Intelligent Dataspaces for e-Science

IBRAHIM ELSAYED, ADNAN MUSLIMOVIC and PETER BREZANY
University of Vienna
Department of Scientific Computing
Nordbergstrasse 15/C/3, A-1090 Vienna
AUSTRIA
{elsayed|am|brezany}@par.univie.ac.at

Abstract: This work focuses its effort on dataspaces and workflow management, two complementary technologies, which, if applied in conjunction, can provide a highly efficient and powerful scientific data management solution for e-Infrastructures. Key contributions are: (1) a hierarchical and iterative metamodel providing a life cycle view of scientific data showing what ideally should happen to data in e-Infrastructures is presented generally and by the means of two pilot application. (2) An ontology based dataspace model with strong regard on the key dataspace concept - managing relationships among participants - is developed, providing intelligent creation, representation, and searching of semantically rich relationships among primary and derived data sets in e-Science applications. (3) The concept of dataspaces is extended to support the data life cycle in e-Science experiments. At first, supported by the ontology, an e-Science application independent metamodel is set up, which is then applied to describe application-specific e-Science experiments. This profound knowledge about e-Science life cycles, consolidated within instances of the ontology will highly contribute to the development of high productivity e-Science frameworks.

Key–Words: Dataspace, Scientific data management, e-Science, Workflow management, e-Infrastructure, OWL

1 Introduction and Context

Scientific data are being collected to a great extent in various research domains. They are stored on multiple national sites in various databases. Scientific collaborations are targeting to provide access to these *primary data* by the means of an e-infrastructure. Through portals scientists are able to undertake these data for significant analyses in the context of their interest. The output of these analyses aims at defining a large number of predictions and might provoke further experimentation, which in turn may take days or weeks, depending on computational and human resources available. However, the resulting data – called *derived data* – that have arisen from the research task represents valuable information not only to the acting research group, but also to other groups with respect to other research areas.

Main objective is to link those derived data with their corresponding primary data by providing semantically rich relationships. Further, to make both relationships and data available within a space of data for people from various groups of organizations who might have use of it and who want to collaborate by the means of virtual organizations in the context of an e-infrastructure. To fulfill these goals, we further develop a dataspace paradigm introduced in [10].

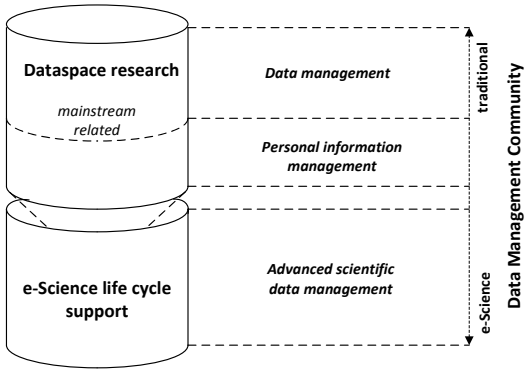

Figure 1: Dataspace research extension

A dataspace consists of a set of participants and a set of relationships among participants [10]. A participant can be any element containing data in some way. Relationships describe how two participants are related to each other. Relationships can be expressed by single word-relationships, such as replica-of, related-to, view-of, etc. In the extreme example they can be semantic mappings of database schemas.

The initial ideas on managing dataspaces have started to evoke interests of the data management community, however most effort is related to the mainstream and so far not considered in scientific data

management. In Figure 1 we illustrate our extension to the mainstream dataspace research providing advanced scientific data management.

The challenge is to raise up the level at which data is managed [12]. Systems providing the required services over dataspaces are considered to be Dataspace Support Platforms (DSSPs) [11]. In [6] we have defined such a system as *a set of software programs that controls the organization, storage and retrieval of data in a dataspace. It also handles the security and integrity of the dataspace.*

The success of a dataspace will be highly dependent on the power of the used relationship concept as well as its flexibility. Rich relationships between the participants are going to be the backbone of such a system, with the basic necessity to support semi-automatically creation of them as well as their improvements and maintenance.

The development of a suitable relationship model customizable towards various application needs is therefore an important issue, to be challenged by the *e-Science life cycle ontology*, whose major role is to describe and semantically enrich the existing relationship among primary and derived data sets in e-Science applications. This is the basis for elaboration of intelligent and more powerful paradigms for the creation, representation and advanced searching of relationships among participants of a dataspace.

The rest of the paper is organized as follows. Chapter 2 presents the e-Science life cycle view. Its major activities are described in Section 2.1. In Section 2.2 the life cycle metamodel is described. Scientific Dataspaces are discussed in Section 2.3 and its search and query features in Section 2.4. Related Work is addressed in Chapter 3. Application scenarios are described in Chapter 4 and finally in Chapter 5 the paper is concluded .

2 Life Cycle View of Scientific Data

In order to elaborate how dataspace concepts can support e-Science, we have investigated what happen, or better what should ideally happen to data in e-Science applications. The result of this investigation is an iterative and hierarchical metamodel with five main activities, represented in Figure 2, which we define as following: *The e-Science life cycle - a domain independent ontology-based iterative metamodel, tracing semantics about procedures in e-Science applications. Iterations of the model - so called e-Science life cycles - organized as instances of the e-Science life cycle ontology, are feeding a dataspace, allowing the dataspace to evolve and grow into a valuable, intelligent, and semantically rich space of scientific data.* First

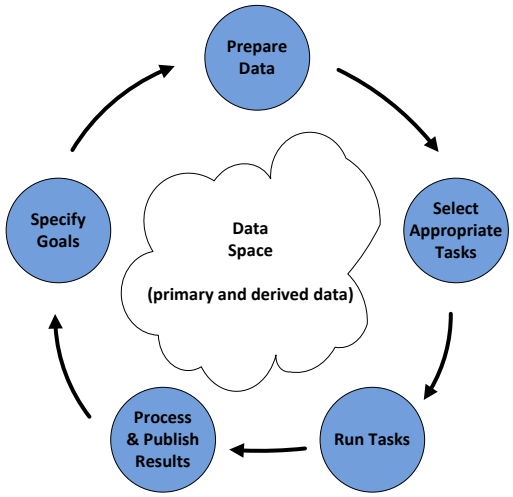

Figure 2: The e-Science Life Cycle

we provide an overview of these activities and then in Section 2.1 a more detailed discussion.

At the beginning of the life cycle targeted goals are specified, followed that a data preparation step including pre-processing and integration tasks is fulfilled. Further appropriate data analysis tasks are selected and applied on the prepared dataset of the previous step. Finally achieved results are processed and published, which might provoke further experimentation and consequentially specification of new goals within the next iteration of the life cycle. The outcome of this is a space of primary and derived data with semantically rich relationships among each other providing (a) easy determining of what data exists and where it resides, (b) searching the dataspace for answers to specific questions, (c) discovering interesting new data sets and patterns, and (d) assisted and automated publishing of primary and derived data.

Each activity in the life cycle shown in Figure 2 includes a number of tasks that again can contain a couple of subtasks. For instance, the activity *Prepare Data* covers, on a lower level of abstraction, a data integration task gathering data from multiple heterogeneous data resources that are participating within an e-Infrastructure. This task consists of several steps that are organized into a workflow, which again is represented at different levels of abstraction - from a graphical high level abstraction representation down to a more detailed specific workflow language representation, which is further used to enact the workflow.

2.1 e-Science Life Cycle Activities

1 *Specify Goals* - Scientists specify their research goals for a concrete experiment, which is one iteration of the entire life cycle. This is the start-

ing activity in the life cycle. A textual description of the objectives, user name, corresponding user group, research domain and other optional fields like a selection of and/or references to an ontology representing the concrete domain is organized by this activity.

2 *Prepare Data* - Once the objectives for this life cycle are either specified or selected from a published life cycle that was executed in the past, the life cycle goes on with the data preparation activity. Here it is specified which data sources are used in this life cycle in order to produce the final input dataset, by the data integration process. For example, the resource URI, name, and a reference to the OGSA-DAI[1] Resource File is recorded. The final dataset as well as the input data sets are acting as participants in the dataspace and are referenced with an unique id. Additionally, the user specifies a short textual description and optionally some keywords of the produced data set.

3 *Select Appropriate Tasks* - In this activity the data analysis tasks and to be applied on the prepared dataset are selected. In e-Science applications it is mostly the case that various analytical tasks, for instance the widely used data mining techniques, are executed successively. The selected tasks, which are available as Web and Grid services, are organized into workflows. For each service, its name and optionally a reference to an ontology describing the service more precisely is captured. Also for the created workflow, its name, a short textual description, and a reference to the document specifying the workflow are recorded.

4 *Run Tasks* - In this activity the composed workflow will be started, monitored and executed. A report showing a brief summary of the executed services and their output is produced. The output of the analytical services used is represented in the Predictive Model Markup Language (PMML) [3] which is a standard for representing statistical and data mining models. PMML documents represent derived data sets, thus they are managed as participants of the scientific dataspace and considered as resources by this activity.

5 *Process and Publish Results* - This is the most important activity in order to allow the underlying dataspace to evolve and grow into a valuable, powerful, semantically rich space of scientific data. Based on the settings of the user, one automatically publishes the results of the data mining tasks, represented in PMML as well as all semantical information captured in the previous activities. Different publishing modes allow to restrict access to selected collaborations, user groups, or research domains.

[1]OGSA-DAI [13] is the de facto standard for data access and integration for relational and xml data as well as file resources.

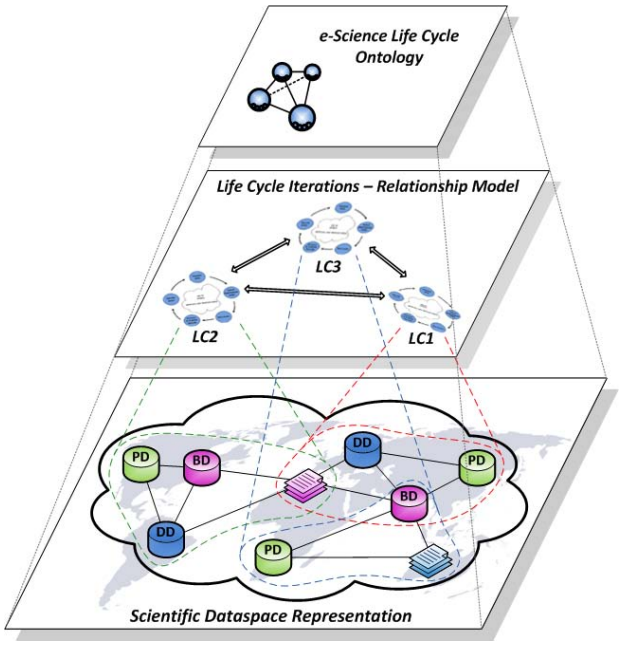

Figure 3: Abstraction Layers of Scientific Dataspaces *(PD - Primary Data, DD - Derived Data, BD - Background Data)*

2.2 Life Cycle Metamodel

Ontological knowledge is sharable, understandable to machines, and supports the enrichment of data sources and relationships at the semantic level. Therefore we have developed the *e-Science life cycle ontology*, which organizes the concepts and coherences of the above described e-Science life cycle activities. Strong regard was put on considering input (primary) and output (derived) data sets as well as relevant background data (e.g. domain ontologies, data statistics, OGSA-DAI resource files, workflow descriptions, etc.) for modeling an intelligent relationship paradigm.

At first, supported by the ontology, a metamodel independent from the various e-Science domains is set up. Then this metamodel is applied to describe domain-specific iterations of the e-Science life cycle, which describe the relationship among data participating within the scientific dataspace, illustrated as different abstraction layers in Figure 3. One iteration

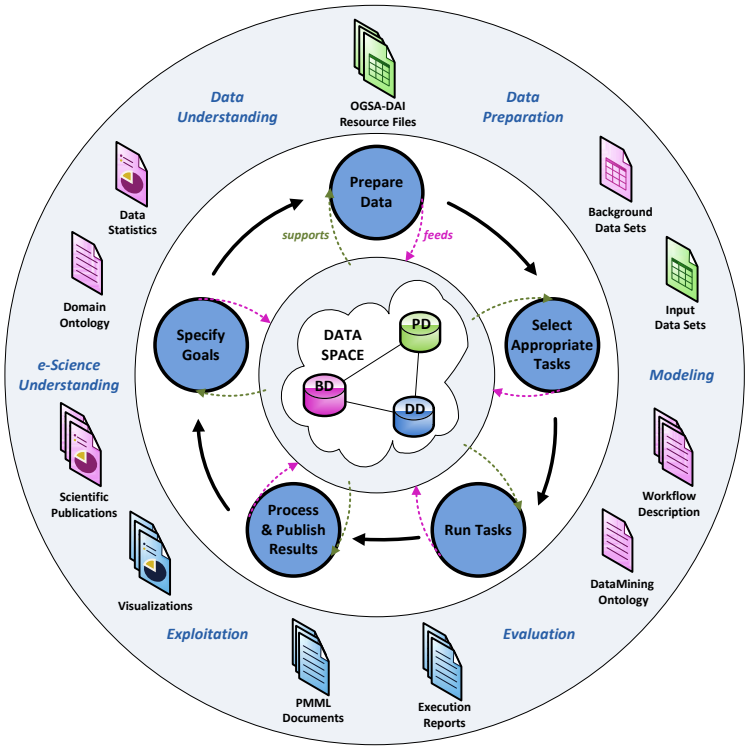

Figure 4: Environment of the e-Science Life Cycle

of the e-Science life cycle has, in short, a goal specification, a set of input data (primary data), a set of output data (derived data), a set of background data, and a set of activities describing what has been done to the input data sets in order to produce the output data sets. These data sets are populating the scientific dataspace, enriched with semantic relationships among each other, described by its corresponding life cycle iteration. We can see from this, that the dataspace is evolving with an increasing number of life cycles.

This profound knowledge about iterations of the e-Science life cycle, consolidated within instances of the ontology represents an intelligent relationship model for scientific dataspaces, because it provides (a) creation, (b) representation, and (c) searching of semantically rich relationships among dataspace participants. Realization of a scientific dataspace paradigm will highly contribute to the development of high productivity e-Science frameworks.

With the help of the e-Science life cycle ontology, it is made possible for scientists to describe, execute and share their e-Science experiments with others in an efficient manner. Further, it is feasible to search for published instances of the life cycle or even for instances of single activities of the life cycle. In such a way, a scientist could search for all published goal specifications corresponding to his research domain, by searching for a given domain name. The dataspace will then provide not only the published instances of the activity, but also the complete instance of the e-Science life cycle, including the inputs of other activities and its corresponding results. In addition, it will give hints about similar life cycle iterations by using the semantically rich relationships described by the ontology. With this in mind, it will be easier for research groups to engage collaboration, provide knowledge transfers within collaborations and among different research groups with respect to different research areas. In conclusion, the e-Science life cycle metamodel is likely to unify the process of publishing primary, derived, and background data sets as well as the their interconnection and make it easy for scientists to register, describe and execute new e-Science experiments and for users to find, explore and understand these applied experiments. The e-Science life cycle ontology that we have developed is available at http://www.gridminer.org/e-sciencelifecycle/.

2.3 Scientific Dataspaces

Scientific dataspaces will be set up to serve a special subject, which is on one hand to semantically enrich the relationship of primary and derived data in e-Science applications and on the other hand to integrate e-Science understandings into iterations of the

life cycle model allowing scientists to understand the objectives of applied e-Science life cycles. Figure 4 shows the environment of e-Science life cycle. In particular there is a set of participants participating to one ore more activities of the e-Science life cycle. Each activity feeds the dataspace with new participants, as for example the activity *Specify Goals* adds new domain ontologies, the activity *Prepare Data* adds new final input data sets as well as OGSA-DAI resource files, and the activity *Select Appropriate Tasks* adds new workflow description documents, while the activity *Run Tasks* adds new PMML documents describing the data mining model applied, and finally the activity *Process and Publish Results* adds new documents visualizing the achieved data mining outputs. All these participants belong to at least one or more e-Science life cycles, expressed as instances of the ontology describing its relationship and interconnection to a great extend.

Each iteration of the life cycle metamodel will produce a new instance of the ontology. Based on the publishing mode, set by the scientist who accomplished the life cycle, the whole instance will automatically be published into the dataspace and thus is available to other users of a wider collaboration with respect to other research areas. We distinguish between four publication modes, (1) *free access*, (2) *research domain*, (3) *collaboration*, and (4) *research group*. Users will have access to sets of participants available in the scientific dataspace, depending on their assigned role. By this, the concept of managing sub-dataspaces is realized. A sub-dataspace contains a subset of participants and a subset of relationships of the overall dataspace. There can be sub-dataspaces setup for different domains, then for different research collaborations and even for single research groups. e-Science experiments that were published using the *free access mode*, will participate in the overall dataspace, thus its participants and the life cycle instances are accessible for every one having access to the scientific dataspace. In order to access data of a specific life cycle iteration, that was published using the *research group mode*, it will be necessary to be member of that specific research group, as the data will be participating only in the corresponding sub-dataspace.

2.4 Search and Query Scientific Dataspaces

Based on an instance of our unified e-Science metamodel, search and query services can be provided for all the participants of the corresponding scientific dataspace. Hence, it is possible to forward a keyword query to all participants, which has the aim to identify relevant data sets. However, each query submitted to the scientific dataspace, will receive not only the matching data but also data of its followed e-science activities. For instance it will be possible to receive what mining task were applied on a discovered dataset, the concrete workflow, the workflow report, the results presented in PMML and its corresponding visualizations.

Using *SPARQL* query language for RDF [14] and semantically rich described e-Science life cycles, consolidated within instances of the ontology, keeping relationships among each other, the dataspace is able to provide answers to specific questions, such as the following:

A *"I have detected a model error and want to know which derived data products need to be recomputed."*

B *"I want to apply a NIGM-analysis on meridian HE GU. If the results already exist, I'll save hours of computation."*

C *"Is there any experiment done on meridian BA XIE"*

Through portals and advanced user interface scientists are supported with the needed tools, which enable users to express search queries visually and in an simple way. The output is simply *SPARQL*, which allows to query instances of an ontology efficiently. This is part of our ongoing work, currently under investigation. However, the basis for intelligent dataspaces for e-Science is developed and have cleared the way towards developing high-productivity e-Science frameworks.

3 Related Work

So far dataspace paradigms have been mainly considered in terms of personal information management. In [10] the concepts of dataspaces are introduced in a visionary way. Influenced by this vision a personal dataspace management system with an own data model and query language is presented in [5]. Other projects like iRods [15], SRB [1], and the IBM's commercial product Websphere Information Integrator [8] have considered some dataspace concepts in their architecture; however the key dataspace paradigm, which is to provide semantically rich relationships among participants, is not taken into consideration. A first approach towards realization of dataspaces regarding the Grid is given in [6]. VDS, the GriPhyN Virtual Data System, formerly known as Chimera [9], provides a virtual data system for managing and tracking different aspects of various data transformations and its results in workflow environments stored in a

virtual data catalog, where the produced data and the steps being used to produce the data can be later retrieved for further analysis. myExperiment [4] is a system designed to support scientific collaborations and the life cycles of workflows. It creates a virtual research environment allowing the scientific community to share and execute scientific workflows in the context of their research forming a distributed community of scientists. Both systems, Chimera and myExperiment are targeting to model relationships from primary and derived data through collecting provenance data of executed workflows. However, dataspace search and query features are not tightly focused.

4 Application Scenario

A first application highly profiting from the above described life cycle of scientific data is located in the field of Traditional Chinese Medicine (TCM). According to the basic TCM theory, the human body has 14 acupuncture meridians, which are a secret to our biological and medical knowledge. Within the China-Austria Data Grid project [2] investigations on how high-tech measurement and it technologies can support the exact estimation of the meridian status are observed. Therefore an e-Infrastructure supporting computation and data management services as well as access to meridian measurement databases is set up among the nine participating research institutions in China and Austria. The analytical techniques used (electro signal and subcutaneous impedance measurement) have collected huge amounts of data referred to as meridian measurement data, which again, as a result of followed data analysis, have produced a large number of derived data products. In order to use this large amount of valuable information, it is necessary to make available a space of data accessible for other research groups targeting different research areas. Data published in the scientific dataspace set up for this application, is available for further data mining studies aiming at further improvements in diabetic care and meridian theory resulting in higher patient comfort. An e-Health service aiming in the treatment of diabetic patients [7] is the first output of commonly achieved research results within the scientific collaboration. Consequently, the scientific technological basis for extension to other domains will be established as part of targeted ongoing work.

In Figure 5 we show an simplified extract of an instance of the activity *Run Tasks*, taken from the scientific dataspace which was set up for the participating researchers of the above introduced scientific collaboration.

The corresponding OWL class is named *taskExecution*. This example defines three outputs corresponding to the experiment, (1) the PMML document representing the output of the neural network model executed within the experiment, (2) its corresponding visualization document, and (3) a report summarizing the workflow execution; all considered as derived data.

```
<taskExecution rdf:ID="LC1_TE1">
...
 <hasOutput>
  <visualization rdf:ID="visualisations_001">
   <hasRef rdf:datatype="http://.../XMLSchema#string">
    http://.../CADGrid_BPNNVisualization_001.svg
   </hasRef>
   <hasType rdf:datatype="http://.../XMLSchema#string">
    pmml visualisation
   </hasType>
  </visualization>
 </hasOutput>
 <hasOutput>
  <pmmlDocument rdf:ID="pmmlDocument_001">
   <hasRef rdf:datatype="http://.../XMLSchema#string">
       http://.../cadgrid/CADGrid_BPNN_4711.pmml
   </hasRef>
  </pmmlDocument>
 </hasOutput>
...
</taskExecution>
```

Figure 5: Instance of a taskSelection Activity

A second application is currently in the beginning. The breath-gas analysis for molecular-oriented detection of minimal diseases project, in short BAMOD project [16] is focused onto the diagnosis of minimal disease and early stages of lung and oesophageal cancer. Relevant source datasets include data produced from diverse mass spectrometers and corresponding patient data. The volume of these datasets is growing daily as new experiments are done continuously at different breath gas research centers. The breath-gas analysis community is investigating and screening for hundreds of compounds in the exhaled breath. The analytical techniques used include various statistical and data mining techniques supporting identification of specific disease markers. The scientific community interested in the analytical results of breath-gas analysis is also geographically distributed. The scientific dataspace model described in this paper allows to fully utilize this large amount of scientific data available at each breath gas research center.

5 Conclusions

In this paper we have presented a novel methodology and associated informatics to support the interaction among specific research groups by the means of advanced scientific data management in e-Infrastructures. Key contributions are: (1) a hierar-

chical and iterative metamodel providing a life cycle view of scientific data showing what ideally should happen to data in e-Infrastructures while they are processed is presented generally and by the means of one pilot e-Science application. (2) The e-Science life cycle ontology, organizing the concepts and coherences of e-Science life cycle activities as classes and properties, is developed. (3) The dataspace paradigm presented in [10] is further developed by considering its major research challenge "managing relationships among participants" in order to explicitly support the existing relationship among primary and derived data in scientific collaborations.

The intelligence of the proposed e-Science life cycle model lies in its capability as customizable relationship model for scientific dataspaces, as it covers the creation, representation and searching of semantically rich relationships among participants of a dataspace. It enables researchers to find not only relevant primary data in connection with its derived data, but also lot of semantics about what was initially done with the data, such as which data preprocessing methods have been applied, which data mining and analysis models have been used, which result visualizations are available etc. Further it points to relevant background and ontological data, such as descriptions of applied services, models, research domains etc.

All these information is meant to be the semantically rich relationship among primary and derived data described by the e-Science life cycle ontology. Additionally scientists will retrieve information about the goals specified, which domain it corresponds, and whom to contact in case of interest for engaging collaborations, in short, users will understand for what reason a specific e-Science life cycle was applied, which we summarize by the meaning of *e-Science Understanding*.

Acknowledgements: The work described in this paper is supported by the "Austrian Grid Phase 2" project, funded by the Austrian BMWF (Federal Ministry for Science and Research).

References:

[1] A. Rajasekar et al. Storage resource broker - managing distributed data in a grid. *Computer Society of India Journal, special issue on SAN* 2003.

[2] CADGrid. The China-Austria Data Grid project. *http://www.par.univie.ac.at/project/cadgrid*, 2007.

[3] Data Mining Group. The Predictive Model Markup Language (PMML). *http://www.dmg.org/v3-2/* July 2008.

[4] D. De Roure, C. Goble, and R. Stevens. Designing the myexperiment virtual research environment for the social sharing of workflows. *Proceedings of the International Conference on e-Science and Grid Computing*, Dec. 2007.

[5] J.-P. Dittrich. iMeMex: A platform for personal dataspace management. *Proceedings of SIGIR Workshop on Personal Information Management (PIM)*, August 2006.

[6] I. Elsayed, P. Brezany, and A M. Tjoa. Towards realization of dataspaces. *Proceedings of International Conference on Database and Expert Systems Applications (DEXA)*, 2006.

[7] I. Elsayed, J. Han, T. Liu, A. Woehrer, F. A. Khan, and P. Brezany. Grid-enabled noninvasive blood glucose measurement. *Proceedings of the International Conference of Computational Science (ICCS)*, June 2008.

[8] F. Erfan. Maintain federated data using websphere information integrator autonomic monitoring tools. *IBM Technical Article*, 2005.

[9] I. Foster, J. Vckler, M. Wilde, and Y. Zhao. Chimera: A virtual data system for representing, querying, and automating data derivation. *Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM)*, September 2002.

[10] M. Franklin, A. Halevy, and D. Maier. From databases to dataspaces: A new abstraction for information management. *ACM SIGMOD*, December 2005.

[11] A. Halevy, M. Franklin, and D. Maier. Principles of dataspace systems. *Proceedings of the ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS)*, Dec 2006.

[12] A. Halevy. Why your data won't mix. *Queue*, 3(8):50–58, 2005.

[13] M. Antonioletti et al. OGSA-DAI 3.0 - the whats and the whys. *Proceedings of the UK e-Science All Hands Meeting* 2007, September 2007.

[14] E. Prud'hommeaux and A. Seaborne. SPARQL query language for RDF. *http://www.w3.org/TR/rdf-sparql-query/*, January 2008.

[15] A. Rajasekar et al. A prototype rule-based distributed data management system. *HPDC workshop on "Next Generation Distributed Data Management*, 2006.

[16] Breath-gas analysis for molecular-oriented detection of minimal diseases. European Research Area SIXTH FRAMEWORK PROGRAMME. *PRIORITY 1 - Life Science, genomics and biotechnology for health*,, Proposal/Contract no.: LSHC-CT-2005-019031.