Management and Access of Biomedical Data in a Grid Environment

Cracow Grid Workshop 2006
October 15th – 18th, 2006
Cracow, Poland

Authors

Assel Matthias (assel@hlrs.de)

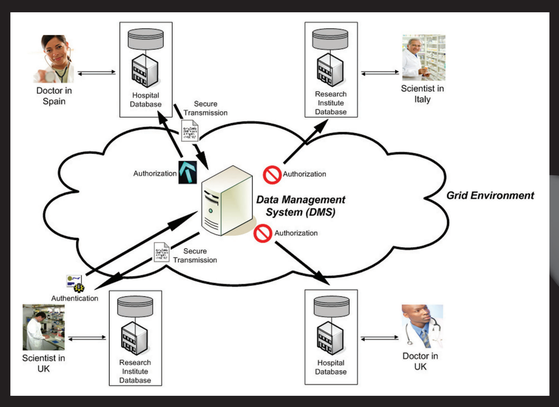
Krammer Bettina (krammer@hlrs.de)

Loehden Aenne (loehden@hlrs.de)

HLRS – High Performance Computing Center of University Stuttgart

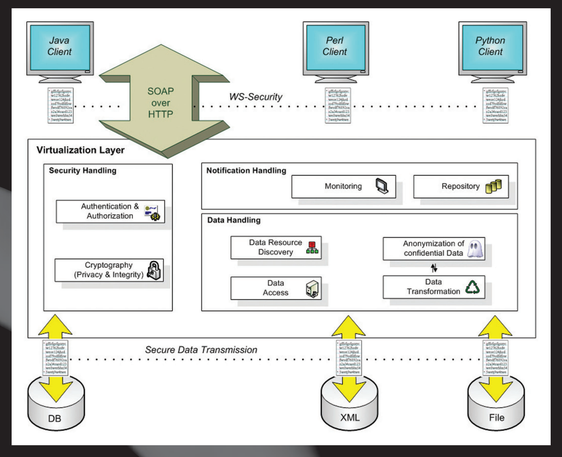
Nobelstr. 19, 70569 Stuttgart, Germany

Security Handling



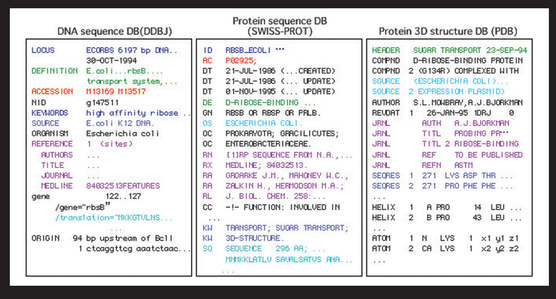
A Data Management System (DMS) controlling access rights for distributed biomedical databases

Data Management Virtualization Layer



A virtualization layer for dealing with distributed confidential data resources

Data Heterogeneity



Example of the RbsB protein, a ribose binding protein, which is differently represented in the DDBJ, SWISS-PROT and PDB databases
(Each color stands for the same information)

Data Management on a Grid

- Applications require data at very large scale, both in size and distribution
- Complexities of data management on a grid arises from
  - Scale
  - Dynamism
  - Autonomy
  - Heterogeneity
  - Distribution

→ Layer of virtualization services that is transparent to users and that guarantees access in a consistent, data resource-independent way

Data Management Virtualization Layer

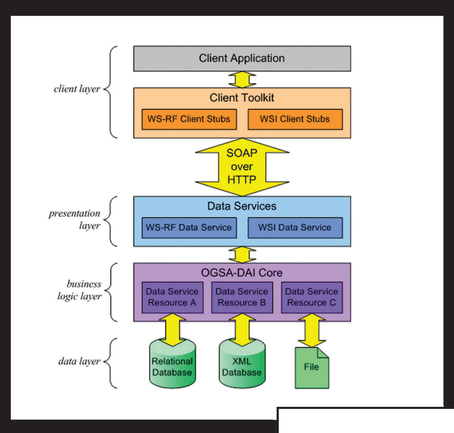
- Data Access: OGSA-DAI as the core providing standard mechanisms for discovering and querying distributed data resources
- Transformation: Load dynamically data definition schemas into the transformation engine to transform the queried data into a unified „biomedical“ data format
- Security:
  - Extend the OGSA-DAI security mechanisms with more sophisticated functionalities
  - Use a policy driven model like the TrustCoM project
  - Anonymization as part of the transformation process
- Notification: Use the WSRF notification mechanisms for any internal message exchange including information as well as error messages

Requirements to virtualization services in the context of biomedical data

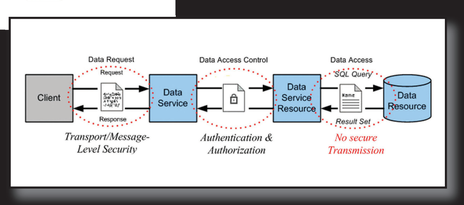
- Standard mechanisms for data access & discovery of distributed data resources at different locations
- Transformation of heterogeneous data
- Security Handling
  - Authentication & Authorization
  - Secure Transmission / Encryption
  - Anonymization of confidential data
- Monitoring & Notification Handling

Existing Data Management Technologies

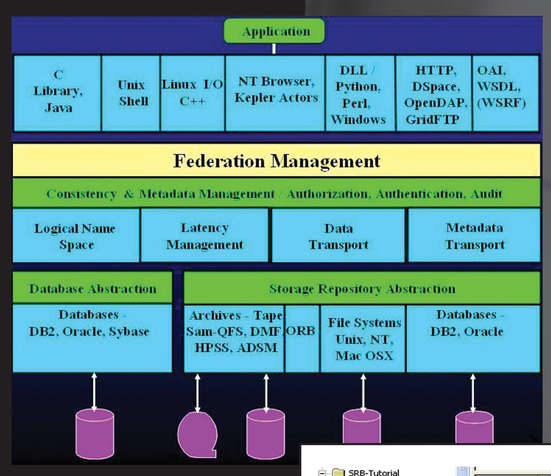
OGSA-DAI
(Open Grid Services Architecture Data Access and Integration)



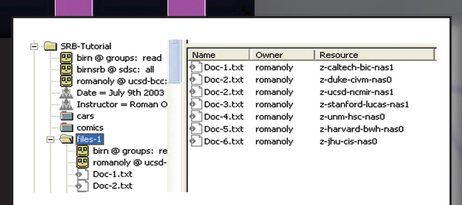
OGSA-DAI is a middleware product that allows different data resources to be accessed via Web Services



San Diego Supercomputer Center
Storage Resource Broker (SRB)



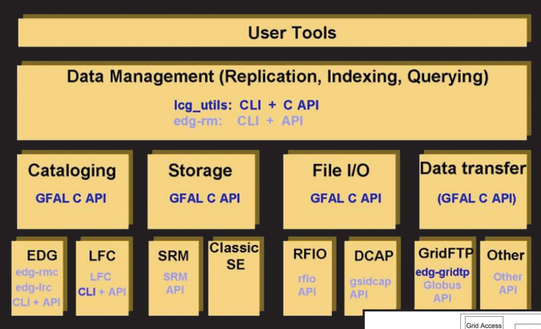
The SRB is a Data Grid Management System (DGMS) or simply a logical distributed file system which presents the user with a file hierarchy



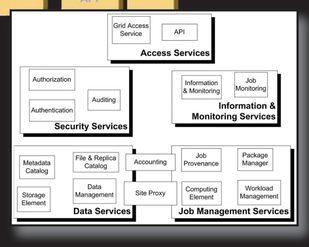
Conclusions

- Data access to distributed biomedical data resources must be controlled by a sophisticated Data Management System (DMS)
- High sensitivity of biomedical information demands strong security mechanisms
- Privacy and confidentiality of patients must be kept and protected before sharing medical data
- Anonymized information needs to satisfy the requirements under the Data Protection Act (Legal and ethical issues)
- Usability of different end-user applications should be considered and simplified by using established standards, e.g. SOAP over HTTP in the form of Web Services
- Common solutions do not strongly fulfill the specific prerequisites
- OGSA-DAI provides a good basis but different services need to be adapted

EGEE – gLite Data Management Subsystem



The gLite Data Management component uses standard Grid Services to provide a virtual file system



Comparison

	OGSA-DAI	gLite	SRB
Interfaces for Data Access & Discovery	Web Service Standard (SOAP over HTTP)	Library Functions (Perl, Java)	Library Functions for Higher Level Software
Transformation of (heterogeneous) Data	Data can be transformed, but needs to be adapted	Not provided	Not provided
Authentication & Authorization	X.509 Certificates	User Certificates	Password Authentication
Secure Transmission / Encryption	Message/Transport-Level Security	Transport-Level Security/ WS-Security	Mechanisms for data encryption
Anonymization of confidential Data	Not provided	Not provided	Not provided
Monitoring & Notification Handling	Simple logging but no notification system	Logging & Bookkeeping Subsystem	Simple logging functionalities
Extensibility	Source available - designed to be extendable	Source not available	Source available for academic research
Basic Applications	Mainly used in Grid projects around the globe	Virtual File Systems	Virtual File Systems (Digital Libraries)

The Data Management Technologies with respect to the fulfillment of requirements for biomedical data resources

References

Vijayshankar et al. Data Access and Management Services on Grid. IBM Research Center San Jose, USA. 2002

Paul Watson. Databases and the Grid. University of Newcastle, UK

Lingfen Sun, Emmanuel C. Ifeachor. The Impact on Healthcare. University of Plymouth, UK

The BioGrid Project. http://www.biogrid.jp/

The DNA Data Bank of Japan (DDBJ). http://www.ddbj.nig.ac.jp/

The UniProtKB/Swiss-Prot Protein Knowledgebase (SWISS-PROT). http://www.ebi.ac.uk/swissprot/

The RCSB Protein Data Base (PDB). http://www.rcsb.org/pdb/Welcome.do

The OGSA-DAI Project. http://www.ogsadai.org.uk/index.php

The SDSC Storage Resource Broker. http://www.sdsc.edu/srb/index.php/Main_Page

gLite – Lightweight Middleware for Grid Computing. http://glite.web.cern.ch/glite/

The Globus Toolkit Homepage. http://www.globus.org/toolkit/

The TrustCoM Project. http://www.eu-trustcom.com/

The ViroLab Project Website. http://www.virolab.org:8080/virolab