Struct Bond (2010) 134:37–58
DOI:10.1007/430_2009_5

Published online: 1 September 2009

Data Bases, the Base for Data Mining

Christian Buchsbaum, Sabine Höhler-Schlimm, and Silke Rehme

Abstract Data collections provide a basis for solving numerous problems by data mining approach. The advantages of data mining consists in the retrieving of a new knowledge from existing information. The comprehensiveness of the data collection, the structure and quality of the data, and the selection of relevant data sets are extremely important to get correct results. In the crystallographic field, scientists will find several databases dealing with crystal structures of inorganic and organic compounds, or proteins. Usually databases have detailed data evaluation mechanisms integrated in their database production process and offer comprehensive and reliable data sets. The CIF standard enables the scientists to exchange the data. As an example, the Inorganic Crystal Structure Database (ICSD), a source of information for crystallographers, mineralogists, physicists, and chemists will be presented here. The ICSD contains about 120,000 entries (March 2009) of fully determined crystal structures. This chapter gives a detailed description of data collection, the contents of the data fields, data evaluation, and finally search the functionality of the ICSD database.

Keywords: Crystallographic databases · ICSD · Data collection · Data evaluation · Database functionality · Database design · Search strategies

Contents

C. Buchsbaum, S. Höhler-Schlimm, and S. Rehme (✉)
Fachinformationszentrum Karlsruhe, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen
e-mail: crysdata@fiz-karlsruhe.de

1 Introduction

With the abundance of all kinds of extensive data collections and the emergence of new technical possibilities, data mining has become an important issue in recent years. In the field of crystallography, databases like ICSD, CSD, ICDD, and others offer unheard of possibilities, e.g., in the fields of molecular design and crystal engineering. For this, the available data are evaluated using different criteria. In all fields of science, data mining would be impossible without extensive databases. Since the data mining is a powerful algorithm to analyze data and predict properties, the structure of database standardized data are of prime importance. Any analysis of selected data fields will be useless if the information in these data fields is not available in standardized form. Looking at the content of crystallographic databases, standardization starts with the unification of units and ends with the calculation of standardized crystal data or the classification of structures like structure types. A standard data exchange format, the Crystallographic Information File (CIF), was defined in the early nineties in order to enable the exchange of data between the most varied applications. Consequent application of this standard has made it possible to evaluate data deposited at FIZ, CCDC, or publishing houses without difficulty. Completeness of the data material is the second important aspect. The more extensive the data material, the more satisfactory the results of the statistical methods will be in the end. Thirdly, the data quality is a decisive factor. Prior to the data mining process, the data should be checked and evaluated carefully. Last but not the least, the choice of datasets is an important concern. The higher the amount of relevant content in the selected clusters, the more accurate the results of the data mining process will be. Databases like ICSD, CSD or PCD offer manifold search fields and combination potentials in their databases which leads to precise search results. Some aspects of retrieval are dealt with in this article in more detail.

The chapter will focus on the above mentioned aspects. As a producer of the ICSD database, FIZ Karlsruhe has decades of experience in compiling data, processing data, and making them available to customers. This chapter describes the FIZ Karlsruhe procedures for data collection and data processing, which are probably similar to those used by other producers of crystallographic databases.

Table 1 An (incomplete) overview of available crystal structure databases.

Provider	Raw Data	No. of entries (year)	Search fields	Software
FIZ	ICSD	120,000 (2009)	Structure	FindIt, WWW
ICDD	ICSD + LPF	285,000 (2008)	Powder pattern + structure	PDF4+/PDF2
CSD	CSD	470,000 (2009)	Structure	ConQuest
Crystal Impact	PCD	165,000 (2008)	Structure + Powder pattern	PCD
PDB	PDB	57,000 (2009)	Structure	N/A
COD	AMS + COD	48,000 (2006)	Structure	Online, free
NISTMet	Metals	40,000+	Structure	N/A
Toth	Metals File/CrystMet	126,000 (2008)	Powder pattern	TothToolkit
PCD	PDB	52,000 (2009)	Structure	Online, free
crystaleye	SI from publishers	100,000 (2007)	Structure	Online, free

The above-mentioned databases provide very good coverage of the various aspects of chemistry. Crystal structures and powder patterns of organic and inorganic compounds, metals, and proteins can be searched. Each database is unique in its character but there are overlaps resulting, e.g., from similar scopes of long years of cooperation between producers. Many crystal structures are also available on the servers of publishing houses as supplementary information (SI) contained in scientific publications. These have grown into extensive free data collections during the past few years. On the other hand, databases that are available only with costs still have an advantage in the completeness of coverage. Most of these databases date back many years, which means many years of data collection and deep analysis of publications. Also, the evaluation process is more complex, and the software is more sophisticated and offers more and better functionalities. Of course, the above-mentioned CIF format is a useful tool for combining data from various databases and use them for scientific studies.

Databases of crystal structures (see Table 1 for an overview) are usually regarded as archives of unit cells and atomic coordinates. Coordinates have proved themselves as valuable input for e.g., Rietveld refinements [18, 19] or similar refinements. Besides this, crystallographic databases contain a lot more valuable information, which can enhance high-quality research easily. In order to get an idea of the concepts of a database, this chapter will explain certain details of the Inorganic Crystal Structure Database (ICSD).

The Inorganic Crystal Structure Database (ICSD) [5, 6] contains information on all structures:

- Which have no C–C and no C–H bonds
- Which include at least one of the nonmetallic elements H, He, B–Ne, Si–Ar, As–Kr, Te, I, Xe, At, Rn
- Whose atomic coordinates have been fully determined or were derived from the corresponding structure types

Recently, crystal structure data of metallic and inter-metallic compounds were introduced into ICSD.

Each structure determination reported in the literature forms its own entry. The earliest entry comes from Bragg's paper on sodium chloride in 1913 [7,8]; the most complex is the structure reported in 2006 of the mineral Johnsenite from Grice and Gault which contain 22 different elements [11].

At present (March 2009) the ICSD contains 120,000 entries. The database is maintained by the *Fachinformationszentrum (FIZ) Karlsruhe* with the assistance of several contributors around the world.

The database is provided as a stand-alone version for installation on a local computer (*FindIt*, in cooperation with NIST), as a local intranet version, suitable for small groups of users, and as a WWW version, which is hosted and maintained by FIZ Karlsruhe [1].

2 Contents

2.1 General

The ICSD contains records of all inorganic crystal structures published since 1913, including the very first crystal structure determination by X-rays (NaCl by Bragg and Bragg). Inorganic compounds in the context of ICSD are defined as compounds without C–C and/or C–H bonds and containing at least one of the following non-metallic elements: H (D, T), He, B, C, N, O, F, Ne, Si, P, S, Cl, Ar, As, Se, Br, Kr, Te, I, Xe, At, Rn. To appear in ICSD a crystal structure has to be fully characterized, i.e., it has to contain unit cell data and all atomic coordinates as well as a fully specified composition.

Crystal structure data are analyzed and categorized by experts with the help of sophisticated computer programs. During the recording process additional information, such as Wyckoff sequence, Pearson symbol, molecular formula and weight, calculated density, ANX formula, mineral name, structure types, etc. are generated and added to the crystal structure data.

Currently (ICSD 2009/1) the database contains 120,000 entries. More than 90% of entries in ICSD are represented by compounds with 2–5 elements (Fig. 1). A statistical view of binary metal–element compounds, sorted by groups of the periodic table, can be seen in Fig. 2.

2.2 Description of Data Fields

The entries stored in the ICSD give full structural and bibliographic information including:

- Chemical name and phase designation
- Special name record

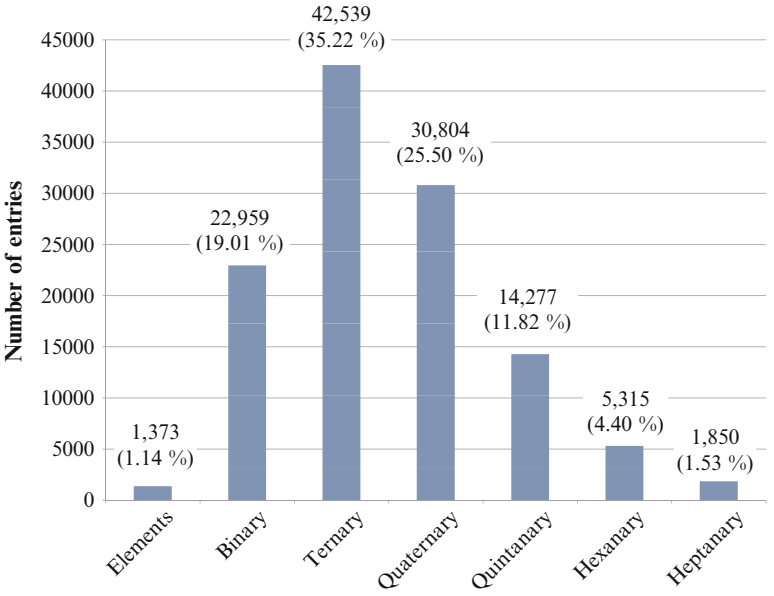

Fig. 1 Overview of multinary compounds. Compounds with 2–5 elements make more than 90% of ICSD

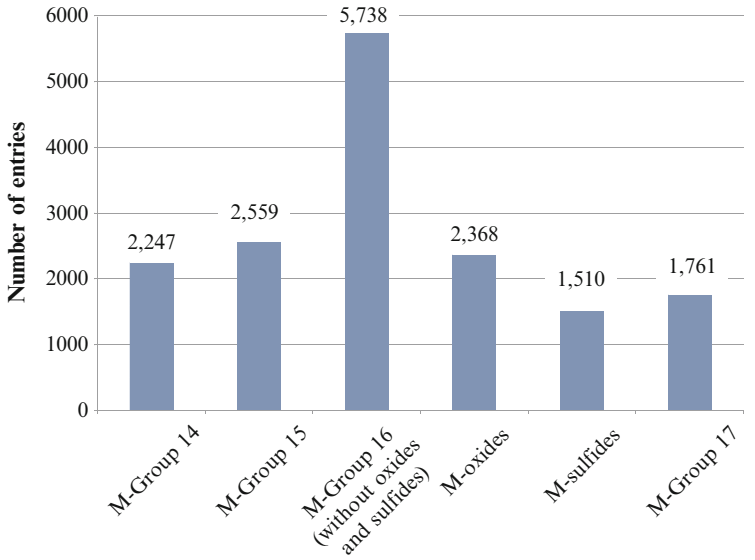

Fig. 2 Overview of compounds with metal (M) and elements of groups 14–17 of the periodic table

- CAS Registry Numbers
- Mineral name and origin
- Chemical formula
- Unit cell dimensions and measured density
- Number of formula units
- Hermann–Mauguin space group symbol
- Atomic coordinates and site occupation
- Oxidation state of the elements
- Thermal parameters
- Temperature and pressure of measurement
- Reliability index
- Method of measurement
- Author, journal, volume and page, year of publication
- Title of paper

Chemical names follow the IUPAC rules; further electronic processing might result in difficulties, though. Modifications to the rules have been made to produce a standardized name suitable for computer treatment. The name should reflect the composition and structure of the compound, if possible.

Phase designations are taken from the paper and standardized. Both the originally published as well as the standardized crystal structures are available.

The *special name record* provides additional identification of the material. It contains the substance numbers from Landolt–Börnstein where the user may find further information and references to the compound in question.

The *mineral name* follows the conventions of Strunz. However, some older names are included and in some cases the names of families of minerals such as feldspars and zeolites are given, even when the entry describes a synthetic member. The origin of the sample has been given if possible.

The *chemical formula* is given in the normal structured form; atoms on identical sites, chemical building units or complexes are given in parentheses. The atomic symbols follow the normal sequence, e.g.,

$$(NH_4)_2(PO_3F)(H_2O)_2 \qquad \text{or} \qquad (Nb_{0.55}Zr_{0.45})(O_{1.1}F_{1.8})$$

For solid solutions or nonstoichiometric compounds the formula of the actual sample investigated is given. Minor constituents which cannot be observed by X-ray diffraction may be omitted from the formula and thus from the table of atomic coordinates. In these cases a second formula record containing the analytical composition is given, e.g.

```
Mineral: Cyprine
Chemical formula: Ca29 Al4 Cu Al8 Si18 O68 (O H)10
2nd formula record: (Ca28.28 Mn0.68) Al4 (Fe0.29 Cu0.71)
   (Al6.36 Mg0.56 Ti0.03 Zn0.97) (Si17.51 Al0.49) O68
   ((O H)8.5 F1.5)
```

Because the number of formula units Z in the unit cell is always given as an integer, it is occasionally necessary to multiply the formula unit by a decimal fraction.

The *unit cell dimensions* are given in Ångström units and degrees. The number of formula units Z in the unit cell is also given in this record.

The *space group* is given by the Hermann–Mauguin symbol. More than 400 different settings of the space groups have been used to report inorganic crystal structures, and to avoid introducing errors by transforming these to a standard setting, the authors' original settings have been retained. To make the space group symbol setting consistent some conventions and additions have been adopted:

- The bar always follows the number.
- The letter Z has been added at the end to indicate that the origin is at the centre of inversion.
- The letter S at the end of the symbol indicates a special origin (e.g., the second setting given in International Tables or a setting indicated by the symmetry operators that follow).
- Monoclinic space groups are always given the full symbol.
- Rhombohedral settings are indicated by the letter R at the end of the symbol.
- The obverse hexagonal setting of rhombohedral space groups is indicated by the letter H and the reverse setting by HR.

Symmetry records are included to give the symmetry operators of the special position when a nonstandard space group setting has been used.

The *atomic coordinates* are preceded by the element symbol (which may include D for deuterium) followed by an atom identifier which is always a simple integer regardless of the identifier used by the author.

This is followed by the oxidation state which is the formal charge of the atom in the most probable ionic formulation. Standard rules for determination of oxidation states are applied. In some cases it may not be possible to assign individual oxidation states and all the atoms of that element will be assigned the same (possibly nonintegral) oxidation number. If an oxidation state for some atom cannot be assigned by any of these methods it is set to zero. In any case the sum of all the oxidation numbers in the unit cell must be zero.

After the oxidation number, the multiplicity and the Wyckoff symbol of the site, coordinates (in decimals) with the standard deviations (as given by the author), and the site occupation factor follow. It always relates to the number of positions and only occurs if it is different from one.

All the elements occupying the same site are listed separately each, with its own site occupation factor (unless this is less than 0.01). Hydrogen (and occasionally other) atoms whose coordinates have not been determined are included in the atom list without coordinates but with an appropriate site occupation factor (which in this case may be greater than 1.0) in order that the formula calculated from the atom list agrees with the structured chemical formula. The number of hydrogen atoms attached to each anion is indicated in those cases where the coordinates are not given (e.g., OH_2, NH_3).

Thermal parameters (atomic displacement factors) are stored in the form given by the authors:

$$\exp(-B\sin^2\theta/\lambda^2),$$

$$\exp(-8\pi^2 U \sin^2\theta/\lambda^2).$$

Temperature factors are often deposited (indicated by *L*) but wherever possible are included in the database. Anharmonic temperature factors are not included but their existence in the original paper is indicated by the remark "AHT."

The *reliability index R* serves as a rough measure of the quality of the structure determination. Authors use different definitions and the lowest is usually included in the database.

Comments on the structure and its determination are included in the *remarks*, the more common of which have been coded into three letter symbols. Unless otherwise stated it is assumed that the structure determination was carried out at room temperature and pressure using single crystal X-ray diffraction. Any other method is indicated by the use of standardized remarks:

A full description of each record is described in the Coding Instructions which can be obtained from the authors.

2.3 Structure Types

Structure types were introduced into ICSD in 2005 [2]. For this purpose new standard remarks (labels) had to be integrated into the database; their names are TYP and STP, and they can be assigned to each entry. Every entry that belongs to a certain TYP label is represented by a member of this group of entries. This arbitrarily chosen member will then serve as the structure prototype and is additionally marked with the STP label. Methods had to be developed to overcome the difficulties of determining structure types automatically and assigning crystal structures to their structure prototype [20].

Two definitions, as mentioned in an IUCr report [10], proved to be suitable and fully sufficient as a theoretical concept for the task, namely isopointal and isoconfigurational structures. Two structures should be described as isopointal, if they have the same spacegroup type or if they belong to two enantiomorphic space group types. They are isopointal as well, if the atomic positions are the same in both structures, i.e., the sequence of occupied Wyckoff positions is identical for both structures after standardization.

For differently standardized crystal structures, the Wyckoff sequence may depend on the selected cell origin. For example, in spinels that crystallize in space group $Ia\bar{3}d$ (two standard settings: origin at $\bar{3}$ or $\bar{4}$), an origin shift by $\frac{1}{2}\,\frac{1}{2}\,\frac{1}{2}$ will change the Wyckoff sequence from *e d a* into *e c b*.

Isoconfigurational structures include a subset of isopointal structures in a way that isoconfigurational structures have to be isopointal and that the crystallographic orbits for all Wyckoff positions have to be similar.

It has to be noted that the second definition is difficult to deal with, since the exact meaning of "similar geometric interrelationships" is not specified. For the introduction of structure types, novel methods that combine different criteria needed to be introduced. According to the definition of Lima-de-Faria, we use an a priori definition of geometric criteria for the distinction of structure families.

3 Acquisition of Information

The ICSD database was created in 1978, with annual updates comprising about 1,000–2,000 structures. During the past decades, the number of publications on inorganic crystal structures has increased steadily, and a more professional approach to database processing was called for. Today, about 7,000 new structures are recorded annually for ICSD, and the existing structures are regularly revised, corrected, and updated (see Fig. 3). This was made possible, among other things, by software enhancements that enabled today's largely automatic data acquisition and processing procedure. The number of modified structures has been raised extensively during the past four years. This included also the deletion of many duplicates, which had been incorporated into ICSD in the 1980s as a result of distributed update procedures among the partners. These were problems of a more or less technical nature; they could be solved completely, and the database quality has increased dramatically during the past few years. In the early days of database development, crystal structures were recorded manually and checked only intellectually. Today, many records are incorporated electronically as CIF records, and verifications and checks are made automatically (see Sect. 4).

A look at Fig. 4 shows that about 40% of all structures contained in ICSD are taken from the 12 most relevant journals. In addition to these, ICSD also contains structures that derive from more than 1,000 journals published worldwide. One can easily imagine that it is much more difficult to find these 60% of structures and incorporate them in the database. On the other hand, it is this 60% that decides the completeness of coverage and thus the quality of the ICSD database. Information on the existence of relevant structures is obtained in various ways, e.g., by regular scanning of expert journals and also by regular searching of bibliographic databases

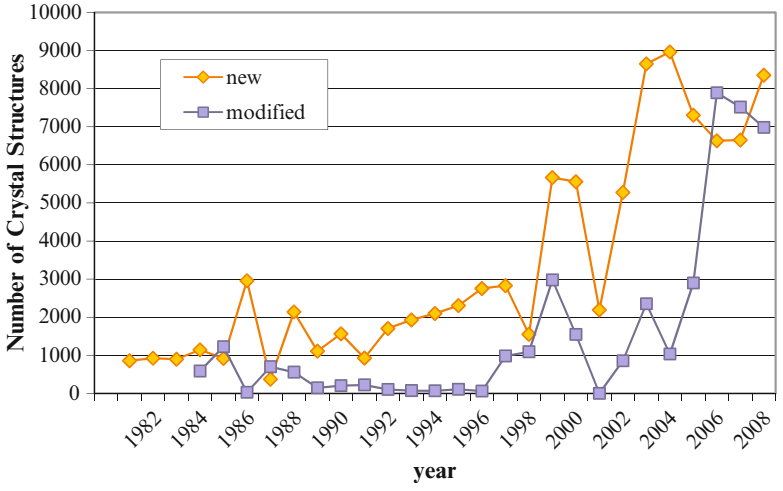

Fig. 3 Input development in ICSD over the years 1981–2008

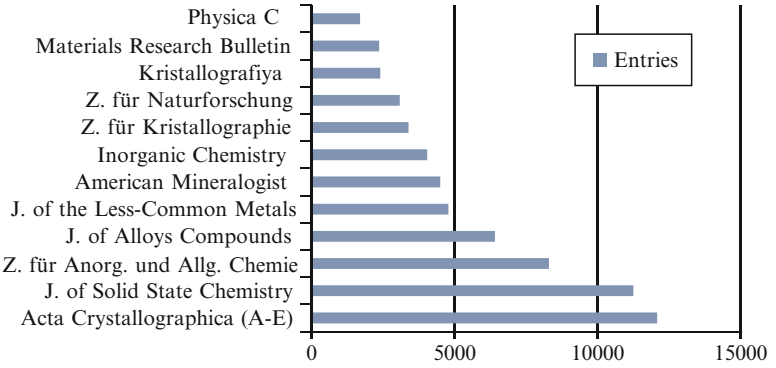

Fig. 4 The 12 most productive journals in ICSD

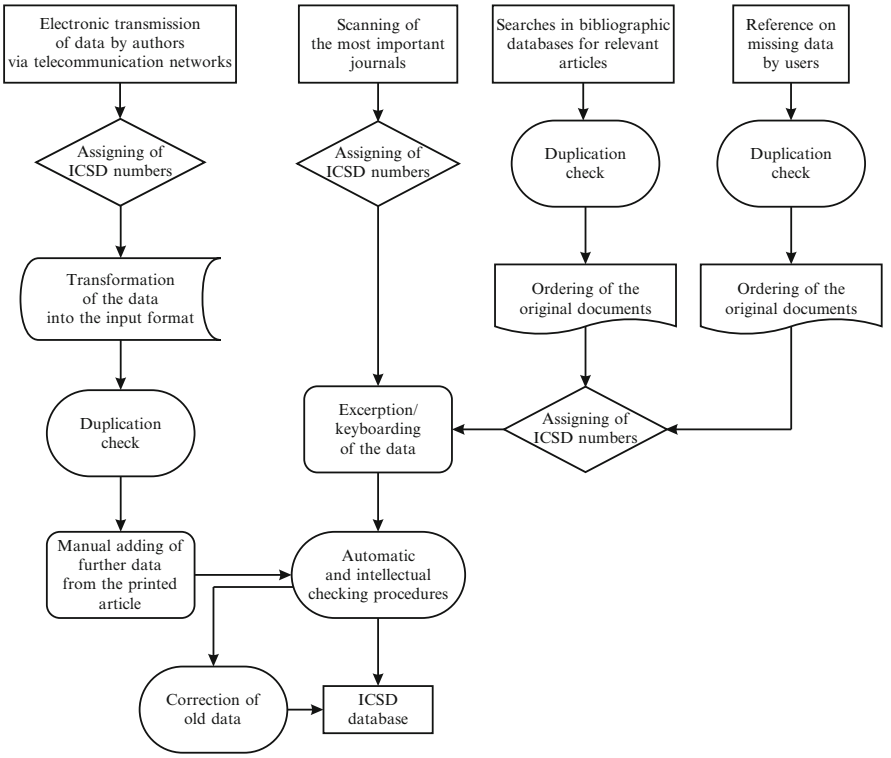

Fig. 5 Procedure of data input for ICSD. After [3]

and Internet publications. FIZ Karlsruhe also receives information from queries to our crystal structure depot and by scientists informing the authors of structures that should be incorporated.

As soon as FIZ Karlsruhe receives information on publications that may contain relevant structures, FIZ experts start to check the contents of the original publication (see Fig. 5 for a schematic overview). If the information is up to the standard of

the ICSD quality criteria, the publication will be processed for incorporation in the database. If the information is incomplete, the publishers' sites will be analyzed for any further structural information that may be obtained. If necessary, the authors will be contacted as well.

There are two ways of data acquisition: Parts of the structures are still processed manually; on the other hand, increasing numbers of crystal structures are imported as CIF files. Electronic import is, of course, much more efficient and less prone to errors.

Most of the CIF files can be generated from the crystal structure depot that is available at FIZ Karlsruhe. FIZ Karlsruhe started this depot of crystal structure data more than 25 years ago, first in printed form and, since the early nineties, also increasingly in electronic form. Today, only electronic data are stored in the depot. The crystal structure depot enables authors to store their extensive data and to refer to the stored record(s) in their publications. Access to the crystal structure depot is free on request for scientists (see also http://www.fiz-karlsruhe.de/request_for_deposited_data.html).

4 Revision and Evaluation

To maintain a high quality standard of ICSD, it is absolutely necessary to validate the compiled data prior to publication in the database. This step is essential also in the interest of the database user; it is the central element of database production.

As mentioned earlier, many checks today are made automatically. This comprises formal checks, plausibility checks, and checks of certain constraints resulting from mathematical and physical laws. The sections below will present some examples.

4.1 Formal Checks

Formal checks are independent of the content; they comprise, e.g., duplication checks of new records, missing field entries, correctness of bibliographic data, standardization of authors' names, syntax, etc.

4.2 Verification of Contents

These checks refer to the correctness of the chemical and crystallographic information; they are of decisive importance for the scientist user. Among others, they comprise the following checks:

- Plausibility and validity of cell
- Matching of cell and space group

- Validity of oxidation state
- Multiplicity
- Site occupation
- Electroneutrality
- Molecular formula
- Plausibility of isotropic/anisotropic temperature factors
- Interatomic distances.

These verifications are regular background processes during data processing; their results are evaluated intellectually by experts. Validation rules are, e.g.,:

- Plausibility and validity of cell:
  The following rules apply to the validation for the plausibility of cell:

$$\alpha + \beta + \gamma \leq 360^\circ,$$

$$\alpha \leq \beta + \gamma; \qquad \beta \leq \alpha + \gamma; \qquad \gamma \leq \alpha + \beta$$

  and for the validity of cell

$$0.5\,\mathrm{g\,cm^{-3}} < \rho < 20\,\mathrm{g\,cm^{-3}}.$$

- Validity of sum formula:
  The sum formula is calculated from atomic parameters, site occupation, and site multiplicity, and compared with the corresponding formula given by the author. There are cases in which uncertainties remain even after automatic verification and intellectual checking. It is a basic principle of ICSD that original data should be retained as far as possible unless the database expert finds errors of a fundamental and obvious nature. In such cases, the original data will be corrected, and the record will contain a comment written by the database expert to show where changes were made. If no corrections are made, the so called test flags are introduced instead. Examples are:

  – Difference between the calculated formula and the formula in the formula-field is tolerable.
  – Deviation of the charge sum from zero is tolerable.
  – Calculated density is unusual but tolerable.
  – Displacement factors are those given in the paper but are implausible or wrong.
  – A site occupation is implausible but agrees with the paper.
  – Lattice parameters are unusual but agree with the paper.
  – Coordinates are those given in the paper but are probably wrong.
  – Reported coordinates contain an error. Values corrected

In addition to the automatic checks described above, there are some aspects of ICSD that are verified and, if necessary, corrected only intellectually. This includes, e.g.:

- Relevant decision
- Chemical nomenclature

- Mineralogical nomenclature
- General, free remarks added to the structures

5 Database Design

5.1 Access to ISCD Data

The ISCD database is offered via various channels:

- WWW Version
  The ICSD Web version is available since June 2009. It offers both, the flexibility of a browser based interface and the functionality of a graphical user interface. The new Web Version was developed in 2008/2009 by FIZ Karlsruhe in order to meet both the increased requirements of the user community (user friendly Interface, easy to navigate, up-to-date retrieval interface and visualization, flexible export of data), and the requirements of modern software development. For a schematic view of the database design see Fig. 6.

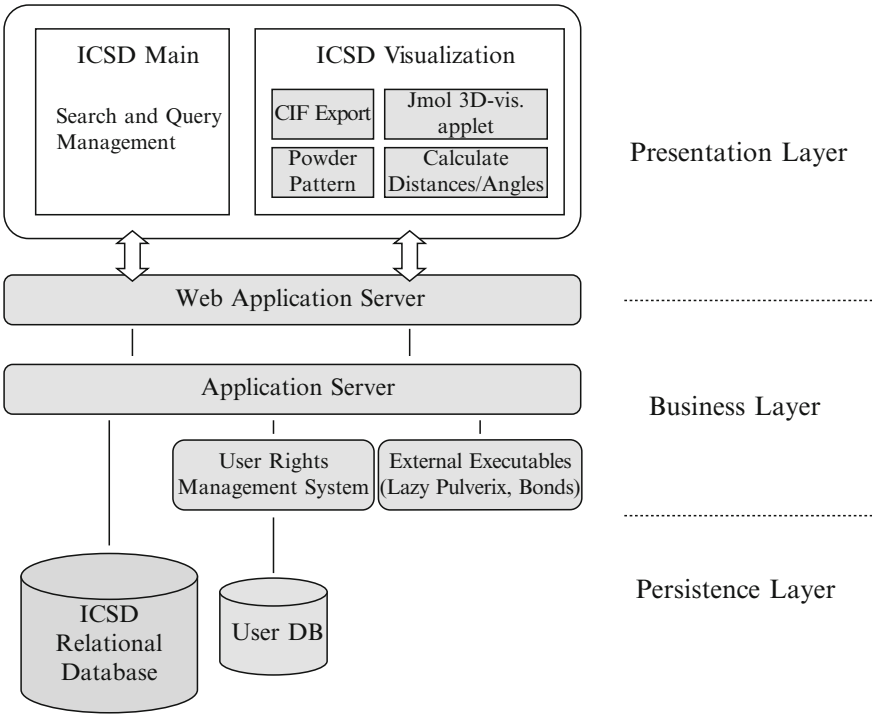

Fig. 6 Modules of the new ICSD Web version

- *STN Version*. The ICSD STN Version is fully integrated in STN, the Scientific and Technical Information Network. The graphical user interfaces for this version of ICSD are STN Express and/or STN on the Web. The visualization components of the ICSD Web version are also available for the STN version.
- *Intranet Version*. The ICSD Intranet version is available since June 2009. It is based on the previous ICSD Web version developed by A. Hewat & FIZ Karlsruhe. It can be installed locally in corporate or campus network. It will be replaced as soon as possible by the new ICSD WEB version components.
- *CD-ROM Version*. The PC-based Search and Retrieval interface was developed by NIST.

The Web system is based on a Java Enterprise edition (JEE) multitier architecture. Each layer encapsulates a specific set of functionalities. The session management located in the Web application server tracks and authorizes the user's activities utilizing the services of a User Rights Management System. From there it is connected asynchronously to an Enterprise Resource Planning (ERP) system. The business process components run on a multithreaded application server. Some scientific executables are bound externally to fulfill the requirements of the user communities.

This new Web version of ISCD will be used in the following chapters for screenshots and examples.

6 Retrieval Examples

6.1 General Strategies

ICSD Web provides various retrieval options, from classic bibliographic searches through simple or combined searches to complex search strategies. The user is offered clearly structured search masks in which bibliographic data, chemical compositions, selected symmetry properties, measuring conditions, etc. can be entered, combined logically, and stored. The complex search strategies comprise all published data that are contained in the database, as well as data derived or calculated from these (e.g., structure types, Pearson symbol, ANX formula, interatomic distances, etc.). All records are recorded twice on principle, i.e., both in the published form and in a standardized form; they can be compared directly with each other or with the original data in a synoptic view. Searching via the reduced cell is possible as well. Depending on the query, the search result can be used for ordering the document in full text, e.g., in the case of reviews. Further, structured data can be exported in CIF format [12] for use as input data for all common visualization programs (Mercury [15], Diamond [17], Jmol [16]). The PDF number provides a reference to the powder pattern database of ICDD (PDF2/PDF4+) [13].

The following sections contain information for choosing the optimum search strategy, which should be both highly selective and comprehensive. Often, there are

several ways to arrive at the desired search result. In many cases, the Basic Search will be sufficient; this strategy is the most user friendly and permits intuitive use.

If this mode is not precise enough, there is also the Advanced Search mode, which offers several options. For example, the user can search all settings of the desired space group; he can search groups in the periodic table of the elements instead of single elements, or he can open a browser window to get a general idea of the available data. This may be useful, e.g., if the correct spelling of a name or structure type is not known and if wild card searching would produce too many undesired results.

If a cell or space group is used as a start, the user must choose the data source he would like to access, i.e., standardized data, published original data, or the reduced cell data (Niggli).

Search in standardized data is recommended whenever a comparison of several structures in a synoptic view is intended. It is also possible here to compare the published data with the standardized data. Unusual cell settings in a publication may not be of particular relevance; on the other hand, some settings may illustrate developments in phase changes or contain information on a group theory context. A look at the published data is particularly useful in these cases.

Searching in reduced cell data has the advantage that, e.g., in the case of pyroxenes, the search is independent of the number of formula units per cell in a record (which is unknown to begin with). For example, via the reduced cell one finds both $FeSiO_3$ ($Z = 16$) and $Fe_2Si_2O_6$ ($Z = 8$), which would not be possible when searching via the formula with stoichiometric coefficients. Another advantage of reduced cell searching is the fact that the search is independent of the chemical composition and may provide unexpected results (e.g., isomorphic compounds, mixed crystal series). On the other hand, when searching in the reduced cell one needs to choose the optimal tolerance and to know how to interpret the results.

Many users prefer searches for isotypic compounds in ICSD Web in order to obtain initial values for a new structural search.

For classification of simple inorganic structures, the ANX type was used in the "Strukturberichte." The letters A–M represent the cations, N–R the elements with oxidation number zero, and S–Z the anions. Further details like the treatment of hydrogen, partially occupied sites, elements with different oxidation numbers, etc. are contained in the scientific reference manual of the ICSD Web database.

The ANX type "AX" comprises structure types that are as different as NaCl, NiAs, ZnS (wurtzite type or zinc blende type), which can be derived from different packing patterns and also have different types of gaps (octahedral gaps, tetrahedral gaps) and different coordination polyhedra. The characteristic "AX" therefore is not sufficient for accurate classification. ICSDWeb offers the additional option of searching for the Pearson symbol, which is described in detail in the scientific manual. To give an example: A search for *mP*4 (monoclinic, primitive cell with four atoms) alone is too unspecific to yield satisfactory results. More detailed searching, e.g., with space group number = 11 and the number of elements permitted = 2, will find only records of the compound NiTi on the one hand and records of the high-pressure phases of AgCl, AgBr, and AgI on the other hand.

If one applies the ANX concept to these two groups, AgCl belongs to "AX" (cation–anion) while NiTi belongs to "NO" (elements, oxidation number = 0). The Wyckoff sequence is 2*e* in both cases, i.e., it cannot be used for further differentiation. This is where ICSD Web offers the possibility of viewing both structures side by side in a synoptic view. ICSD Web also gives the user the option of further differentiation of the two groups by entering a structure type according to Allmann. As described in Sect. 2.3, Allmann developed a structure type concept based on 11 criteria, which enables further detailing of the structure description and a combination of these structures in the ICSD database. The concept is described in detail in the publication by Allmann and Hinek [2]; it regularly comprises the space group, Wyckoff sequence, and Pearson symbol as characteristics. For further differentiation, the ANX type, $(c/a)_{\min}$, $(c/a)_{\max}$, $\beta_{\min}$, $\beta_{\max}$, space group number as well as the occurrence or absence of certain elements are introduced. New records are automatically selected by these criteria and are assigned to structure types accordingly. In the above case, the high-pressure phases of AgCl, AgBr, and AgI are of the AgCl (*mP*4) type while NiTi belongs to the NiTi type.

On the other hand, if we reverse the search by looking for the NiTi structure in ICSD Web, we will find that NiTi is the only representative of this type so far in the ICSD Web database.

Another feature for visualization of crystal structures is the so-called Movie Display, where two structures are displayed alternatingly in a single window. This is an interesting feature, e.g. for illustrating slight differences between structures. One example of this is the alternating display of the superconductor $YBa_2Cu_3O_{(7-x)}$ in the space groups *Pmmm* and $P4/mmm$ as described in [9].

6.2 The Black Tar Mystery

A spectacular example of a problem solved with ICSD is found in a publication by Kaduk [14].

In a BP refinery, a pump valve was clogged by an unknown black, viscous mass. It was feared that sulfuric acid might be released and react with the aluminium casing, although it was unclear if this would happen at the onset or in the course of the process.

The powder diagram (Fig. 7) indicated the presence of $Al_4H_2(SO_4)_7(H_2O)_{24}$, $FeSO_4(H_2O)$, and $Al_2(SO_4)_3(H_2O)_{17}$; the tar-like substance could be removed by washing with acetone. The crystal structure of $Al_4H_2(SO_4)_7(H_2O)_{24}$ was unknown at the time. For refining by Rietveld methods, it was necessary first to identify the structure.

The search for the ANX formula (in this case A4B7X52; Al, S: cations, O: anion) resulted in a single hit, i.e., the compound $Cr_4H_2(SO_4)_7(H_2O)_{24}$. The two compounds were found to be isostructural. Refining by Rietveld methods with Al in the positions of the Cr atoms provided an excellent result.

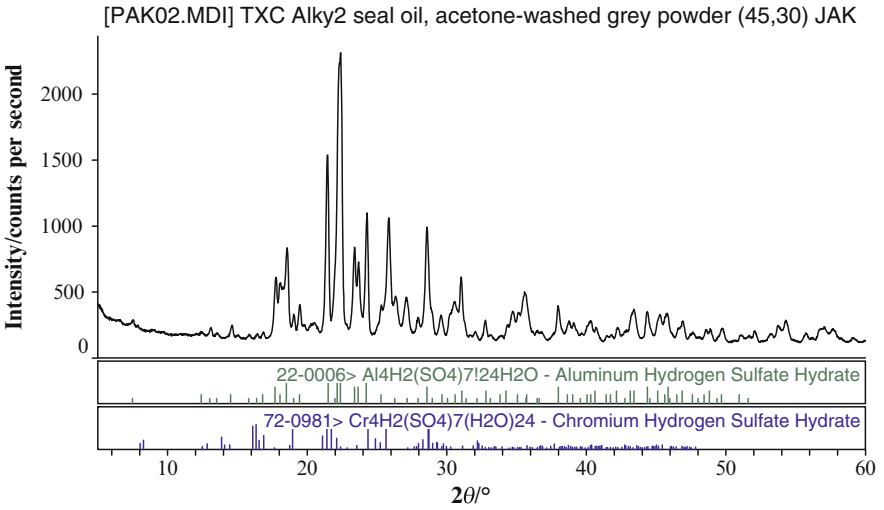

Fig. 7 Powder pattern of the unknown black substance

It was found that sulfuric acid was indeed released, which reacted with parts of the pump to form metastable $Al_4H_2(SO_4)_7(H_2O)_{24}$. Corrosion occurred in the course of the process and not (only) at the onset. The problem could be solved and the production process was optimized. Since then, the crystal structure of $Al_4H_2(SO_4)_7(H_2O)_{24}$ has been given a record in the ICSD database (Collection Code 77310).

6.3 Example: Searching for Ice I_c

The following example shows how to search and find crystal structures of Ice I_c.

As mentioned in Sect. 6.1, the first approach is to use the *Basic Search* function of ICSD Web. Entering H_2O in the appropriate field and limiting the number of elements to 2 gives the first set of results (Fig. 8).

Unfortunately, among the results are entries about crystal structures of H_2O_2 as well as $H_2O_2{\cdot}H_2O$, which are not interesting in this special search case (Fig. 9).

A more suitable approach to solve the problem is the *Advanced Search* function (Fig. 10). Since we know that we are looking for ice there is probably a structure type for it in ICSD. Opening the list of structure types and entering "H_2O" we find five results, among these is also the desired ice I_c.

The results can be seen in Fig. 11. The result set includes one crystal structure of D_2O (entry 5), although we did not explicitly include any chemical element in the search.

For a comprehensive overview it is possible to display the set of crystal structures or the simulated powder patterns in a synoptic view (Fig. 12).

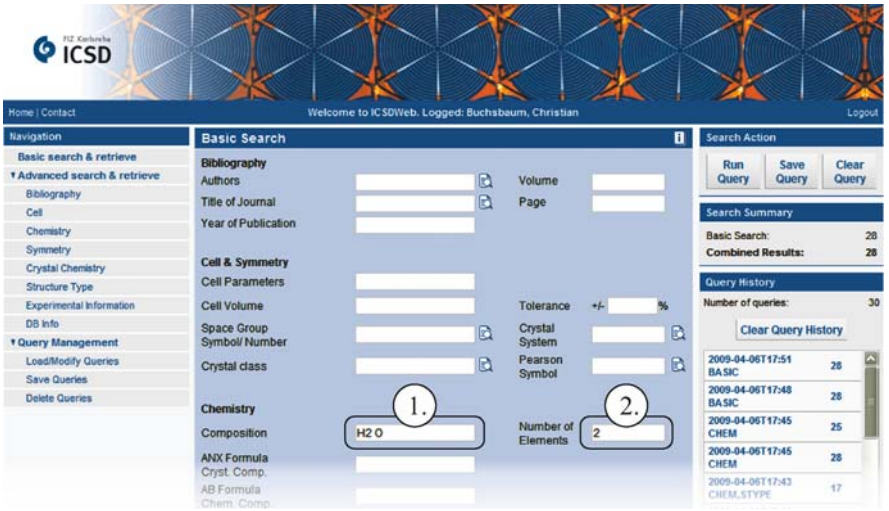

Fig. 8 Basic search. Enter H_2O into *Composition* field (**1**) and 2 into the *Number of Elements* field (**2**)

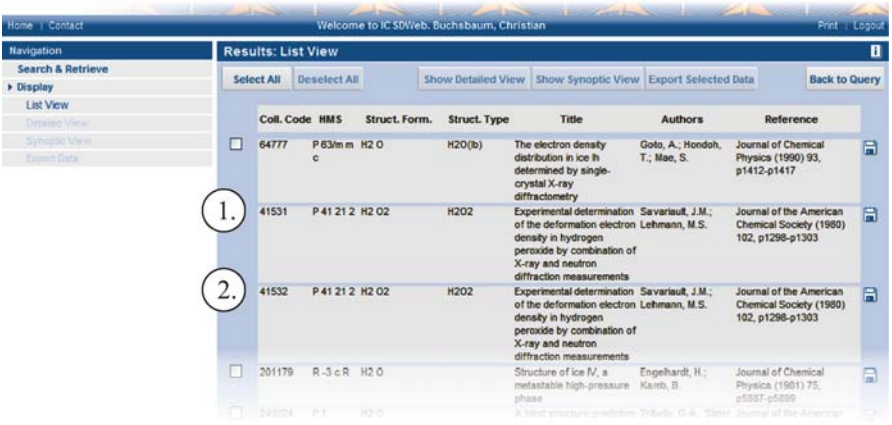

Fig. 9 Basic search results. Entries of H_2O_2 (**1** and **2**) are not desired in this search

7 Outlook

As mentioned before, ICSD has undergone significant changes during the past decades. On the one hand, there is an exponential increase in the structures added annually. Bibliometric analyzes have shown that the number of structures contained in ICSD has doubled every 10–11 years (Fig. 13) [4]. On the other hand, database experts are working continually on filling in the gaps in the older data (Fig. 14). Of

Fig. 10 Advanced search. Possible candidates for structure types beginning with "H_2O" (**1**). We enter the appropriate structure type into the corresponding search field (**2**)

Fig. 11 Advanced search results. All entries with structure type of ice I_c are included

course, this dramatic increase in ICSD structures made it necessary to have a more efficient search functionality for quick and selective retrieval of the desired results.

In 2001, ICSD changed over to a relational database system, which was a decisive step ahead. The contents were now presented in 25 tables and about 200 database fields. The new Web interface permits searches for an even larger number of criteria; as a result, ICSD today offers more than 35 tables and far more than 300 fields. The most important innovations were the introduction of structure types and the calculation of standardized data which resulted in a new search option.

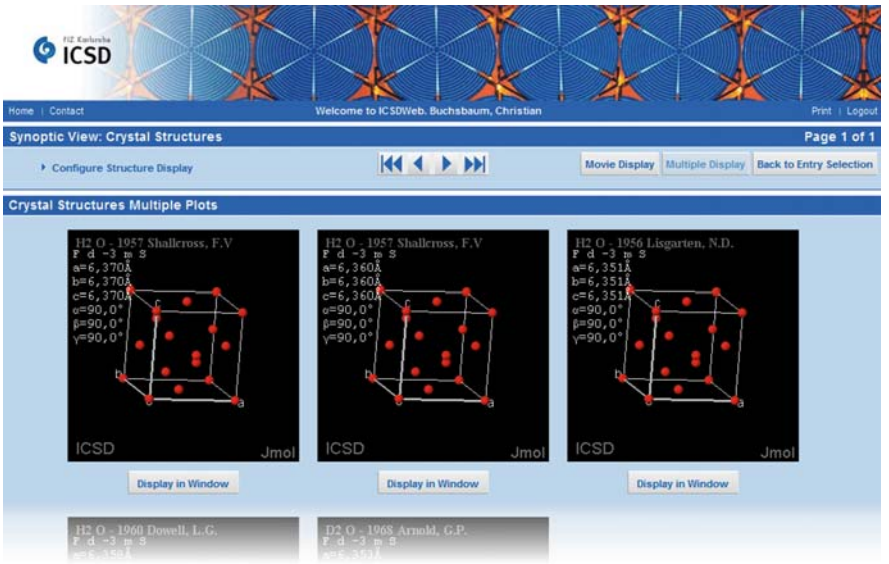

Fig. 12 The synoptic view allows a quick comparison of crystal structure

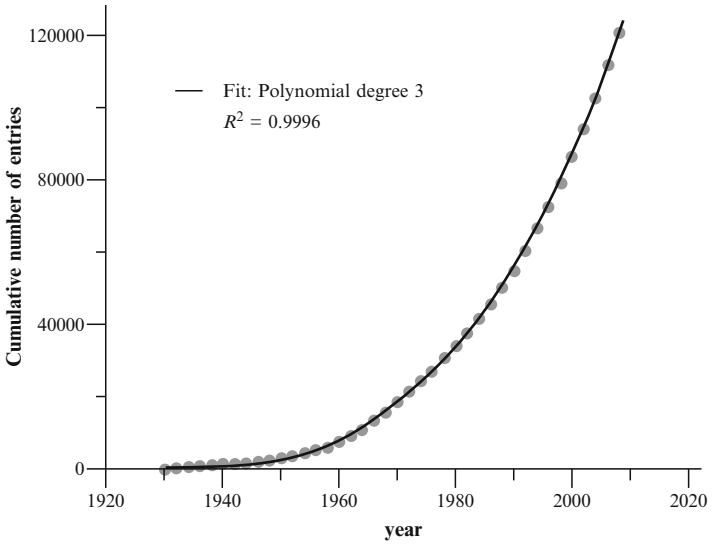

Fig. 13 Cumulative number of database records added per publication year. Fit with polynomial of degree 3; parameters: $a_0 = -16,30,88,372$, $a_1 = 25,38,524.873$, $a_2 = -1,317.0895$, $a_3 = 0.227785$, $R^2 = 0.9996$

Further continuous enhancement of the content of ICSD is envisaged also for the future. This includes on the one hand continuous updating for completeness and on the other hand the introduction of new data fields and contents as well as constant

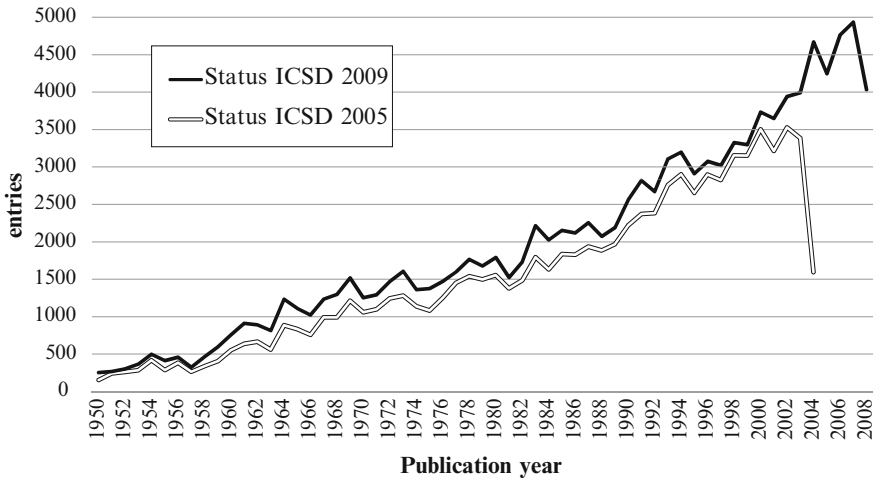

Fig. 14 Gaps in old data are continuously filled

improvement of the database functionality. The focus is on high database quality as this is the only way to enable extensive and complex data analyzes and therefore data mining. The search strategies and examples presented here may give an idea of the many possibilities offered by a crystal structure database to solve problems of materials science. Search functionalities will develop further in the next years. New technologies as Web 2.0 and semantic web will have their impact also on databases like the ICSD. We can expect that the future databases will offer more possibilities for interaction, e.g. easy integration of data in existing data collections or ways to annotate or comment data. Standard exchange formats like CIF even today make it possible to switch to other databases and consult them with regard to the problem at hand. It may be possible to develop networks of different databases to link different types of content. This may lead to advanced solutions in fields like data mining.

References

1. ICSD is available at FIZ Karlsruhe at http://www.fiz-karlsruhe.de/icsd.html or http://icsd.fiz-karlsruhe.de (2009)
2. Allmann R, Hinek R (2007) The introduction of structure types into the inorganic crystal structure database icsd. Acta Crystallogr Sect A 63:412–417
3. Behrens H (1996) Data import and validation in the inorganic crystal structure database. J Res Natl Inst Stand Technol 101:365–373
4. Behrens H, Luksch P (2006) A bibliometric study in crystallography. Acta Crystallogr Sect B 62:993–1001
5. Belsky A, Hellenbrandt M, Karen VL, Luksch P (2002) New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. Acta Crystallogr Sect B 58:364–369

6. Bergerhoff G, Brown ID (1987) Crystallographic databases. International Union of Crystallograhy, Chester, pp 77–95
7. Bragg WH, Bragg WL (1913) The reflection of X-rays by crystals. Proc Roy Soc London Ser A, Containing Papers of a Mathematical and Physical Character 88:428–438
8. Bragg WL (1913) The structure of some crystals as indicated by their diffraction of X-rays. Proc Roy Soc London Ser A, Containing Papers of a Mathematical and Physical Character 89:248–277
9. Cava RJ, Hewat AW, Hewat EA, Batlogg B, Mareziod M, Rabe KM, Krajewskia JJ, Peck Jr WF, Rupp Jr LW (1990) Structural anomalies, oxygen ordering and superconductivity in oxygen deficient $BA_2YCU_3O_x$. Physica C 165:419–433
10. de Faria JL, Hellner E, Liebau F, Makovicky E, Parthé E (1990) Nomenclature of inorganic structure types. Report of the International Union of Crystallography Commission on Crystallographic Nomenclature Subcommittee on the Nomenclature of Inorganic Structure Types. Acta Crystallogr Sect A 46:1–11
11. Grice JD, Gault RA (2006) Johnsenite-(CE): A new member of the eudialyte group from Mont Saint-Hilaire, Quebec, Canada. Can Mineral 44:105–115
12. Hall S, McMahon B (eds) (2005) Definition and exchange of crystallographic data. International Tables for Crystallography, vol G. Springer, Dordrecht
13. Kabekkodu SN, Faber J, Fawcett T (2002) New Powder Diffraction File (PDF-4) in relational database format: advantages and data-mining capabilities. Acta Crystallogr Sect B 58: 333–337
14. Kaduk JA (2002) Use of the inorganic structure database as a problem solving tool. Acta Crystallogr Sect B 58:370–379
15. Macrae CF, Bruno IJ, Chisholm JA, Edgington PR, McCabe P, Pidcock E, Rodriguez-Monge L, Taylor R, van de Streek J, Wood PA (2008) Mercury csd 2.0 – new features for the visualization and investigation of crystal structures. J Appl Crystallogr 41:466–470
16. McMahon B, Hanson RM (2008) A toolkit for publishing enhanced figures. J Appl Crystallogr 41:811–814
17. Pennington WT (1999) Diamond – visual crystal structure information system. J Appl Crystallogr 32:1028–1029
18. Rietveld HM (1967) Line profiles of neutron powder-diffraction peaks for structure refinement. Acta Crystallogr 22:151–152
19. Rietveld HM (1969) A profile refinement method for nuclear and magnetic structures. J Appl Crystallogr 2:65–71
20. Villars P, Cenzual K (eds) (2004) Structure types. Part 1: space groups (230) $Ia\bar{3}d$ – (219) $F\bar{4}3c$, Landolt–Börnstein – Group III Condensed Matter, vol. 43A1. Springer, Berlin