J. Faber, C.A. Weth & J. Bridge, "*A Plug-in Program to Perform Hanawalt or Fink Search-Indexing Using Organics Entries in the ICDD PDF-4/Organics 2003 Database*", Adv. X-Ray Analysis, v. 47 (2004) pp166-173

A PLUG-IN PROGRAM TO PERFORM HANAWALT OR FINK SEARCH-INDEXING USING ORGANICS ENTRIES IN THE ICDD PDF-4/ORGANICS 2003 DATABASE

J. Faber, C. A. Weth and J. Bridge*

International Centre for Diffraction Data (ICDD)
Newtown Square, PA 19073, USA
**West Chester University, Department of Computer Science,*
West Chester, PA 19380

ABSTRACT

In an attempt to fill a gap between fully automatic search/match programs and purely manual methods based on paper products, a relational database plug-in has been developed that functions as a PC-based Search/Index program for extracting information from PDF-4 powder diffraction databases. The plug-in provides an adjustable search window and match window to account for experimental errors. Both Hanawalt [1,2] and Fink [3] search methods are incorporated. In this paper, we report search-indexing results obtained with the new PDF-4 plug-in applied to a new relational database, the PDF-4/Organics 2003. This database has 24,385 experimental entries and 122,816 calculated patterns derived from the Cambridge Crystallographic Database (CSD). We introduce a Goodness of Match (GOM) parameter to describe the relative agreement between the experimental input data and selected reference patterns from the PDF-4/Organics 2003. The relevance of the GOM is illustrated in several example problems. Multiphase samples can be treated on a phase-by-phase basis.

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 [4-5] 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 locating 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. A completely new relation database (RDB) was used to house the new PDF-4/Organics 2003. The principal classes of database compounds are organic and organo-metallic. The properties of new PDF-4 databases are illustrated in Table 1.

We could anticipate that some search-indexing problems may arise using the PDF-4/Organics database:

- Only organic entries are present in the PDF-4/Organics 2003 database. However, note that 1,117 inorganic compounds are present in the PDF-4/Organics 2004. Inorganic excipients are particularly relevant for pharmaceutical R&D.
- There could be larger uncertainties in the lattice parameters since single crystal experiments are often not optimized for high-resolution d-spacing determinations. The focus is on integrated intensities. Only 623 calculated pattern entries from the CSD indicate that cell constants were obtained using powder diffraction methods.
- Organic entries are often done at low temperature. Comparison between low-temperature reference data and room temperature powder diffraction data does not account for thermal expansion in the reference pattern.
- Organic powder diffraction patterns often contain substantial preferred orientation effects. However, as we shall see, we have implemented rotation operators that permute the strongest lines in the pattern (in both Hanawalt and Fink analyses), which has the effect of taking preferred orientation into account. Severe preferred orientation effects cannot be overcome since this would completely distort these strong line/long line methods. We will discuss this issue in more detail.

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 for PDF-4/Full File 2002 (a predominantly inorganic database) has been given [6].

	PDF-4/ Full File 2003	PDF-4/ Organics 2003	PDF-4/* Organics 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
Calculated patterns from CSD		122,816	191,468
Drug Activity Index		4,508	6,343
Pharmaceuticals	2,039	1,192	2,039
Excipients	801	184	1114
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 of the PDF-4 databases. Please note that PDF-4/Organics 2004 will be released in November, 2003. 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 2003

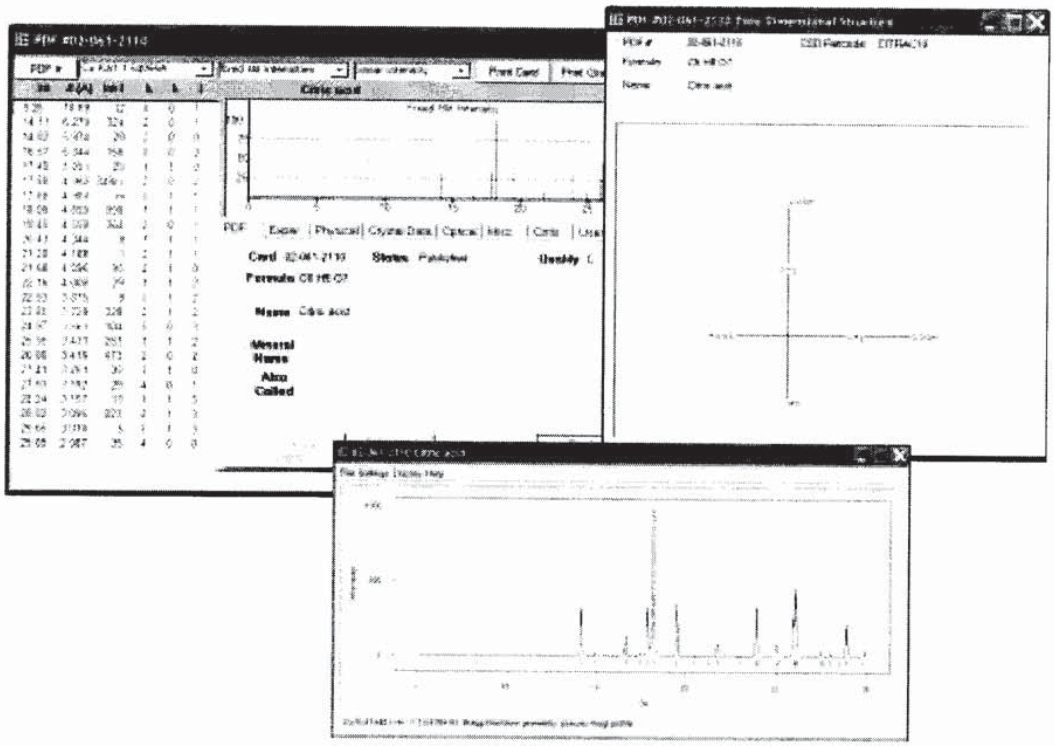

Figure 1. Example data from the PDF-4/Organics 2003 for Citric Acid. Note the 2D structure display and the on-the-fly digitized pattern.

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. [5]). 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 8000 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 [6-17]. 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 and these persist for a variety of reasons. The search-indexing plug-in 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 and Fink methods can be followed in great detail as search-indexing proceeds. The educational benefit of this approach is also realized.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.