Coupling the PDF-4/Organics Relational Database to SIeve+, a Hanawalt and Fink Search Indexing Plug-In

J. Faber, J. Blanton and C. A. Weth
International Centre for Diffraction Data (ICDD)
Newtown Square, PA 19073, USA

Introduction

The International Centre for Diffraction Data (ICDD) has been the reference source 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 data mining [1-2] capability as well as "fingerprint" identification 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 locating matching d-I data in the PDF with the d-I pairs obtained from an unknown specimen.

The PDF has exhibited recent dramatic growth in entry population over the past 5 years and the new PDF-4/Organics results from a collaboration with the Cambridge Crystallographic Data Centre (CCDC). An illustration of organic entry growth for the PDF-4/Organics RDB is given in Table 1.

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. In addition, we wish to demonstrate the use of the PDF-4 relational database as a filter for search-indexing. In this case, we will use a new plug-in for PDF-4 databases, SIeve+, to illustrate search results using Hanawalt and Fink methods.

Category	PDF-4/ Full File 2003	PDF-4/ Organic 2003	PDF-4/ Organic 2004
Organic Compounds	**25,609**	**147,201**	**217,077**
Inorganic Compounds	**133,370**		**3,048**
Both Organic and Inorganic	**1,931**	**1,776**	**1,931**
Only Inorganic	**131,439**		**1,117**
Calc. patterns from CSD		**122,816**	**191,468**
Drug Activity		**4,508**	**6,343**
Pharmaceuticals	**2,039**	**1,192**	**2,352**
Excipients	**801**	**184**	**1,114**
Forensic Materials	**3,767**	**2,015**	**2,113**
Pigments	**342**	**284**	**296**
I/Ic	**73,087**	**125,342**	**195,316**
Total Entries	**157,048**	**147,201**	**218,194**

Table 1. *Selected entry counts for PDF-4 databases. Please note that because entries can be listed in both the inorganic and organic collection, the total number of distinct entries is obtained from the organic and only inorganic rows in the table.*

PDF-4/Organics

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. [2]). 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 search criteria is very useful for retrieval of relevant information from the database.

Examples of searchable fields are illustrated in Table 2.

Property	Entry Population
Drug Activity	**6,343**
Pharmaceuticals	**2,352**
Excipients	**1,114**
Merck	**1,554**
Pigments	**296**
Color	**214,501**
Density	**201,084**
Melting Point	**64,733**
Organic Functional Group Designations	**41,552**
Empirical Formula	**All**
# of Elements	**All**
Periodic Table	**All**
# Searchable Fields	**>30**

Table 2. *Selected searchable fields in the PDF-4/Organics 2004 that can be used as pre-filters for Search-Indexing, using the RDB plug-in, SIeve+.*

Search-Indexing using the Plug-in, SIeve+ in the PDF-4/Organics 2004

Manual techniques were first discussed by Hanawalt and these persist for a variety of reasons. The search-indexing plug-in, SIeve+, discussed here follows a traditional path to act as a replacement for paper search manuals published by the ICDD. An advantage to this approach is that Hanawalt [3,4] and Fink methods [5] can be followed in great detail as search-indexing proceeds. Educational benefits accrue following this approach. A more detailed report concerning PCSIWIN, the predecessor of SIeve+ has been given [6].

The Hanawalt search method has been implemented for many years at the ICDD. The method involves sorting the patterns in the PDF according to the d-spacing value of the strongest line. This list is partitioned 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. All reference lines must then be compared to the unknown. The Fink method was designed as 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. To help account for preferred orientation effects, the lines need to be permuted to remove limiting intensity criteria. This permutation process is easily computerized (in contrast to paper manual searches).

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, quality-mark filters and any of the searchable properties in the database (see Table 2).

SIeve+

The basic idea of the plug-in is to provide d,I pairs as input to the program. The d-spacings are in Å and the I's are peak intensity values from the X-ray powder diffraction experiment. The principal input is from an ASCII file that contains the d,I pairs. However, the plug-in can also accommodate 2θ,I pairs if the first record also contains the wavelength. The uncertainty in d, Δd, can be obtained by taking the derivative of Bragg's Law:

$$\Delta d = d \cdot \cot\theta \cdot \Delta\theta . \qquad \text{Eq. 1}$$

In the case of the Hanawalt method, the search window, $SW = \Delta(2\theta_{SW})$ defines the Hanawalt group; the match window, $MW = \Delta(2\theta_{SW})$ defines the criteria used to judge matches for individual lines for each member of the group. The angular

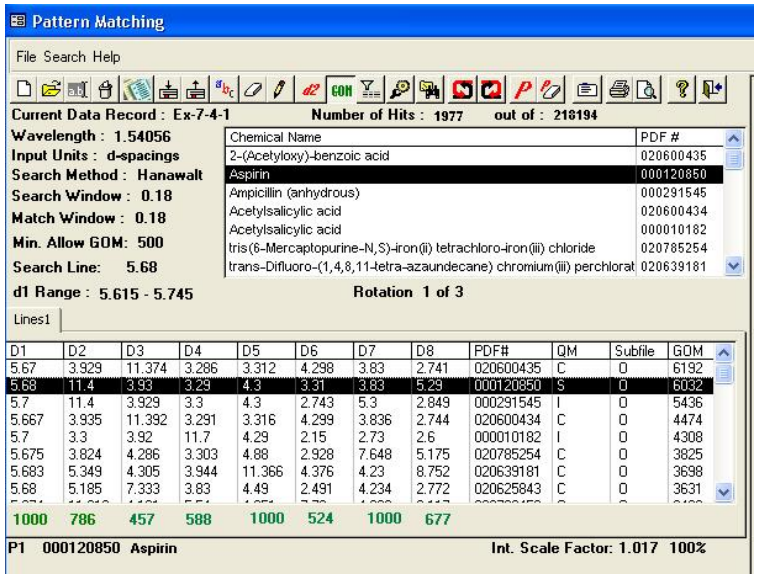

Figure 1. *Hanawalt method applied to an over-the-counter medication. The drug is aspirin. The tablets were ground and standard XRD experiments were performed. A peak-listing program was used to define d-spacings and peak intensities for all Bragg lines detected. A wide search window and match window were employed for this analysis.*

dependence of the search window and match window are defined by Eq. 1. The SIeve+ results illustrated in Figures 1 and 2 show that aspirin is the unknown phase.

In Figure 1, wide search and match windows were chosen and the result is a Hanawalt group that contains 1,977 entries. Each of these reference patterns are then compared line by line against the "unknown".

In contrast, Figure 2 shows the results of coupling the PDF-4/Organics RDB as a pre-filter to the search indexing functions in SIeve+. Only 6 reference patterns need be compared to the unknown phase. It important to recognize that the Pharmaceutical subfile search in the PDF-4 was executed and the search results were conveniently transmitted to SIeve+.

Summary

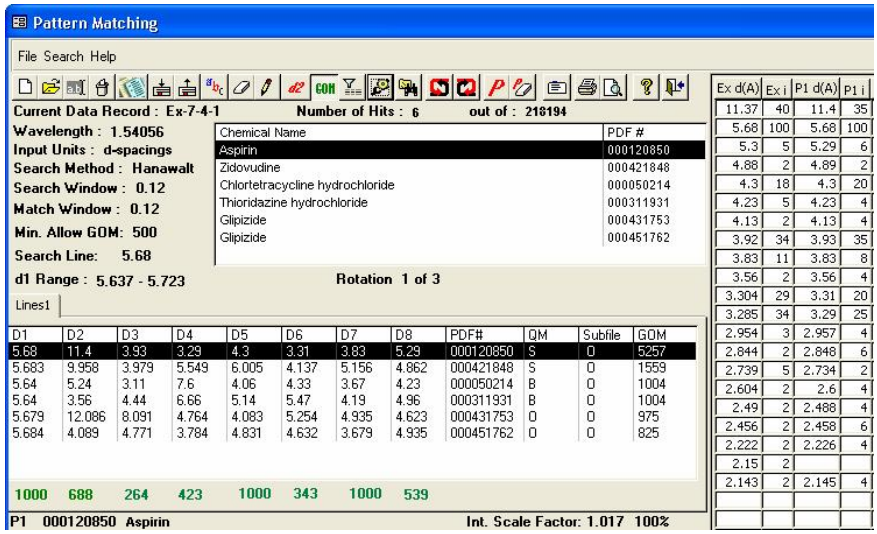

Figure 2. *Hanawalt method applied after pre-filtering using the PDF-4/Organics 2004 RDB. Notice that the Hanawalt group has been reduced to just 6 candidate reference materials.*

We have outlined how the PDF-4/Organics can be used as a pre-filter to search-indexing using classical Hanawalt and Fink methods. The advantage of pre-filtering centers on the ability to use *any* combination of the searchable properties in the RDB. These can be applied to search-indexing using SIeve+. There were no filtering protocols available with older search-index paper manuals. In contrast for the plug-in, SIeve+, the effort required to examine large numbers of entries in the Hanawalt or Fink group is minimized and relevant solutions to search-indexing are illustrated. This streamlines the search-indexing process.

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, 333-337 (2002).
[2] Faber, J. and Fawcett, T., Acta Cryst., Vol. B58, 325-332 (2002).
[3] Hanawalt, J. D. and Rinn, H. W., Ind. Eng. Chem. Anal., 8, 244 (1936); Hanawalt, J. D., Advances in X-Ray Analysis, 20, 63-73 (1976).
[4] Hanawalt, J. D., Cryst. in North America, Apparatus and Methods, American Crystallographic Association, Chapter 2, 1983, 215-219.
[5] Bigelow, W. and Smith, J.V., ASTM Spec. Tech. Publ. STP 372, 54-89 (1965).
[6] Faber, J., Weth, C. A. and Jenkins, R. (2001), Materials Science Forum, Vol. 378-381, 106-111 (2001); Faber, J. and Weth, C. A., J. Powder Diffraction, Vol. 19, 26-30 (2004).