



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

***D.10.2.2 Documentation of advanced implementation of
SIMDAT Pharma Prototypes for Interoperability Phase
including evaluation of underlying technologies***

Start date of project: 1 September 2004

Duration: 48 months

Due date of deliverable: March 01, 2007

Actual submission date: 23. April 2007

Lead contractor for this deliverable: NEC Europe Ltd.

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
29.01.	0.1	Falk Zimmermann	Template
16.03	0.1.1	Falk Zimmermann	Revised Template
22.03.	0.2	Kai Kumpf	Progress on TUAM developments
30.03	0.3	Changtao Qu	Progress on Semantic Broker
03.04	0.4	Richard Kamuzinzi and Joseph Mavor	Progress on Federated Prototype
05.04	0.5	Richard Kamuzinzi	Minor changes
10.04	0.6	Falk Zimmermann	Executive Summary
11.04	0.7	Falk Zimmermann	Introduction and Conclusions
11.04	0.8	Thomas Fuhrmann	Progress on IGOR File system
20.04	0.9	Falk Zimmermann	Incorporated requested changes from internal reviewers
23.04	1.0	Falk Zimmermann	Incorporated final comments from internal reviewers

Copyright

This report is property of the partners of the SIMDAT consortium 2007. Its duplication is restricted to the personal use within the consortium, funding agency and project reviewers

Copyright © NEC Europe Ltd and other members of the SIMDAT consortium,
www.simdat.eu, 2007

Table of contents

1	Executive Summary	4
2	Introduction	4
3	Technology Developments.....	5
3.1	Semantic Service Registry	5
3.2	Service Annotation	6
3.3	IGOR File System	7
3.3.1	Security and Reliability Mechanisms.....	7
3.3.2	Progress on the overlay network IGOR	8
3.3.3	Interaction with Pharma partners	8
3.3.4	Practical issues in the development process.....	8
3.3.5	Theoretical Foundations for Future Developments.....	8
3.3.6	Integration of IGOR File System in Pharma Prototype	8
3.3.7	General Requirements	9
4	Advanced Business to Business Prototype.....	9
4.1	Resulting Requirements	10
5	Advanced Federated Prototype	13
5.1	Evaluation of Technology	14
5.2	Resulting Requirements	16
6	Final Prototypes.....	17
6.1	Overview	17
6.2	Resulting Requirements	20
7	Conclusion.....	24
8	References	24

1 Executive Summary

The work during the reporting period focuses on the uptake of the latest software releases in the areas of integrated Grid technology, Ontologies and distributed data access within the Pharma prototypes. Based on the prototype architectures that have been described in the last Pharma deliverable [5] the Pharma partners evaluated the underlying technologies with respect to constraints of robustness and usability and analyzed the potential of the semantic framework along the lines of the knowledge model. The results of these evaluations will be fed back to the corresponding technology layers in terms of modified requirements with the expectation to receive adequate solutions in time for the implementation of the final prototype.

Concerning the technological developments which are defined in the Pharma work-package, significant progress has been achieved. The TUAM tool (cf. section 3.2) is now able to deal with GRIA application lists and is using NEC's E2E framework for securing the communication paths and, most importantly, it proved to interact with the latest version of the Semantic Broker. The Semantic Broker (cf. section 2.1) is now based upon extended domain ontology and provides a variety of new features. In particular, the Semantic Broker has been completely integrated into the GRIA5 middleware and represents GRIA's PBAC access control policies via semantic rules and ontologies. The maturity and performance of the IGOR file system (cf. section 3.3) was greatly enhanced and encryption as well as versioning was achieved.

The architectural design of the final Pharma prototype (cf. section 6) is still in line with the projection we have included in [5]. Regarding the test-case which will be demonstrated in the final installation of the integrated Pharma prototype, the industrial partners have agreed on the so-called Master Sequence Analysis Pipeline. This pharmaceutical workflow will encompass compute resources and data spread across different industrial and academic sites, each operating under different security and QoS regulations.

Current work on the prototypes revealed the necessity for regular updates of the underlying technologies. Apart from improvements in maturity and performance of the software packages, extensions and new functionalities are the main drivers for this activity since certain essential demands arising in real scenarios cannot be adequately addressed otherwise. Thus we strongly believe that SIMDAT's incremental development approach was well chosen, predominantly due to the continued dialog between the application and technology layers on the basis of qualified requirements resulting from limitations of the current implementations to be overcome in the next release.

2 Introduction

In this deliverable we outline the ongoing technical developments of the Pharma partners in areas of Ontology (Semantic Broker, TUAM) and Distributed Repository Data Access (IGOR-File System). We also report on the recent advances of developments of the Federated and the Business-to-Business prototype and briefly summarise the architectural design of the final Pharma prototype. Connected to this is a corresponding list of resulting requirements that should be addressed by the technology partner in charge during the next project phase. Of particular interest are the paragraphs on the evaluation of underlying technology which have been integrated into the discussion of the Federated Prototype.

We do not discuss or introduce activities that have been started with the knowledge service group since we agreed to integrate this contribution into their current deliverable (cf. D8.2.2) and not duplicate any text across deliverables as has been recommended by the reviewers.

3 Technology Developments

3.1 Semantic Service Registry

During the reporting period several major technology improvements have been made to the Semantic Broker (SB).

1. *SB is integrated with the new SIMDAT Pharma domain ontology and service annotation format.* Based on the new domain ontology and service annotation developed by SCAI Bio, we integrate several important service features, in particular some dynamic service features such as databank release date, databank version, and databank entries into the matchmaking algorithm. In a Grid computing environment like the SIMDAT Pharma Grid where service instances are typically duplicated to obtain a reasonable degree of redundancy, the dynamic features are usually of major interests for biologists during the service selection phase. Thus, including dynamic service features into the service annotation/discovery model can enable biologists to better refine their requirements in order to find optimal services.
2. *Advanced ranking functionality is newly implemented in SB.* Concerning a new user requirements from ULB, an advanced ranking functionality has been newly implemented in the SB to support user customized result ranking. Besides semantic matching degree as the primary ranking criterion users can additionally choose any preferable dynamic service features, e.g., databank release date, as the secondary ranking criteria. If the matching degrees of service instances are equivalent, these services will be subject to an additional ranking procedure ensuring optimal services always go first in the ranking list. Such an advanced ranking functionality makes special sense for the workflow adviser to automatically select service instances in terms of user preferences in addition to functional requirements.
3. *SB is seamlessly integrated with the SIMDAT annotation toolkits, i.e., TUAM and Dynamo.* The SB knowledge base is partly re-factored to better accommodate modularized service annotations. The service annotations generated by TUAM can thus directly be stored in the knowledge base after verification, and sequentially used by SB and Dynamo. A new SB operation *retrieveAnnotation* is newly developed in order to coordinate asynchronous manipulation on service annotations through SB, TUAM, and Dynamo.

Besides these two SIMDAT specific annotation toolkits, we also explore SB's interoperability with other OWL-S annotation tools, in particular the open source OWL-S editor. Based on the OWL-S standard, we encountered no problem to make directly use of annotations generated by OWL-S editor in SB.

4. *SWRL rules are tentatively introduced into the knowledge base, which are in the first instance used for annotating Access Control (AC) policies of services.* As most of SIMDAT services are deployed as GRIA services and further leverage GRIA PBAC to secure their access, we explore the possibility to represent PBAC AC policies in SB using semantic rules and ontologies. In general, PBAC AC policies can logically be separated into two parts: the "static" part representing the role-permission relationships; and "dynamic" part represents the user-role relationships. For annotating the static part of PBAC AC policies, we develop sets of security ontologies to model entities in the role-permission relationships. For annotating dynamic part of PBAC AC policies, we adopt semantic rules to dynamically infer user-role relationships on the fly based on users' X.509 certificates or SAML attributes. As SWRL provides a natural way for combining OWL and semantic rules, it is used in SB as the semantic rule language to represent dynamic part of PBAC AC policies. The investigation of semantic rules in SB is still at an early stage, and more results are expected to be reported in the next major Pharma deliverable.

-
5. *Pellet[3] ontology reasoner is introduced to support the reasoning over OWL-DL and SWRL rules.* As OWL-DLP may not be enough to represent a complex biological domain [2], we introduce Pellet as another ontology reasoner into SB with the purpose to support full set of OWL DL reasoning and hence addressing more complex requirements on SB. As Pellet uses a different API set the Pellet reasoning part has been newly implemented in the SB. As all ontology reasoners share a common interface in SB (c.f. Deliverable 10.2.1), the semantic matchmaking part does not need new implementation for Pellet. Also all ontology reasoners can smoothly be switched between each other at run time.

As Pellet can also support the reasoning over part of SWRL rules, i.e., DL safe rules as defined in [1], it is also used as the semantic rule engine for the AC policy based “controlled discovery” in SB. “Controlled” here means that SB is able to discover service instances based on pre-annotated AC policies, thus only those services where users have access permission will be returned in the result set. In order to implement “controlled discovery”, the semantic matchmaking algorithm is extended to include AC policy based matchmaking, using Pellet as the semantic rule reasoner. Though our first experience reveal some performance issues with rule reasoning, we recognize that SWRL rules are essential if dynamics come into play. Dynamic parts of policies cannot be sufficiently addressed by ontologies themselves and therefore we will continue to explore their possibilities in the SB framework.

6. *Improved SB usability:* The client side API of SB is again re-designed to reflect changes on both the domain ontology and service annotation model. An SB wrapper class has been newly developed to facilitate the invocation of SB at the client side. Additionally, the integration of SB into GRIA5 has been initiated. A secured access to SB will be implemented based on GRIA PBAC in the next project phase.

In summary, our initial evaluation of semantic rules indicates that the task on “Evaluation of the usage of ontology rules in the matchmaking process” has to be continued in the next project phase since the potential of semantic rules for the bioinformatics service discovery is not yet fully revealed.

3.2 Service Annotation

FhG-SCAI’s annotation tool TUAM[1] was adapted so that it now reads GRIA 4.x application lists from GRIA servers through an NEC’s E2E-secured communication line and displays them in tree form. The application list contains information from the application wrappers, in particular the plain name of the application, a short description, and two lists of data types for input and output, respectively. Since those qualifiers are invariably hand-written, they usually contain no controlled vocabulary and so inferences as to the type of tool can only be achieved by human users.

As before, TUAM also reads the domain ontology in RDF/OWL and displays it in graph form, but now OWL class instances, which should be used for annotation of applications, are displayed as stars. A few other minor enhancements to the user interface concerning usability were also implemented. Annotations are produced by TUAM first as regular TUAM mappings in a specialised subject-predicate-object format, which in the SIMDAT Pharma case will contain

```
<GRIA_application_name>  
    <is-a>  
<OWL_domain_ontology_class_instance>
```

e.g.

```
http://utility.scaibit.de:8080/our_blastx_implementation  
    is-a  
http://www.scai.fhg.de/bioinformatics/simdat_pharma_ontology#blastx
```

On demand, those are transformed into OWL-S containing the same logical structure as the OWL-S annotations produced by Protege's OWL-S editor *plugin*. Special adaptations had to be made to account for the fact that GRIA services do not come with individual WSDL descriptions and thus the OWL-S "grounding part" was filled in non-standard form using several dummy values (that are currently also ignored by the Semantic Broker).

The generated OWL-S can be uploaded to the Semantic Broker directly from TUAM on mouse-click through NEC's SB publication interface that expects one annotation per provider, in string form. The necessary public certificates have been generated on both sides and distributed as needed. Special provisions had to be made to accommodate the requirements for time-variable instance attributes in the annotations, i.e. the fields that should hold information about database versions, their no. of entries, etc. for use with Blast databases. While this is strictly out of scope of the domain ontology that at the core only captures time-invariant and general properties, ULB's DYNAMO filter is thought to take care of modifying individual annotations. It will do so by first importing one annotation out of SB, replace a reference to one database instance with reference to another, newly generated instance and write that one along with the just mentioned updated fields back to the SB via the same publication interface as the one TUAM uses.

Logically, all necessary further modifications should only have to take place inside the domain ontology, because the OWL-S only references service class instances from there and all the relevant descriptions and relations should be taken from there for semantic query resolution.

3.3 IGOR File System

The development of the IGOR file system has significantly progressed over the reporting period. Primarily, the overall system stability and performance has been improved. This is partly due to general bug fixes and code improvements. Many of these improvements have been facilitated by the valuable input from SIMDAT partners. Other parts of the performance gains are due to advances in the peer-to-peer overlay network that forms the basis of IGOR FS. Besides these straightforward issues in the IGOR FS core, many practical side topics have been discussed with the SIMDAT partners and third parties. We believe that the respective ideas and results are important for the future deployment of IGOR FS. Finally, all this practical work was supported by further theoretical work which we expect to form the basis for future progress in the development of the IGOR file system.

In the following sections we briefly discuss the technological core issues of the reporting period.

3.3.1 Security and Reliability Mechanisms

Encryption, versioning and the Grid interface are premier building blocks for the secure and reliable use of IGOR FS.

The encryption stage of IGOR FS encrypts all user data blocks. It operates immediately after the block cutting stage and right before any network activity. Thus an adversary would have to attack the client machine directly to obtain illegitimate access to the data. All data in the network is fully protected by encryption.

IGOR FS encryption uses state of the art cryptographic primitives. The algorithms are used in a way that guarantees security and safety as designed, while still preserving the efficiency of the system. For example, it is possible for arbitrary untrusted nodes to cache encrypted data without any way for these nodes to gain knowledge about the cached data. Furthermore, all data items are kept in encrypted form during network transit and permanent and temporary storage. Decryption happens only if the data is required and the use of the data is authorized.

The versioning system is closely related to the encryption and hashing subsystem. It computes a unique identifier for each revision of the file system. An IGOR FS instance can retrieve the state of the file system at any time for which such an identifier has been computed. The storage system as well as the data transport system efficiently detect similarities between different revisions and eliminate the redundancies. This property is important for large data sets that change often but the changes are relatively small compared to the whole data set.

The interface to the Grid component has been designed as means to integrate IGOR FS into the business side of the SIMDAT system. It handles encryption and versioning keys of IGOR FS and

provides them to customers who want to integrate the data into their workflow. It records the state or revision of the file system at any desired point in time. This process yields an identifier and decryption key. Both can be transported via the Grid system. For this transportation, the Grid system guarantees for confidentiality, authenticity and is responsible for authentication, audit and accounting.

3.3.2 Progress on the overlay network IGOR

The overlay network IGOR is an essential part for a working IGOR file system. Its implementation has been revamped and the code is much more modular now. This directly leads into an easier development for the features mentioned below.

Integration proximity neighbour selection and proximity route selection has been designed completely and is currently under implementation. Once these features are fully working, IGOR will be able to adapt to the Internet much better. This will directly lead to further performance improvements for IGOR FS.

Furthermore, the service oriented concept for IGOR has been designed, too. Meanwhile it is already implemented to a great extend. Service orientation will allow other services besides IGOR FS to co-exist with each other.

3.3.3 Interaction with Pharma partners

The interactions both with the SIMDAT partners and third parties have provided valuable insight for the development of IGOR FS. Additionally to the progress in the Pharma activity, new use cases from non-Pharma activities have been taken into consideration, too.

The originally envisaged extensions towards multiple write access instances have been postponed in accordance with the SIMDAT partners because performance and stability has been ranked as more important. As a result of this shift of focus, significant progress has been made stability and performance-wise.

Workflow extensions have been discussed with EMBL but have also been postponed until EMBL can commence its participation in SIMDAT.

3.3.4 Practical issues in the development process

The already running system of regular software tests has been extended by load tests. These load tests show only rare malfunctions which could typically be resolved soon after being reported. Benchmarks show that the performance is now within an order of magnitude of the performance of that of a local file system.

In the design goal use case, IGOR achieves order-of-magnitude better results than state-of-the-art data distribution in the pharmaceutical industry. Further benchmark results are expected once the collaboration with EMBL begins.

In order to keep track of the issues that arose during testing with EMBL and ULB, an issue tracking system has been established.

3.3.5 Theoretical Foundations for Future Developments

Finally, the so called "push-mechanism" has been discussed with EMBL. Such a mechanism will allow subsequent data exchange between users without requiring further key exchanges once a successful transfer has been established. This push mechanism will require further cryptographic primitives which are not yet implemented. Furthermore, there are some usability aspects to consider because the frequency of such updates without key exchange is important. The issue of multiple writers was also pursued, but as mentioned above, with lower priority. In the current state, multiple writers can use the file system at once, as long as they write in independent parts of the file system.

The redundancy avoidance system (see above) works as long as multiple writers save (at least in parts) the same content to IGOR FS. Note that the number of readers is not limited.

3.3.6 Integration of IGOR File System in Pharma Prototype

The IGOR file system provides the data distribution system underlying the analysis tools. An analysis tool can access the required databases as if they had just been downloaded from their original source. In particular, the analysis tool does not need to download data in advance.

The IGOR file system has two major interfaces to other components in the SIMDAT Pharma prototype.

1. The data access of the analysis tools such as the MRS system uses the POSIX interface. This interface directly connects the analysis tool to IGOR FS via the Linux kernel. It covers all the standardized POSIX file operations and thereby guarantees full interoperability, not only between IGOR FS and MRS but to all analysis tools used in the Linux environment. This interface has been completely implemented in the reporting period. Tests at ULB and EMBL have demonstrated its principal fitness for the purpose. However, the tests have also revealed bugs which are now being fixed.
2. The authorization and authentication credentials interface grants or blocks access to the data stored in the IGOR file system. Originally, NEC's DAC component was foreseen as interface to IGOR FS. Meanwhile the focus of the DAC component has been changed and GRIA's functionality has been taken over.

3.3.7 General Requirements

In order to uptake the IGOR file system the following requirements have to be met. Obviously, they are not imposing any sensible restriction on a particular installation, since one of the major design goals of IGOR file system has been compliance to off-the-shelf Linux distributions:

- Linux operating system; (first tests on MAC OS X have been conducted.)
- One arbitrary open TCP network port for the interconnection of the IGOR daemons.
- Fuse module installed. Note that FUSE is included in all major Linux distributions.
- OpenSSL libraries for cryptographic operations; also included in all major Linux distributions.
- CppUnit libraries; only required for testing.

4 Advanced Business to Business Prototype

Within the reporting period we have been concentrating on the construction of the final proof of concept model with InforSense and GSK which was successfully demonstrated in November at the Second Annual review. This technical proof of concept model highlighted three important issues:

1. GRIA v4 has some significant limitations within a business context.
2. Galapagos's current GRIA server is not suitable for a fully functional prototype system.
3. For its size and complexity bioclip application has some significant technical overheads and additionally its target audience is limited.

One of the central limitations of the GRIA v4 is clearly the way user authentication works. Before a user at the client site (GSK in the B2B scenario) can access an individual service at the supplier site (Galapagos) the account owner must upload the corresponding user certificates. Thus the actual authentication decision will be done at the supplier site. In contrast, the token model of GRIA v5 allows the client management to provide arbitrary users with a signed token that is sufficient for authentication at the remote site. Thus a greater flexibility is achieved. In addition there is no SLA management service available, which is of significant importance in industrial settings. Both functionalities are covered by the current GRIA v5 so it was decided to upgrade to this middleware version. At the same time Galapagos upgraded their GRIA system to a server grade dual CPU system. The merger of Inpharmatica and Galapagos introduced modification of the underlying operational

security model which also had a great impact on GRIA's security configuration. Up to February the required adaptations were partially complete.

The bioclip application only has a limited audience and requires substantial number of servers and database services to be available. In order to address these limitations Galapagos and GSK designed and agreed on an improved service which will enhance our ability to integrate our systems more effectively. Additional Galapagos staff have become involved SIMDAT project and briefed on the concepts GRIA based services. They were able to contribute to development and design of these new services. The service has been designed from a scientific/business perspective thus producing the concept of a product that can be seen by interested parties as an exciting offering rather than the technical bioclip approach which produced a product looking for a market. This prototype has been designed to include both chemical and biological information thus penning the SIMDAT concept to a wider audience.

The emphasis of bioclip on large databases and few servers has been redesigned. Our new application will use a smaller database with a set of redundant farm nodes to undertake the processing. This will mean that the development of the prototype will not be reliant on one single system therefore providing a greater chance of success. Overall the amount of hardware will remain the same but the risk of failure should be greatly reduced. A high level design of this application has been completed and the standard for communication between GSK and Galapagos has also be produced and reviewed. The proof of concept for the new prototype application has been tested.

4.1 Resulting Requirements

The following list of requirements is modifications of requirements that have been already reported in [5].

Name	<i>B2B Security of Customer Intellectual Property (IP)</i>		
Business Requirements	To be able to demonstrate that security of data and access is such that IP is not put at risk.		
Date created	2006-03-22	Source	Galapagos/GSK
First Implementation	2006-08-31	Priority	Medium
Application Activity	Pharma	Related Prototype	B2B
Technology Component	Integrated Grid Infrastructure	SIMDAT Modules targeted	GRIA, DAC
Detailed description of the requirement			
The client should be able to confidently and seamlessly connect to a service, provided by a trusted organisation, over a secure external link that ensures data integrity end-to-end. This will be achieved by using a simple trust model based on a common CA (Certificate Authority) shared and trusted by all system participants. A Single Sign-On model will ensure that the process is seamless for the user. This functionality will be provided by utilising the components available in GRIA and DAC			
Relation to the prototype			
A core component of the B2B scenario. Confidence in this component is crucial to the success of the Pharma B2B prototype.			
Requested functionality			
Once the client has signed up for the service and has an agreement in-place then the process of authorisation, accounting and the secure exchange of inputs and results will be seamless from the user's perspective.			
Validation of the requirement			
Testing using invalid or unsigned certificates which should be rejected. Correctly assigned keys should be tested. Additional testing to be agreed and fully defined during the prototype.			

Name	<i>B2B Service Provision (Supplier)</i>		
Business Requirements	To accept service requests from a range of customers with low overhead and low lead time.		
Date created	2006-03-22	Source	Galapagos/GSK
First Implementation	2006-08-31	Priority	Medium
Application Activity	Pharma	Related Prototype	B2B
Technology Component	Integrated Grid Infrastructure, VO	SIMDAT Modules targeted	GRIA, DAC
Detailed description of the requirement			
<p>The Supplier will define one or more Analysis/Annotation services to be provided to Customers. Having published the service, the Supplier will be able to contract with one or more Customers to supply the service via the GRIA components. A sample request will be made by the Customer and the Supplier will deliver results using the authentication and associated security components to be used in the Service. In the proof of concept, the Service will consist of the analysis and annotation of a set of Protein Sequences. The sequences may be identified by a public accession number (e.g. Swissprot ID) or by the Protein Sequence. The resulting annotation will be delivered in the form of XML.</p>			
Relation to the prototype			
<p>The purpose of the prototype is to show a real Pharma B2B scenario. This component is the first step in the process so that subsequent requests for the Service can be executed and additional services offered by the supplier.</p>			
Requested functionality			
<p>The functionality in this requirement is a simple request and delivery. It will include the initial steps of Authentication, Authorisation and access control and Accounting</p>			
Validation of the requirement			
<p>The validation of the requirement will be the receipt of the expected analysis and annotation by the Customer. During the prototype stage the customer will confirm by e-mail to the Supplier that the expected results have been received and that the Service can be considered to be ready for use.</p>			

Name	<i>B2B application and technology reuse by Supplier</i>		
Business Requirements	To reuse applications and technology on behalf of customers via a) multiple services, and b) provision of the same service to multiple customers.		
Date created	2006-03-22	Source	Galapagos/GSK
First Implementation	2006-08-31	Priority	Medium
Application Activity	Pharma	Related Prototype	B2B
Technology Component	Integrated Grid Infrastructure, VO, Analysis services	SIMDAT Modules targeted	GRIA, KDE
Detailed description of the requirement			
<p>The Supplier will define one or more Analysis/Annotation services to be provided to Customers. Having published the service, the Supplier will be able to contract with one or more Customers to supply the service via the GRIA components. The service must be capable of being used as is or adapted with negligible effort to be suitable for use by multiple clients. Accounting processes implemented must be able to identify and authorise appropriate access for multiple clients</p>			
Relation to the prototype			
<p>The purpose of the prototype is to show a real Pharma B2B scenario. This component</p>			

demonstrates reuse, shared use of components and appropriate controls over access to services.
Requested functionality
Provision of multiple identical and/or differentiated services to multiple or single clients with controls on access dependent on user.
Validation of the requirement
Customers will be able to select from multiple services. Services for which they have contracted will be accessible, those they haven't will not. The Supplier will be able to track usage and ensure this matches contracted usage.

Name	<i>B2B Security of Supplier IP</i>		
Business Requirements	To be able to guarantee the integrity of data and resulting analysis for each customer to ensure their IP is not put at risk.		
Date created	2006-03-22	Source	Galapagos/GSK
First Implementation	2006-08-31	Priority	Medium
Application Activity	Pharma	Related Prototype	B2B
Technology Component	Integrated Grid Infrastructure	SIMDAT Modules targeted	GRIA, DAC
Detailed description of the requirement			
The supplier should be able to confidently publish a service that allows trusted organisations with prior agreements in place to seamlessly connect to the service over a secure external link that ensures data integrity end-to-end. This will be achieved by using a simple trust model based on a common CA (Certificate Authority) shared and trusted by all system participants. Once the client has been authorised and accounting perform then the required process will be called. The supplier should be able to trace and account for all connections to this service. This functionality will be provided by utilising the components available in GRIA and DAC.			
Relation to the prototype			
A core component of the B2B scenario. Confidence in this component is crucial to the success of the Pharma B2B prototype.			
Requested functionality			
The suppliers service should seamless accept client's connection to the service they've signed up for. The process of authorisation, accounting and the secure exchange of inputs and results will be seamless from the user's perspective.			
Validation of the requirement			
Testing using invalid or unsigned which should be rejected. Correctly assigned keys should be tested. Additional testing to be agreed.			

5 Advanced Federated Prototype

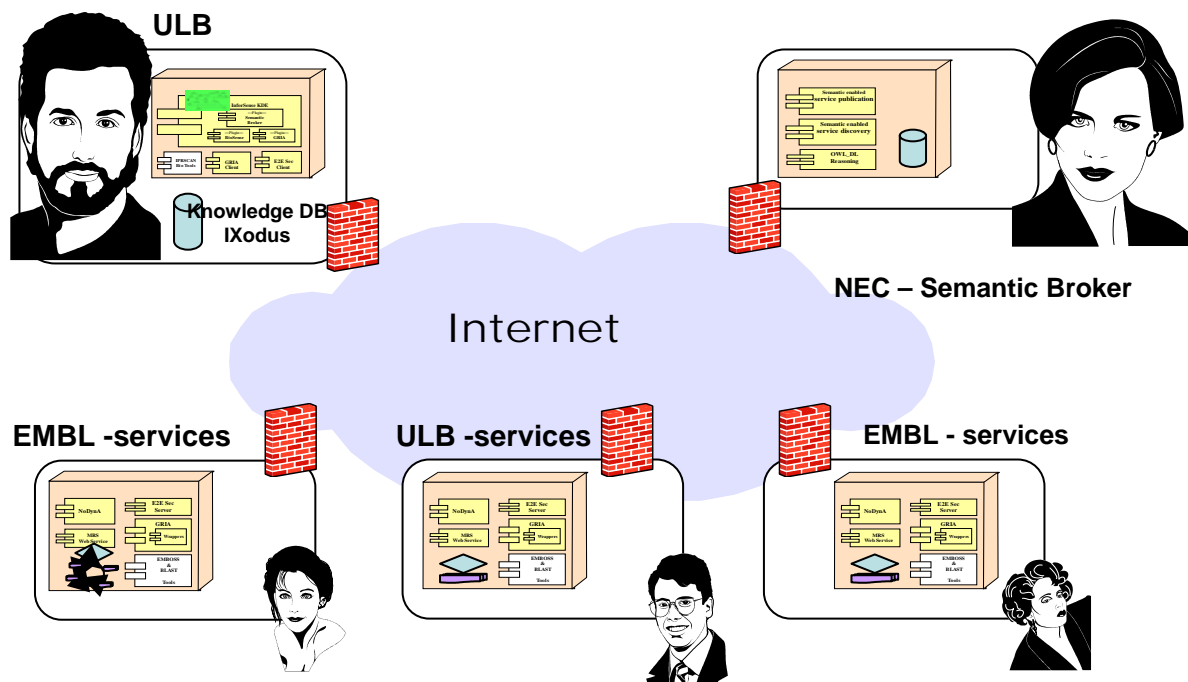


Figure 2. Overview of the advanced Federated Prototype

Last developments around the Federated Prototype activity have led to an advanced scenario of cooperative e-Science. The prototype is driven by the IXodus process, an *in silico* experiment which requires to operate in a GRID environment for many reasons such as

- to benefit from the diverse applications deployed
- to benefit from the resource redundancy for 24/7 operational requirement
- to benefit from GRID infrastructures capabilities to manage security and service provider's relationships

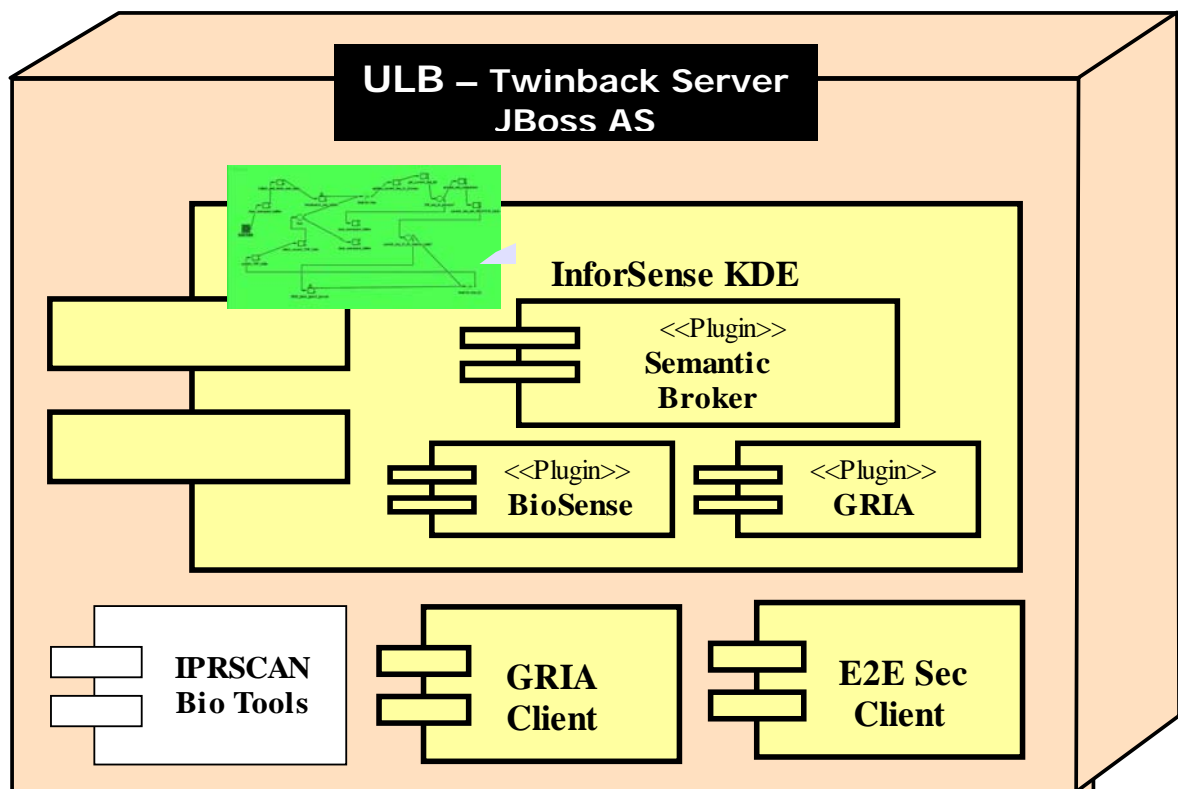


Figure 2. The PM24 deployment diagram at the ULB site integrating the IXodus orchestration

To improve the implementation of the IXodus workflow, we have migrated from a hard-coded version to version using the InforSense's KDE workflow system. The workflow is leveraging the SIMDAT developments in the KDE and thus accesses the various GRIA wrapped applications through an End-to-End Security framework provided by NEC. The discovery process of the deployed applications (BlastX, BlastN, InterProscan and Getorf) is fulfilled by the KDE enactment engine thanks to the semantic broker API which enables to query for specific service provider. Besides the analysis services deployed in the testbed, a databank management system, namely MRS has also been integrated in the architecture to support the knowledge discovery aspect of the IXodus workflow.

To support an operational semantics-enabled service discovery in the frame of SIMDAT-Pharma federated scenario, we need a tool to update the information stored in the semantic broker. This task is supplied by the Dynamic annotator module (Dynamo), new acronym of the former Nodyna, which consults with both the MRS system and the GRIA services. The MRS system is used to gather the databanks status on a given service provider where the GRIA infrastructure is accessed to discover the applications currently available.

5.1 Evaluation of Technology

Although the current Federated prototype approach proves to be functional and of significant relevance in a knowledge discovery process such as IXodus, we can point out some limitations that should be addressed for the final prototype implementation. The current limitations cover the following evaluation aspects: the robustness, the usability and the knowledge model.

The robustness limitation

Whenever a workflow is executed, it is the responsibility of the enactment engine to ground each of the "abstract" analysis steps of a process to the most accurate server. This implies that the enactor should follow a methodology we defined as "the node selection mechanism". The current implemented mechanism is error prone as, for instance, it does not consider the user account status on

the remote servers. As an example, let's consider a user A who has exceeded his/her resources limits on the server B, if the enactment engine routes any particular analysis step of a workflow launched by the user A, it will fail.

We would like the workflow enactor to ground any GRIA applications specified in an analysis process to real services according to an advanced GRID service selection mechanism. As a first approach, we could suggest the enactment system to perform the following actions:

1. To query the semantic broker to discover services associated to specified application ID and capable to operate on a specified databank. Example: "Which are the freshest BLASTN service coupled to the up-to-date EMBL databank",
2. To retain, among the returned list of services, only services where the user has been registered (GRIA accounts),
3. To retain, among the returned list of services, the services that present the cheapest execution cost (GRIA cost model),
4. To select, in any ex-aequo case, the first alive (heart-beat) service in the remaining list.

The usability limitation

The Federated prototype showed also some usability limitations from the workflow designer perspective as well as from the end user perspective.

During the workflow design process, the bioinformatician should be able to easily access all the arguments of a GRIA-enabled application as it is already the case for any other KDE node. To launch an application using the KDE system, the workflow designer deals for the moment not only with the design workbench but also with a GRIA work.xml file to specify the arguments. To ease the use of applications node at the design phase we need the KDE to discover application descriptions including the required arguments. These descriptions could be provided by the GRIA infrastructure as a part of the application metadata.

To ease the exploitation of the workflow from end user perspective, the IXodus front end is entirely based on the KDE Portal system. The system provides a means to present a designed workflow to the end user as any other application. The IXodus workflow is deployed as web interface in the KDE Portal environment and exposes an input area where the user enters sequences to process. For some reasons, this approach of workflow deployment proves not flexible enough:

- Not all desired arguments are exposed at workflow level: for instance, we did not succeed to present some relevant arguments of BLAST tools like the E-value
- The link between the workflow execution and the workflow result is not easy to establish
- The construction of the pages for the user is not intuitive enough

As a consequence, we plan to adopt an existing and more familiar Portal environment, namely wEMBOSS as future deployment approach.

The knowledge model limitation

By using the current knowledge model we foresee several future pitfalls, from the point of view of the end user. Knowledge models tend to grow in size and complexity in time. Both redundancy of data types and ambiguity of the data type are nearly certain to occur eventually. From the end users point of view can present problems when choosing the proper data type. Although several solutions exist for these types of classical problems in the domain of knowledge models, they are not sufficient to completely resolve them. For example, the Gene Ontology project is forced to rely on manual reviews of the ontology to eliminate redundancy and they have not been able to completely eliminate this problem. To solve the pitfall of data type ambiguity, one well-known project offered long descriptions

of each data type. This solution turns out to be incomplete as long data type descriptions cannot guaranty that the user will choose the correct data type.

From a technical point of view, the last SIMDAT-Pharma prototype was the last to use an ontology not formatted in OWL. As of PM30, all ontology annotations are formatted in OWL. The semantic broker now houses the ontology in OWL format. Furthermore, work is in progress by the ULB to upgrade Dynamo so that all MRS updates are annotated in OWL format as well. In that perspective, the MRS information, such as 'release date' and 'number of entries' will be annotated in OWL.

The advantage of annotating the ontology in OWL is that the data can be better structured and controlled. OWL allows data to be hierarchically structured and therefore better organized. Furthermore, the syntax and structure of any new data that an external source is trying to add to the OWL based semantic broker can be verified and controlled by software components that ensure data consistency.

Some difficulties and pitfalls can be expected on the path to upgrading all data sources to OWL-based annotation compatibility. For one, the correct syntax and structure of the OWL ontology is centralized under the auspices of the ontology designer. Until all data sources learn to understand this new ontology some efforts in communication must be increased. This difficulty is quite unique to OWL due to its complexity. The same challenge extends also to the software components that annotate data based on the ontology. When written in Java code, these OWL annotation components tend to be difficult to maintain. This problem is inherent in the fact that OWL and Java are not based on the same data structures. Indeed OWL is not exactly Object Oriented in the same way as Java and therefore the Java code needed to manipulate OWL tends to be quite obscure and disorganized relative to good programming practices.

5.2 Resulting Requirements

Name	<i>From workflow to GRIA application</i>		
Business Requirements	To be able to deploy a workflow as a GRIA-enabled application		
Date created	2007-04-02	Source	ULB
First Implementation	2007-08-31	Priority	High
Application Activity	Pharma	Related Prototype	Federated Prototype
Technology Component	Integrated Grid Infrastructure, Workflow	SIMDAT Modules targeted	GRIA, KDE
Detailed description of the requirement			
The scientist in Life Science research has long experience in accessing various bioinformatics applications through web-based PSE such as wEMBOSS. To provide a unified access to all SIMDAT-Pharma applications as GRIA-enabled application, we need to access workflows as GRIA applications. In particular, we need to define a mechanism to port a workflow authored by the InforSense's KDE tool into a GRIA-enabled application.			
Relation to the prototype			
This requirement is related to the Portal development activity			
Requested functionality			
A way to launch a KDE workflow using the command line interface.			
Validation of the requirement			
We will test the provided functionality with the IXodus workflow			

Name	<i>To expose arguments of GRIA wrapped applications in the Workflow authoring tool</i>		
Business Requirements	To be able to calibrate applications arguments at workflow design phase		
Date created	2007-04-02	Source	ULB
First Implementation	2007-08-31	Priority	Low
Application Activity	Pharma	Related Prototype	Federated Prototype
Technology Component	Integrated Grid Infrastructure, Workflow	SIMDAT Modules targeted	GRIA, KDE
Detailed description of the requirement			
When designing a workflow it is important to adjust some applications arguments. Workflow tools such as KDE from InforSense already expose arguments for non GRIA applications. There is no justification to make this difference provided that from the user perspective the GRID infrastructure should be transparent and applications always presented in the same manner. This requirement is likely imposing the GRIA infrastructure to provide extended metadata information of deployed applications. Thus, the KDE can use this information to present the applications properties in the workbench.			
Relation to the prototype			
This requirement is obviously related to the Federated Prototype			
Requested functionality			
The presentation of applications arguments in the KDE workbench a workflow design phase.			
Validation of the requirement			
The IXodus workflow will be refined by using this new functionality			

6 Final Prototypes

6.1 Overview

The goal of the final Pharma prototype is to demonstrate the usefulness of Grid technology in the area of life sciences. We will implement an industrial strength pharmaceutical workflow across a Grid test-bed comprised of both academic and industrial partners. Naturally, these different types of partners are operating their Grid node with different security and quality of services policies and obviously have different commercial interests. The prototype will implement all infrastructure elements required to establish trusted relationships between academic and industrial partners. The infrastructure will allow an academic service provider to accept outsourcing of high value applications or databases from a research group in order to give controlled and managed access to this application for other partners, both academic and industrial. The underlying model guarantees the preservation of intellectual property rights among the partners.

Regarding workflow management we will integrate scientific problems of the life science area into the SIMDAT Grid infrastructure and further evaluated its strength and weaknesses. In particular, we will demonstrate the execution of GSK's Master Sequence Analysis Pipeline(MSAP) in a GRIA5 based Pharma test-bed. Of particular interest for the owner/initiator of the workflow is the availability of remote resources and their secure access. This redundancy will be achieved by integrating equivalent servers into the test-bed.

The development of the MSAP is driven by the need to get high quality state of the art analysis for GSK genes of interest. These analyses are currently confined to systems within GSK or uploaded at cost within the intranet. We envisage that SIMDAT technology will allow companies like GSK to

broaden the scope of their analysis and be able to import the best of breed analysis from both Academia and Vendors at costs appropriate to the type of analysis. The MSAP will initially focus on decorating sequence data with annotation from sources designed to better validate sequence structure and function as well as annotation on orthology and species availability. These we believe will be perfect for Pharma partners like Galapagos and ULB.

For each individual component of the workflow the end-user has the capability to define the set of providers which are allowed to satisfy the request. We call this technique *scoped services*. Especially large Pharma companies like GSK won't access services for sensitive calculations at providers where they didn't establish a certain level of trust beforehand.

Beyond the demonstration of the MSAP workflow we aim at developing a portal on top of the SIMDAT infrastructure to allow customers to run either individual bioinformatics services or complete preconfigured workflows. Advanced users will be able to load existing workflows from a workflow repository and reconfigure them to the resolution of the new bioinformatics problems. Extensions coming from the knowledge discovery of the workflow analysis will provide the customer with fully documented and extensively decorated results helping to ease the decision making process. These results will appear as extensions to the raw workflow analysis results.

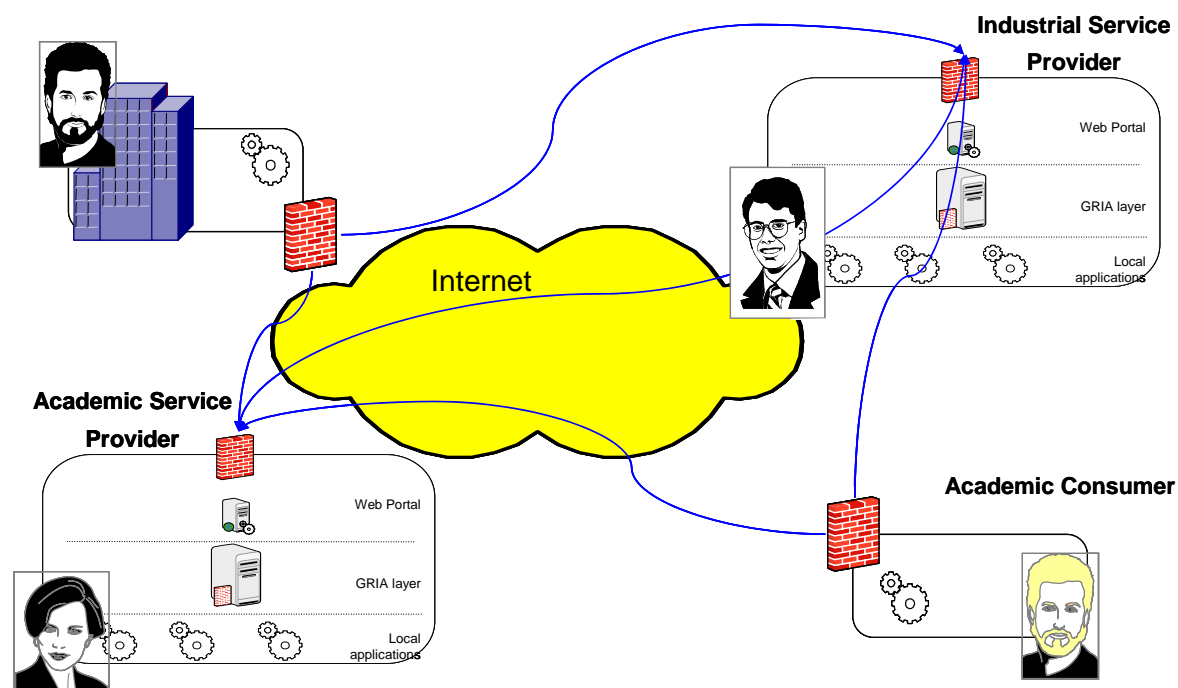


Figure 1: Business to Academia Scenario

This scenario encompasses all elements of the former M12 prototype and B2B scenarios. Services are created and hosted at various sites and are consumed by industrial and academic customers, secured over the internet via the GRIA framework components. Exact usage of the service market is determined by providing client sided brokerage. Through active relationship management a repertoire of scoped services would be available locally to customer workflow. The workflow tools will dynamically reflect the enactment options available, facilitating guaranteed availability, fail-over, parallel processing and optimisation and meta scheduling.

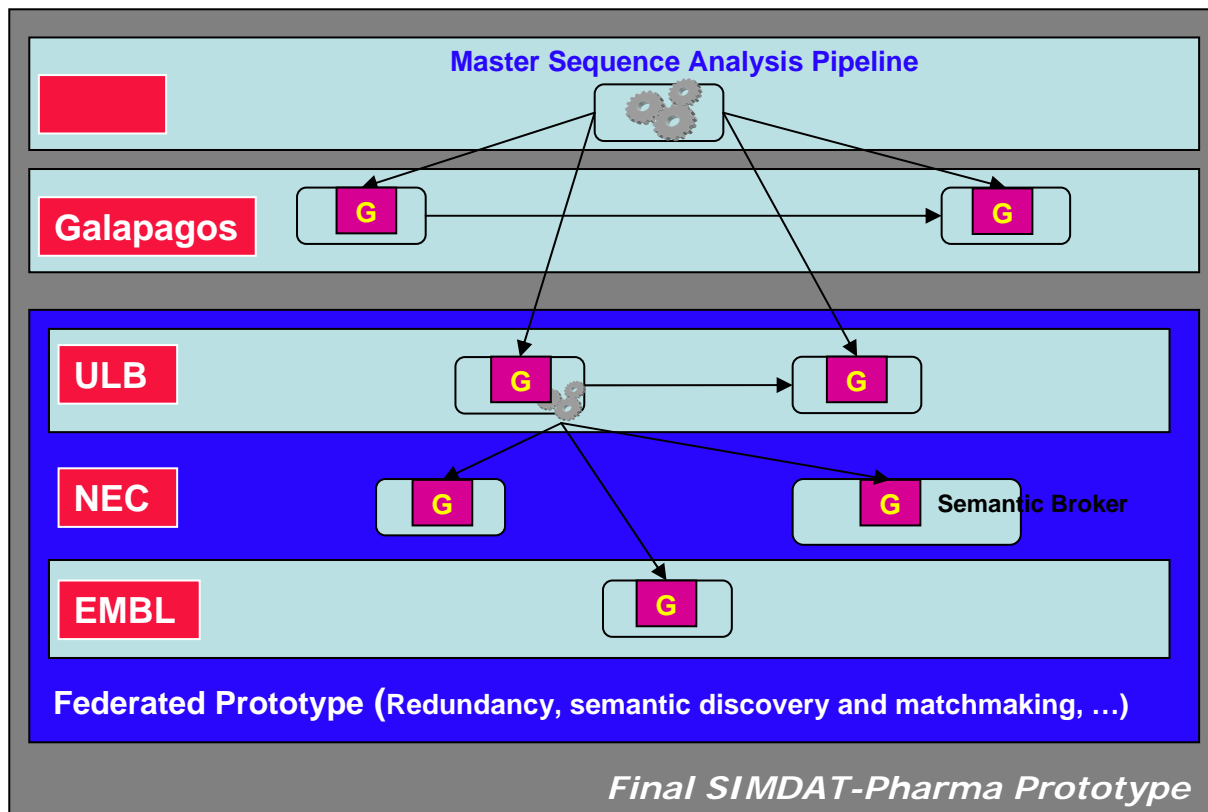


Figure 2: Final Pharma Testbed Installation

The deployment diagram of the final prototype sketches the central idea of the activity: merge the existing installations of the B2B and the Federated scenario. From the perspective of the B2B customer (GSK in our setup) this set-up looks very similar to the initial B2B scenario the only difference lying in the fact that more than one remote service is required to run the MSAP workflow. These additional services can be requested at one of their associated commercial partners (Galapagos in this example) or at arbitrary locations that are part of the Federated Prototype. Important issues to mention here are that the B2B customer neither has any influence how the academic service provider deals with his request nor what kind of technology is used for data processing or data handling. With respect to the technologies that are developed in the Pharma activity like the IGOR-file system or the Semantic Broker it means that there would transparently become part of the GSK workflow execution environment although GSK had never directly accessed or used them due to a variety of reasons. We conclude this chapter with a list of requirements that have to be addressed by the technology partners in the upcoming phase.

6.2 Resulting Requirements

Name	<i>Automated take-up of new applications in workflow authoring systems and GRIA wrappers</i>		
Business Requirements	To automatically generate, based on a complete metadata description of a new application, the workflow components useable in various workflow management systems. To automatically generate corresponding GRIA wrappers.		
Date created	2006-09-28	Source	ULB
First Implementation	2008-03-01	Priority	High
Application Activity	Pharma	Related Prototype	B2A prototype
Technology Component	Grid Infrastructure, Ontology, Workflow	SIMDAT Modules targeted	GRIA, Semantic Broker, KDE, Taverna, TUAM
Detailed description of the requirement			
Take-up of a new application in the SIMDAT-Pharma infrastructure involves several tasks. These are related to an ontology-based description, the publishing of it to the semantic broker, the development of GRIA wrappers and the integration of corresponding workflow components into the various workflow design software. A software module is to be developed to produce these components provided the availability of a complete and rational metadata description of the new application to integrate. A precondition for the development of this module is to gather all prerequisites of the systems for which the generation is targeted, notably GRIA, InforSense and Taverna development APIs, or other.			
Relation to the prototype			
During the lifetime of the SIMDAT-Pharma infrastructure, it is very likely that new applications will require integration. Rather than ad-hoc developments, a generic and comprehensive approach is very desirable.			
Requested functionality			
Presentation of metadata properties to fill in for the characterisation of a new application; automated generation of corresponding GRIA wrapper files and workflow components (nodes) for a set of workflow design systems; publishing of the ontology-based description to the semantic broker			
Validation of the requirement			
Use of this development for existing applications to verify conformity with the targeted systems; use of new applications enabled by this component.			

Name	<i>Customised management of a GRIA-based Pharma service provider</i>		
Business Requirements	To customise account creation, fine-grained accounting and SLA definition		
Date created	2006-09-28	Source	ULB
First Implementation	2007-06-30	Priority	Medium
Application Activity	Pharma	Related Prototype	B2A Prototype
Technology Component	Integrated Grid Infrastructure	SIMDAT Modules targeted	GRIA

Detailed description of the requirement
The service provider shall be in the position to provide its customers with a user-friendly interface for requesting a GRIA account on the SIMDAT-Pharma infrastructure. The server administrator should be able to define the information data fields associated with account templates. We suggest five levels of templates corresponding to various levels of requirements from the user side (trial user, simple user, advanced user, premium user, “à la carte”). The service provider should also be able to define the level of accounting policy connected to the kind of template chosen. We can imagine strong accounting policy for premium (industrial level) users where a simple log can be sufficient for many typical trial users. At the level of the SLAs, the GRIA system administrator should be able to define fine-grained SLA properties for any particular service.
Relation to the prototype
These components are required to build a flexible B2A infrastructure.
Requested functionality
<ul style="list-style-type: none"> • Custom definition of account fields • Custom definition of the information stored service logging and accounting • Custom definition of SLA properties to consider at the level of an individual service
Validation of the requirement
Implementation of modulated access to an existing workflow by various kind of users, while taking into account their respective levels of IPR, security and QoS

Name	<i>Ontology-based workflow authoring and repository</i>		
Business Requirements	To assist the scientist in composing a workflow using domain-specific concepts		
Date created	2006-09-28	Source	ULB
First Implementation	2008-03-01	Priority	Medium
Application Activity	Pharma	Related Prototype	B2A prototype
Technology Component	Ontology, workflow management	SIMDAT Modules targeted	Semantic Broker, InforSense, Taverna, TUAM

Detailed description of the requirement
Natural language description of complex workflows would very fast become useless if some conformity to a rational ontology is not complied to. A workflow creation advisor should take into account domain ontological concepts in order to support the discovery of adequate components at any stage during a workflow design. For instance, it would be useful to restrict the choice of valid components based on the current stage and status reached during the design of the workflow. It is also highly desirable that any completed workflow be described according to a robust ontology of the domain.
Relation to the prototype
The B2A prototype will be based on the execution of complex workflows that involve various partners' resources. These workflows will be designed using this ontology-based advisor.
Requested functionality
Ontology description of bioinformatics workflows and components thereof
Validation of the requirement
Implementation of this advisor should prove to ease the authoring of complex workflows. Development time and implied knowledge from the part of the workflow designer should be lowered.

Name	<i>Client side local service brokerage and workflow enactment diversification</i>		
Business Requirements	To allow a GRIA-enabled service consumer to dynamically utilise remote applications from a trusted partner		
Date created	2006-09-28	Source	GSK
First Implementation	2008-03-01	Priority	High
Application Activity	Pharma	Related Prototype	Fully integrated B2B scenario
Technology Component	Integrated Grid Infrastructure	SIMDAT Modules targeted	GRIA, InforSense workflow
Detailed description of the requirement			
A service consumer needs to build workflow that will have decision points in it's enactment that will be dependant upon the business models and relationship management of the consumer entity.			
Relation to the prototype			
This component is required to provide information on the current repertoire of services scoped locally for the consumer. The available services need to have dynamic representation as nodes for workflow construction.			
Requested functionality			
The Semantic Broker needs to function as a client side application taking the global service availability information and applying negotiated business models and relationships as filters to provide a locally scoped service menu. The possible enactment contingencies must be reflected from this locally scoped service menu as the repertoire of GRIA wrapped nodes in the InforSense workflow engine			
Validation of the requirement			
Implementation of modulated enactment to an existing workflow by various kind of users, while taking into account their respective levels of IPR, security and QoS, demonstrating interoperability and robustness			

Name	<i>Academic Broker based Service Provision</i>		
Business Requirements	To allow a commercial service provider to act as a broker to high value academic services and data sources		
Date created	2006-09-28	Source	ULB
First Implementation	2008-03-01	Priority	Low
Application Activity	Pharma	Related Prototype	B2A prototype
Technology Component	Integrated Grid Infrastructure	SIMDAT Modules targeted	GRIA
Detailed description of the requirement			
Academic research institutions have developed high value services and data sources that they want to exploit commercially within the pharmaceutical industry. However, academic institutions do not have the expertise or inclination to market and manage relationships with potential paying customers. Academic institutions want to out-source these functions to another organisation that is responsible for maintain relationships and brokering requests from a customer to different academic service providers. The service level agreements between the customer/broker and the broker/academics will be different and the broker may need to adaptively select academic resources to fulfil customer requests. This situation matches frequent real life scenarios.			
Relation to the prototype			
This component is required to provide access to resources provided by a brokering organisation			

and high value, low resource research groups.
Requested functionality
A brokering service provider
Validation of the requirement
Implementation of the existing customer workflow (e.g. IX-odus) bound to a broker service provider, whilst meeting the customer requirements for security, QoS and IPR protection

7 Conclusion

With the technology developments in the areas of Grid security, Ontology, and distributed data access the Pharma activity could successfully enhance existing technologies and demonstrate their uptake in two typical life science scenarios. Of particular importance here is the conformity to well accepted standards in the corresponding areas allowing not only smooth integration with SIMDAT's base technologies GRIA (integrated Grid infrastructure) and KDE (workflow) but also paving the way for uptake beyond the boundaries of the SIMDAT project. Feed-back received from various collaboration activities in the context of technical concertation underpinned this.

With the move to GRIA 5 and the availability of InforSense nodes to access this component the industrial partners are on the verge of successfully deploying the B2B prototype demonstrating the usefulness of Grid technology in the area of life sciences. These enhancements will implement an industrial strength pharmaceutical workflow across a Grid test-bed comprised of both academic and industrial partners. Once this has been successfully demonstrated with the new technologies we will be able to further develop our scenario to the Final Prototype allowing Pharma, Academia and Vendors to be able to interact in a controlled and secure way and developing a novel method for eBusiness in this domain. We can see from other vendors in the space an interest in "on demand" services and we see the framework we have in place being of interest to these organisations and plan to demonstrate these tools to those groups as we progress into the later stages of SIMDAT. Also, as the GRIA and Inforsense technology partnership matures we hope to be able to widen the use of the Semantic Broker into the Pharma area and demonstrate the value of such a technology from both an external and internal perspective.

8 References

- [1] Kumpf, K., TUAM: A new tool for universal annotation and mediation, Journal of Web Semantics, 2005
- [2] [B. Motik](#), [U. Sattler](#), [R. Studer](#). Query Answering for OWL-DL with Rules. Proc. of the 3rd International Semantic Web Conference (ISWC 2004), Hiroshima, Japan, November, 2004, pp. 549-563.
- [3] C. Golbreich, "Web rules for Health Care and Life Sciences: use cases and requirements," in Proc. of Reasoning on the Web Workshop at WWW2006, Edinburgh, UK.
- [4] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur and Yarden Katz. Pellet: A practical OWL-DL reasoner, Journal of Web Semantics, 2006.
- [5] Falk Zimmermann et al., Pharma Prototypes for Interoperability Phase, D10.2.1 , <http://bscw.scai.fraunhofer.de/bscw/bscw.cgi/d78249/D.10.2.pdf>