hIT$_e$Q: A New Workflow-Based Computing Environment for Streamlining Discovery. Application in Material Science

Laurent A. Baumes*[*], *Santiago Jimenez, and Avelino Corma

Instituto de Tecnologia Quimica, UPV-CSIC, Universidad Politecnica de Valencia,

Avda de los Naranjos s/n, 46022, Valencia, Spain.

Fax: +34 9638 7789; Tel: +34 9638 77800; e-mail: baumesl@itq.upv.es

Abstract. This manuscript presents the implementation of the recent methodology called Adaptable Time Warping (ATW) for the automatic identification of mixture of crystallographic phases from powder X-ray diffraction data, inside the framework of a new integrative platform named hIT$_e$Q. The methodology is encapsulated into a so-called workflow, and we explore the benefits of such an environment for streamlining discovery in R&D. Beside the fact that ATW successfully identifies and classifies crystalline phases from powder XRD for the very complicated case of zeolite ITQ-33 for which has been employed a high-throughput synthesis process, we stress on the numerous difficulties encountered by academic laboratories and companies when facing the integration of new software or techniques. It is shown how an integrative approach provides a real asset in terms of cost, efficiency, and speed due to a unique environment that supports well-defined and reusable processes, improves knowledge management, and handles properly multi-disciplinary teamwork, and disparate data structures and protocols.

Keywords: *Integrative Platform, Workflow, Streamlining Process, Data Mining, Materials, Diffractograms*

1. Introduction

The increasing pressure on chemical industries for enhanced operational efficiency is a persuasive driver for innovation, and new technologies receive a stronger attention by companies and academic laboratories due to their potential impact on current research and development (R&D). On the other hand, the absorption of new techniques and methodologies is not straightforward. Thus, the effectiveness of workflow technology to acquire, analyze, store and integrate disparate apparatus, software, and data in materials science is examined. Such a problematic turns into a real challenge when combined with high throughput (HT) experimentation due to *i*) the corresponding reliance on informatics, and the overwhelming amount of data, *ii*) the intricacy in describing a solid catalyst,[1] and the corresponding use of new computational methodologies since previously developed ones within the life sciences are not flawlessly transferable, *iii*) the necessity to cope with various data structure and exchange protocols because very few standards[2] are available. For all these reasons, a reliable and exhaustive capture of related data and processes remains extremely tough and requires tailored tools. Together with Inforsense©[3], the Instituto de Tecnologia Quimica[4] (CSIC[5]-UPV[6]) has dedicated important resources inside TopCombi[7] ("Towards Optimized Chemical Processes and New Materials by Combinatorial Science"), a 5 years integrated project of the 6th PCRD in Europe, started in March 2005 with a budget of 23 M€, and composed of 22 partners, for the harmonization and integration of this information. hIT$_e$Q, pronounce "high tech", is the resulting integrative platform created by the Instituto de Tecnologia Quimica (ITQ) which encapsulates all new development done in ITQ for R&D in materials science and other chemistry fields. As an example, this manuscript presents the implementation of the recent methodology called Adaptable Time Warping[8] (ATW) for the automatic identification of mixture of crystallographic phases from powder X-ray diffraction data, inside the framework of the platform. The methodology is encapsulated into a so-called workflow, and we explore the benefits of such an environment for streamlining discovery research.

Beside the fact that ATW successfully identifies and classifies crystalline phases from powder XRD for the very complicated case of zeolite ITQ-33 high-throughput synthesis process, we stress on the numerous difficulties encountered by both academic laboratories and companies when facing the absorption of new techniques. It is shown how an integrative approach provides a real asset in terms of cost, efficiency, and speed due to its environment that supports well-defined and reusable processes, improves knowledge management, and handles properly multi-disciplinary teamwork, and disparate data structures and protocols. ATW is chosen as application because of zeolites are versatile materials for an increasing number of applications,[9] including catalysis; and the discovery

of new structures or enlarging the synthesis space, and optimization of existing ones require a considerable experimental effort which can be diminished by using high throughput (HT) technology (usually parallelization and miniaturization)[10] as also done for heterogeneous catalysts.[11] The increase of the amount of experiments implies the data to be automatically analyzed not to slow down the whole process, see Figure 1. One of the challenging tasks is the automatic determination of the crystalline phases contained in each new sample from its powder diffraction data (XRD) because an adapted methodology is still lacking, considering that a specific structure can present differences in the XRD diffractogram, for both the intensity of peaks and the 2θ diffraction angles, depending on its level of crystallinity and its chemical composition.

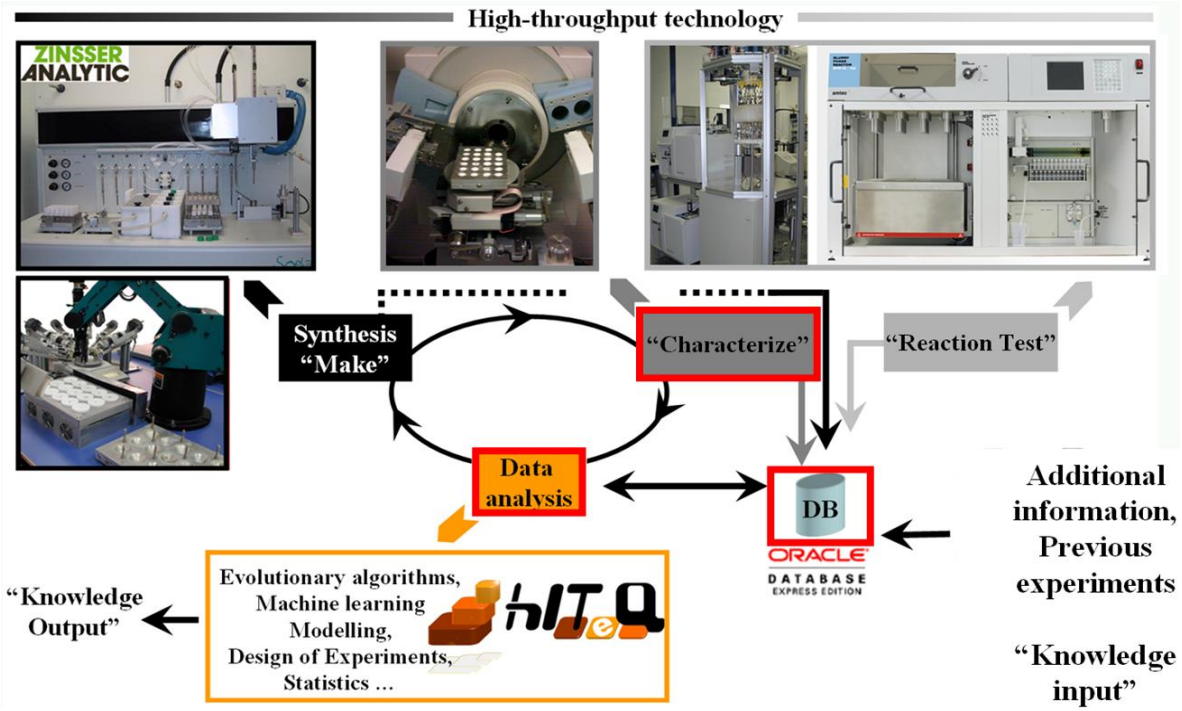

Figure 1. *The presented application encapsulated into the integrative platform hIT$_e$Q aims at directly treating characterization data, i.e. X-ray diffractograms while handling database connection and high-throughput apparatus file formats.*

Firstly, a general description of integrative platforms is given. Then, hIT$_e$Q functionalities are described and the ATW methodology is detailed through the examination of the supporting workflow. It is shown how versatile such kind of solution is. Finally, the corresponding benefits are clearly demonstrated.

2. Integrative and workflow-based platforms

Numerous papers denote new problems and new strategies in various aspects of materials science. Obviously, there is no software that handles all these problems, and therefore it is normal that research centres and companies rely on several software systems with overlap functions. The lack of an integrated architecture motivates the drive to search for new solutions. Workflow technology is a mechanism to integrate data, applications and services, enabling scientists to dynamically construct their own research protocol for scientific analytics and decision making by connecting various resources and software applications together in an innovative way. The principal objectives are the following: 1) to eliminate both the time spent to transfer data from one system to another one, and the related problems of data formats and conversion; 2) to integrate existing tools and to communicate between these different systems; 3) to merge treatments and analysis usually performed by various specialists in different locations by handling distributed or remote functionalities allowing an effective and fast collaborative work; 4) to capture the knowledge of the whole business model through workflows conception; 5) to unify and consolidate data in an homogeneous way making the use of databases easier while opening the way toward broader analysis or data mining; 6) and to retrieve and publish results effortlessly. Ignoring these issues typically results in delays in getting research results, the data is prone to getting lost and error prone, the discovery process is often extended, effort duplication. The ITQ provides an integrative analysis platform, hIT$_e$Q which is a workflow-based computing environment, permitting various tools to be integrated. Will show how the introduction of new technologies that create agility on top of less flexible environments.

A general description of workflow platforms and existing efforts are first presented. Then, a closer look at technologies and standards such as XML and Web services allows explaining how workflows are designed, managed and executed. Finally, we focus on some of the hITeQ functionalities.

2.1. Overview and general description

Data and processes flow within a single environment where each so-called “task” or “node” performs a set of operations. Data processing is constructed from connections of elementary functionalities contained into the tasks which must be dragged and dropped onto the “workspace” and then be linked to form the desired entire data-processes. After execution of a task, results are passed to the connected ones. Workflow can be described as a process, in whole or part, during which information or data are passed from one node to another for action, according to a set of

procedural rules, i.e. the data analysis is broken down to separate standard processes, which are abstracted to separate nodes. A general workflow node is basically defined by the following parameters: meta-data (input and output), algorithm, and user parameters. If the workflow platform is flexible enough, it can be easily and fully customized allowing users without programming knowledge to match with their requirements, while programmers take advantages of previously coded functionalities, and the ease of adding new ones. Thus, the workflow mode facilitates the re-use of data models and algorithms: one routine being capable of working on different specified data sets and one data set being able to fit different routines. This is the main reason that workflow technology is being increasingly applied in discovery informatics to organize and analyze data.

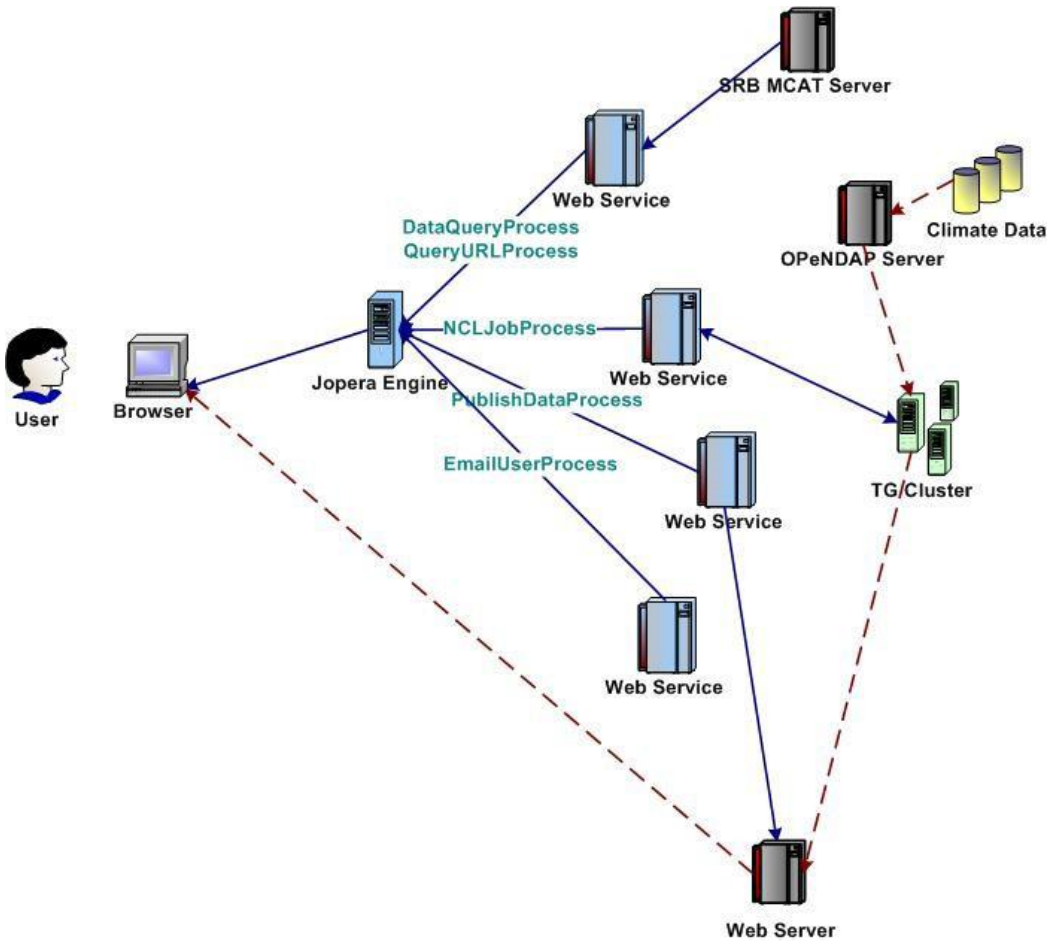

Figure 2. *Mapping the workflow to distributed physical infrastructure in JOpera.*

Workflow technology is generic so analytics workflow can be built for any areas like gene expression analysis, sequence analysis, proteomics, system biology and so on. Here, we quickly review different platforms that allow scientists to construct and execute workflows using

components that encapsulate many cheminformatics- and bioinformatics-based algorithms. For example, SciTegic's[12] Pipeline Pilot, InforSense[3] KDE, BioLog,[13] Vision,[14] KNIME[15] are some of the chemically intelligent implementations of the workflow technology. Other platforms have focussed only on bioinformatics, some existing efforts including Biopipe,[16] BioWBI,[17] Taverna,[18,19] etc. All of them provide mechanisms to integrate bioinformatics programs into workflows. Biopipe is based on programming language perl, and a user-friendly interface for building workflow is lacking so far. BioWBI and Tarverna use web-services for components to construct workflows. However, to convert a 3rd-party program into web-services, they lack of integrative GUI environment. Wildfire,[20] aims at using workflow to provide huge computing capability to bioinformatics application. However, there is no integrative environment provided for multiple users to collaborate in the same large-scale bioinformatics project, see also JOpera [21] in Figure 2 for distributed calculations. A non-exhaustive list focussed on bioinformatics should also include Science Factory,[22] Amadea Biopack[23] (Figure 3), Ptolemy II,[24] Kepler[25] (Figure 4), MIGenAS,[26] INCOGEN[27] (Figure 5), GeneBeans,[28] and IBM's WsBAW.[29] Readers interested on neuroscience application are referred to Data-MEAns,[30] Spike,[31] MEA-Platform.[32]

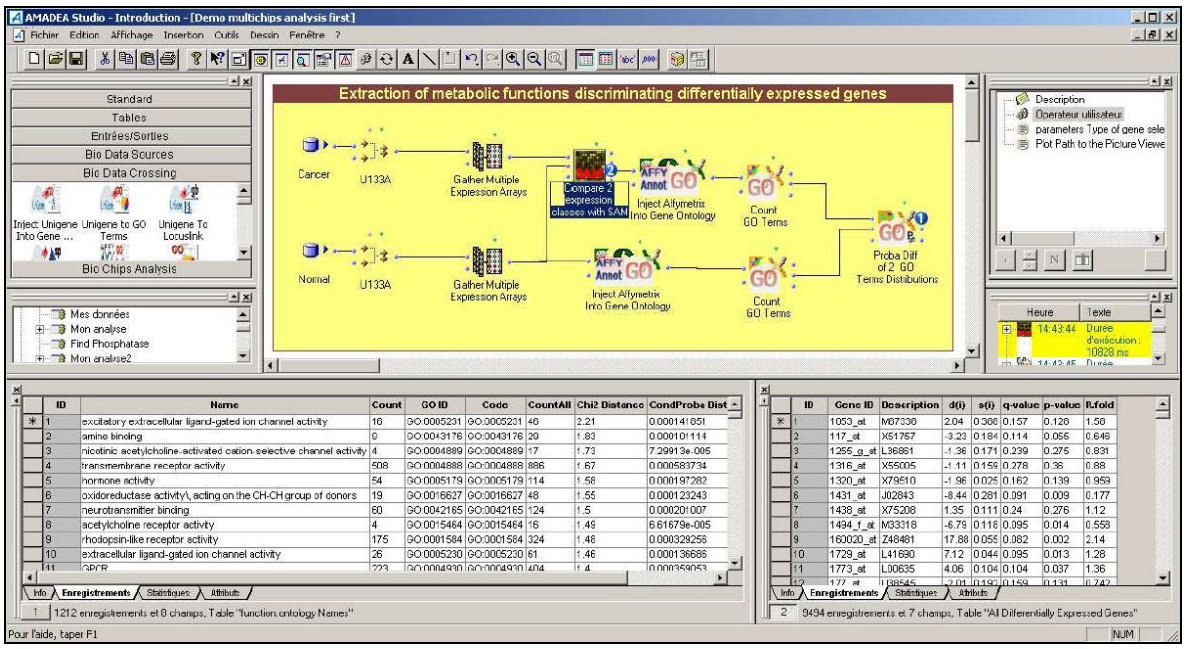

Figure 3. *The Amadea Biopack Workflow Development Studio*

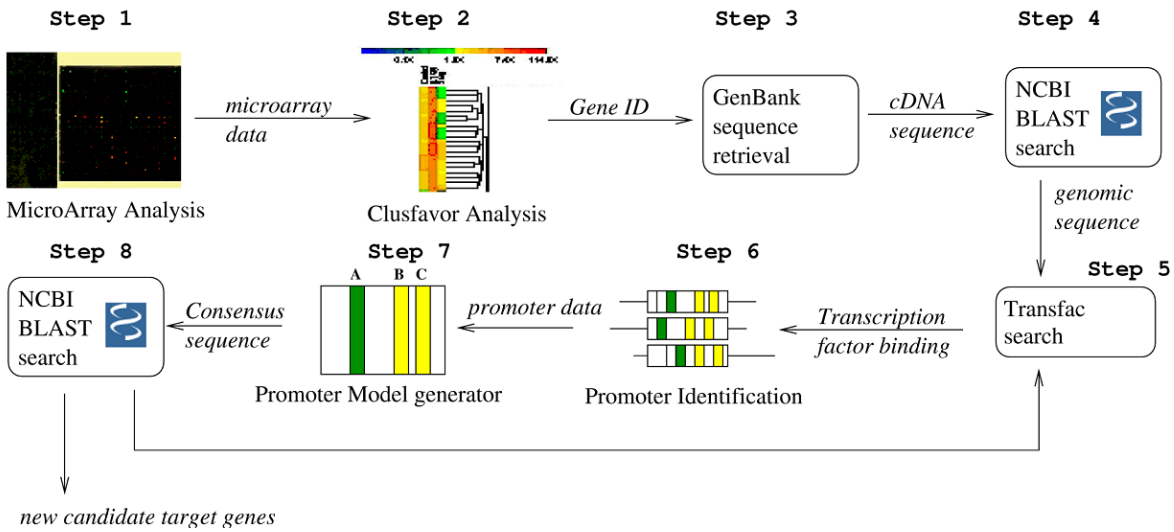

Figure 4. *The Kepler platform*

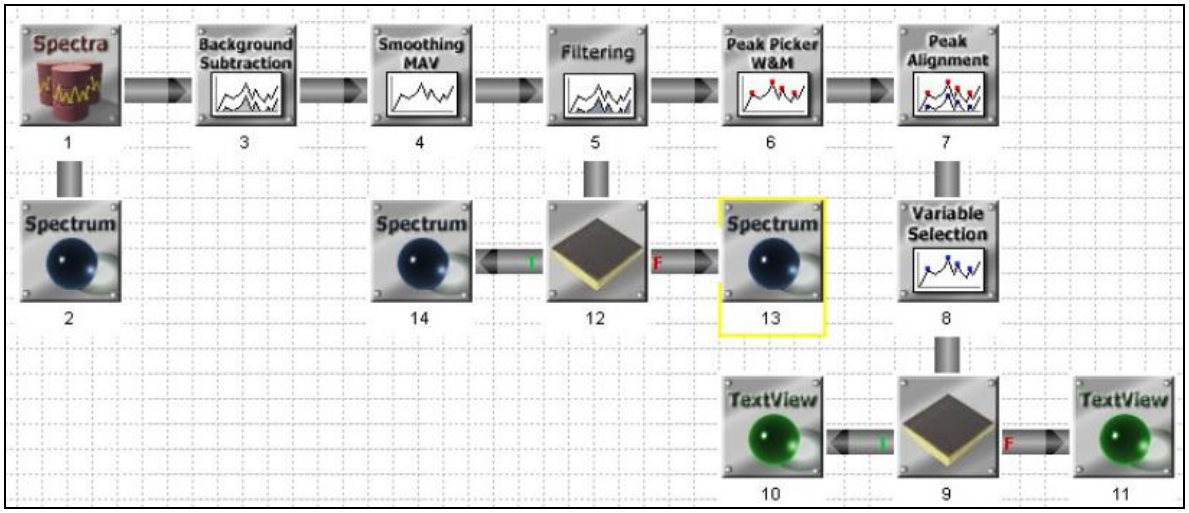

Figure 5. *The Workflow System for Mass Spectrometry in Cancer Diagnosis Using the INCOGEN VIBE Software*

2.2. How it works?

From the viewpoint of informatics, the requirements of an integrative informatics platform consist of four parts: integration of data, algorithms, computing hardware, and human interaction. hIT$_e$Q platform is a workflow system which defines, manages, and executes workflow processes. The data model and algorithm interface are designed independently, so it is more effectively to integrate or modify corresponding data sets and routines. In order to integrate different tools, hIT$_e$Q, like most

of the previously cited platforms, uses two principal strategies. The first one refers to the principle of embedding, and the second one to web services.

The problem of integration is complex in that programs are implemented using a variety of programming languages, on different operating systems, operated in different ways using input and output data in a wide range of formats. The second option to facilitate research through integration uses the principles of embedding with a common input–output data format. The input, for example a table, is exported in the right format for the tool to embed, then the tool is executed with an automatically computed command-line, *e.g.* a very simple node, and finally, the files produced by this execution are imported as result tables. hIT$_e$Q embedding toolkit is sufficient to embed the majority of external tools. As an example, the data-mining program Weka[33] and the chemistry library CDK[34] written in Java have been totally integrated into hIT$_e$Q. Embedding 3rd part components is interesting considering external software which permit adequate compatibility and interoperability options effort which make the re-implementation of all functionalities unnecessary. As mentioned earlier, data exchange between hIT$_e$Q and external executable programs can be implemented by creating a new instance dynamically of the program. As such, distributed (local or remote) algorithm/function can be incorporated as a pluggable component (a node). This has been intensively used in order to run Fortran code for crystallographic calculation on clusters inside ITQ. However, other ways are possible such as the following: using DDE (Dynamic Data Exchange), or performing calls to libraries such as to DDLs, or calling a remote process (RPC) using one of the existing protocols (RMI, CORBA, ...).

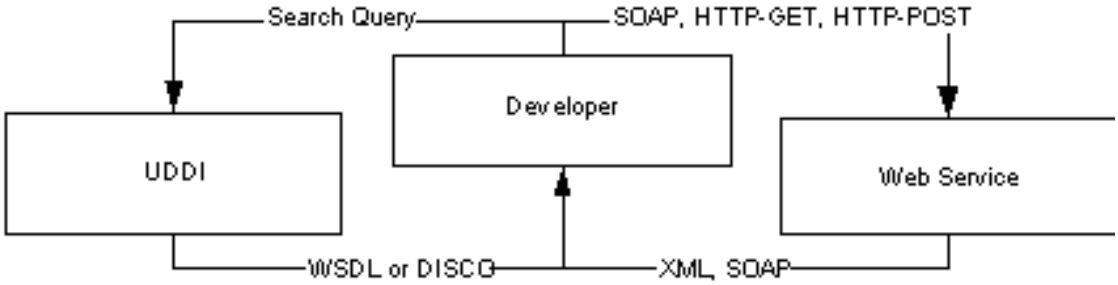

Figure 6. *Computer networking and web services. By allowing data interchange in the standard XML format, anybody can pick up the data and use it. Web services use a standard protocol called SOAP to transfer messages over HTTP. SOAP makes it possible for applications written in different languages running on different platforms to make remote procedure calls effectively. The Web Service Description Language (WSDL) is an XML-based, i.e. both machine and human-readable, grammar for describing web services, their functions, parameters, and return values.*

Microsoft has invested in an initiative called .net. Though it evolved differently, Sun Microsystems has created a similar technology base through their "J2EE" initiative. Both efforts relate to a new area of computer networking known as web services. .net contains some interesting networking technology that hIT$_e$Q employed. When dealing with multi-platform interoperability, the use of new languages and common standards like XML and SOAP (see Figure 6) enables disparate systems to communicate with each other. Therefore, such technological advancements, more specifically web services, have become the new rallying cries of most integrative platform vendors.

A web service is a standard platform for building interoperable distributed applications. hIT$_e$Q provides a framework that makes building web services easy. Its intuitive wizard allows us to create web services with ASP.NET without any skills in programming. Strong from its intensive use of .net, rather than creating an actual WSDL file, hIT$_e$Q generates the WSDL information dynamically, which is then returned to the client and displayed in the web browser. As the file is generated on request, clients can be sure that the WSDL contains the most current information.

"Tasks" or "Nodes" that performs calls to Web Services can be considered as a variant embedding 3rd part component, with the addition that ensures some features: 1) Time and resources for calculation are used from external machines potentially more powerful than clients' ones. 2) Only input parameters and output responses types are necessary to perform a call to a Web Service. 3) Web Services improves the independence between organizations while allowing collaborative work. For example, considering that a first organization has a Web Service that offers descriptors for molecules taking as input the CAS number, a second one can employ a node that performs a call to the Web Service and get the result. The information exchange is done without software compatibility problems and with little expense in time and resources.

2.3. I/O, Parameters, GUI

Integrative platforms reflect the up-to-the-minute progress in research as a consequence of the lack of standardization. However, hIT$_e$Q makes a strong use of XML, e.g. for integration purposes, a XML file descriptor encapsulating the knowledge about each algorithm properties and features (name of its parameters, allowed values for them, etc.) has been setup. Such an approach, already used in previous solutions,[35] is based on an advanced conception where XML is used as a fundamental basis for modelling processes and workflows in a unified and integrated environment.

Such platforms enable easy plug-and-play of components, ease of integration, fast reconfiguration of processes.

Another important trait of workflow platforms is the consistency of the workflow and the way this concept is handled. hIT$_e$Q provides a mechanism for metadata processing that executes before the workflow operates on the actual data. Since metadata provides extra information, the data controls such as logical constraints can be implemented, for example for the verification of the compatibility of a particular algorithm to a given dataset such as the use of a classifier for data without any classes. Table metadata specifies the column names and types, and hIT$_e$Q includes various types which are not restricted to basic types shared by most programming languages, such as matrix, molecule, crystal, graph, charts, images, etc…, but also sub-types such as XRD which belongs to series type with special features. Numerous types have been pre-defined in the system and are extendable. They allow a better control and the use of native functions, *e.g.* a matrix differs from a data table from the view point of mathematical functions such as determinant or inverse which is not always possible with the table since alpha-numerical data are allowed. The early integration of types permits the creation of algorithms on a common basis, *i.e.* data structure, even by various programmers. Nodes can be plugged together only if the output of one, previous node represents the mandatory input requirement of the following node. Thus, the essential description of a node comprises only input and output that are described fully in terms of data types. Once the workflow conception has been verified by the core engine, the execution may be launched generally starting from import nodes while export nodes or saving are usually the end points. However, other nodes may stand as input such as an "event", for example the creation of a new file by a given apparatus which automatically launches calculations or storing functions.

The core implementation of .net includes C#, pronounce C Sharp, a programming language very close to JAVA syntax, and the Common Language Runtime (CLR) for support of other programming languages. The system development environment, Visual Studio .net, is the tool used to build hIT$_e$Q written in C#. The principal reasons of this choice are: *i*) to provide a consistent object-oriented application, *ii*) to simplify deployment and versioning, *iii*) to provide a code execution environment that eliminates the issues faced by scripted environments with respect to performance, and *iv*) to provide a common programming model where the choice of a programming language becomes a matter of choice.

2.4. Some integrated functionalities

Repetitive tasks - Typical common exigencies of most scientists are: the possibility of executing the same algorithm on different input files through a single submission; executing the same algorithm on the same input letting the value of a parameter change at each run; executing in an appropriate order analysis processes consisting of more than one algorithm; choosing the algorithm to execute on the basis of the input data type; suspending the execution of an analysis process in order to verify the effectiveness of the partial results or for tuning different parameters while the process proceeds.

Loops – Related to "repetitive tasks", another particularity is that iterative treatments can be designed in the workplace by producing circuit of bricks. Few software available on the market permits this special requirement which is inherent to combinatorial and high throughput approach. For a direct use of hIT_eQ for real experiments (*i.e.* no benchmarks) in laboratories, each cycle of an iterative treatment will wait the user for typing, loading new data or directly catch new experimental values from the updated DB. Considering evolutionary algorithms (EAs), there is an unlimited number of configurations possible. The actual algorithms themselves are specified by a reproductive plan which describes how to get the first generation, and thenceforth how to get the new generation from the old. To be useful, a framework for EAs, on which researchers can hang ideas and quickly implement them, must show a balance. The framework has to be general enough to cope with arbitrary genome representations, but at the same time it needs to be simple to implement a new plan. This balance is also shown by a general framework for GAs developed by Russo[36] and extended by Jones[37].

Monitoring - Workflow monitoring is an important feature of our system. Scientific data driven workflows typically involve data processing tasks that could take a long period of time, coupled with the possible waiting time on execution queues. Thus, users would like to be aware of the status of execution and the stage at which the process is currently at. A web interface is automatically updated specifying all executed and under execution actions, results, etc…

Immediate publication on a web server - Once a data analysis workflow has been designed, it is possible to make it immediately accessible through web pages, e.g. on an http server, allowing people who do not need nor want to understand the technical details of the workflow, to have an interactive access to results but also to introduce new experimental conditions.

Plug-in - Plug-ins are the pieces of software independent of the hIT_eQ Core, and that can contain groups of "tasks" or "actions". The advantages of plug-ins are: 1) code independence from the

platform core, 2) sharing without giving sources, 3) organization "*à la carte*": users simply select the plug-ins they need and algorithm objectives are clearly separated and organized, e.g. crystallography, Design of Experiments (DoE), Evolutionary Algorithms (EAs), etc..., and 4) all code contained into plug-ins can be executed without recompiling the whole software.

3. Case study: Automatic analysis of X-ray diffraction for crystallographic phase's recognition

3.1. Introduction

Few solutions that aim at identifying crystalline materials from the analysis of powder X-ray diffraction (XRD) data have been reported so far.[38,39,40] A careful inspection of the latter and the corresponding highlight of specific failures when used for the discrimination of crystallographic phases of zeolites among mixtures, has allowed the creation of the recently proposed strategy called Adaptable Time Warping (ATW).[41] As the use of high-throughput techniques for the discovery or the synthesis space enlargement of new microporous crystalline structures makes the trustworthiness of search-match methods a critical point to be assessed, a meticulous evaluation of the reliability and the robustness is of great importance. Will show how such methodology is integrated in hIT$_e$Q, and we focus on the benefits of the platform in terms of time and efforts for the evaluation of the methods with numerous datasets, parameters, parameters variations, etc…

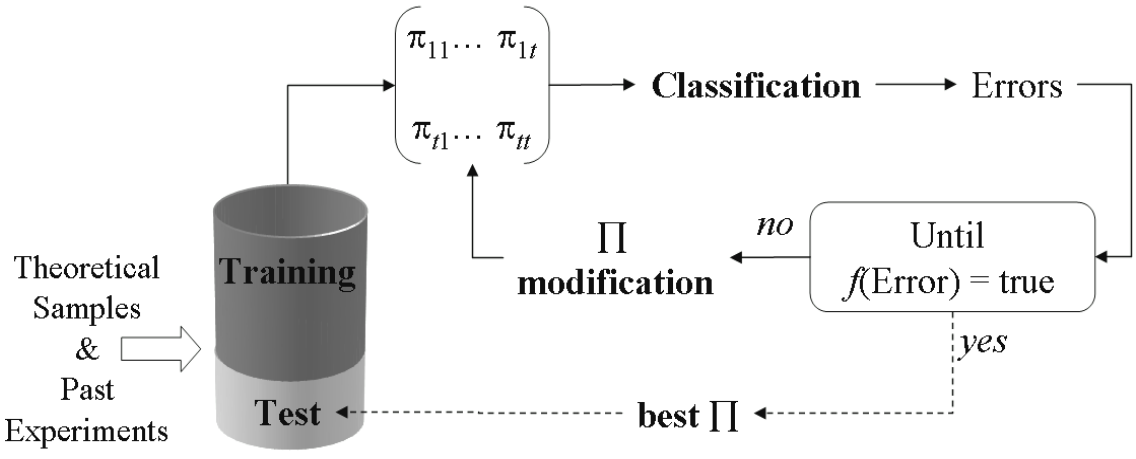

***Figure* 7**. *ATW conception. Based on classification errors of training samples, the $\prod$ matrix is modified until reaching a given termination criterion, i.e. the convergence to a minimum. Finally the best matrix is used for unseen test samples.*

The basic idea of ATW is the careful inspection of XRD data from past or theoretical experiments in order to modify the distance measure (using a Pi matrix) between diffractograms. Once trained, the measure is able to better associate XRDs with their corresponding phases due to the previous recognition of their particular traits into the diffractograms, see Figure 7.

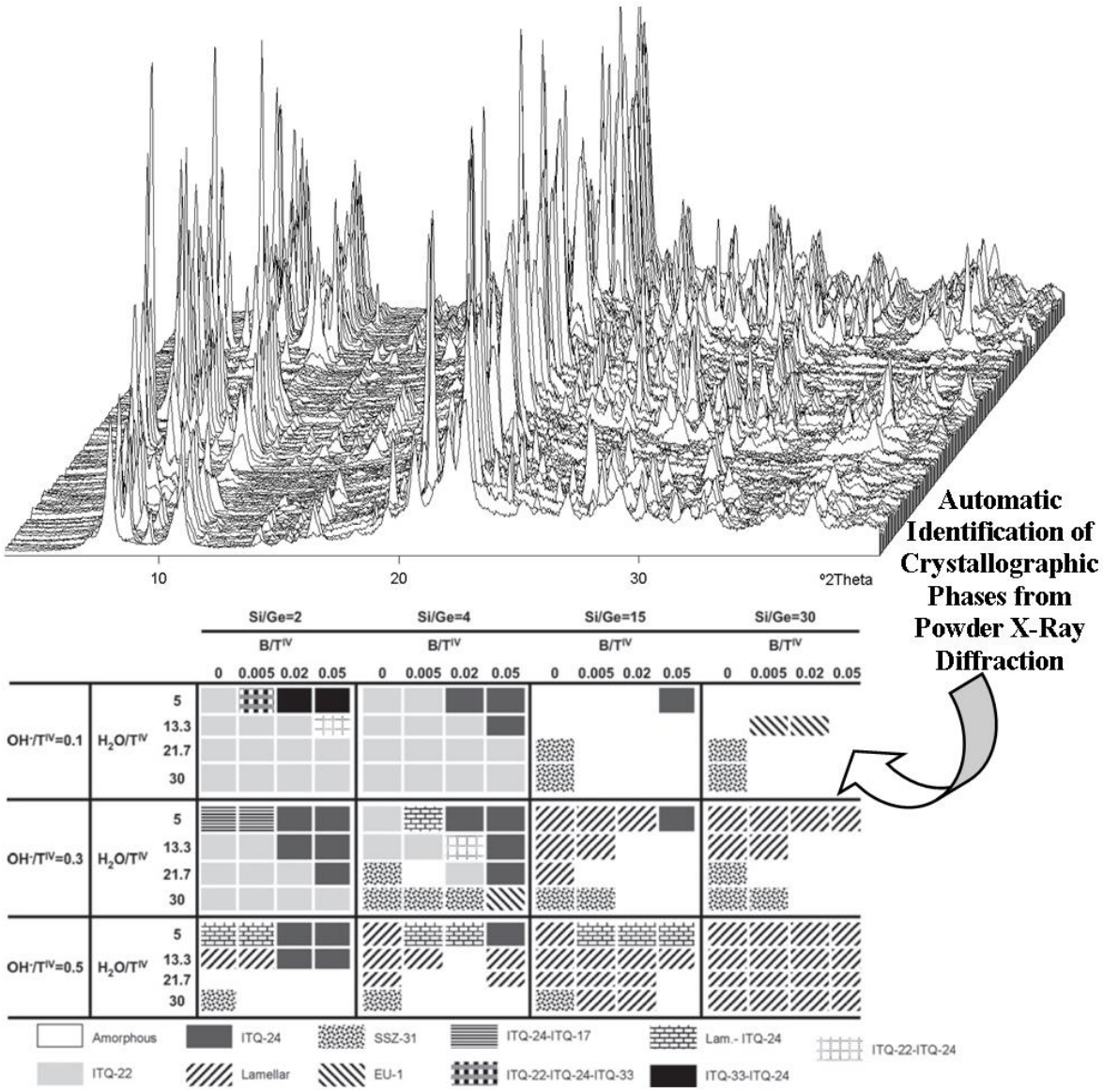

Figure 8. *The automatic analysis of the 192 X-Ray diffraction data of the discovery program of the zeolite ITQ-33 is expected to identify the different crystallographic phases present in each sample, allowing the creation of the corresponding phase diagram (8 different phases, numerous mixtures, and among them the presence of a new material has to be established).*

3.2. Experimental data

The microporous framework structures, the wide range of chemical composition and surface acidity, and the possibility of tuning their electric fields are key factors that render zeolites versatile materials for an increasing number of applications.[42,43] The discovery of new structures or enlarging the synthesis space, and optimization of existing ones require a considerable experimental effort. This can be diminished by using high throughput (HT) technology (usually parallelization and miniaturization)[44] as also done for heterogeneous catalysts.[45] As a recent example, this technique has allowed the synthesis of a very unique zeolite structure that includes extra-large 18MR connected with medium 10MR pores (ITQ-33).[46] The unusual conditions needed to synthesize this new material were found by using HT techniques to cover a wide range of possible synthesis conditions. That study required the generation of 192 diffractograms to follow the formation of the different crystallographic phases and mixtures of phases depending on the synthesis conditions, and from those mixtures of phases the presence of a new zeolite structure had to be established, see Figure 8. To achieve this is not obvious when complex mixtures of several crystalline phases exist in the same samples and consequently, the identification of the phases is time consuming. At the same time when increasing the amount of experiments to cover a broader synthesis range, it becomes mandatory to develop an automatic data treatment not to slow down the whole process.[47] Therefore an important step in this direction is the automatic classification of crystalline structures through XRD data, as it has been previously pointed by Takeushi et al.[48] Three datasets of experiments have been selected: "Zeolite β", "ITQ-21/30", and "ITQ-33" respectively described in Ref.44 and Ref.46.

"Zeolite β" - In order to optimise the preparation conditions for the synthesis of beta zeolite, a factorial design ($3^2 \times 4^2$) is employed for exploring different molar gel composition. The molar gel composition is explored by varying the following molar ratios: Na/(Si+Al), TEA/(Si+Al), OH/(Si+Al) and H_2O/(Si+Al), while Si/Al ratio is fixed. The experimental design considers the following four molar ratios (level): Na/(Si+Al) (3) ranging from 0 to 0.5; TEA/(Si+Al) (4) from 0.1 to 0.6; OH-/(Si+Al) (4) from 0.15 to 0.52; and H_2O/(Si+Al) (3) from 5 to 15. The total number of samples synthesised considering this factorial design is 144. The materials obtained in the studied area are amorphous materials, zeolite beta, and another dense material (named as UDM).

"ITQ-21/30" - The gel composition is explored by varying the following molar ratios: Al/(Si+Ge), MSPT/(Si+Ge), F^-/(Si+Ge) and Si/Ge. A factorial experimental design ($4 \times 3^2 \times 2^2 = 144$) is selected for studying simultaneously the crystallization time and the composition of the gel varying the

molar ratios of the components. Two zeolites are found in the area of the phase diagram studied: ITQ-21 and ITQ-30.

"ITQ-33" - Hexamethonium is used as structure directing agent (SDA). An initial experimental factorial design (3×4^3) is selected. Si/Ge, T^{III}/(Si+Ge), OH^-/(Si+Ge), and H_2O/(Si+Ge) are the synthesis variables. This experimental design considers the following four molar ratios (level): Si/Ge (4) ranging from 2 to 30; B/(Si+Ge) (4) from 0 to 0.05; OH^-/(Si+Ge) (3) from 0.1 to 0.5; and H_2O/(Si+Ge) (4) from 5 to 30. The number of synthesized samples is 192. Hexamethonium is a promising structure directing agent (SDA), since this molecule has a high number of degrees of freedom. Such flexibility allows different conformations that stabilize several competing structures, such as EU-1, ITQ-17, ITQ-22, ITQ-24, SSZ-31, a lamellar phase, and the new structure, ITQ-33.

3.3. Workflow Implementation

When integrating a new technique, the level of granularity must be defined by the programmer taking into account that splitting the technique into numerous sub-tasks gives more flexibility to modify the methodology but also it increases the complexity of the workflow due to the growing number of tasks. Here, we have chosen to encapsulate the ATW methodology into two different building blocks. We consider a set of *n* unknown diffractograms, *i.e.* time series, and a given classification algorithm. From the training step, we expect optimal settings for ATW distance considering the given problem and related data. The instance-based learning (IBL) method called *k*-nearest-neighbors (*k*-nn) is chosen as classification strategy. However, one very important trait of the methodology is its independence from the classification algorithm, *e.g.* another one more powerful could have been employed. The aim of the training is to obtain the best matrix noted Π which minimizes the classification error rate with ATW distance and the chosen learning strategy. Then, each time a new query instance will be presented, *i.e.* one of the unknown samples, its relationship to the previously stored examples is examined to assign a target function value and thus, the corresponding crystallographic phases it contains. An accurate description of the methodology will be given. The main reason for providing such a new strategy is that using other software, either commercial or freeware, it can be clearly observed that on this kind of problems, the methodology is not adapted. The reason why the performances of most classical algorithms initially created for multi-dimensional data suffer is mainly due to the lack of knowledge of the intrinsic ordering data particularity. Due to the relative importance/impact of the similarity expression, even very sophisticated approaches may failed, and the lack of stability when facing

various problems remains their principal drawback. Among the whole set of distances applied for the treatment of series, the Dynamic Time Warping[15] (DTW) appears as the most famous one.

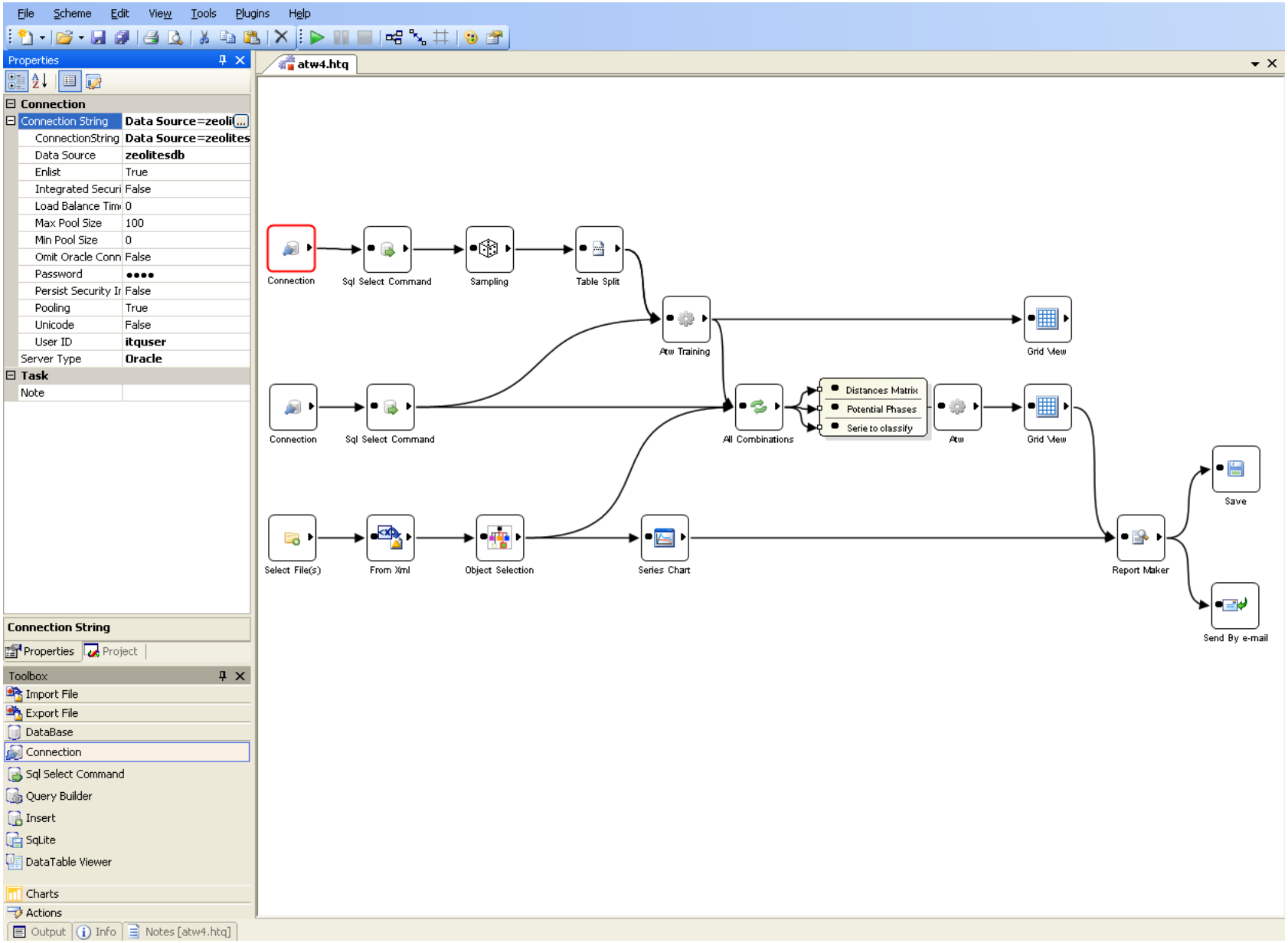

Figure 9. *Implementation of ATW methodology in hIT$_e$Q*

- Starting points: “select files” and “database connection”

Once the workflow is launched, each task is initialized and waits for its execution until it receives all required inputs. Usually, the tasks without input stand for starting points, their internal parameters being set at design time, see property grid placed on the top-left of Figure 9. hIT$_e$Q making use of multi-threading, numerous tasks can be executed at the same time. Thus, the workflow in Fig. 9 is initiated from three different starting points: two different database connections are automatically performed using the connection string previously defined at design time, while an “open” window for file selection is displayed and waits for the user to select the XRD files to be treated. Let us first focus on the upper part of the scheme. The two database connections allow retrieving the diffraction data from both the local ITQ database in which experiments conducted in the institute have been stored, and the second one which is composed of theoretical structures for which the composition can be controlled and the corresponding X-ray diffractograms modeled. With both real and theoretical XRDs, ATW will be trained in order to catch the inherent diffraction traits from all the presented crystallographic phases.

- SQL commands-sampling-table split

Every task has a set of typed inputs and outputs, and its implemented functionality that relates inputs and outputs can be of any kind, like sorting, making a database query, saving files, displaying charts, etc. During workflow construction, output-input type constraint is verified for consistency of the execution, as mentioned earlier. Here, each SQL select command receives one connection string as input and allows retrieving data from database. The SQL sentence can either be given as an optional input or be defined at design time, for example if a list of queries must be executed. The output will be composed of the resulting records. In order to train the ATW methodology, each of the queries will provide training examples, *e.g.* diffractograms and corresponding crystallographic phases. Then, the content of the resulting data table is ordered randomly (sampling task) and then is divided horizontally into two groups of x% and (100-x)% where x is the training set size (table split, variables are in columns and cases are in lines).

- ATW training

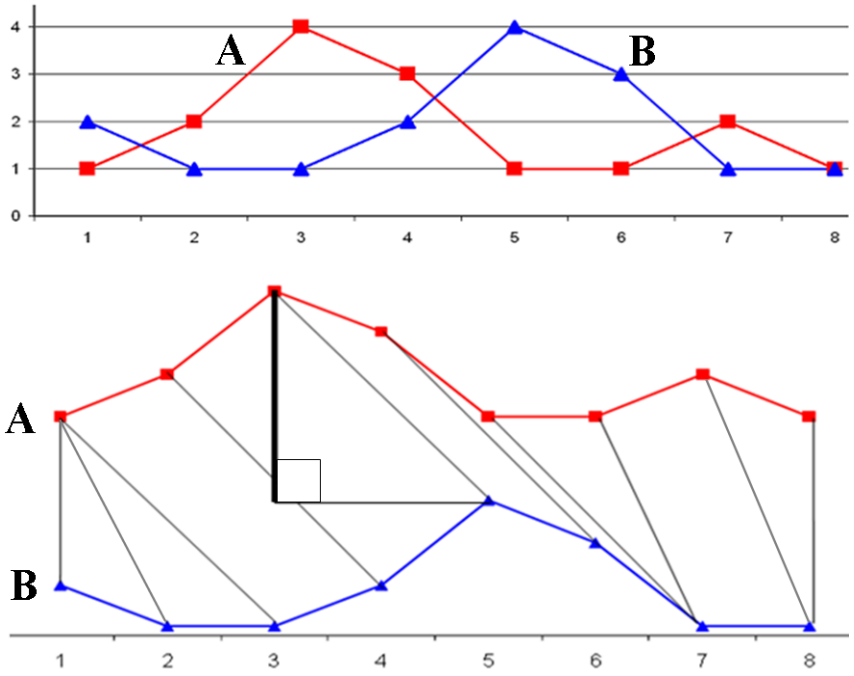

Figure 10. *The two series A (square) and B (triangle) with the matching using DTW and $\delta=\pm 2$. (A3,B5) are matched together, i.e. they belong to the warping path, as indicated by the grey line between the points. The distance which is retained corresponds to the black bold vertical line.*

The operation of non-warping distances over a pair of series consists in matching each point from one time series with the point from the second one that occurs at the same time. The easiest non-warping distance is the 1-norm distance $1\text{-}norm(A,B) = \sum|a_k - b_k|$ with $A=\{a_1,\cdots, a_t\}$ and $B=\{b_1,\cdots,b_t\}$. On the other hand, warping distances aim at managing temporal shift; they allow matching two time series by computing distance between points that do not occur at the same moment without rearranging the sequence of the elements, see Figure 10. The final sequence of pairs of points that are matched together is called the warping path $W(A,B) = w_1,\cdots,w_k,\cdots,w_K$ with $w_k=(a_i,b_j)_k$ and $i,j\in[1..t]$. w_k is the k^{th} pair of the warping path. w_k is composed of the i^{th} point of the series *A* and the j^{th} point of *B*. *K, i.e.* the length of the warping path, can be superior or equal to the length *t* of the series involved. The warping path $W(A, B) = (a_1, b_1), \ldots, (a_i, b_i), \ldots, (a_t, b_t)$, *i.e.* $\forall\{k,i,j\}, k=i=j$ corresponds to the 1-norm value. A point that occurs at instant *t* can be matched with points from the other series that occur in $[(i-\delta)..(i+\delta)]$. δ is generally set as constant value. Let's defined $A_i^{'}=(a_1,\ldots,a_i)$ and $B_i^{'}=(b_1,\ldots,b_j)$ with $1\le i,j\le t$, $A_i^{'}$ and $B_j^{'}$ being sub-sequences of *A* and *B* starting at 1 and finishing respectively at *i* and *j*. We note

$\gamma_D(A'_i, B'_j)$ the recursive function that defines the distance between the two sub-sequences. DTW is defined following by the recursive Equation 1. The computation of DTW needs a dynamic programming approach consisting in creating a $t \times t$ matrix, see Figure 11. Inside the cell (i, j) of the matrix the value $|a_i - b_j|$ is stored. After the cells are filled, the search for the best warping path begins. It is the path beginning in position (1,1) and finishing in (t, t) that minimizes the sum of the cells it goes through.

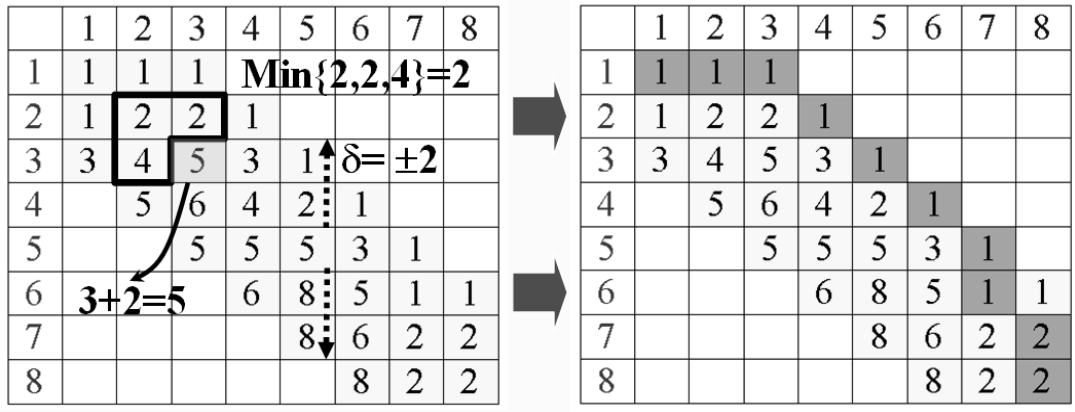

Figure 10. *The matrix on the right hand side is filled with the values corresponding to the DTW recursive function (Eq.1). For i=3 and j=3, /A3-B3/=/4-1/=3, and the minimum of the cells {(Ai-1,Bj);(Ai,Bj-1); (Ai-1,Bj-1)}=2. The best warping path which minimizes the sum of the cell it goes through is displayed in the matrix on the left with dark cells. It can be noticed that there is not always a unique best warping path, (A6,B7) can be substituted by (A6, B8).*

$$DTW(A,B) = \gamma_D(A_m, B_n) \text{ with} \qquad \text{(Equation 1)}$$

$$\gamma_D(A_i, B_j) = \left|A_i - B_j\right| + \min\left\{\gamma(A_{i-1}, B_j), \gamma(A_i, B_{j-1}), \gamma(A_{i-1}, B_{j-1})\right\}$$

Time warping efficiently handles the problem related to shifting of diffractograms. However, it must be merged with the integration of additional knowledge such as preliminary inspection of potential crystallographic phases, previous experiments, and theoretical calculations. The aim of analyzing additional data is to modify the similarity between diffractograms in order to improve recognition of mixtures, with or without the presence of amorphous. A way to achieve this is to modify the influence of the intensities for each angle position. Since the matrix which is employed for DTW calculation is a squared matrix of dimension *t, i.e.* the total number of angles steps, it can be used to integrate weights. A possible way to correctly assign the values of each weight is to "learn" how intensities must be interpreted or modified in order to get a correct answer from the labelling procedure. Due to the large number of angle positions, a machine learning (ML) approach

is employed. By presenting various cases either from past experiments or theoretical calculations, the setting of weights improves the classification results.

$$ATW\left(A,B,\Pi\right)=\gamma_A\left(A_t^{'},B_t^{'},\pi_{tt}\right) \qquad \text{(Equation 2)}$$

$$with\ \gamma_A\left(A_i^{'},B_j^{'},\pi_{ij}\right)$$

$$=\begin{cases}\infty \quad if\ \ \pi_{ij}=\infty \\ \left|a_i-b_j\right|\times\pi_{ij}+\begin{cases}0, if\ \ i=j=1\\ \gamma_A\left(A_i^{'},B_{j-1}^{'},\pi_{i(j-1)}\right), if\ \ i=1\ and\ \ j>1\\ \gamma_A\left(A_{i-1}^{'},B_j^{'},\pi_{(i-1)j}\right), if\ \ i>1\ and\ \ j=1\\ \min\begin{cases}\gamma_A\left(A_{i-1}^{'},B_j^{'},\pi_{(i-1)j}\right)\\ \gamma_A\left(A_i^{'},B_{j-1}^{'},\pi_{i(j-1)}\right)\\ \gamma_A\left(A_{i-1}^{'},B_{j-1}^{'},\pi_{(i-1)(j-1)}\right)\end{cases}, else\end{cases}\end{cases}$$

$$with\ \Pi=\begin{pmatrix}\pi_{11} & \cdots & \pi_{1t}\\ \vdots & \ddots & \vdots\\ \pi_{t1} & \cdots & \pi_{tt}\end{pmatrix}\ \ \pi_{ij}\in\square^{+},\ \ \forall i,j\in\left[1,t\right]$$

Let Π a $t\times t$ matrix, be the set of parameters required to compute ATW, $\Pi=\left[\pi_{ij}\right]\in\square^{+}\ \ \forall i,j\in\left[1,t\right]$. If $\pi_{ij}=\infty$, a_i is not allowed to be matched with b_j. $\gamma_A(A'_i, B'_j, \pi_{ij})$ is the recursive function that computes the distance between A'_i and B'_j. ATW is defined by Equation 2. Since the aim is to improve labelling results, the best matrix Π is defined as the matrix that minimizes the classification error rate with ATW and the chosen learning strategy, see Figure 7. Genetic algorithms (GAs) are selected in order to generate such matrix and optimize it. The reader is referred to Ref[26] for an introduction and advanced discussions. Let X be a set of n time series, $X = \{x_1,\ldots,x_n\}$, and Π a population of p individuals $\{\Pi_1,\ldots, \Pi_p\}$. Each individual is a candidate matrix Π_i that owns t^2 variables. The terms "individuals", "population", etc… are used following the GA terminology. Each cell of the matrix is a variable that the GA optimizes. The GA (*i.e.* all individuals Π_i) is initialized by assigning random values to every cell π_{ij}. The evaluation, then the selection, and finally the mutation/recombination are applied to each generation (Figure 7).

- From XML and Object Selection

ITQ and TOPCOMBI are investigating how XML[49] schemas – existing and in development – and web services suit the stringent requirements for data standardization, accessibility, portability and

modularity with new computational techniques. The underlying ontology defined in the XML schema facilitates the transformation of data into tangible knowledge. Here, diffractograms are automatically generated as XML files following XRDML format from PANalytical,[50]. This gives users complete control over their XRD measurement data. ASCII format XRD measurement data are stored in XML based files (.XRDML files) that contain all measurement data as well as all information required to reproduce the data, including instrument type and settings. The data are written in accordance with the XML schema. The information contained in the schema gives users the possibility to develop tailor-made programs to process their measurement data in any way they need. In hITeQ, a XML viewer and editor have been integrated. For each type of content the viewer automatically displays using the corresponding format: images (for example for SEM and TEM), 2D (XRD) or 3D charts, etc…

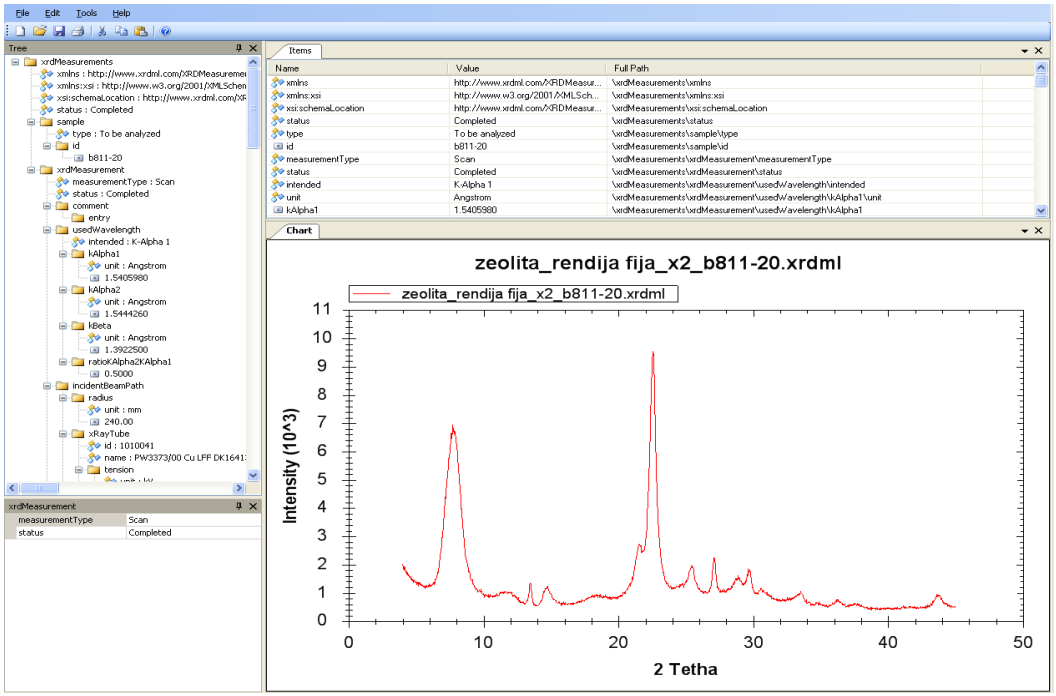

Figure 12. *XML view of an XRDML file produced by Philips X-ray diffractometer.*

An example, in Figure 12 we can observe that data is organized and labeled allowing data to be treated properly and without any confusion on the meaning of the information provided. Moreover, this allows for example to rescale diffraction intensities due to the use of different X-ray tubes and their corresponding wavelengths. This would be impossible if XRD data would have been stored as simple XY files without such additional and crucial information. Note that previously from storing the XRDML files in the databases, their content is split and stored giving faster access to each peace

of information. For example and as mentioned before, the data of interest considering past experiments and theoretical structures is directly retrieved from their corresponding DBs without opening XRDML files. For the new experiments, only the XRD series are used and consequently

- ATW - Grid - Charts - Report

Considering a set of *n* time series and a given classification algorithm, we expect optimal settings for ATW distance considering the given problem and related data. Two different algorithms are required: a classification technique, and another one for the automatic weight setting. In order not to make the methodology too complex, the instance-based learning (IBL) method called *k*-nearest-neighbors (*k*-nn) is chosen as classification strategy. Our approach is a learning process which uses a heuristic, and allows reaching near-optimal solutions. According to a given classification problem, ATW approach adapts itself for better capturing data specificity. ATW can be associated to both warping and non-warping distances as demonstrated in Ref.41. Therefore, whatever the temporal classification problem, optimal parameters *Π* imply ATW to give results at least equal or superior to other distances, and consequently ATW is always at least equivalent or superior to all other distance use. Once the unknown samples, *i.e.* test samples are treated, ATW calculates using the optimized, *i.e.* trained, matrix Pi, to which class or phase they belong to. It can be observed how the recognition rate is clearly improved compared to PolySnap use. One of the good points using ATW is that there is no false negative as indicated in the Table1. The training of the matrix has allowed to capture the specificities of each phase making lower peaks more important than the highest ones where no clear decision can be done while handling the shifting of the diffractograms. Once the presence of the phases is done, solving the problem of the respective part of each phase remains trivial. Quantitative analysis is carried out using all the measured data points with singular value decomposition (SVD) as the tool of matrix inversion to ensure computational stability. The percentage composition of the sample is determined. Then, results are automatically displayed in a grid, while series are plotted in order to get a visualization of the classification. Other functionalities may be integrated as building reports where series, tables, data, etc… are merged in a pre-designed way defined by the user. As the output can be multiplexed, *i.e.* an output can be send to different inputs bricks, while one input only receive data from one output, the information can be sent by e-mail at the same time, printed or other kind of actions.

Table 1. *Confusion matrix with the real phases in the experimental design "ITQ-33" versus predicted classes obtained with PolySnap2© indicated on the left hand-side of the cell, and versus ATW indicated on the right hand-side of the cell (in italic and between brackets). [a] ITQ-24 + amorphous. [b] SSZ-31 + amorphous.*

	Real phases										
Predicted phases	Am.	ITQ-22	ITQ-24	EU-1	SSZ-31	Lam.	ITQ-17	Lam. + 24	24 + 33	17 + 24	22 + 24	22 + 24 + 33
Amorphous	55 *(55)*	2	3	2	3	28		2				
ITQ-22		36 *(36)*	2								1	1
ITQ-24			8 *(17)*					2				
EU-1				1 *(3)*								
SSZ-31					13 *(15)*							
Lamellar						18 *(46)*		1				
ITQ-17							0 *(0)*			1		
Lamellar + 24								0 *(8)*				
24 + 33									0 *(2)*			
17 + 24										0 *(2)*		
22 + 24		*(2)*									0 *(2)*	
22 + 24 + 33												0 *(1)*
Other mixtures		6 *(2)*[a]			*(1)*[b]			3	2	1	1	
	55	38	19	3	16	46	0	8	2	2	2	1

131/192=68.2% *(187/192=97.4%)*

4. Conclusion

hIT$_e$Q is a workflow tool which facilitates the design and implementation of workflows, includes a graphic tool for drawing a workflow using icons that represent steps in a process, and a workflow engine that drives the workflow. In contrast to pure automation, workflows assumes and, in fact, requires manual intervention through a work process, such as “input files”, “initial parameters set up”, “selection among results”, etc… In contrast to workflow tools, application automation tools or so-called job schedulers begin with the assumption of full automation where human intervention must be eliminated wherever possible. Vendors such as Oracle™ provide such automation tool with the aim to open organizations to the most efficient means of completing processes. The engine schedules the execution of entire business processes based on dates, times, day of weeks, business and fiscal calendars, dependencies, conditions, and events, connect individual jobs together into meaningful business processes that model the business goals. Despite research laboratories are dedicated to scientific activities for which the elimination of all human tasks is not conceivable, a large part of their activity rely on data, reports, events, rules which could be automated increasing the process reliability, data accessibility, overall efficiency… hIT$_e$Q which was first designed as a workflow tool, integrates now functionalities of automation tools. Fully automated processes without human interventions can be designed at a higher level than individual data treatments. For example when coupled with a high throughput apparatus where series of predefined treatments are executed on all incoming files responding to some rules (*e.g.* run report if property_X is above a given threshold) and/or filters (*e.g.* file extension) trying to mimic the flow of information between agents inside the institute. In this example, the same scheme is re-executed for each new set of diffractograms. The classification for the other applications takes only few minutes and the results (grid, charts, and reports) are displayed without additional intervention. As such, the workflow processing tool combines an ensemble of tools which clearly facilitates the rapid prototyping of solutions and results/documentation sharing. However, building integrated platforms is one challenging task which scientific community is dealing within recent years. Facing this task, a number of specific problems arises connected to data integration, integration of specialized tools and algorithms. The solution described in this paper goes in the direction to solve this challenge. The presented tool constitutes the basic components of a much more general materials science e-workplace where webservices are tools of excellence. These can be seen as plug-and-play components dealing automatically with on-demand processing, in a collaborative and

transparent fashion within the concept of virtual organization such as the large European consortiums, for example TopCombi.

Acknowledgements

EU Commission FP6 (TOPCOMBI Project) is gratefully acknowledged.

References

[1] C. Klanner, D. Farrusseng, L. Baumes, C. Mirodatos & F. Schueth, *QSAR Comb. Sci.* **2003**, *22*, 729-736; C. Klanner, D. Farrusseng, L. Baumes, M. Lengliz, C. Mirodatos & F. Schueth, *Angew. Chem., Int. Ed*. **2003**; L.A Harmon, A.J Vayda & S.G. Schlosser,. Abstr. Pap. 221st ACS Meeting, **2001**, BTEC-067

[2] F. Gilardoni, V. Curcin, K. Karunanayake & J. Norgaard , QSAR & Comb. Sci. **2005**, *1*, 24, 120-130

[3] http://www.inforsense.com/

[4] http://www.upv.es/itq

[5] http://www.upv.es

[6] http://www.csic.es

[7] http://www.topcombi.org/

[8] L.A. Baumes, M. Moliner, A. Corma. ChemEngComm, **2008**, 10, 1321-1324. This article has been chosen to be highlighted in the RSC news magazine Chemical Science (www.rsc.org/chemicalscience). Chemical Science provides a snapshot of some of the most significant new papers published.

[9] a) H. Lee, S.I. Zones, M.E. Davis, *Nature* **2003**, 425, 385-387. b) A. Corma, *J. Catal.* **2003**, 216(1-2), 298-312.

[10] a) M. Moliner, J.M. Serra, A. Corma, E. Argente, S. Valero, V. Botti, *Micropor. Mesopor. Mat.* **2005**, 78, 73-81. b) O. B. Vistad, D.E. Akporiaye, K. Mejland, R. Wendelbo, A. Karlsson, M. Plassen, K. P. Lillerud, *Stud. Surf. Sci. Catal.* **2004**, 154, 731-738. c) A. Cantin, A. Corma, M.J. Diaz-Cabanas, J.L. Jorda´, M. Moliner, *J. Am. Chem. Soc.* **2006**, 128, 4216-4217. d) A.Corma, M. Moliner, J.M. Serra, P. Serna, M.J. Díaz-Cabañas, L.A. Baumes. *Chem. Mater.* **2006**, *18, 3287-3296.* e) J.M. Serra, L.A. Baumes, M. Moliner, P. Serna, A. Corma. *Comb Chem High Throughput Screen.* **2007**, 10 (1), 13-24

[11] a) D. Farrusseng, L.A. Baumes, C. Hayaud, I. Vauthey, P. Denton, C. Mirodatos. Kluver Academic Publisher, Nato series, edited by E. Derouane. Proc. NATO Advanced Study Institute on Principles and Methods for Accelerated Catalyst Design, Preparation, Testing and Development, Vilamoura, Portugal, 15-28 July **2001**. eds. E. Derouane, V. Parmon, F Lemos, F. Ribeiro. Book Series: NATO SCIENCE SERIES: II: Mathematics, Physics and Chemistry. Vol. 69, 101-124, Kluwer Academic Publishers, Dordrecht. Hardbound, ISBN 1-4020-0720-5. July **2002**. b) L.A. Baumes, D. Farruseng, M. Lengliz, C. Mirodatos. *QSAR & Comb. Sci.* **2004**, vol. 29, Issue 9, 767-778. c) C. Klanner, D. Farrusseng, L.A. Baumes, M. Lengliz, C. Mirodatos, F. Schüth. *Angew. Chem. Int. Ed.* **2004**, 43, N° 40, 5347-5349. d) F. Schüth, L.A. Baumes, F. Clerc, D. Demuth, D. Farrusseng, J. Llamas-Galilea, C. Klanner, J. Klein, A. Martinez-Joaristi, J. Procelewska, M. Saupe, S. Schunk, M. Schwickardi, W. Strehlau, T. Zech. *Catal. Today*. Vol. 117, **2006**. 284-290. e) L.A. Baumes. **2006**, 8, 304-314. *J. Comb. Chem.*

[12] http://www.scitegic.com/

[13] http://www.biolog-tech.com/

[14] http://mgltools.scripps.edu/

[15] http://knime.org

[16] Hoon S, Ratnapu K, Chia J, Kumarasamy B, Juguang X, Clamp M, *Genome Res* **2003**;13:1904–15.

[17] Leo P, Marinelli C, Pappada G, Scioscia G, Zanchetta L. Bioinformatics Italian Society Meeting (BITS 2004), Padova; **2004**

[18] Oinn T, Addis M, Ferris J, Marvin D, Greenwood M, Carver T, *Bioinformatics* **2004**;20:3045–54

[19] http://taverna.sourceforge.net/

[20] Fang F, Chua C, Ho L, Lim Y, Issac P, Krishnan A. *BMC Bioinform*at **2005**;6:69
[21] L. Zhao, T. Park, R. Kalyanam, S. Goasguen, SRB Workshop, Vol. 1, pp. 611, February 2006.
[22] http://www.science-factory.com/iindex.html
[23] http://www.isoft.fr/bio/biopack_en.htm
[24] http://ptolemy.eecs.berkeley.edu/ptolemyII/
[25] http://kepler-project.org/
[26] http://www.migenas.org/home/index.jsp
[27] http://www.incogen.com/
[28] http://www.uncw.edu/csc/bioinformatics/
[29] http://www.alphaworks.ibm.com/tech/wsbaw
[30] Bonomini MP, Ferrandez JM, Bolea JA, Fernandez E. J. Neurosci Meth **2005**;148:137–46
[31] Vato A, Bonzanoa L, Chiappalone M, Cicero S,Morabito F, Novellino A, Stillo G. *Neurocomputing* **2004**;58–60:1153–61.
[32] Tiwari A, Sekhar AKT. *Comput Biol Chem* **2007**;31:305–19.
[33] http://www.cs.waikato.ac.nz/ml/weka/
[34] a) Steinbeck C., Han Y., Kuhn S., Horlacher O., Luttmann E., Willighagen E.L., *J. Chem. Inf. Comput. Sci.* **2003** Mar-Apr; 43 (2):493-500. b) Steinbeck C., Hoppe C., Kuhn S., Floris M., Guha R., Willighagen E.L., *Curr. Pharm. Des.* **2006**; 12 (17):2111-2120
[35] a) Letondal, C., in *Bioinformatics*, 17(1), 73-82, 2001. b) D'Elia, D. et al., in ECCB 2003 Proc. Paris, France, September 27-30, **2003**.
[36] Russo, Report "A general framework for implementing genetic algorithms", Edinburgh Parallel Computing Centre, Univ. of Edinburgh. EPCC-SS91-17, **1991**
[37] Jones, Report "Parallel reproduction plan language", Edinburgh Parallel Computing Centre, Univ. of Edinburgh.**1992**
[38] a) I. Takeuchi, C. J. Long, O. O. Famodu, M. Murakami, J. Hattrick-Simpers, G. W. Rubloff, M. Stukowski, K. Rajan. *Rev. Sci. Instrum.* **2005**, 76, 062223; b) X'Pert HighScore Plus software from PANalytical, see website visited Nov. 20th **2008**, http://www.panalytical.com/index.cfm?pid=547; c) D. L. Bish, S. A. Howard, *J Appl Crystallogr*. **1988**, 21, 86-91; d) D. L. Bish, S. J. Chipera, in *Advances in X-ray Analysis* (Ed.: C. Barrett, et al.), Plenum Press, New York **1988**, 31, pp. 295-308; e) D. L. Bish, S. J. Chipera, in *Advances in X-ray Analysis* (Ed.: P. Predecki, et al.), Plenum Press, New York **1995**, 38, pp. 47-57; f) S. J. Chipera, D. L. Bish, *Powder Diffr*. **1995**, 10, 47-55; g) F. H. Chung, *J. Appl. Crystallogr*. **1974**, 7, 519-525; h) RockJock – Uses Microsoft Excel Macros and the Solver function to perform a whole-pattern modified Rietveldtype refinement to perform quantitative analysis. Written by Dennis D. Eberl, the software was published in **2003** as U.S.G.S. Open-File Report 03-78, "Determining Quantitative Mineralogy from Powder X-ray Diffration Data". Available via FTP from ftp://brrcrftp.cr.usgs.gov/pub/ddeberl/RockJock. i) See JADE from http://www.rigaku.com/index_world.html.
[39] a) C. J. Gilmore, G. Barr, J. Paisley, *J. Appl. Crystallogr*. **2004**, 37, 231–242; b) G. Barr, W. Dong, C. J. Gilmore, *J. Appl. Crystallogr*. **2004**, 37, 243–252.
[40] G. Barr, W. Dong, C. J. Gilmore, J. Appl. Crystallogr. 2004, 37, 658–664.
[41] a) L.A. Baumes, M. Moliner, A. Corma. *Chem. a Eur. J.* **2009**, 15, 4258-4269. b) L.A. Baumes, M. Moliner, A. Corma, *CrystEngComm*, **2008**, 10, 1321–1324
[42] A. Corma, *J. Catal*. **2003**, *216*, 298-312.
[43] H. Lee, S. I. Zones, M. E. Davis, *Nature*. **2003**, *425*, 385-387.
[44] a) A. Corma, M. J. Díaz-Cabañas, J. Martínez-Triguero, F. Rey, J. Rius, *Nature*. **2002** *418*, 514-517. b) O. B. Vistad, D. E. Akporiaye, K. Mejland, R. Wendelbo, A. Karlsson, M. Plassen, K. P. Lillerud, *Stud. Surf. Sci. Catal.* **2004**, *154*, 731-738. c) A. Cantin, A. Corma, M. J. Diaz-Cabanas, J. L. Jordá, M. Moliner, *J. Am. Chem. Soc*. **2006**, *128*, 4216-4217. d) A. Corma, M.J. Diaz-Cabañas, M. Moliner, C. Martinez. *J. Catal.* 241, 2, 312-318, **2006**. e) A.Corma, M. Moliner, J. M. Serra, P. Serna, M. J. Díaz-Cabañas, L. A. Baumes, *Chem. Mater*. **2006**, *18*, 3287-3296.
[45] a) J. M. Serra, A. Corma, D. Farrusseng, L. A. Baumes, C. Mirodatos, C. Flego, C. Perego, *Catal. Today*, **2003**, 81, 425-436; b) C. Klanner, D. Farrusseng, L. A. Baumes, C. Mirodatos, F. Schüth, *QSAR & Comb. Sci.*, **2003**, *22*, 729-736; c) D. Farrusseng, C. Klanner, L. A. Baumes, M. Lengliz, C. Mirodatos, F. Schüth, *QSAR & Comb. Sci.* **2005**, *24*, 78-93; d) F. Schüth, L. A. Baumes, F. Clerc, D. Demuth, D. Farrusseng, J. Llamas-Galilea, C. Klanner, J. Klein, A.

Martinez-Joaristi, J. Procelewska, M. Saupe, S. Schunk, M. Schwickardi, W. Strehlau, T. Zech, *Catal. Today*. **2006**, 117, 284-290.

[46] a) A. Corma, M. J. Díaz-Cabañas, J. L. Jordá, C. Martínez, M. Moliner, *Nature*. **2006**, *443*, 842–845; b) M. Moliner, M. J. Díaz-Cabañas, V. Fornés, C. Martínez, A. Corma, *J. Catal*. **2008**, *254*, 101-109.

[47] a) L. A. Baumes, D. Farruseng, M. Lengliz, C. Mirodatos, *QSAR & Comb. Sci.* **2004**, *29*, 767-778; b) C. Klanner, D. Farrusseng, L. A. Baumes, M. Lengliz, C. Mirodatos, F. Schüth, *Angew. Chem. Int. Ed.* **2004**, *43*, 5347-5349; c) L. A. Baumes, J. M. Serra, P. Serna, A. Corma, *J. Comb. Chem.* **2006**, 8, 583-596; d) L. A. Baumes, *J. Comb. Chem.* **2006**, *8*, 304-314; e) J. M. Serra, L. A. Baumes, M. Moliner, P. Serna, A. Corma, *Comb. Chem. High Throughput Screen*. **2007**, 10, 13-24; f) L. A. Baumes, M. Moliner, A. Corma, *QSAR & Comb. Sci* **2007**, *26*, 255-272; g) P. Serna, L. A. Baumes, M. Moliner, A. Corma. *J. Catal.* **2008**, *258*, 25-34; h) L. A. Baumes, P. Collet, *Com. Mat. Sci*. **2008,** doi:10.1016/j.commatsci.2008.03.051.

[48] C. J. Long, J. Hattrick-Simpers, M. Murakami, R. C. Srivastava, I. Takeuchi, V. L. Karen, X. Li. *Rev. Sci. Instrum*. **2007**, 78, 072217.

[49] B. Schaefer, L.A. Baumes, A. Corma. LabAutomation **2008** Palm Springs CA, 2633 Monday, 01/28/2008 1:00PM - 3:00PM , Room MP94

[50] http://www.xrdml.com