EGI-InSPIRE

Services for the Life Science Community

EU MILESTONE: MS611

Document identifier:	EGI-MS611-683-v1.2
Date:	**07/09/2011**
Activity:	**SA3**
Lead Partner:	**CNRS**
Document Status:	**FINAL**
Dissemination Level:	**PUBLIC**
Document Link:	https://documents.egi.eu/document/683

Abstract

This report details the services offered to the Life Sciences community and how the intended users can access them.

I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner/Activity	Date
From	Johan Montagnat	CNRS/SA3	16/07/2011
Reviewed by	**Moderator:** Steve Brewer **Reviewers:** Rebecca Breu	EGI.eu	21/08/2011
Approved by	**AMB & PMB**		6/09/2011

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1.0	16/07/2010	First draft	Johan Montagnat
1.1	24/08/2010	Revision after review	Johan Montagnat
1.2	30/08/2010	Last typos	Johan Montagnat

IV. APPLICATION AREA

This document is a milestone for the European Commission, applicable to all members of the EGI-InSPIRE project, beneficiaries and Joint Research Unit members, as well as its collaborating projects.

V. DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the authors. The procedures documented in the EGI-InSPIRE "Document Management Procedure" will be followed: https://wiki.egi.eu/wiki/Procedures

VI. TERMINOLOGY

A complete project glossary is provided at the following page: http://www.egi.eu/results/glossary/.

VII. PROJECT SUMMARY

To support science and innovation, a lasting operational model for e-Science is needed − both for coordinating the infrastructure and for delivering integrated services that cross national borders.

The EGI-InSPIRE project will support the transition from a project-based system to a sustainable pan-European e-Infrastructure, by supporting 'grids' of high-performance computing (HPC) and high-throughput computing (HTC) resources. EGI-InSPIRE will also be ideally placed to integrate new Distributed Computing Infrastructures (DCIs) such as clouds, supercomputing networks and desktop grids, to benefit user communities within the European Research Area.

EGI-InSPIRE will collect user requirements and provide support for the current and potential new user communities, for example within the ESFRI projects. Additional support will also be given to the current heavy users of the infrastructure, such as high energy physics, computational chemistry and life sciences, as they move their critical services and tools from a centralised support model to one driven by their own individual communities.

The objectives of the project are:

1. The continued operation and expansion of today's production infrastructure by transitioning to a governance model and operational infrastructure that can be increasingly sustained outside of specific project funding.
2. The continued support of researchers within Europe and their international collaborators that are using the current production infrastructure.
3. The support for current heavy users of the infrastructure in earth science, astronomy and astrophysics, fusion, computational chemistry and materials science technology, life sciences and high energy physics as they move to sustainable support models for their own communities.
4. Interfaces that expand access to new user communities including new potential heavy users of the infrastructure from the ESFRI projects.
5. Mechanisms to integrate existing infrastructure providers in Europe and around the world into the production infrastructure, so as to provide transparent access to all authorised users.
6. Establish processes and procedures to allow the integration of new DCI technologies (e.g. clouds, volunteer desktop grids) and heterogeneous resources (e.g. HTC and HPC) into a seamless production infrastructure as they mature and demonstrate value to the EGI community.

The EGI community is a federation of independent national and community resource providers, whose resources support specific research communities and international collaborators both within Europe and worldwide. EGI.eu, coordinator of EGI-InSPIRE, brings together partner institutions established within the community to provide a set of essential human and technical services that enable secure integrated access to distributed resources on behalf of the community.

The production infrastructure supports Virtual Research Communities (VRCs) − structured international user communities − that are grouped into specific research domains. VRCs are formally represented within EGI at both a technical and strategic level.

VIII. EXECUTIVE SUMMARY

This milestone report lists the services delivered by the Life Science Heavy User Community, and describes how each of them can be accessed. All services have not been developed or deployed already and some sub-sections refer to work planed rather than actual instances of deployed services.

TABLE OF CONTENTS

1 INTRODUCTION

This document lists the delivered (or partially delivered) services by the Life Sciences Heavy User Community from EGI-InSPIRE. This includes:

- TSA3.2.1: Services, Dashboards (section 2.1). The Life Sciences Dashboard aims at integrating community monitoring and administration tools into a single, comprehensive interface.
- TSA3.2.3: Services, GReIC (section 2.2) and Hydra (section 0). GRelC is a generic database access interface that is useful for any grid application requiring metadata management. Hydra is a file encryption/decryption service used to protect sensitive data stored on grid storage resources.
- TSA3.2.4: Services, Taverna workflow engine (section 2.4). Taverna is a popular workflow workbench geared towards execution of flows of Web Services, especially popular within the Bioinformatics community. Taverna is developed completely independently from the EGI-InSPIRE project but it is planed to be delivered as a service to the user community.
- TSA3.4: CoreBio database services (section 3.1) and tools for Life Sciences community (section 0). CoreBio database services are bioinformatics database management and access services planed to be delivered for the Bioinformatics community. Life Sciences community tools are needed to manage a multi-Virtual Organization (VO) scale community.

Beyond the EGI-InSPIRE project's own definition of a Heavy User Community, members of the Life Sciences area self-organized into a multi-grid users community named Life Sciences Grid Community (LSGC) and legally represented by the HealthGrid association (http://www.healthgrid.org). The LSGC implements a Virtual Research Community (VRC). This document lists the services delivered or being developed by the Life Sciences VRC within EGI-InSPIRE. It does not provide an exhaustive list of all services developed by and available to the community outside of the EGI-InSPIRE project. The Life Sciences Grid Community (LSGC) wiki page [R4] is another important source of information for users from the LS community.

This document provides the bootstrapping information needed for any user from the community who would like to make use of the services deployed. Most services address end-user needs, except the Life Science VRC management tools described in section 0 that are geared towards VO administrators.

2 SHARED SERVICES AND TOOLS

2.1 Dashboard

Virtual Research Communities (VRCs) is a new concept in the EGI-InSPIRE user communities' organization. The middleware developed so far is predominantly relying on a separation of users through various Virtual Organizations (VOs). VRC management is a complex duty for which appropriate support tools still need to be developed. In particular, integrated tools to provide vision of the whole VRC activity and facilitate its administration are needed.

The Life Sciences Grid Community (LSGC) dashboard will integrate various VO management services into a single portal. From the experience gained in operating the LSGC during the project first year the following services are currently considered for integration:

1. User management tools, covering users registration and grid activity follow-up (see section 3.2 for details).
2. VRC-wide grid monitoring and operating tools, based on a Nagios server [R8] currently deployed by the NGI-France. The Nagios interface requires improvement to provide a relevant VRC-view.
3. Community files management, needed to deal with storage resources, which often cause difficulties (filled-up or decommissioned SEs) and would benefit file migration procedures.
4. VRC-wide accounting, needed to deliver statistics at the VRC level. The current EGI accounting portal [R9] only provides per-VO accounting information.

The "VO Admin Dashboard" (https://vodashboard.lip.pt), which similarly aims at providing an integrated VO portal, is a promising basis for the Dashboard and is currently investigated. The precise Dashboard development roadmap is currently being settled. In addition, the UCB evaluates a VO-oriented version of the EGI Central Operations Portal [R14], which would cover several important operational needs from the LSGC point of view such as integration of monitoring sources (Nagios alarms), reporting system (GGUS and GOC tickets), as well as technical team organization. We are currently liaising with both the VO Admin Dashboard development team and the UCB to coordinate efforts in tools development and avoid redundant work.

2.2 GRelC database access and integration service

The management of databases plays a crucial role in different scientific domains like Earth Science, High Energy Physics and Life Science. In the EGI context, the GRelC project has been supporting the Life Science community from several points of view through the development use cases, user support, training and tutorial activities stemming from the biosciences domain:

A key activity related to this task (as part of the user support) is the GRelC interface. It aims at providing a web access interface to the database resources available in the EGI production grid. Such a "registry" will complement the functionalities provided by the EGI Application Database [R10] and will support the users using social network functions, e.g. a message board on the DashboardDB platform (similar to shoutboxes found on public community websites), create discussion groups, rate existing resources, etc. The users will be able to publish their data resources and discover new ones already deployed through search & discovery functionalities.

During the first year, a questionnaire has been sent to the users to identify the available database resources and identify specific needs like the porting in grid of new databases. The questionnaire is part of the user support activity and represents a good starting point to start the interaction with the end user community. Starting from it, new use cases have been identified (the most relevant one relates to some biological databases ported in grid and moved from a flat file model to a relational one).

From an end-user perspective, the GRelC project provides support through the official project website (www.grelc.unile.it) where users can find information about the status of the project, new releases, installation guides, software development kit, rpms, etc. The website represents the best way to bootstrap the activities related to the GRelC software, find news related to next events, available papers, tutorial sessions, new features, releases and so on.

The server and the client parts rely on 1 rpm each, which must be installed on a UI-like machine. The server can be managed either through the GRelC Portal or through the Command Line Interface.

The CLI and the GRelC Portal provides a set of functionalities (secured through the adoption of the Grid Security Infrastructure) to interact with the GRelC server and to register a new grid-database, define user access policies, add/remove grid-users, submit queries in grid.

The complete documentation is available on-line [R5]. It includes the following guides:

1) grelc-dais-admin-guide (to manage new grelc server instances)
2) grelc-dais-authorization-guide (to manage authorization tasks)
3) grelc-dais-bdii-extension-guide (to publish database oriented information on a bdii)
4) grelc-dais-freetds-guide (to configure new resources through freetds)
5) grelc-dais-install-guide (to install a new grelc server instance)
6) grelc-dais-user-guide (end user functionalities)

Support is also provided through some tutorials available on the GILDA [R6] website where people can learn more about the GRelC service, the GRelC Portal, the Command Line Interface, and so on. Tutorials can be found on the GILDA Wiki [R7] and are organized to provide different skills according to different users' needs and requests. Basically, the GRelC service provides support to the HUC in terms of grid database management and in terms of grid metadata management. Most of the use cases already implemented in the past fall in one of these two classes. In both the cases, it is possible to define common use cases that can act, at the same, time both as best practices and effective and tested architectural/usage patterns. Such knowledge can be easily exported in several contexts in a shortened time scale. For example, in terms of grid metadata management, the solution adopted in the Climate-G [R15] testbed (environmental domain), is quite general and could be exploited in completely different domains to provide grid-enabled distributed search & discovery capabilities. On the other hand, in terms of grid database management, the solution adopted to access to the UNIPROT biological databank (see Project Quarter 5) could be easily exploited to port in grid other data sources related to the Life Sciences domain.

2.3 HYDRA file encryption service

Sensitive data (related to individuals' health and possibly identifiable data) is often manipulated in the Life Sciences area. In an open environment such as a grid, only encryption of data through robust cryptography methods can ensure confidentiality of patient data. The HYDRA is a secure file encryption service designed in the context of the EGEE series of project to fulfil that requirement.

2.3.1 Service overview

Hydra is a file encryption/decryption tool developed at CERN as part of the gLite middleware, enabling encryption of sensitive files stored on storage resources. It implements the Shamir's secret key sharing algorithm [R1], a key-based encryption solution where encryption keys are stored in a distributed key store, ensuring availability and robustness against attacks. An overview of the Hydra software is available here [R2].

The Hydra software is part of the regular gLite User Interfaces and Worker Nodes through the command line interface (glite-data-hydra-cli package).

2.3.2 Hydra Client configuration and usage

From the user's point of view, the service may be configured following two methods: a local-file based method, or through the service unique id published in the BDII. The latter is the preferred and most simple solution. To do so users should set the following variables in the user' profile:

```
export GLITE_SD_PLUGIN=bdii
export GLITE_SD_VO=biomed
```

The command-line interface is available under regular gLite User Interface hosts as binaries prefixed with “glite-eds-”: glite-eds-chmod, glite-eds-getacl, glite-eds-rm, glite-eds-decrypt, glite-eds-key-register, glite-eds-setacl, glite-eds-encrypt, glite-eds-key-unregister, glite-eds-get, glite-eds-put.

A user with a valid proxy certificate usually interacts with the Hydra service through the glite-eds-put and glite-eds-get commands. The former encrypts and uploads a file to a Storage Element. Conversely the latter downloads and decrypts the file (if the user is authorized to access the file encryption keys).

The encryption keys access control is ACL-based. The glite-eds-get/setacl commands are used to control these rights. While the glite-eds-put/get commands coherently manipulate a file and its associated encryption key, an atomic operation on the hydra server can be performed through the glite-eds-key-register/unregister commands (to register/unregister new encryption keys) and glite-eds-encrypt/decrypt (to locally encrypt or decrypt a file). These interactions shall typically follow the examples below:

1. Create a key in Hydra, e.g. with id fmichel-id:

```
# glite-eds-key-register -v fmichel-id
A key has been generated and registered for ID 'fmichel-id'
```

2. Encrypt a local file:

```
# echo test > test.txt
# glite-eds-encrypt fmichel-id test.txt test.txt.encrypted
File 'test.txt' has been successfully encrypted
        with key 'fmichel-id'
        and written to 'test.txt.encrypted'.
```

3. Decrypt a local file:

```
# glite-eds-decrypt fmichel-id test.txt.encrypted text_decrypted.txt
File 'test.txt.encrypted' has been succesfully decrypted
        with key 'fmichel-id'
        and written to 'text_decrypted.txt'.
```

2.3.3 Service delivery progress

An experimental Hydra service has been successfully deployed within the first year of the EGI-InSPIRE project on a gLite release 3.1 UI, and is usable for test purposes. Work is currently on going to migrate Hydra to gLite release 3.2 at CERN by gLite experts from the EMI project. See [R3] for details.

Beside deployment of the key store services, it will be necessary to install and publish the Hydra client on all sites where Worker Nodes may be required to access the Hydra service (presumably all sites accessible to the LS HUC VOs).

A 3-servers based Hydra key-store will be deployed as a service for the life sciences community. This task, due in the first year of the EGI-InSPIRE project, has been delayed as it is deemed preferable for long-term maintenance issues, to wait for the gLite 3.2 version of Hydra to be released, and install the Hydra service on recent gLite 3.2 servers from the beginning, rather than have to move installed key stores on new machines later on.

2.4 Taverna workflow engine

The EBI (European Bioinformatics Institute) project partner, who is in charge of this task, provided no information on its progress.

3 SERVICES FOR LIFE SCIENCES

3.1 *Core Bioinformatics Services*

The EBI (European Bioinformatics Institute) project partner, who is in charge of this task, provided no information on its progress.

3.2 *Life Science Virtual Research Community*

Following the recommendations of the Life Sciences Grid Community (LSGC) board, effort was invested in the organization and the development of appropriate tools to better manage the users' community. The LSGC board organizes monthly phone meetings whose minutes are available at the community wiki [R4]. Most effort during the project first year has been invested in the set-up of community communication channels and technical assistance to the end users. The development of user management tools is planed to ease the task of VO administrators and VRC leaders.

3.2.1 Services delivered

The HealthGrid association maintains per-VO and VRC-wide mailing lists. These communication channels are essential in maximizing the impact of the tools and services developed within the community, as well as negotiating and managing resources with the NGIs. The mailing lists are automatically updated daily by querying the VRC-affiliated VOMS servers.

A redundant VOMS server has been provisioned and installed by the HealthGrid association for the biomed VO. Its production use will help shielding the users from downtimes of the main server.

The LSGC receives support from the NGIs involved and the HealthGrid association in term of manpower and grid resources. Part of this manpower is used to operate a technical team of members from the biomed VO [R11] to assist the LS users. The function of the team is to address problems reported by the community, usually through the GGUS front-line support system; the support is organized using duty shifts. The technical team also anticipates problems by actively probing the most critical services for the proper VO operation through a dedicated Nagios server [R12]. Procedures have been defined to react to regular maintenance events such as SE decommissioning operations.

3.2.2 Plans for the coming year

The LSGC is currently designing a user management database, which will facilitate liaising with hundreds of users registered in the affiliated Virtual Organizations. The database schema was specified, and state diagrams have been drafted. This service will interface to Virtual Organization Membership Service (VOMS) servers as well as the EGI applications database, to avoid replicating existing information. It will complement the VOMS and applications database with extra-information on the users and their affiliations.

The delivery of a redundant LFC server is planed. Technical solutions for setting up such an alternate service are currently being investigated with the support of UCST.

4 CONCLUSION

Within the context of the EGI-InSPIRE project, some services are provisioned for the exploitation by the Life Sciences Heavy Users Community: the GRelC database access interface, the Hydra file encryption services, the Taverna workflow management workbench and the CoreBio bioinformatics database management service. This effort is complementary to the work endorsed by the Life Sciences Grid Community, which similarly aims at provisioning community-level services to ease grid applications development and grid exploitation [R4].

In addition, the emergence of Virtual Research Communities which encompasses multiple VOs triggers the need for new community monitoring and management tools, and proper interfaces to such tools.

The Life Sciences Grid Community covers a very wide area of health-related computational activities. An exhaustive list of services that would be useful for such a broad and heterogeneous community cannot be provided. However, the community has invested significant work in identifying the fundamental requirements for the underlying grid infrastructure and middleware. This effort resulted in a comprehensive list of core technical requirements that have been endorsed by the User Community Board. This list of requirements is indexed on the LSGC wiki [R13].

5 REFERENCES

R1	A. Shamir. "How to share a secret" in Communications of the ACM (CACM), vol. 22, pages 612-613, 1979.
R2	Hydra service overview. https://twiki.cern.ch/twiki/bin/view/EGEE/DMEDS
R3	MS607: Hydra service deployment on a multi-servers configuration, https://documents.egi.eu/document/327
R4	Life Sciences Grid Community, http://wiki.healthgrid.org/LSVRC:Index
R5	GRelC documentation, http://grelc.unile.it/grelc_source/grelc-dais-all-guides.zip
R6	GILDA (Grid INFN Laboratory for Dissemination Activities), https://gilda.ct.infn.it/
R7	GILDA Wiki, https://grid.ct.infn.it/twiki/bin/view/GILDA/GRelCProject
R8	Nagios IT infrastructure monitoring tool, http://www.nagios.org/
R9	EGI accounting portal (CESGA), http://www3.egee.cesga.es/gridsite/accounting/CESGA/egee_view.php
R10	EGI Application Database, http://appdb.egi.eu
R11	Biomed VO, http://wiki.healthgrid.org/LSVRC:Biomed
R12	Biomed Nagios server, https://grid04.lal.in2p3.fr/nagios (accessible to EGI certificate owners only)
R13	LSGC list of requirements, http://wiki.healthgrid.org/LSVRC:Index#EGI_requirements_from_the_LSGC
R14	EGI Central Operation Portals, https://operations-portal.egi.eu
R15	S. Fiore, G. Aloisio, P. Fox, M. Petitdidier, H. Schwichtenberg, S. Denvil, J. D. Blower, A. Cofino, “The Climate-G testbed: towards large scale distributed data management for climate change”, Proceedings of the International Conference on Computational Science ICCS 2011, June 1 - June 3, 2011, Nanyang Technological University, Singapore, Procedia Computer Science, Elsevier, pp. 567-576.