



SIMDAT

Data Grids for Process and Product Development using Numerical Simulation
and Knowledge Discovery

Project no.: 511438

Grid-based Systems for solving complex problems – IST Call 2
Integrated project



Deliverable

***D3.1.3. SIMDAT Distributed Data Repository Access
Infrastructure (Second Version) &
D3.1.4 SIMDAT Report on SIMDAT Distributed Data
Repository Access Infrastructure. Evaluation and
validation***

Start date of project: 01 September 2004

Duration: 48 months

Due date of deliverable: 01 April 2006

Actual submission date: 9 May 2006

Lead contractor for this deliverable: Intel

Revision: 1.0

| Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006) | | |
|--|---|---|
| Dissemination level | | |
| PU | Public | X |
| PP | Restricted to other programme participant (including the Commission Services) | |
| RE | Restricted to a group specified by the consortium (including the Commission Services) | |
| CO | Confidential, only for members of the consortium (including the Commission Services) | |

Revision history

| Date | Version | Author | Modification |
|-------------|----------------|----------------------|--|
| 06-03-26 | V 0.1 | Michael Krüger | Initial draft version |
| 06-03-31 | V 0.2 | Michael Krüger | Updated and revised |
| 06-04-04 | V 0.3 | Dr. Thomas Kentemich | Corrections of logic errors |
| 06-04-06 | V 0.4 | Michael Krüger | Improvements according to internal reviewers' comments |
| 06-04-06 | V 0.5 | Michael Krüger | Added technology-state-of-the-art and new requirements section |
| 06-04-10 | V 1.0 | Hans-Christian Hoppe | Final version |

Copyright

Copyright © Intel and other members of the SIMDAT consortium, www.simdat.org, 2006

Table of contents

| | | |
|-----|--|----|
| 1 | Introduction | 4 |
| 1.1 | Purpose | 4 |
| 1.2 | Scope | 4 |
| 1.3 | Definitions, acronyms and abbreviations | 5 |
| 1.4 | References | 5 |
| 2 | Technology improvements | 7 |
| 3 | State-of-the-art surveys and the contribution/position of SIMDAT | 8 |
| 4 | Cooperation with and feedback from the application activities | 10 |
| | Investigation of Peer-to-peer techniques | 11 |
| 5 | Specific and modular requirements documentation | 13 |
| 5.1 | Aerospace prototype requirements | 13 |
| 5.2 | Automotive prototypes requirements | 16 |
| 5.3 | Meteorology prototype requirements | 17 |
| 5.4 | Pharmaceutical prototypes requirements | 18 |
| 6 | Implementation and test plans for PM 24 and PM 30 | 19 |
| 7 | Conclusion | 21 |

1 Introduction

1.1 Purpose

This document represents the deliverables D3.1.3 “SIMDAT Distributed Data Repository Access Infrastructure (Second Version)” and D3.1.4 “SIMDAT Report on SIMDAT Distributed Data Repository Access Infrastructure. Evaluation and validation” of the Integrated Project IST-2002-511438 (SIMDAT), as specified in the Annex 1- “Description of Work” [1]. It is the combined third and fourth deliverable for work package 3 and follows the consolidated requirements report D3.1.1 and the first version of the Distributed Data Repository Access infrastructure D3.1.2.

The intended audience for this document is the application and technology partners within the SIMDAT consortium, as listed in Section 3 of Annex 1 [1] as well as the European Commission Services. SIMDAT partners have a broad range of deep expertise in both application sectors and horizontal technology activities. The document is structured to give a view of the state of the Distributed Data Repository Access (DDRA) activity from the perspective of the technology partners and each of the four application activities as of project month 18 (PM18).

The document presents and discusses the second version of the DDRA infrastructure designed, developed and implemented in conjunction with and on top of the Integrated Grid Infrastructure (as put forward by work package 2). It integrates input from application and technology activities, and is the result of numerous discussions with the respective SIMDAT partners.

1.2 Scope

The main challenge for SIMDAT is to develop and deploy technology and techniques that improve the ability of industrial organizations to collaborate in a flexible and dynamic fashion. This collaboration has to take place at a deep technical level, with applications, databases and resources communicating directly with one another in a controlled and secure fashion. The complex problems to be solved involve multiple data repositories describing many aspects of the product and process development, typically hosted in different departments and at different sites, and not currently linked with each other in a direct fashion. To make the SIMDAT vision reality, these data repositories have to be made accessible in a transparent and reliable way regardless of their location. In some cases, this will also involve replication and synchronization of data repositories. In most industrial sectors like the automotive and aerospace industries, implementations for interlinking the distinct distributed data repositories involved in product design, development and production did not exist prior to the SIMDAT project. The concepts for DDRA services have been developed in PM 1–6 and are reported in Deliverable D3.1.1. An initial implementation of a subset was developed in PM 7–11 and rolled out at the end of PM 11. That implementation has been refined in PM 12-17. In close cooperation with the application activities, detailed feedback on the existing and potential future use cases was harvested, including requirements for the next generation of prototypes, and a list of potential shortcomings of the current DDRA infrastructure.

1.3 Definitions, acronyms and abbreviations

DAIS: database access and integration services

FTP: file transfer protocol

GGF: Global Grid Forum

GRIA: GRID Resources for Industrial Applications (<http://www.gria.org>)

HPC: high performance computing

OASIS: Organization for the Advancement of Structured Information Standards (<http://www.oasis-open.org>)

OGSA: Open Grid Services Architecture (<https://forge.gridforum.org/projects/ogsa-wg>)

OGSA-DAI: Open Grid Services Architecture – Database Access and Integration (<http://www.ogsadai.org.uk>)

RDBMS: relational database management system

SIMDAT: Data Grids for Process and Product Development using Numerical Simulation and Knowledge Discovery

SRB: Storage Resource Broker (<http://www.sdsc.edu/srb>)

VGISC: Virtual Global Information System Centre, a virtual node of a world-wide meteorological information system

VO: virtual organization

WS-I: web service interoperability <http://www.ws-i.org/>

WSRF: web services resource framework
http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrf

XML: extended markup language

1.4 References

1. SIMDAT Annex 1, second version (covering PM 1 to 30)
2. IBM DB2, <http://www-306.ibm.com/software/data/db2/>
3. Oracle, <http://www.oracle.com/index.html>
4. OGSA-DAI Overview, <http://www.ogsadai.org.uk/docs/current/doc/DAIOverview.html>
5. Globus Toolkit Reference, <http://www.globus.org/>
6. Global Grid Forum GGF, <http://www.ggf.org/>

-
7. M. Surridge, M. Boniface: SIMDAT Architecture Issues, Technical Note SIMDAT T02
 8. Stephen C. Phillips, Mike Boniface, GRIA and OGSA-DAI Integration Design, Technical Note SIMDAT T14

2 Technology improvements

Complex data repositories are key to the modern design, development, and production of complex products and services. Different data repositories store product design data, results from physical tests, and numerical simulation results characterizing the functional properties of products and processes, as well as knowledge about the design process itself, material properties and development strategies. These data repositories might be a collection of data bases potentially combined with a collection of flat files. The Meteo activity uses some custom database / archive systems that are neither relational database systems or flat file repositories; these are accessed by using a bespoke request language. Effective correlation of data generated in different departments or at different sites within a global organization is a crucial problem for all industries represented in SIMDAT. This solution requires DDRA services that interact with the semantic definition of the data models and storage formats involved, and enable the retrieval of all relevant information even though the data might be distributed across heterogeneous data repositories.

A first release of the DDRA prototype was derived from existing work, mainly the OGSA-DAI (Open Grid Services Architecture – Data Access and Integration) tools before PM 12. OGSA-DAI specifies a service-based interface for transparent access to data repositories (which can be relational or XML data bases or file systems) across Grid nodes and provides a reference implementation supporting many of the major database engines and packages (IBM DB2, Oracle, MySQL, XIndices, and many more). To accommodate the needs of most of the application activities of SIMDAT for an integrated Grid and DDRA solution, the adaptation and integration of OGSA-DAI with the WS-I based GRIA layer chosen as the Integrated Grid Infrastructure in WP2 was performed within that phase of SIMDAT. GRIA provides an access control model based on business-to-business service provision, supporting business models and processes that allow collaborative resource sharing for industry. The GRIA/OGSA-DAI integration enables dynamic management of database resources whereby users are able to create, manage and destroy databases located at remote service providers in accordance with their application needs.

The integration solution did require a new GRIA service to be written: the “OGSA-DAI Service”. This service is a wrapper for OGSA-DAI, communicates with OGSA-DAI through direct Java calls, and provides access control through the PBAC mechanisms of GRIA. As a GRIA service it makes use of the GRIA’s WS-Security infrastructure providing user authentication, and data confidentiality and integrity. For a detailed description of the integration please refer to Deliverable D3.1.2. Working with the application activities, it was found that the GRIA/OGSA-DAI integration fulfills all requirements of the PM 12 prototypes. Therefore in the months PM12-18 the focus was not on extending functionality, but on improving stability and robustness. Business processes were and are being developed to assign, control access and account for database resources, i.e.:

- Tendering for resources
- Accounting and billing for usage
- Delegation based on database privileges

Generally integration with the WP2 Integrated Grid Infrastructure and support for the higher level grid technologies, i.e. those building on top of the DDRA services was provided. This enables the application activities to use the available DDRA services in real-world scenarios. Additionally a lot of effort was put into mutually consulting with the application activities.

3 State-of-the-art surveys and the contribution/position of SIMDAT

For the DDRA activity I SIMDAT, the most important external source of technology is the UK eScience OGSA-DAI project. Here the work package partners Intel and IT Innovation are in direct contact with the OGSA-DAI team.

Availability and functionality of the previous releases of OGSA-DAI including the version currently used as a base for SIMDAT DDRA (release 6) have been described in previous deliverables. Additional information is available from the OGSA-DAI web site at <http://www.ogsa-dai.org.uk/>. For this chapter, the OGSA-DAI roadmap starting at release 6 is relevant. It should be noted that there are two flavors of OGSA-DAI (one for fully WSRF/OGSA compliant Grids, and one for WS-I), both of which will be continued and receive the same set of changes and improvements.

The current release 7 is concerned with the integration and provision of a comprehensive set of facilities that can sustain the existing user community. A first implementation of the DAIS specification as endorsed by the GGF was done and distributed query processing (OGSA-DQP) is now integrated. Additional featured in this release include:

- Transparent access to different varieties of Grid data services
- Refactored SQL activities
- Refactored data resource configuration
- New support for sessions and concurrency

The extension interfaces within OGSA-DAI were revised and improved; they enable external developers to customize and extend OGSA-DAI for their own applications and have been a key factor behind OGSA-DAI's widespread take-up.

Future releases of OGSA-DAI are scheduled up to 2007, with preliminary definitions of new features and improvements available: Release 8, which is scheduled for the second quarter of 2006, will introduce improvements to the core OGSA-DAI engine:

- Performance enhancements
- Extended functionality of the Grid data services (support for transactions, monitoring and management)
- Security improvements
- Integration of additional data resources

Scheduled for late 2006, release 9 aims to integrate Federation and implementation of Notification Specifications for Web Services (FINS) from OMII and additional data resources like the Storage Resource Broker (SRB) that has been analyzed in D3.1.1.

After that the release 10 will focus on workflow integration, including the Taverna and BPEL components used within SIMDAT.

Additionally the OGSA-DAI developers are continuously working on number of items regarding:

- Functionality of services and containers

-
- Support for more types of data resources
 - Additional data resource functionality
 - Support for application developers
 - Support of workflow scenarios
 - Support for data integration

To guide the further development of OGSA-DAI, and ensure that features and properties relevant for SIMDAT are receiving sufficient attention, WP3 partners Intel and IT Innovation have established close ties with the OGSA-DAI team. In discussions with that team, it will be determined which improvements/extensions will be done by the OGSA-DAI project, and which ones will be done within WP3 and integrated into the OGSA-DAI codebase. To ensure quid-pro-quo, WP3 plans to donate OGSA-DAI extensions developed within SIMDAT to the OGSA-DAI project, provided that they are not inextricably intertwined with specific application activities.

This style of open collaboration will serve both projects: OGSA-DAI gets the “straight talk” from industrial use cases as well as key extensions, while SIMDAT profits from the effort put into the OGSA-DAI development.

4 Cooperation with and feedback from the application activities

A suitable data access layer integrated with the Grid infrastructure prototype was made available to the SIMDAT partners in PM12 and refined in the months until PM18. After transitioning from initial (proprietary) data access platforms, the application activities are now able to rely on the data access services available through the DDRA layer. The work in these areas benefits from having a stable, reliable platform and being able to get support during installation and operation. Emphasis was put on portability and interoperability by relying on de-facto standard interfaces (OGSA-DAI as an implementation of the GGF-endorsed DAIS standard), thereby protecting the SIMDAT developers' and users' investments.

In the UK e-Science Core programme, the OGSA-DAI project has produced a toolkit that enables the transparent, remote access to data repositories, and contains support for implementing data federation and other extensions. OGSA-DAI (data access and integration) implements an interface for read and write access to remote databases and for distributed queries, with bindings for standard WS-I Web Services, and for the WSRF Web Services extensions standard approved by OASIS. An implementation of OGSA-DAI working with the Globus Toolkit 4 is available from the University of Edinburgh; it supports other WSRF-compliant Grid systems, and provides a WS-I based interface that can work with purely Web service based Grid implementations (like GRIA, which is the base of the WP2 Integrated Grid Infrastructure). In the first phase of SIMDAT the focus was on integrating these components into a comprehensive data access layer based on Web Services and hardening/extending it to match the requirements of the application activities.

In the European EGEE project, components for the management of data catalogues and replicas have been developed and deployed as parts of the gLite Grid system. Prime user is the CERN LHC (Large Hadron Collider) Grid, which will distribute the enormous data volumes generated by the LHC experiments starting in 2007. The data components will be evaluated on whether they could supply the replication and caching functionality for the third and fourth phase of SIMDAT.

The aerospace PM12 prototype data management strategy was based on a combination of OGSA-DAI, GRIA data services and bespoke data integration components all of which are incorporated into the prototype as part of the workflow. The data files of each application were in most cases proprietary and to allow for workflow composition bespoke data translation components had to be developed. Given the advanced state of the aerospace activity, the partners in the DDRA as well as in the aerospace activity have relied on the use case from this activity to evaluate the combination wrapper OGSA-DAI/GRIA. The feedback and additional requirements resulting from the work here will lead to further advancement of the DDRA usage and technology in SIMDAT.

The three SIMDAT Automotive prototypes developed for month 12 and the underlying use cases did focus on different aspects. They made use of advanced distributed data repository access to various degrees. With the Audi prototype (Auto-1) the main focus is on data access. The RENAULT/IDESTYLE use case covers mainly virtual organizations (security) aspects. The LMS/NOESIS prototype's main focus is on workflow. During the further course of the project, Auto-1 and Auto-2 will be combined into a single prototype, merging the data access, workflow and security solutions into a single architecture. In addition, it will be

especially important to have an option to federate OGSA-DAI accessible databases through a semantic layer hiding the database specifics and mapping between them, all the while fully accommodating the security requirements of the industrial partners. Within the automotive activity a lot of technical discussions on the use of OGSA-DAI for the Auto-1/2 prototype have taken place. It was found that for the future generation of that prototype a key issue is the integration of semantic mediation and OGSA-DAI, as this will address security concerns of the industrial partners.

The Meteo activity partners were concerned that the Grid middleware relevant standards are still a moving target (WSRF, WSI,). To avoid architectural dead-ends, the partners decided to define the architecture and implement the components and protocols required for the PM12 prototype themselves. The data repositories in Meteo do expose an OGSA-DAI interface to the VGISC components, which communicate through XML messages to retrieve data and harvest metadata from their repositories. The prototype was based on standard Web Services hosted in a J2EE environment. Internally OGSA-DAI was used to access the DB available locally at the partner sites. This made it possible to support the wide variety of specialized repositories used in the Meteo field with minimal effort: all that was needed to bring a new type of repository online was to code an OGSA-DAI compatible interface.

The access to distributed data using Grid mechanisms is not efficient enough, according to the high standards of the Meteo community. The main reason is that current Grid data transport mechanisms are not efficient enough. The second (and larger) challenge is to ensure the consistency of data across the VGISC nodes (at least the metadata catalogues need to be consistent across all nodes). In the Meteo activity, the future work will focus on developing an efficient engine for synchronization and replication of meta-data, as well as improving the data transfer and messaging protocols of the current prototype. Expected results include:

- Software to manage the mesh network allowing for dynamic insertion, node to node message routing in a firewall protected environment with limited connections, while still having highly reliable features. Here some relation to the peer-to-peer techniques developed in / for the Meteo activity can be found.
- Replication software to synchronize the metadata efficiently and reliable. All the while this Distributed Data Repository Access component should be generic to maybe allow to be used by the other activities.

Within the Pharma activity a database access layer (in the form of Lion's SRS system) did already exist at the project beginning. Hence, the original project plan did forgo the development and/or deployment of DDRA services in favor of exploiting SRS. To facilitate the creation of cross-Grid indices, a distributed file system (IGOR) that leverages peer-to-peer (P2P) techniques has been designed.

As a consequence of the reduced role of partner Lion, new scenarios have been evaluated and the Pharma plans have been amended in the SIMDAT Annex I. The new scenarios are much more inline with regular business-to-business (B2B) scenarios, and early discussions strongly indicate that DDRA technologies will play a much larger role. At the time of writing this Deliverable, discussions about the DDRA implications and requirements for the Pharma B2B prototype are underway.

Investigation of Peer-to-peer techniques

As described in detail in Pharma application activity Deliverables, partner University of Karlsruhe (UKA) has continued its development of the peer-to-peer file system IGOR. A few modules have been released as alpha versions for general testing, and reported bugs have been

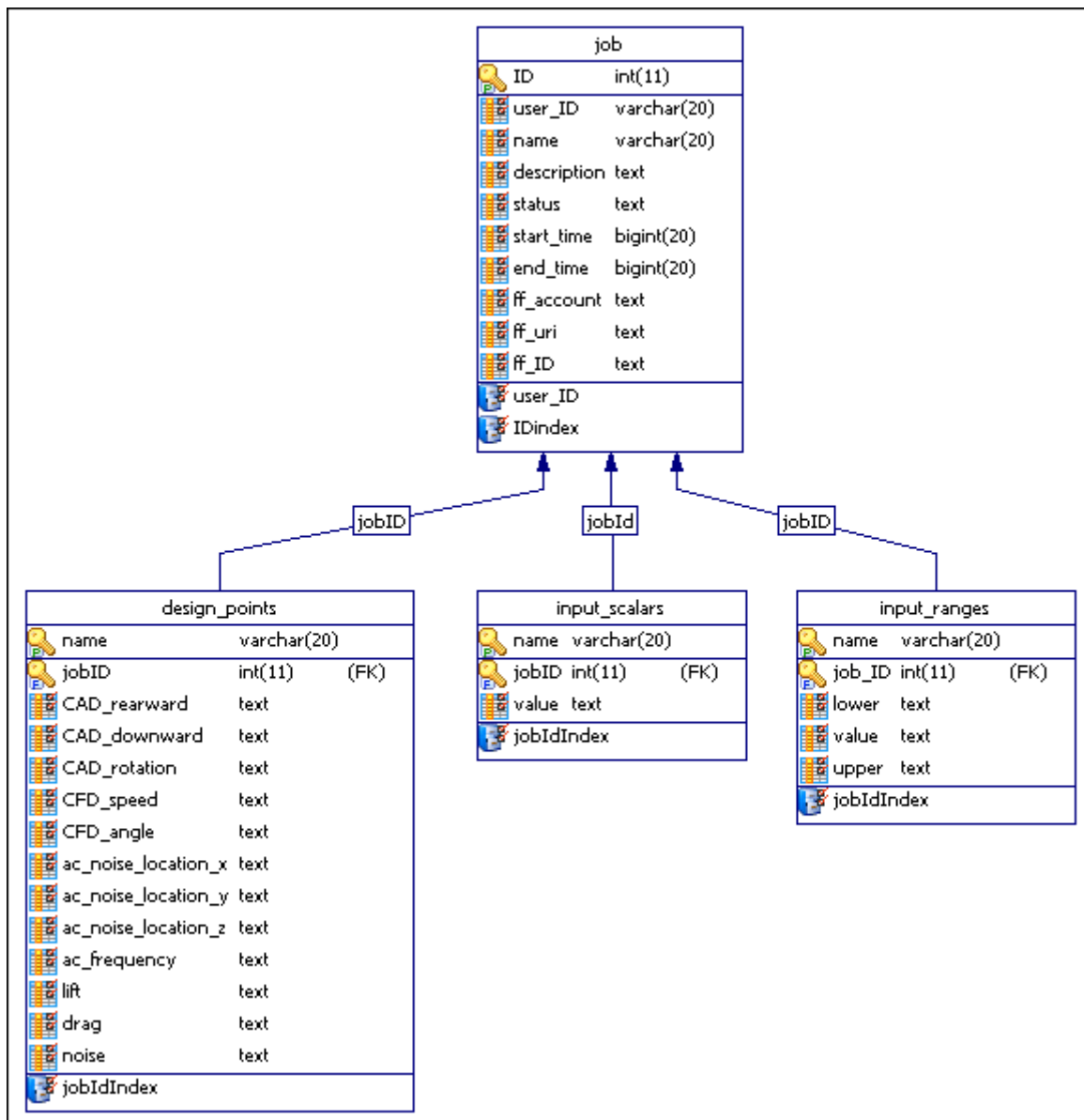
identified and resolved. Other aspects of IGOR are still in the design phase. UKA has tried to accommodate the perceived needs of the non-Pharma partners; however. In the second phase of SIMDAT, detailed discussions on the role of IGOR for the other application activities and on opportunities to use IGOR technology inside the DDRA layer to facilitate replication and caching will take place.

5 Specific and modular requirements documentation

In this chapter the updated and more concrete requirements from the application activities towards the Distributed Data Repository Access work package are captured.

5.1 Aerospace prototype requirements

The integration of OGSA-DAI WS-I with GRIA was inspired in large parts by the aero prototype and its need for a solution for distributed data access for relational databases and simulation files. OGSA-DAI WS-I is used to store metadata with references to data files stored by GRIA data services. A database schema has been developed for the aerospace application that is used to store design of experiments along with the resulting data for each design point and sufficient metadata to determine how the results were derived. This schema is shown in the figure below.



The aerospace prototype will consist of the architecture described in the aerospace Deliverable and will be implemented on the SIMDAT Aerospace Grid with the services deployed at each partner site to enable the scenario to be run at any time. A portal to the system will be developed to allow the remote demonstration of the system. Based on that the following requirements have been derived by the partners in the aerospace activity and the work packages on Integrated Grid Infrastructure and Distributed Data Repository Access. Requirements are on a joint basis for both technology work packages since in this activity the technologies are entwined especially closely.

5.1.1 Functional Requirements

| Requirement | Description | Priority | Implemented | Remaining |
|-------------------------------------|--|----------|-------------|----------------------------|
| WS-I compliant services | Standards compliant Web Service interfaces | High | ✓ | |
| Identity management services | CA, Revocation Service | High | Partial | Revocation service |
| Authentication service | Ability to authenticate users (X509 based) Authenticated transactions Ability to handle federated identity for exploitation phase | High | Partial | Federated identity |
| Authorization Service | Policy driven access control to resources Dynamic policy management for exploitation phase | High | Partial | Dynamic policy management |
| Access to legacy applications | Service container for legacy applications including submission access to compute service | High | ✓ | |
| Data transfer / access service | Transfer/access of flat files Database access including schema publishing in exploitation phase Policy based access control | High | ✓ | |
| Single interface to compute service | Ability to access compute services with different scheduling requirements through single interface Reservation of compute resources Sandboxing runtime in compute service including ability to specify sandbox environment | Medium | Partial | Reservation and sandboxing |
| Resource discovery | Ability to discover alternate services due to service failure or unavailability | Low | × | |

5.1.2 Performance Requirements

| Requirement | Description | Priority |
|----------------------------|--|----------|
| Fast data transfer service | Timely transfer of data of the order of Gigabytes comparable to performance of the ftp protocol | High |
| Scalability | The infrastructure needs to be scalable to enable exploitation by at least 100 simultaneous users | High |
| Manageability | The infrastructure and policy specification management overhead should not grow exponentially with scale | High |

5.2 Automotive prototypes requirements

The automotive application activity partners are working on the PM24 and PM30 prototypes. Here the requirements towards the Distributed Data Access work package can in general be condensed to a need for consulting and support for the implementation of a secure protocol with data encryption for the transfer of Meta and bulk data and support for the implementation of authentication and authorization for data access via OGSA-DAI. In more detail as agreed upon with the automotive partners:

| Requirement | Description | Priority |
|---------------------------------|---|----------|
| OGSA-DAI activity | OGSA-DAI Services for LMS Tec.Manager/ MSC SimManager | High |
| Means of transfer for mass data | Transfer of mass data between MSC SimManager and LMS Tec.Manager | High |
| Fast data transfer | Fast transport layer and a reliable underlying protocol | High |
| Secure data transfer | Encrypted communication channels for metadata requests and mass data transfer | High |
| Access rights handling | Provision of the most suitable solution for handling access rights when accessing a database through Web services | Low |
| User credentials delegation | Secure delegation of user credentials in WSRF and OGSA-DAI based environments | High |

5.3 Meteorology prototype requirements

The Meteo partners have stated the following requirements towards the Distributed Data Repository Access work package:

| Requirement | Description | Priority |
|---|---|----------|
| Discovery and request functionality | Functionality to search and identify relevant database Human to formulate queries directly to catalogue (interactive access) Machine-to-machine queries (batch access) | High |
| Catalogue synchronization | Provide an unified view of all shared data through a synchronized catalogue Implement a reliable synchronization service on a mesh network topology | High |
| Data replication / caching | Provide mechanisms to replicate and cache real-time data across the nodes | High |
| Universal interface for Meteo databases | Provide a uniform Interface with existing meteorological databases (flat file repositories, RDBMS database, off-line archives, XML DB). Extend the Data-Repository component Provide the ability to easily add new data source | High |
| Virtual database administration tool | Administration of the existing local database must be independent of the administration of the Virtual Database | High |
| Simple adding of sources | Provide the ability to simply add new data sources | |
| Fast announce / notification mechanism | Provide mechanism for a node to quickly announce to the other nodes the availability of some special datasets like notification, push mechanisms, QoS, prioritization | |

5.4 Pharmaceutical prototypes requirements

The partners involved in the Pharma activity are working on two different scenarios. The first scenario has been largely developed in the first phase of the project, and is based on partner Lion's SRS product. An additional scenario is the so-called the business-to-business scenario. Both are described in detail in the respective deliverable by the activity. Due to the quite recent developments especially in the B2B scenario the requirements stated here might very well be adapted slightly during the later phases of the project. Since the project partners working on the peer-to-peer technology are members of the Pharma activity not all the requirements cited below will be worked on in the Distributed Data Repository Access WP. Indeed most issues will probably be solved within the Pharma work package itself. For the time being the requirements are stated as given by the Pharma work package.

| Requirement | Description | Priority |
|--|--|-----------------|
| B2B service provision | To accept service requests from a range of customers with low overhead and low lead time | Medium |
| B2B service acquisition and cost | To buy in services in a timely fashion at lower cost than would be achieved with in-house solution alone | Medium |
| B2B security of customer intellectual property | To be able to demonstrate that security of data and access is such that IP is not put at risk. | Medium |
| B2B service provision (supplier) | To accept service requests from a range of customers with low overhead and low lead time | Medium |
| B2B security of supplier intellectual property | To be able to guarantee the integrity of data and resulting analysis for each customer to ensure their IP is not put at risk | Medium |

6 Implementation and test plans for PM 24 and PM 30

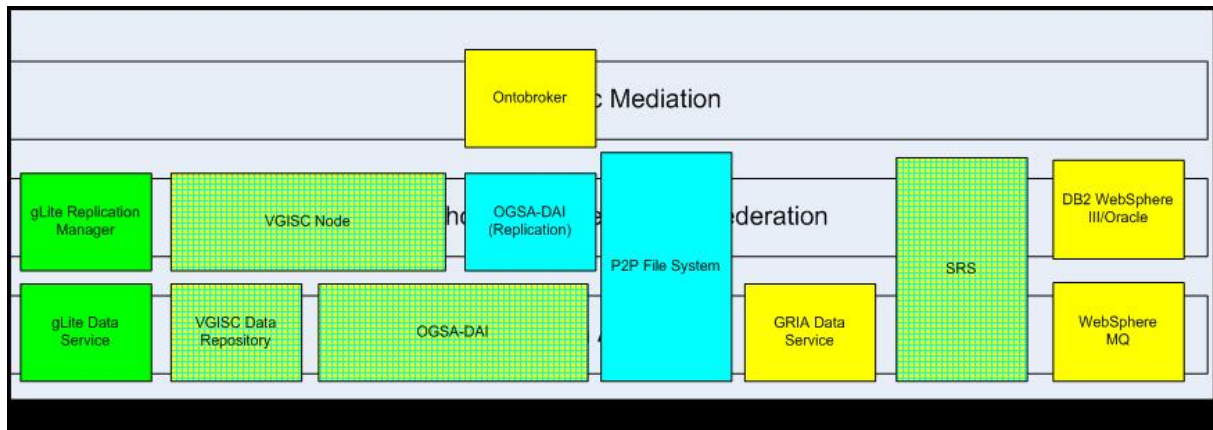
At the PM12 and PM18 respectively, a number of data access interfaces were used in the project:

- OGSA–DAI Web Services interface (WS–I, WSRF)
- GRIA data interface (WS–I)
- SRS interface (native, ad–hoc WS encapsulation)
- WebSphere MQ and Information Integrator (WS–I)
- VGISC interfaces (OGSA-DAI OGSi with specific activities for access handling and archive requesting)

Additionally the project partners continuously evaluated and developed new components over the span of PM12 to PM30:

- LHC Grid data services (based on gLite)
- IGOR file system (proprietary P2P technology)
- OGSA–DAI extended access control for DB access

This profusion of DDRA interfaces clearly presents a problem. Scenarios with similar architecture and requirements should use a common interface. The following indicates the plans for consolidation up to PM30; existing components are pictured in yellow, existing components to be significantly enhanced till PM30 are colored light green, components to be developed until PM30 in blue and finally external components to be evaluated till PM30 in green.



The WebSphere MQ and Information Integrator (WS–I) focus on data access within one organization, and will thus be “deprecated”. To further consolidate the interfaces, WP3 partners propose to use the OGSA–DAI interface for most data access components, and extend it where necessary. OGSA-DAI supports a variety of file transfer mechanisms including GridFTP, HTTP and URL references. The requirements of data replication and synchronization as OGSA–DAI activities will have to be considered as well if required by the partners.

OGSA-DAI being the main component of the DDRA services in SIMDAT, there are some gaps that need to be closed:

- Regarding the security model there is the need to have fine–grained database access control. The current model is too coarse. The project will cooperate with the

OGSA-DAI team on this problem amongst other issues. As well the partners will evaluate the integration of an End-to-End security model here.

- On the topic of replication and synchronization SIMDAT will develop components here which eventually might contribute to OGSA-DAI. Here the project will especially consider the mesh network / synchronization component built in / for the Meteo activity which will have a WSRF and potentially an OGSA-DAI interface (see below).
- Regarding the caching and/or replication of data partners will evaluate LHC DataGrid approach (and cooperate with EGEE project), evaluate P2P file system techniques (Pharma/Univ. Karlsruhe) and of course continue to closely cooperate with OGSA-DAI team (EPCC).
- Finally on the issues of robustness and performance, optimizations in OGSA-DAI code base are looked for. Some work will hopefully be done in SIMDAT, and contribute to OGSA-DAI. As well the project will also look at “low-level” optimizations / optimized managed runtime environments

A special component that will be built is the synchronization / replication engine for metadata and data which relies on the mesh network component for the Meteo activity. It will use a WSRF interface (potentially OGSA-DAI) is going to be an innovative component as it brings specially derived peer-to-peer technology in the Grid world.

Specific issues that will need to be tackled are:

- The interfacing of OGSA-DAI and semantic mediation
- Security and access rights management
- The integration between data access and job execution
- Role of DDRA/OGSA-DAI in new Pharma B2B showcase

To be a little bit more detailed, the concrete problem posed by the need for security and access rights management is that SIMDAT requirements ask for fine grained access control to data stored in relational, OO databases and flat files based on the credentials of the user invoking the access service. The OGSA-DAI access control currently gives users access to the whole database or directory structure. There is no finer-grain access control (rows, columns, fields) provided. The first proposed solution is to use map from the credentials presented to the access service to db access rights using ACL. The second option is to rely on the DB built in access control mechanisms; means to have a DB user defined for each grid user/VO.

It has to be investigated if the partners can rely on transport-level security for data transport or if they need complete end-to-end security including message-level security and encryption. Here is a clear relation to the VO activity and its results. Each alternative will have to be evaluated.

7 Conclusion

The work on DDRA in the first SIMDAT period has focused on providing distributed read and initial update capability, and basic support for distributed queries. Also, both relational and/or XML databases and regular file systems are supported. Initial provision for attached Metadata to records has been included, and the access-control and authorization mechanisms of the current basic Grid infrastructure (based on GRIA) are fully enabled.

Due to time pressure for completing the PM12 prototypes, most of them actually used a simpler way of accessing remote databases. After the PM12 milestone, this work package cooperated with the application activities to define the best use of the DDRA services based on OGSA-DAI, and to leverage the capabilities of the DDRA services in their applications/demonstrators. Architecturally, the simplified scheme described in section 6 fits in very well with OGSA-DAI, since that layer can use SQL embedded in XML documents to define the desired queries/data access operations, and provides Web Service compliant ways to return the result data.

From the experiences up to PM12, it has become clear that information security in industrial scenarios needs to be given prime attention. The academic model of federating databases at the level of SQL queries raises concerns about protecting data from unauthorized access: each partner is only allowed access to certain parts of a model, which are stored in a highly complex subset of rows/columns. To validate the access rights at the level of raw SQL queries is difficult, and the OGSA-DAI components have to be trusted to not open any backdoors to protected information. One possible solution is to federate at a higher level (parts of a model), and the DDRA work package will work with the application activities (in particular, the automotive sector) to define a workable solution and include the necessary support into the DDRA services.

Regarding functionality of the DDRA services the focal points currently are:

- Provide support for the emerging requirements of the “higher-level” technology activities, in particular Workflow (for distributed queries), Ontologies and Knowledge Services.
- Extend the functional envelope to accommodate effective metadata management, and support distributed coherence protocols (driven by the Meteorology activity).
- Investigate and remedy performance bottlenecks within the DDRA services implementation, and generally increase the reliability and robustness.
- Cooperating with the VO technical activity and the application activity, to accommodate the requirements on authentication and authorization, security and trust management within the DDRA services. An important part of this is to support industrial-quality mechanisms for authentication and authorization, and to produce audit trails for post-mortem analysis.
- Compliance to and interoperability with emerging standards, such as in the data area of the GGF (DAIS, OGSA-D, GridFTP, OREP, others ...).

Of course, this WP will continue to liaise very closely with WP2 to ensure seamless integration with future versions of the Basic Grid Infrastructure, and to leverage the capabilities of this infrastructure. Conversely, feedback on the basic Grid infrastructure

releases will be given, and input/requirements for the ongoing architecture and development work will be given.