DATA MANAGEMENT AND INTEGRATION WITHIN COLLABORATIVE WORKING ENVIRONMENTS

Assel Matthias and Kipp Alexander
High Performance Computing Center - HLRS, University of Stuttgart
Nobelstr. 19, Stuttgart, Germany

Keywords: Distributed Data Management, Data Integration, Data Sharing, Collaborative Working Environments, Data Exchange Security, Virtual Laboratory, ViroLab, CoSpaces.

Abstract: With increasingly distributed and inhomogeneous resources, sharing knowledge, information, or data becomes more and more difficult and manageable for both, end-users and providers. To reduce administrative overheads and ease complicated and time-consuming integration tasks of widely dispersed (data) resources, quite a few solutions for collaborative data sharing and access have been designed and introduced in several European research projects for example in CoSpaces and ViroLab. These two projects basically concentrate on the development of collaborative working environments for different user communities such as engineering teams as well as health professionals with a particular focus on the integration of heterogeneous and large data resources into the system's infrastructure.
In this paper, we present the two approaches realised within CoSpaces and ViroLab to overcome the difficulties of integrating multiple data resources and making them accessible in a user-friendly but also secure way. We start with an analysis on systems' specifications describing user and provider requirements for appropriate solutions. Finally, we conclude with an outlook and give some recommendations how those systems can be further enhanced in order to guarantee a certain level of dynamicity, scalability, reliability, and last but not least security and trustworthiness.

1 INTRODUCTION

Today's B2B[1] relationships are not limited on local or regional collaborations any longer but the international orientation of enterprises in fact combines various organisations scattered all over the world to cooperate with each other. Due to the necessity of exchanging information and/or confidential data among the business partners involved, easy, dynamic and especially secure ways of sharing certain data sets and information need to be considered and applied before cross-organisational collaborations can take place, and, in order to prevent any abuse by third parties while communicating over untrusted networks like the Internet.

From an eBusiness perspective, the concept of Virtual Organisations (VO) is widely and often used to approach similar issues, namely to make (data) resources and products available dynamically and on-demand. The main purpose pursuit in such a concept consists in making ad-hoc collaborations (Schubert et al., 2005) possible that meet specific business goals, respectively address a (temporary) market niche. To achieve this, VO frameworks (Wilson et al., 2005) allow for identification of resource providers according to current business needs/goals, and integration of these so as to enable collaborative workflow execution.

The CoSpaces[2] project elaborates a framework to support world-wide distributed engineering teams by developing an environment that supports dynamic, ad-hoc collaborative working sessions. In particular, this framework shall support on-demand selections of participants, documents, and data for a collaboration session as well as the easy integration of partners regarding both, the support to ease the access to and from partners and their applications.

However, within ViroLab[3], a virtual laboratory

[1] Business-To-Business

[2] http://www.cospaces.org/

[3] http://www.virolab.org

Matthias A. and Alexander K. (2008).
DATA MANAGEMENT AND INTEGRATION WITHIN COLLABORATIVE WORKING ENVIRONMENTS.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - DISI*, pages 258-263
DOI: 10.5220/0001704302580263

for HIV[4] research and medication support (Assel and Krammer, 2007) is being developed that allows several experts in this field to share their knowledge and results interactively while working together on the same data and information sets, which are currently widely dispersed over Europe and without cross-national or even cross-institutional collaboration.

To face real end-users' needs and requirements, concrete scenarios in cooperation with industrial partners in CoSpaces and respectively hospitals in ViroLab have been developed and will be evaluated against the defined concepts. Sharing data or documents between partners stresses security issues to be of the utmost priority and importance (Assel et al., 2007) while developing such collaborative working environments for business partners and/or academic institutions. Those issues include several levels of security implying trustworthiness among participants to meet appropriate collaboration goals without harming legal issues as well as keeping the user's privacy.

In the following, we come up with general systems' specifications including requirements on both, user and provider sites that need to be considered while developing environments for dynamic (business) collaborations, as well as two concrete examples demonstrating appropriate dynamic and secure infrastructures realised within the CoSpaces and ViroLab project.

2 SYSTEMS' SPECIFICATIONS

Talking about distributed collaboration environments between organisations across different countries, the specific and even dynamic requirements for integrating and accessing services, applications, and data resources in particularly regarding security issues differ from "normal" local collaboration federations.

The immense complexity of the different technologies involved as well as the heterogeneity of present infrastructures and resources together with their spatial distribution, requires lots of effort for the design and development teams. Lots of progress has been made in the recent past and quite a few solutions have been developed in other project like Akogrimo[5] and TrustCoM[6]. The currently running Integrated Project BREIN[7] extends the eBusiness approach by merging semantics, agents, and Grid technologies to provide an intelligent, self-manageable infrastructure. But all these approaches do not consider the very specific needs in case of dynamic collaboration sessions, respectively distributed data handling.

Basically, to ease and reduce management and configuration overheads during the setup phase but also during the runtime of such sessions, several issues including amongst others usability, performance, scalability, reliability, flexibility and especially security need to be taken into account, and should be carefully evaluated before deciding on specific technologies, respectively designing and developing appropriate systems and/or infrastructures.

The following list briefly highlights the most important requirements for collaborative working environments in general by taking the recent project results and extending them accordingly.

- The overall infrastructure shall be highly flexible - not limited to one specific operating system, middleware, and technology (Elmroth et al., 2007) but rather interoperate with different widely-used solutions including standard technologies and specifications, as well as following the latest approaches of the so-called Software as a Service (SaaS) paradigm;
- End-user friendliness including easy to deploy and run components/services as well as application and resource transparency;
- Virtualisation of services and corresponding components to allow on-the-fly modifications or extensions without affecting the current session(s);
- The easy usage of the collaboration platform should be supported through a decentralized authentication and authorisation model (Assel and Kipp, 2007) based on a Single-Sign On (SSO) procedure across and within organisational boundaries;
- The dynamic setup of collaboration partners including the on-demand modifications of firewalls in order to guarantee that only trusted people are allowed to execute corresponding operations;
- Dynamic management and control of attribute-based access policies required to authorise users before accessing services, applications, and resources;
- Keeping the user's privacy and protecting his/her confidentiality by impersonalising data or excluding irrelevant information;
- Satisfying requirements under the Data Protection Act as well as explicit consent from all parties concerned;
- Secure data transmission based on data encryption on several levels (e.g. encrypted messages versus

[4]Human Immunodeficiency Virus
[5]http://www.mobilegrids.org
[6]http://www.eu-trustcom.com.
[7]http://www.gridsforbusiness.eu

secure protocols) ensuring trustworthiness and integrity of exchanged information;

- Additional security mechanisms and policies for storage of confidential data sets;
- Recording of relevant user interactions for auditing, accounting, and pricing;
- Monitoring of critical infrastructure components as well as services/applications to react on suddenly intermittent failures;
- Flexible system(s) for distributing interesting events or pre-defined topics of interest to foreseen users/components (notification support);
- Methods for defining and negotiating Service Level Agreements (digital contracts) (Hasselmeyer et al., 2006) based on QoS[8] parameters in order to guarantee a certain level of reliability, performance, and scalability to customers/users and to facilitate the individual pricing of single service capabilities.

3 COSPACES SHARED DATA SPACE

Beside the dynamic setup of a collaboration session the management of sensible data has to be considered as a critical aspect. Within industrial collaborations critical data has to be shared between all collaboration partners. So the CoSpaces framework has to provide an infrastructure to support the secure distribution of this data to partners being foreseen for a specific collaboration. As one of the most important aspects for industrial partners regarding the sharing of data, it has been identified that the control of the corresponding data, i.e. *who* is allowed to access or modify *which* specific data sets, must remain by the corresponding data owner. The following current practise has been identified within current industrial collaborations:

- If companies want to share data, the data being foreseen for a specific collaboration session is stored in a dedicated shared data space within the Demilitarised Zone (DMZ) of the corresponding company;
- Access rights are just granted on data within that shared data space;
- The DMZ is protected by another firewall and secured with an authentication and authorisation system;
- Access to the corresponding data is realised ensuring encryption of the entire data traffic;
- The access rights are fully controlled by the corresponding data owner.

Since within such a collaboration different tools are going to be used by the collaboration partners, the framework shall also provide a data transformation and integration support in order to allow the participating users to exchange and integrate their local data sets with the ones being published by other collaboration partners. As a result, the users involved benefit by additional knowledge resulting in the combination and integration of the available data sources. Figure 1 reflects this entire CoSpaces Data Space approach.

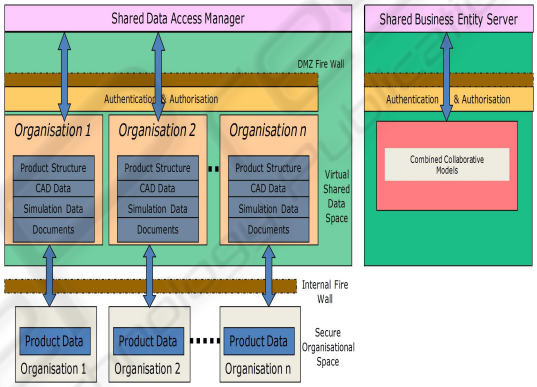

Figure 1: CoSpaces Data Space.

To face real user requirements as mentioned before, the CoSpaces approach realises the data management processing as follows:

Every collaboration partner willing to share data with other partners has to provide a *Virtual Shared Data Space* within their DMZ. Within this Virtual Shared Data Space every partner is able to upload data from the *Secure Organisational Space* and assigns access control policies for the corresponding data sets and partners. Within CoSpaces, this Virtual Data Space is going to be realised by providing dedicated databases as well as a modified version of the BSCW shared workspace system (Horstmann and Bentley, 1997).

This modified version allows the usage of the Shibboleth approach for authentication and authorisation tasks, since for the realisation of the entire authentication and authorisation processing, Shibboleth's federated Single-SignOn and attribute exchange framework has been selected (Assel and Kipp, 2007).

To allow for the need to combine and integrate data artefacts to get additional knowledge, a *Shared Business Entity Server* is mentioned in the CoSpaces approach as well. This Shared Business Entity Server

[8]Quality of Service

provides functionality to convert data according to specific formats and integrate these data sets to a new, global one. It also provides the same security, authentication and authorisation infrastructure as the Virtual Shared Data Spaces and can so consequently be hosted by one of the data providers or by an external trusted third party. The access rights of the combined, collaborative models are defined by all data providers themselves.

4 VIROLAB VIRTUAL LABORATORY

The mission of the ViroLab project is to provide researchers and medical doctors in Europe with a virtual laboratory for infectious diseases for HIV drug resistance.

The virtual laboratory integrates the biomedical information from viruses (proteins and mutations), patients (e.g. viral load) and literature (drug resistance experiments) resulting in a rule-based decision support system (Sloot et al., 2006) for drug ranking. In addition, it includes advanced tools for (bio)statistical analysis, visualisation, modelling and simulation, enabling the prediction of temporal virological and immunological response of viruses with complex mutation patterns for drug therapy.
The virtual laboratory is basically used by medical doctors to review previous results and rankings of recent HIV drug resistance interpretations or by scientists to conduct new experiments and simulations starting from pre-defined process flow templates, which allow an interactive selection of available bioinformatics applications to be combined into one explicit workflow for analysing individual HIV drug resistance. Furthermore, the virtual environment offers different capabilities such as on-demand requests to people more involved or real-time data sharing, in order to allow easy collaborations with other medical professionals for studying and discussing previous results and experiments.

To achieve a smooth integration of the distributed and heterogeneous resources into the overall laboratory infrastructure, a set of virtualisation services that guarantees access to resources in a consistent, resource-independent, and efficient way is being developed to facilitate a direct and on-demand interaction with all available biomedical databases, thus enabling collaborative research and workflow execution.

In order to meet the specific requirements for exchanging the confidential biomedical data sets within such a virtual environment, the solution introduced in ViroLab is built on existing Grid technologies providing the core for the so-called *Data Access Services (DAS)* (Assel et al., 2008).

These services implement standard user interfaces realised as basic Web Service capabilities to guarantee an easy interoperability with different end-user systems, and to support various user groups such as researchers, medical doctors, etc. for accessing the distributed data in a user-friendly way. Furthermore, the DAS also allow the integration of several data resource types to be exposed within the virtual infrastructure. With only one central entry point acting as the only ”visible” and accessible system, users are unaware that they are dealing with a federation of different data resources rather than a single one.

Thus, when answering requests for data, the services need to transform and translate heterogeneous data according to application-dependent formats, access heterogeneous technologies, consolidate data gained from several resources simultaneously, and assure the availability of new/current data while observing data confidentiality and ownership. The latter ones are very crucial in an eHealth scenario and can be seen as one of the most essential and extremely important parts. Therefor, the DAS are equipped with sophisticated security mechanisms based on established technologies like Shibboleth and the Grid Security Infrastructure (GSI) provided by the Globus Toolkit (Barton et al., 2006) to protect the sensible sources of data and to keep the privacy of single data sets (patients). The following figure depicts the overall data access and integration architecture of ViroLab’s virtual laboratory.

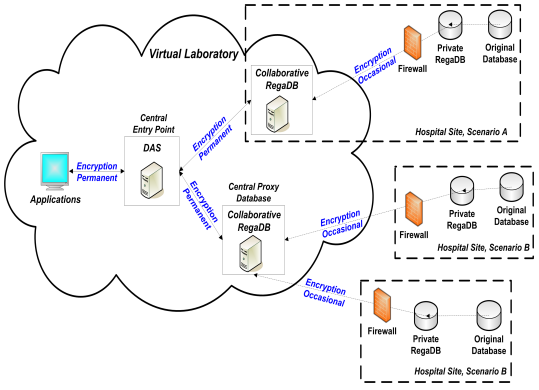

Figure 2: ViroLab’s Data Access and Integration Architecture.

At every data provider site, within the hospitals’ security regions (behind their firewalls), data from the original database(s) are transferred into a private RegaDB installation. RegaDB is a specific HIV data and analysis management environment (Libin et al., 2007) developed by the Rega Institute of the Katholieke

Universiteit Leuven that enables an easy storage and management of biomedical data sets of HIV-treated patients. The transfer is done by exporting data from the original database and converting that data extract through a custom script into the latest RegaDB schema. The transfer can be conducted repeatedly over time at the discretion of the database administrator(s).

Data anonymisation can occur either while transferring data from an original database into a private RegaDB or alternatively when transferring data from a private RegaDB onto a collaborative one.

To contribute data to the ViroLab virtual laboratory, there are two alternative scenarios, both including the upload of data from a private RegaDB into a collaborative RegaDB. The main difference between both solutions is the physical location of the collaborative RegaDBs. Data providers can either host their own collaborative RegaDB installation within a trusted region outside their institutional firewall (within their Demilitarised Zone or DMZ), or they utilise one of the "centrally managed" collaborative RegaDBs hosted by some trusted third parties via a secure connection.

Currently, both of the described scenarios are in the scope of ViroLab. Since some hospital policies basically prohibit an installation of additional server machines and/or software components within their administered networks, the only way to contribute data to the project's workspace is limited to the second possibility as described above.

5 PRELIMINARY RESULTS

At present, both projects are in their implementation phases developing and realising the mentioned approaches. While CoSpaces just finalised the conceptual design phase of the entire data management infrastructure and recently started working on a first prototype to be available until the end of this year, ViroLab already released a first version of the Virtual Laboratory that supports data access and integration of the heterogeneous resources in a limited way. One can deal with these resources as a federated data space, which can be queried by submitting multiple and concurrent requests for gathering any kind of biomedical data sets that still reside in an inhomogeneous state. Basic security features including the encryption of transferred data messages as well as the support for user authorisation based on Shibboleth's authentication and authorisation infrastructure (AAI) are also in place. These data requesting activities are applied within several pre-defined experiments that can be used by virologists but also clinicians to estimate the possible drug resistance for a particular virus mutation. A more detailed description of corresponding experiments can be found in (Gubala et al., 2008). Future releases of ViroLab's Data Access Services will enhance the services' capabilities with respect to performance and scalability, and facilitate the interaction with the services through a specific query language based on natural language terms instead of providing common SQL statements.

6 CONCLUSIONS

Dynamic (business) collaboration, as an exciting and promising field of an interdisciplinary cooperation, will provide new working environments that ease cross-organisational data exchange and communication. It has attracted worldwide attention and several international research projects have already designed and implemented first prototypes for appropriate infrastructures.

Actually, today's systems often stick to static environments instead of developing flexible and dynamic solutions enabling ad-hoc collaborations with on-demand application and data sharing.
Future developments need to take service and resource virtualisation more into account to hide the complexity of the underlying technologies from the users/customers but also to allow on-the-fly modifications of internal interfaces and/or enhancement of existing functionalities while simultaneously keeping and guaranteeing dynamicity, scalability, and performance of the available services/resources. To develop applicable environments for daily business processes, current systems need to be enhanced with reliable models and tools to monitor user activities including data requests or service invocations, and in fact, with sophisticated mechanisms to price and account corresponding user interactions according to current market variances or surprisedly changing user requirements.

In this paper, we have presented the approaches of two running European research projects CoSpaces and ViroLab, which are trying to overcome the complex problem of building dynamic infrastructures for secure data management and data integration. We have identified some key features and requirements, which should be considered during the design and development phase of such environments and which may lead to better and much easier to handle solutions not limited to research but also for the next generation of intercontinental business collaborations.

ACKNOWLEDGEMENTS

The results presented in this paper are partially funded by the European Commission under contract IST-5-034245 through the project *CoSpaces* and through the support of the *ViroLab* Project Grant 027446. The authors want to thank all who contributed to this paper, especially the members of both consortiums.

REFERENCES

Assel, M. and Kipp, A. (2007). A secure infrastructure for dynamic collaborative working environments. In *Proceedings of the 2007 International Conference on Grid Computing & Applications, GCA 2007, June 25-28 2007, Las Vegas, Nevada, USA*. CSREA Press.

Assel, M. and Krammer, B. (2007). Towards innovative healthcare grid solutions: Virolab - a virtual laboratory for infectious diseases. In *Proceedings of the German e-Science Conference 2007, May 02-04 2007, Baden-Baden, Germany*. Max Planck Digital Library.

Assel, M., Krammer, B., and Loehden, A. (2007). Data access and virtualization within virolab. In *Proceedings of the 6th Cracow Grid Workshop 2006 (CGW06), Oct. 15-18 2006, Krakow, Poland*. ACC-Cyfronet AGH.

Assel, M., Krammer, B., and Loehden, A. (2008). Management and access of biomedical data in a grid environment. In *Proceedings of the 7th Cracow Grid Workshop 2007 (CGW07), Oct. 16-18 2007, Krakow, Poland*. ACC-Cyfronet AGH.

Barton, T., Basney, J., Freeman, T., Scavo, T., Siebenlist, F., Welch, V., Ananthakrishnan, R., Baker, B., Goode, M., and Keahey, K. (2006). Identity federation and attribute-based authorization through the globus toolkit, shibboleth, grid- shib, and myproxy. In *Proceedings of 5th Annual PKI R&D Workshop, April 04-06 2006, Gaithersburg, USA*. NIST Interagency Reports.

Elmroth, E., Gardfjll, P., Norberg, A., Tordsson, J., and stberg, P.-O. (2007). Designing general, composable, and middleware independent grid infrastructure tools for multi-tiered job management. In *Proceedings of the CoreGrid Symposium, Aug. 28-31 2007, Rennes, France*. Springer-Verlag.

Gubala, T., Balis, B., Malawski, M., Kasztelnik, M., Nowakowski, P., Assel, M., Harezlak, D., Bartynski, T., Kocot, J., Ciepiela, E., Krol, D., Wach, J., Pelczar, M., Funika, W., and Bubak, M. (2008). Virolab virtual laboratory. In *Proceedings of the 7th Cracow Grid Workshop 2007 (CGW07), Oct. 16-18 2007, Krakow, Poland*. ACC-Cyfronet AGH.

Hasselmeyer, P., Qu, C., Schubert, L., Koller, B., and Wieder, P. (2006). Towards autonomous brokered sla negotiation. In *Proceedings of the eChallenges 2006 (e-2006) Conference, Oct. 25-27 2006, Barcelona, Spain*. IOS Press.

Horstmann, T. and Bentley, R. (1997). Distributed authoring on the web with the bscw shared workspace system. *StandardView*, 5(1):9–16.

Libin, P., Deforche, K., Laethem, K. V., Camacho, R., and Vandamme, A.-M. (2007). Regadb: An open source, community-driven hiv data and analysis management environment. In *Proceedings of the Fifth European HIV Drug Resistance Workshop, March 2007, Cascais, Portugal*. Reviews in Antiretroviral Therapy.

Schubert, L., Wesner, S., and Dimitrakos, T. (2005). Secure and dynamic virtual organizations for business. In *Proceedings of the eChallenges 2005 (e-2005) Conference, Oct. 19-21 2005, Ljubljana, Slovenia*. IOS Press.

Sloot, P., Boucher, C., Bubak, M., Hoekstra, A., Plaszczak, P., Posthumus, A., van de Vijver, D., Wesner, S., and Tirado-Ramos, A. (2006). Virolab - a virtual laboratory for decision support in viral diseases treatment. In *Proceedings of the 5th Cracow Grid Workshop 2005 (CGW05), Nov. 20-23 2005, Krakow, Poland*. ACC-Cyfronet AGH.

Wilson, M., Chadwick, D., Dimitrakos, T., Doser, J., Giambiagi, A. A. P., Golby, D., Geuer-Pollmann, C., Haller, J., Ketil, S., Mahler, T., Martino, L., Parent, X., Ristol, S., Sairamesh, J., and Schubert, L. (2005). The trustcom framework v0.5. In *Proceedings of the 6th IFIP Working Conference on Virtual Enterprises (PRO-VE '05), Sep. 26-28 2005, Valencia, Spain*. Springer-Verlag.