and the Internet. In addition, each platform has a separate address on the commercial packet-switched networks to accommodate clients limited to that means of access. Most clients work interactively, but some, depending on the priority or complexity of the problem, choose to work in batch mode.

Several gateway arrangements are used to facilitate access and to avoid duplication of effort. Thus, clients of the Microbial Strain Data Network (MSDN) in the U.K. or clients of the Computer Assisted Organic Synthesis and Computer Assisted Molecular Modelling Center (CAOS/CAMM) in The Netherlands simply select CAN/SND as a menu item on their respective services, and they are automatically connected to the database of their choice on CAN/SND. Similarly, those wishing to use FACT on CAN/SND are passed through to McGill University in Montreal, Canada, where the master copy of FACT is maintained. Provision is also made for those wishing either additional computing resources or special search and retrieval capabilities in the sequence databases to access them interactively via Gopher, an Internet-based distributed information system.

To be truly useful tools, of course, databases must not only be searchable but be searchable in as many scientifically interesting ways as feasible. With this in mind CAN/SND has developed, and continues to enhance, a proprietary, custom-designed search, retrieval, and analysis (SRA) system (GENSEARCH) for the molecular structure databases. Acclaimed for its easy to learn, intuitive syntax, GENSEARCH is extremely economical in machine time and powerful in the scope of question that can be addressed. Several SRA systems are used for the molecular biology databases, but the primary one is that developed and maintained by the Genetics Computer Group in the United States.[2]

The Brookhaven Protein Data Bank (CRYSTPRO in Figure 5) is also made available via a fileserver mode. Scientists interested in the atomic coordinates of a particular protein need only send a 'mail' message suitably identifying that protein to the Internet address of the CAN/SND fileserver. By way of Internet, the coordinates arrive shortly at the requester's computer.

Analogously, interactive users may send files of results to their 'home' Internet address for local geometric or graphic analysis.

A new feature to be active shortly, indicated in Figure 6, will be the capability to order documents online. With this facility, users who have completed a search and wish to see the actual work from which the data were abstracted will be able to request a photocopy of the relevant document and have it delivered to their location. If a user were to have an urgent need, the document could be delivered within hours by means of telefacsimile; less urgent needs would be met by courier service or conventional post.

B. Licensing of Databases. For those scientists having suitable local computing resources, combined with the time to devote to database maintenance, the option of leasing a private copy of a database is made available. An agreement laying out the conditions for use must be signed, and an annual license fee must be paid. At this time, only the Cambridge Structural Database and CRYSTMET can be obtained in this fashion. Those wanting licenses for the other databases must deal directly with the producers.

C. Customized (Valet) Searches. Finally, provision is made for those scientists who, for various reasons, cannot or will not avail themselves of the two means of database access just described. For a customized search, the client outlines the problem in consultation with CAN/SND staff who then execute the required searches and analyses and return the results. As equipment for accessing computers online becomes more easily available and as more and more scientists become comfortable with the procedures, the need for this service is diminishing.

IV. DATA EXPLOITATION: RESEARCH AND DEVELOPMENT

To remain relevant and useful to its clients, CAN/SND is committed to carrying out research and development, not only in areas related to the storage, retrieval, and analysis of data but also on the informed exploitation of the wealth of information contained in these databases. Two examples of these activities, both carried out in collaboration with external colleagues, will be described.

A. Regularities and Prediction.[3] Databases like CRYSTMET represent an immense deposit of information waiting to be mined by those who take the appropriate approach. Looking up values for properties or seeking details of crystal structure of intermetallics are certainly legitimate and important uses of CRYSTMET. What is often ignored is the knowledge latent in the data themselves about, for example, how compounds form or how properties relate systematically to structure. Some of the investigations that may be carried out on the data themselves are

1. *Data evaluation and registration*: checking the crystallographic consistency of newly reported compounds with those of known, related materials[4]
2. *Property prediction*: inferring physical properties of yet-to-be formulated compounds from a knowledge of related compounds with similar stoichiometry, chemical composition, and structural features[5]
3. *Gap identification in intermetallic systems*: discerning regions in 'compound formation space' where no elemental combinations are known in order to focus experimental efforts and obviate attempts to combine elements with a low probability of compound formation[6]
4. *Relationship determination among structure/properties/chemistry*: seeking insights concerning these affinities that may be used to develop more astute identification strategies or to test fundamental theories of compound formation[7]
5. *Scientific intelligence gathering*: detecting trends in the types of materials being studied and observing how these trends vary with time with a view to understanding how scientific activity evolves or even to appraising a potentially useful field of investigation[3]

B. Extended Identification Techniques. For several years, CAN/SND has offered a facility on the CRYSTDAT database (see Figures 2 and 5) for identifying unknown compounds on the basis of their cell parameters a, b, c, α, β, and γ where a, b, and c are the edges of the cell of the unknown and α, β, and γ are the associated interaxial angles. As long as these data are reasonably accurately known, the identity of the unknown compound can usually be readily deduced by retrieving entries from the database having similar lattices.

Once the errors in the parameters exceed a few percent, however, too many candidate entries are retrieved for this identification method to be practicable. Ignoring the cell angles on the ground that small errors in the lattice of the unknown could lead to significant errors in the cell angles reduces the 'noise', but still too many candidate hits are found. Clearly there is a need to supply more information about the unknown to increase the discrimination of the search.

Experience has shown that the addition of any chemical or physical properties of the unknown into the search specification decreases spurious hits considerably. Even something as rudimentary as specifying one element known to be present or giving a range of densities can reduce 'noise' by at least 50%.

Recent research reveals that further significant reductions are obtained by introducing powder diffraction lines to the search strategy should the user have them available. Only three to six lines are required; it is not necessary to know the intensity of the lines, just the spacings. The effectiveness of this technique is perhaps best illustrated by considering an example:

> The lattice parameters for an unknown compound are determined to be a = b = c = 3.57 Å and α = β = γ = 90°; because of experimental problems, these values have an uncertainty of ±0.2 Å. Its density has been measured to be in the range 7.2–7.7 g cm^{-3}. From a powder pattern taken from the same compound, lines have been located at 3.62, 1.78, 1.45, and 0.98.

Using GENSEARCH on the CRYSTDAT database on CAN/SND, the following results were obtained:

1. Searching on a, b, c, α, β, and γ yielded 214 hits
2. Intersecting those 214 hits with all the entries in the database having densities in the range 7.2–7.7 g cm^{-3} gave 14 hits
3. Entering the four powder lines resulted in 2513 hits
4. Intersecting the sets of 14 and 2513 hits yielded a tractable collection of six hits, one of which corresponded to the 'unknown'

It is evident that such a capability represents an important addition to the tools available to the analytical laboratory, and CAN/SND encourages its clients to exploit it. To the writers' knowledge, CAN/SND is the only publicly available service providing these complementary search capabilities on one system.

Finally, it should be observed that promising extensions to this approach are anticipated from work recently reported by Le Page and one of the authors (J.R.R.).[8] Their studies of lattice retrieval utilizing the three shortest direct and reciprocal lattice parameters show great potential for searching efficiently even when the axial lengths of the unknown cell contain up to 5% experimental uncertainty.

V. CONCLUSION

This brief overview has attempted to demonstrate that an international data center as CAN/SND consists of a number of complementary and synergistic components. Clearly, and in contrast to the passive role to which some might relegate it, a data center is more than a gatekeeper or a collector and disseminator of information. It is a place where the intrinsic value of data is appreciated, enhanced, and proclaimed.

APPENDIX I: DATABASES AVAILABLE ON CAN/SND

Molecular Biology. Databases in this family feature, in addition to the sequence data listed for each, bibliographic information and technical comments.

PIR	*Protein Identification Resource* (National Biomedical Research Foundation), sequence listings for 40 000 proteins
EMBL	*European Molecular Biology Data Library*, sequence listings for 63 000 nucleic acids (many in common with those in GenBank)
NRL_3D	*Naval Research Laboratory-3D* (U.S. Naval Research Laboratory), 1200 sequences from protein structures contained in the Brookhaven Protein Data Bank (see CRYSTPRO below)
GenBank	*Genetic Sequence Data Bank* (National Center for Biotechnology Information), sequence listings for 65 000 nucleic acids
SWISSPROT	*Swiss Protein Sequence Database* (European Molecular Biology Data Library), sequence listings for 23 000 proteins (some duplication with those in PIR)
HYBRIDOMA	*CODATA/IUIS Hybridoma Databank*, 25 000 cloned cell lines and their immunoreactive products

Molecular Structure. As the name implies, databases in this group feature structural data of crystalline solids. In addition, they contain related bibliographic, unit cell, and chemical information as appropriate. (Though mentioned earlier in the text, the names are repeated here for completeness.)

CRYSTPRO	*Brookhaven Protein Data Bank* (Brookhaven National Laboratory), information on about 800 biological macromolecules, e.g., proteins, nucleic acids, viruses, and polysaccarides
CRYSTOR	*Cambridge Structural Database* (Cambridge Crystallographic Data Centre), information on about 91 000 organic and organometallic compounds
CRYSTMET	*NRC Metals Crystallographic Database* (National Research Council of Canada), information on some 40 000 metal phases
CRYSTDAT	*NIST Crystal Data File* (U.S. National Institute of Standards and Technology), information on all crystalline solids for which the basic cell parameters are known; currently about 185 000 entries
CRYSTIN	*Inorganic Crystal Structure Database* (Fachinformationszentrum, Karlsruhe), information on about 31 000 inorganic substances

Analysis.

SPIR	*Search Program for Infrared Spectra* (American Society for Testing and Materials), a collection of about 140 000 infrared spectra of some 96 000 compounds; entries consist of peak locations with a minimum of intensity information
FACT	*Facility for Analysis of Chemical Thermodynamics* (Thermfact), data on over 3800 inorganic stoichiometric compounds are available for use with a number of modeling and analysis programs for thermochemistry

REFERENCES AND NOTES

(1) Rodgers, J. R.; Wood, G. H. In *Crystallographic Databases*; Allen, F., Bergerhoff, G., Sievers, R., Eds.; International Union of Crystallography: Chester, U.K., 1987; p 96.
(2) Genetics Computer Group, 575 Science Dr., Madison, WI 53711.
(3) Rodgers, J. R.; Villars, P. Statistical Studies for Intermetallics—Stressing the Validity of 3 Principles. *J. Alloys Compd.* **1993**, in press.
(4) Rodgers, J. R.; Mighell, A. D. The Use of Lattice and Empirical Formula in the Registration of Crystalline Materials. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 42–7.
(5) Villars, P.; Phillips, J. C.; Chen, H. S. Isocahedral Quasicrystal and Quantum Structural Diagrams. *Phys. Rev. Lett.* **1986**, *57*, 3085–8.
(6) Villars, P.; Mathis, K.; Hulliger, F. In *Structures of Binary Compounds*; de Boer, F., Pettifor, D., Eds.; North-Holland: Amsterdam, 1989; Vol. 2, p 1.
(7) Rabe, K. M.; Phillips, J. C.; Villars, P.; Brown, I. D. Global Multinary Structural Chemistry of Stable Quasicrystals, High T_c Ferroelectrics and High T_c Superconductors. *Phys. Rev.* **1992**, *B42*, 7650–76.
(8) Le Page, Y.; Rodgers, J. R. Robust and Discriminating Search Parameters for Lattice Retrieval from Crystallographic Databases in the Presence of Large Experimental Uncertainties, American Crystallographic Association 1992 Annual Meeting, 1992; Poster PA 106, p 107; J. Electr. Diffr., submitted for publication.