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

D14.1.3 Test report and statement of requirements for phase 2 technology developments from Pharma application area

Start date of project: 1 September 2004 Duration: 48 months

Due date of deliverable: 01 April 2006
Actual submission date: 09 May 2006

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
05.05.	1.0	Falk Zimmermann	

Copyright

Copyright © NEC Europe Ltd. and other members of the SIMDAT consortium, www.simdat.org, 2006

Revision history 2
Copyright 2
1 Introduction 4
2 Prototype Description and Validation 4
2.1 Prototype Introduction 4
2.1.1 Role of semantic Service Discovery 5
2.1.2 Role of the Federated Portal 6
2.2 Validation Criteria and Testbed 7
2.3 Performance Analysis 8
2.3.1 Manual analysis 8
2.3.2 Workflow analysis 10
2.3.3 Overhead Analysis 12
2.3.4 Functional Review 12
2.4 Industrial Relevance 12
2.5 General Remarks 13
2.6 Resulting Requirements 13
3 Requirements for Phase 2 Technical Developments 15
3.1 B2B Scenario 15
3.1.1 Motivation 15
3.1.2 Requirement List 18
3.2 The B2B Prototypes (M24 & M30) 25
Demonstration Activity 27
3.2.1 Peer-to-Peer Filesystem 27
3.2.2 Requirements for Optimisations 28
3.2.3 Dynamic Access Control (DAC) for IGOR-FS 28
3.2.4 Semantic Service Annotation 33
3.2.5 Knowledge enrichment of SIMDAT-Pharma workflow reports 36
4 Conclusion 37
5 References 37

1 Introduction

During the first year of the project, the Pharma Activity enhanced the widely used SRS software with Grid technologies like semantic service discovery and extended security mechanisms. The key problem addressed is the discovery and secure use of distributed databases and analysis tools that are made available via the so-called Federated Portal. Whereas the work during this phase concentrated mostly on the pharmaceutical requirements capture, design and implementation of the prototype, the focus of the reporting period (M13-18) has been on the performance evaluation of the IX-odus workflow: we compared the effort needed to perform a well defined biological task using our testbed environment with the conventional way as it is currently being done in many research institutes (see section 2.2). We include a detailed analysis of each involved step in order to identify every significant performance hotspot. On that basis we could draft requirements needed to further improve the performance of the testbed.

Besides the prototype validation the Pharma partners, in particular GSK and Inpharmatica, designed the so called Business-to-Business scenario, which will constitute the major development task in the second phase of the project. As outlined in detail in section 3.1.1, the central objective, besides introducing innovative Grid and security technology, of this testbed will be to see commercial benefits from the transactions for B2B partners and thus reflecting SIMDAT's strategic goals: raising awareness for the advantages of Data Grids in important industrial sectors and to foster their uptake.

The Pharma demonstration activity will be based on the technology that has been developed during the first phase of the project. It takes the revised M12 testbed environment, where most of the requested changes from the commission will get integrated, as the foundation of a far more sophisticated Grid framework. The key contribution will be the replacement of SRS's data-modules with the newly developed peer-to-peer filesystem (see section 3.3.1).

2 Prototype Description and Validation

In order to test the performance of the prototype, it was agreed during the meeting from 24 to 25 January 2006 NEC premises in Sankt Augustin that UKA will set up and host the Federated Portal with help from ULB and NEC. ULB and GSK will do the validation tests, focussing on efficiency gains.

2.1 Prototype Introduction

This Pharma prototype intends to evaluate the merits of a distributed GRID-enabled infrastructure where centralised components orchestrate in the background the solving of a specific bioinformatics problem without imposing knowledge by the end user of the underlying architecture. As an example, the first approach in this area is focussed on the identification of all genes expressed during the early phase of the interaction of the parasite *Ixodes ricinus* and its host. The goal is to try to identify all the proteins synthesised in the salivary gland of the tick, with the hope to derive a novel way of interfering with the pathology induced by the parasite. While presently involving mainly academic partners, the project included all components required for an industry-level approach. This project had been nicknamed "IXodus".

The experimental approach inside IXodus consists in extracting complete messenger RNA populations from the salivary gland of the tick, before and various times after the initiation of the interaction of the parasite and its host. As several thousands of messenger RNAs are potentially produced at any time, the approach consists in sequencing by automated techniques as many as possible of these sequences after conversion to cDNA. The identification of known candidates by sequence similarity is based on a workflow where several bioinformatics comparison steps are chained inside a decision tree. Sequences producing no hits with the public databanks are sifted out and put forward for further analysis. Till now, the first three steps of the complete analysis workflow have been hard coded in a Java software module.

The architecture of the M12 prototype consists of

- several public analysis servers that hold up-to-date versions of several relevant public biological sequence databanks (essentially EMBL and UniProt)
- a federated portal that provides a user interface and holds the workflow to be executed when requested by the end user
- a project-specific server that holds the confidential project data, including a database where all analysis results are collected for further reference
- a semantic broker where each analysis server is described using XML description files holding the dates and versions of the various databanks and software analysis tools

The different elements of the architecture reside on various partner's computers (NEC, UKa, ULB, SCAI-Bio, Lion Biosciences) that are interconnected using the E2E strong encryption protocol developed by NEC. The Federated Portal has been developed by Lion Bioscience while the workflow has been produced at ULB.

A typical situation inside the IXodus project is that new sequences emerge from the automated sequencing at the rate of about 20 per day and a total of several thousands of sequences are expected to require analysis before it can be considered that completion has been reached.

2.1.1 Role of semantic Service Discovery

The semantic broker (SB) plays a central role in the architecture as every step of the workflow involves a call to the broker in order to identify the most adequate server to accomplish any analysis step. The semantic broker receives, in principle on a regular basis[1], an XML file describing the current status of any of the partner's computers, i.e.,

- Date of last update of each hosted databank
- Version of each databank
- "Availability" flag
- Version of each installed analysis tool

The syntax of these description files is based on a rigorous semantic description and ontology of the various elements. When called by the Federated Portal (see next paragraph), the SB returns the URL of the computer known to host the databank or the analysis tool with the most recent version.

[1] Till now the annotation still requires a manual intervention

2.1.2 Role of the Federated Portal

The Federated Portal (FP) provides a user interface that allows the end user to identify him/herself relative to the private project server. The FP presents four categories of actions:

- Simple search
- Extended search
- Various Blast similarity search options
- Access to the IXodus workflow

Figure 1: Initial Search Page

The first three functionalities are of no direct relevance to the workflow. However, they allow e.g. to compare visually the status of the databanks on the various analysis servers. They also allow for manually checking the presence of some specific sequence(s) in the public databanks.

Launching the IXodus workflow is definitely the central function of the FP. Opening this section of the software exposes an interface designed for copy/pasting cDNA sequences and prepares them for workflow analysis.

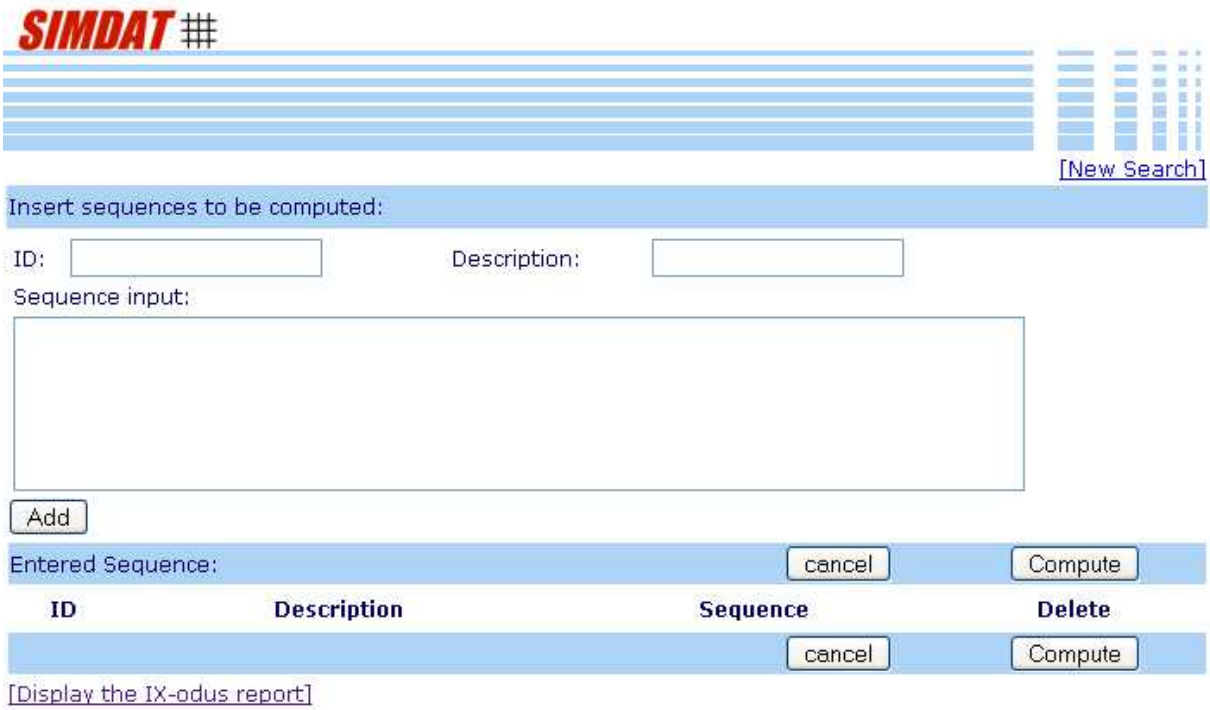
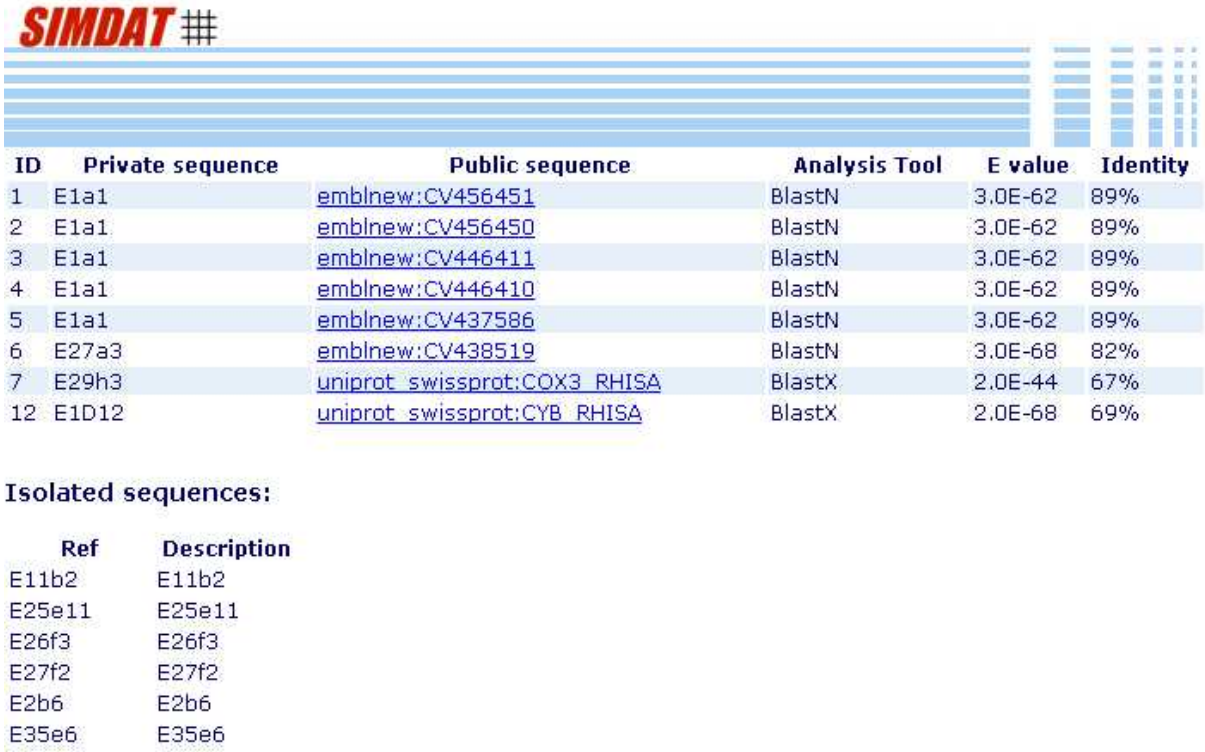


Figure 2: Input Page for Sequences

The interface also proposes to look at the current status of the project. In this mode, it opens a report table where the results about all submitted sequences to the workflow can be witnessed. This table essentially presents six columns:

- The ID of the cDNA (this is automatically attributed by the RDBMS
- The name of the sequence as given by the end user
- The entry of the public databank that produced the best match with this sequence ; this entry is hyperlinked to the actual record in the public databank
- The analysis tool that produced this result (currently BlastN or BlastP)
- The "E- value", i.e. the evaluation of potential relevance of the reported match
- The percentage identity that has been found

Further down on the same page, the user gets a list of all sequences that did not produce any hit when compared with the DNA and Protein databanks. These are reported as "isolated sequences" until eventually a further step of the workflow will return a hit with some known entry. In that case, the entry will enter the upper table with the corresponding fields filled in.


SIMDAT

ID	Private sequence	Public sequence	Analysis Tool	E value	Identity
1	E1a1	emblnew:CV456451	BlastN	3.0E-62	89%
2	E1a1	emblnew:CV456450	BlastN	3.0E-62	89%
3	E1a1	emblnew:CV446411	BlastN	3.0E-62	89%
4	E1a1	emblnew:CV446410	BlastN	3.0E-62	89%
5	E1a1	emblnew:CV437586	BlastN	3.0E-62	89%
6	E27a3	emblnew:CV438519	BlastN	3.0E-68	82%
7	E29h3	uniprot_swissprot:COX3_RHISA	BlastX	2.0E-44	67%
12	E1D12	uniprot_swissprot:CYB_RHISA	BlastX	2.0E-68	69%

Isolated sequences:

Ref	Description
E11b2	E11b2
E25e11	E25e11
E26f3	E26f3
E27f2	E27f2
E2b6	E2b6
E35e6	E35e6

Figure 3: Report Table

2.2 Validation Criteria and Testbed

The main criterion is the efficiency gain by using the Federated Portal instead of the classic SRS Web interface (measured in reduced time consumption per complete analysis workflow).

UKA hosts the Federated Portal[2] which accesses the Semantic Broker and the Node Broker, both located at NEC. Queries are dispatched to the SRS [18], [19] and private project servers hosted by ULB, NEC and SCAI.

[2] All tests reported here were effectively run on a system where the Federated Portal was hosted at ULB, because of configuration difficulties on the UKa computers

For each query the user enters, the FP will query the Node Broker for the most suitable server and dispatch the query to that machine. The Node Broker together with the Semantic Broker [3] is responsible for the selection process. All message transfers between the Federated Portal and the Node Broker are secured by NEC's E2E Security framework [1]:

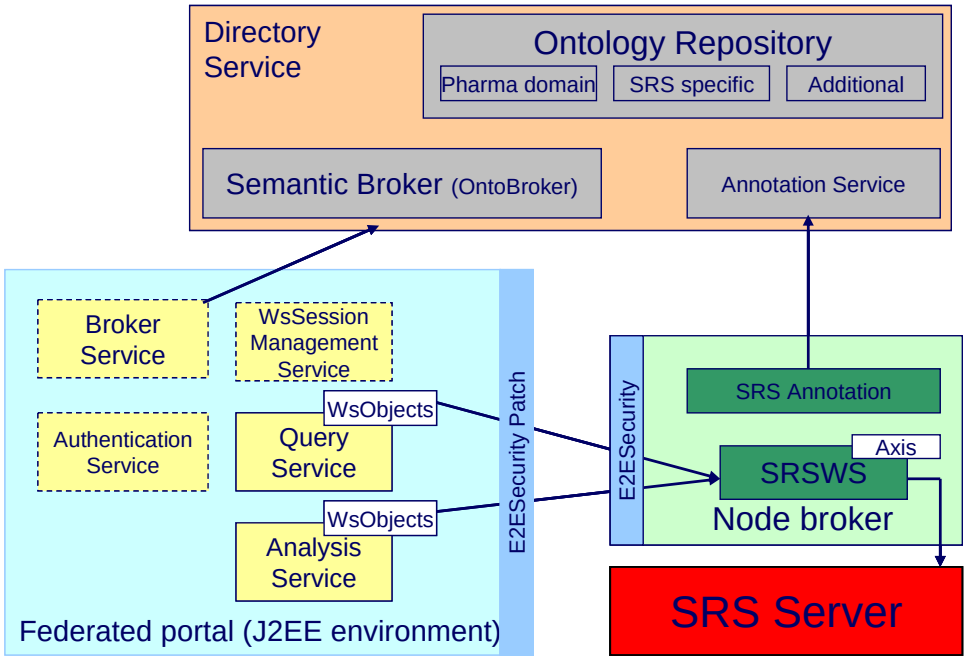


Figure 4: M12 Pharma Prototype

ULB and GSK stress test the system using the workflow integrated in the Federated Portal.

2.3 Performance Analysis

For the performance comparison, we initially took at random a set of 20 cDNA sequences from the research group on tick pathogenicity from the Laboratory of Applied Genetics[3] of the ULB; these sequences came from the automatic sequencing machine. In this set of sequences, about one third had previously been characterised by the research team. The other sequences where considered as "unidentified".

In order to evaluate the relative performance of the SIMDAT Pharma prototype compared with a "manual" approach, we used, besides the SIMDAT servers, the biocomputing infrastructure provided by the Belgian EMBnet Node (BEN). On this infrastructure, the public DNA and protein databanks are kept up-to-date essentially in a similar way as on the project public servers. This set of computers also hosts the various bioinformatics software suites required in this study, and notably the BLAST[4] programs (BlastN, BlastP, BlastX, TBlastn, TBlastX) and the EMBOSS[5] sequence analysis suite.

2.3.1 Manual analysis

The "manual" approach imposes to run each individual analysis step separately, look with an expert eye at each result, estimate its relevance and, depending of this decision, go to the next analysis step or eventually drop the candidate cDNA sequence because of confirmed

[3] Edmond Godfroid, Bernard Couvreur and Maryse Vandenbulcke, Laboratoire de Génétique Appliquée, ULB, Gosselies
[4] Altschul SF; Gish W; Miller W; Myers EW; Lipman DJ - J Mol Biol 215: 403-10 (1990)
[5] Rice,P. Longden,I. and Bleasby,A. - Trends in Genetics *16*, (6) pp276—277(2000)

irrelevance[6]. Whereas only the examination of the results requires expert advice, the manual approach essentially requires expert presence throughout the whole of the procedure as some analysis steps complete in a very short time. Another not to be overlooked aspect of the manual approach is that it requires accurate accounting of the completion of each step. In our case, we ran the manual process with one person manipulating the computer interface while the second person kept an accounting of what had been done at every moment.

We undertook the manual analysis for only the first 8 entries as the time and effort involved in this was already rather overwhelming. Two persons were involved full time for a period of 45 minutes and reached the conclusions presented in the following table. Initially, the research group considered analysing the daily set of 20 sequences by hand on a day-by-day basis, a definitely overwhelming task on the long run.

Sequence identifier	BlastN private databank	BlastN public databank (EMBLNEW)	BlastX public databank SwissProt	Note	Relevance to subject of research	clock time
E11b2	first of batch	no hit - E-value too high	hit with P63174	was previously identified by research group	confirms previous work	17:24
E25e11	no hit	no hit - E-value too high	no hit - E-value too high	yet unknown	confirms previous work	
E26f3	no hit	no hit - E-value too high	no hit - E-value too high	yet unknown	confirms previous work	
E27a3	no hit	CV438519	n.a.	was previously identified by research group	confirms previous work	
E27f2	no hit	no hit - E-value too high	no hit - E-value too high	was previously identified as similar to Q01042	***reveals threshhold too low in workflow***	
E28e3	no hit	no hit - E-value too high	Q97319	was previously identified by research group	confirms previous work	
E29h3	no hit	CV455153	n.a.	was previously identified by research group	confirms previous work, is a tick protein	
E2b6	no hit	no hit - E-value too high	no hit - E-value too high	yet unknown	confirms previous work	18:07

[6] The most frequent case of irrelevance comes from the identification that the current cDNA is already completely, or essential completely present in the existing project dataset. Indeed, it can be expected from any high throughput cDNA sequencing project that many copies of a limited number of sequences is found in great abundance

Table 2-1: Performance Evaluation

The manual procedure requires great attention in order to avoid errors and erroneous entries. For instance, when a BLASTN run must be started, the level of the E-value has to be chosen in order to filter out trivial and meaningless similarities. The repetitive character of the approach is very error-prone. Besides this, the results of the various analysis steps are stored at various places depending on which computer which task was run. This makes further reference to the results of the analysis rather cumbersome.

The average treatment time of a sequence is 322 seconds.
The average sequence-per-manpower time is 644 seconds.

2.3.2 Workflow analysis

The workflow approach only requires the user to copy/paste the sequences to the interface and click a button to trigger the whole procedure. In order to follow the progress of the workflow, we checked permanently the log file of the *Jboss* server. In this way, the dialog between the FP, the SB and the execution nodes can be observed.

The copy/paste part of the work is clearly the most cumbersome one as the user has to make sure to type in or copy/paste precisely each sequence and sequence identifier. We ran the set of 20 sequences in three batches:

- The first batch consisted of only one sequence
- The second batch consisted of 15 sequences
- The third batch consisted of the remaining 4 sequences

The copy/paste part of the work took about 12 minutes for the set of 15 sequences. Thereafter, the execution of the workflow took about 6 minutes (detailed times are available in the following table).

SIMDAT prototype validation

N.B. Some sequences do not require BlastX execution

Sequence identifier	cDNA analysis launch time	time for local BlastN comparison and call of SB	time when SB replied	time when BlastN effectively starts (end of E2E negotiation)	time to produce BlastN result (EMBLNEW)	time for E2E transport of BlatsN result	call SB for BlastX execution	time when SB replied and E2E negotiation starts	time when BlastX effectively starts	time for BlastX execution (SwissProt)	time for E2E transport of BlatsX result	time to report similarity in project database	**Total execution time**	**Time spent in method**	**time spent in negotiation and secured transport**
E1a1.		2	5	6	26	3						0	**42**	**28**	**14**
E27a3.		5	3	3	12	1						0	**25**	**17**	**8**
E27f2.		5	4	2	12	1						n.a.	**24**	**17**	**7**
E28e3		5	3	2	12		0	3	3	8	0	n.a.	**37**	**25**	**12**
E29h3		6	3	3	13	0	0	3	2	7	4	0	**41**	**26**	**15**
E2b6		7	3	3	11	0	0	3	2	9	0	n.a.	**38**	**27**	**11**
E33d4		6	4	2	14	0	0	3	2	8	13	n.a.	**52**	**28**	**24**
E35e6		5	4	2	10	0	1	3	2	6	0	n.a.	**33**	**21**	**12**
E42c6.		8	3	5	11	0	0	4	7	9	0	n.a.	**47**	**28**	**19**
E43a2		7	3	4	13	0	0	3	3	7	0	n.a.	**40**	**27**	**13**
												Average times	**0**	**37,9**	**24,4**	**13,5**
												Standard Deviation Average		**6,52**	**3,64**	**3,6**

Table 2-2: Prototype Workflow Analysis

The job consisting of 20 sequences was started at about 16.30 and finished with all results reported in the project database at 16.48. Note that the manual copy/paste tasks consume more than 60% of the total time! This confirms our impression that a *batch* mode for sequence introduction is required for the future steps of development of the system.

The average treatment time of a sequence is about 40 seconds.
The average sequence-per-manpower time is also 40 seconds.

In order to confirm the average time required by a typical sequence, we took another batch of 17 sequences coming from the automated sequencing procedure and submitted it as a single job. The entry time of the sequences in the FP was about 13 minutes. The execution time of the workflow time was 11 min. 17 sec. minutes or about 39 sec. per sequence.
From these tests, we can conclude that the average time to analyse any sequence from the project in the workflow mode is close to ***40 seconds***.

2.3.3 Overhead Analysis

Given that the BLASTN and BLASTX analysis steps require approximately 12 and 8 seconds, respectively, the overall overhead imposed by the distributed architecture is reasonably low. The E2E dialog is the second most time consuming step with an average E2E negotiation period of 6 seconds. The impact of this on an overall real life analysis that would imply the complete EMBL public databank, rather than the much reduced set we were compelled to use here would significantly reduce the impact of the E2E protocol on the system. We plan to implement a revised AXIS SRS client in the near future in order to be able to run the workflow on the complete dataset.

2.3.4 Functional Review

As a whole, the prototype installation fulfils all the requirements put forward at the time of its conception. Each major component like the Federated Portal, the Semantic Broker and the SRS public databank and public service providers acts as designed. One significant addition to the original architecture is the private project server, where all project sequences are entered and all analysis results are collected for further reference. The addition of this component mimics in several ways the situation of a pharmaceutical partner, who holds large sets of private data that may be required at some step of an analysis workflow.

2.4 Industrial Relevance

The SIMDAT-Pharma PM12 prototype addressed essentially the security aspect of common industrial requirements. The use of the E2E security framework is enforced for all dialogs between the Federated Portal and the execution nodes. To reach this functionality we included in each execution node a proxy component (Axis) to access the SRS server. As such, the Pharma prototype embeds several elements of a valid architecture, subject to further developments to include the concepts of VO in the frame of B2B scenarios and SLA management.

Nevertheless, the current security implementation requires improvements at least at two levels:

- The dialog between the end user and the FP shall be encrypted
- The results of every analysis run on a public server have to be removed from the server after completion

- The issue of the need to encrypt the calls to the Semantic broker has to be discussed

The IXodus project, merely an academic example implementation of the PM12 prototype, consists of essentially the same steps that might be required in a cDNA bank analysis in any pharmaceutical company. While in this project, only a few analysis services were effectively implemented (BlastN against the private databank, BlastN against the public EMBLNEW databank and BlastX against the SwissProt databank), nothing prevents extending the workflow with an arbitrary number of analysis steps. These analysis steps could be specialized functions provided only by experts on any particular domain and depending on specific SLA status or by classical public servers.

2.5 General Remarks

The limited IXodus prototype proved that a so-called workflow approach can significantly improve the speed of a moderate or high throughput bioinformatics analysis, and at the same time improve its quality by removing the very cumbersome and error-prone judgement by the scientist at every single analysis step. However, the prototype also proved that a hard-coded workflow is not a reasonable approach, as every minor change in any of the numerous parameters requires both programmer's intervention and, subsequently, redeployment of the application. This further stresses the need to incorporate a workflow engine into the process, and preferably a dynamic kind of workflow where most parameters remain under control. It might even be nice that an expert-designed workflow allows the end user to modify the values of any of the parameters within a range of valid values provided by the expert.

During the installation of the hard-coded workflow, decisions had to be taken about which links would be presented to the user about the hits identified in the public databanks. We identified among other the importance to provide links to generic datasets, rather than to specific subsections of these. Typically, a link identified at one moment of time to a sequence residing in EMBLNEW will inevitably enter the EMBLRELEASE section within the upcoming three months at the most. The mechanism built-in SRS thankfully provides this level of "generic ness".

Evidence from our experience is that an analysis run on any sequence from a project has to be repeated on a regular basis until eventually a similarity gets revealed. Such repeat should be scheduled or at least whenever a new release of a databanks is published. A negative result at one time may turn into an essential discovery days or weeks later. Our limited hard-coded workflow did not pretend to take care of this, but a full-blown system must certainly present this provision.

2.6 Resulting Requirements

The following lists summarises all issues that will be addressed in the near future to improve the current M12 prototype. As already mentioned in the introduction, this work will be done in the context of the demonstration activity (see section 3.3), where the M12 prototype will be recycled as the basis for the developments. Some level of encryption between the end user and the FP interface

- The project computer and possibly all server computers should be properly protected by advanced firewalling technology
- The system requires an on-the-fly annotation of the FP whenever a change in one or more of the datasets or tools happened

- The system requires to build in redundancy, in such a way that no single point of failure remains in the system. At this stage, the FP and the SB are critical components
- All computers serving the project should be properly hidden behind firewalls (which can be an issue for some public data providers)
- The architecture should involve GRID enabling technologies like GRIA
- The hard-coded workflow proved its limits, as the change of any of the parameters (e.g. the threshold E value applied to a BlastN output to take a workflow decision) requires manual intervention and redeployment of the FP
- At this time only a single workflow is available. Provision is required for any number of distinct workflows, each of them protected by user identification
- All analysis results should be collected in the project server, rather than reside in the remote execution node. While this is not vital, it makes further human analysis of the results cumbersome. Moreover, this situation is potentially insecure, as analysis results remain scarcely protected in remote computers.

3 Requirements for Phase 2 Technical Developments

3.1 B2B Scenario

The design and the initial implementation of the Business-to-Business prototype will be the central task the Pharma Activity within the second phase of the project and be demonstrated at the next project review. While initially conceived for the management of relations among industrial partners, this scenario is also of immediate relevance for any EMBnet service provider like ULB. Its academic customers occasionally require entering VOs to support R&D partnership, including both academic and industrial stakeholders.

We start this section with an motivation of the scenario, outlining the relevance of the addressed work within the pharmaceutical domain. Subsequently we specify detailed requirements on various technology components and attach priorities to them to indicate their individual importance. The plan is to develop several iterations of the B2B scenario. Starting with a simple functioning Gria framework that only allows for straightforward access to resources, each further instance should add technology developed within the Pharma (or other application or technology) sectors, e.g., semantic service discovery and the P2P filesystem.

3.1.1 Motivation

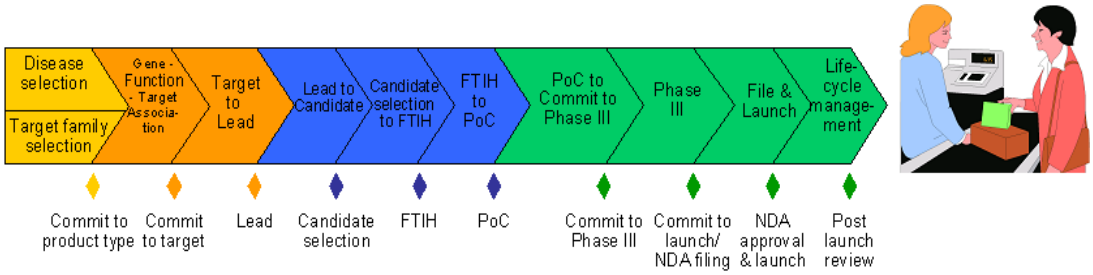


Figure 5: Drug Discovery Pipeline

Drug Discovery is a complex high risk process requiring the use of information technology throughout the lifecycle. One challenge for Pharmaceutical companies is how to reduce cost and risk whilst increasing the probability of successfully developing new drugs. Innovative solutions to assist this process exist both within large pharmaceutical (Big Pharma) and outside in the hands of hundreds of biotechnology businesses and academic institutions. Increasingly, it is the use of technology to enable successful partnerships between these entities that will achieve the Drug Discovery goal.

The possibility of GRID enabled partnerships opens up potentially valuable options that would not be possible using traditional business relationships. This will be of benefit to large pharmaceutical companies as customers and their (generally) smaller partners (suppliers) whether commercial or academic. At each stage of discovery it should be possible to use proven, secure, GRID enabled technology to partner with specialists who can add value and reduce risk at a specific stage of the discovery pipeline. The integrity of the data and results will be key, in order to protect Intellectual Property (IP) and the validity of results. This implies effective security that operates in an unobtrusive manner. The wide success of such a solution will depend on the use of trusted but open standards and open tool sets. This will make it possible for the widest range of technology partners to enable their platforms and

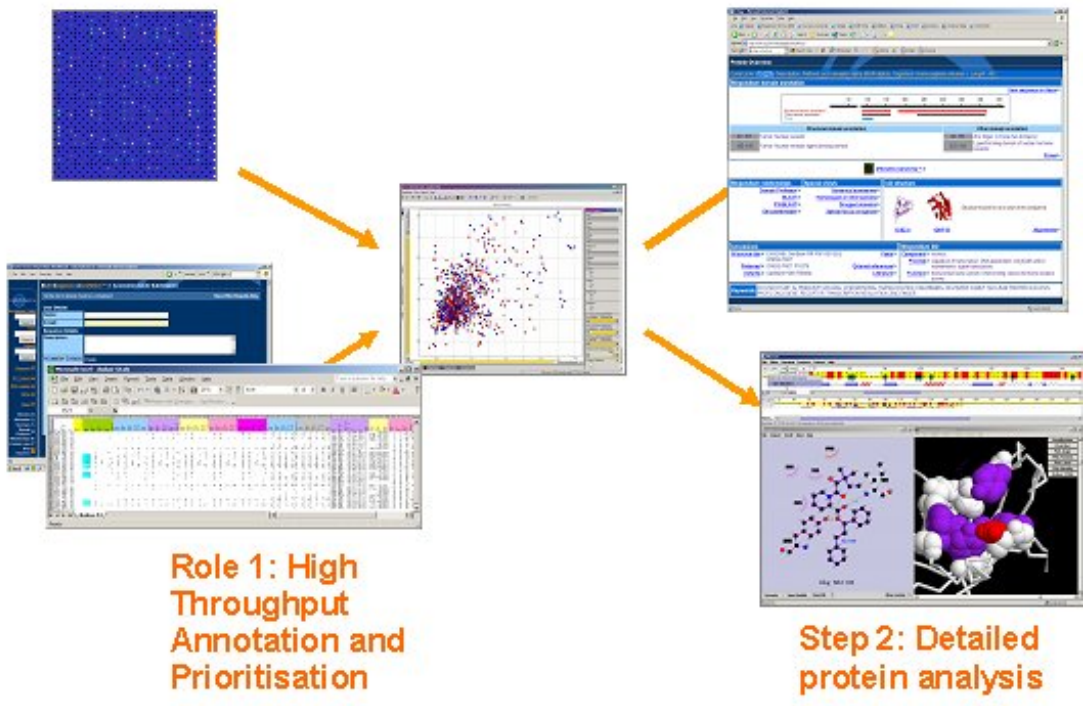
services to connect to a set of such GRID enabled technologies. The GRIA documentation provides further background on many of the associated issues and business drivers[7].

In defining initial requirements to demonstrate a real B2B scenario within the SIMDAT Pharma sector, Inpharmatica and GSK have worked together to define requirements from the supplier and customer perspective. We suggest that for success to be achieved it has to be possible for both B2B partners to see commercial benefits from the transaction. By demonstrating these mutual benefits, we believe that a wider take-up of this technology will be more likely.

- GSK have defined a technology and business model that will establish a GRID enabled portal. Serviced can be delivered using workflow enabled processes or standalone process. The SIMDAT infrastructure is required to provide the link to third party supplies.
- Inpharmatica delivers a wide range of data, applications and services to pharmaceutical customers and will implement access to these using the GRID enabled services based on the SIMDAT model. In this case these are mostly based around biological data[8].

The specific business case that will be demonstrated will be based around the analysis and annotation of selected sets of proteins.

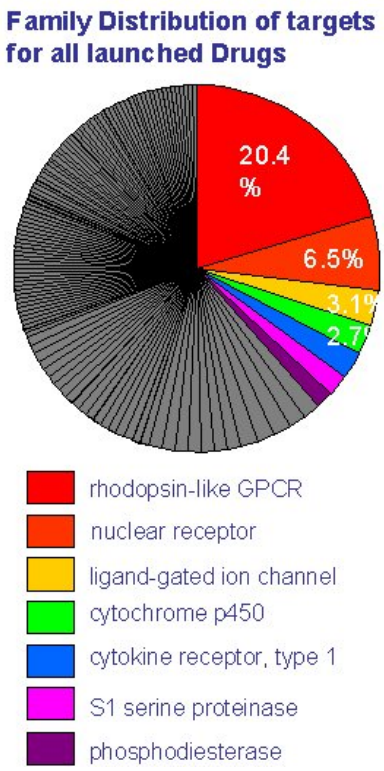
1. Inpharmatica will publish a service to provide aggregate and detailed annotation of protein data including similarity based on structure, ligand binding sites and sequence annotation in various predetermined formats.
2. GSK will subscribe to, and request results using the service.
3. Each request for the GRID services will identify a set of data that require additional annotation and the required format for the resulting annotation.
4. Inpharmatica will deliver the resulting annotation to complete the transaction.



[7] http://www.gria.org/docs/DolphinDPSN04.pdf
[8] http://www.inpharmatica.com/biopendium.htm

Figure 6: Protein Annotation

This scenario is typical of collaborations in the Pharma sector. It recognises that whilst large Pharmaceutical companies are able to develop and utilise complex internal technical platforms, they will often wish to draw on specialisations that exists outside the organisation. By making it easier to request services from best of breed suppliers and to establish a role for services on demand they are able to leveraging in-house vs. in-sourced solutions. The scale of the technology, the need to distribute data and the services used to process this data makes GRID related technologies a preferred option.



Analyses of protein data to support lab work using MicroArrays (Chips) or to identify protein families that are suitable drug targets are examples of analysis that may be the subject of collaborations.

Figure 7 - Druggable protein families - Inpharmatica Ltd

Inpharmatica undertakes research in collaboration with academic bodies and Pharmaceutical companies. A recent example would be a contribution to work leading to new understanding of Crohn's disease [27].

Detailed Business Requirement for the SIMDAT B2B Pharma prototype are provided on the following pages. In summary:

1. **(Customer Side)**
    a. To select and partner with service providers able to deliver specialised annotation/analysis processes.
    b. To buy in services in a timely fashion at lower cost than would be achieved with in-house solution alone.
    c. To be able to demonstrate that security of data and access is such that actual or potential Intellectual Property (IP) value is not put at risk.
    d. To use Workflow technologies integrated with existing solutions to manage the flow of data and process through the Drug Discovery pipeline.
2. **(Supplier Side)**
    a. To accept service requests from a range of customers with low overhead and low lead time.
    b. To reuse applications and technology on behalf of customers and demonstrate appropriate controls over access and usage etc.
    c. To be able to guarantee the integrity of data and resulting analysis for each customer to ensure their IP is not put at risk.
    d. To make use of Workflow solutions to offer secure and straightforward GRID based integration with customers and other third party collaborators including

academic partners. (Will require partnering with Workflow library providers to develop and deliver libraries and components)

3.1.2 Requirement List

The descriptions of the requirements follow the structure that has been requested by the SIMDAT project management for application activities.

Name	***B2B Service Provision***		
Business Requirements	To accept service requests from a range of customers with low overhead and low lead time		
Date created	2006-03-22	**Source**	Inpharmatica/GSK Steve Tringham
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			
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. Subsequent iterations of the prototype are intended to use Distributed Files System as part of the data delivery process.			
Relation to the prototype			
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			
Requested functionality			
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			
Validation of the requirement			
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			

Name	***B2B Service acquisition and cost***		
Business Requirements	To buy in services in a timely fashion at lower cost than would be achieved with in-house solution alone.		
Date created	2006-03-22	**Source**	Inpharmatica/GSK Steve Tringham

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			
The GRIA components will be used to establish the service and validate the existing of a valid contract as part of the delivery of the Service to the Customer. In addition both the customer and Supplier will set expectations of the effort expended and costs of achieving the Analysis/Annotation service via the GRID in comparison with costs and alternative models of supply. During the prototype the Customer and Supplier will assess these in order to determine the success of the Service.			
Relation to the prototype			
The purpose of the prototype is to show a real Pharma B2B scenario. This component uses the GRIA technology to establish and manage commercial transactions through the GRID technology.			
Requested functionality			
Documented expectations and assessment of the outcome.			
Validation of the requirement			
The validation of the requirement will be the ability to show the success of the prototype.			

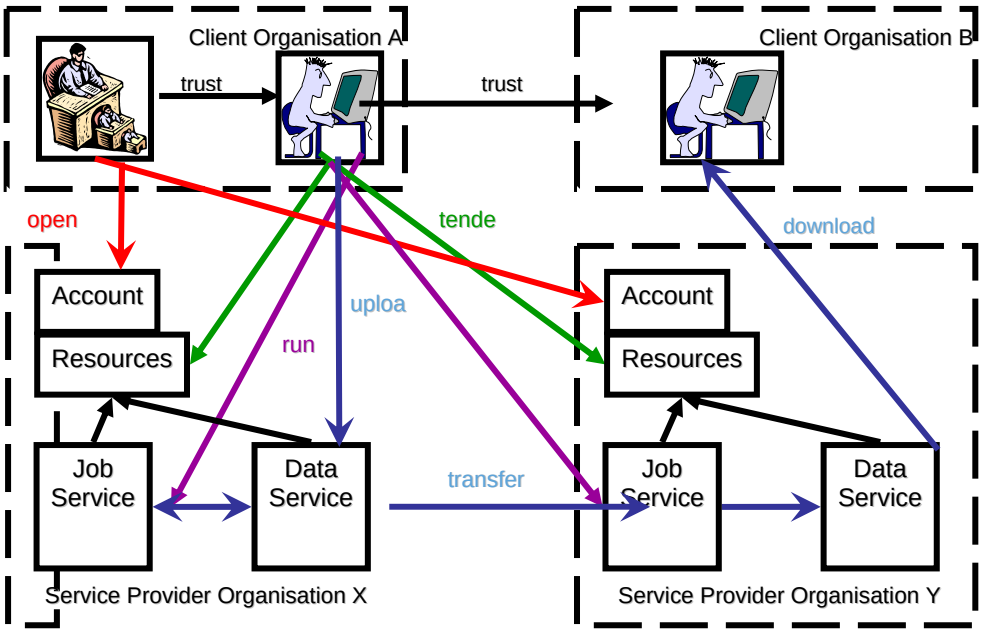


Figure 8: GRIA core services

Name	***B2B Security of Customer Intellectual Property (IP)***		
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**	Inpharmatica/GSK Paul McNair
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	***Workflow technology integrated flow of data and process through the Drug Discovery pipeline.***		
Business Requirements	To encapsulate and enact service requests from a range of business related workflow to achieve integrated data and process flow in support of the drug discovery pipeline		
Date created	2006-03-22	**Source**	Inpharmatica/GSK John Armstrong
First Implementation	2006-08-31	**Priority**	High
Application Activity	Pharma	**Related Prototype**	B2B
Technology Component	Workflow	**SIMDAT Modules targeted**	KDE
Detailed description of the requirement			
Relation to the prototype			
Core system. Workflow will be used to underpin all client side operations facilitating data collection, analysis and integration			

Requested functionality
InforSense Workflow must allow seamless integration with in-house systems allowing data exchange and control flow passing InforSense Workflow must provide tools to allow the integration of GRIA framework components as Nodes. GRIA Nodes must functionally map to the GRIA component architecture. Publication of InforSense Workflow meta-Nodes must ensure encapsulation of any required GRIA functionality. Additional framework components to facilitate service negotiation or enhanced security should have concomitant development in InforSense Workflow
Validation of the requirement
InforSense Workflow will allow full use of the GRIA framework. InforSense Workflow will allow Drug Development Pipeline support(within the context of the defined B2B scenario) InforSense Workflow will satisfy the agreed Client acceptance criteria(within the defined B2B scenario)

Name	***B2B Service Provision (Supplier)***		
Business Requirements	To accept service requests from a range of customers with low overhead and low lead time.		
Date created	2006-03-22	**Source**	Inpharmatica/GSK Steve Tringham
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			
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.			
Relation to the prototype			
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.			
Requested functionality			
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			
Validation of the requirement			
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.			

..

Name	***B2B application and technology reuse by Supplier)***		
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**	Inpharmatica/GSK Steve Tringham
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			
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			
Relation to the prototype			
The purpose of the prototype is to show a real Pharma B2B scenario. This component 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	***B2B Security of Supplier IP***		
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**	Inpharmatica/GSK Paul McNair
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.

Name	***Knowledge service encapsulation.***		
Business Requirements	To encapsulate business relevant knowledge services to facilitate distributed access using GRID infrastructure		
Date created	2006-03-22	**Source**	Inpharmatica/GSK John Armstrong
First Implementation	2006-08-31	**Priority**	High
Application Activity	Pharma	**Related Prototype**	B2B
Technology Component	Workflow	**SIMDAT Modules targeted**	KDE
Detailed description of the requirement			
To make use of Workflow solutions to offer secure and straightforward GRID based integration with customers and other third party collaborators including academic partners. ((will require partnering with Workflow library providers to develop and deliver libraries and components))			
Relation to the prototype			
Core system. Workflow will be used to underpin supplier side operations facilitating service based resource access			
Requested functionality			
InforSense Workflow must allow seamless integration with in-house systems allowing data exchange and control flow passing. InforSense Workflow must provide tools to allow the integration of GRIA framework components as Nodes. GRIA Nodes must functionally map to the GRIA component architecture. Publication of InforSense Workflow meta-Nodes must ensure encapsulation of any required GRIA functionality. Additional framework components to facilitate service negotiation or enhanced security should have concomitant development in InforSense Workflow InforSense workflow must allow rapid and easy encapsulation of layered knowledge services as Nodes callable from the client interface.			
Validation of the requirement			

InforSense Workflow will allow full use of the GRIA framework. InforSense Workflow will allow Drug Development Pipeline support (within the context of the defined B2B scenario) InforSense Workflow will satisfy the agreed Client acceptance criteria (within the defined B2B scenario) InforSense Workflow will allow timely provision of business relevant knowledge services (as required within the define B2B scenario.

3.2 The B2B Prototypes (M24 & M30)

From an architectural point of view, the Pharma prototype of the interoperability phase will be based on the following diagram:

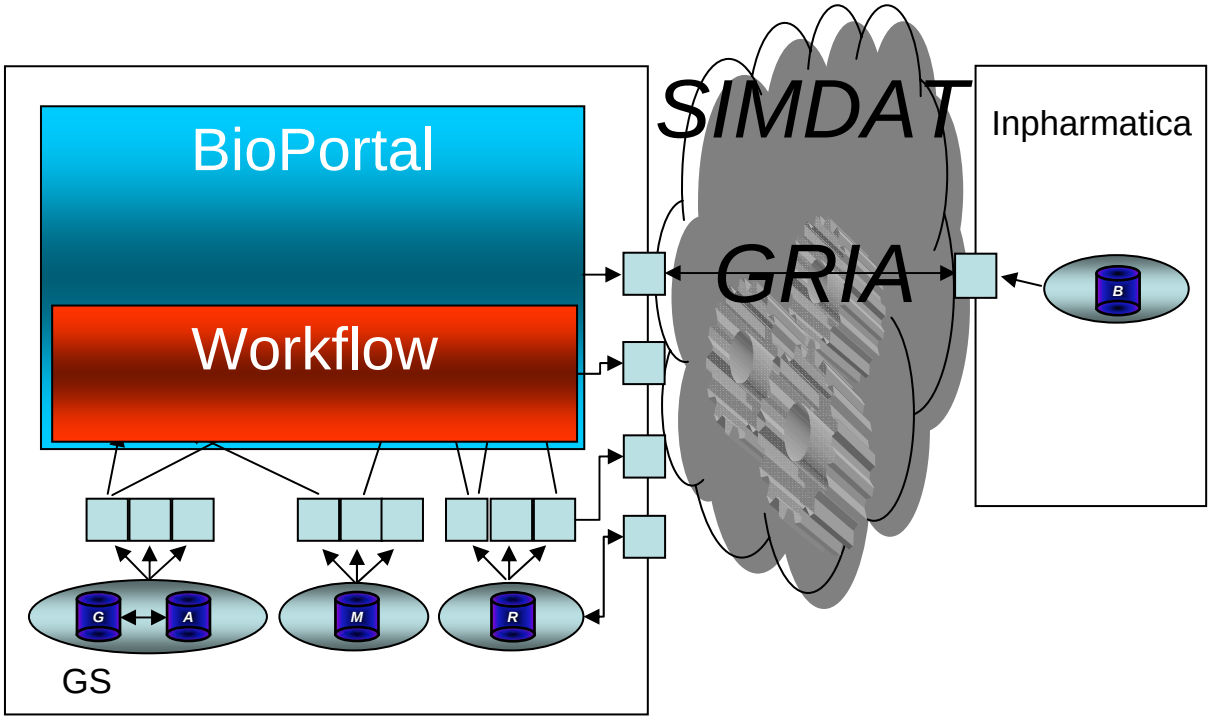


Figure 9: M24 B2B Prototype

As already outlined in the previous sections, the first prototype (M24) will allow for straightforward access to resources, employing technology like GRIA (Integrated Grid Middleware, DAC (Dynamic Access Control - Security), and KDE (Workflow). In general, we will adopt an incremental development plan that allows us both the integration of new technology components like semantic service discovery and the P2P filesystem and the extension of the underlying testbed. In particular, the integration of additional remote sites (from the perspective of GSK) will increase the complexity of the VO administration and, thus, the involved security mechanisms.

The ultimate goal, however, would be to have a final scenario that comprises the B2B M30 prototype with the Pharma demonstrator. Although it is quite unclear at the moment whether enough resources are available for such an activity, the underlying pharmaceutical scenario would allow the integration of SRS technology in the analysis pipeline.
The Pharma Activity is confident to provide a detailed description of the M30 scenario in the upcoming M24 deliverable, where the experiences of the developments until this milestone are carefully considered.

The following diagram shows the envisaged layout of the M30 B2B prototype.

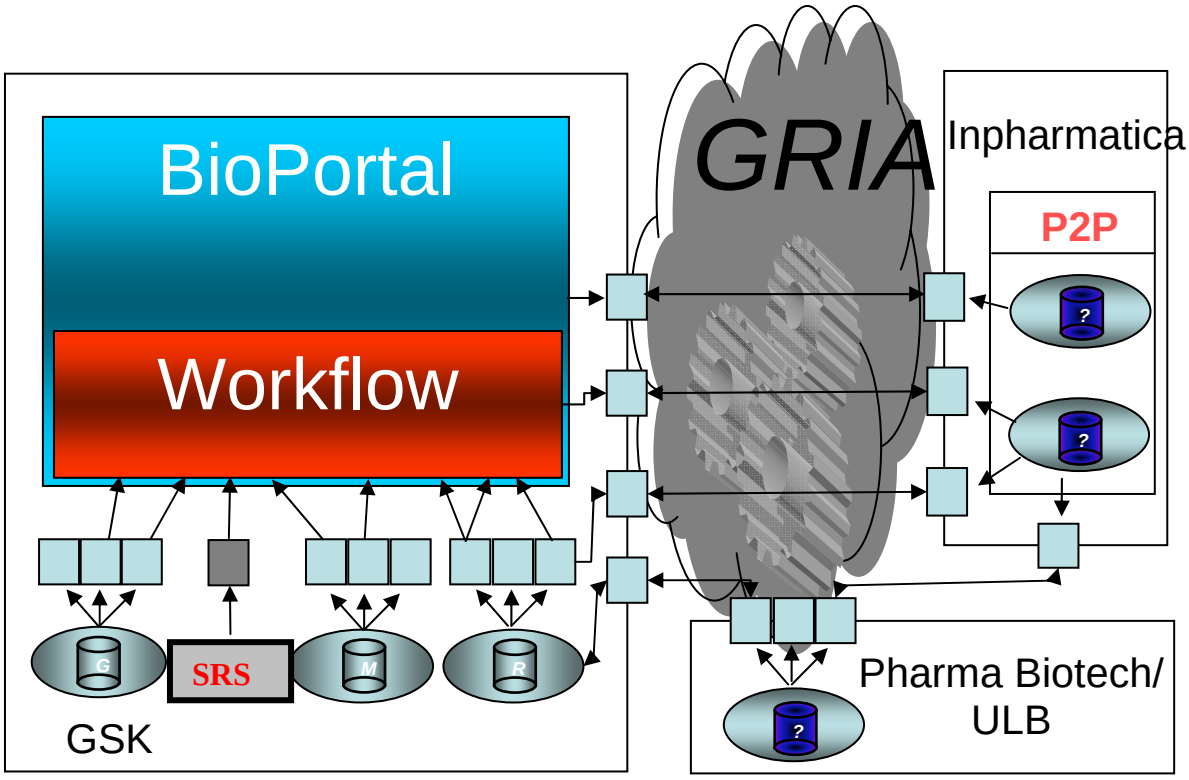


Figure 10: M30 B2B Prototype

3.3 Demonstration Activity

The demonstrator resulting from this activity bases on the technology which has been developed during the first phase of the project and integrated into the M12 prototype. In particular the new components for semantic service discovery (Semantic Broker) and transport security (E2E security framework) will be either re-used or further developed (Ontology framework, see section 2.1.1 and 3.3.4) within this activity. The new key contribution in this area, however, will be integration of the peer-to-peer filesystem IGOR[27] which should significantly improve the efficiency of the underlying data-handling. The IGOR filesystem is being developed at the University of Karlsruhe (UKA) and constitutes one of Pharma's central technology developments.

Whereas information on the M12 prototype has been given earlier in this document (see section 2, [1]), the description of the demonstrator in this chapter concentrates mainly on the new technology components. Similar to the B2B prototype security plays an important role also in this context. The section on the dynamic access control outlines in detail design and implementation strategies for the P2P based architecture.

Rather than specifying requirements in the format as it is requested for prototypes (with the exception of one requirement for Knowledge Services) the resulting requirements for the demonstrator are directly integrated into the specific sections. This is mostly motivated by the fact that the majority of all requirements will be addressed from the Pharma partners themselves. However, with respect to security, NEC and IT Innovation will team up to develop and integrate appropriate modules.

The major difference between a SIMDAT prototype and the demonstrator is the maturity of the underlying technology. All extensions that we're adding to the core system, in our case the SRS software, should have "industrial strength". That's why we are restricting the technology uptake for the demonstrator to functionality that we believe has the maturity and quality to be integrated into a commercial product.

3.3.1 Peer-to-Peer Filesystem

Many Pharma applications, including SRS, work with large collections of semi-structured data. Often, the data is stored in flat files, so that a remote site needs to download the entire collection before researchers can access the respectively interesting part of such a file. The IGOR file system makes such large files easily accessible without prior download. With IGOR, all files of a collection appear to be part of the locally available file system. Only when the application accesses a file, the respective part is loaded into the local machine. There it is cached for future use. Moreover, this cache serves as replica server for other sites accessing the same piece of data.

The Pharma demonstrator will be based on a first version of IGOR that offers full read access to the published data, but that limits the system to a single publisher. In detail, we will show several machines at various sites (e.g. Brussels, St. Augustin, Karlsruhe, and Cambridge) that have access to a selection of typical bioinformatics data bases (e.g. the EMBL and UNIPROT collections) so that they can be processed by the SRS installations at the respective sites. Data will be fed into the IGOR file system at one source site and will be readable at all the other sites. We will demonstrate the efficiency of the distribution by showing an example that publishes a small modification to a large file. This will be contrasted by the performance of state-of-the-art data distribution mechanisms which have to download the entire file.

Technology-wise, the demonstrator will contain (1) the IGOR-FUSE module that connects the IGOR file system to the LINUX operating system, (2) the IGOR structured routing overlay, which conveys IGOR protocol messages between the participating sites, and (3) the IGOR file system cache, including the access control and crypto modules. The routing overlay will exploit network proximity and respect machine load, so that for each piece of data the most responsive cache is queried. The details of the file system cache are described in a working draft of IGOR. In particular, the cache will allow high-speed access to previously read data and provide replica for the data distribution process.

3.3.2 Requirements for Optimisations

The IGOR file system (FS) adheres to a subset of the general POSIX file system interface, so that any application will be able to productively work with the IGOR FS, too. Although, the IGOR FS will thus provide efficiency gains to unmodified applications, even more improvements can be achieved by slightly modifying the applications that use the IGOR FS. These modifications are optional and independent of the regular file system functionality.

The following two options have been identified in extensive discussions between UKA and LION:

1. Block size hints. Similar to most other file systems, the IGOR FS stores data in blocks. These blocks may have variable size, but certain limits must apply to not degrade system efficiency. If the application hints to the file system which is the typical size of those data chunks that are typically accessed as an application data unit, the performance of the IGOR FS can be optimized accordingly. This should include the hinting of typical start and end delimiters for such blocks. For example, XML documents use special tags to structure the data. If the IGOR FS is aware of the position of these tags, it can create blocks that respect the document structure. This reduces both cache and network load when individual application data units are modified in a large file.

2. Block offset identifiers. According to the file system semantics, a file is a linear data structure where the offset from the beginning of the file is the only means of accessing individual data chunks in the file. But, as said above, IGOR should be aware of a potentially existing file structure. In that case the offset is not very well suited to address the application data units in the file. The application should rather use a different block identification mechanism. Especially, this would keep the block identifiers of unmodified blocks constant even when individual preceding blocks change in size. This could greatly increase the performance and efficiency of the IGOR FS.

UKA and LION have discussed how these two options can be included into the SRS application. So far, both partners do not yet have enough experience with the IGOR FS to make a justifiable decision. Hence, these options will not be included in the first version of the demonstrator release. However, the first demonstrator will be used as a basis for evaluating the options.

3.3.3 Dynamic Access Control (DAC) for IGOR-FS

Drug discovery in the pharmaceutical industry has to deal with a huge amount of data, ranging from several hundred Gigabytes to a few Terabytes, today. Moreover, the amount of data is quickly increasing. Therefore, it is important to provide a resource-efficient way to share this data among the various sites using that data, be it globally scattered labs within one

legal organization or virtual organizations, e.g. collaborations between academic and industrial partners.

Hence an important issue for Pharma middleware is the ability to build shared working spaces, where different types of users can share resources, data and computation results and at the same time protecting proprietary and private resources and information. The Dynamic Access Control (DAC) mechanisms will target these requirements. The DAC component will allow users to build working groups dynamically, according to attributes and user identities and will allow moreover for fine grained authorization to both services and resources.

A potentially very efficient way to accomplish a resource-efficient way to share the huge amount of data is currently based on peer-to-peer technologies. In order to simplify the integration of these new technologies with existing GRID systems, SIMDAT Pharma has decided to employ the 'file system paradigm'. This means that a module will be developed that is hooked into the respective operating system and that represents the interface to the rest of the GRID system.

This peer-to-peer files system module (P2P-FS) will be based on the peer-to-peer data distribution infrastructure that is being implemented at UKA (called IGOR file system, IGOR-FS). All data written to IGOR-FS will be encrypted and integrity protected. In order to mount and read from IGOR-FS the necessary decryption keys and unique identifiers of the files and folders have to be available. The initial data set for mounting an IGOR-FS containing a decryption key and a unique identifier for the file system root is called henceforth "entry point". The mounted IGOR-FS will support the standard UNIX access rights within the local operating system's domain, i.e. different users on one machine (or local cluster of machines) can be granted different access to the resources (e.g. read-write, read-only, etc.).

The basic course of events, when using IGOR-FS in a Publisher/Subscriber scenario, is as follows (see Figure 11):

- The Publisher wants to provide his data to his Subscribers.
- The Publisher writes his data into IGOR-FS so that it can be transferred to his Subscribers. IGOR-FS provides a confidential and integrity protected communication channel.
- After writing his data to IGOR-FS the Publisher gets some entry point(s) back from IGOR-FS.
- His local authorization policy states which of his subscribed customers is allowed to access which of the entry points (and thus which data).
- If a Subscriber requests to access some data he authenticates to the Publisher and in case the Subscriber is authorized to access the data he gets the associated entry point.
- With the obtained entry point the Subscriber can then access (decrypt) and read the data.

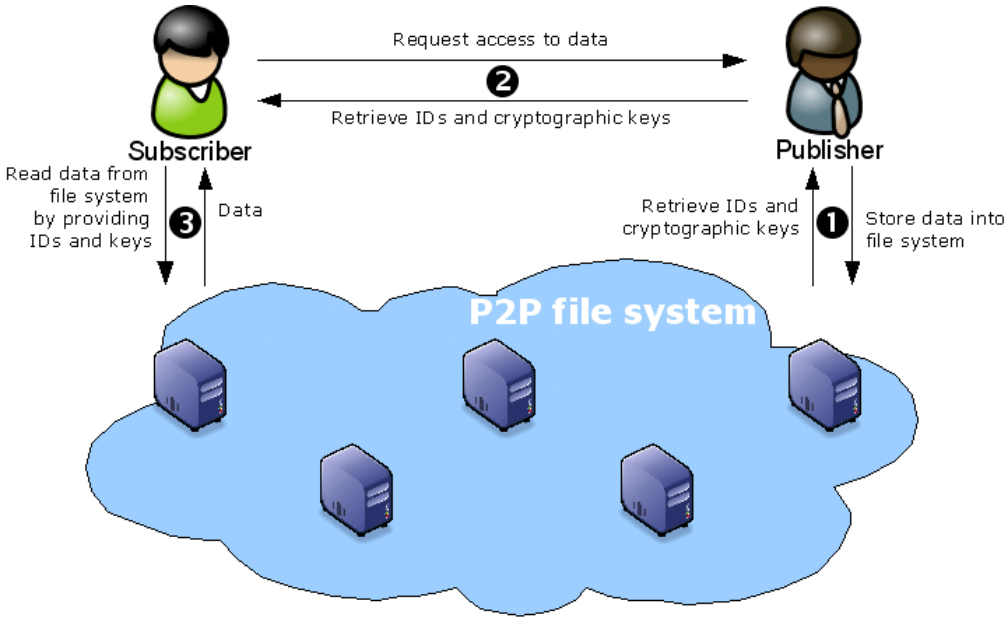


Figure 11: Description of general workflow and Publisher/Subscriber interaction

Although, IGOR-FS supports – among others – the storage and distribution of data in a confidential and integrity protected manner, it does not support the distribution of entry points between Publishers and Subscribers (as necessary in step 2 shown in Figure 11). To this end, the IGOR-FS will interface to the DAC and corresponding components. Only if a user has been successfully authenticated by the DAC and only if the user is authorized to access a certain file (or folder of files), the user will obtain the corresponding entry point and will afterwards be able to see and read file (or folder). As a result, virtual organizations will be capable of accessing their huge distributed data resources as if they were stored locally at each participating node.

The following paragraphs will focus in more detail on the demonstrator concerned with the integration of IGOR-FS with SRS. A first design of the DAC component has been developed according to the P2P-SRS use case and its requirements. The main characteristics and features of the developed design are:

- The exchange of security tokens on a personal basis during contract signing
- The use of this security token to federate different trust domains on a technical level
- A centralized point of administration of relationships
- The local management of users at the each federated side without the interaction of a system administrator
- Privacy issues by not revealing the concrete individuals accessing the services on the service-side

The following security modules are needed for realising the DAC component:

1. Relationship Manager (RM)
   Service which is used to manage relationships between different trust domains. The RM is deployed on both the Subscriber- and the Publisher-side and therefore allows managing access rights on both sides locally in a centralised manner and without the interaction of system administrators.

2. Identity Provider (IP)
   Service which issues, maps and manages identities and pseudonyms. It furthermore provides security token issuing and validation functionalities.
3. Attribute Provider (AP)
   Service that manages attributes assigned to system entities. Attributes can change and be set dynamically.
4. Authorization Provider (AzP)
   Service that guarantees fine-grained access control to data and resources. Dynamic access control decisions can be taken on the bases of the attributes associated to the identity in the given relationship and the requested action.

When subscribing to the service of a Publisher the different trust domains of the Publisher and the Subscriber have to be federated. The basic course of events is as follows (see Figure 12):

- After the contracting phase, the Subscriber and the Publisher exchange their security tokens
- The Publisher and the Subscriber insert an entry on the relationship into their site's Relationship Manager and add the partner's security token to the relationship
- After these steps the Relationship Managers on both sites know about the relationship and the security tokens. A trust relation and federation between these partners have been established.

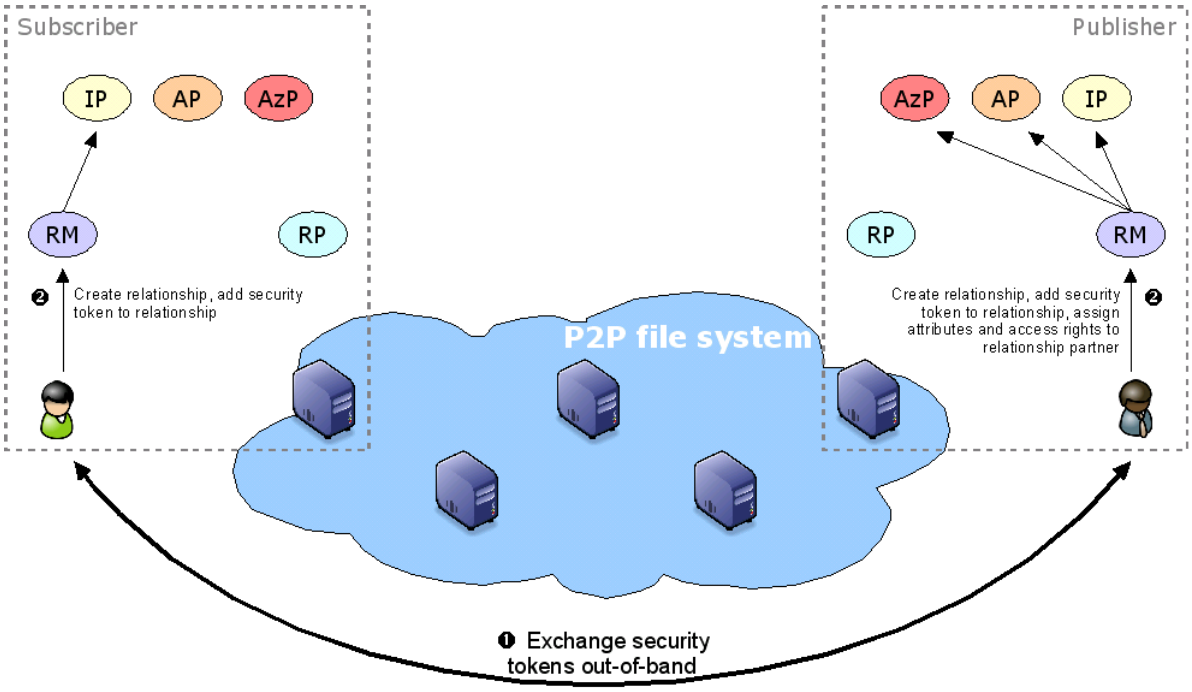


Figure 12: Federation setup – bootstrapping trust

The Subscriber is capable to add local identities to the relationship without the interaction of both the local administrator and the Publisher. The local identities as well as organisation and employee structure stays therefore opaque from the Publisher.

To store and provide data to subscribed customers the Publisher performs the following steps (see Figure 13):

- The Publisher writes data into IGOR-FS.

- IGOR-FS encrypts the data and passes the entry point(s) to the RP along with some meta data concerning the data's category, cost, version, and so forth.
- The RP stores the entry point(s) and associated data accordingly and provides them to authorized Subscribers.

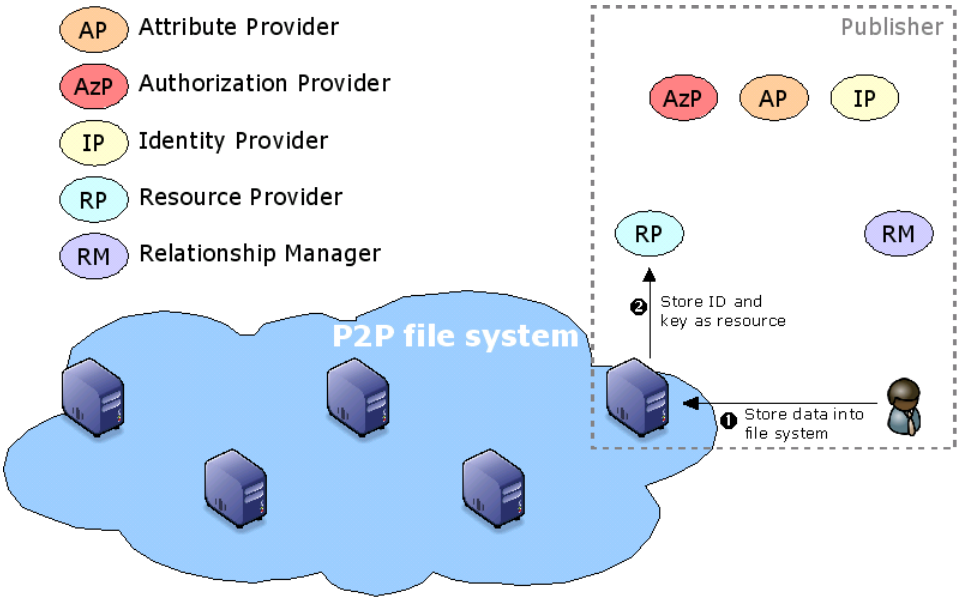


Figure 13: Writing data into IGOR-FS

To read and work with the paid for data, the Subscriber uses the well known SRS portal/interface which will use the IGOR-FS as data source. To be able to mount IGOR-FS the corresponding entry point(s) have to be obtained as follows (see Figure 14):

- The RP on the Subscriber's site requests the RM on the Subscriber's site for a given entry point.
- The RM verifies if the request is authorized and if so gets appropriate assertions to retrieve the requested entry point from the addressed Publisher.
- The RM on the Publisher's site checks the authenticity, the identity and the freshness of the request and then check's if the request is authorized.
- If the request is authorized the RP on the Publisher's site sends the requested entry point to the Subscriber.
- The RP on the Subscriber's site checks in turn the integrity, authenticity and the freshness of the response and uses the obtained entry point to mount the P2P file system.

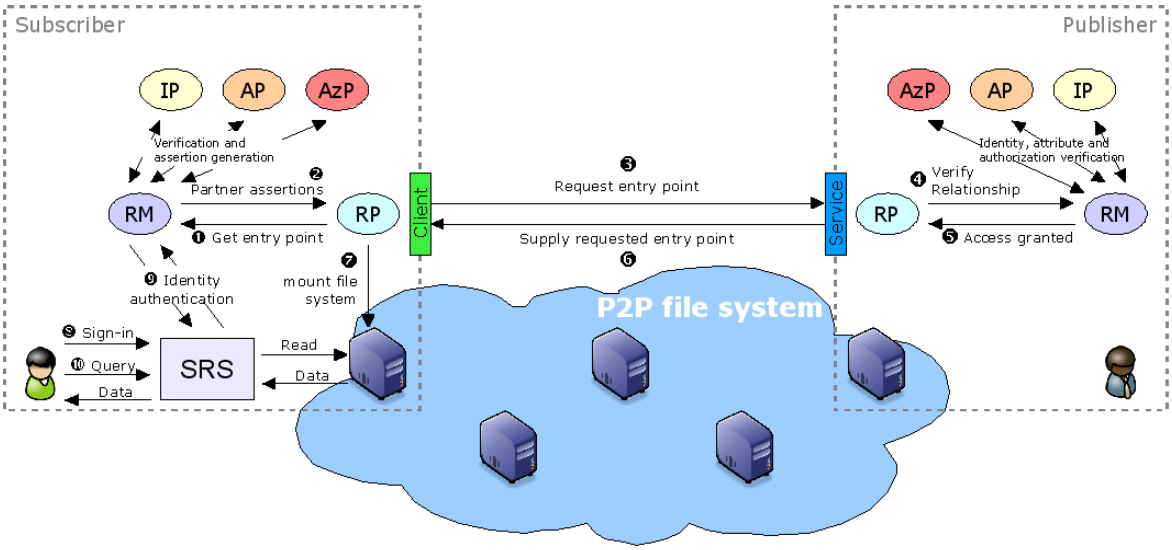


Figure 14: Reading data from IGOR-FS (pull entry point)

Note that the aforementioned pull of entry points can easily be substituted by a push-like behaviour.

During the SIMDAT Pharma meeting on January 24th – 25th hosted by NEC in St. Augustin (Germany) this design has been presented to the attending partners. It appears that the DAC developments proposed and planed by NEC have points of contact with the GRIA development roadmap. In order to avoid similar developments by different project partners, NEC and IT Innovation will be joining forces to implement a DAC module which suffices completely the requirements from the Pharma Activity

3.3.4 Semantic Service Annotation

The ultimate goal of a semantic grid is to retrieve grid services on the basis of their domain-related content. This is additional information that has to be attached to standard grid services in a similar manner as in Semantic Web [23] showcases: Services have to be annotated with appropriate semantic network structures (RDF-based) [21] in order to become "semantically visible" and thus queryable.

WSRF-conformant grid service descriptions must be annotated using terms from the knowledge domain they are targeted for in order for the domain user to retrieve only those services accurately fitted for the required task. Service description data in structured format must therefore be linked to concepts from a semantic network defined in RDF which includes classes for description of biological sources and data types as well as a description hierarchy of bioinformatics tools. All services to be retrieved semantically have to receive a semantic annotation before publishing them. This annotation will have to be done by the service providers. Annotation supplying more details can be added in the course of the life cycle of services. One important promise of semantic grid will be demonstrated: the retrieval of services by service content instead of service name. Interfaces from the annotation tool and repository to the ontology-repository, the grid service provider and the semantic broker will have to be clearly defined and implemented.

The functionality in this requirement is a service query and delivery. The client user (usually a biologist/bioinformatician) defines constraints on the type of data or computation that he wants to use in terms of domain-specific terms relating to biological taxonomy, data type and –if looking for computation services the type of data manipulation required. The choice has to

be based on controlled vocabularies capturing the most prominent concepts in all three subdomains. The resolution of the service must be based the constraints defined by the user.

As preliminaries we state that

- Web/Grid bioinformatics services need semantic enhancement in order to provide the user with domain specific information as the basis for their selection (as opposed to free-text string descriptions or service-internal information that should not concern the user)
- The semantic annotation has to take place somewhere sometime in between setting up the service and its display on a user interface. As a rule, semantics will have to be attached by the domain expert and / or service provider.
- Basic services descriptions should be provided through an application server / middleware in standard format (WSDL or equivalent)
- Semantic additions can either be inlined (as RDF) physically or logically. We do not actually want to modify the original WSDLs, so we have to insert annotations on the way - after service setup but well before the services are presented to the user or brokered by an SOA broker component.
- The additional semantic information should ideally capture everything that is not SOA- or grid specific but everything that the tentative end user is likely to look for - domain knowledge concepts and attributes.

We will extend TUAM [20] so as to be capable of displaying Web/Grid services accessed via some host/port combination and letting them instantiate concepts from OWL domain ontology. The sole additional development necessary within TUAM would then be a read-only structured and easily navigable viewer module for services. For reading in an ontology, an RDF or OWL interface would be required from Ontoprise. In effect the differences between the prototype solution and the current proposal are twofold:

- TUAM reads (only) OWL/RDF and WSDL
- TUAM will not write a proprietary service description file but simply pass on
    - the service names plus
    - the associated RDF class / instance names via HTTP or a database view

Apart from that, TUAM will be positioned just as before, serving as a non-invasive annotation tool for the service provider /domain expert.
Most of what is already there as ontology framework from SRS', from Ontoprise's, myGrid's and from SCAI's side can be reused based on the following observation. A biologist / bioinformatician when looking for a "service" will invariably be interested in being able to set the following constraints:

- taxonomic (which organisms are we dealing with)
- data specifics (are we dealing with sequences, structures, amino acids, nucleic acids, proteins, profiles, mutations, ...)
- tool specific (are we looking for DB queries or alignment, phylogenetic tree, etc. tools)

Using those constraints, queries like the following will be possible:
Search queries for services from the software-inexperienced biologist can then be formulated roughly like this:

- "Give me all compute services for aligning some DNA sequence data from rodents with similar data from humans"
- "Give me all data services (=DBs) for protein structure data from arachnoids species"

We therefore have to extend the classic SOA (service oriented architecture) in a way that leaves its core components untouched but provides semantic extensions.

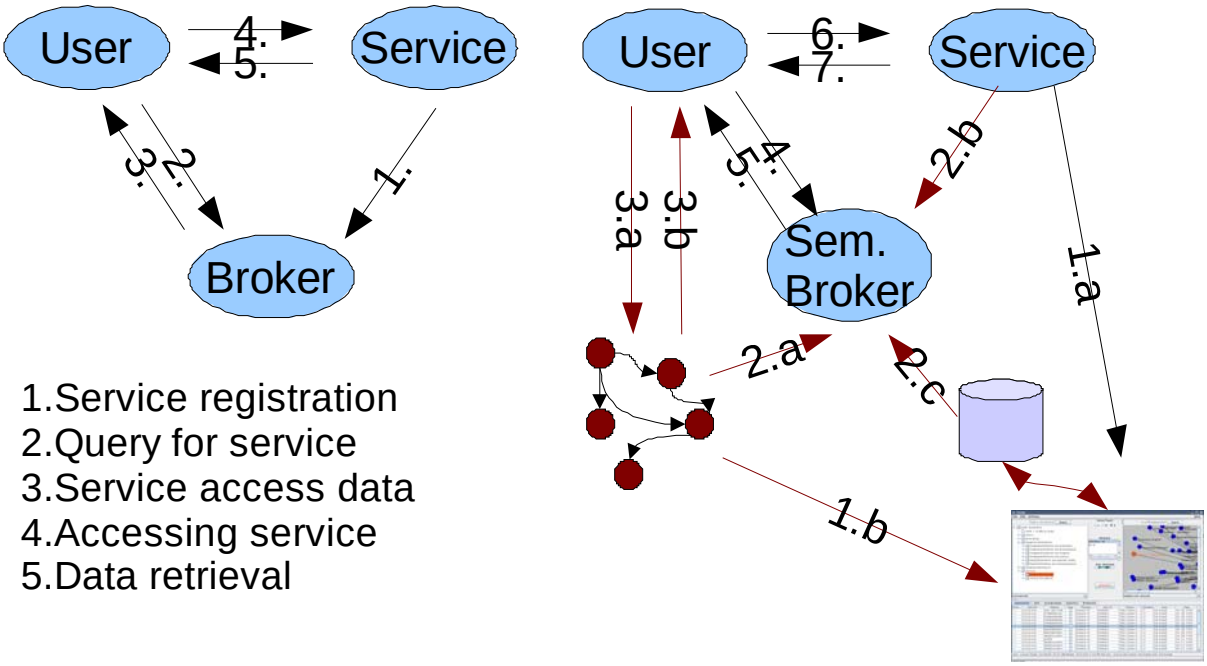


Figure 15 From SOA to semantic grid SOA (sgSOA). The basis of the extended architecture is TUAM, the annotation tool (lower right), that brings together grid service handles and ontology (indicated as a directed graph in brown) class descriptions. Note that the basic SOA is still hidden in this model and can thus represent a fallback solution. Semantics will provide navigation possibilities beyond the classical model.

The data passed between the components in figure 1 can be identified as:

- 1/2.a WSDL
- 1/2.b ontology class names (plus properties + namespace)
- 2.c WSDL IDs+associated class data
- 3./4. class specifications OR traditional service query (plain SOA shortcut)
- 5. service handle
- 6. service request
- 7. service data

For using the semantic retrieval, there is the need for a basic annotation amount to identify all existing services. Successive additions and refinement of semantic detail can go on throughout the lifecycle of the services, enabling ever better semantic navigation. On the other hand a missing semantic annotation means that a service can not be found or rather only found by „conventional“ means via the plain SOA shortcut. Annotations must be provided by the application domain specialists (bioinformaticians, biologists) who should also contribute and understand the underlying domain ontologies. TUAM will have to be installed at the domain specialists' sites, but need *not* be a Grid service itself, only its mappings have to be available for the semantic broker.
Semantic querying is a still unaddressed issue that will have to be implemented in a very basic way and providing the two steps:

1. Coarse navigation in (hyperbolic) concept / class graph or forest (= many inheritance trees)
2. Specific search via attributes (properties) using drop down menus. Individual triple constraints have to be combined by Boolean operators

The following tasks have been identified:

- Redesign of Domain Ontology (possibly split up into BioTaxonomy, ToolTaxonomy and DataTypeTaxonomy
- WSDL viewer for TUAM (early implementation exists)
- OWL/RDF repository (possibly versioned!)
- (OWL/RDF viewer for TUAM exists)
- Redesign of Semantic Broker, esp. concerning its TUAM mapping interface and its user input as RDF resource constraints
- Design and implementation of a semantic navigation interface as sketched here

3.3.5 Knowledge enrichment of SIMDAT-Pharma workflow reports

The envisaged Pharma demonstrator produces annotation of the cDNA sequences that are submitted to the workflow by the scientist. This annotation essentially reflects the contents of the "description" field present in the public sequence databank(s) in which hits were identified by the BLAST procedures (blastn against the EMBL nucleic acids databank or blastx against the protein databank Swissprot). Till now, no effort has been done to exploit the large knowledge present in the complete documentation of these public sequence databanks, notably the cross-references of each entry to PubMed, the Gene Ontology database (GO), the Ensembl genomic databank, and several others. This correlated information is an ideal candidate for an effort in data mining as mentioned in the objectives of WP8.1 of the Annex 1 ("annotation and indexing of heterogeneous resources that are stored on distributed servers"). The data mining procedure would consist of steps like:

- Collect the complete sequence entries corresponding to each significant hit identified by a blast workflow step, as well as their counterparts in the related databanks (protein ↔ nucleic acids)
- Index the documentation of these entries
- Evaluate the frequency of words and co-occurring words
- Collect from their respective Internet (GO, PubMed, Ensembl, etc.) servers the various cross-referenced entries
- Index these datasets and their ontology-based contents
- Produce a statistical evaluation and clustering of relevant terms
- Produce a searchable report, integrated with the raw workflow result

Name	Addition of mapping of relevant biological terms in sequence analysis reports		
Application Activity	Pharma		
Prototype(s)	Pharma		
Date Created	03-16-2006	Priority	high
Created By	Richard Kamuzinzi, Jörg Kindermann	Technology component	Workflow, Knowledge Services
First Implementation Date		SIMDAT module targeted	Information Services
Description			
Sequence analysis reports can be enriched by the addition of mapping of relevant biological terms related to discovered similarities.			
Relation to prototype			
• Each databank similarity search, which produces a ranked list of relevant hits, is submitted to data mining searches in peripherals data sources (Medline, GO, Ensembl, …) • The results enrich the documentation of each entry in the project database			
Requested functionality			

• This data mining product is a module that can be triggered by a corresponding workflow step • A set of specific terms (or keywords) provided by the scientist might orient the data mining process towards his/her specific area of interest. • The report integrates global statistics on terms discovered by the data mining processes.
Validation
• A rate of relevance provided by an expert of a specific domain has to be reached

4 Conclusion

In this deliverable we've presented a validation of the first Pharma prototype. The intention of this validation was to demonstrate the performance gain of a relevant biological work process integrated into a Grid-based environment could have over an ordinary manual approach. The achieved results of this *early prototype* are very promising already now and we are quite confident that the additional deployment of technologies, e.g., the peer-to-peer filesystem and professional workflow tools, will further underline the benefits of the Grid approach in the Pharma domain.

As the activities around the B2B prototype indicate, the Pharma sector clearly puts the emphasis on the development of industrial relevant technologies and scenarios during the upcoming project phase. We hope to demonstrate that specific business cases around the analysis and annotation of selected sets of proteins can be implemented with SIMDAT technologies in a way that the participating partners see commercial benefits from the transactions. This has been explicitly defined as a central success criterion.

Especially for a Bio-tech partner like Inpharmatica this could serve as an ideal platform to demonstrate the mutual benefits for participating partners which could then in turn lead to a wider take-up of this technology also in other business areas.

5 References

[1] Falk Zimmermann et al., Pharmaceutical Scenario Construction, Deliverable 14.1.2 , http://bscw.scai.fraunhofer.de/bscw/bscw.cgi/d33663/SIMDAT_D.14.1.2%20.pdf

[2] Grosof , B. N., I Horrocks, R. Volz, and S. Decker, Description Logic Programs: Combining Logic Programs with Description Logic , in Proc. of the 12th International World Wide Web Conference (WWW 2003), Budapest, HUNGARY, May 2003.

[3] Qu, C. and F. Zimmermann, Application of Standard Semantic Web Services and Workflow Technologies in the SIMDAT Pharma Grid, Position paper presented on W3C Workshop on Frameworks for Semantics in Web Services, Innsbruck, Austria, June, 2005.

[4] Paolucci, M., T. Kawamura, T. R. Payne, and K. Sycara, Semantic Matching of Web Services Capabilities, in Proc. of International Semantic Web Conference (ISWC), Sardinia, Italy, June 2002.

[5] Qu, C. Semantic Broker User Manual, http://bscw.scai.fraunhofer.de/bscw/bscw.cgi/d28001/SB_Manual.pdf

[6] Sun Microsystems, Java API for XML-Based RPC (JAX-RPC), http://java.sun.com/webservices/jaxrpc/index.jsp

[7] Sun Microsystems, Java Web Services Developer Pack (Java WSDP), http://java.sun.com/webservices/jwsdp/index.jsp

[8] The Apache Software Foundation, WebServices – Axis, http://ws.apache.org/axis

[9] Grid-Enabled Medical Simulation Services, GEMSS, IST Project, IST-2001-37153, http://www.gemss.de

[10] PKIX Working Group, IETF. http://www.ietf.org/html.charters/pkix-charter.html.

[11] W3C, SOAP, http://www.w3.org/TR/soap
[12] OASIS, Web Services Security: SOAP Message Security 1.0. OASIS Standard 2004401, March 2004.
[13] S. Anderson et. al. Web Services Trust Language (WS-Trust), Specification, February 2005. ftp://www6.software.ibm.com/software/developer/library/ws-trust.pdf
[14] S. Anderson et. al. Web Services Secure Conversation (WS-SecureConversatoin), Specification, February 2005. ftp://www6.software.ibm.com/software/developer/library/ws-secureconversation.pdf
[15] ITU-T Recommendation X.509. Information Technology – Open Systems Interconnection – The Directory: Public-key and attribute certificate frameworks. March 2000.
[16] A. O. Freier, et. al. The SSL Protocol Version 3.0. IETF Internet Draft, November 1996.
[17] Zdobnov, E. M., Lopez, R., Apweiler R., Etzold, T., *Using the Molecular Biology Data*. Biotechnology, vol. 5b: Genomics and Bioinformatics, C.W. Sernsen (eds); pp. 281-300, Wiley-VCH (2001).
[18] Etzold, T., Harris, H., Beaulah, S., *SRS: An Integration Platform for Databanks and Analysis Tools in Bioinformatics.* Bioinformatics: Managing Scientific Data, ISBN: 1-55860-829-X, pp109-145.
[19] Public SRS servers and databases. http://www.lionbio.co.uk/publicsrs.html.
[20] Kumpf, K., TUAM: A new tool for universal annotation and mediation, Journal of Web Semantics, 2005 (submitted)
[21] RDF Semantics http://www.w3.org/TR/rdf-mt/
[22] Lee, R: Triple Store comparisons http://simile.mit.edu/reports/stores/, July 26, 2004
[23] W3C Semantic Web http://www.w3.org/2001/sw/
[24] JPOX: Java Persistent Objects JDO http://www.jpox.org/
[25] JUNG Java Graph library http://jung.sourceforge.net/
[26] The Lancet on Feb 28, 2006, 367: 668–78 http://www.thelancet.com/journals/lancet/article/PIIS0140673606682652/abstract
[27] Kutzner, K., et al, “Design Sketch of Distributed Filesystem”, Technical Report 2006-6, Fakultaet fuer Informatik, Universitaet Karlsruhe (TH), 2006