EGI-InSPIRE

Services for the Life Science Community

EU MILESTONE: MS604

Document identifier:	EGI-MS604-v6
Date:	**07/01/2011**
Activity:	**SA3**
Lead Partner:	**CNRS**
Document Status:	**FINAL**
Dissemination Level:	**PUBLIC**
Document Link:	https://documents.egi.eu/document/236

Abstract

Report detailing the services offered to the Life Sciences community and how they can be accessed.



I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner/Activity	Date
From	Johan Montagnat	CNRS/SA3	11/10/2010
Reviewed by	**Moderator:** Steve Brewer **Reviewers:** Rebecca Breu	EGI.eu	9/11/2010
Approved by	**AMB & PMB**		7/1/2011

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1.0	11/10/2010	First draft	Johan Montagnat
2.0	27/10/2010	Second draft	Johan Montagnat
3.0	16/11/2010	Third draft	Johan Montagnat
4.0	1/12/2010	Final version after reviews	Johan Montagnat

IV. APPLICATION AREA

This document is a formal deliverable 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

1.1 Purpose

This document lists the services planed for delivery by the Life Sciences Heavy User Community in the EGI-InSPIRE DoW. This includes:

- TSA3.2.1: Services, Dashboards (section 2.1)
- TSA3.2.3: Services, GReIC (section 2.2) and Hydra (section 2.3)
- TSA3.2.4: Services, Taverna workflow engine (section 2.4)
- TSA3.4: CoreBio database services (section 3.1) and Life Sciences VRC (section 3.2)

The 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 3.2 which are geared towards VO administrators.

1.2 Application area

This document is a formal deliverable 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.

1.3 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

1.4 Terminology

A complete project glossary is provided in the EGI-InSPIRE glossary: http://www.egi.eu/results/glossary/.

2 SHARED SERVICES AND TOOLS

2.1 Dashboard

The dashboard functions are to support statistical analysis of the grid usage and follow up of the tools used by the communities. HealthGrid is to deploy the dashboard in the coming months but due to the actual changes from biomed VO into the LSVRC (see section 3.2), the functions and tools to be integrated into the dashboard will have to be chosen in accordance to the new user requirements and the new tools that will be deployed. HealthGrid and the LSVRC are currently undertaking the review process and will propose the dashboard's new services and functions in the coming months, as a result of the LSVRC new organizational scheme.

2.2 GReIC database access and integration service

GRelC (Grid Relational Catalogue) is a Grid database access and integration service. The GRelC service allows users to interact with different Database management systems, both relational (PostgreSQL, MySQL, Oracle, DB2, SQLite, etc) and non-relational (eXist, XIndice, XML flat files). It provides a uniform access interface to heterogeneous data sources in a grid environment. The GRelC middleware is part of the EGEE RESPECT Programme since it works well with the EGEE software by expanding the functionality of the grid infrastructure (with regards to database management in the grid). During the project, the GRelC system will be enhanced to support the EGI communities with a new set of functionalities. These new features will be available to the end-users through the GRelC Portal, a seamless, ubiquitous and web-based environment for the management of geographically spread and heterogeneous grid data sources.

In the first quarter of the project, a new version of the GRelC service has been released. A new service instance has been deployed to support life sciences communities activities, to test new use cases, collect further requirements, validate new functionalities. A mailing list is also available (grelc-user@sara.unisalento.it) to receive news about the software, to post comments, questions and interact with the GRelC team. Moreover, on the project website (www.grelc.unile.it) users can find rpms, documentation, news, etc.

2.3 HYDRA file encryption service

2.3.1 Service overview

Hydra is a file encryption/decryption tool developed as part of the gLite middleware. Hydra is a special secure metadata catalog designed to hold encryption keys. The Hydra functionality is accessible in the regular gLite User Interfaces and Worker Nodes through the command line interface (glite-data-hydra-cli package). Hydra may be deployed as a single key store or as a distributed key store, implementing the Shamir's secret key sharing algorithm, for improved availability and higher robustness against attacks [R1].

An overview of the Hydra software is available here [R2]. The installation and deployment procedures for the Hydra keystore are described in this gLite documentation [R3].

2.3.2 Hydra Client

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.

The client requires some information on the hydra server used that can be provided either in a local file (GLITE_SD_PLUGIN variable value set to file and GLITE_SD_SERVICES_XML variable pointing to the configuration file), or in a BDII (GLITE_SD_PLUGIN variable value set to bdii).

A user usually interacts with the Hydra service through the glite-eds-put and glite-eds-get commands. The former encrypt and upload a file to a Storage Element. Conversely the latter download and decrypt the file (if the user is authorized to access the file encryption keys). The access control the to encryption keys 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, 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).

2.3.3 Service delivery progress

The client for hydra is currently not available yet in gLite v3.2 but it is part of the former releases and work is on going at CERN to update it for the latest gLite version.

A Hydra catalog will be deployed within the first year of the EGI-InSPIRE project as a service for the life sciences community. Support for the Hydra software is available through the GGUS global gLite support.

2.4 Taverna workflow engine

EBI plans to provide plugins for Taverna to support life science services. Taverna is the most widely used workflow system for users of bioinformatics services. Originally developed at EBI through the myGrid projects in the UK e-Science programme, Taverna is now an open source project coordinated by the OMII-UK team in Manchester with many third party contributions.

The Taverna developers support two versions, the original Taverna 1.7 and the new Taverna 2.2 release. Unfortunately, the changes to the Taverna internal architecture require separate plugins to be developed for each version. This situation should be resolved in the architecture changes planned for Taverna version 3.

The current status of Taverna plugins is documented on the project website at http://www.taverna.org.uk/documentation/taverna-plugins/ .

EBI intends to work on Taverna plugins when the new Taverna 3 OGSi platform is released. If this is delayed for any reason we expect to be able to work with the Taverna 2 architecture and then migrate to OGSi Taverna 3.

3 SERVICES FOR LIFE SCIENCES

3.1 Core Bioinformatics Services

Bioinformatics services are widely available using data and tools, which are freely available. The user community is very large and geographically dispersed. The underlying data changes and grows rapidly - for example, the rate of growth in the amount of nucleotide sequence data in public repositories is faster than Moore's Law for transistors growth in integrated circuits.

To provide access to individual data items and to tools that search large data sets, a variety of interfaces is in common use, mostly based on REST and SOAP protocols. Standardisation efforts including the EMBRACE network and the BioCatalogue project have produced common service definitions that facilitate the porting of services, tools and data resources to other service architectures. Many individual data resources and tools have been ported to grid systems for specific projects.

3.2 Life Science Virtual Research Community

3.2.1 Life Science VRC set up

The Life Science community invested most effort of this first quarter in setting up a sustainable operation model for the Virtual Research Community (VRC). Indeed, the fragmentation of the former EGEE infrastructure and management into National Grid Initiatives calls for a new decentralized model, that serves the community well while remaining compatible with the overall European Grid managerial structure.

An important moment in this organization process was the Life Science VRC meeting organized in Paris in conjunction with the HealthGrid 2010 conference on June 28. A broad consortium representative of the community decided to push forward the emergence of an international VRC implemented through a large-scale pan-European Virtual Organization (rather than relying on national-scale structures and VOs) to foster international collaborations and facilitate grid adoption. This operational model requires defining:

1. Governance policies;
2. Secure sustainable funding; and
3. Design technical tools for daily operations.

3.2.2 Current status

A first proposal for the Life Sciences VRC governance has been written and distributed to all stakeholders for feedback and formal declaration of support. An LS VRC wiki [R4] has been set up to collect and publish practical and technical information related to the community. The LS VRC is currently representing 4 VOs (biomed, enmr, lsgrid, and vlemed). It receives support from 6 NGIs (Dutch, French, German, Italian, Spanish and Swiss NGIs) and one ESFRI project (LifeWatch – discussions are on going with the ELIXIR ESFRI). A monthly phone conference is being organized to address the managerial and technical work involved with representatives from each of these NGIs, VOs and ESFRIs. Minutes of the meeting are available from the wiki [R4].

To ensure sustainability, funding models are being investigated by the HealthGrid association, founded in 2003 to promote and facilitate the use of grid technologies in Life Sciences. Technical work on VO administration tools has started with the design of a VO users and application database

and associated tools to monitor and manage the population of VRC members exploiting the grid infrastructure.

To assist the LS user communities, a Technical Team of members from the biomed VO was set up [R5]. This team defines duty shifts to ensure that there is always a pair of people in charge of addressing problems reported, usually through the GGUS front-line support system (see teams shift schedule in [R5]). The technical team also anticipates problems by actively probing the most critical services for the proper VO operation (currently VOMS and LFC which are centralized single-point-of-failure, as well as SEs which are critical to retrieve data). Procedures have been defined to react to regular maintenance events such as SE decommissioning operations.

3.2.3 Services delivery plan

The technical team monitoring tool is currently based on a lightweight Hudson integration server (see Hudson's dashboard in Figure 1). It is meant to evolve towards a Nagios server maintained by the operations.

The LS VRC is also currently designing a user management database, needed to maintain up-to-date and liaise with the hundreds of users registered to the LS VRC Virtual Organizations. This database will interface to VOMS servers as well as the EGI application database, to avoid replicating existing information. It will complete the VOMS and application database with extra-information on the users and their affiliations. It will be used to manage the user community and to produce sub-themes mailing lists (per-NGI, per-project, per-scientific domain) to liaise with the end users.

All | Biomed VO Storage Elements | Biomed VO general tests

S	W	Job ↓	Last Success	Last Failure	Last Duration
		All Storage Elements	6 mo 0 days (#6)	6 mo 0 days (#5)	4 hr 26 min
		Failure	N/A	6 mo 1 day (#1)	0.35 sec
		Init jobs	3 days 19 hr (#67)	5 mo 18 days (#21)	2 min 33 sec
		LFC response time	1 hr 10 min (#45)	N/A	0.94 sec
		Proxy infos	1 hr 10 min (#17)	N/A	0.85 sec
		Proxy init	1 hr 10 min (#16)	N/A	4.7 sec
		SE agh15.atlas.unimelb.edu.au	3 mo 11 days (#29)	N/A	4.6 sec
		SE agh3.atlas.unimelb.edu.au	1 mo 2 days (#28)	1 mo 1 day (#29)	2 min 42 sec
		SE aliserv1.ct.infn.it	2 mo 13 days (#29)	2 mo 15 days (#23)	3.7 sec
		SE axon-g05.ieeta.pt	1 hr 10 min (#16)	N/A	1 min 57 sec

Figure 1. Hudson dashboard used by LS VRC technical team for infrastructure monitoring.

4 CONCLUSION

This document lists the services planed for delivery by the Life Sciences Heavy User Community in the EGI-InSPIRE DoW. It does not provide an exhaustive list of all services developed by and available to the community outside of the EGI-InSPIRE project. Other services are of general interest and regularly used though, such as the vBrowser virtual catalog browser and grid front-end [R6] developed by VLe-Med, and the MOTEUR2 gLite-interfaced data-centric workflow engine [R7] developed by CNRS.

5 REFERENCES

R 1	A. Shamir. "How to share a secret" in Communications of the ACM (CACM), vol. 22, pages 612-613, 1979.
R 2	Hydra service overview. https://twiki.cern.ch/twiki/bin/view/EGEE/DMEDS
R 3	Hydra service installation. http://glite.web.cern.ch/glite/packages/R3.0/R20060502/doc/installation_guide_3.0-2.html#_Toc135537608
R 4	LS VRC wiki. http://wiki.healthgrid.org/LSVRC:Index
R 5	Biomed technical team. http://wiki.healthgrid.org/Biomed-Shifts:Index
R 6	vBrowser virtual catalog browser. http://www.nikhef.nl/~ptdeboer/vbrowser
R 7	MOTEUR2 workflow engine. http://modalis.polytech.unice.fr/moteur2