IWPLS'09

Pages 1–8

Towards Realization of Scientific Dataspaces for the Breath Gas Analysis Research Community

Ibrahim Elsayed [1,*], Thomas Ludescher [2], Konrad Schwarz [3], Thomas Feilhauer [2], Anton Amann [3], and Peter Brezany [1]

[1]Department of Scientific Computing, Faculty of Computer Science, University of Vienna, Nordbergstrasse 15/C/3, 1090 Vienna, Austria

[2]Research Center for Process and Product Engineering, Vorarlberg University of Applied Sciences, Hochschulstrasse 1, 6850 Dornbirn, Austria

[3]Department of Operative Medicine, Innsbruck Medical University, Innsbruck, Austria and Breath Research Institute of the Austrian Academy of Sciences, Innsbruck, Austria

September, 2009.

Associate Editors: Sandra Gesing and Jano van Hemert

ABSTRACT

Motivation: Scientific dataspaces aim at providing associated mechanisms for managing semantically rich relationships among scientific data sources (primary data) and its corresponding findings (derived data). The latter result from a set of activities defining concrete preprocessing and analysis methods (background data) that were applied to a source dataset. To keep track of scientific experiments that are being conducted by members of a scientific community these experiments should be linked with user information i.e. institutional affiliation, email address, working field, etc. of the scientist who conducted the experiment. This paper deals with the development of a scientific dataspace for the breath gas analysis community. Breath gas analysis is an emerging new scientific field with a growing international scientific community addressing many different breath gas studies in terms of investigating and screening for hundreds of compounds in exhaled breath gas. The purpose of *breath gas analysis scientific dataspaces* is to enable collaborating scientists and institutions several important activities, which include: (a) access to distributed breath gas data and analytical resources collected and developed at different research institutions around the world and (b) to easily contribute to and leverage the resources of an international- and national-scale, multi-institutional environment.

Results: This paper describes the conception and prototypical implementation of the scientific dataspace paradigm for breath gas analysis. The scientific dataspace is evaluated on top of applications from breath gas analysis.

Availability: We discuss the scientific dataspace paradigm in the context of applications from breath gas analysis, however the concepts introduced can also be deployed in other research domains.

Contact: elsayed@par.univie.ac.at

1 INTRODUCTION

The project *Austrian Grid 2* is the second phase of the national grid initiative funded by the Austrian Federal Ministry of Science and Research[1]. The Austrian Grid consortium combines Austria's leading researchers in advanced computing technologies with well-recognized partners in grid-dependant application areas. An overview of the Austrian Grid project is provided in 4. In 2007, three partners of the Austrian Grid consortium have started a research collaboration with the intention to realize a *Secure Infrastructure for Scientific Data Life Cycle Management* on top of applications in the field of breath gas analysis. This project is described in 6. That paper presents motivation, an overview of the architecture, and the mechanisms involved in providing such a secure infrastructure. It also introduces use case scenarios depicting the current sequence of events in conducting breath gas analysis experiments. Our data management approach is based on the Grid and *dataspace* concepts and a new positioning of the MATLAB® language and computing environment. The Grid is an infrastructure that enables flexible, secure, and coordinated resource sharing among dynamic collections of individuals and institutions 11. The idea of a future data management paradigm called *dataspace* was introduced by Franklin et al. 12 and also addressed by authors of this paper in 8, 9. The goal is to manage a dataspace, rather than a database or other dataset type. Dataspaces are modeled as participants (datasets) and relationships. The concepts of a scientific dataspace paradigm are described in 9. It introduces a specific model of the e-Science life cycle, which the authors defined as:

> *... 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.*

The major role of the e-Science life cycle ontology 15 is to describe and semantically enrich the existing relationship among primary, background and derived datasets in e-Science

*elsayed@par.univie.ac.at

[1] BMWF (Federal Ministry of Science and Research) Contract: GZ BMWF-10.220/0002-II/10/2007 www.austriangrid.at

applications. Scientific experiments described by the e-Science life cycle ontology are referred to as *Life Cycle Resources* (LCR). This paper deals with an implementation of a scientific dataspace paradigm on top of the e-Science life cycle ontology for the breath gas analysis research community.

Breath gas analysis is an emerging new scientific field with a large scientific community spread all over the world and with a promising significant impact on many application domains. Recent results suggest that detection of different kinds of cancer is possible by means of breath gas analysis beyond the scope of available diagnostic methods. There is strong evidence that specific cancers can be detected using the concentration pattern of volatile compounds in exhaled air 17.

The growing international community of breath gas researchers is addressing many different studies including endogenously-derived volatile compounds such as emitted by exhaled breath, from skin, urine, faeces, and flatulence. They are currently at the stage of developing new analytical methods, collecting pilot data for cancer and other diseases and identifying marker compounds. Breath gas researchers are investigating and screening for hundreds of compounds in the exhaled breath gas. The analytical instruments and techniques used include GC-MS (Gas Chromatography Mass Spectrometric), PTR-MS (Proton-Transfer-Reaction Mass Spectrometry), SIFT-MS (Selected-Ion-Flow-Tube Mass Spectrometry), IMS (Ion Mobility Spectrometry), laser spectrometry - as well as various statistical and data mining techniques supporting identification of specific markers. Currently, during the investigations new different sampling and analytical techniques for breath gas measurements are being developed. More information about the breath gas research community and their research work is available in 3 and online at the website of the International Association for Breath Research (IABR - *www.iabr.li*), which was founded in May 2005 in Vienna.

The purpose of *breath gas analysis scientific dataspaces* is to enable collaborating scientists and institutions several important activities, which include: (a) access to distributed breath gas data and analytical resources collected and developed at different research institutions around the world, (b) to easily contribute to and leverage the resources of an international- and national-scale, multi-institutional environment. This will strongly support global collaborations of scientists, improve decisions and increase the chance and scope of discoveries in the breath gas research domain. In this context there is a need for a supporting information infrastructure providing advanced data management and analytical methods and their composition into scientific experiments allowing the scientist to keep track of their e-Science activities and to publish corresponding results of breath gas analysis linked together with their source data and semantics about the purpose of the experiment.

Source data obtained from the previously mentioned analytical methods are referred to as breath gas measurement data and are saved, together with corresponding patient data, locally at each research center. These data are the fundament for simulation and modeling by the acting research group, e.g. observation of the correlation between isoprene breath content and cholesterol level in blood. Such breath gas experiments, if evaluated on large amount of real data, allow a more detailed analysis including e.g. gender-specific relation with respect to age-dependency 14. 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 *derived data*, that have arisen from the research task represent valuable information not only to the acting research group, but also to other groups with respect to other main focuses.

Breath gas research specific dataspaces will be set up to serve a special subject, which is on one hand the relationship of source data (exhaled breath gas measurement data) and its derived data (e.g. specific cancer markers) in breath gas analysis experiments and on the other hand to integrate scientific knowledge into these experiments.

This paper discusses the design of a scientific dataspace paradigm for the breath gas analysis scientific community. The rest of this document is set out as follows: Section 2 introduces selected use case scenarios, lists the currently known uses of the scientific dataspace and describes the information a breath gas analysis scientific dataspace should contain and how this information is organized; Section 3 describes the operations the dataspace must support with an indication of how breath gas researcher will interact and take advantage of the services the scientific dataspace provides. In Section 4 we discuss the iterative approach of conducting breath gas experiments on top of the e-Science life cycle model and in Section 5 we conclude the paper.

2 APPROACH

In our previous work we have introduced the e-Science life cycle ontology 9, whose major goal was to semantically enrich the existing relationship among primary, derived, and background datasets that emerge during the life cycle of scientific data. The goal is to implement a data management solution based on the concepts of dataspaces for large-scale and long-term management of scientific data. Our approach is to preserve both, relationships and data together within a dataspace to be reused by owners and others. To enable their reuse, data must be well preserved. The effects of data loss can be economic, because the experiments have to be re-run, but in some cases data loss represents an opportunity lost forever 18. To look on the bright side of things, well preserved data represent valuable information, which can lead to fruitful scientific findings in the acting and also in related research areas. Preservation of scientific data is therefore a major requirement, which can best be established if the full life cycle of data is addressed. This is achieved in our approach by the e-Science life cycle ontology. In the following sections we describe how breath gas analysis experiments are consolidated, what actions are involved in terms of data access, analysis, and publication. We also show how the full life cycle of data of breath gas experiments is organized by the breath gas analysis scientific dataspace.

2.1 Breath Gas Analysis Dataspace Usage

The use case scenario depicting the current sequence of events in conducting breath gas analysis experiments is illustrated in Fig. 1, which is reproduced from the position paper 6. It presents an overview, with the *Server* being an isle system gathering the data. Import and export activities on this system are protected by smart card security measures.

The implementation of a secure infrastructure, which provides the needed services for breath gas researchers to efficiently and securely

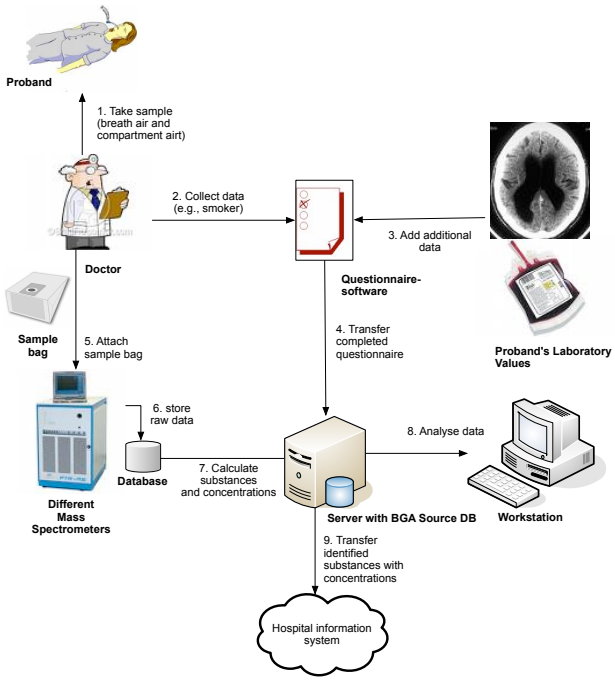

Fig. 1. The use case depicting the current sequence of events 6 - Steps 2 to 4 present the collection and subsequent transfer (i.e. manual import) of personal data to the server hosting a breath gas analysis database. Steps 5 to 7 involve the collection and preprocessing of the probands analysis data. Step 8 is the actual analysis done using a workstation employing Matlab.

perform steps 1 to 7 is part of the project described in 6. However, the present document focuses on the realization of a scientific dataspace for the breath gas scientific community. We assume that a secure and isolated database storing masspectrometer and patient data is already set up and administered by the corresponding *Regional Head Service* as described in 6. In the following we refer to this database as the *source database*.

We are aware that in order to successfully establish a large-scale scientific dataspace for the breath gas analysis community with a large amount of well described experiments of exhaled breath, we rely on active participation of members of the scientific community. Therefore we have - in cooperation with leading breath gas researchers - defined a number of actions that a researcher is conducting during the process of performing breath gas studies. We have then mapped the specific actions to activities of the e-Science life cycle. We also indicate where (on the community portal or within MATLAB) these actions should be taken. The actions, its corresponding e-Science life cycle activities, and the"place" where they are taken are listed in Table 1.

Based on this common understanding we are designing and implementing the tools that support the breath gas researchers in conducting their breath gas studies according to actions described above.

2.2 Breath Gas Analysis Dataspace Content

The breath gas analysis scientific dataspace consists of a set of databases set up by a *dataspace designer* and administered by a *dataspace manager*. In particular, there are:

- *primary databases* for storing the *final input dataset*. This kind of datasets are being created after steps 1 to 7 of the use case described above. They can be retrieved by submitting a query to one of the breath gas research *source databases* available to the acting researcher. However, in order to handle the issue that data and probably also structure of the data might change over time, we store this created *final input dataset* into a separate database called *primary database*, which is designed particularly for this purpose. A typical final input data set is less than 1 MB in its MATLAB structure (`.mat` file), which is a binary data container format used by MATLAB. It may include arrays, variables, functions, and other types of data. It is organized in three blocks as follows (1) patient data - includes all collected data of different test persons such as proband value (e.g. height, weight), burden (e.g. smoker/nonsmoker), labor value (e.g. blood parameter), etc. (2) the system information block manages all system settings for the two databases like all users with their corresponding user groups, studies with their questionnaires, different mass spectrometers with status, container types with status, and (3) the analysis data part includes all information on a specific measurement of a sample such as mass spectrometer type, used container, collection date, measurement date, data (substances with concentration and additional information), etc.
- *background databases* for storing the analytical methods used to analyze the *final input dataset* i.e. MATLAB functions in M-files (ASCII-textfiles containing MATLAB commands and functions). Typical size of a single M-file is less that 500 KB.
- *derived databases* for storing the results of analyses tasks. Once the breath gas researcher has accomplished his experiment, he can publish the results of his analysis. Therefore we take advantage of MATLAB's publishing function, which lets you export results as plots or as complete reports. Using the MATLAB editor, researchers can automatically export their MATLAB results, including the code into XML and various other file formats, e.g. HTML or LaTeX. Since typical breath gas experiments include plotting functions, this dataset usually includes jpg images. Typical size is less that 2 MB.
- *other databases* i.e. *volatomics databases*, which contains data from studies of exhaled breath gas as well as other sources of endogenously-derived gases such as skin, urine, faeces, and flatulence. There is a small number of mandatory data fields, which will record basic metadata of studies of exhaled breath gas. Records in this database will be based on a specific report from some published source, such as a journal, conference proceeding, or on-line publication. This means that records in this database will be made only after the study has shown some significant research results that were already scientifically published.

In addition there are special databases set up for storing the instances of the e-Science life cycle ontology, which are defined

Table 1. Definition of breath gas analysis actions and their mapping to e-Science life cycle activities

action	description	e-Science life cycle activity	place
1	Login to the system.	*GoalSpecification*	Portal
2	Definition of the goals of the study.	*GoalSpecification*	Portal
3	Collection of the probands analysis data (masspectrometer and patient data; this action covers steps 1-7 of the current sequence of events depicted in Fig. 1).	*DataPreparation*	Lab/Questionnaire Software
4	Formulation and submission of a query to the *source database*. This action generates the *final input dataset*, which will be included into the dataspace as participant marked with type "primary data".	*DataPreparation*	Portal
5	Selection/development of the analytical method for analyzing the prepared dataset. This action generates the analytical methods, which will be included into the dataspace as participant marked with type "background data".	*TaskSelection*	MATLAB
6	Execution of the analytical methods.	*TaskExecution*	MATLAB
7	Process the results and export them using MATLAB's publication function into XML. This action generates the results report, which will be included into the dataspace as participant marked with type "derived data".	*ResultPublishing*	MATLAB
8	Set publication mode to conducted experiment.	*ResultPublishing*	Portal

in RDF 2. This is based on using OGSA-DAI 19 to present an RDF store. OGSA-DAI is the de facto standard for data access and integration for relational and XML data as well as file resources.

In contrast to data warehouse systems, data owners don't lose control over their data stored in above introduced participants of the scientific dataspace. This is due to the publication concept that is provided by the e-Science life cycle ontology (see Section 3). Breath gas researcher can limit access to their recources by assinging different publication modes. Also, single breath gas experiments and even whole studies can still be removed from the dataspace by the corresponding data owner. Another important distinction to data warehousing is in user management. The scientific dataspace provides multiple research groups and types of users to share their scientific resources without the time consuming preparation of separate data marts as required in data warehousing. Furthermore, in data warehousing systems data regularly need to be extracted, cleaned, and loaded into the system; even though this process is fully automated, it still consumes time as does regular maintenance. This time commitment required is not the case with the scientific dataspace paradigm because the data is loaded into the corresponding databases, during execution of the experiment.

Since patient data is involved and due to legal requirements on such highly sensitive personal data, security and privacy issues are of utmost importance. Thus within applications in the breath gas analysis research domain, all participating databases including the RDF store are isolated, monitored, and restricted to a single point of access using the OGSA-DAI interface, hence implement strict access control. An instance of the breath gas scientific dataspace and its user environment is illustrated in Fig. 2. The figure clearly shows that the dataspace consists of five databases of types described above. Each is deployed as an OGSA-DAI resource running on a secure virtual host in a virtual machine resource. The recently developed logging framework inside OGSA-DAI is used to detect data abuse. The OGSA-DAI services itself are protected by standard security mechanisms supported by the underlying container, which is in our case the Globus Toolkit container 1 . We use mapping files in order to restrict access to OGSA-DAI resources to specific users. Details about the access control mechanisms and the security considerations in general are described in 5, 6. There are many different tools available for analyzing breath gas data (e.g. MATLAB, Octave, GridMiner, etc.). Breath gas researchers are already used to some tools and therefore might not want to change their analysis software. Therefore, we were considering in our architecture to make the interface to the scientific dataspace independent of any analysis tool. Users access breath gas source data from their source database via the community portal, which transforms the source dataset, using wrappers into the format, which the corresponding tool can parse. We provide interfaces for accessing data from and publishing data into the dataspace by following the strict access control mechanisms introduced above. The aim of the user environment is to guide breath gas researcher through their experimentation and their data publication by considering the e-Science life cycle ontology. Once the scientific dataspace has grown into a large scientific resource, we suspect it will be widely used by analysis, diagnostic and visualization tools.

Each instance of the above listed types of databases is defined as *dataspace-participant*. On a lower level of abstraction the contents of the databases itself are distinguished as *dataspace-participants*, i.e the final input dataset or a single analysis function (M-file) used within a specific breath gas analysis experiment. The different levels of abstraction are illustrated in Fig. 3. On a higher level of abstraction there might be whole dataspaces that act as participants of an interconnected large-scale *hyper-dataspace*. Such scenarios are important when multiple research organizations, each having deployed their own dataspace, engage collaborations and agree on sharing their scientific dataspaces and life cycle resources, respectively.

3 METHODS

Breath gas researchers will interact with the scientific dataspace via the community Web portal as depicted in Fig. 2. A Web portal is a website that

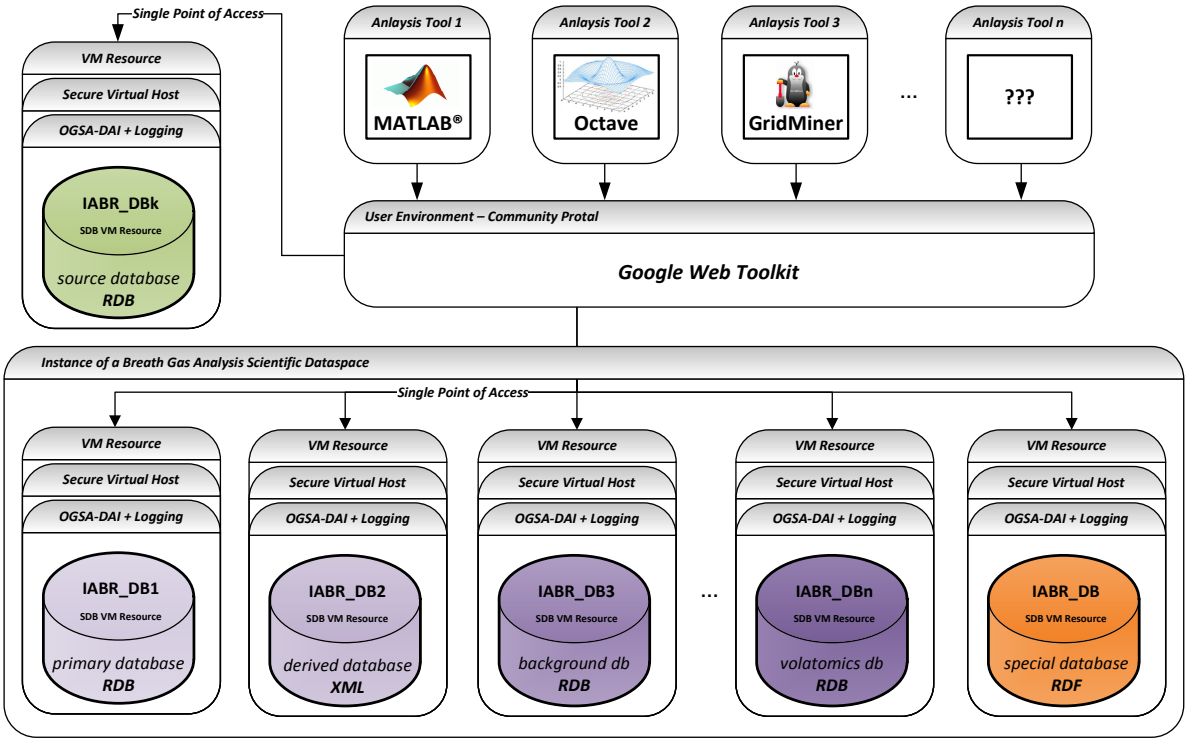

Fig. 2. An instance of the breath gas analysis scientific dataspace and its user environment - An instance is consolidated of at least five separate databases, which might be geographically distributed. Each database stores dataspace-participants of corresponding types (primary, derived, background, and other) as described above. The *special database* stores the relationships among participants in the form of individuals of the e-Science life cycle ontology. Analysis tools are independent of the user environment, where data is accessed from and loaded into the dataspace and corresponding external source databases hosting breath gas raw data.

acts as a point of access for a wide variety of information. The community portal consists of various portlets, which are pluggable components that are managed by the portal. The portlets of the Web portal are categorized into three major categories: (1) Infrastructure - Resource Management Portlets, (2) Monitoring - Semantic Logging Portlets, and (3) Action - Scientific Dataspace Portlets. The first two categories are primarily for administrators and therefore not further described in this paper. The latter category is designed for the breath gas researcher. It consist of two portlets (1) the e-Science life cycle manager and (2) the dataspace browser. In the following, the purpose and functionality of these two portlets are described and an overview of the user interaction is given. The community portal is built using the Google Web Toolkit (GWT) 13, which allows to build and maintain complex yet highly performant JavaScript front-end applications in the Java programming language, especially when combined with the Google Plugin for Eclipse.

3.1 The e-Science life cycle manager

The e-Science life cycle manager is a tool that enables the breath gas researcher to describe the experiment he is currently conducting according to the activities of the e-Science life cycle model. The model consists of five phases representing typical activities a researcher is carrying out when performing scientific experiments. Thus the most important steps in performing scientific experiments in e-Science applications are classified into the five activities, which we named e-Science life cycle activities (*GoalSpecification, DataPreparation, TaskSelection, TaskExecution, ResultPublishing*) 9. The e-Science life cycle ontology defines the concepts of the e-Science life cycle model as OWL-classes and properties.

The e-Science life cycle manager guides the breath gas researcher through the above mentioned five e-Science life cycle activities. It creates new instances of the e-Science life cycle and corresponding activities and allows the researcher to refine them throughout his study. Already defined instances of e-Science life cycle activities can be reused in other iterations or even in other studies. For instance, a created *final input dataset*, which represents the output of the activity *DataPreparation* might be reused in different iterations of the same experiment or even in different studies. In order to find the corresponding instance or other related instances for reuse we provide an integrated search and query interface as described below.

Search & Query the dataspace - Dataspace systems must enable users to interact with dataspaces through a search and query interface. However, we should keep in mind that much of the interaction with such a system is of exploratory nature. The implemented scientific dataspace model is based on the e-Science life cycle ontology. Data about conducted breath gas experiments is semantically rich described within instances of the ontology and is stored in RDF format. Using *SPARQL* query language for RDF, the scientific 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 check if insopiration is different to expiration of breath gas dataset x. If the results already exist, I'll save hours of computation."*

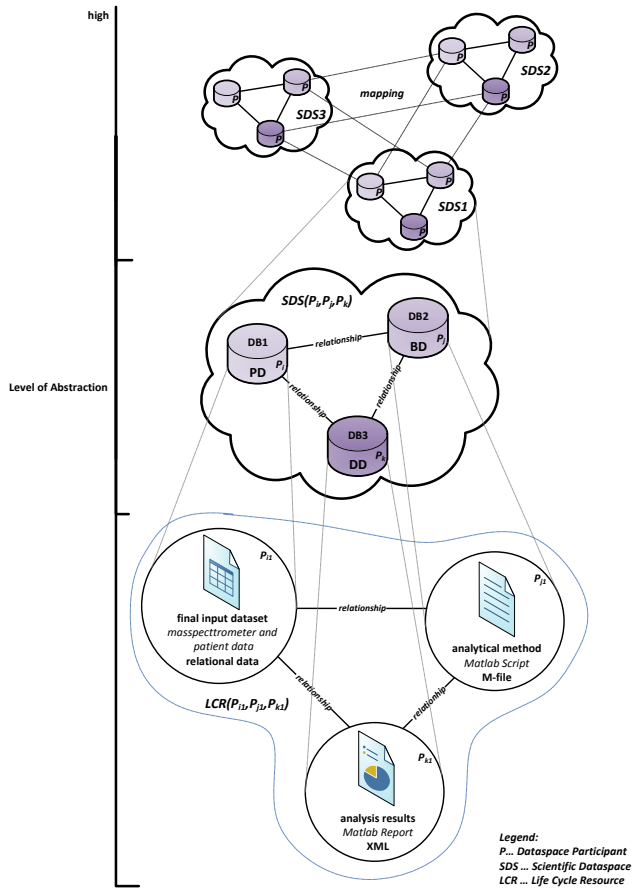

Fig. 3. Levels of abstraction of dataspace participants - in the lowest abstraction level, participants of the scientific dataspace represent concrete datasets that were used within a breath gas experiment. They form together a Life Cycle Resource (LCR). In a particular breath gas experiment, these participants are P_{i1} - the final input dataset (*primary data*), P_{j1} - the analytical method as MATLAB commands and functions organized in a m-file (*background data*), and P_{k1} - the resulted analysis report generated using MATLAB's publish function (*derived data*). These participants are stored in corresponding databases, as depicted in the figure. The relationship among these participants is semantically rich described by individuals of the e-Science life cycle ontology. In order to simplify matters we don't show in this figure the database that stores the relationships (RDF). On the next abstraction level the databases DB1, DB2, and DB3 represent participants forming the scientific dataspace $SDS(P_i, P_j, P_k)$, which is an instance of a breath gas analysis dataspace as was deployed as experimental framework for the Breath Research Institute of the Austrian Academy of Sciences. On the top level we illustrate a large-scale hyper dataspace, that arise when multiple dataspaces deployed for different organizations are interconnected e.g multiple breath gas analysis research institutions engage research collaborations, each running their own breath gas dataspace. In this scenario the dataspaces itself represent participants of the large-scale hyper-dataspace.

C *"Is there any experiment done on the volatile organic compound isoprene on exhaled breath gas in the context of cholesterol level in blood?"*

SPARQL is a powerful query language for RDF data. However, strict and complex grammar is a common characteristic of any query language being proposed for structured data including SPARQL, SQL, XQuery, etc. We are aware that it is not easy for breath gas researcher to accept these languages. Keyword search is widely used in Web search engines, but it cannot efficiently support semantic query. Therefore we implement a search and query interface providing the breath gas researcher an easy way how to express queries that will be transformed into a SPARQL query behind and submitted to the RDF store.

Neighborhood keyword queries - as introduced in 7, they have also the goal to explore associations between data items, can be provided by the life cycle ontology. For example searching for "endogenously-derived gases" returns not only the *goalSpecification* instances available in the dataspace that mention "endogenously-derived gases", but also the instances of its neighbor activities informing the user what proband data was used and where it resides, which analytical methods were applied, and what results of the corresponding experiment were achieved. Information about the scientist who conducted the experiment is also returned. This will strongly enforce new research collaboration within the breath gas research community.

Publication Modes - Different research groups are more or less collaborating within the breath gas scientific community. A research group belongs to a *Regional Head Service* (RHS). The RHS is responsible for managing a country's secure databases, which are under its ultimate control. The RHS again belongs to a *Project Head Service* (PHS), which implements the access decisions services, authentication, user management, and the dataspace services implementing the functions described above 6.

Users can publish their conducted e-Science life cycle experiments using five different publication modes:

- *researcher* - access to the instance of the life cycle can only be accessed by a specific researcher i.e. the researcher who conducted the experiment.
- *investigator* - the instance will be accessible for investigators i.e. the supervisor of a researcher.
- *research group* - the life cycle instance will be available for members of the research group the publisher is working with.
- *project member* - the life cycle instance will be available for members of the project the publisher is involved in.
- *IABR member* - the life cycle instance will be available for all members of the IABR. In this case there will be an entry made into the *volatomics database*

3.2 The Dataspace Browser

The dataspace browser is a tool that allows the user to navigate trough the e-Science life cycle resources available in the dataspace in a visual way. It is implemented as a portlet for easy integration into a community portal. It submits SPARQL queries attached with the role of the requesting user to the RDF Store. Based on the role of the user, he will see more or less e-Science life cycle resources. For instance the scientific dataspace may contain life cycle resources to which the publication mode *Researcher* was assigned. Such life cycle resources should be only accessible for the researcher who created the resource. Therefore, the request from the *Dataspace Browser* will include the role of the user. The response represents RDF-data and is used as input for the dataspace browser.

There are a number of tools available that visualizes RDF data. Some example projects include Welkin 20, multiple plugin tools for the Protege environment 10, and Semantic Analytics Visualization 16. These tools could

also be used by breath gas researcher in order to browse the contents of the dataspaces.

4 DISCUSSION

In data warehousing data typically is extracted, transformed from multiple data sources, and loaded into a separate database, called a data warehouse. This is not the case with the scientific dataspace paradigm discussed in this paper. Although, we load the final input dataset prepared for an analysis task into a separate primary database that is participating the dataspace. This is in fact one step of the whole data life cycle in e-Science, which we are trying to semantically enrich and preserve within the scientific dataspace.

Experiments on exhaled breath gas are being successively refined [*iterations of action 4-6 in Table 1*], by the acting researcher until the study either shows a significant result (i.e. definition of accurate methods for estimation of blood gas levels of certain biomarker values from breath gas samples) [*prepare action 7*] or ends up in a modification of the intended defined goal specification for that experiment [*modify goals and restart action 2*]. However, in both cases several iterations of the e-Science life cycle model are being performed. Some instances of e-Science life cycle activities might be reused in another iteration of the life cycle, for instance when a breath gas researcher executes the same *final input dataset* on a slightly refined analytical method. In this case the goals defined and the data prepared for that experiment did not change, therefore its corresponding instances of the e-Science life cycle activities are being reused within a new iteration of the life cycle.

During this iterative process of refinement, it might be the case that the acting researcher needs to share the experiment with his supervisor, who might decide to further share it with the research group or even with members of a collaboration project. Finally, once the study has shown some significant results that are published in some journal or in proceedings of a conference etc., these results represent valuable information not only to the acting research group and it might be worth to publish the experiment using the *IABR member* publication mode. Thus it will be accessible for all members of the breath research association.

4.1 Evaluating Studies

During our investigations on the e-Science life cycle model, we have cooperatively (computer scientists and breath gas researcher) conducted several sample breath gas experiments on top of the scientific dataspace. In the following we summarize the major activities of those experiments and describe their outputs and how data is organized by the scientific dataspace. At first the breath gas study is described in textual form. Users can define their own attributes and add values to it, e.g. attribute *Description* contains a textual description of the goals of the sutdy. This data is, together with information about the acting researcher (research group, department, publication mode) saved as individuals of the life cycle ontology in the RDF Store. A snapshot of a concrete RDF graph of a sample breath gas experiment is illustrated in Fig. 4.

Data access is done using the OGSA-DAI client and a certificate. In most scenarios, the researcher first loads all values of a the dataset he is investigating into the MATLAB structure. This is done using an implemented data load MATLAB-function (`loadData.m`), which communicates with an OGSA-DAI client. Then the breath gas researcher selects the values he is interested in for the current experiment, e.g. selection of all ex-smokers. The outputs of this selection process, which is done within MATLAB are twofold: (1) A new MATLAB MAT-structure containing the selected data (`fids.mat`). This represents the final input dataset and is therefore stored in the primary database as Binary Large Object (BLOB). (2) The MATLAB function itself (`selectData.m`), which is responsible to select the required data. This M-file represents the background data and is therefore saved in the background database.

Serveral analysis functions might be implemented by the breath gas researcher and applied to the final input dataset. Such MATLAB functions also represent background data, thus are saved correspondingly. For easier handling we provide an empty MATLAB file named `executeStudy.m`, which will be used in all experiments. The breath gas researcher may add their own implementations into it or import external analysis functions. These functions calculate the derived data, which are usually input to a plot function. Again an empty template (`plotData.m`) is prepared to be used by the researcher. Finally we provide a publish function (`dsPublish.m`), which generates an XML report of the conducted experiment including plots, if used, by taking advantage of MATLAB's publishing feature. This report is then zipped and saved as BLOB in the derived database of the scientific dataspace.

Based on this investigations, we have created guidelines for breath gas researchers, defining concrete activities and documentation policies, which guide users through the e-Science live cycle. An empty template with named files is also created for a better comprehension.

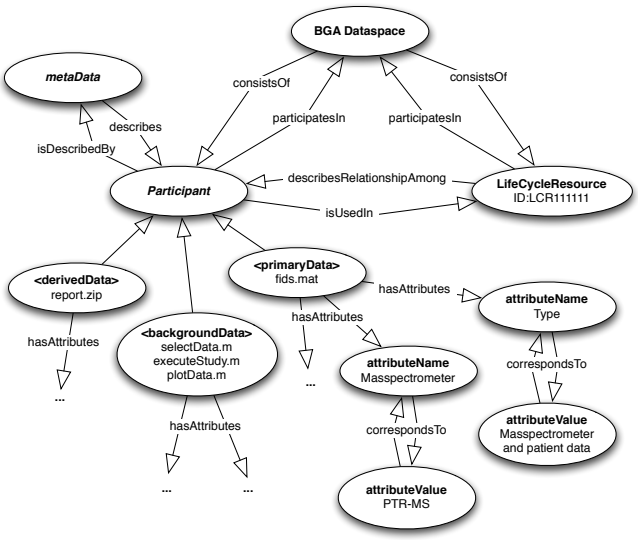

Fig. 4. Snapshot of an RDF graph of a sample breath gas experiment - This snapshot shows a life cycle resource, which describes the relationship of the participants depicted in the figure. The participants represent primary, background, and derived data of the breath gas experiment. Participants are described with attributes and their values, which the acting breath gas researcher defines, while conducting the experiment. In the breath gas analysis domain, there is a set of attribute names predefined, to be followed by scientists for a consistent description of breath gas analysis experiments.

4.2 Implementation Status

A first experimental framework has been developed in order to early evaluate the concepts of the e-Science life cycle ontology, which is the basis of the scientific dataspace paradigm discussed in this paper. This framework builds on top of Protege. We used the built-in individual editor to create several life cycle resources that represent breath gas analysis experiments and the SPARQL query panel in order to retrieve those resources. Furthermore, we deployed an experimental breath gas source database, restricted to a single point of access using the OGSA-DAI interface and implemented a MATLAB function that allows to communicate with that database using a certificate. We then conducted several breath gas experiments using the MATLAB environment and manually saved its background and derived data into corresponding databases. Even, if this first experimental framework has no interfaces implemented, it allowed us to proof the concepts of the introduced breath gas analysis scientific dataspace and to elaborate a feasibility study. Based on this first experimental evaluations we have finalized and documented the architecture of the system including the design of all necessary OGSA-DAI workflows in an UML-based design document to be available as technical report.

As the next step we addressed the interface design, and started to build a first prototype using the above introduced Google Web Toolkit (GWT).

5 CONCLUSION

This paper describes the proposal and implementation of a scientific dataspace paradigm for the breath gas analysis scientific community. The authors, computer scientists and leading breath gas researcher have investigated and evaluated a novel data management and scientific data preservation approach based on the concepts of dataspaces for the scientific community of breath gas researcher. First breath gas experiments were collaboratively conducted on top of the experimental dataspace framework, which allowed us to evaluate the e-Science life cycle ontology. This first evaluation represents the bases for an extension and deeper investigation of the scientific dataspace paradigm introduced.

ACKNOWLEDGEMENT

The Austrian BMWF (Federal Ministry for Science and Research) funding of the Austrian Grid 2 project (Contract: GZ BMWF-10.220/0002-II/10/2007) is key to bringing the partners together and to undertaking the research. The Department of Scientific Computing, University of Vienna brought the e-Science life cycle ontology to the project and the Research Center for Process and Product Engineering, University of Applied Sciences, Dornbirn, set up the meetings that led to the research collaboration with the Breath Research Institute of the Austrian Academy of Sciences, Innsbruck, forming the breath gas analysis application. The entire research team contributed to the discussions that led to this paper and provided the environment in which the ideas could be tested. Finally we would like to thank the support from the e-Science Institute for the hosting of the IWPLS09 workshop.

REFERENCES

[1]The globus alliance. http://www.globus.org.

[2]Resource description framework (RDF). http://www.w3.org/RDF, February 2004.

[3]A. Amann and D. Smith. Breath analysis for clinical diagnosis and therapeutic monitoring. World Scientific, Singapore, 2005.

[4]M. Baumgartner, C. Glasner, and J. Volkert. An overview of the austrian grid infrastructure. Proceedings of the 1st Austrian Grid Symposium, 2005.

[5]M. Descher, P. Masser, T. Feilhauer, A. M Tjoa, and D.Huemer. Retaining data control to the client in infrastructure clouds. In Proceedings of International Conference on Availability, Reliability and Security, IEEE, Fukuoka, Japan, 2009.

[6]M. Descher, P. Masser, T. Ludescher, B. Wenzel, T. Feilhauer, P. Brezany, I. Elsayed, A. Wöhrer, A. M Tjoa, and D.Huemer. Position paper: Secure infrastructure for scientific data life cycle management. In Proceedings of International Conference on Availability, Reliability and Security, IEEE, Fukuoka, Japan, 2009.

[7]X. Dong and A. Halevy. Indexing dataspaces. In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 43–54, New York, NY, USA, 2007. ACM.

[8]I. Elsayed, P. Brezany, and A M. Tjoa. Towards realization of dataspaces. In DEXA '06: Proceedings of International Conference on Database and Expert Systems Applications, IEEE, 2006.

[9]I. Elsayed, A. Muslimovic, and P. Brezany. Intelligent dataspaces for e-science. In Proceedings of International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics (CIMMACS '08), World Scientific and Engineering Academy and Society Press, Cairo, Egypt, 2008.

[10]Stanford Center for Biomedical Informatics Research. Protege ontology editor and knowledge-base framework. http://protege.stanford.edu/, 2009.

[11]I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the Grid: Enabling scalable virtual organizations. International Journal of Supercomputer Applications, 15(3), 2001.

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

[13]Google Code. Googel web toolkit (gwt). http://simile.mit.edu/welkin/, 2009.

[14]Kushch I, Arendack B, Stolc S, Mochalski P, Filipiak W, Schwarz K, Schwentner L, Schmid A, Dzien A, Lechleitner M, Witkovsk V, Miekisch W, Schubert J, Unterkofler K, and Amann A. Breath isoprene - aspects of normal physiology related to age, gender and cholesterol profile as determined in a proton transfer reaction mass spectrometry study. Clin Chem Lab Med. 2008;46(7):1011-8.

[15]I. Elsayed et al. The e-Science Life Cycle Ontology. www.gridminer.org/e-sciencelifecycle, 2008.

[16]A. P. Sheth L. Deligiannidis and B. Aleman-Meza. Semantic analytics visualization. In Proceedings of the International Conference on Intelligence and Security Informatics (ISI-2006), IEEE, San Diego, CA, USA, 2006.

[17]M. Ligor, T. Ligor, A. Bajtarevic, C. Ager, M. Pienz, M. Klieber, H. Denz, M. Fiegl, W. Hilbe, W. Weiss, P. Lukas, H. Jamnig, M. Hackl, B. Buszewski, W. Miekisch, J. Schubert, and A. Amann. Determination of volatile organic compounds appearing in exhaled breath of lung cancer patients by solid phase microextraction and gas chromatography mass spectrometry. Clinical Chemistry and Laboratory Medicine 47, 550–560., 2009.

[18]C. Lynch. Big data: How do your data grow? *Nature*, 455(7209):28–29, 09 2008/09/04/print.

[19]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.

[20]S. Mazzocchi and P. Ciccarese. Welkin - a graph-based rdf visualizer. http://code.google.com/intl/en/webtoolkit/, 2004.