Journal of the Ceramic Society of Japan, Supplement 112-1, PacRim5 Special Issue, 112 [5] S1434-S1438 (2004)

Review

PDF-4/Organics—A New Relational Database Format for Powder Diffraction, Data Mining and Materials Characterization

J. FABER

The International Centre for Diffraction Data (ICDD), 12 Campus Boulevard, Newtown Square, PA 19073

The Powder Diffraction File (PDF), published by the International Centre for Diffraction Data (ICDD), is widely used for phase and polymorph identification by x-ray powder diffraction methods. In collaboration with the Cambridge Crystallographic Data Center (CCDC), calculated x-ray powder patterns have been integrated into the new PDF-4/Organics. Nearly 150,000 patterns are present in this database. The PDF-4/Organics is distributed with integrated search and retrieval software that provides Boolean search logic on subfiles (e.g., Drug Activity, Pharmaceuticals, Excipients, Forensics, Pigments, Polymers, and Explosives). All entries are available as fully digitized diffraction traces and 2D structures are displayed. The relational database (RDB) format provides for exceptionally powerful data mining capability.

[Received August 6, 2003; Accepted December 16, 2003]

Key-words: *Powder diffraction, Search/index, Hanawalt method, Fink method, PDF (Powder Diffraction File), Goodness of match*

Introduction

The International Centre for Diffraction Data (ICDD) has been the primary reference for X-ray Powder Diffraction (XRPD) data for over 50 years. The primary information in the PDF is the collection of d-I data pairs, where the d-spacing (d) is determined from the Bragg angle of diffraction, and the peak intensity (I) is obtained experimentally under the best possible conditions for a phase-pure material. These data provide a data mining[1),2)] capability as well as "fingerprint" of the compound because the d-spacings are fixed by the geometry of the crystal and the intensities are dependent on the contents of the unit cell. Hence, d-I data may be used for identification of unknown materials by matching d-I data in the PDF with the d-I pairs obtained from the unknown specimen. Identification is the most common use of the PDF, but the presence of considerable supporting information for each entry in the PDF allows further characterization of the specimen. Examination of the crystal data, Miller indices, intensity values, scale factors, physical property data and the comprehensive literature reference data provide extraordinarily useful information concerning the specimen under study. For pharmaceutical R & D, XRPD and the PDF have been used for example as an indispensable tool in phase identification (both qualitative and quantitative), in the identification of unknowns, evolution of polymorphism and solvate structures, and crystallinity determinations. The impact of the PDF as a reference pattern database has been used in patent disclosures and as such has immediate impact for pharmaceutical R & D.

The PDF has exhibited recent dramatic growth in entry population over the past 5 years. Historically, the PDF-2 has been a flat file database that contains powder patterns of inorganic compounds. However, in late 1998, the ICDD reached an agreement with the Cambridge Crystallographic Data Center (CCDC) that allows for the calculation of x-ray powder patterns from the structural information in the Cambridge Structural Database (CSD). The resultant explosion in the organic population from 25,000 to 150,000 entries in 2003 is a direct result of this agreement. Also in 1998, the ICDD began the development of a successor relational database (RDB) design for the PDF. The ICDD RDB format is designated PDF-4. All inorganic entries and just the experimental organic entries will be found in the PDF-4/Full File 2003. The calculated patterns from the CSD and the experimental organic entries will be found in a separate database, designated the PDF-4/Organics 2003. Properties of these two new databases are illustrated in **Table 1**. A completely new database, the PDF-4/Organics 2003 contains mainly organic and organo-metallic entries. In this paper we

Table 1. Selected Entry Counts of the PDF-4 Databases

	PDF-4/ Full File 2003	PDF-4/ Organics 2003	PDF-4/ Organics 2004*
Organic Compounds	25,609	146,256	~207,000
Inorganic Compounds	133,370	1,695	~1,000
Calculated patterns form the CSD		122,816	~182,000
Drug Activity Index		4,292	
Drug Activity		4508	
Pharmaceuticals	2,039	1,171	
Forensic Materials	3,767	2,008	
Pigments	342	277	
I/Ic	73,087	~125,000	
Total Entries	157,048	146,256	~208,000

*Please Note that PDF-4/Organics 2004 will be Released in November, 2003

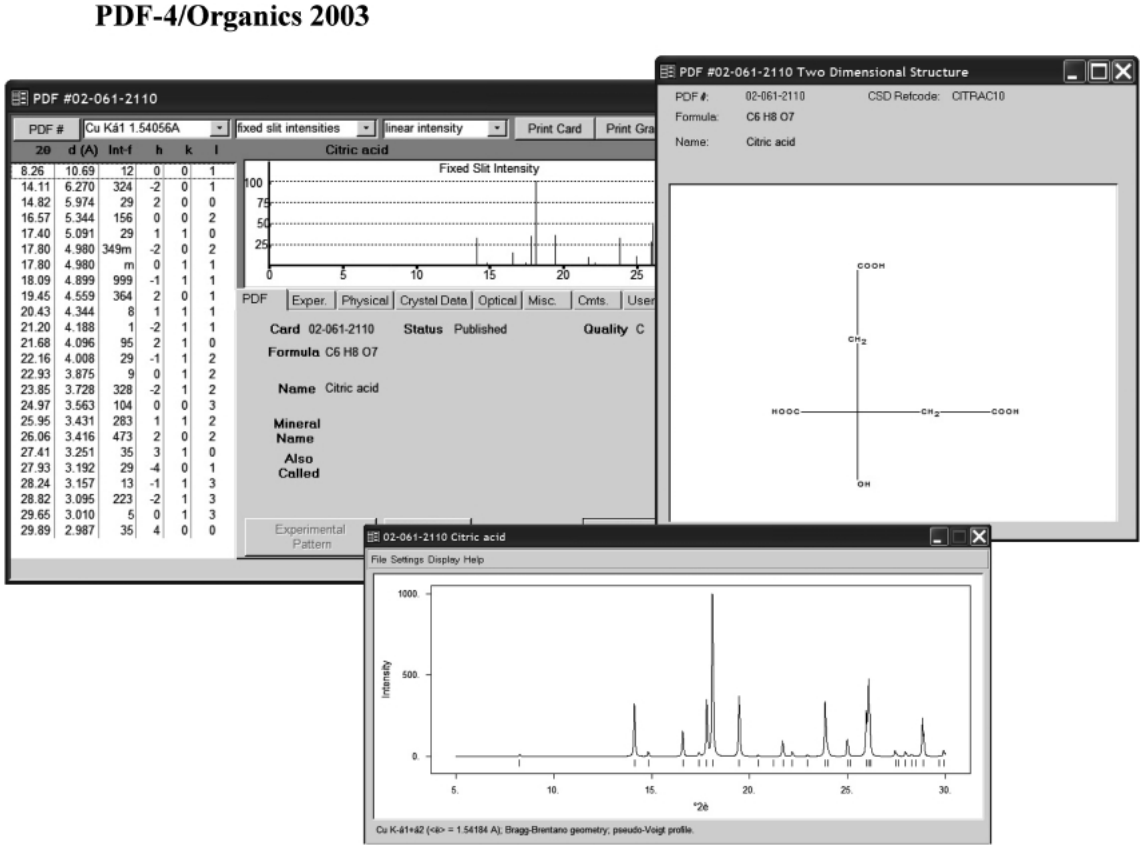

Fig. 1. Example data from the PDF–4/Organics 2003 for Citric Acid. Note the 2D structure display and the on-the-fly digitized pattern.

give a brief overview of the PDF–4/Organics 2003. The focus of this paper is to present applications that demonstrate the power of a PDF–4/Organics 2003. We shall demonstrate this analytic power by illustrating results obtained from phase identification and search-indexing, using Hanawalt and Fink methods. A preliminary report has been given.[3)]

The PDF–4 database contains interplanar spacings (d) and relative intensities (I). However, other useful data such as synthesis, physical properties and crystallographic data are also stored in the database. With this new format, we will provide a broader range of analyses, for example, improved quantitative analyses, full pattern display, bibliographic cross referencing, etc. The PDF–4 uses relational database technology that provides pliable access to the database to carry out data mining studies and enhances the pursuit of conventional materials characterization using diffraction techniques (see Faber et al.[1)]). In addition to better access to some of the RDB fields, users can also build search criteria by combining individual search conditions using Boolean operators. The availability of logical operators for combining the search condition is very useful in arriving at the desired information from the database

The CSD database is being used to calculate entries in the PDF. Thus, to derive d-spacing and peak intensity data requires the synthesis of full diffraction patterns, i.e., we use the structural data in the CSD database and then add instrumental resolution information. In addition to the peak intensities, $|F(hkl)|^2$, the square of the structure factor magnitudes will also be calculated. Thus, calculated powder patterns are obtained for all CSD entries in the PDF–4/Organics 2003 RDB. For example, we can calculate (on-the-fly) a selected profile function to describe paracrystallinity, or particle size and/or strain effects. PDF data for an ideally random crystal distribution in the absence of preferred orientation may also be obtained. In the future, preferred orientation models will be developed. The main focus is to provide tools that can be used for materials design.

The CSD contains bitmap control integers that can be used to project out specific categories of entries in the CSD. Of particular interest for pharmaceuticals is the drug activity flag. There are approximately 250,000 entries in the CSD and of these, approximately 6000 have the drug activity flag set. The PDF–4/Organics 2003 contains calculated patterns for 4, 292 of these entries. The process of calculating PDF data is an ongoing task; we will calculate powder patterns for all entries in the CSD when the ICDD editorial review has been successfully completed.

Search-Indexing using the PDF–4/organics RDB: Hanawalt and Fink Search/Match Procedures

Most of the commercial software packages for qualitative phase identification have been designed to implement fully automatic search/match sequences.[4)–16)] On the other hand, traditional methods of search/match (based on *d*-spacings, intensities and chemistry) are mainly manual techniques using paper-based search/indices. Manual techniques were first discussed by Hanawalt[17),18)] and these persist for a variety of reasons. Some of these reasons have been explored by Jenkins,[19)] Jenkins and Hubbard,[20)] and Schreiner and Jenkins.[21)]

Traditional methods for search/match in powder diffraction are based upon combinations of d-spacing, intensities and chemistry. **Table 2** lists the types of search indices currently

Table 2. Types of Data Search Indexes

Index	**Entry Method**	**Search Parameters**
Alphabetic	Chemistry, chem. formula fragments	Permuted chemical formula fragments
Hanawalt	d,I pairs, sorted in decreasing intensity	3 strongest lines
Fink	d,I pairs, sorted in decreasing d-space	8 longest lines (longest of the strongest)

being used.

The Hanawalt search method has been implemented for many years at the ICDD.[22] The method involves sorting the patterns in the PDF according to the d-spacing value of the strongest line. This list is broken into discrete d-space intervals defined as Hanawalt groups. A small overlap in d-intervals is employed to reduce the probability of missing powder pattern entries due to uncertainty in the d-space accuracy. Each Hanawalt group is sorted in order of decreasing d-spacing of the second most intense diffraction line. Subsequent lines are listed in order of decreasing intensity. The analysis rests on the three most intense lines, but the eight most intense lines are listed. Considerable redundancy exists in this method because patterns appear twice for the (1,2) and (2,1) pairs when $I_2/I_1 > 0.75$ and $I_3/I_1 > 0.75$. Patterns appear three times (1,2), (2,1), and (3,1) when $I_3/I_1 > 0.75$ and $I_4/I_1 < 0.75$. The rationale for multiple entries is to minimize problems of preferred orientation, especially when these affect the three strongest lines. In summary, the Hanawalt method relies on the d-spaces for the three strongest lines; further confirmation of a search hit is taken from matches on the eight strongest lines.

The Fink method was designed by Bigelow and Smith[23] and was named after William Fink. This method creates an index based on the eight strongest d-spaces in the experimental pattern, but these are ordered in decreasing d-spacing. In short, the Fink method considers the 8 longest of the strongest diffraction lines. Creating permutations of these is not practicable for large databases as the corresponding paper manuals become enormous.

As we shall see, permuting the Hanawalt three strongest lines or the Fink eight longest lines is straightforward using computer methods. For both the Hanawalt and Fink methods, the problem is that the associated paper manuals have grown cumbersome and difficult to use. In addition, the integration of elemental composition and other important ancillary information is not easily accomplished with these methods. Filtering criteria need to be "remembered" while carrying out the search/indexing process. We have developed a "plug-in" for the PDF-4 databases that implements the Hanawalt/Fink strategy, including chemistry, subfile and quality-mark filters.

Search-Index Plug-In

The basic idea of the plug-in is to provide d,I pairs as input to the program. The d-spaces are in Angstroms and the I's are peak intensity values from the x-ray powder diffraction experiment. As additional input, P is a phase parameter associated with each d,I pair; if P = 1, the peak is included in the analyses, otherwise the peak is ignored. As we shall see, this is quite helpful in multiphase problems. Also, contaminant peaks can be easily excluded in the analysis by adjusting P. The principal input is from an ASCII file that contains the d,I pairs. However, the plug-in can also accommodate 2-theta pairs if the first ASCII record also contains the wavelength.

$$\Delta d = d \cdot \cot\theta \cdot \Delta\theta \qquad (1)$$

In the case of the Hanawalt method, the search window defines the Hanawalt group. The match window is defined by Eq. (1). The match window defines the PDF entry lines that match with the experimental data. Up to eight strongest experimental lines appear as d1–d8 just above the match list box in **Fig. 2**. The best match is assigned to PDF # 000161157. For this match, 4 or more of 8 lines fall in the match window.

Match window hits are selected and an algorithm is used to obtain the hit that best matches the experimental data. This best match is obtained by calculating the GOM, defined by

$$GOM = 1000 \cdot \Sigma[(1 - |\Delta d|)/SW]^2, \qquad (2)$$

where the sum is taken over the experimental lines and their corresponding match lines in the selected PDF entry, and SW is the search window defined from Eq. (1). Notice that a perfect match between experiment and the PDF for 8 lines would yield GOM = 8000. Also, GOM values < 1000 are not significant since this corresponds to the identification of only a Hanawalt Group. GOM < 2000 means that no single reasonable entry has been identified within the Hanawalt Group. For the analysis presented in Fig. 2, GOM = 4,390. The Intensity Scale Factor seen in Fig. 2 is calculated based on all matched lines in the analysis.

The GOM = 4247 (**Fig. 3**) indicates that nearly 4 of the 8 strongest lines have been matched. In this case, all except one weak line (d = 14.092, I = 2) have been account for. Notice also in Figure 3 that relative scaling factors are indicated for each detected phase. However, these are only approximate since RIRs (reference intensity ratios) are not used in this analysis. Obtaining semi-quantitative analyses using reference intensity ratios is under development. One of the powerful features of the plug-in is that the data grid on the right side of Figs. 2–3 can be filled by any selection from the hit list. The data grid lists all experimental lines and all lines from the selected entry in the hit list. Thus, a detailed comparison between the match hit and the input data can be seen at a glance.

Enhanced search indexing is realized by allowing for permutation of the order of the experimental lines, thus removing the need for redundant (and complex) indexing of reference data from the database. In the Hanawalt search method we have implemented rotation operators to permute the order of the three strongest lines. This is indicated by the rotation counter: "Rotation: 1 of 3," seen in Figs. 2 and 3. The rotation operator is also available for the Fink method (not shown), however, in this case since we are ordering 8 strongest of the longest d-space lines, the range is 1–8.

Summary

We have illustrated several examples of the complementary use of XRPD techniques coupled with a new organic database, the PDF-4/Organics 2003. An example, citric acid, was used to show some of the powerful features available in this new PDF-4. In particular, calculated on-the-fly powder patterns were generated and 2D structures are available. The search-indexing example, Alka-Seltzer Plus, was used to show search-indexing results using Hanawalt and Fink methods for phase

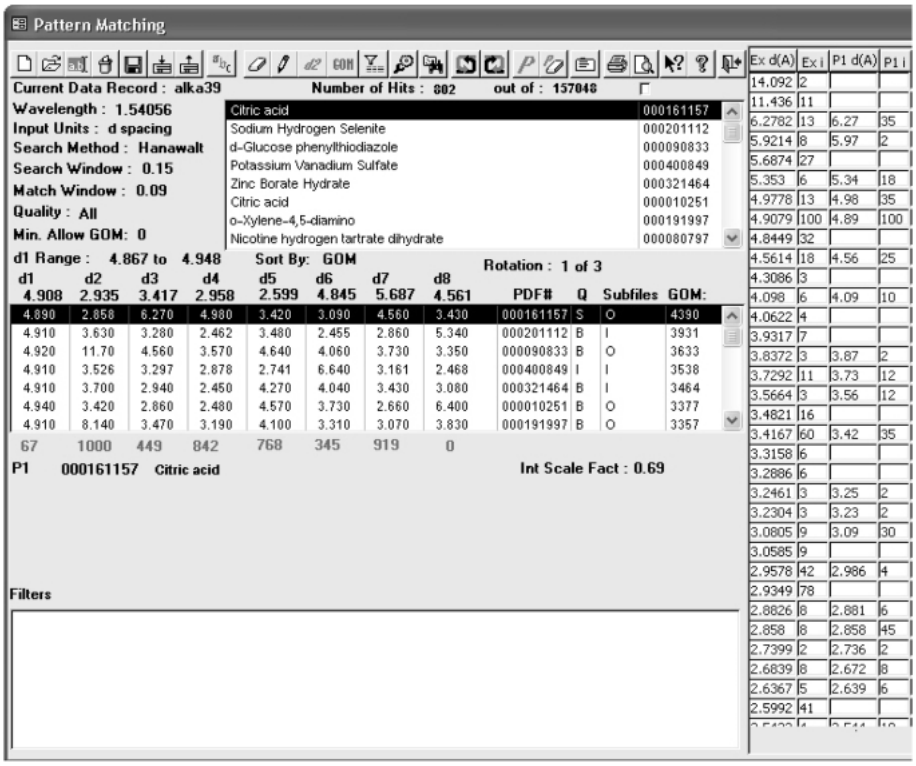

Fig. 2. Hanawalt method applied to an over-the-counter medication. The drug is Alka-Seltzer Plus, ingested after dissolution in water. The tablets were ground and standard XRD experiments were performed. A peak-listing program was used to define d-spaces and peak intensities for all Bragg lines detected.

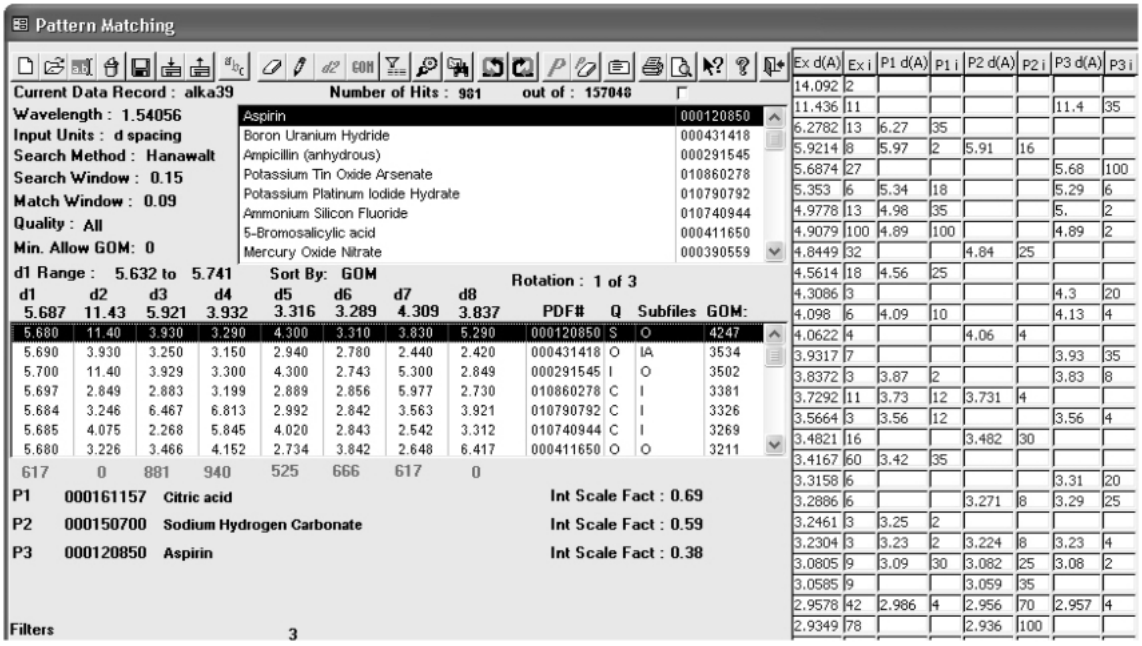

Fig. 3. Two additional phases were identified in this analysis: a second and third search were performed. The GOM = 5140 for the second phase (not shown) and GOM = 4247 for the third phase.

ID. In this case, we were able to identify the three most abundant components in the tablet. We feel that the importance of the PDF-4/Organics RDB will grow as its use becomes commonplace in the pharmaceutical community.

Acknowledgments The author wishes to thank C. A. Weth and Jerome Bridge (student intern from West Chester University) at ICDD for valuable help in implementing the "plug-in" ideas presented in this paper.

References

1) Faber, J., Kabekkodu, S. N. and Jenkins, R. (2001), *International Conference on Materials for Advanced Technologies*, Singapore, unpublished; Kabekkodu, S. N., Faber, J. and Fawcett, T., *Acta Cryst*., Vol. B58, pp. 333–337 (2002).
2) Faber, J. and Fawcett, T., *Acta Cryst*., Vol. B58, pp. 325–332 (2002).
3) Faber, J., Weth, C. A. and Jenkins, R. (2001) *Materials Science Forum*, Vol. 378–381, pp. 106–111 (2001).
4) Johnson, G. G., Jr. and Vand, V., *Ind. Eng. Chem*., Vol. 59, pp. 19–19 (1965).
5) Nichols, M. C., Lawrence Livermore Lab. Report UCRL–70078 (1966).
6) Frevel, L. K., Adams, C. E. and Ruhberg, L. R., *J. Appl. Crystallogr*., Vol. 9, pp. 300–305 (1976).
7) Marquardt, R. G., *J. Appl. Crystallogr*., Vol. 12, pp. 629–634 (1979).
8) Snyder, R. L., *Advances in X-Ray Analysis*, Vol. 24, pp. 83–90 (1980).
9) Jobst, B. A. and Goebel, H. E., *ibid*., Vol. 25, pp. 273–282 (1981).
10) Parrish, W., Ayers, G. L. and Huang, T. C., *ibid*., Vol. 25, pp. 221–229 (1981).
11) Jenkins, R., Hahm, Y., Pearlman, S. and Schreiner, W. N., *ibid*., Vol. 23, pp. 279–285 (1979).
12) Goehner, R. P. and Garbauskas, M. F., *X-Ray Spectrom*., Vol. 13, pp. 172–179 (1984).
13) Toby, B. H., *Powder Diffraction*, Vol. 5, pp. 2–7 (1990).
14) Caussin, P., Nusinovici, J. and Beard, D. W., *Advances in X-ray Analysis*, Vol. 31, pp. 423–430 (1987); *ibid*., Vol. 32, pp. 531–538 (1988).
15) Nusinovici, J. and Winter, M. J., *ibid*., Vol. 37, pp. 59–66 (1993).
16) Hanawalt, J. D. and Rinn, H. W., *Ind. Eng. Chem. Anal*., Vol. 8, pp. 244–244 (1936); Hanawalt, J. D., *Advances in X-ray Analysis*, Vol. 20, pp. 63–73 (1976).
17) Hanawalt, J. D., Cryst. in North America, Apparatus and Methods, American Crystallographic Association, Chapter 2, pp. 215–219 (1983).
18) Jenkins, R., *Advances in X-ray Analysis*, Vol. 20, pp. 125–137 (1976).
19) Jenkins, and Hubbard, C. R., *ibid*., Vol. 22, pp. 133–142 (1978).
20) "Automatically Correcting for Specimen Displacement Error During XRD Search/Match Identification," *ibid*., Vol. 25, pp. 231–236 (1981).
21) Jenkins, R., *ibid*., Vol. 37, pp. 117–121 (1994).
22) "Powder Diffraction File Hanawalt Search Manual for Inorganic Phases," published each year, ICDD. The current volume is Release 2003; see also Jenkins, R. and Snyder, R. L., "Introduction to X-Ray Powder Diffractometry," Volume 138 in Chemical Analysis, John Wiley & Sons, pp. 335–339 (1996).
23) Bigelow W. and Smith, J. V., *ASTM Spec Tech Publ. STP*, Vol. 372, pp. 54–89 (1965).