2009 International Conference on Availability, Reliability and Security

Position Paper: Secure Infrastructure for Scientific Data Life Cycle Management

M. Descher, T. Feilhauer
T. Ludescher, P. Masser, B. Wenzel
Research Center for
Process and Product Engineering
University of Applied Sciences Vorarlberg
[mde|tf|lt|maph|web]@fhv.at

P. Brezany, I. Elsayed, A. Woehrer
Institute of
Scientific Computing
University of Vienna
[brezany|elsayed|woehrer]@par.univie.ac.at

A. M. Tjoa, D. Huemer
Institute of
Software Technology
Vienna University of Technology
[amin|dhuemer]@ifs.tuwien.ac.at

***Abstract*—Within the Austrian Grid project phase 2, three different groups, each allocated to a different workpackage, join their efforts to implement a grid infrastructure for the european research project "Breath Gas Analysis for molecular oriented diseases". This position paper provides background on the task and the resulting requirements, a presentation on solutions developed during related projects in the application domain, identifies problems that have not yet been solved, and finally presents the intended solution to be developed.**

I. Introduction & Context

This position paper describes the current state, the intended realisation and a discussion of the project Grid Breath Gas Analysis (BAMOD-Grid) carried out within the "Austrian Grid Project Phase 2". The Austrian Grid consortium join Austrias leading researchers in advanced computing technologies with well-recognized partners in grid-dependent application areas. An overview of the Austrian Grid is provided in [1].

A. The BAMOD Project

The Breath-Gas Analysis for molecular-oriented diseases (BAMOD) [2] project is focused on the diagnosis of minimal diseases and early stages of lung and oesophageal cancer. It aims to apply breath-gas analysis techniques, that is the molecular analysis of exhaled air, to realize a non invasive method of diagnostics.

BAMOD is the mother-project of BAMOD-Grid described in this paper. The work-packages to be delivered are separated to several project partners, one of them being the University of Applied Sciences Vorarlberg (FHV).

FHV workpackage: The task appointed to the FHV is the development of a database to host the project's data. This task contains the following subtasks:

- Definition of a common annotation file format for experimental data
- Design and construction of a database for experimental data
- Programming of interface scripts and applications, that is
  - *Questionnaire application*: Collect personal data from a test person
  - *Server software*: Interface to the database system (webservice)
  - *Matlab interface*: Import data from server into matlab for analysis

Security and Privacy: Security and privacy issues are of utmost importance within this project, due to legal requirements on such highly sensitive personal data. Therefore the database is currently implemented as an isle system, due to the decision of the group responsible for security and privacy in the BAMOD project.

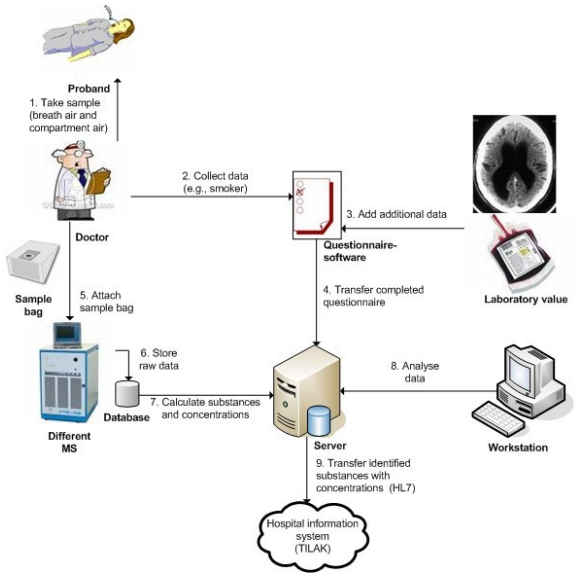

Figure 1. Use case depicting the current sequence of events

Fig. 1 presents an overview use-case, with the *Server* being the aforementioned isle system. Import and export activities on this system are protected by smart card security measures.

Steps 2 to 4 present the collection and subsequent transfer

978-0-7695-3564-7/09 $25.00 © 2009 IEEE
DOI 10.1109/ARES.2009.99

(i.e. manual import) of personal data to the server. 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.

Identified roles: Fig. 2 shows an access matrix of the roles currently applied in project BAMOD and their respective allowed tasks. Each entry in the database (generated from a proband) is logically[1] divided into three different parts. These are the

- *personal data*, containing e.g. the unique identifier of the hospital information system (cf. Fig. 1) being a subset of the
- *patient data*, containing e.g. the patients anamnesis, and the
- *analysis data*, containing the preprocessed data generated from the mass spectrometry.

Access to these logical subsets is granted based on a user's roles, where a single user typically holds several different roles that complement each other.

B. BAMOD-Grid

As the BAMOD project is intended to support stakeholders from several european countries, there are, among other points, legal issues to be taken into account. Using techniques provided within grid computing as the base infrastructure for the project proves to be advantageous in several respects:

- Certificate security infrastructure (through the use of GSI) satisfies the different legal requirements of the participating countries, as certificates are chartered to every entity (human and machine) involved. This allows remote access, as opposed to local access to the isle system.
- Interfaces are built upon agreed standards to prevent uncontrolled growth of proprietary solutions
- Reliable file transfer techniques
- Load-leveling and off-site backup between the involved databases
- Tracing of every task executed by every user, through application of certificates allows consistent accounting
- Advanced methods of data management resp. analysis and results management

II. Requirements

Identification of the relevant issues of BAMOD-Grid issued the following requirements:

A. Secure Database Resource

Databases to be used within the project have to be centrally controlled and monitored. This is due to ownership issues of medical data, see Sect. IV-B for further elaboration. A secure and isolated database resource hence has to be assured. Every access to such secure database resources has to be completely side-access free, that is, there has to exist only one interface to the database management system (DBMS) hosted on the resource, namely OGSA-DAI [3], which is the de facto standard for data access and integration for relational and xml data as well as file resources. This means that a local administrator must not be able to circumvent this single point of access by e.g. using an available root login on the system and the respective database administration tools. The database has to implement and enforce strict access control, as determined by the project management.

[1]The database schema, especially the analysis part, is still under development. Hence we denote the access to be logical, as to circumvent implementation details.

B. Access control on datasets

Access to single sets stored in the aforementioned secure database resource shall support role based access including fine grained control, i.e. access per cell. Depending on the role of the user (cf. Fig. 2), only a subset of the data stored in the database has to be visible, resp. accessible.

Additional rights to a particular user might be granted irrespective of the role of this user, e.g. creator of a data set always gains full access to this specific data set.

A user must not be allowed to extract data from the database to which he does not have the corresponding access rights. The enforcement of these access rights therefore needs to be implemented on the lowest level possible, in this case the datasets itself.

C. Federated management of Roles

As BAMOD-Grid is ultimately designed for pan-european usage, extension of the basic roles for country-specific usage need to be made. The overall system has to reflect the operations and transfers that are in correspondence with the respective prevalent laws. That is e.g. a researcher from country X accessing data sampled and stored in country Y needs to pass a "border gateway" where the respective call is evaluated and permitted or denied.

D. Scientific Data Management

In the breath gas research domain 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 different compounds in the exhaled breath. The analytical techniques used include various statistical and data mining techniques supporting identification of specific disease markers. The output of theses analyzes 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.

Role \ Access right	Collect data from new test person	Preprocess raw data (read/write)	Read self collected patient data and analysis data	Read all patient data and analysis data	Read personal data of test person	Administrate data (read/write)	Import data
Sampler[1]	■						
Analysis Preprocessor[2]		■					
PhD – Researcher			■				
Investigator – Researcher			■	■			
Elevated Researcher					■		
Administrator[3]						■	
Importer[4]							■

[1] generates data of a new test person using the Questionnaire software
[2] accesses different mass spectrometry raw data and calculates the substances with their concentrations
[3] can add or modify existing questionnaires, mass spectrometries, users, substances, etc.
[4] imports data on to the isle system.

Figure 2. Access matrix showing the roles and respective access rights currently employed in project BAMOD

However the resulting data, called the derived data, that has arisen from the research task represents valuable information not only to the acting research group, but also to other groups with respect to other main focuses.

The scientific community interested in the analytical results of breath-gas analysis is also geographically distributed. In order to fully utilize this large amount of scientific data available at each breath gas research center, it is necessary to provide a novel data management paradigm taking into account the above described requirements in terms of security and privacy considerations.

III. Lessons learned from related projects

Many of the requirements will be covered by learning from, resp. applying techniques developed in related projects:

A. VOTES

Within the "Virtual Organisations for Trials and Epidemiological Studies" (VOTES) [4] project, carried out in the UK, problems like role management and fine grained access management have been dealt with. In VOTES a central transfer grid was established, connecting databases from different universities. Each node in this transfer grid administers a pool of databases containing the crucial patient data. Users may issue queries to this grid through a portal. Depending on the users role, only a subset of the available database fields are included in the result that is returned by the transfer grid. The role to field assignment is done using an access matrix maintained at VO level. Other issues like anonymisation and statistical inference enabling patient identification on various non-identifying queries are also discussed.

B. MammoGrid

The MammoGrid project [5], funded by EU FP5 from 2002 to 2005, provides information about operating pan-european medical grid systems.

It defines a hierarchical (two level) structure of VOs where all child VOs are controlled by a super VO. The idea was to let the child VOs represent the hospitals of the project partners while there is a single super VO that coordinates the data accesses between the different child VOs.Therefore the super VO is responsible for inter-VO communication which includes filtering the accessed data from external VOs according to specified rules and restrictions that are based on the specific regulations of the hospitals and on the different legislations of the participating countries.

MammoGrid does however not apply an explicit role based access control, as there is no technical separation between the stakeholders involved (mammograms are created and investigated by physicians only). The project is implemented for use within the EGEE network, hence its technical realization is not applicable to the existing Austrian Grid infrastructure. MammoGrid is now a commercial product, called "MammoGrid+".

C. BRIDGES

The BRIDGES project [6], funded by the UK e-Science program from 2003 to 2005, developed a bio-medical research infrastructure supporting data sources, bioinformatics analysis and visualization tools, and research activities combining shared and private data. As the inclusion of patient records and animal experiments was planned, privacy and access control were particular concerns.

IBMs commercial product WebSphere Information Integrator [7] is used to provide access to heterogeneous and distributed data sources. Wrappers are implemented for each type of resource accessing the data and feeding a warehouse, thus BRIDGES does not limit access to resources by a single point of access.

IV. Unsolved Problems

By learning from the projects presented in Sect. III, we are not able to cover all the requirements set upon the project.

The following problems are yet to be solved:

A. Isolated database resource

To ensure resp. provide an accountable database resource all input and output needs to be routed through a single point of access while being protected against any way of side-access. As the databases are distributed over different locations, the physical resources hosting them are subject to the respective local control.

A local administrator however, providing the respective physical grid resource, is equipped with full access. A solution for a securely isolated and integrity preserving hosting of a grid database on a remote resource hence has to be developed.

B. Ultimate data control

Samples generated on a proband are owned by the sampler, which has full access on his generated data. Ultimately the owner of the data is however the proband the data was generated from, according to [8, Sect. 2.4] "the participating patient has the right of withdrawal at any time".

The system to be created needs to take care of such a scenario, where we have distributed databases, on which we need to be able to ensure the removal of a selected entry. That is, ultimate control of all data available within a project has to be retained to the project management in order to be empowered for any such request.

C. Central logging and access control

Distributed databases allow distributed attacks, e.g. full data download by partial retrievement of data on single resources such that the single logging resp. access decision is not aware of.

Another important aspect is to ensure that local administrators do not tamper data or hosting machines and are only able to access data with valid credentials through OGSA-DAI, as all other users do. Here a federated logging has to be applied, keeping semantically enriched information about overall data access.

The system implemented has to be able to assist in detecting any kind of violation of access restrictions, be it even by means of distributed partial attacks.

D. Advanced scientific data management

Today, there are powerful systems for managing data at the level of a single database system (whether relational, XML, or in some other model). While the commercial world has standardized on the relational data model and SQL, no single standard or tool has critical mass in the scientific community [9].

There are many parallel and competing efforts to build scientific data management solutions. Data interchange outside each group is problematic, therefore being an open research issue and probably a never-ending story faced by the data management community. A much greater challenge facing the data management community is to raise the abstraction level at which data is managed, in order to provide a system controlling different data sources, each with its own data model [10].

V. Intended Solution

Fig. 3 provides an overview by an exemplary usage of project resources within Austria and Germany.

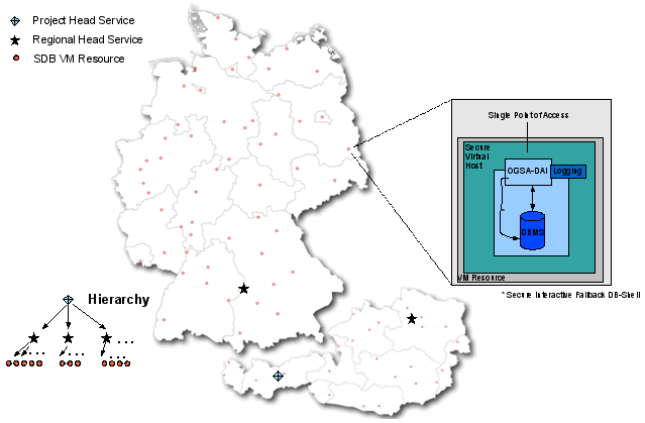

Figure 3. Planned three layer architecture. An exemplary participation of Germany and Austria is shown with the respective situated resources.

A. Infrastructure

We plan on implementing a three layer architecture for the infrastructure. The *Project Head Service* (PHS) represents the root of the graph. It is the head service of the project implementing the access decision services, authentication, user management, etc.

Each of the countries involved has a respective *Regional Head Service* (RHS). The RHS is responsible for managing the secure databases available for its country, which are under its ultimate control. This will be realized by applying Globus Nimbus [11] resources for hosting the *secure database virtual machine resources* (SDB-VM), where the life-cycle of the machine will be centrally controlled by the RHS. These SDB-VM Resources are allocated to the respective regional head.

Before a user can get access to a SDB-VM resource, a grant has to be retrieved from the RHS. The RHS first registers the respective access with the SDB-VM and then passes the access ticket to the user (analogous to the kerberos system). The SDB-VM may then be directly accessed.

If a user in a specific country wants to access the data available in another country, the ultimate access decision is made by the PHS.

Every access grant (resp. decision) made by a RHS or PHS is logged at the respective location (SDB-VMs log to their associated RHS and so on). Audit trails will be realized through log files and by indemnifying that every data access

will be protocolled. This can be guaranteed as communication through the OGSA-DAI interface is the single point of access to the data. Therefore concepts mentioned in [12], where a data centric logging is presented, will be modified and extended.

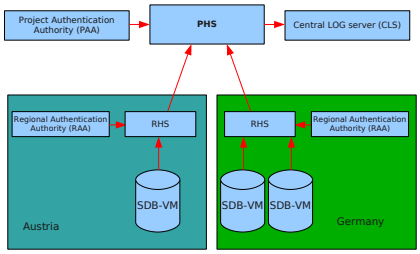

Figure 4. Central log server: depicted flow of the logging data from the involved entities

[12] applies the eXtensible Access Control Markup Language (XACML) where data owners define logging policies explaining their audit needs. Although this solution is flexible and performant, it lacks the specification of the logging data format and the protection mechanism used on the way from its actual production to the central log server.

The logging framework, shown in Fig. 4, intended for BAMOD-Grid will provide a universal logging solution, utilizing semantics with a defined but extensible schema and secure transmission of log data to a central log server. The concept of a standard interface, already known from OGSA-DAI will be used for logging purposes, too. A parser for each log producing module (i.e. a program) has to be written, mapping logging data to a schema, defined by the framework. The schema presents the universal interface to which all logging data is parsed to. Those semantically enriched log files will then be securely transferred in a batch job mode to a central log server.

As a result semantically enriched log events of all components of BAMOD-Grid are located at the central log server, enabling analysis and reporting of details in arbitrary granularity and to detect distributed attacks. It is even possible to define patterns describing secure workflows and to ensure that those patterns are followed. For example a user first has to authenticate himself at the PHS before he may use the issued credentials to access the resp. database(s).

As each interaction of the user with BAMOD-Grid components results in a log entry and all logs get concentrated at the central log server enriched with semantics, a violation of this workflow can be detected by combining the information of the according logs.

It is also possible to produce reports of arbitrary content for different stakeholders by utilizing EXtensible Stylesheet Language Transformation (XSLT) or to build an IDS/IPS (intrusion detection/prevention) system without false positives/negatives based on such semantically enriched log data.

B. Scientific dataspaces

The challenge is to manage a dataspace, rather than a database. Dataspaces [13] are modeled as participants and relationships. Relationships can be expressed by single word-relationships, such as replica-of, related-to, view-of, etc. They can even be semantic mappings of database schemes. Systems providing the required services over dataspaces are considered to be Dataspace Support Platforms (DSSPs) [13]. In [14] 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 a dataspace.

The initial ideas on managing dataspaces have started to evoke interests of the research community [10], however most effort is related to the mainstream and so far not considered in scientific data management. In Fig. 5 we illustrate our extension to the mainstream dataspace research providing advanced scientific data management.

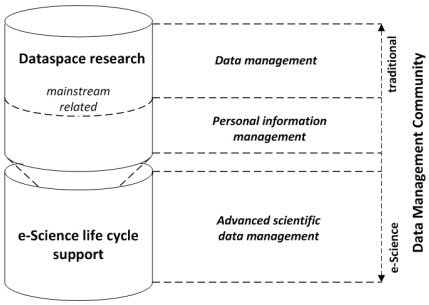

Figure 5. Extension to the mainstream dataspace research providing advanced scientific data management.

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-automatic creation creation of these relationships as well as their improvement 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* [15], 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.

C. Project partners

Project BAMOD-Grid is carried out by three groups. The groups and their respective allocated workpackages within the *Austrian Grid Phase 2* project are:

- (A) Prof. Feilhauer (Application): BAMOD-Grid
- (B) Prof. Tjoa (Infrastructure): Security in the Grid
- (C) Prof. Brezany (Infrastructure): Datagrid

The goal is set by the intended implementation as described. The application requires several modifications and enhancements to be made to the existing infrastructure, as well as the integration of a data and analysis lifecycle management system.

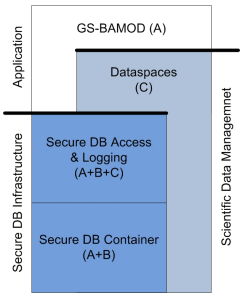

Figure 6. Layer representation of the work and assigned teams

VI. Discussion

We intend to implement the system in a way that it will provide a secure centrally controlled data grid infrastructure where ultimate data access is retained to the project head service. We think that such a system is applicable to lots of different domains. With the ever growing application of techniques like virtualization and cloud computing where data is handed over to a third party, and ultimate control is lost.

The system is hence intended to be developed in a more generic way, as to support any such infrastructure requirement. In such a system the RHS and SDB-VMs will be available at any participating country (or location) but not bearing any “intelligence” or providing any functionality until assigned to a PHS, where the infrastructure is actually “instantiated”.

Acknowledgment

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] M. Baumgartner, C. Glasner and J. Volkert, *An Overview of the Austrian Grid Infrastructure*, Proceedings of the 1st Austrian Grid Symposium, pp 277-287, Dec. 2005, Schloss Hagenberg, Austria

[2] *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

[3] M. Antonioletti, M.P. Atkinson, R. Baxter, A. Borley, N.P. Chue Hong, B. Collins, N. Hardman, A. Hume, A. Knox, M. Jackson, A. Krause, S. Laws, J. Magowan, N. W. Paton, D. Pearson, T. Sugden, P. Watson, and M. Westhead. *The Design and Implementation of Grid Database Services in OGSA-DAI.*, Concurrency and Computation: Practice and Experience, Volume 17, Issue 2-4, Pages 357-376, February 2005

[4] R. O. Sinnott, A. J. Stell and O. Ajayi, *Initial Experiences in Developing e-Health Solutions across Scotland*, Paper presented at the Integrated Health Records Workshop, 9th March 2006, Edinburgh, UK. http://labserv.nesc.gla.ac.uk/projects/votes/edinIHR.pdf, Last visit: Oct. 29, 2008

[5] R. H. McClatchey, D. Manset and A. E. Solomonides, *Lessons Learned from MammoGrid for Integrated Biomedical Solutions*, Presented at the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS 2006). Salt Lake City, USA. June 2006

[6] R. O. Sinnott, M. Bayer, D. Houghton, D. Berry and M. Ferrier, *Development of a Grid Infrastructure for Functional Genomics*, Springer Verlag, 2004, Kanazawa, Japan

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

[8] Federal Office for Information Security (BSI) of the German Government, *Vorstudie Grid Sicherheits-Infrastruktur (GSI), Arbeitspaket 4: Zusammenfassung der Ergebnisse*, Version 3.4, http://www.bsi.de/literat/studien/GSI/index.htm, Last visit: Oct. 29, 2008

[9] J. Gray, D. T. Liu, M. Nieto-Santisteban, A. S. Szalay, D. DeWitt and G. Heber, *Scientific data management in the comnig decade*, Microsoft Research Technical Report MSR-TR-2005-10,January 2005

[10] Alon Halevy, *Why your data wont mix*, Queue, 3(8):5058, 2005

[11] K. Keahey, I. Foster, T. Freeman, X. Zhang and D. Galron, *Virtual Workspaces in the Grid*, Europar 2005, Lisbon, Portugal. September, 2005

[12] W. Zhang, D. Del Vecchio, G. Wasson and M. Humphrey, *Flexible and Secure Logging of Grid Data Access*, 7th IEEE/ACM International Conference on Grid Computing, Sept. 2006, Barcelona, Spain

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

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

[15] I. Elsayed, A. Muslimovice and P. Brezany, *The e-Science life cycle ontology*, 2006, http://www.gridminer.org/e-sciencelifecycle, Last visit: Nov. 1, 2008