



## ***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**

***D10.3.1 Documentation of Software Design, Initial  
Implementation and Underlying Technologies for Final  
Pharma Prototype***

Start date of project: 1 September 2004  
Due date of deliverable: 01/10/2007  
Actual submission date: 24/10/2007

Duration: 48 months

Lead contractor for this deliverable: NEC Europe Ltd.  
Revision: 1.2

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

---

## ***Revision history***

<b>Date</b>	<b>Version</b>	<b>Author</b>	<b>Modification</b>
16.08.	0.1	Falk Zimmermann	Template
13.09	0.2	Changtao. Qu	Semantic Broker
18.09	0.3	Rob Gill	Validation and Evaluation
21.09	0.4	Kai Kumpf	Service Annotation
22.09	0.4.1	David Gee	Validation and Evaluation
25.09	0.5	Falk Zimmermann	Executive Summary
25.09	0.6	Jörg Kindermann	Textmining services
28.09	0.7	Thomas Fuhrmann	IGOR
28.09	0.8	Richard Kamuzinzi	Federated Portal
04.10	0.9	Falk Zimmermann	Proof Reading
06.10	1.0	Nabeel Azam	Grid Architecture of B2A Prototype
08.10	1.1	Falk Zimmermann	Requirements
09.10	1.2	Nabeel Azam	New Diagrams

## ***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

---

## ***Table of contents***

1	Executive Summary .....	4
2	Introduction .....	4
3	Technology Developments.....	5
3.1	Semantic Service Registry .....	5
3.2	Service Annotation.....	7
3.2.1	Integration with Prototype Scenario.....	8
3.3	IGOR File System .....	8
3.3.1	Introduction .....	8
3.3.2	Implementation.....	9
3.3.3	Integration with Prototype Scenario.....	9
3.4	Textmining Service .....	10
4	Final Pharma Prototype - Business to Academia .....	10
4.1	Introduction of B2A Scenario .....	10
4.2	Grid Architecture of B2A Prototype .....	10
4.2.1	Involved SIMDAT Components .....	10
4.2.2	Testbed Deployment .....	10
4.2.3	Role of Academic Partners within B2A Scenario .....	12
4.3	Evaluation and Validation .....	12
4.3.1	End-user Success Criteria - GSK .....	12
4.3.2	Evaluation and Validation - GSK.....	13
4.3.3	Service Provider Success Criteria – Inpharmatica .....	14
4.3.4	Evaluation and Validation – Inpharmatica.....	15
4.3.5	Evaluation and Validation – ULB .....	15
5	Pharma Portal and Federated Prototype .....	19
5.1	Motivation .....	19
5.2	Introduction of wEMBOSS .....	19
5.3	wEMBOSS-SIMDAT Integration.....	21
6	Requirements.....	22
7	Conclusions .....	23
8	References .....	24

---

## 1 Executive Summary

Basing on the design patterns that had been developed during the last project phase major parts of the underlying pharmaceutical workflow, the so-called Master Sequence Analysis Pipeline have been implemented and embedded into the final Pharma testbed environment. This scenario allows GSK to initiate and control the execution of the targeted workflow on their local compute environment. At various steps of the workflow enactment remote resources are being consumed reflecting the increasing trend for industry to outsource functions that can easily be addressed by external organisations. In our particular case we extend the notion of a classical B2B scenario in a sense that we allow academic institution to integrate their resources into such a scenario, where IPR protection of their know-how as well as security constraints is sufficiently addressed. We call this the Business to Academia (B2A) prototype and it denotes final prototype in our application activity.

The current implementation bases on the latest version of the integrated middleware software (GRIAv5.1) and workflow (KDE 5.0). This framework is complemented by the recent version of the semantic service registry (Semantic Broker) and the databank integration tool MRS, which is being used for maintaining all databases that are provided from the academic partners of the testbed.

Recently bio-informaticians are making use of a web-based access to applications in order to learn, evaluate or run analysis services. The number of experts requiring direct command-line access is rapidly shrinking. Reflecting this trend, EMBnet has co-developed a free web portal framework to access EMBOSS-like analysis services. This system is now quite popular in the bioinformatics community, especially in the academics organizations. During the last reporting phase the ULB has initiated an activity aiming at integrating SIMDAT-Pharma technologies into the wEMBOSS framework. This task forms the so-called Federated Portal of the Pharma activity and can be considered as a preliminary stage of the demonstration activity which will be addressed during the last phase of the project.

## 2 Introduction

This deliverable describes the progress the Pharma application activity has made in terms of technical developments in the areas of semantic service discovery and annotation, distributed data access and knowledge services. As most of these tasks are being executed since the beginning of the project we avoid recapping the complete development history and restricting ourselves to describing the advances since the last milestone. We, however, include references to earlier deliverables to allow for consistent and complete coverage of the performed work. With respect to the Business to Academia prototype we concentrate our discussion to evaluation and validation criteria, which play a distinguished role at this phase of the project. This is mostly due the fact that the operation of a Grid testbed has already been successfully demonstrated during the last review meeting. At this point in time the technology has to prove whether the strict requirements on quality and operation are being met. Although problems at GSK (cf. Pharma Activity report) has induced an unexpected delay in the set-up of the final prototype, however, it was still possible for the involved industrial partners to evaluate the existing installation under these constraints, although an in depth evaluation and validation will be postponed to the major Pharma deliverable due at M42.

The developments around the Federated Pharma Portal have been started during the reporting period and complement somehow the work on Grid technology in the Pharma sector. The portal approach offers users a straightforward way to consume resources provided by sophisticated Grid environments without prior configuration of their infrastructure or the installation of specialised client software. Following this path, the involved Pharma partners will further exploit the portal approach within the demonstration activity that is targeted for the end of the project.

### 3 Technology Developments

#### 3.1 Semantic Service Registry

Until PM36, the Semantic Broker (SB) has been released as a software component on the SIMDAT BSCW server, and also deployed in the SIMDAT Pharma Grid testbed as a GRIA service based on both GRIA 5.0.1 and GRIA 5.1 middleware. It has successfully been used by project partners in the SIMDAT workflow toolkit (InforSense) and service annotation toolkits (Dynamo from ULB and TUAM from SCAI).

During the reporting period, the major improvements on the SB functionalities can be concluded as:

1. The migration of the SB to the DL platform is fully completed. The SB is now based on the domain ontology represented through OWL-DL/OWL-DLP, which has also integrated different bioinformatics ontologies such as the Sequence Ontology, myGrid ontology, etc (cf. Figure 1 ). In the mean time, SB's interoperability with different ontology development tools, e.g., Protégé and OWL-S service annotation tools such as OWL-S Editor has also been explored (cf. Figure 2). In particular, for testing SB's support for multiple languages, we have constructed a small SB prototype, which uses Chinese language to represent the domain ontology and corresponding service annotations. The results of above interoperability tests are rather positive; proving that SB can achieve a good interoperability with other DL-based ontologies, DL-based bioinformatics projects, and DL based toolkits.

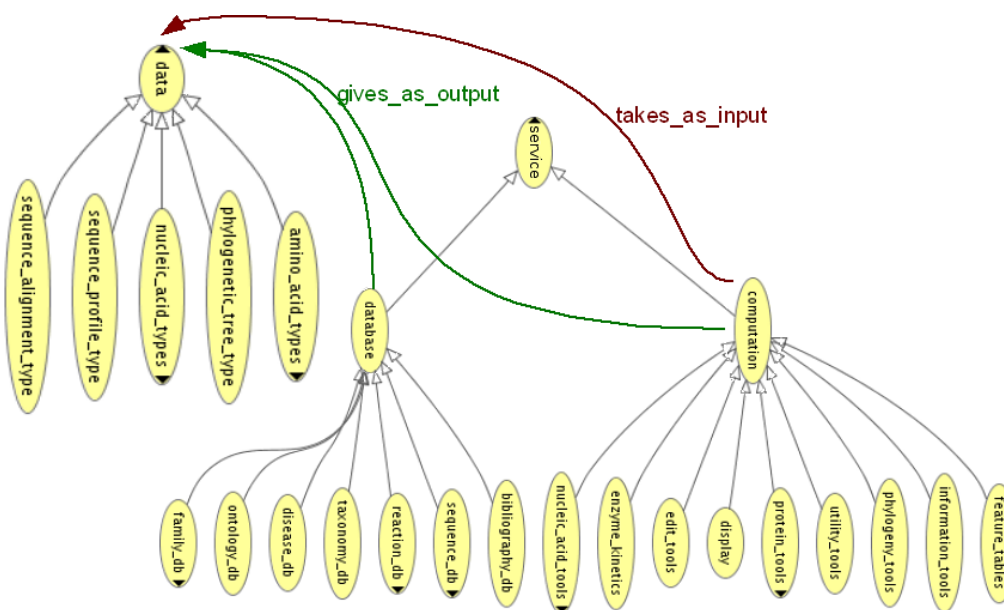


Figure 1: Domain ontology with integration of SO and myGrid ontology

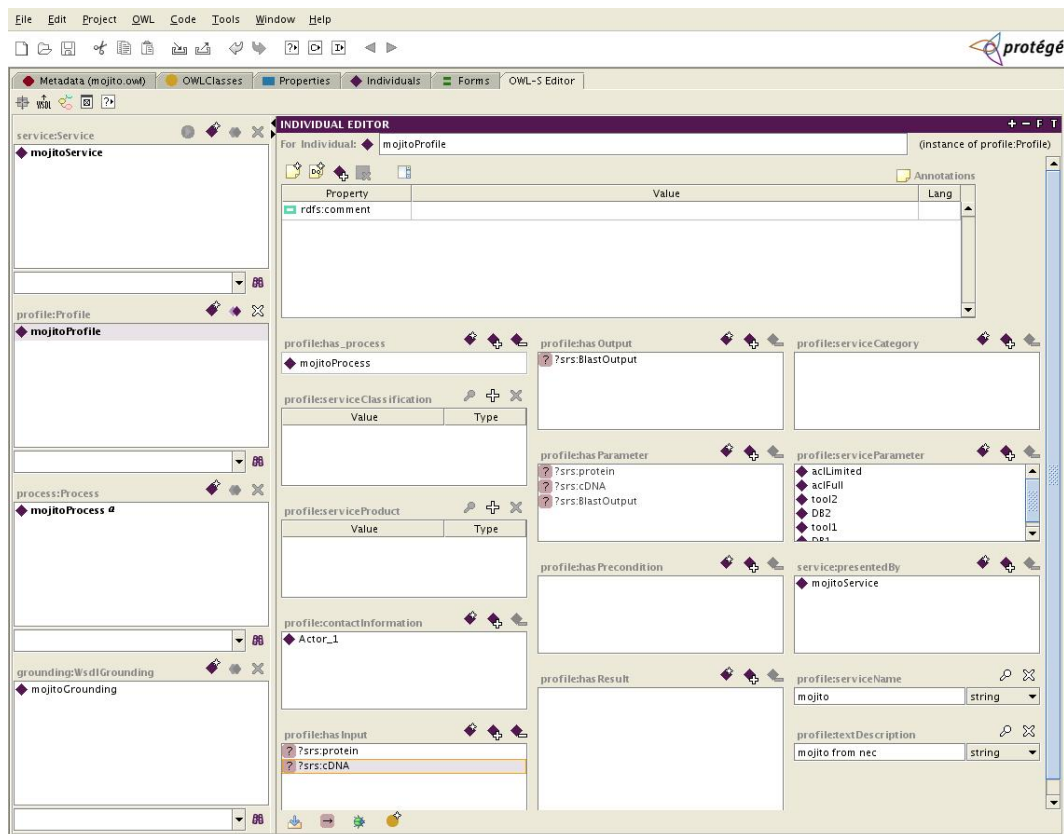


Figure 2: Service annotation using OWL-S editor

[Main](#) | [Check Axis installation](#) | [View logs](#) | [Access control](#) | [List of services](#) | [Atom feed](#) | [Send support request](#)

## PBAC Admin

**Resource: 2c9d19d6-14970a87-0114-970c9874-0001**

Resource status: **ready**

Resource type: <http://www.simdat.org/pharma/gria/SBResourceType>

Access control list for this resource (to have a process role, a user must match at least one *Sufficient* rule, and **no** *Deny* rules.)

Process Role	Rule Type	Match rule	Authority	Action
SB_Admin	Sufficient	Subject DN: Changtao Qu	Issuer DN: SIMDAT Testbed Root-CA [ <a href="#">Download Cert</a> ]	<a href="#">Delete</a>
	Sufficient	Subject DN: Changtao Qu	Issuer DN: SIMDAT Testbed Root-CA [ <a href="#">Download Cert</a> ]	<a href="#">Delete</a>
SB_Annotator	Sufficient	Subject DN: GRIA Application Server	Issuer DN: SIMDAT Testbed Root-CA [ <a href="#">Download Cert</a> ]	<a href="#">Delete</a>
	Sufficient	Subject DN: Joseph Mavor	Issuer DN: SIMDAT Testbed Root-CA [ <a href="#">Download Cert</a> ]	<a href="#">Delete</a>
	Sufficient	Subject DN: Lehong Ding	Issuer DN: SIMDAT Testbed Root-CA [ <a href="#">Download Cert</a> ]	<a href="#">Delete</a>
owner	Sufficient	Subject DN: Changtao Qu	Issuer DN: SIMDAT Testbed Root-CA [ <a href="#">Download Cert</a> ]	<a href="#">Delete</a>

Figure 3: SB server side security setup

- The SB has fully been integrated with the GRIA middleware. The access to the SB is now secured through the GRIA PBAC module, which can authenticate users and sequentially allow them to access authorized services and service operations. In SB we have defined three different user process roles according to the typical usage of the SB, respectively *SB\_Admin*, *SB\_Annotator*, and *SB\_User*. Whereas *SB\_Admin* has full access to the SB service, and *SB\_Annotator* can access *getServiceMatchings*, *publishAnnotation*, *removeAnnotation*, and *retrieveAnnotation*, the *SB\_User* is only allowed to access the *getServiceMatchings* operation for the query purpose. This fine-grained access control over SB service operations can effectively secure the completeness of the service repository by ensuring that only authorized users can modify the service annotations.

---

Based on the PBAC, different users can dynamically be assigned different process roles, which can also further be delegated among users. In Figure 3 we illustrate the current SB server side security setup to achieve fine-grained access control based on the GRIA PBAC module. More details can also be found in WP2.2.3.

3. For the GRIA based SB service, the SB client side APIs are greatly improved, which can simplify the usage of the SB at the client side. In particular, we develop a SB GRIA client wrapper class, which can wrap the invocation details of the SB service, thus can efficiently hide users from the complexity of the GRIA middleware. As shown in Figure 4, by means of the wrapper class, users can use only one line Java code to setup the connection with the SB service. According to the new service annotation entries added, the client side APIs are also improved with regard to the client side query construction support. As a result, although the semantic queries and query responses are all in OWL-S syntax, the users do not need any OWL-S knowledge to deal with the query and query results. For the purpose of demonstrating the client usage of SB for invoking SB functionalities as well as constructing queries, a standalone SB client is delivered, which provides all corresponding Java codes.

```
SemanticBrokerGriaClient gria = new SemanticBrokerGriaClient
("sbclient.state", "https://mojito.ccrl-
nece.de:5443/simdat/services/SemanticBrokerService");
// the file sbclient.state stores client side persistence info.
```

**Figure 4: SB GRIA client wrapper class**

4. With regard to the functionality improvement, a new SB operation *retrieveAnnotation* has newly implemented on request of the service annotation requirements. This operation enables the retrieval of service annotations directly from the service repository. This is of importance for distributed service annotation within the SIMDAT testbed, where both static annotation tools such as TUAM and dynamic annotation tools such as Dynamo are involved.

In addition to above work done, SB's new functionality to support the controlled service discovery based on semantic rules is also under development. We expect to report the research in more details in the next project deliverable.

### **3.2 Service Annotation**

Annotation of GRIA wrapped application services has been introduced as a means to supplement the semantically sparse application wrapper descriptions with controlled terms from domain ontology.

The domain ontology in RDF/OWL possesses top level classes for two principal kinds of service (computational and databases) and data types that serve as in/output descriptions for the individual service types. Generic classes for data and services types are supplemented with concrete instances that serve as the basis for service annotation. In particular, a complete set of relevant bioinformatics databases and –tools have been included in the ontology.

Interfacing TUAM[2] with GRIA 5.1: Because the bioinformatics-applications in wrapped in GRIA do not possess a WSDL interface of their own, a TUAM-viewer for GRIA uses particular GRIA-API calls to access them along with their basic i/o-descriptions, all taken from the handwritten wrapper XML document. Thus, a specialised viewer was implemented to supported opening and display of the application information from individual providers. Migrating from GRIA 4.x to GRIA 5.1 brings with it a change in the API functions that necessitated considerable modifications to the code originally written for 4.x.

---

Interfacing TUAM with NEC's Semantic Broker based on GRIA 5.1: A direct interface to NEC's Semantic Broker was implemented allowing publication of semantic annotations for one provider in OWL-S format which is actually grounded on WSDL services, thus requiring adaptations for the non-WSDL GRIA application descriptions. Annotation is now possible on either one of the following three aspects of a service-wrapped application: 1. Name 2. Input and 3. Output. A complete annotation requires at least the application name being annotated, input and output annotations usually being redundant, because those qualities are already represented in the domain ontology. An annotation project involves one GRIA application listing and the domain ontology and mappings between the two. The annotation project can be uploaded to the Semantic Broker in OWL-S format (when the annotator possesses a valid public key certificate) through the GRIA viewer in TUAM.

Because manual service annotation proves cumbersome when dealing with large numbers of matching concepts, a support for approximate syntactic matching was introduced as a prototype interface that will eventually help improve the sensitivity of semantic annotations as well as the efficiency with which annotations can be done.

### **3.2.1 Integration with Prototype Scenario**

The integration of TUAM, the annotation process and publishing of annotations to the semantic broker involves deployment of TUAM at the sites where the domain experts reside (usually at the service provider site). The annotator has to load the GRIA application list via a secure HTTP URI and the domain ontology from a local file or an insecure HTTP URL. After exhaustive annotation, the SB is contacted, the mappings uploaded and the TUAM annotation project can be closed for possible later reuse. Since the domain ontology provides "dynamic" fields for certain types of service (variable descriptions for databases), Dynamo can retrieve the OWL-S annotations from the broker, add up-to-date information (version, date, etc.) and republish this information through the same channel as TUAM. Since the annotation is basically a one-time process, all further processes involve only the SB and the client applications.

## **3.3 IGOR File System**

### **3.3.1 Introduction**

IGOR-FS is being developed in order to provide an application independent means to share large data files that are frequently updated, but where those updates typically modify only minor parts of these files. The latter is a typical feature of bio-informatics data bases and many other application areas beyond the pharmaceutical industry. IGOR-FS is particularly useful when an application reads only small fragments of data from a large file. Again, this can be expected to be a common case.

IGOR-FS achieves its speed-up as compared to classic file-sharing approaches by splitting files up into variably sized chunks. Chunk boundaries are computed based on the data in the boundary vicinity. As a result, small modifications in a file typically modify only one or two chunks rather than requiring re-writing the entire file. Similarly, reading a small fragment means reading one or two chunks only.

All chunks of a file are identified by a hash that is computed from the respective data content of the chunk. Thus, chunks are independent entities that can be transmitted and cached without the need for any global synchronization. The file system is inherently consistent even though different users can work with different versions. This feature is particularly useful for a gradual role-out of an update. Moreover, storing multiple versions is comparably cheap in terms of disk space because due to its chunked nature, IGOR-FS needs to keep only the modified chunks. Additionally, the chunks are protected by cryptographic keys that are maintained in the file system hierarchy. This allows easy access to the data for authorized users while guaranteeing that unauthorized users cannot view or open these protected files. Different versions may be protected by different keys so that the access to an

---

updated version can easily be restricted. This allows publishers to enforce different licenses models such as monthly subscriptions, etc.

### **3.3.2 Implementation**

IGOR-FS has been implemented in C++ using the POSIX file system specification. It primarily aims at the Linux operating system, but it can be expected to run – with some adaptation – on other operating systems as well. For example, IGOR-FS has been successfully tested under Mac-OS X, too.

IGOR-FS is based on the FUSE framework so that it can be mounted directly into the operating system's file system tree. FUSE directs file access operations from the application to IGOR-FS where they are handled accordingly. Owing to this event driven architecture the entire IGOR-FS system is message based. It can therefore be expected to scale well to the future multicore architectures.

Besides the FUSE interface and the file operation handling engine, IGOR-FS contains a cutting facility that breaks the files up into said chunks, a crypto engine that encrypts and decrypts these chunks using the provided credentials, and a two-level cache that stores these chunks in main memory or on the local disk. Data chunks that are not available locally can be fetched from remote IGOR-FS instances using the IGOR overlay network. Access to remote data is a two-stage process: the first stage uses the chunk's globally unique ID to retrieve a pointer to the actual data. Thereby, it is easily possible to store multiple copies of the data and to use nearby replica.

So far, that is in the release of the previous project year, IGOR-FS was able to mount one static version of a given file system. Data that was written subsequently was not accessible by the subscriber. Only a remount at the subscriber site would have revealed the newly written data. Now, this behaviour has been changed so that a publisher may publish content continuously. To this end, update notifications are automatically pushed from the publisher to the subscribers.

More in detail, this newly implemented PUSH mechanism provides an IGOR-FS instance with updated IDs for a mounted file system when this file system is modified at the publisher site. However, at any time, the local administrator can revert her IGOR-FS to an earlier ID so that the file system reflects the respective earlier version of the data.

The PUSH mechanism is a substantial modification because the previous version was entirely subscriber driven: Data was pulled from the publisher upon read requests by the subscriber. Now the publisher must be able to actively push information to all its subscribers. This means that many concepts within IGOR-FS needed to be adapted to the newly developed PUSH mechanism. Most notably, this required significant modifications in the security system because the publisher must be able to control who may receive updates (i.e. forward and backward secrecy), and it must do so even when there is a very large subscriber set. To achieve this IGOR-FS now supports application layer multicast trees and publisher certificates.

### **3.3.3 Integration with Prototype Scenario**

The intended prototype scenario will use IGOR-FS to provide bioinformatics databases to those machines that run the respective analysis services. Each part of the database shall be mounted from its respective publishing site, for example, the EMBL. Thus all the data will be accessible from any IGOR-FS instance without the need for a prior download. If the data is updated at the published site all the subscribe sites will automatically obtain timely information about this update, so that subsequent analyses use the updated data.

---

### 3.4 Textmining Service

IAIS provides a series of knowledge services to fulfil the requirement KNOWLEDGE-005 „Addition of mapping of relevant biological terms to sequence analysis reports. “ The Web services of IAIS provide the following functionality:

- **Compile a list of keywords obtained from abstracts:** a list of primary keywords is retrieved from the ULB project database of the IXodus workflow. Then a literature search engine (PubMed) is queried for abstracts of relevant publications, which contain (some of) the keywords. The abstracts are retrieved by Webcrawler. Then the abstracts are annotated by the IAIS named entity recognition Web service. A list of relevant and frequent keywords is collected and stored at the ULB project database.
- **Retrain the annotation model by user queries:** this workflow can be used to increment the training data, from which the named entity recognition model for sentence annotation is computed. The IAIS annotation service can determine the plausibility of its annotations. The servers then feeds back the least plausible annotations to the scientific user, who is asked to correct the annotations. Then the annotation model can be re-computed, using the original training data plus the corrected annotations. This results in a personalization effect of the service, and added value compared to the bioinformatics tagging services available on the Internet

The services are described in detail in deliverable D8.3.1. The original plan was to integrate the IAIS text mining services as part of the Pharma IXodus prototype. In the current reporting period it was however decided to integrate the IXodus workflow into the SIMDAT Pharma Portal (section 5). Therefore the IAIS services will also be accessible through this portal

## 4 Final Pharma Prototype - Business to Academia

### 4.1 Introduction of B2A Scenario

The B2A scenario has been already introduced and discussed in the last Pharma deliverable D10.2.2 [5].

### 4.2 Grid Architecture of B2A Prototype

#### 4.2.1 Involved SIMDAT Components

The Master Sequence Analysis Pipeline involves GSK invoking annotation services that are distributed locally and between ULB and Inpharmatica. The remote services at ULB and Inpharmatica are wrapped and deployed as GRIA services. These services are co-ordinated using a workflow deployed at GSK. InforSense will provide the workflow tool and portal used by the end user to execute the MSAP. The GRIA client is installed at GSK and used by an administrator to manage GRIA conversations and monitor usage.

#### 4.2.2 Testbed Deployment

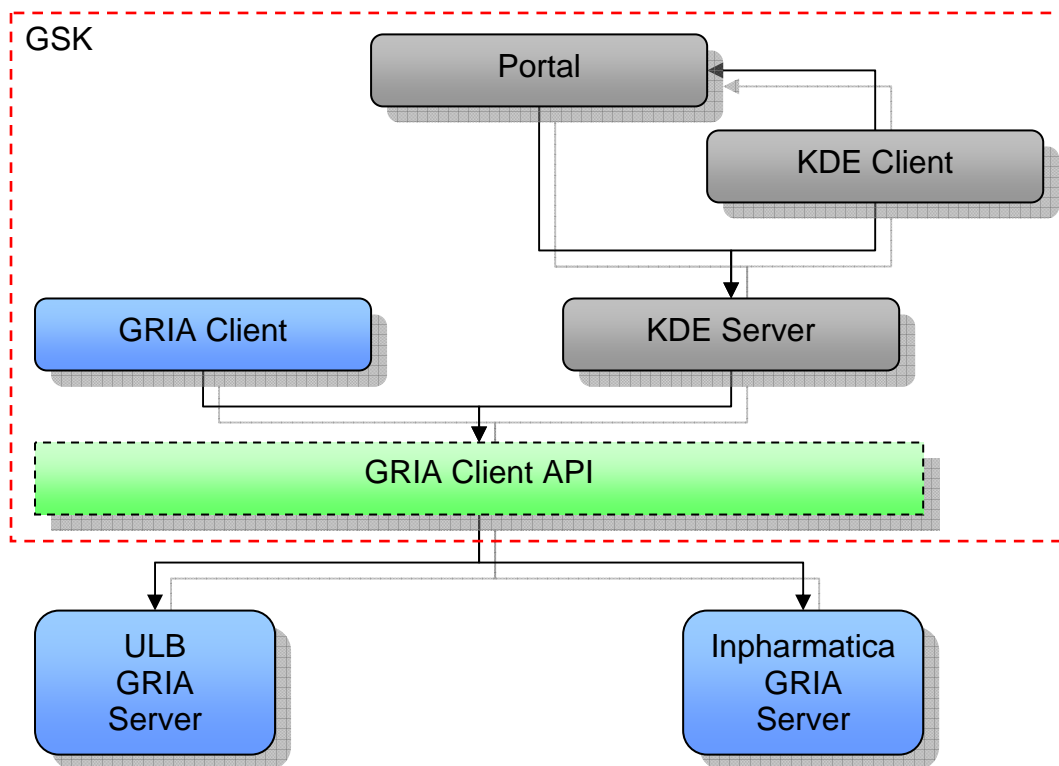
To make their applications accessible to GSK, Inpharmatica and ULB both have installations of GRIA servers that host their annotation services. To communicate with these servers there is an installation of a GRIA Client at GSK. The Client application is only accessible by an administrator and is used to set up and manage authentication and authorisation to external services from GSK. Due to stringent security policies at GSK, access to the ULB and Inpharmatica services had to be specially granted through a secure proxy.

The workflow consists of 3 layers: server, authoring client and portal. For this scenario a single InforSense server installation is used. This is responsible for making the GRIA calls to the remote services and is therefore located on the same machine as the GRIA client, accessible by an administrator. Both the GRIA Client and InforSense Server access the same properties and

configuration files to allow the administrator to monitor and interact with the service calls made from the workflow server. The administrator provides a list of available GRIA servers that can be accessed by end users within GSK.

Multiple workflow clients can be used within GSK to access the InforSense server. These client applications are in essence a visual programming tool used by workflow authors to co-ordinate service calls, as used for the MSAP. To construct the workflows using the GRIA components, the application metadata for each of the GRIA services are required. Web Service calls are made at authoring time to retrieve the metadata, which is then used to parameterise the GRIA components for use within a workflow. These workflows are then deployed and published, exposing certain parameters and outputs for use by domain experts.

The portal is accessible by an administrator to manage the InforSense installation, configure the GRIA settings and assign roles to other users, i.e. workflow author, portal user. Domain experts can also access the portal, however they have access to less functionality and are only allowed to browse and execute workflows published as services. Making the services available in this way means the domain experts are not required to have any knowledge about the underlying workflow or GRIA.



**Figure 5: Interaction between components**

Figure 5 shows the Interaction between the components used for this test bed. The red box indicates the components located within GSK. An administrator uses the GRIA client to manage the connection to the services at ULB and Inpharmatica. The KDE Client is used to author the workflow and submits the workflow to the server for execution. The Client can also be used to publish workflows as services accessible from the portal. An administrator can use the portal to configure the GRIA settings within KDE and specify which service providers can be used by the other users. The portal can also be used by domain experts to execute published services which again, go through the KDE Server.

---

### 4.2.3 Role of Academic Partners within B2A Scenario

The role of academic partners within the B2A scenario is to provide free academics services and manage them using SIMDAT technologies. ULB will make sure that the services are semantically described and published to the semantic broker. According to the master sequence analysis pipeline (MSAP) defined and suggested by GSK, ULB has currently identified three EMBOSS analysis services that are of relevance. The three services are successively *antigenic*, *coderet* and *ePrimer3*.

The three services have been chosen to support the *primers/peptides* annotation part of the MSAP. In a subsequent process, the services will be managed using the GRIA's service level agreement (SLA) in order to negotiate and fulfil specific execution requirements of an industrial partner such as GSK.

The deployment of Grid technology is of crucial importance in such an environment as it is a means of guaranteeing a high degree of resource availability as it has been discussed in earlier Pharma deliverables. Currently public provisions of non-commercial resources from public institutions like the EBI have, due to a variety of reasons, periods where their servers are not on-line. Clearly, industrial end-users cannot take the risk having no access to resources they require in their daily work and hence academic institutions could only play a minor role here. In addition, industrials impose a certain degree of quality on the resources they integrate into their analyses. Thus they are particularly interested in having easy means to check the quality of external resources. The developments around the Federated Portal reflect this requirement appropriately and therefore complements the B2A scenario. Here, end-users have the possibility to upload their test-data to the portal and evaluate the implemented services with them. Depending on the generated results the end-users can decide to integrate, for example, the service directly into their targeted workflow, bypassing the portal.

## 4.3 Evaluation and Validation

In this section we outline the major success criteria coming from the industrial (pharmaceutical) end-users of our final prototype system. Since the central goal of the SIMDAT project is to bring Grid technologies to the targeted application areas, these criteria have to be sufficiently addressed to assure their uptake beyond the life-time of the project

### 4.3.1 End-user Success Criteria - GSK

GSK's involvement in SIMDAT is primarily focussed on supporting the development of technologies to support the goals of Virtualisation and Globalisation of the organisation. Fundamentally this requires the development of tools that are:

1. Usable by scientists within GSK with minimal IT support
2. Flexible that they may be able to change and scale to the changing environment
3. Able to lower the barriers for working with external partners in a virtual organisation
4. Compliant with GSK's stringent security policies
5. Responsive enough to be usable in a timely fashion
6. Create options for the reduction in support currently necessary for Biological and chemical analysis

#### Usability

Scientific analysis and eventual Target Discovery requires the ability to integrate, collate and derive knowledge from disparate data types. A successful implementation of SIMDAT technology would allow laboratory scientists to directly generate analysis workflow and pipelines outside of the current GSK software development process. A framework of tools where scientists could integrate with a visual scripting language to gain access to data and analysis algorithms wrapped as independent services which can be run either internally or externally without high levels of IT involvement delivering valuable analysis pipelines would be seen as a successful application of the technology.

#### Flexibility

Target Discovery is a rapidly changing environment, a successful SIMDAT architecture must allow for rapid development of new workflow and knowledge management processes. This must include

---

access to both internal and 3<sup>rd</sup> party applications. It would be expected that use of the technology would reduce the development time and cost of prototype systems by 50% and make the transition from prototype to production a simpler and more complete process. We focus here on prototype systems purely to allow identification of real value and due to the longer term goal of changing GSK process for production development.

### **External partner access**

Current procurement and access policies create a barrier to rapid technology development and can lead to sub optimal choice of vendor solutions. SIMDAT technology should demonstrate the ability to make vendor and academic data more available for validation and analysis. Also SIMDAT needs to allow for the development of new paradigms for the relationship between Pharma and 3<sup>rd</sup> parties, lowering the boundaries for interaction. Ideally demonstrator data sets should be readily available to Pharma for analysis by real data controlled and secured appropriately.

### **Security Policies**

Absolutely essential to the success of the project is the ability of the SIMDAT architecture to pass the rigorous security requirements of GSK. All components must be compatible with the processes in place for both outbound and inbound transactions.

### **Responsiveness**

A balance needs to be found between responsiveness to user queries and flexibility within the system. User acceptance testing will be carried out using both fully internal systems and those implemented using grid enabled services. We expect a reduction in responsiveness in the grid systems but wish to test if this is acceptable by the user base.

### **Reduction in support cost**

GSK's current support burden per application are considerably. One major benefit of SIMDAT will be the opportunity to reduce this cost by outsourcing this support to vendors and Academics. The SIMDAT architecture must be seen as fit for purpose to support this requirement and will be evaluated by architecture groups at GSK for this.

## **4.3.2 Evaluation and Validation - GSK**

The SIMDAT project has been an incredibly interesting and valuable experience for all involved. Learnings on GRID technology have been very valuable and contacts made and solutions installed due to SIMDAT have been very valuable. Biggest downside of SIMDAT has been the excessive needs for documentation which may well temper any future involvement in EU projects. That's said as described below we feel the overall requirements and deliverables have been achieved.

To the requirements of usability, our implementation of InforSense as our client facing application as part of SIMDAT has been very successful. Workflow development by non IT staff have delivered a number of primitive but valuable workflow and the Master Sequence Analysis workflow demonstrator has been validated by the business. We continue to monitor the generation of business value with the use of InforSense and this has come back with very positive figures. The wrapping of GRIA nodes into simple analysis workflow and their deployment as standard analysis nodes has also been very successful allowing users to chose between in-house and external analysis services with a minimal learning curve. Further developments around workflow discovery and use of the semantic broker would be both interesting and valuable.

Flexibility of the SIMDAT architecture has again been successfully demonstrated within the prototyping area of IT development. Prototype cycles have been reduced significantly with business partners receiving usable prototypes far faster than previously managed. The reduction has been in the

---

region of 30-40% and we perceive this to reduce further as we gain access to more (reusable) components accessing both GSK and external sources.

The Partner Access paradigm has been one area of incredible movement within SIMDAT. The current “all or nothing” model has now been replaced by one where vendors and academics can control access to their data and analysis in a very detailed manor using the GRIA framework. Vendors can control access to their systems in a very flexible and granular manor allowing relationships around actual data sharing and analysis to occur then develop into more structured formal relationships.

Security issues have been a constant problem within the project, however scaling back to outbound access from GSK in the first part has lead to an acceptance of the tools and the demonstrations to be successful. We will continue to review inbound access from 3<sup>rd</sup> parties as we progress and hope to have agreements on this in due course.

User acceptance on responsiveness is still ongoing. As expected we have seen a reduction in response time using grid tools; however the flexibility gained using the SIMDAT tools is seen as very valuable. Current trends suggest user acceptance will be positive and will continue to improve as internal software is modified to fulfil the virtual requirements.

Reduction in support is also an ongoing discussion. The technologies SIMDAT bring to the table for GSK are certainly interesting and we have spawned off a number of future project opportunities to test this outsourcing of support. Architectural approval and addition to the roadmap are needed prior to any full implementation but the demonstrators and training opportunities will be useful in this endeavour.

We feel this has overall been a successful project implementation and a very encouraging opportunity for adoption by GSK architecture for its technology roadmap. There are still a number of improvements and enhancements needed to the technologies; however these can be seen as future projects.

### **4.3.3 Service Provider Success Criteria – Inpharmatica**

Inpharmatica's role has primarily focused on the delivery of services to the other pharmaceutical partners. Our perspective of the SIMDAT project is that it provides a framework which can provide a secure, robust and low overhead method of providing our services to an electronic market place.

This requires the following features:

#### **1. Accessible**

In order for this approach to work it has to be accessible to our customers via a standard method that they are used to and confident doing business with.

#### **2. Secure**

With the possibility of financial transactions and the exchange of extremely sensitive intellectual property. It is essential that the transactions can be shown to be highly secure.

#### **3. Accountable**

From an accountancy perspective the system will be a set of business transactions thus requiring all of the associated auditing for both customers and providers.

#### **4. Stable**

A key requirement for any system is that it is stable and fails in a known and manageable way. In an online market place stability is a key factor of success. Even limited downtimes can be enough for customers to find another supplier.

#### **5. Inexpensive**

---

As far as a Business is concerned this provides an opportunity to enter a market however due to the slim margins in the Pharmaceutical sector the overheads on any system have to be limited.

#### **6. Simple Integration**

From a technical perspective the system will have to integrate to a diverse set of back end systems and thus will require flexible and simple APIs in order to achieve this.

#### **7. Scalable**

If successful a simple service will grow therefore it is vital that the architecture can support this growth without significantly changing the services.

### **4.3.4 Evaluation and Validation – Inpharmatica**

The SIMDAT project has provided an architecture that has previously been impossible to attain. Fundamentally this may have been due to no specific software solution having a large enough market share alternatively it could be perceived that the concepts and software had not reached a mature enough stage. The experience of SIMDAT project has shown that we now have mature enough software and the very accessible software has allowed the partners and others to try and use the architecture. Overall the requirements and deliverables have been met.

Using the Web provides a highly accessible medium, added to this the banking grade encryption provided by GRIA/NEC the SIMDAT platform provides a highly accessible and secure environment. GSK have been using this environment with InforSense to successfully to run jobs. No specific integration has been required from Inpharmatica. From a service provider perspective this has been very successful as it has enabled us to provide a service to GSK by just providing details of our GRIA service. In principle this now gives us the ability to sell our service to anyone with InforSense. However to provide a more professional service significant effort would have to be placed in the usability of the clients general account creation and management facilities. These facilities are provided with the GRIA Java front end however their usability should be addressed in future releases.

The system has been extremely stable and an uptime of >90% is easily achievable. The costs to run a GRIA server are relatively modest in contrast to the costs to run the service. One of the key issues has been the learning curve (costs) with the complexity of installing and administering the service. This has greatly improved since version 4, but the added business options do greatly increase the complexity of the installation. we would suggest that this is an issue that should be addressed in future versions.

Once the service is running the time diminishes and the week to deploy a service is a lot less than the 18 months it would take to implement a system with these features. Integration is relatively painless and after the initial learning curve we have found that a simple service can be deployed in a few days.

One of the features this has highlighted is how exposed the back end service is to being overloaded by the front end server this is vital to enable the systems to scale both in size at the back end and provide responsiveness or provide reasonable customer expectations to the front end.

### **4.3.5 Evaluation and Validation – ULB**

From ULB perspective, SIMDAT technologies have been evaluated according to three requirements defined for the third year project. The first requirement is entitled “from workflows to GRIA applications”, the second is “to expose arguments of GRIA wrapped applications in a Workflow

authoring tool” and finally the last requirement is “customised management of GRIA-based service offering”.

## 1. From workflow to GRIA applications

Motivation: whenever a workflow is designed, we need to access it using different interfaces. The web interface is proven to be very useful for basic and inexperienced users but a programmatic interface is also required. According to real life cases, a workflow is considered as a new application and thus subject to be wrapped as a GRIA service. The added value of this approach is to make developed workflows being callable from any other workflow or application.

This requirement has been tested using the deployment mechanism provided by InforSense KDE. It has proven to be very easy and useful when applied to data flows. The first workflow we tested is the main analysis part of the IXodus knowledge discovery process. It has been successfully transformed into a command-line application and further wrapped into a GRIA application.

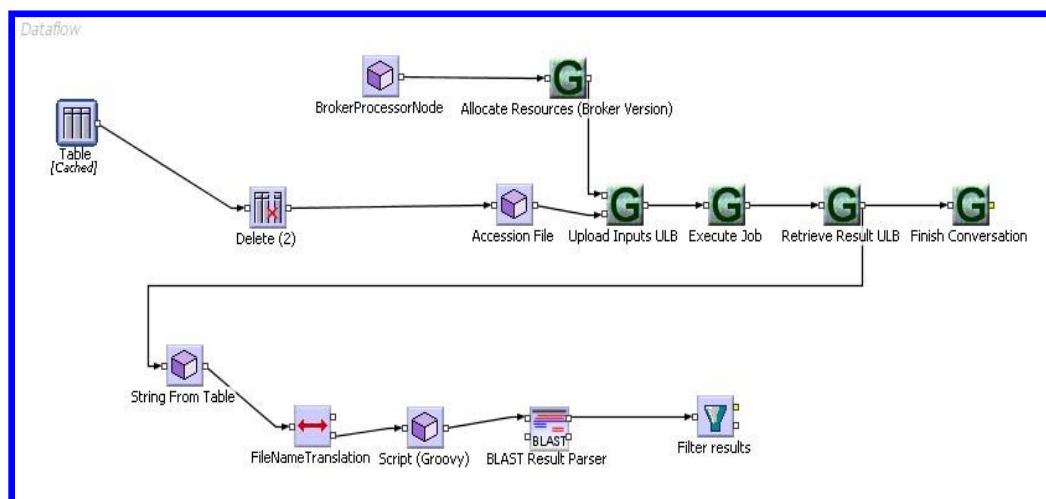


Figure 6: Main analysis sub-process of the Ixodus workflow

The second workflow we have tested, the IXodus process itself, appeared to be very difficult to deploy. At the current stage, we believe it is mainly due to the recursive usage of many control flows. Actually, the operation of promoting a given parameter at the workflow level is hampered by the presence of nested control flows. At the current stage, we are holding discussions with the technology provider in order to address the difficulties we have experienced.

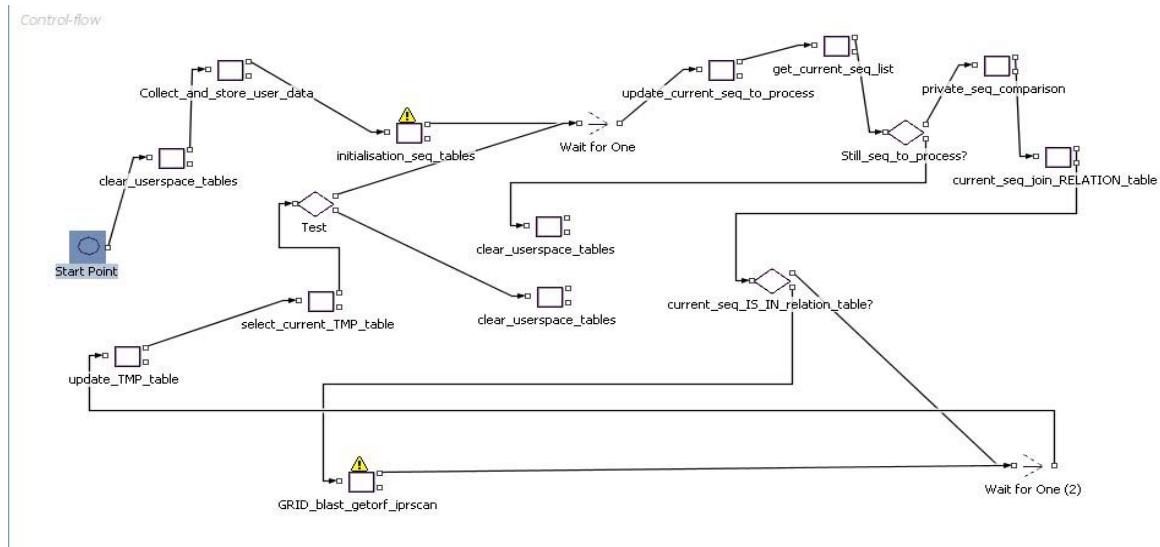


Figure 7: IXODUS Workflow

## 2. To expose arguments of GRIA wrapped applications in a workflow authoring tool

Motivation: to ease the workflow designer task of creating workflows, analysis process authoring tools such as InforSense KDE provide, for each application node, a way to visualise and adjust parameters. Since SIMDAT integrate GRIA wrapped applications, this feature should be similarly applicable to GRIA nodes. Actually, from user experience point of view, it does not make sense to get different behaviours of the workflow authoring tool environment when accessing GRIA-based applications, just as when accessing non-GRIA applications nodes.

In order to enable parameters exposition in a workflow authoring tool, ULB has identified two different approaches. The first approach is to extend the GRIA meta data description of deployed applications in order to include all required and optional parameters that need to be set for an effective execution of the service. The second potential solution is to take advantage of fine grained semantic description of available services in the Semantic Broker (SB). In other words, we need to choose between a GRIA-centric or a SB-centric solution. ULB has finally retained the GRIA-centric approach due to the fact that compared to the GRIA technology, the semantic technologies are such complex to manage and master. Thus, if any partner decides, for any reason, to avoid the implementation of the semantics web capabilities of SIMDAT, there is still some room to discover applications parameters/arguments thanks to the GRIA Job Service interface.

To fulfil this requirement, the GRIA partners have suggested an extension mechanism to their application meta data model. With this extension, we could now integrate for instance the EMBOSS language for parameters/arguments description. This language is known as ACD (Ajax command line description language).

---

```

<?xml version="1.0" encoding="UTF-8" ?>

<GriaApplicationDescription
xmlns="http://www.it-innovation.soton.ac.uk/2007/grid/application">

<JobServiceMinVersion>5.2</JobServiceMinVersion>

<Application>
<Description>Blends two images together</Description>

<ApplicationName>http://it-innovation.soton.ac.uk/grid/imagemagick/blend42</ApplicationName>
<ApplicationVersion>2.0-1</ApplicationVersion>
<Group>graphics</Group>
<Keywords>imagemagick, example</Keywords>
</Application>

<DataStagers>
<DataStager type="input" name="inputImage" minOccurs="2"
maxOccurs="2" defaultSize="2">
<Description>Input images to be blended
together</Description>
<MimeType>image</MimeType>
</DataStager>

<DataStager type="output" name="outputImage">
<Description>Resulting blended
image</Description>
<MimeType>image</MimeType>
</DataStager>
</DataStagers>

<SomeCustomElement xmlns="http://www.example.com">
<SomethingElse>
Anything is allowed here!
</SomethingElse>
</SomeCustomElement>

</GriaApplicationDescription>

```

**Listing 1 Simple job service application meta-data example.**

### **3. Customised management of GRIA-based service offering**

Motivation: An EMBnet service provider such as ULB-BEN (Belgian EMBnet node) has to be able to customise the service offering, to comply with the internal management policy. Initially, ULB has identified three kinds of users' profiles; the candidate, basic and the advanced users profiles. The candidate user profile refers to a scientist who requires a little amount of resources to launch some services and evaluates them along predefined attributes of QoS. We need to allocate a reduced amount of resources to the candidate user as he/she will often interact with the system anonymously. The basic user profile is associated to a registered user who accesses the complete set of free services available at the BEN. Finally, the advanced user profile is, in the same way, applicable to a registered user who is allowed to access free as well as commercial applications portfolio.

To enable such a resource management policy at BEN using SIMDAT technologies, we have been testing the creation and the management of two Service level agreement (SLA) profiles thanks to the SLA management service shipped with the GRIA version 5.1. We, actually, deployed the BLAST service as an example using two different SLA profiles. The first profile, a basic one, is limiting the

---

user to access 1CPU, 500 CPUs per day and the disk space allocated has been affected to 100MB. The second profile, which is more demanding in terms of resources- 4CPUs, 1000CPUs per day and 300MB of disk space – has been conceived to test a different policy and observe how the system reacts. We have used the GRIA client application to interact with the two SLA profiles and it has been working very fine. With respect to the evaluation we have been carried out, we do not see any difficulties to apply this technology in a production environment.

## **5 Pharma Portal and Federated Prototype**

### ***5.1 Motivation***

Today, bioinformaticians mostly prefer a web-based access to applications in order to learn, evaluate or run analysis services. Only few expert users require direct command-line access. Considering this user requirement, EMBnet and especially both the Belgian EMBnet node and the Argentinean EMBnet node have co-developed a free web portal framework to access EMBOSS-like analysis services. This system is now quite popular in the bioinformatics community, especially in the academics organisations.

To enhance the usability of the current status of the Federated prototype, ULB have been developing the integration of the SIMDAT-Pharma technologies into the wEMBOSS framework.

### ***5.2 Introduction of wEMBOSS***

A few years ago the EMBnet nodes adopted the EMBOSS suite as their main sequence analysis package. Professor Marc Colet (manager of the Belgian EMBnet node) devoted himself to the development of wEMBOSS, a web interface to the command-line based EMBOSS applications. While Web interfaces tend to be featureless (they keep no record of what the user did before) or otherwise store results of previous program runs in some temporary or custom area, he conceived the idea of storing results in (and taking input data from) a classic UNIX home directory, so that users can work on the same data over the Web as well as in a terminal session. Remained the challenge to generate HTML pages with boxes and selectors from the ACD files (each EMBOSS program has an ACD file that defines in a precise syntax the input data and other information needed by the program). For this purpose, code was borrowed from Luke McCarthy's "EMBOSS GUI". Note by the way that wEMBOSS parses the ACD files on-the-fly, so that changes in the EMBOSS installation are recognised immediately, without need to re-run some installation script.

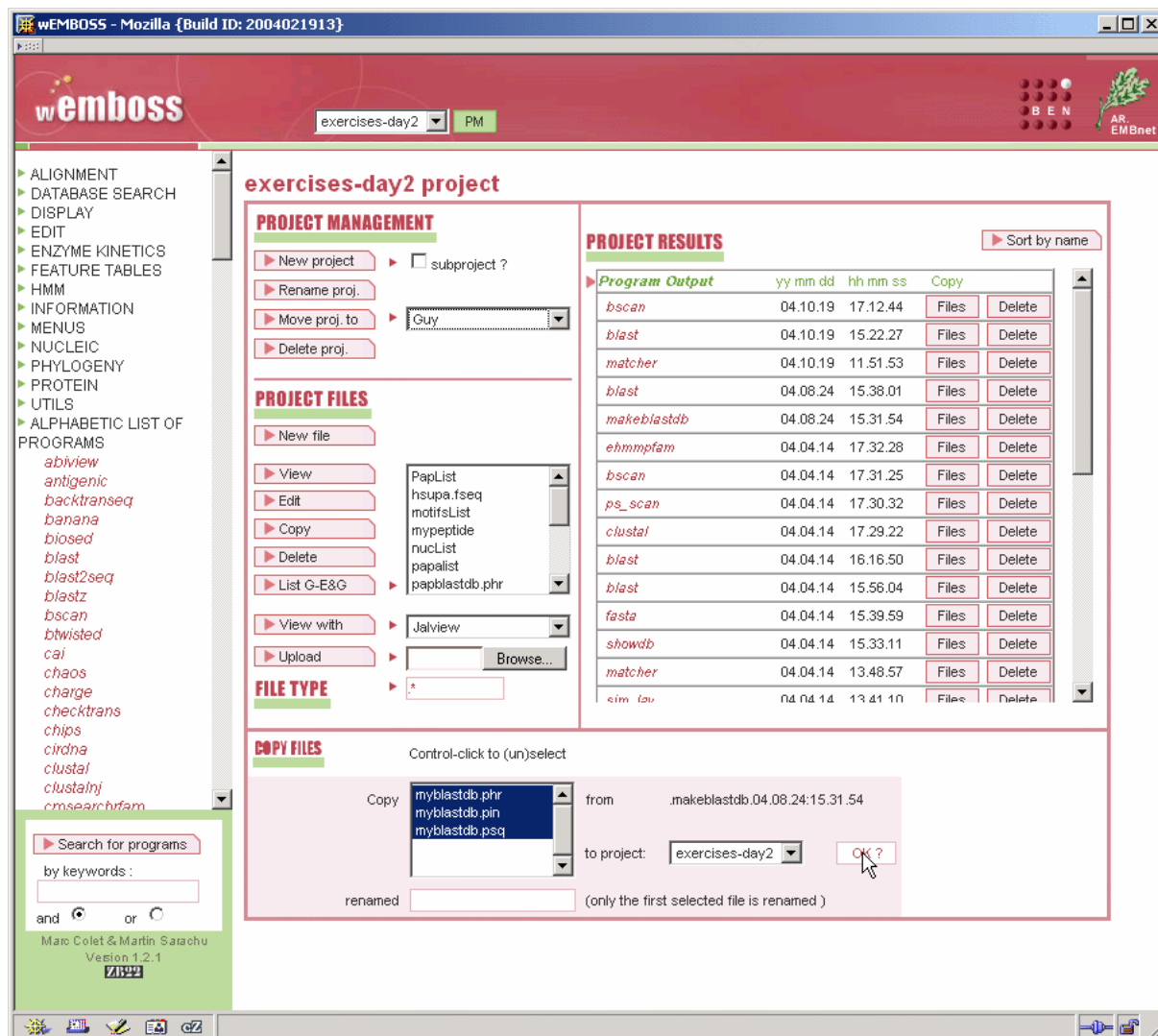


Figure 8: Project Management in wEMBOSS

JavaScript was added to the pages to make them "alive", letting elements appear/disappear and default values change, so that when the user moves from top to bottom through the page while making his selections, he will be prevented from making inappropriate choices. While developers of GUI's sometimes complain about the complexity of ACD syntax, M. Colet made it a point of honour to make sure that wEMBOSS handles all the formulas correctly.

End 2003 a Web design company was hired to give the interface a look more in agreement with modern standards. In the same period, Martin Sarachu from the Argentinean EMBnet node joined the wEMBOSS development team. A Web site <http://www.wembooss.org> was set up and in May 2004 the first non-field test release was made available for public download.

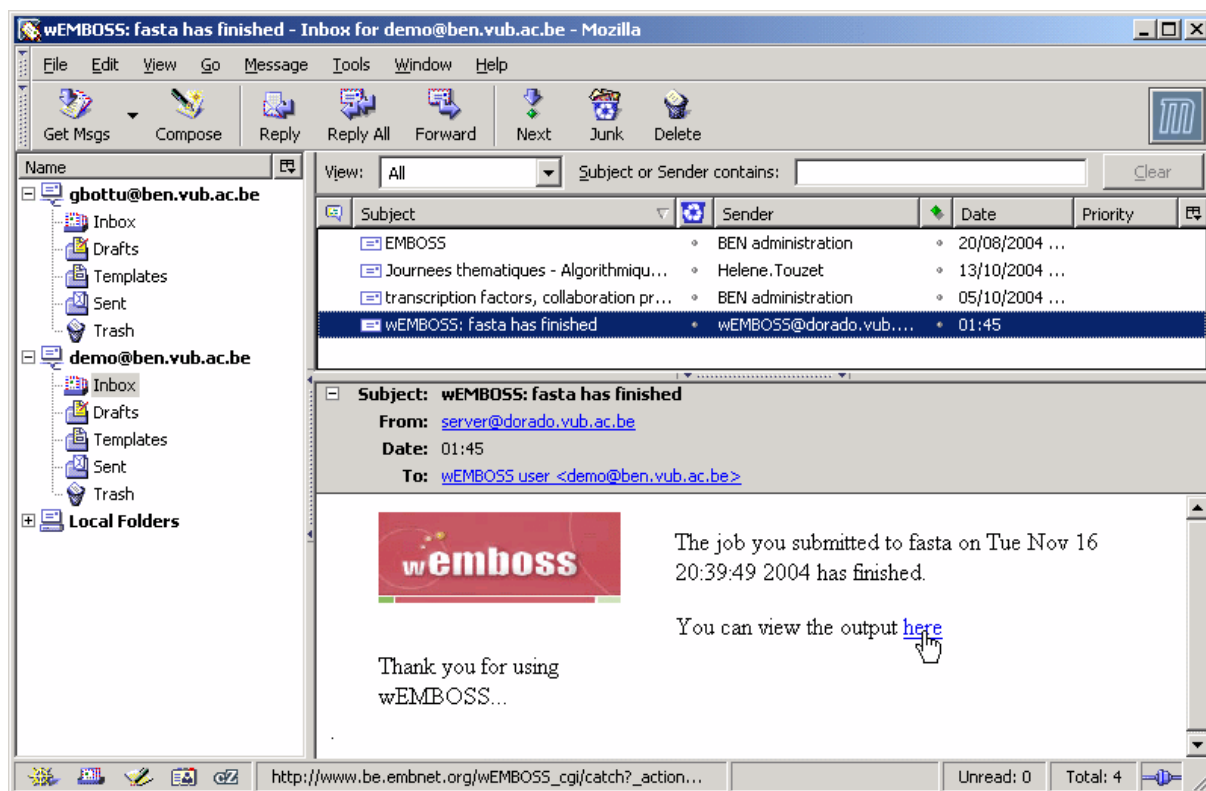


Figure 9: e-mail Notification for Analysis Results

### 5.3 wEMBOSS-SIMDAT Integration

To progress from the Federated Prototype to the Federated Portal, we need to integrate SIMDAT-Pharma technologies and the wEMBOSS framework. Because wEMBOSS generates web pages on basis of ACD files, ULB has adopted an integration strategy that is ACD-centric. It means that any analysis service available as GRIA service should be exposed as a regular EMBOSS program. The EMBOSS API enables to deploy new applications and expose them as EMBOSS program by first creating the corresponding ACD file and then writing an emboss wrapper program. From wEMBOSS point of view, whenever an ACD file is available in the ACD's repository, it is dynamically transformed into a web interface in order to interact with the corresponding analysis program. As an example, the BLAST application available at <https://srsfed.ulb.ac.be:8443/JobService> as a GRIA service has been integrated in the wEMBOSS portal located at <http://akagera.ulb.ac.be/wEMBOSS>. To make the *gria\_blast* application accessible through wEMBOSS, ULB has written a command-line GRIA client code for this application and an EMBOSS wrapper code described by an ACD file.

This integration path will be followed for all analysis services that have been retained in the framework of the SIMDAT-Pharma developments. Similarly, all workflows will also be presented in the wEMBOSS portal. Furthermore, the access to the NEC's Semantic broker will also be embedded in the same way in order to support the search and the browsing of annotated services.

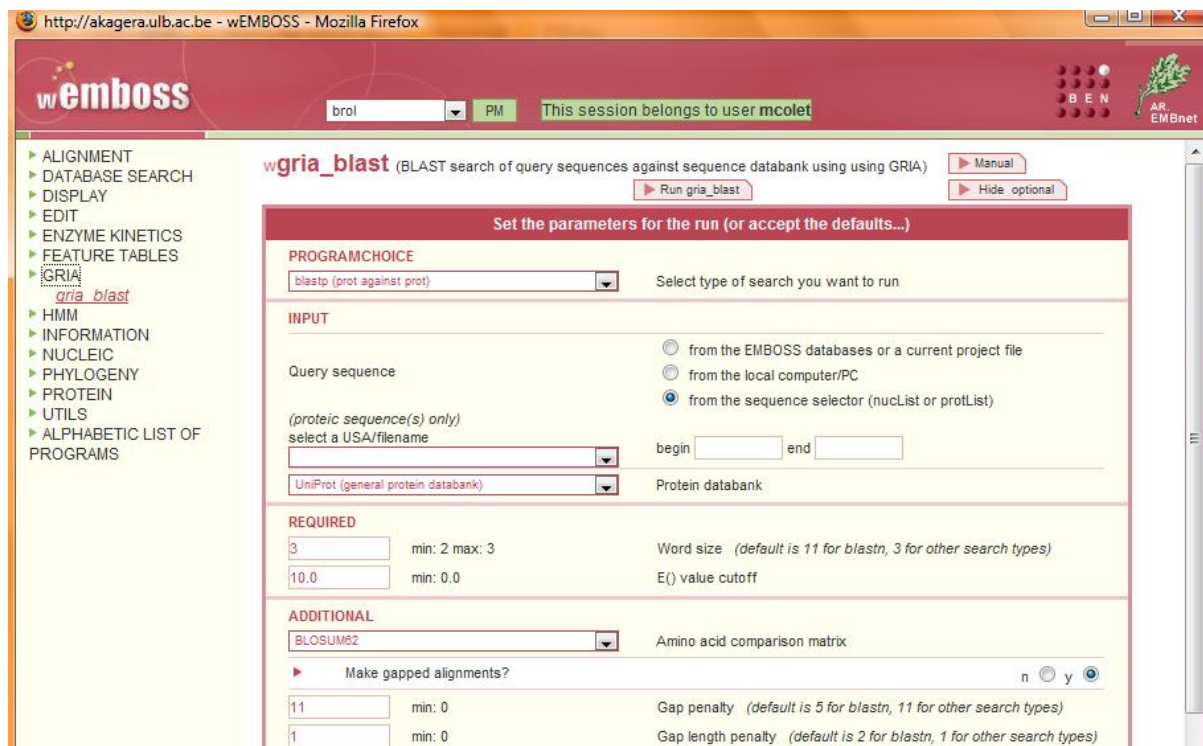


Figure 10: Web-Interface to GRIA BLAST Service

## 6 Requirements

<b>Name</b>	Web services of text mining services		
<b>Business Requirements</b>	To be able to call knowledge services through web service API		
<b>Date created</b>	2007-10-08	<b>Source</b>	Richard Kamuzinzi ULB
<b>First Implementation</b>	2008-01-31	<b>Priority</b>	Medium
<b>Application Activity</b>	Pharma	<b>Related Prototype</b>	Federated Portal
<b>Technology Component</b>	Knowledge Services	<b>SIMDAT Modules targeted</b>	wEMBOSS
<b>Detailed description of the requirement</b>			
The IAIS text mining services need to be integrated into the Federated Portal. To access those services, a standard web service is required. It has to be evaluated whether the text mining services would be GRIA-based. Whenever the web interface is available, a client application targeting the text mining services will be developed and integrated into the wEMBOSS portal framework.			
<b>Relation to the prototype</b>			
The set of the user interactions in the Federated portal needs to be extended to relevant SIMDAT services			
The scientist needs to manually interact with the annotation service in order to refine its accuracy			
<b>Requested functionality</b>			
Web service access to text mining services that require human interaction			
<b>Validation of the requirement</b>			
The web services developed will be considered valid if they enable the creation of the corresponding wEMBOSS-based web interface.			

<b>Name</b>	To simplify the use of GRIA nodes in KDE		
<b>Business Requirements</b>	To make easier the creation of workflows that launch GRIA services		
<b>Date created</b>	2007-10-08	<b>Source</b>	Richard Kamuzinzi ULB
<b>First Implementation</b>	2008-01-31	<b>Priority</b>	Low
<b>Application Activity</b>	Pharma	<b>Related Prototype</b>	B2A, Federated Portal
<b>Technology Component</b>	Integrated Grid Infrastructure, Workflow	<b>SIMDAT Modules targeted</b>	KDE
<b>Detailed description of the requirement</b>			
When designing workflows using the InforSense KDE, it might be useful to simplify the “GRIA interactions” into one GRIA node. For instance, data staging should be hidden to the designer. Thus, the workflow designer would have just to specify parameters about files to upload/download etc. Additionally, the simplified GRIA node should expose all applications parameters thanks to the Job Service interface.			
<b>Relation to the prototype</b>			
This requirement is obviously related to the B2A prototype as we need to design workflows that launch GRIA services It is also related to the Federated Portal as the IXodus work flow could benefit from the simplification			
<b>Requested functionality</b>			
To have one GRIA node (per application) that encapsulates required sub-interactions to launch a GRIA-wrapped application			
<b>Validation of the requirement</b>			
The validation will be carried out by simplifying the IXodus work flow			

## 7 Conclusions

We have seen that the current technologies which are being developed in the SIMDAT project already have the potential to implement an industrial strength pharmaceutical workflow across a Grid test-bed comprised of both academic and industrial partners. During the reporting period this has been successfully demonstrated with a substantial subset of the MSAP running across a test-bed comprising five different remote sites. The central requirement from the industrial perspective to have an environment allowing the participating to interact in a controlled and secure way has been sufficiently addressed. Thus large Pharms like GSK have now the possibility to scale their business relationship with biotech companies like Inpharmatica, i.e., they can restrict themselves on exactly those resources they are interested in and are not forced to subscribe to a complete and costly product. Additionally and of similar importance is the time required to set up a new relationship via the establishment of a new virtual organisation. In the conservative way of implementing such relationships time periods in the order of months are not unlikely the Grid paradigm, however, can reduce this to weeks or even days. Biotechs, on the other hand, have the opportunity to get access to a new market and, hence, are in the position to increase their commercial offer by implementing a finer granularity of their product portfolio. In summary, Grid technology provides a new business model in the life science sector, which can be considered as a success of the SIMDAT project as a whole.

## 8 References

- [1] [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.
- [2] Kumpf, K., TUAM: A new tool for universal annotation and mediation, Journal of Web Semantics, 2005
- [3] 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>
- [4] Qu, C. F. Zimmermann, K. Kumpf, R. Kamuzinzi, V. Ledent, and R. Herzog, “Semantics enabled Service Discovery Framework in the SIMDAT Pharma Grid”, IEEE Trans. on Information Technology in Biomedicine, to appear.
- [5] Falk Zimmermann et al., Documentation of advanced implementation of SIMDAT Pharma Prototypes for Interoperability Phase including evaluation of underlying technologies , <http://bscw.scai.fraunhofer.de/bscw/bscw.cgi/d100700/D.10.2.pdf>