Knowledge Grid and Grid Intelligence 2004

Proceedings of the Second International Workshop on Knowledge Grid and Grid Intelligence

Edited by Hai Zhuge, William K. Cheung and Jie Liu

Sponsors

Organizer

China Knowledge Grid Research Group (http://kg.ict.ac.cn)

In Conjunction with 2004 IEEE/WTC/ACM International Conference on Web Intelligence / Intelligence Agent Technology

ISBN 0-9734039-8-5

Semantic Description and Tracking of Analysis of Chemical Data

Hongchen Fu, Jeremy G. Frey
School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K.
{h.fu, j.g.frey}@soton.ac.uk

Abstract

We investigate the use of semantic technologies to describe chemical data in a manner that facilitates the subsequent reuse of both the data and the analysis procedure used to produce the data. We first propose an approach to describe and record the analysis process for a physical chemistry laser spectroscopic experiment that serves as an e-logbook for people and provides a sufficiently rich semantic description suitable for programmatic processing of the entries. We then developed a semantic-aware enactment system to perform the analysis using the information in the RDF description of the analysis process. A generic RDF browser was developed to browse and reuse the analysis process. The raw and analyzed data is displayed graphically in web browser using interactive SVG tools. This approach demonstrates that semantic technology is an appropriate technology to record and manage the complicated, flexible analysis procedures used to interpret physical chemistry experimental data.

1. Introduction and motivations

There is increasing interest in using grid and semantic technologies to enhance scientific research [1,2]. Examples of the areas were such technology can be expected to improve and extend research are, the remote control and automation of experimental facilities, accessing and managing generated experimental data, acquiring, publishing and managing scientific knowledge acquired process from experimental raw data using computational and network resources in a distributed environment [3-6]. In Comb-e-Chem [7], one of the UK e-science pilot project, we have been investigating a number of aspects of a "smart laboratory" that would support the remote control of chemical experiments, automatic collection and management of experimental raw data using database technology and web services [8], developing the e-logbook which enables us to record and manage experimental process using semantic technology [9]. In this paper we shall focus on acquisition and reuse of knowledge obtained from experimental raw physical chemistry data using analysis and simulation methods.

Scientific data simulation is a dynamic, complicated and flexible process. It is normally decomposed into many steps and each step may use computing resources distributed in different organisations and locations and supplied by different vender. In the environment of grid computing and e-sciences, we use a service-oriented approach in which the computing resources are wrapped as web or grid services and can be accessed transparently using their WSDL [10] or GWSDL [11] description and be discovered via UDDI [12] or grid information service [13]. The recently proposed WSRF supplies a unified framework for web services and grid services [14].

In this paper we shall propose an approach to describe and record the process of (chemical) data analysis which allows us to (a) record how the analysis is performed and how the chemical knowledge is acquired from the experimental raw data, (b) perform the analysis using web/grid services, (c) reuse and tracking the analysis process from the viewpoint of the publication@source [15], and (c) manage the acquired knowledge and analysis process in a unified manner (For detailed requirements see Sec.2.2).

We need to record and manage not only the final chemical knowledge but also the intermediate results and the analysis process itself. Significance of recording and managing the process of knowledge acquisition is as follows:

- Re-use the experiment data check previous analysis and to provide the basis for new investigations and derive new knowledge using for example new theoretical models and assumptions to re-interpret the same set of experimental raw data. In this case it is important to be able to describe and store all the details of the way the analysis was performed.

- Publication@source as a methodology for disseminating data and process tracking [15]. Due to limited space it is impossible to publish the detailed process of knowledge acquisition and the full set of experimental raw data in research papers. So linking back the experimental raw data, held for example in an institutional archive, from a published paper is important.

- Data Sharing with collaborators: Global collaboration is the main motivation of e-science

[16]. Information about chemical process should be recorded by a language that all parties can understand and process.

The approach taken here also enables the enactment and execution of an experimental data analysis or simulation. Compared with other approaches using XML technology [17], our semantic approach enables inference and describes resources (analysis steps) with labeled graphs of relationships [18] and supplies a global identification of analysis process using URIs, an aspect that is essential in global collaboration.

Scientific research is a knowledge-intensive process. RDF [19] and ontology [20] has been generally regarded as a formalized knowledge abstraction in a domain of interest. They can be used by both human and machine for knowledge sharing and communication, as well as common understanding of domain knowledge and metadata. From above requirements we see that semantic technologies (RDF and ontology) are appropriate candidate for our purpose.

This paper is organized as follows: we first model the process of data analysis and then investigate how to record it using the RDF in Sec.2. In Sec.3 we investigate the implementation of enactment system and RDF browser and their use in the tracking and reuse of process. In Sec.4 we apply the approach to statistical process of SHG experiment. We conclude in Sec.6.

2. Semantic description of analysis process

In this section we suggest methods by which the analysis undertaken in the laser spectroscopy experiment could be described using RDF and an ontology. We first model the analysis process and then analyze the requirements the approach should meet. General description of analysis process using RDF model is presented and the hierarchy structure of ontology is discussed in more detail.

2.1. Characteristics of data analysis process

A process of analysis of chemical data is normally decomposed into many steps which depend on subject, computing facilities (hardware and software, operating system, algorithms), theoretical models and assumptions, and chemist's experience, cultural background and habits and usually have the following common characteristics:

- **Decomposition into a series of steps.** Scientists usually divide a complicated analysis process into a series of ordered steps.
- **Serial relevance of steps.** Each step in the decomposition of a process may use results from previous step. The order of steps is thus important. It is possible that there is a *loop* in a step. For example, the result of a step does not meet the requirement due to an unreasonable assumption. In this case, this step will be repeated using a different assumption until the good result is achieved. In this case, we only need to record how the good result is achieved and the loop in this step is not necessary to be recorded. If we have to record the loop, the loop itself can be regarded a multi-step process, namely, it itself can be further decomposes into many steps and each run in the loop is a step. For convenience, we refer to this kind decomposition as the *linearization* or *serialization* of a step.
- **Relative independence of each step.** Each step is relative independent in the sense that it may have its own theoretical model and assumptions and use web/grid services supplied by different organizations and distributed in different locations. It is also possible that a step in a process has more than one next step depending on different conditions and/or the result of current step.

Generally speaking the multi-step decomposition of a chemical process is not unique. Depending on chemist's research experience, the theoretical models adopted and assumptions, different chemists may decompose a process into steps in different ways. On the other hand, in the scenario of *loop* process, the good result may be achieved in a number of steps and the number of steps obviously depends on the initial assumption and algorithm for retrying the computation. In this case, we may not even know how many steps are needed to achieve the good results in advance. In other words, the simulation process is very dynamic.

This multi-step decomposition of data analysis process can be illustrated in the following figure:

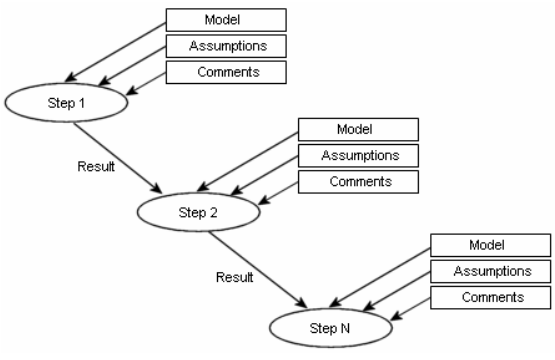

Fig.1. Model of chemical process. In this figure, the

2.2 Requirements

The approach we propose in this paper should meet the following requirements:

- **Process logging.** .
  As described in the Introduction it is important to record how the analysis is performed. The log is for scientists to understand the assumptions, theoretical models used, computing facilities etc. and to be able to automatically process this information if required. These details can hopefully be expressed in terms of scientific notations which take many different formats such as pictures, figures, videos, mathematical formulae, chemical structures etc. Fortunately, those notations can be represented by the XML-based technologies, e.g. XML for structured data and text [21], SVG (scalable vector graphics [22]) for pictures and figures, MathML (Mathematical Makeup Language [23]) and CML (Chemical Makeup Language [24]). The formalism should be able to integrate those XML-based technologies easily.
- **Enable to enact service-oriented data analysis**. The purpose of this work is to propose a formalism to describe and enact a data analysis process. The analysis is performed using WSDL-based web/grid services in a distributed environment. This requires that the description of the analysis is machine-understandable and processible and that the associated enactment system can automatically perform the analysis according to the service information retrieved from the description. Multi-step decomposition of data analysis process is flexible and not unique. But we do require that in the decomposition each step can only invoke at most one operation of the analysis web service. This means that in each step we can only perform one computation task.

- **Built-in logic**. The system should be capable of interpreting the built-in logic such as the structure of multiple step decomposition, without using any hard-coded tools and determining automatically what is the next step. It can also describe whether pause when this step is finished before performing the next step. This feature is import in some scenarios such as computational chemistry where each step may take days to finish. In this case it may be necessary to check whether the result from current step is good or not and then decide if carry on to next step.

- **Self-definition**. In chemical analysis, chemists have to use some scientific terminologies, define some variables and introduce some notations. Those vocabularies needs to be carefully defined/explained and are available and accessible in the proposed approach.

- **Unified management of process and knowledge**. Knowledge management is an important subject in computer science [25]. We require that our approach can supply an easy management of acquired knowledge as well as the process and intermediate results in a unified manner.

2.3. RDF description of analysis process

The requirement of unified management of chemical knowledge and analysis process suggests us using RDF and Ontology to record the chemical data analysis processes. RDF is designed to be used in *applications that require open rather than constrained information models*, processing *machine processable information (application data)* and *combining data from several applications to arrive at new information*.

The RDF model that we have generated to describe a chemical process has the following properties.

- **Represent a step by a RDF resource**.
  Each resource consists of a number of statements and each statement has its *subject*, *predicate* and *object*. An important feature of RDF model is that the object of a statement can be another resource. For convenience, we call the current resource the *parent resource* and the resource as the object of a statement of the current resource the *child resource*. A resource is identified by a unique URI (Uniform resource identifier). We use a resource to describe a step in a chemical process and represent next step by a child resource. The previous step is described by the parent resource of current resource. The feature that the object of a statement can be another resource gives us build-in implementation of multiple steps of chemical process.

- **Definition the vocabularies by an ontology**.
  In recording chemical process, chemists have to use some specific scientific terminologies, define some variables and introduce some notations. Those *new* vocabularies are used as predicates in the RDF description of chemical process and can be defined in the corresponding ontology. In this work we develop ontology for SHG experiment using OWL. This arrangement allows us to process both the chemical process and the schema in the way in the Dolphin RDF browser.

- **Integrate other XML-based representation of chemical knowledge.** This enables us to link and embed complicated scientific notations such as figures (SVG), mathematical formula (MathML) and chemical structures (CML) into the RDF. This can be easily achieved by indicating the object of a statement the RDF literal, namely *rdf:parseType="literal*".

We will discuss this in detail with SHG statistical analysis as an example in next section.

2.4. Service invocation

In each step we need to record information for enactment system to invoke the web services to perform the data analysis. So we introduce the following (Datatype) properties which are defined in the ontology:

hasEndpoint: gives the endpoint URI of the analysis web service.
hasWSDL: gives the WSDL URI of the analysis web service.
useOperation: gives the name of operation you want to invoke.
useParameters: gives the input parameter of the operation you want to invoke.

2.5. Next Step

We define two predicates *hasNextStepURI* and *isPause* to give the URI of the next step and indicate whether the analysis pauses before executing analysis for next step, respectively. The predicate *isPause* takes a Boolean value: true (yes) or false (no).

Use the predicate **result** to indicate the result should be inserted to the RDF file.

2.6. Ontology

All predicates/vocabularies used to describe the analysis process are defined in the ontology. The process, which is referred to as the *OntoProcess*, is implemented using OWL (Web ontology Language [26]). It has the following hierarchy structure:

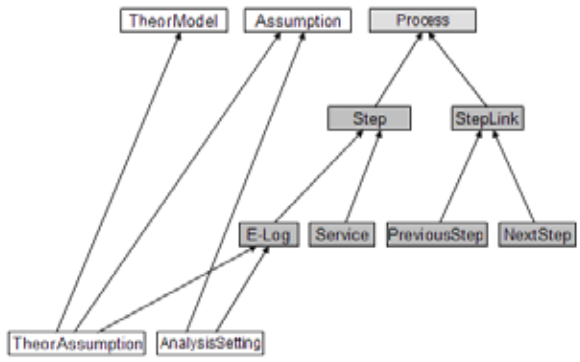

Fig 2. Class hierarchy of OntoAnalysis ontology

Here the grey background classes are essential to describe the analysis process. TheorModel (theoretical model) and Assumption are used mainly by e-logbook and may be slightly different for different subjects.

3. Implementation

We now turn to the development of the semantics-based system of analysis. This system comprises of the following components:

- Analysis designer;
- Analysis enactment system;
- Analysis management system;
- e-bank.

The architecture of the system is as follows. Some components are from other projects and we only need to integrate them into our analysis system.

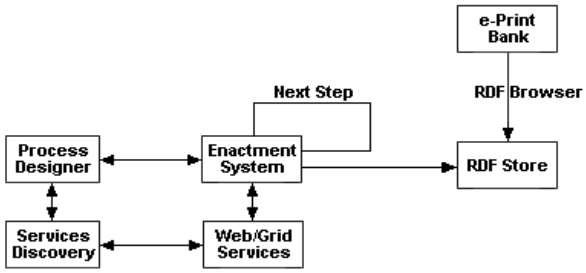

Fig. 3. Architecture of analysis system.

3.1 Analysis enactment system

This component is used to retrieve the web/grid service information from the RDF description of the analysis and then invoke the services to perform the analysis. We developed a generic process enactor, which includes three components: RDF interpreter, web service invoker and RDF updater. The RDF interpreter is used to retrieve the services information from RDF description. Service invoker is used to invoke the web service to perform the analysis and then get the result and submit the result to updater. The RDF updater is used to insert the result from web services to the original process RDF model. The updated RDF model with results is saved on the secure server for review and reuse or is sent to the 3Store for management.

The enactor is implemented as a web application running on the Tomcat server and it enacts the analysis by indicating the URI of the process RDF model. The enactor can automatically find the URI for the first step by checking whether that resource/step has parent resource.

3.2. Triscape RDF browser

To review and repeat an analysis process recorded in RDF, one need a generic RDF browser to review the analysis step and step to check the result in various

format. The RDF browser, which is referred as the Triscape RDF browser, has the flowing features:

1. The Triscape RDF browser is generic and it can be used to browse any valid RDF resources.

2. The Triscape RDF browser is implemented as a web application (Servlet) running on the Tomcat server. Here the Jena API [27] from HP laboratory plays important role. The controller (such as a link or XLink in an e-print) sends a HTTP request to the server and then RDF model is called. The RDF browser shows the content of RDF model describing the analysis process in the web browser according to the structure and logic rooted in the RDF model and schema.

3. For each statement in the resource model, find its predicate (property) and object, as well as the corresponding property of vocabulary defined in the ontology. If the object is not a resource, display the label of properties of vocabulary from the ontology and the object as the value of that vocabulary. The property label is linked to the definition of that property in the ontology. If the object is another resource, display the property label with link to its definition, and the URI of that resource. The URI includes the HTTP request to that resource.

4. The TRB also supplies a link to its parent resource. The parent resource can be obtained by find a resource whose object of one of its statements is the current resource. In the scenario of chemical process, parent resource is the previous step.

5. For its use as a process tracking system, a plug-in component system is added which supplies a menu of all steps in the process and displayed in the left of the page. So one can view details of any step from the menu.

3.3. Data/process tracking from publication

The main is that the acquired chemical knowledge is published in a journal or proceedings. Due to the limited space, it is impossible to present the detailed process how the knowledge is acquired and the full set of experimental raw data in publications. So it is important to link the knowledge published in journal (in electronic form) back to the associated RDF description of the process to enable authorised users to track the original experimental raw data. This process is called the *data/process tracking* or *pub@source* in Comb-e-Chem project. In this subsection we shall integrate the available components we developed to demonstrate the data/process tracking of statistical analysis.

Let us take a typical *semantic* e-print from experiment. It is in HTML format.

Fig. 4. E-print in HTML format.

It also has a semantic description which can be viewed using our RDF browser

Fig. 5. RDF counterpart of e-print.

The main result of the paper is Fig.7 showing the intensity of laser against the input angle for different output angles (0°, 45° and 90°). The continuous curves are fitted from the experiment data using which the users refer to as the ABC-Chi-RD model indicating the different steps in the analysis of the experimental data first to determine coefficients A,B, & C and then further model fitting with these coefficients.

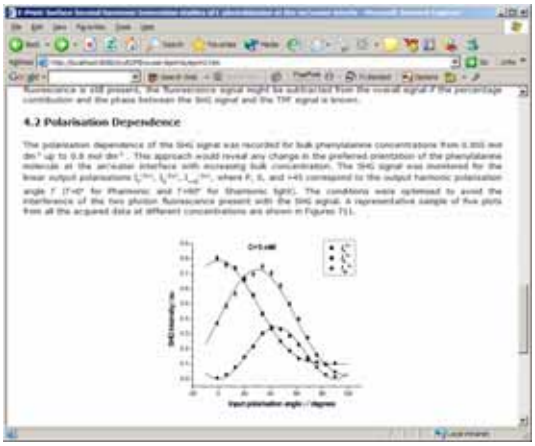

Fig.6. Presentation result in a e-print report

This figure is linked to the RDF description of analysis using the analysis ID. One can simply click this figure to track how the statistical analysis is performed. The following figure (Fig.6) show the first step of the analysis, retrieving the experimental raw data from the database. Actually, each point in the figure represents the average value of 1000 laser runs. So it is for statistical package to fit the data, but not the original raw data in the database. This figure is generated by the interactive SVG API we developed on-the-fly, that means each point on it is active and can be clicked to track further down to the original raw data, namely 1000 lines of the experimental data.

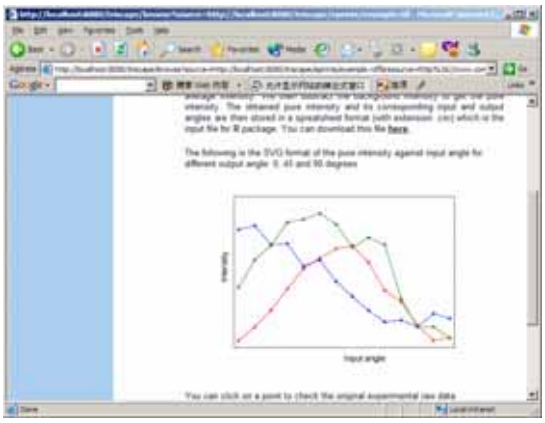

Fig.7. Last step of analysis.

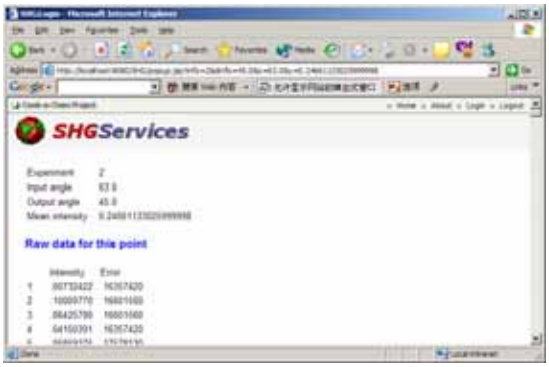

Fig.8. Experimental raw data.

Above example shows who a tracking system works and how an analysis process can be tracked and reused.

4. Application to SHG statistical analysis

Second harmonic generation (SHG) experiment is a surface-specific nonlinear laser experiment used to investigate the surface properties of compounds. The SHG experiment generates a set of raw data which is flowed to the SHG database automatically and can be retrieved through a web service. Then chemists can investigate the surface property of the sample using statistical analysis methods. Depends on their preferences chemists use different statistical package, different assumptions and theoretical models to perform the statistical analysis. There are a number of theoretical models used in the statistical analysis of SHG experiment. Here we only discuss the ABC-Chi-RD model as an explicit example. In this model, the non-linear susceptibility tensor and the hyperpolarisability tensor are expanded to the second term and the laser intensity depends on input and output polarization angles with three unknown complex variables *A*, *B*, *C* to be determined by the statistical fitting from the experimental data. One can then calculate the second order susceptibility tensor and obtain the orientation of molecules at the interface.

For this model the statistical analysis is decomposed into five steps:

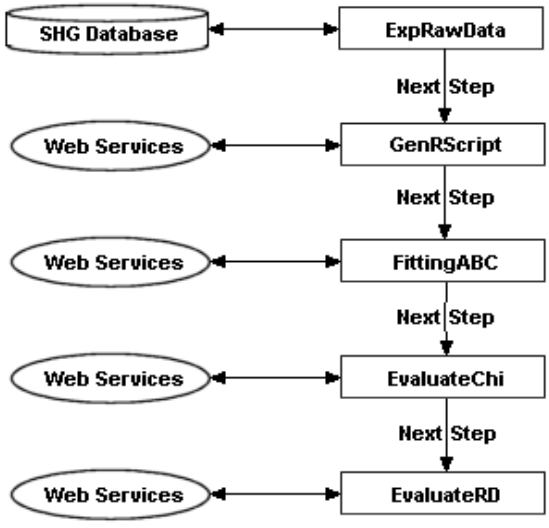

Fig.9. Multi-step decomposition of SHG analysis

Here the step name is the URI fragment, namely the full URI for the first step is http://www.combechem.org/shg/analysis/analysisid.rdf#ExpRawData.

Step 1: **Prepare experimental raw data**.

We first retrieve experiment raw data from the SHG database by specifying the experiment ID and then evaluate the mean value of laser intensity for each run of experiment. When the raw data is retrieved and processed, it is displayed in the web browser graphically using interactive SVG technology (see Fig.6). The original raw data can be obtained by clicking the relevant point on the figure as we indicated in last section. We normally set *isPause=false* as we would like to check the quality of the data.

Step 2. **Generate R script for the analysis**

We use the statistical package R, which has been wrapped up as a web service, to perform the analysis. We should decide which complex parameter (*A*, *B*, or *C*) is assumed to be real and what their initial values are. Then the service generates the R script dynamically.

Step 3. **Fitting complex parameters *A*, *B* and *C***

In this step the web service calls R web services and generated script generated in last step to fit the parameters *A*, *B*, and *C*. We note that the assumption that *A* is real may be not reasonable and leads to an unsatisfactory result. In this case, we will call R program to repeat fitting again by assuming other parameters are real until a good result is achieved. We only record the assumptions leading to the final good result. The fitted curves (intensity against input polarization angle) for different output angles, along with the experimental raw data are plotted in the SVG format and are embedded into the RDF description.

Step 4. **Evaluate Susceptibility tensor**

Susceptibility tensor is linearly related to the parameters *A*, *B* and *C* and five other coefficients that depend on the details of the experimental geometry and some assumptions made about the interface (the Fresnel factors). This step calls the service to evaluate susceptibility tensor.

Step 5. From the susceptibility components and with knowledge of which are the dominant hyperpolarisibility components, one can derive an orientation parameter *D*, which provides useful information about the orientation of the surface molecules in the system.

5. Conclusion and further work

In this paper we have investigated some issues in the statistical analysis of chemical data. We first proposed a formalism to describe and record the analysis process which serves as an e-logbook for humans and analysis descriptor for machine. This formalism is self-defined via ontology and suitable for common understanding and knowledge sharing among collaborators in different visual organizations. We then developed the associated semantic-aware enactment system to perform the analysis from the information in the RDF description of the analysis process. A generic RDF browser is developed to browse and reuse the analysis process which plays important role in the pub@source project. The retrieved raw data is displayed graphically in web browser using interactive SVG tools we developed. Our approach shows that semantic technology is an appreciate technology to record and manage the complicated, flexible and dynamic chemical process and acquired chemical knowledge. We believe that this approach can be further used in other scientific disciplines.

As further works, we shall address the following issues:

*Visual simulation process designer***.** The RDF description of a chemical process should be generated automatically when the process is finished. For a fixed model, the RDF description has the same structure and computer can use a template RDF model to generate the RDF description of a process by inserting the values and results from the statistical packages. Then it is an issue to make a template for a model. As our further work we are going to design a visual template designer to help chemists to design the RDF template for a chemical process.

Process management. We initially record analysis process as the RDF files on the server and those files can be accessed by their URIs. This is problematic in reality when we have large amount of processes each day in a large research group. On the other hand, the management of vocabularies defined in ontology is also an issue as scientific research is so dynamic and flexible that scientists may add or update vocabularies constantly. Jena API has supplied mechanism to store RDF model in mySQL database and to manage the resources via RDQL, a query language for RDF. But it is problematic to deal with large amount of RDF models [28]. 3Store is a management system of triples which supplies efficient mechanism to handle large amount of triples [28]. Integration of our systems with 3Store is currently under development.

Integration of security. Security is always a big issue for web-based applications and our service-originated simulation system is not an exception. A security suit GRIA [29] and enforcement system for Comb-e-Chem, the GRIA has been developed and the integration of our system with it has been in consideration.

Acknowledgements

Authors would like thank EPSRC for the award of the Comb-e-Chem grant.

References

1. I. Foster and C. Kesselman, *The Grid: Blueprint for a New Computing Infrastructure*, 2nd Edition, Morgan Kaufman, 2004.
2. D. de Roure, N.R. Jennings and N.R. Shadbolt, "The Semantic grid: A future e-Science Infrastructure", http://www.semanticgrid.org/documents/semgrid-journal/semgrid-journal.pdf
3. F. Berman and T. Hey, "The scientific imperative", Chapter 2 in *The Grid: Blueprint for a New Computing Infrastructure*, 2nd Edition, Morgan Kaufman, 2004.
4. M.B. Hursthouse, J.G. Frey, S.J. Coles, M.E. Light, M. Surridge, K.E. Meacham, D.J. Marvin, D.C.De Roure and H.R. Mills, "Grid/Web enhancements to the National Crystallographic Service: experiments with an interactive e-science demonstrator", Euroweb 2002 - The web and the Grid: from e-science and e-business.
5. P. Murray-Rust, H.S. Rzepa, M. J. Williamson and E.L. Willighagen, "Chemical Markup, XML and the World-Wide Web. 5. Application of chemical metadata in RSS aggregators", *J.Chem.Inf.Comput.Sci.* and references therein.
6. J.G. Frey, D. De Roure, M.C. Schraefel, H. Mills, H. Fu, S. Peppe, G. Hughes, G. Smith and T.R. Payne, "Context Slicing the Chemical Aether", In D. Millard, Eds. *Proceedings of First International Workshop on Hypermedia and the Semantic Web*, Nottingham, UK, 2003.
7. Comb-e-Chem web site: http://www.combechem.org
8. Smart Lab.
9. Smart Tea project web site: http://www.smarttea.org.
10. WSDL: http://www.w3.org/TR/wsdl
11. GWSDL: http://www.globus.org/
12. UDDI: http://www.uddi.org/
13. http://www-unix.globus.org/toolkit/mds/
14. WSRF: http://www.globus.org/wsrf/
15. S.Coles, J.G. Frey, M.B. Hursthouse, L.A.Carr, and C.J. Gutteridge, "Crystal Structure EPrints: Publication @ Source Through the Open Archive Initiative", *In, British Crystallography Association Spring Meeting 2004*, 6-8 Apr 2004, *Manchester, UK*.
16. J. Taylor, "e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it", http://www.e-science.clrc.ac.uk/web
17. http://taverna.sourceforge.net
18. J. Futrelle, *Emerging Tools for Building Integrated Scientific Data Resources*, http://www.ncsa.uiuc.edu/People/futrelle/ppt/NIH0106.ppt
19. RDF, Resource Description Framework, is a W3C standard. See http://www.w3.org/TR/rdf-concepts/
20. Ontology
21. W3C standard: http://www.w3.org/XML/
22. http://www.w3.org/Graphics/SVG/
23. http://www.w3.org/Math/
24. http://www.xml-cml.org/
25. S. Schreiber, H. Akkermans, A. Anjewierden, R. Hoog and N. Shadbolt, Knowledge Engineering and Management, The MIT Press, London, 1999.
26. http://www.w3.org/2004/OWL/
27. http://www.hpl.hp.com/semweb/jena.htm
28. http://www.aktors.org/technologies/3store/
29. S. Taylor, M. Surridge and D. Marvin, "Grid Reources for Industrial Application", Preprint.