research papers

Full Profile Search-Match

Luca Lutterotti,[a,c]* Henry Pillière,[b] Christophe Fontugne,[b] Philippe Boullay[c] and Daniel Chateigner[c]

[a]Dipartimento Ingegneria Industriale, Università di Trento, Italy, [b]INEL SAS, Artenay, France, and [c]CRISMAT - ENSICAEN, Université de Caen-Basse Normandie, France. Correspondence e-mail: luca.lutterotti@unitn.it

A new search-match procedure has been developed and tested that, in contrast to previously existing methods, does not use a set of lines identified from a pattern, but an optimised Rietveld fitting directly on the raw pattern. Modern computers with multicore processors allow the routine to be fast enough to perform the entire search on a reasonable time using as a database the Crystallography Open Database (COD). The search-match is done using the crystal structure for all phases, the instrumental geometry, and as such can be applied to every kind of diffraction experiment, using X-rays, thermal or TOF neutrons or electrons. The methodology is particularly efficient when dealing with nanocrystalline samples for which peak identification may be a problem. A web interface has been developed to permit an easy testing and evaluation of the procedure. The goodness of the results mainly depends on the availability of the searched phase in the structure database. The method permits not only the phase identification but also a rapid quantification of the phases and their gross microstructural features provided the instrumental function is known.

1. Introduction

The history of search-match is at most coincident with the history of Hanawalt and the International Center for Diffraction Data (ICDD), and a good reading for the early days is the paper by Hanawalt (Hanawalt, 1986).

After the "Hanawalt" manual search the appearance of the personal computer has revolutionised the search-match making it faster, more powerful and easy to use (Smith & Gorter, 1992; Langford & Louër, 1996). More lines could be used, subsets and chemical information helped the identification of the correct phases and errors were taken into account. Most of the work has been devoted to correctly determine peak positions in an automated way and use, as much as possible, the intensity information. Some search-match methods are working also in a double step: the first step is a classical search-match using the identified lines, then in a second step the possible structures are compared directly with the powder pattern to further refine the choice.

A large part of the search-match work has then been devoted to improve the algorithms and software to provide more and more reliable and automated results limiting human intervention. In parallel another big work has been dedicated to increasing the number of "cards" from Hanawalt days, using experimental data from different labs and collected by ICDD into the Powder Diffraction File (PDF) databases. In the last ten years a large improvement in both number and reliability of cards have been obtained by including calculated "cards" or patterns. The crystal structures coming from different databases, notably the Inorganic Crystal Structure Database (ICSD) and the Crystallography Open Database (COD) (Gražulis et al., 2009; Gražulis et al., 2012), has been added in search-match procedures. As these database provide the crystal structures for the identified phases, one can proceed after with phase quantification using the Rietveld method (Rietveld, 1967; Rietveld, 1969) that does not require any standard or knowledge of reference intensity ratio (RIR) factors.

In 2002, Le Meins et al. (Le Meins et al., 2002) organised a Round-robin and the overall result was satisfactory, but if it proved that "a good search-match software, in expert hands, combined with an up-to-date PDF can perform efficiently in solid state phase identification" (from authors words), it showed also that only one participant was able to succeed in identifying all compounds and still the "in expert hands" gives already a warning for a general applicability.

But the investigation of nanostructured compounds, or in general not perfectly crystallized samples, was not given any concern in this round-robin. For such compounds, the increased peak broadening makes reliable peak finding and indexing more difficult to perform, and unfortunately increased instrument resolutions are of no help. Another general problem is that the traditional search-match, in certain cases, identifies a large amount of phases that make it difficult for the non-expert to discern which ones are really present and which are not. This is in part due to the ranking mechanism that uses mainly the list of identified peaks, with more emphasis on number and positions as the true overlapped intensity cannot be exploited.

As a general trend, actual search-match procedures are very efficient in identifying phases in compounds coming from synthesis or chemical labs, with well defined and crystallized structures. But this efficiency decreases for less defined structures, either because of poorer crystallinity, or/and with strong inter- and intra-phases overlapping, e.g. in soil samples or raw industrial materials.

In 2009 a project named "Nanoair" was founded by the Euro-

pean Community (Nanoair, 2009) to develop a fully automatic and remotely controlled instrument to collect nano-powders from air and perform a complete analysis for particle dimensions and composition (phase composition by diffraction). The analysis part of the project focused on performing an automatic phase identification and quantification using the Rietveld method. As the result of the phase identification would have been automatically restricted to the case of known crystal structures, we decided to work on a new procedure to merge the phase quantification and search-match in a unique step using the Rietveld method. In that project the software/procedure FPSM (Full Profile Search Match) has been developed. Some early results about the Nanoair project have been presented in a poster at the EPDIC 13 in Grenoble (Lutterotti *et al.*, 2012). From that time the method and software has been tested and improved and the full theoretical basis, implementation and some testing results are presented in the following.

2. Theory

2.1. Automatic Rietveld refinement

The entire searching and profile fitting algorithm is based on the Rietveld method (Rietveld, 1969). In particular we use a special implementation, as in the Maud software, that differs from the classical Rietveld implementations for three main points:

(1) There are no scale factors for phases, but the volume fractions are refined directly.

(2) The peak profile function is composed by two different contributions (explicitly): the instrument broadening profile and the sample broadening. The two contributions are directly convolved for the resulting peak broadening and in the analysis we do not refine the instrument broadening contribution but only the crystallite sizes and r.m.s. microstrains.

(3) The refinement strategy, e.g. the parameters refined in successive order, is automatically decided by the algorithm and does not require human intervention or decision.

Only the last point is strictly necessary for the Rietveld search-match algorithm but the two formers are convenient as they provide directly the quantities of interests with a correct evaluation of their e.s.d. values.

The general formula used for the Rietveld computation is the following:

$$I(d_i) = bkg(d_i) + I_0 \sum_{j=1}^{N_p} \frac{f_j}{V_j^2} \sum_{k=1}^{Nr_j} L_k |F_{k,j}|^2 S_{k,j} (d_i - d_{k,j}) P_{k,j} A_{k,j} \tag{1}$$

where we use the variable d_i as coordinate for the pattern instead of 2θ because we use here for both angular dispersive and energy dispersive (e.g. neutron TOF) techniques. For the other symbols: $bkg()$ is the background function, I_0 the incident intensity, N_p the number of phases; f_j is the volume fraction of phase j, V_j its unit-cell volume and Nr_j the number of reflections of phase j inside the pattern range considered, L_k the

Lorentz-polarization factor for reflection k, $F_{k,j}$ its structure factor, $S_{k,j}()$ the peak shape function of reflection k of phase j, $d_{k,j}$ its position in d-space coordinates, $P_{k,j}$ the texture or preferred orientation factor and finally $A_{k,j}$ the absorption factor.

Now we are going to detail some of the functions and models used in the Rietveld algorithm implementation.

For the background function we use a simple polynomial function plus an additional gaussian decay at zero 2θ or d-spacing (for TOF/energy dispersive patterns) that we can find in some experiments close to the incident beam. In the case of electron diffraction in Transmission Electron Microscopy (TEM) such a strong contribution is always present, resulting in a strong increase of the background at low d-values.

For the peak profile we use asymmetrical Pseudo-Voigt function as described by Enzo et al. (Enzo *et al.*, 1988) and Lutterotti et al. (Lutterotti & Scardi, 1990). The instrumental broadening part should be determined before using a method like in (Scardi *et al.*, 1994) and we made it compatible with the one in the Maud software, so the latter can be used for the instrument broadening determination and then fixed in the search-match procedure. As reported later, in the web interface we prepared some general categories for instrumental broadening to represent most of the available cases. An imperfect definition of the instrument contribution does not prevent the search-match to work correctly (unless a low resolution instrument broadening is used for a high resolution instrument and well crystallised sample), but it will give wrong results for the crystallite size and microstrain values, and makes the search more difficult.

The broadening due to the crystallite size and microstrain for each phase is computed using a simple inversion of the model described by De Keijser et al. (De Keijser *et al.*, 1983; Delhez *et al.*, 1983). In the Rietveld algorithm we compute first the integral breadth (both the Gaussian β_g and Cauchy β_c parts) for the Pseudo-Voigt from the value of crystallite size $\langle D_V \rangle$ and r.m.s. microstrain $\langle \epsilon^2 \rangle^{\frac{1}{2}}$ according to:

$$\beta_c = \frac{d^2}{\langle D_V \rangle} \tag{2}$$

$$\beta_g = \sqrt{2\pi} \langle \epsilon^2 \rangle^{\frac{1}{2}} d \tag{3}$$

The integral breadth is then convoluted analytically with the Pseudo-Voigt, symmetric instrument broadening part and finally a numerical convolution with the asymmetry function is operated as described in (Lutterotti & Scardi, 1990). For the analytical Pseudo-Voigt convolution we started from the formulation of De Keijser et al. (Delhez *et al.*, 1983; De Keijser *et al.*, 1982) but respect to their work we need to invert their formula to obtain the convolution result. We used a least squares fitting routine to determine the proper coefficients for the convolution formula for Pseudo-Voigt.

The Lorentz-polarization correction is computed according to four different cases: Bragg-Brentano, Debye-Scherrer, transmission geometry (TEM or images in transmission integrated around the diffraction rings) and neutron TOF. No correction for monochromator or highly polarized beams like on a synchrotron is used as in nearly all cases it will not greatly affect

the results. In general for a better refinement, after the phase identification and first quantification using this method, a more accurate complete Rietveld refinement may be ideally practiced using all corrections and required proper models.

We chose not to use a preferred orientation model, for two reasons. The first one is that the only automated and efficient method for texture correction using one pattern is the spherical harmonic model (Popa, 1992) as it does not require to establish a preferred orientation plane like the March-Dollase (Dollase, 1986), but its use will decrease quite a lot the execution speed. Second, it is proved by common practice and testing (Lutterotti, 2012) that in general a preferred orientation correction, when not based on a true ODF (this requires the complete texture coverage with several measured patterns (Ferrari & Lutterotti, 1994; Wenk *et al.*, 1994; Lutterotti *et al.*, 1997)), may improve the quality of the fitting, but might end up usually in a wrong estimation of the phase quantities. So, especially as we are targeting a phase quantification in addition to the search-match, and we do not contemplate human judgment, we decided it is safer not to use any texture correction.

For the absorption-volume correction, the formulation used depends on the diffraction geometry and 4 were implemented: 1) classical Bragg-Brentano with $\theta/2\theta$ measurements for which the absorption factor is a constant and can be ignored for bulk samples, 2) Bragg-Brentano with 2θ measurement only and the absorption factor is calculated as a function of 2θ from the incident beam angle and phases composition assuming a bulk sample, 3) Debye-Scherrer camera with the absorption-volume correction for a cylinder and 4) transmission and TOF neutron geometry with no absorption correction.

Finally, we explain how an automatic Rietveld refinement is performed in the code. The execution strategy is very similar to a Rietveld analysis focused on phase quantification and not on crystal structure refinement. We remind the reader that the goal is to obtain a quick correct analysis and avoid as much as possible a solution divergence. Thus we reduce the number of refined parameters to the minimum possible and privilege physically meaningful parameters/formulation even if it does not result in the best possible fit of the pattern.

The refinement strategy is reduced to three cycles and those correspond to the first three steps performed by a Rietveld expert performing a pattern fitting. The first step consists in refining only the background and scale parameters, the second step additionally refines peak positions and finally in the third (final) step the peak profile parameters are introduced. There is no forth cycle to refine the crystal structure or further fitting as explained before.

To go into more details considering our Rietveld specific formulation, for each cycle we perform the least possible iterations to reduce the execution time. We actually use two different types of refinements, one during the search or scan through all the selected structures (as explained in the following paragraph) that is focused on speed and we do not use more than 3 iterations per cycle or even less. The second type of refinement is performed to assess the phase quantification and requires a better fitting. It is done only at the end of the selection of a new

phase and thus it does not need to be as fast as possible being executed only few times. In this case more iterations are performed to achieve a better convergence and more parameters can optionally be refined to improve the fitting.

During the first cycle, 3 to 5 background parameters are refined in the polynomial function (depending on the computation range) plus the incident intensity and a decay parameter for the diffuse scattering halo around zero. This step is also used to get a first estimate of all the volume fractions. As the volume fractions should amount to a total of 100%, at the end of each iteration, the volume fractions are normalised and the normalisation factor is used to change the incident intensity (see Eq. 1). We keep this particular procedure to remain compatible with the calculations in Maud, in which working with volume fractions (not simply phase scale factors) is necessary when dealing with multiple data sets (from different instruments) or layered and multilayer samples.
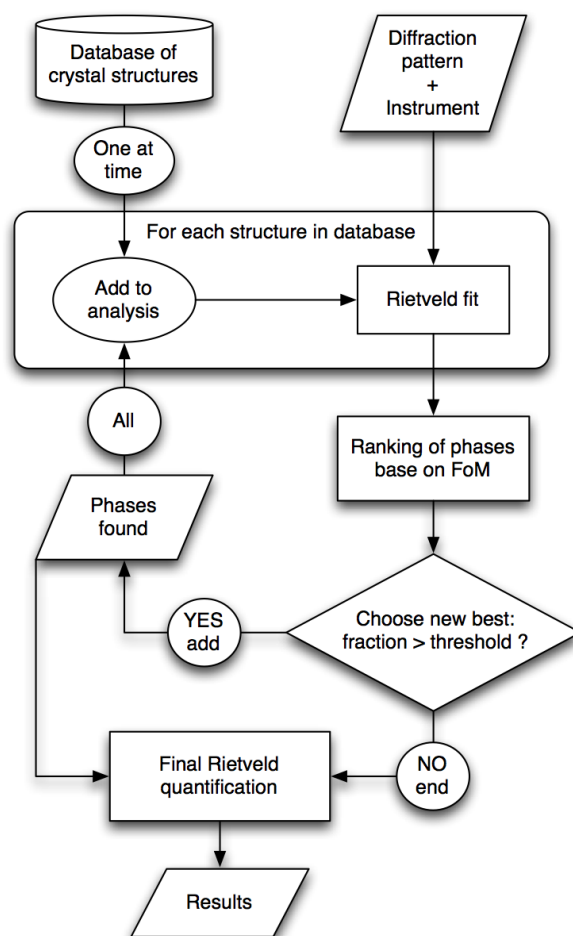

Figure 1
Schematic layout of the FPSM algorithm.

After the first cycle, if there are multiple phases and one or more of them results in a volume fraction below a minimum threshold (normally between 0.5 and 0.1%), these are eliminated from the refinement. For the remaining phases, in the second cycle we add to the refinement the cell parameters only of

those with a volume fraction over a sensible threshold (around 5 and 10%). In addition, a classical 2θ or constant d-space parameter is refined to account for goniometer or measurement errors. Regarding the cell parameters it is not advisable in our case to refine all of them especially for triclinic or monoclinic phases and during the search algorithm we use instead a volume expansion or contraction parameter, thereby reducing the total number of refinable parameters added in this cycle to $(1 + N_p)$. In the third cycle we add to the refinement the crystallite size and microstrain parameters for each phase with a volume fraction over a similar threshold as before. During this cycle we refine also an overall Debye-Waller thermal factor common to all atoms of all phases.

2.2. Search and Match algorithm structure

We describe in this section the main algorithm used for the Full Profile Search Match (Fig. 1). We start from a crystal structure database restricted to match the composition conditions. In principle the method may work with all the possible phases without any composition restriction, but we have to consider the computing time, and limiting the number of phases using some restrictions on the composition is actually important to perform the search in few seconds or minutes. Such an elemental input can be provided from any spectroscopic knowledge on the composition.

Lets call S_1 the element-restricted COD subset, composed of N potential phases and S_2 the subset containing the identified phases initially empty. Each phase of S_1 is then evaluated against the experimental pattern by performing an automatic Rietveld refinement as described in the previous section. At the end of this run the N phases are ranked based on a Figure of Merit (*FoM*, defined later) and the rank-1 phase is moved to S_2. All phases of S_2 (at the moment just one) are then added in turn to each of the $(N - 1)$ remaining phases of S_1 for another automatic Rietveld run over these $(N - 1)$ combination of phases. The combination $(S_2 + 1$ phase from $S_1)$ providing the best *FoM* indicates the next phase that has to be moved to S_2, letting in S_1 only $(N - 2)$ phases. At this point a more accurate Rietveld refinement is performed to evaluate the volume fractions for all the phases found up to now. If one of the S_2 phases falls below a critical threshold (between 0.5 and 0.1%) the phase is moved out of S_2 and another cycle starts to evaluate the next phase to add to S_2. The process is carried out until the phase removed from S_2 is also the last one added or there are no more phases left in S_1 or the number of phases in S_2 is over an (optional) imposed limit. We keep the last possibility as a precaution to stop the algorithm in case of a completely failing identification giving raise to several phase additions without a real fit improvement. In our still brief experience with the algorithm the latter is happening when i) severe errors are present in the pattern (including wrong data formats etc.); ii) the pattern does not correspond to the declared experimental conditions; iii) neither the phase nor a closely related structure are present in the database.

Optionally, after each cycle, we can eliminate from S_1 not only the phases moved to S_2 but also the ones giving the lowest

FoM to speed up the next Rietveld-ranking having less structures in S_1.

2.3. Figure of Merit

The simpler Figure of Merit that can be used for the proposed Full Profile Search Match is based on the R_{wp} factor (Prince, 1983) computable directly by the Rietveld fitting. As we are using it to test the suitability of a phase to reproduce a pattern and we aim to decide if the phase is present or not, it may be reasonable to use the more general and best overall Rietveld index.

In such case the Figure of Merit becomes:

$$FoM = \frac{1}{R_{wp}} \tag{4}$$

This *FoM* was also the first one used during testing and tuning of the algorithm. But we found it not the best one to search and sort out the structures composing the pattern for several reasons of which the main ones are:

(1) The R_{wp} criteria does not distinguish between two similar phases from the database mainly differing by their cell parameters values. This may happen for two structures originally determined at different temperatures or compositional modifications where the structure factors do not change too much in relationship to the pattern quality (other overlapping phases may be present and would still not be identified).

(2) When the sample is composed by light and strong scattering phases, the R_{wp} privileges the strong scatterer. Sometimes, the light scattering phases may contribute very little to the pattern and R_{wp} in term of intensity and in such cases they never come out. Wrong phases may be picked up if they contribute to just fit better the background.

To overcome these problems we propose a modified *FoM*.

$$FoM = \left(\frac{1}{R_{wp} + a\frac{\rho_r - \rho_0}{\rho_0}} + bf_r \right) \delta \left(\langle D_V \rangle, \langle \epsilon^2 \rangle^{\frac{1}{2}} \right) \tag{5}$$

where ρ_0 is the original density, ρ_r the refined one and f_r the volume fraction of the structure for which we are evaluating the *FoM* from the Rietveld fit. The quantities a and b are weighting coefficients to be tuned for optimal performance. In addition we add a strong penalty function δ if the mean crystallites of the phase is below a customisable threshold value (default is 20Å) or the microstrain too large similarly.

The term weighted by the coefficient a, is a penalty factor to give more importance to structures with the closest cell parameters to the observed ones. The term weighted by b, allows low scattering phases to be selected in virtue of the non negligible volume fraction they are representing.

A correct tuning of a, b and δ is of customary importance for the success of the algorithm, as the only R_{wp} criteria is not sufficient to cover all the possibilities. The algorithm tuning is a work under progress and it may happen to find cases in the future requiring further refinement of the FoM. The method is still on the early stages and we are going to compare it with

much more mature search-match procedures refined and tested in the last 20-30 years.

3. Implementation and testing

During the implementation we tested different possibilities on how to do the calculation, to store and parse the database and if/when to use also the GPU or just the CPU. This full profile search match algorithm is based mainly on pure speed or brute force to find the structures more likely present in the sample. Thus its adoption as a routine tool to make quick automatic analyses depends mainly on the speed of execution of the code. For the implementation and testing we used both a portable computer with an i7 quad-core 2.8 GHz processor and a workstation with 2 x 3.33 GHz 6-core Xeon resulting in 178 Gigaflops and 2 x Tesla 2050 GPUs giving 856 Gigaflops raw computation power. Both systems are equipped with SSD. The GPU computation speed was tested only on the workstation. The language used was C++ for the main program and OpenCL for the GPU calculations. For the database we used mysql for the format and for parsing. All the crystallographic computations (space group interpretation, symmetry operations, reflections list and structure factors calculation) were made using the excellent cctbx library (Grosse-Kunstleve & Adams, 2002). The latter proved very efficient in all the calculations and we used only the C++ implementation without the Python bindings. We added to the cctbx library the scattering factors for electron diffraction from (Peng *et al.*, 1996) to enable a simple kinematical calculation of structure factors in such case.

The structure factors are computed only once at the loading of the structures from the database and the atomic structure is never refined apart from a global Debye-Waller or thermal vibration factor. We tested one implementation in which the intensities were stored (thus limiting the implementation to one radiation and wavelength) in the database instead of being computed, but we found that loading all the peak intensities for each phase was taking the same or more time than calculating them once from the crystal structure. This behaved identically even using a quick SSD to store the database. In any case, Even if the database can be further optimised and the loading made quicker, the structure factors computation is not a limiting factor in speed and we choose the latter as more flexible, accurate, more general (many different radiations/wavelengths can be used without changing the database) and compactness of the structures database. A smaller database can be parsed more quickly to load only the interesting structures.

For the non-linear least squares algorithm we use a Marquardt (Caceci & Cacheris, 1984) algorithm which implementation was derived from the Maud software (Lutterotti, 2010). The derivatives are computed numerically and this has some advantages, like the possibility to use numerical models without an analytical function, but also some disadvantages like calculation speed and intrinsic accuracy. As in our formulation we have some numerical only models we were obliged to use numerical derivatives. We tested both the speed and robustness of the algorithm against some open source and commercial non-linear least squares packages with the aim to find something better, but

we did not succeed, neither for speed nor especially for robustness.

Figure 2
Web interface to test the algorithm using the COD database of crystal structures. The user uploads a pattern in a double column format, selects the instrument type and experiment conditions, restricts the search using composition and/or COD subsets and finally runs the analysis. There are two mirror sites at the moment: http://nanoair.dii.unitn.it:8080 and http://cod.iutcaen.unicaen.fr.

Some long time has been spent on trying to take advantage of the raw power of the GPU but without any success. The reason is mainly in the kind of computation the Rietveld algorithm requires. The GPU does not support a too complex computation or requiring too much memory per computing unit and an efficient refactoring of the computation to avoid frequent memory movements between CPU and GPU has not been found being the latter a strong limiting factor in GPU calculations.

Instead we took maximum advantage of the CPU power exploiting all possible cores/threads and the vector computation. In a modern CPU we can run 2 threads per core, even if 2 threads in a core are going a little slower (per thread) than only one; the gain is estimated to 50%. With our 2 x 6 cores workstation we can use 24 simultaneous threads during the computation. Also, each thread may put in the computation pipeline 4 floating numbers at time, so we used this feature explicitly by using always when possible matrices and vectors multiple of 4 and exploiting the vector computation. Dealing with pattern with a lot of numbers, long list of reflections etc. this gives a gain of nearly 4 in many calculations.

The structures database preparation required more time than expected to solve some of the problems encountered and optimise everything for speed and efficiency of the results. We choose mysql to implement it for its availability and flexibility. Many open source solutions are available as libraries and interfaces for it. The performance of mysql for a medium-big database as the one required in our case are satisfactory.

All the structures imported in our database are coming from the COD (Crystallography Open Database (Gražulis *et al.*, 2009; Gražulis *et al.*, 2012)) in CIF format (Hall *et al.*, 1991).

In principle we can use any other source providing crystallographic structures in CIF format, but for testing and our first database we used the COD as it contains also the AMCSD (American Mineralogist Crystal Structure Database (Downs & Hall-Wallace, 2003)), the zeolites database (Baerlocher *et al.*, 2007) and it is the largest one for which the CIF files are freely available comprising inorganics, organics and organometallics together.

A dedicated interface for OSX only has been written to import the cif files, manage the databases, perform checking and merging operations and producing the final optimised database for the FPSM core program. The OSX program performs the following tasks:

(1) imports all the CIF files from a local repository (in our case we used a clone of the COD at a specific time; we plan to implement later a function to update the database in continuous when structures are added to COD).

(2) Save the database in mysql format. Other practical operations can be done on the sql databases like merging two or more of them, removing or adding new structures, sorting and manual checking or modify the information stored.

(3) Check each structure for consistency: cell content with respect to chemical formula, space group and symmetry operations. At the present implementation we do not make any check for bond distances and we will not do it in the future.

(4) Classify the phases in inorganic, organic or unknown (metallorganic mostly) when no clear distinction can be made from the raw chemical formula. We hope this can be better done in the COD database in the future with the help of the users depositing the structures.

(5) Finally each sqlite database (full, organics + metallorganics, inorganics + metallorganics) is optimised to remove similar structures. We define as similar structures two phases with similar composition (variation in atom quantities are allowed), a density difference smaller than 5% and resulting in two very close patterns (that with one phase we can fit the pattern generated by the other with a Goodness of Fit (*GoF*) smaller than 2).

The last time (27 february 2014) we built up our database from COD, 263327 structures were loaded successfully, 120 discharged because retracted (see (Harrison *et al.*, 2010)), 40 CIFs were duplicates of others and 769 not loaded because of space group, atoms or composition problems. Within the latter ones, most of them are modulated structures that are not supported in our implementation at the present time and thus really few remain with a chemical composition not corresponding to the cell content or other errors. For them, even a check with the original article was not sufficient to solve clearly the problem (most of the time the CIF is a partial solution or a general framework with some sites not occupied by a specific atom) and we prefer not to use them in our method. Indeed we could successfully identify, validate and use more than 99.6% of the structures of the COD. Exception of retracted, duplicated and modulated structures, the percentage of problematic CIFs in COD drops to nearly zero and it is actually a good sign of the quality reached by COD in the last few years.

The core program FPSM instead is devoted to the full profile search match only and the implementation was based on the layout of Fig. 1. We chose to keep it independent from any graphical user interface for easy compilation and running on multiple platforms. This core part can be compiled and used as a library by any graphical interface and as such was included in the previously described OSX program to build the database and test the algorithm. But to permit everyone to try and use it, at the end of the testing period we have built an interface accessible through a web page.

The web interface is reported in Fig. 2, and has a minimal input to keep it simple and focus mainly on usability.

It has 3 sections for the input. In the first section we load the pattern to be analysed as a simple text format with two columns: first column is the diffraction coordinate (2θ for everything except for TOF measurements for which d-spacing is expected) and the second the intensity. After loading the pattern the user may specify a possible composition, and if chemical species are entered, the program will use phases containing only such atoms and exclude all the phases having atoms not specified there. This may be helpful to restrict the number of phases to search from and it is strongly advised to use it. In case of doubt many additional atoms may be included to lower the restrictions. Two others restrictions can be set by the web interface: the first is a threshold density to avoid lower density phases to enter the search (gases or low density organics), the second to limit the search to a maximum number of phases as explained previously. Other threshold values are not changeable through the web interface. Finally for a nanocrystalline sample, a check box can be ticked. This will change the initial value used for the crystallites, to a smaller one that provides a better starting situation.

In the second section the user may specify the instrument geometry and details used in the measurement. Three different radiations are available, X-ray, neutron and electron; for X-rays, one can select conventional tubes for which both $K\alpha_1$ and $K\alpha_2$ are supposed to be present, otherwise a strictly monochromatic radiation can be selected and the wavelength introduced. Three different geometries are available: Bragg-Brentano, Debye-Scherrer or Transmission. For Bragg-Brentano both possibilities of measurement with $\theta/2\theta$ or only 2θ are available. Finally we can specify the instrumental broadening, from very low as in a high resolution synchrotron beamline to very broad as in TEM. The possible choices are just broad categories as the main purpose of the program is to perform the search-match (and rapid quantification) and not an accurate line broadening analysis. A specification of an instrument broadening permits to work in the computation directly with only 2 parameters (crystallite size and r.m.s. microstrain) to refine for peak shapes and reduce the probability of solution divergence. If the instrument broadening specified is sufficiently close to the real one, the program will give also a good estimation of the crystallisation of the compound. For nanocrystallised samples, as the broadening is much larger than the instrument broadening (unless TEM is used), an error on the instrument broadening choice will not cause a big error in the size-strain parameters at end of the anal-

ysis.

Finally in the last section we specify which database to use. At the moment three choices are available: the full COD database or two subsets, the inorganic or the organic only. Actually, due to the impossibility to discern the organometallic structures from either the organic or inorganic structures, we have included them in both. So one subset is organic+organometallic and the other is inorganic+organometallic.

When the input is finished and proper options selected, pressing the "Search and quantify" button will start the computation by submitting the job to a local workstation selecting the structures to be used and launching the FPSM algorithm. Another web page will be presented at the end (like in Fig. 3) carrying the results in term of a table with the phases found, their percentage and their crystallite sizes and microstrains. Along the table a graph will illustrate the fitting of the pattern for visual evaluation.

Found phases and quantification:

Phase ID	name	vol. (%)	wt. (%)	crystallites (Å)	microstrain
9004180	Zincite	20.2448	28.8293	2067.89	0.000321462
1000043	Fluorite	42.5474	33.9434	2163.67	0.00037629
9007498	Corundum	37.2078	37.2273	1981.98	0.000294557

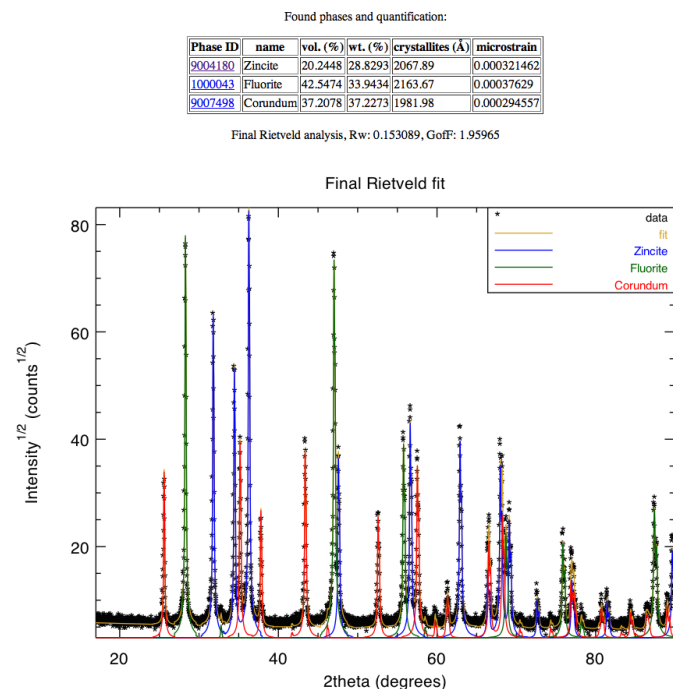Final Rietveld analysis, Rw: 0.153089, GofF: 1.95965

Figure 3
Results obtained after less than 10 secs using datafile cpd-1h provided as an example and composition as on Fig. 2. The cpd-1h diffraction datafile is the one coming from the quantitative phase analysis round-robin.

4. Discussion

Testing has been done using initially the dataset cpd-1h from the round robin on quantitative analysis and available from the CPD/IUCr website. This is a simple analysis with 3 phases (Corundum, Fluorite, Zincite) that was useful also for testing the accuracy on phase quantification. The tests focused on final result and speed of execution. We report the results using the web interface when using some composition restrictions and the standalone program without restriction. This because the

web base interface does not return a result when the computation becomes too long and the standalone program must be used instead. For this sample the analysis was identifying correctly all three phases both using no composition restriction and with restriction. The only difference between the two was the time to complete the analysis. On a workstation with 12 cores at 2.93GHz the computation time was 565 secs without composition restrictions but using the inorganic subset of the COD (which includes also all the organometallics that we are not able to separate automatically, total 174064 phases). When restricting the composition to only phases containing the following atoms: Al, Ca, F, Zn, O, Mg, Na, Si, Cl, the computation time dropped to 19 secs. If we specify only the atoms really present in the sample (Al, Ca, Zn, O, F) the total time fall below 10 secs, most of which is consumed to load the database and filter out the structures to use for the search. We could estimate from this sample that the program is able to perform an average of 30 full automatic Rietveld refinements per second, per thread.

In Fig. 3 the output from the web page is reported directly with the final plot and the agreement indices. The analysis results include also a final restricted Rietveld refinement to quantify the phase amounts and the size-strain from the sample. The size-strain results depend actually on the accuracy of the instrumental function chosen. For the web interface the instrumental broadening must be selected as the one more similar to the one of the sample and it is greatly approximated, but for the standalone program, the true instrumental broadening of the used instrument can be used. In table 1 we have reported the results of the quantitative analysis and compared with the expected results of the round robin. Even if the Rietveld analysis is performed using a lot of restrictions (only one overall B factor, no full refinement of the cell parameters, no structure refinement, very few parameters used and quick calculation in automatic, no absorption contrast and preferred orientation corrections) the results are quite satisfactory and inside the 2% absolute error. Looking at the general results of the round-robin (Le Meins *et al.*, 2002), our automatic analysis scored much better than the average participant result. In fact the standard deviation on the quantitative results for sample 1h is also reported in the table and is bigger than our error. In reality we think that all the restrictions actually help the automatic analysis to avoid possible errors arising from a wrong correction.

Table 1
Results of the FPSM analysis on sample cpd-1h from the quantitative analysis Round-Robin (R-R in the table) (Le Meins *et al.*, 2002). Column 4 contains the mean value from all participants in the Round Robin. In column 5 the mean standard deviation of the analyses performed by all participants using the CPD supplied data is reported for comparison. In column 6 and 7 the minimum and maximum values obtained by the participants is shown.

Phase	Wt.%	FPSM	Mean R-R	Std. Dev.	Min	Max
Corundum	35.12	37.23	35.96	2.71	30.30	43.83
Fluorite	34.69	33.94	35.21	2.40	30.33	40.30
Zincite	30.19	28.83	28.93	2.56	16.49	33.81

The first test was a simple one, mainly used to test the speed, feasibility and accuracy of the analysis. The second part of the tests was instead devoted to compare the method with the classical search-match routines. For this reason we used the datafiles

and tests provided by the search-match round robin and available from (Le Meins *et al.*, 2002). These tests are more tailored on a search-match based on the JCPDS-PDF database and well crystallised samples, so we do not expect our FPSM procedure to outperform the classical routines or even succeed on all of them. The principal reason is that FPSM in the tested version can only perform a search using phases for which the structure not only has been solved and refined, but also deposited in the COD database. Another reason is speed, we expect FPSM to be slower than classical methods in providing the results.

But first we aim to get much more refined results with our procedure, i.e. that the results will include a much more restricted number of phases, in particular only the ones effectively present without the need to further refine the list found; and also a better quantification is obtained through the use of the Rietveld method, where the classical search-match can only give a qualitative estimation. Second, this is performed in a fully automated way without human intervention, up to the final quantification.

Found phases and quantification:

Phase ID	name	vol. (%)	wt. (%)	crystallites (Å)	microstrain
9000098	Siderite	39.5067	45.9923	770.485	0.000874197
9002213	Hydroxylapatite	34.9531	32.5953	1000	0.0004
9010145	Quartz	1.97382	1.53981	1000	0.0004
9005624	Souzalite	12.452	11.1812	1000	0.0004
1011172	Quartz low	11.1144	8.9139	1000	0.0004

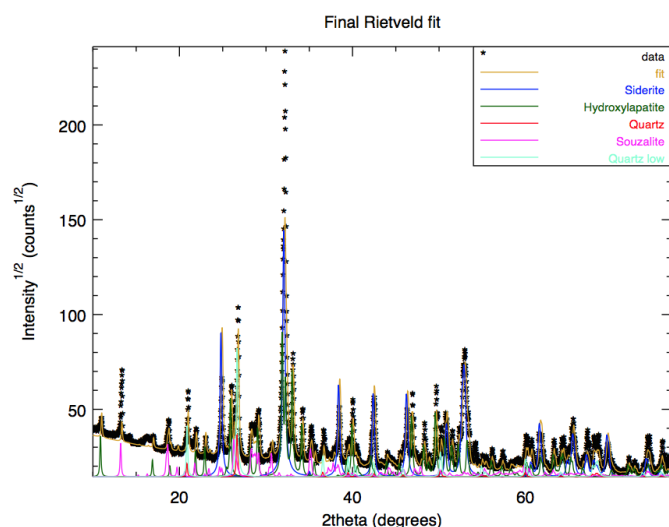Final Rietveld analysis, Rw: 0.317671, GofF: 10.5251

Figure 4
Results of the Full Profile Search Match on sample 1 of the search-match round robin (Le Meins *et al.*, 2002). Quartz and Quartz low have the same crystal structure. The execution time was 282 secs on the nanoair workstation (2x6 cores, 2.93GHz).

The round-robin was contemplating 4 samples and we used two modalities as it was done originally; in the first, no composition restriction are applied, while in the second we used the possible elemental composition given by the round-robin proposers to restrict the number of phases for the search. The analysis couldn't be completed for sample 2, because the only phase

contained has an unknown crystal structure. There is no similar structure in the COD and the phase sufficiently close is only reported in the JCPDS-PDF database as a card. The structure is described as a Silicon Oxide Quinuclidine Fluoride Octadecasil and we haven't found it even in the Cambridge Structural Database. So we will only report the test results on the other 3 samples: two inorganics and one organic.

Sample 1. This sample analysis was performed using the standalone version of the program for the unrestricted composition analysis and the web form for the restricted one. Both gave the same results so we focus on the web version. Fig. 4 is showing the results that are essentially in line with the round-robin solution. In FPSM two quartz structures have been found, but by a check with the COD web site using the link to the structures indicated that these are the same structure, with only some small differences in the cell parameters that probably accommodated better the fit. The main discrepancy in the fit visible in the figure is due to the souzalite structure which is not exactly the same as the one in sample 1. Souzalite, gormanite have a high variability in composition and structure. In the analysis reported, we limited the phases to no more than 5, otherwise FPSM will also find and add several other minor phases all around 1% that are indeed possible, as from the round-robin organiser comments, but we preferred to leave them out from the results.
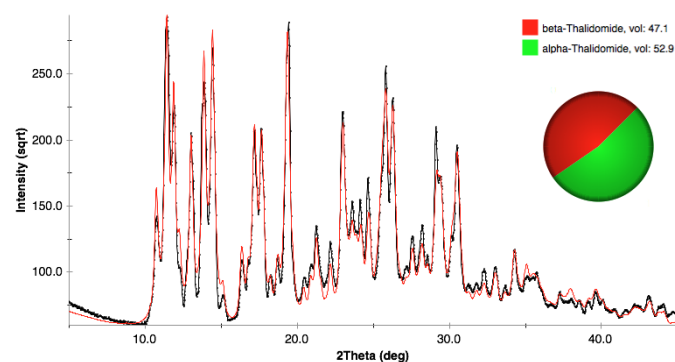

Figure 5
FPSM final fit (red line) of the search-match round-robin sample 3. Both polymorphs have been identified as the only present compounds and quantified at about 50% each as indicated in the caption. The fit this time belongs to the standalone program version of FPSM. Time for computation was 3307 secs. Alpha and beta thalidomide have been fitted with crystallite sizes of respectively 502 and 548Å, microstrain around 0.002.

Sample 3. This sample is an organic and by the round-robin results notes, it should be composed by about 50-50% of two polymorphs, alpha and beta Thalidomide. But the JCPDS-PDF only contains the alpha one and consequently an identification with only the alpha form was accepted, and nobody found the other in fact.

In our case we have both the alpha and the beta in the COD database and FPSM found both correctly. But in this case, even with the composition restrictions, we couldn't run it on the web interface as in the organic sqlite database there are nearly 40000 phases having as atoms C, H, N, O only or a subset of them. So

we only used the standalone version. FPSM used about 3300 seconds to find the solution on the usual 12 cores workstation with the composition restrictions and the final fitting is visible in Fig. 5. The time is relatively high, but if we consider that we went directly from the raw pattern to the final correct phase quantification is not so bad in absolute terms. How many analyses have we completed up to the Rietveld refinement in less than one hour starting from the pattern? And in this case FPSM was able to identify the second polymorph not present in the PDF database.
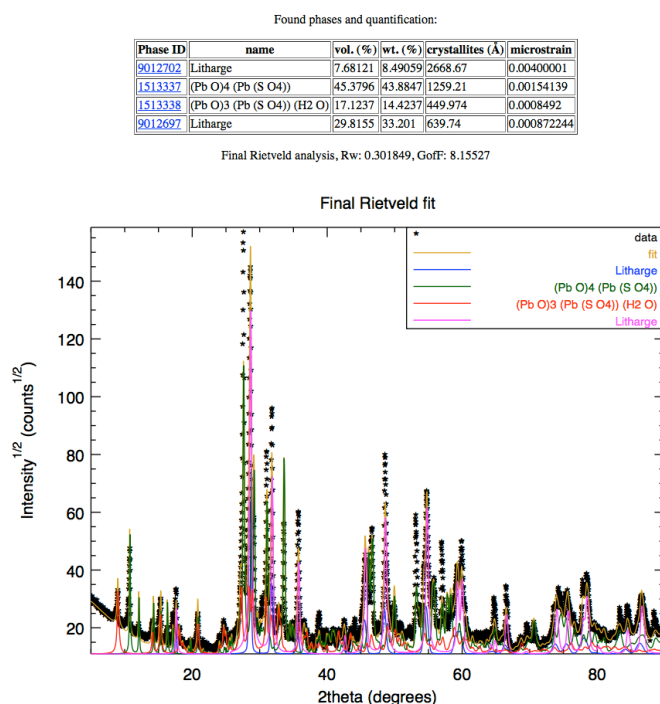
Found phases and quantification:

Phase ID	name	vol. (%)	wt. (%)	crystallites (Å)	microstrain
9012702	Litharge	7.68121	8.49059	2668.67	0.00400001
1513337	(Pb O)4 (Pb (S O4))	45.3796	43.8847	1259.21	0.00154139
1513338	(Pb O)3 (Pb (S O4)) (H2 O)	17.1237	14.4237	449.974	0.0008492
9012697	Litharge	29.8155	33.201	639.74	0.000872244

Final Rietveld analysis, Rw: 0.301849, GofF: 8.15527

Figure 6
Results of the FPSM analysis on round-robin sample 4. Two PbO phases are present with the same structure, one with large and one with small crystallites to account for a better fitting. Not having the real sample we do not know if the PbO does have something like a bimodal distribution of grain sizes that may explain the particular peak profile.

The fourth round-robin sample is the easiest one, being composed of only lead based compounds but it showed some particularities. We refer here directly to the analysis done by the web form using the restricted composition. The unrestricted gave similar results but in a much longer time. The first web analysis by FPSM was completed in 40 seconds, giving the correct phases but in addition several other compounds like sulphur, ice and sulfuric acid were found with small crystallites. These compounds were mainly used by the least squares routine to provide a better fitting of the background and diffuse modulations. Due to their low density they enter at the end of the evaluation despite the modified FoM of Eq. 5. To get rid of them we can use the option to limit the phases to the ones with density not lower than 2 (gr/cm^3) and then the total computation time also drops down to 19 secs. Final results in the last

case are reported in Fig. 6. One may note that FPSM has not found the massicot phase that the round-robin organisers found to be present. Indeed many participants have not found it. So we investigated the massicot presence a bit further by performing a Rietveld refinement with the found phases and including in addition the massicot structure. It turned out that the Rietveld always decreases the massicot quantity to zero. This is due to the fact that the peaks assigned to the massicot are also peaks of other present phases and the massicot has a medium-strong peak at 63° 2θ that is not present in the pattern. The organisers did not use the entire range at their disposal but only up to 62° in 2θ from their plot. We can safely conclude that the massicot, not identified by FPSM, is indeed not present in the sample. This is actually an advantage of the present procedure, because using the pattern fitting it may better evaluate the effective presence of a phase and not only base the identification on the presence of the peaks in a certain range.

In addition to the standard and round-robin tests we have analysed other samples and we will report here a couple of extreme cases. We anticipated that the FPSM procedure may be ideally suited for the case of nanomaterials where the large broadening may cause some difficulties to the traditional procedures as it will be more difficult to identify the peaks, especially without human intervention. One example (TiO$_2$ rutile nanoparticles) is already reported in a paper dealing with TEM diffraction rings analysis (Boullay et al., 2014) and we will not duplicate it here.
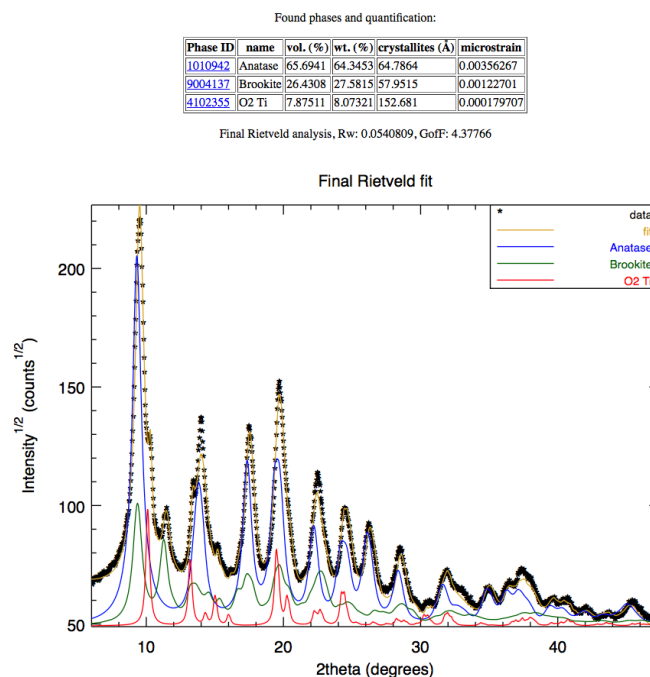
Found phases and quantification:

Phase ID	name	vol. (%)	wt. (%)	crystallites (Å)	microstrain
1010942	Anatase	65.6941	64.3453	64.7864	0.00356267
9004137	Brookite	26.4308	27.5815	57.9515	0.00122701
4102355	O2 Ti	7.87511	8.07321	152.681	0.000179707

Final Rietveld analysis, Rw: 0.0540809, GofF: 4.37766

Figure 7
Results of the FPSM analysis for a nanocrystalline titanium oxide sample. All the three polymorphs have been found. Pattern collected with an INEL CPS 120° detector and Ag source.

Instead we have used another TiO$_2$ sample prepared in dif-

ferent conditions (Ceccato *et al.*, 2003) containing more than one polymorph. The sample has been measured in a capillary using silver radiation and a INEL CPS 120° detector to reach a large range in Q space for PDF (Pair Distribution Fucntion) like analysis. We tested it with the FPSM as if it is easy to identify the anatase phase (nearly 2/3 of the sample), the rutile and brookite are not by an automatic procedure. Using a traditional search-match and even restricting the composition to only Ti and O, the two minor phases are ranked low, as only one peak for each may be discernible. Using FPSM we also need to restrict the composition to Ti and O, as the large broadening otherwise made it possible to other phases (apart the anatase that it is always identified) to enter the fit before the brookite and rutile. We used only a range in 2θ up to 40° to avoid a longer computation time. It already corresponds to a 5.8Å^{-1} in Q or the equivalent of more than 90° 2θ using $CuK\alpha$ radiation. The raw results from the Rietveld search-match routine are visible in Fig. 7. All the three polymorphs have been correctly identified and the quantification is not far from the best fit obtained using the full range (Q up to 16Å^{-1}), an anisotropic model for the size-strain broadening and full refinement of the crystal structures (full results to be presented elsewhere). FPSM on the nanoair workstation took 20 seconds to the final result.
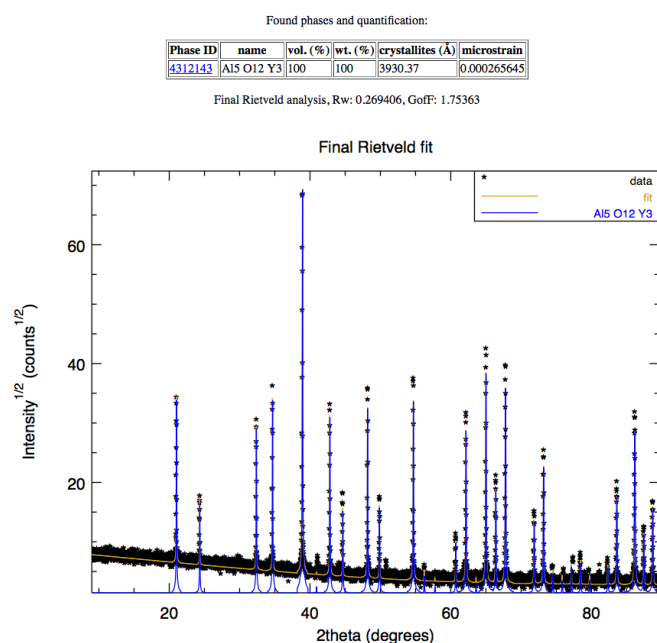
Found phases and quantification:

Phase ID	name	vol. (%)	wt. (%)	crystallites (Å)	microstrain
4312143	Al5 O12 Y3	100	100	3930.37	0.000265645

Final Rietveld analysis, Rw: 0.269406, GofF: 1.75363

Figure 8
Results of the FPSM analysis for a highly crystallised garnet sample. The total filtering and search time was 1 second when restricting composition to only phases with Al, Y and O atoms. Pattern collected with an INEL CPS 590° detector, $CoK\alpha_1$ only source. In the FPSM web page the "Synchrotron" instrument broadening was selected and "high" for crystallisation.

Now, if FPSM may be well suited for nanocrystalline samples, we where questioning ourselves about the limits when we have a highly crystallised sample on a high resolution instrument. In such case, peaks are becoming more similar to delta functions and the profile is less significant with respect to pure reflection positions. Small differences in cell parameters or instrument misalignments may be critical for the procedure presented here.

But surprisingly all the test we have done came to the conclusion that there are no drawback with sharp peaks. The full profile search-match is much quicker and identification easier. As a last example we present the test made on a Yttrium Aluminum Garnet for which the possible atoms were Al, Y, Ca and may be O. The pattern was collected with $CoK\alpha_1$ radiation (using a Johansson Ge111 crystal monochromator), a capillary and this time an INEL 500 mm radius CPS 590° detector to obtain reasonable intensities in a relatively short time (90 minutes) but very sharp reflections. Fig. 8 reports the final results obtained through the web interface. As options we selected "high" for the crystallisation of the sample and "Synchrotron" for the instrument broadening. The entire search and quantification, using the 4 atoms compositions restriction, completed in 1 second.

5. Conclusions

We have described a new method to perform the search-match that is based on the full profile and on the Rietveld algorithm. The method is promising and well suited for nanosized particles or nanomaterials but it works also for well crystallised samples. The method relies heavily on brute force computing and databases of crystal structures. In the present work we used the COD database, but potentially any other database of structures can be used.

The main drawback of the method, at the present time, is the computation time that is indeed much longer than usual search-match programs, unless some composition restriction can be used. But considering that the method provides also a Rietveld quantification through pattern refinement, the total time should be compared to the total time needed to complete not only a search-match but the Rietveld quantification. No human intervention is required during the entire time, and so its amount can also be considered not so problematic. Indeed we have tested it only on a single desktop computer, but for a heavy use it will not be difficult to port the system over a cluster to perform even better calculations and improve its capabilities to detect minor phases or in the case of complex nanocrystalline samples.

One bigger problem arises from the completeness of the database used. If the JCPDS-PDF database is considered quite complete, at least for the inorganic case, including both the classical experimental reflections and the ones calculated from the crystal structure databases, not so can be said for the crystal structure databases, especially the free ones like the COD. Using a more complete database like the ICSD for inorganics or the CSD for organics may well improve its performance, but it is only theoretic up to now.

In the future, as the speed of computers is growing every day, and databases as the COD will become more complete, these initial concerns should not be a problem anymore.

research papers

References

Baerlocher, C., McCusker, L. B. & Olson, D. H. (2007). *Atlas of Zeolite Framework Types*. Elsevier Netherlands.

Boullay, P., Lutterotti, L., Chateigner, D. & Sicard, L. (2014). *Submitted to Acta Cryst. A*.

Caceci, M. & Cacheris, W. (1984). *Byte*, **May**, 340–362.

Ceccato, R., Diré, S. & Lutterotti, L. (2003). *Journal of Non-Crystalline Solids*, **322**(1-3), 22–28.

De Keijser, T. H., Langford, J. I., Mittemeijer, E. J. & Vogels, A. B. P. (1982). *J. Appl. Cryst.* **15**, 308–314.

De Keijser, T. H., Mittemeijer, E. J. & Rozendaal, H. C. F. (1983). *J. Appl. Cryst.* **16**, 309–316.

Delhez, R., De Keijser, T. H., Langford, J. I., Louër, D., Mittemeijer, E. J. & Sonneveld, E. J. (1983). In *The Rietveld Method*, edited by R. A. Young, vol. 5 of *IUCr Monograph*, pp. 132–166. Oxford University Press.

Dollase, W. (1986). *J. Appl. Cryst.* **19**, 267–272.

Downs, R. T. & Hall-Wallace, M. (2003). *American Mineralogist*, **88**, 247–250.

Enzo, S., Fagherazzi, G., Benedetti, A. & Polizzi, S. (1988). *J. Appl. Cryst.* **21**, 536–542.

Ferrari, M. & Lutterotti, L. (1994). *J. Appl. Phys.* **76**(11), 7246–7255.

Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P. & Le Bail, A. (2009). *J. Appl. Cryst.* **42**, 1–4.

Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirós, M., Serebryanaya, N. R., Moeck, P., Downs, R. T. & Le Bail, A. (2012). *Nucleic Acids Research*, **40**, D420–D427.

Grosse-Kunstleve, R. W. & Adams, P. D. (2002). *Acta Cryst.* A**58**, 60–65.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.

Hanawalt, J. D. (1986). *Powder Diffr.* **1**(1), 7–13.

Harrison, W. T. A., Simpson, J. & Weil, M. (2010). *Acta Cryst.* E**66**, e1–e2.

Langford, J. I. & Louër, D. (1996). *Rep. Prog. Phys.* **59**, 131–234.

Le Meins, J.-M., Cranswick, L. M. D. & Le Bail, A., (2002). Search-Match Round Robin. http://www.cristal.org/smrr/.

Lutterotti, L. (2010). *Nuclear Inst. and Methods in Physics Research B*, **268**, 334–340.

Lutterotti, L. (2012). In *Uniting Electron Crystallography and Powder Diffraction*, edited by U. Kolb, K. Shankland, L. Meshi, A. Avilov & W. I. David, NATO Science for Peace and Security Series B: Physics and Biophysics, pp. 233–242. Springer Netherlands.

Lutterotti, L., Matthies, S., Wenk, H.-R., Schultz, A. J. & Richardson, J. W. (1997). *J. Appl. Phys.* **81**(2), 594–600.

Lutterotti, L., Pillière, H., Fontugne, C. & Driss, H. (2012). In *European Powder Diffraction International Conference 13*. Grenoble, France.

Lutterotti, L. & Scardi, P. (1990). *J. Appl. Cryst.* **23**, 246–252.

Nanoair, (2009). Was granted by REA FP7 (Grant agreement N° 222333).

Peng, L.-M., Ren, G., Dudarev, S. L. & Whelan, M. J. (1996). *Acta Cryst.* A**52**, 257–276.

Popa, N. C. (1992). *J. Appl. Cryst.* **25**, 611–616.

Prince, E. (1983). In *The Rietveld Method*, edited by R. A. Young, vol. 5 of *IUCr Monograph*, pp. 43–54. Oxford University Press.

Rietveld, H. M. (1967). *Acta Cryst.* **22**, 151–152.

Rietveld, H. M. (1969). *J. Appl. Cryst.* **2**, 65–71.

Scardi, P., Lutterotti, L. & Maistrelli, P. (1994). *Powder Diffr.* **9**(3), 180–186.

Smith, D. K. & Gorter, S. (1992). *J. Appl. Cryst.* **24**, 369–402.

Wenk, H.-R., Matthies, S. & Lutterotti, L. (1994). *Mat. Sci. Forum*, **157-162**, 473–480.