Data Mining the Powder Diffraction File, Present and Future Capabilities

T. G. Fawcett, S. N. Kabekkodu, C. A. Weth, J. R. Blanton, J. Faber,
International Centre for Diffraction Data

Abstract
The objective of data mining is to sort through large amounts of data in order to find a particular result that the user is searching for. As the size of the database being searched grows the probability of finding the result increases. However the path traveled to find the result may get more difficult. There are many challenges in deciding which turn to take, and information to keep, in the search for a particular result.

The Powder Diffraction File (PDF™) has grown dramatically in the past five years due to a series of strategic collaborations with global database organizations. There are now 157,048 entries in the PDF-2 and PDF-4/Full File Release 2003 databases and 218,194 entries in PDF-4/Organics Release 2004. Each entry contains chemical and empirical formula, structural classifications, bibliographic, crystallographic and physical data as well a diffraction pattern that can be digitally displayed or presented as an indexed listing of d-spacing and intensity pairs. Therefore each entry can contains hundreds to thousands of pieces of information. The total compilation of the PDF has more than 650,000 literature reference citations from over 150,000 authors of x-ray analyses compiled since the earliest published studies. The task facing many users of the Powder Diffraction File™ now becomes the efficient use and searching of this large amount of data to find exactly the information that they are seeking.

In this paper we will describe the many paths that can be taken to explore the perovskite family of materials. In the last several years new database tools have been developed to aid researchers in exploring the database. The database itself has been reorganized in a relational database format which consists of numerous tables that can be easily cross referenced and queried. A new display program, DDView has been developed that allows all data fields to be ordered by the user and sorted by their personal preference. Finally, the utility program SIeve, automates and evaluates data relative to the classical searches and indexes such as Fink, Hanawalt and Long 8, can automatically import various files containing d, I pairs and is seamlessly interfaced to DDView. The use of multiple search paths results in the proper identification and classification of many materials that weren't previously identified as perovskites.

Keyword Search: Data Mining, Powder Diffraction File, Perovskites

Results
Perovskites are a rich class of important materials well known by mineralogists, ceramists and chemists for their unusual chemical and electronic properties. It is well recognized [1, 2] that perovskites have distinct structural characteristic that provide insights into their various physical properties. Because of this fact scientists have frequently tried to collect structural information on this important class of materials in order to examine physical characteristics and properties. Historically several editorial teams, typically consisting of field experts, have tried to define and classify these materials at the international database organizations. Yet if a novice explores these databases they would find various answers to simple questions such as how many perovskites are known? What are their chemical compositions? How are they defined?

The Powder Diffraction File has grown dramatically due to collaborations with other database organizations. A consequence of these collaborations is that the combined output of several different editorial teams from several organizations are now available in a single product. This combined knowledge while powerful can be confusing because the answer you get depends on how you ask the question (definition and convention) and how many questions you ask (i.e. the paths that you take).

Using a relational database format allows the user to use any field parameter as a query. Table I shows the general types of information that are available in the Powder Diffraction File relational database.

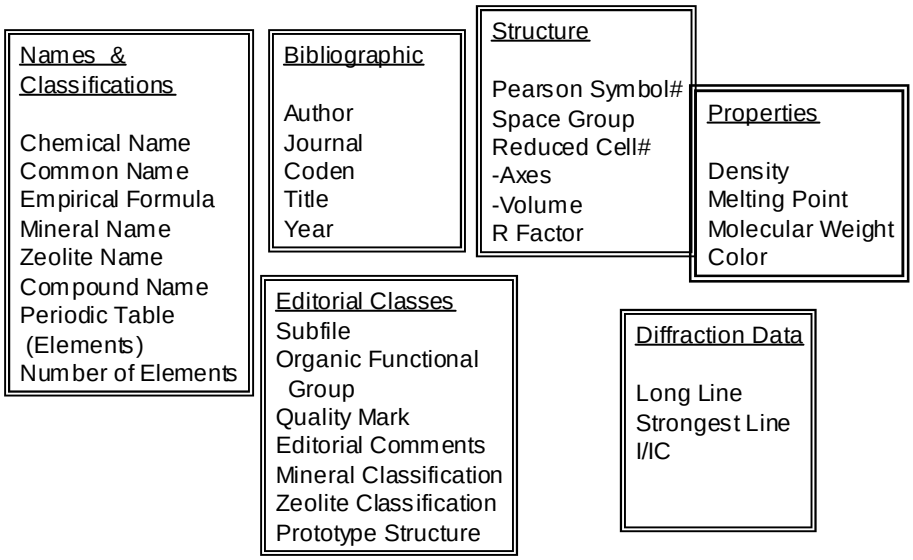

Figure 1. Different fields that can be independently search in the Powder Diffraction File (PDF™) grouped in general categories.

Each of these fields becomes a different way of querying the database. For example one can query for perovskites through an author title search. Using the first five letters of the word "perov"skite a text search will find 1021 authors titles. Perovskite is also a mineral classification; a similar five letter word text search will identify 80 entries. The difference in these populations is "who" defined the perovskite, the original reference author, or an editorial review team. In the latter classification, mineralologists, by definition, classify perovskites that are known minerals in nature; many synthetics are omitted by convention. Editors for the Inorganic Crystallographic Structural Database (ICSD) has used ABX prototyping conventions that have been adopted by the Powder Diffraction File (PDF). In the PDF the ABX3 prototype can be searched by searching the comment fields with a text search. However this convention has not been used in powder diffraction files that were experimental where atomic parameters were not determined. One can also search for perovskites through well known perovskite prototype structures, using the methodology originated by Pearson [3] and latter refined by Villars and colleagues [4]. A search of the classical CaTiO3 prototype yields 593 entries.

The ability to query any of these fields individually results from the relational database architecture. Using the utility program DDView allows us to filter, sort and combine queries. For example, each of the above queries results in a separate population of perovskite structures that depends on the editorial assignments, prototyping and editorial conventions used for that particular classification system. With DDView, the Preferences Module and use of Boolean operators with Global Searches allows the user to string several queries together, eliminate duplicates and save a user-defined number of variables in a composite table. This composite table can then be filtered and searched.

Search	PDF Entry Population
Mineral Perovskites	80 Entries
Author Defined Perovskites	1,021 Entries
ABX3 Prototype	2,973 Entries
CaTiO3 Prototype	593 Entries
Combined Searches	4,182 Entries

Table I. Entry populations for various searches on perovskite related materials

The unusual result was that the searches used such varying criteria that there was only a small overlap in the entry populations. However in this simple exercise, the user has now defined a customized subset of the database specific for perovskites that is a unique compilation. This list could easily be expanded using other known chemical formula prototypes (i.e. BaTiO3) and generic formula searchers AxA1-xBX3 or AA'B2X6. A subsequent combination of 6 different searches yielded 4,528 Entries.

However creative individuals can also define their own criteria and systems. For example previous works [1] have already identified common space groups in the Perovskite family. Chief among these space groups are Pm-3m, R-3c, and Pnma. Using the "Preferences" module in DDView the authors examined the entry populations in Table I and not unexpectedly found very high populations of these space groups in each entry set, an example for the 1,021 author defined perovskites is shown in Figure 2. Several additional high population space groups were also found. The authors then used many of the same tools historically used in references 3 and 4 to compare formula, crystallographic parameters and diffraction data in these population sets. This process is very analogous to the "reverse engineering" process by using known correlation parameters on a set of standards and then use the parameters to find and identify new materials for the perovskite family.

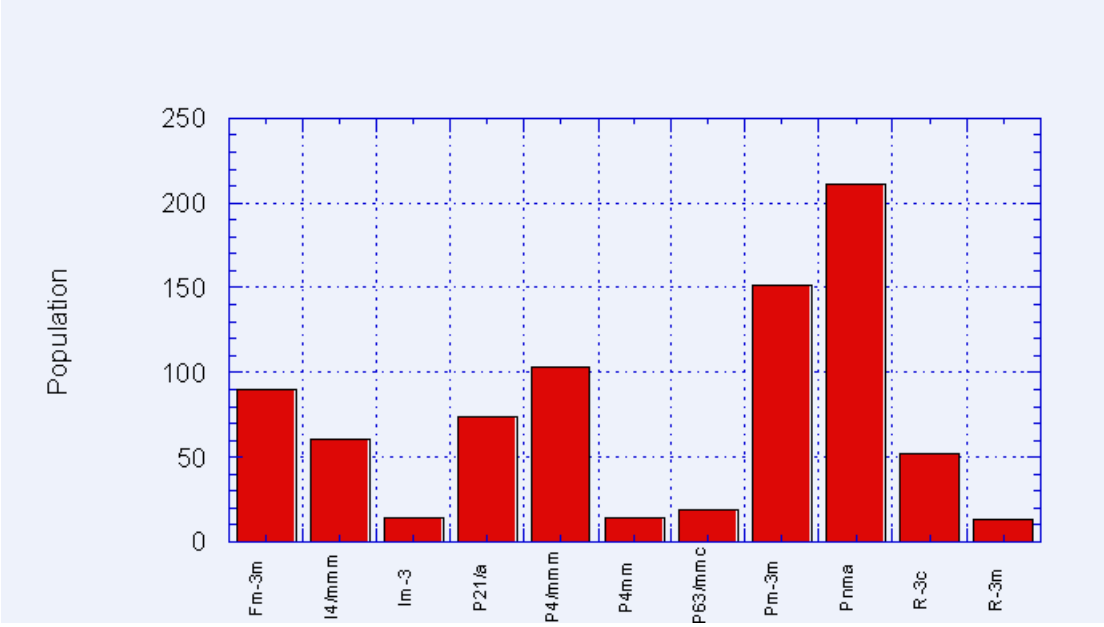

Figure 2. Entry population statistics where the authors has referred to perovskites in the title of the publication and the resulting space group of the perovskite analyzed.

The process is shown in the screen capture in Figure 3. Here we explore the population of the space group Pm-3m in the compilation of the author title search. In this table one can compare empirical and chemical formula, Pearson symbols, prototype designations, the reduced cell, reduced volume, cell edge rations, longest peak and three strongest peaks (Hanawalt).

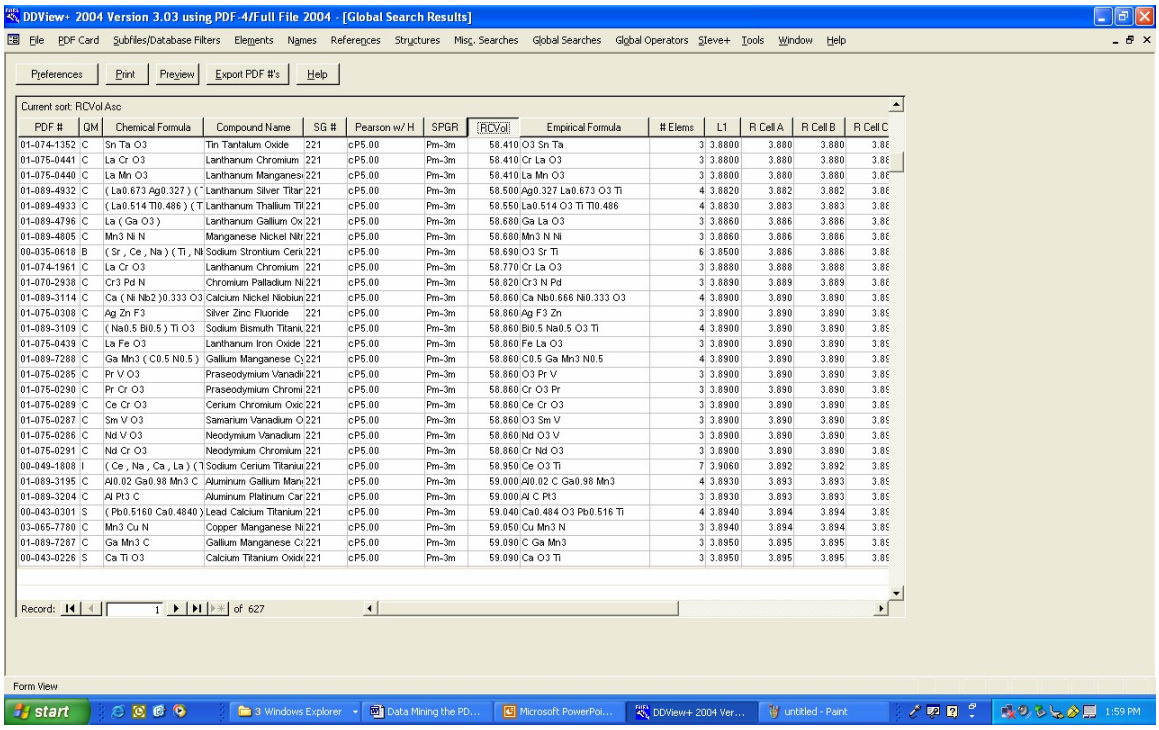

Figure 3. Screen capture from the DDView+ program of the Powder Diffraction File comparing various fields for material entries in the perovskite mineral classification. Dozens of different fields are available and are user selected. Scroll bars on the bottom and right enable the user to view large population sets.

It quickly becomes apparent upon visual inspection of the table that most of the entries can be characterized by their 1) Pearson symbol, 2) small unit cell volumes, 3) empirical formula and 4) a narrow range found in the longest d-spacing (L1) from the x-ray diffraction pattern. These four criteria describe 146 materials found in the 151 data sets of author defined perovskites from space group Pm-3m. Now reverse engineering was applied in that these criteria were then applied to the entire database. Once again each criterion was used in a query and the query was combined to eliminate any duplicates. In this manner 627 data sets were found for space group Pm-3m. Similar processes were applied to a total of 4 different space groups and 3,124 data sets were identified. It should be mentioned that the authors in this specific case used a very broad definition of perovskites that includes many mixed layers and defect structures. For example many mixed and multi-layer perovskites are found in the space group P4/mmm and it would be the user's option to decide whether to use these materials or not. Several of the new entries found that were not identified previously as perovskites were materials analyzed prior to the common use of the term in the 1940's. In reviewing data, such as author's title and year of publication it became readily apparent that the term perovskite became more popular in the last 30 years.

There are other ways of searching for a structure family using the PDF and associated software these include:

Search by empirical formula or formula fragment
Search by pattern matching

In the first case, a Boolean text search in either the empirical formula or the compound name can produce a number of entries that can be related to a structural class or chemical species. For example manganates, titanates, ruthenates and cuprates can be searched using these elements and the ABX3 formula. For example one could search on CaTiO3 and combine that with three or more element searches to get ABX3

structures with calcium titanate chemistries. The search on cuprates produces many superconductors that contain perovskite interlayers.

Searches by pattern matching utilize the program SIeve [5] that is integrated into the Powder Diffraction File (PDF) product line. The concept is to take an unknown sample, statistically match it to materials with similar diffraction patterns and then analyze the collection of matched patterns for trends in structure and chemistry. In this process thousands of candidate structures can be reviewed and reduced to a manageable list of candidate solutions. Under the best of conditions this could lead to phase identification and/or structure solution of a new material. This search takes advantage of the ability to sort, filter, organize and review data with the program DDView [6] and then index and pattern match with SIeve. SIeve uses an 8 d-spacing fitting algorithm [5] with a user option of Hanawalt, Fink or Long 8 search algorithms. A statistical goodness of merit is calculated for each reference pattern [5]. Once a series of candidate materials is identified the matched materials as a collective body can be imported into DDView so that the user can explore the crystallographic and chemical trends present in the group.

As a test case the pattern of LaAlO3 was exported from the PDF and then imported into SIeve and treated as an unknown. As one might expect LaAlO3 was the top candidate identified by pattern matching process. but it is interesting to examine structural trends in the other candidate matches. Figure 4, shows two representative candidate materials taken from the list of 43 materials with a high goodness of merit [6]. These can be graphically compared by their digitized patterns. It is interesting to note that of the candidate materials of high goodness-of-merit, the majority are known perovskites in many of the classic perovskite space groups [1, 2].The second best candidates are cubic rock salts in the AX family. If this sample were a true unknown it would have been reasonable for the researcher to hypothesize that the material was probably in the perovskite or rock salt family with a fairly narrow choice of 4-5 different space groups. By itself this is not sufficient information for accurate phase identification but it could be if combined with either other physical property measurements or a correct structural refinement using one of the candidate structures.

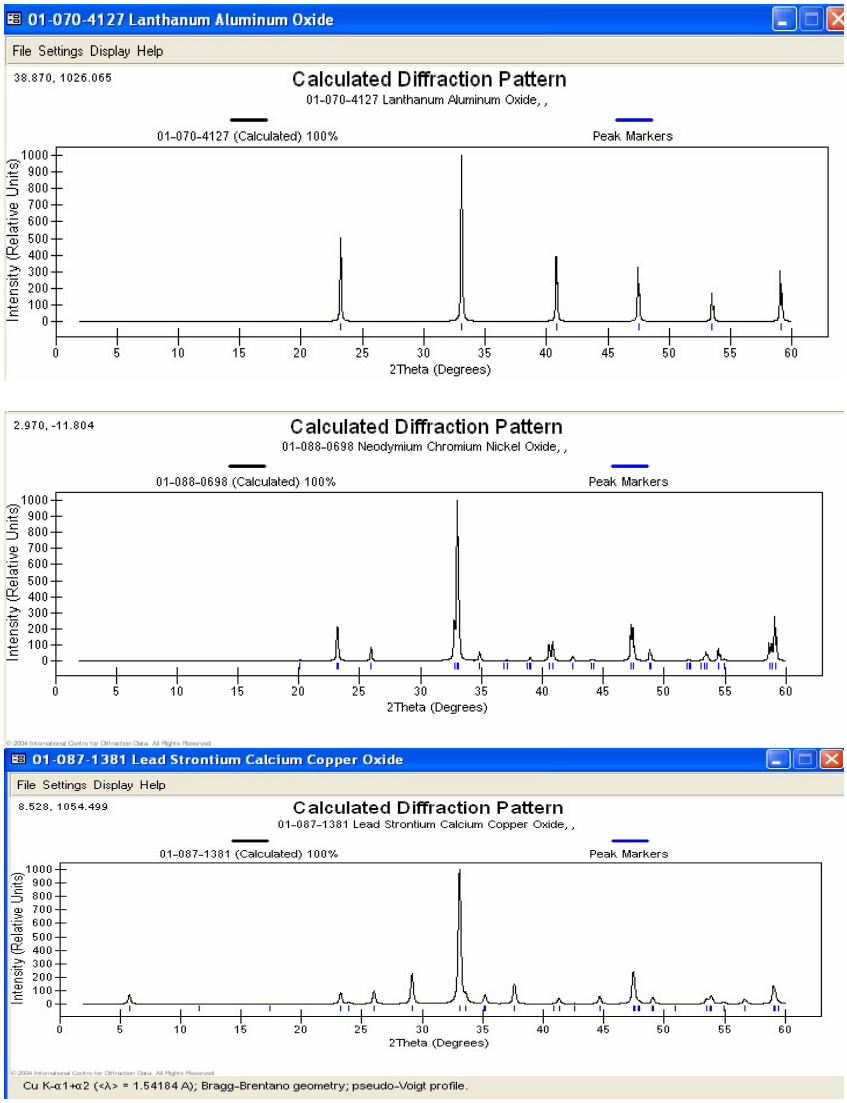

Figure 4. Digitized patterns comparison. Top LaAlO3 reference test case and pattern matches to neodynium chromium nickel oxide and lead strontium calcium copper oxide. The former being a known perovskite and the latter being a superconductor with perovskite layering.

In the above examples we have shown seven different ways that the database can be "mined" to explore perovskite materials. These ways can be combined at the user's discretion, and using user definitions to create a user defined dataset for further structural analysis. Perovskites were used as a test demonstration but analogies can be developed for several different classes of materials or sets of analytical problems. Similar examples have been explored for battery materials and narcotic drugs for the energy industry and law enforcement agencies, respectively. The development of new tools for data mining such as DDView and SIeve offers global scientists enhanced capabilities in materials exploration and identification.

References

[1] Landolt-Bornstein, "Numerical Data and Functional Relationships in Science and Technology, Magnetic and Other Properties of Oxides and Related Compounds", New Series, Group 3, Volume 4, Springer-Verlag Publisher, Berlin, (1970) pp 127-149.
[2] C.N.R. Rao and J. Gopalakrishnan, "New Directions in Solid State Chemistry", Cambridge University Press, Cambridge, Great Britain, 2nd Edition, (1997), pp 437-443.

[3] Pearson's Handbook Desk Edition, "Crystallographic Data for Intermetallic Phases", Volume 1, ASM International Publisher, Materials Park, Ohio (1997), preface.
[4] The Villars prototyping system was first published with the Linus Pauling File, Binaries Edition, Version 1.0, 2002. The prototyping system is described on pages 61-62 of the users manual.
[5] J. Faber, C. A. Weth and J. Bridge, " A plug-in program to perform Hanawalt or Fink search-index using organics entries in the ICDD PDF-4/Organics 2003 database", Powder Diffraction, 19 (1), March 2004.
[6] J. Faber and J. Blanton, "DDView+, a New Interface for ICDD PDF-4 Databases: The Use of User-Selectable Search Preferences for Data Mining and Total Pattern Display", to be presented at the 2004 Denver X-ray Conference and submitted to the conference proceedings, Advances in X-ray Analysis.