PROCEEDINGS OF SCIENCE

The agINFRA Science Gateway for Agricultural Sciences

R.Bruno[1]
INFN Division of Catania
Via S. Sofia, 64, 95123 Catania, Italy,
E-mail: riccardo.bruno@ct.infn.it

G. Allegri
GIS3W s.a.s
Viale G. Verdi, 24, 51016 Montecatini Terme, Italy
E-mail: allegri@gis3w.it

G. Andronico
INFN Division of Catania
Via S. Sofia, 64, 95123 Catania, Italy,
E-mail: giuseppe.andronico@ct.infn.it

R. Barbera
INFN Division of Catania and Dpt of Physics and Astronimy oft he University of Catania
Via S. Sofia, 64, 95123 Catania, Italy,
E-mail: roberto.barbera@ct.infn.it

F. Bitelli
INFN, Division of Roma Tre
Via della Vasca Navale, 84, 00146 Rome, Italy
E-mail: bitelli@fis.uniroma3.it

A. Budano
INFN, Division of Roma Tre
Via della Vasca Navale, 84, 00146 Rome, Italy
E-mail: antonio.budano@roma3.infn.it

A. Calanducci
INFN Division of Catania
Via S. Sofia, 64, 95123 Catania, Italy,
E-mail: antonio.calanducci@ct.infn.it

F. Celli
Food and Agriculture Organisation of the United Nations
Viale delle Terme di Caracalla, 00153 Rome, Italy,

[1] Speaker

PoS(ISGC 2013)034

PROCEEDINGS OF SCIENCE

E-mail: fabrizio.celli@fao.org

E.A.C. Costantini
I Consiglio per la Ricerca e la Sperimentazione in Agricoltura, Centro di ricerca per l'agrobiologia e la pedologia (CRA-ABP)
Piazza M. D'Azeglio, 30 - 50121 Florence, Italy,
E-mail: edoardo.costantini@entecra.it

M. Fargetta
INFN Division of Catania
Via S. Sofia, 64, 95123 Catania, Italy,
E-mail: marco.fargetta@ct.infn.it

A. Fornaia
Consortium GARR
Via dei Tizii, 6, 00185 Rome, Italy,
E-mail: andrea.fornaia@ct.infn.it

G. L'Abate
I Consiglio per la Ricerca e la Sperimentazione in Agricoltura, Centro di ricerca per l'agrobiologia e la pedologia (CRA-ABP)
Piazza M. D'Azeglio, 30 - 50121 Florence, Italy,,
E-mail: gio@soilpro.eu

S. Monforte
INFN Division of Catania
Via S. Sofia, 64, 95123 Catania, Italy,
E-mail: salvatore.monforte@ct.infn.it

A. Puliafito
Faculty of Engineering of the University of Messina
Contrada Di Dio, 1, 98166 Messina, Italy,
E-mail: apuliafito@unime.it

R. Ricceri
INFN Division of Catania
Via S. Sofia, 64, 95123 Catania, Italy,
E-mail: rita.ricceri@ct.infn.it

F. Ruggieri
INFN, Division of Roma Tre
Via della Vasca Navale, 84, 00146 Rome, Italy
E-mail: federico.ruggieri@roma3.infn.it

D. Saitta
INFN Division of Catania
Via S. Sofia, 64, 95123 Catania, Italy,

PoS(ISGC 2013)034

PROCEEDINGS OF SCIENCE

E-mail: davide.saitta@ct.infn.it

M. Villari
Faculty of Engineering of the University of Messina
Contrada Di Dio, 1, 98166 Messina, Italy,
E-mail: mvillari@unime.it

agINFRA (www.aginfra.eu) is a project co-funded by the European Commission under its Seventh Framework Programme that tries to introduce the agricultural scientific communities into the vision of open and participatory data-intensive science. agINFRA aims to remove existing obstacles concerning the data sharing and open access to scientific information and agriculture‘ data as well as to improve the preparedness of agricultural scientific communities to face, manage and exploit the abundance of relevant data that is available and can support agricultural research.

The agricultural domain includes a wide variety of increasingly complex, multi-disciplinary topics. Subjects vary from plant science and horticulture to agricultural engineering and agricultural economics to the environment generally and include an ever-growing array of inter-related research issues such as the linkages between climate change on the one hand and food security, or the loss of agro-biodiversity, or pressure on individual species on the other.

Scientists from all over the world are extensively researching those different subjects and thereby consuming as well as producing large volumes of data.

The integration process of the services accessing those data requires a registry of all the existing systems, a challenge that has started since the beginning of the project (agINFRA started on the 15th of October 2011 and will last three years). Many of those systems will be efficiently and securely accessed through single web entry points by both end users and system/data maintainers.

This contribution aims to demonstrate how the adoption of the Catania Science Gateway Framework (www.catania-science-gateways.it) can have a key role during and also beyond the agINFRA project lifetime providing a unique environment able to deal with this heterogeneity of systems. This work will describe the Science Gateway (http://aginfra-sg.ct.infn.it/) developed by the INFN Dpt. of Catania and registered as a Service Provider of several Identity Federations, which together with the adoption of the CLEVER cloud middleware, can provide a unique interface able to seamlessly access the different services of the project. Among others, the integration and use of the WebGIS-enabled Italian Soil Information System (ISIS), developed by the Agrobiology and Pedology Research Centre of the Italian Agricultural Research Council, will be shown.

This very challenging target could be reached only thanks to the adoption of widely accepted standards such as SAGA and SAML that ensure the sustainability, reliability and scalability of the proposed architecture.

The International Symposium on Grids and Clouds (ISGC) 2013
March 17-22, 2013
Academia Sinica, Taipei, Taiwan

PoS(ISGC 2013)034

PROCEEDINGS OF SCIENCE

1. Introduction

agINFRA is an Integrated Infrastructure Initiative (I3) European project that wants to introduce the agricultural scientific communities into the vision of open and participatory data-intensive science. In particular, agINFRA aims to design and develop a scientific data infrastructure for agricultural sciences that will facilitate the development of policies and the deployment of services that will promote sharing of data among agricultural scientists and develop trust within and among their communities. agINFRA wants to remove existing obstacles concerning the open access to scientific information and data in agriculture, as well as improve the preparedness of agricultural scientific communities to face, manage and exploit the abundance of relevant data that is (or will be) available and can support agricultural research.
Ultimately, agINFRA demonstrates how a data infrastructure for agricultural scientific communities can be set up to facilitate data generation, provenance, quality assessment, certification, curation, annotation, navigation and management.
This document will present first the Catania Science Gateway Framework and then how it has been used in the context of the agINFRA project showing several use cases of applications running on the Grid, or as gateway to access Cloud hosted services.

A final chapter will provide conclusions and final considerations.

1.1 Project architecture

agINFRA has been conceived as a sustainable data infrastructure for Agricultural research. The project addresses mainly services that could be of importance for repository mangers, aggregators and developers and eventually final researchers.

Designing the agINFRA research infrastructure, in order to adapt existing infrastructures that relevant partners brought to the project, different middleware components and services have been integrated and customized.

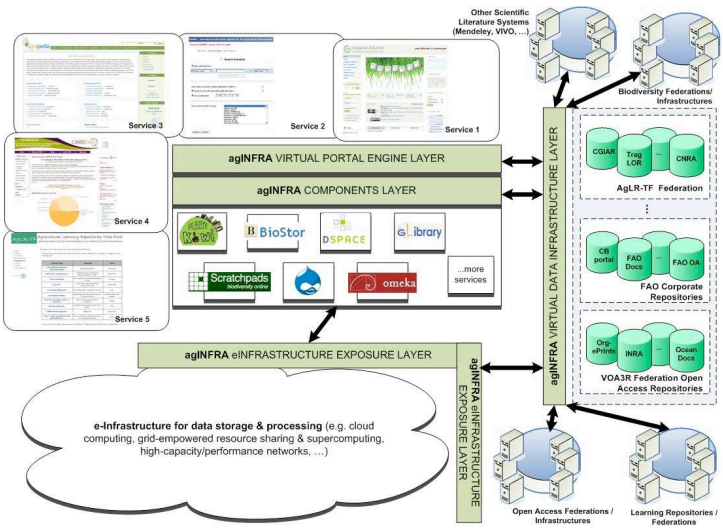


Figure 1-1: Overall agINFRA architecture

The Figure 1-1 represents the technical architecture of the project, this part of the agINFRA technical solution mainly takes care of the back-end components supporting the agINFRA e-Infrastructure Exposure Layer. That is, the layer of middleware components and services that

allow various data-related software components of the agINFRA Components Layer to find, engage, and use processing and computing resources in order to support the data intensive applications. Infrastructure related services are then linked to high level services, which interfaces web portals, or higher level services.
The Grid technology is being used to submit jobs that need quite intensive computational activities that fit batch systems operations. Grid infrastructure offers data and metadata services as well and both of them help to manage, locate and provide huge quantity of information. Cloud computing addresses instead the requirement of interactive/on-line use and/or permanent services that need to be up and running 24 hours/day. On top of Grid and Clouds infrastructures the layer of Science Gateways provide an easy access to the distributed infrastructures for final user and even higher level portals such as the one of FAO [2].
Grid and Cloud services are then harmonized and orchestrated by high-level user interfaces provided by the Science Gateways.
Another architectural view of the agINFRA project seen in the Figure 1-2 where it is possible to distinguish all possible user typologies and all possible interconnections among involved services.

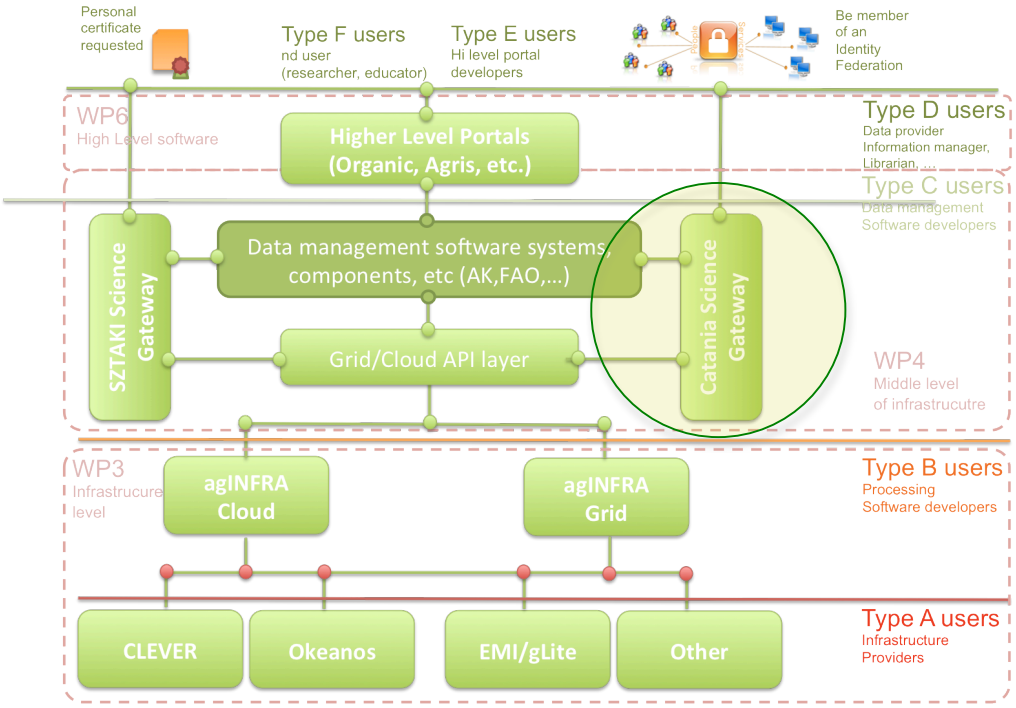


Figure 1-2 The agINFRA architecture and Service Interconnections

This paper will focus on the Catania Science Gateway Framework showing its structure, technology and presenting several use cases that demonstrates the very important potentiality of this framework.

2. Catania Science Gateway Framework

The Catania Science Gateway Framework [1] provides a unique interface able to access any distributed infrastructures seamlessly and without requesting to its user community to get in

PoS(ISGC 2013)034

touch with the specific technicalities of the specific infrastructure implementation such as: Grid certificate management, complex command line commands and even deal with specific application programming APIs.

By definition a Science Gateway is a community-developed set of tools, applications, and data that is integrated via a portal or a suite of applications, usually in a graphical user interface, that is further customized to meet the needs of a specific community [3].

Starting from the definition above the Catania Science Gateway Framework has been designed to easily address any possible specific need from scientific communities and to take care of existing standards in the world of distributed infrastructures and web/internet environments.

This chapter will describe the Catania Science Gateway Framework in its generic aspects highlighting each specific standard adopted while designing it.

2.1 Overall architecture

The Catania Science Gateway Framework adopts a modular architecture well shown by the figure 2.1, which presents three main components:

1. The AAI, the portlets layer and the Grid Engine
2. The Catania Grid Engine
3. The web applications as JSR286 portlets

Each component covers a core functionality of the portal and it is made adopting the most important standards to ensure the technological sustainability of the whole system.

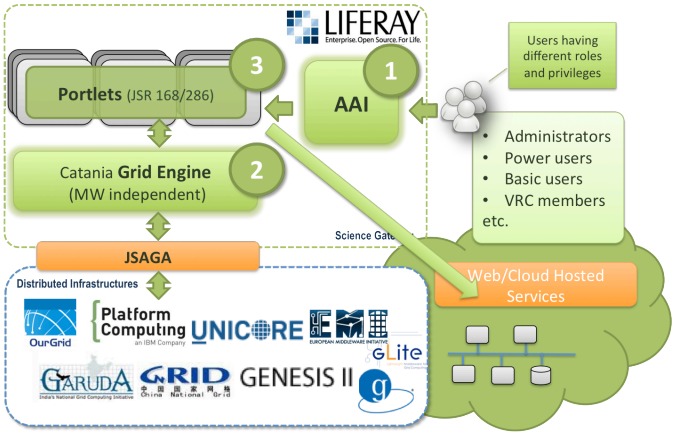


Figure 2-1 The Catania Science Gateway Framework Architecture

The Catania Science Gateway provides a full-featured environment able to access with unique software architecture to many distributed infrastructures and remote resources located on the Web or Clouds. It also allows creating high-level user interfaces able to submit and control Grid jobs, access data and metadata content, manage Cloud node instances, provide both secure and anonymous access to the Cloud hosted services and more in general avoid its user community to deal with all the existing technicalities of each specific technology addressed.

One of the peculiarities of the Catania Science Gateway is to allow an easy access to the distributed resources without compromising the strict security rules of any distributed

architecture. The Catania Science Gateway accomplishes this non-trivial task managing directly the user memberships and tracking all user activities by dedicated sub-components called UserTracking; finally the portal exposes itself while accessing the distributed infrastructure resources. The user-tracking activity answers to precise specifications suggested by the EGI Traceability policies [4]. Another key aspects of the Catania Science Gateway consist of the portal user membership management that makes use of the new concepts of Identity Federations and Identity Providers. The identity providers are those entities able to recognize and then authorize user membership, while an identity federation collects one or more identity providers as a single entity. Thus the main aim of identity federations is to provide the Single Sign-On (SSO) across the supported identity providers in the federation. This is a very powerful feature in the context of the new virtual research communities in e-Science.

The Catania Science Gateway has recently extended the world of Science Gateways to the world of Social Networks creating a dedicated Identity Provider that supports user identities coming from the most known Social Networks such as: Facebook, Google+, etc.

The next chapters will present all the aspects briefly covered by this chapter.

2.2 The AAI module

The Catania' Science Gateway manages its user Authentication and Authorisation with two separate modules each using different background technologies.

The Authorisation part is handled by the use of the Identity Federations (IdF) together with the Identity providers (IdP). An Identity Federation is made of *"[...] the agreements, standards, and technologies that make identity and entitlements portable across autonomous domains (Burton Group)*". Identity Federations have the aim of setting up and supporting a common framework for different organisations to manage accesses to on-line resources. They are already established in many countries and currently gather a number of people which is in the order of $O(10^7)$.

Identity providers are the physical entities able to recognize the user membership of a given community.

The Catania Science Gateway Framework authenticates users relying on several supported Identity Providers (IdPs). Each IdP can be a member of one or more Identity Federations, so that the user SSO will be granted across Services Providers (SPs) registered within the Identity Federations.

The Catania Science Gateway Framework currently supports federations based on the SAML 2.0 [5] standard specifications and on its implementation done by Shibboleth [6] and SimpleSAMLphp [7].

PoS(ISGC 2013)034

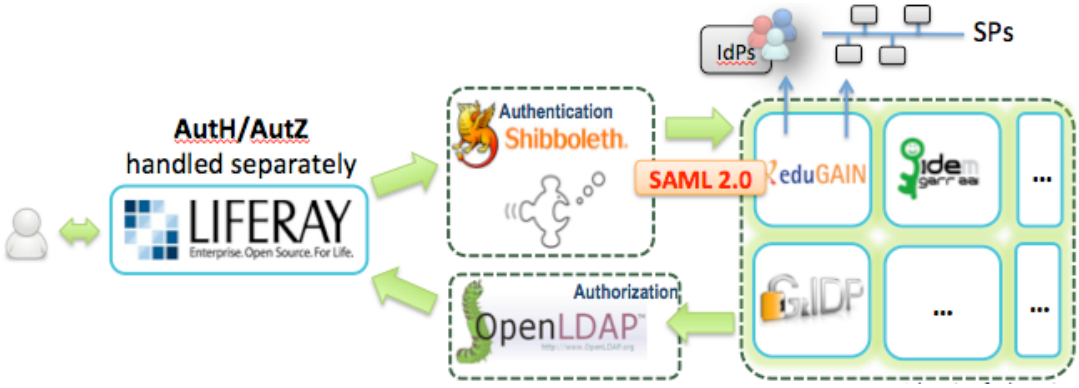


Figure 2-2 Catania Sciecnce Gateway Framework AuthN/AuthZ

In particular the agINFRA implementation of the Catania Science Gateway Framework currently supports several official Identity Federations and it is registered as Service Providers of the eduGAIN [8] inter-federation service within the GÉANT project [9]. The agINFRA portal is also registered as SP of the Identity Federation: Grid IDentity Pool (GrIDP) [10], a "catch-all" Identity Federation that has been expressly created to gather all the IdPs that do not already belong to any official federations and all the users of the Science Gateway who are not (already) registered in any IdPs. This is particularly important and useful in the contexts where it is necessary to authenticate the so-called "citizen scientist" (i.e., people belonging to the general public) and let him/her access the e-Infrastructure for dissemination and self-learning purposes. Inside the GrIDP Federation, there also exists a special IdP, the "Social Networks' Bridge Identity Provider" [11], that allows people to get authenticated with the same credentials they already have with the most known and populated social networks.

Unlike authentication, user authorization is carried out at the level of the Science Gateway: users whose request to register is approved by the managers of the portal, are stored in a LDAP-based registry together with the roles they have and the privileges they are granted.

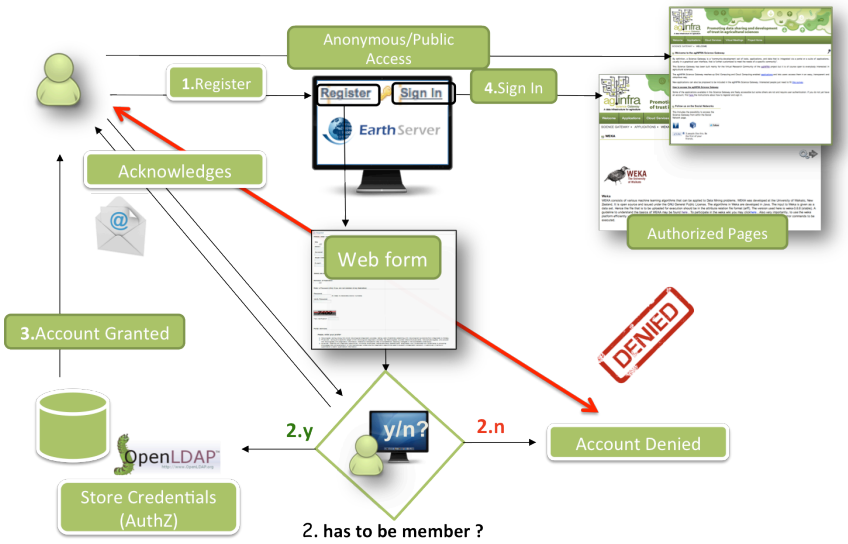


Figure 2-3 The user registration workflow

PoS(ISGC 2013)034

Users wanting to register to the Science Gateway will be automatically recognized and they will be prompted to register to one of the supported IdPs. In case the user already belongs to an IdP, will be automatically redirected to a web form where the user can apply for the foreseen portal rights. During the registration activity, any user transaction will be confirmed by email exchanges. The user rights request will be forwarded to the administrators of the portal. If the access right request will be accepted, the user information will be stored on the LDAP registry and the user is notified that the sign in is now allowed. Otherwise, a notification will be sent and the access will be denied. This procedure has been put in place in order to ensure that authorisations are not provided automatically to everybody and that a check be done on the requests by a human being.

Once a user has been authorised to access the Science Gateway, it will be possible to sign in and run allowed portal applications. The workflow of this phase is the following: when the user signs in, a web page prompts to select user' Identity Federation and the Identity Provider; then it will appear the Identity Provider login page where the user can insert the right credentials. If they are correctly verified, the control returns to the Science Gateway that checks if the user is inserted in the LDAP registry and will map his role with the registered user rights.

2.3 The Catania Grid Engine

The Catania Grid Engine is a generic software module able to interconnect the Scientific Gateway presentation layer with the underlying distributed infrastructures (Grid and Clouds) using standard technologies. It allows the quick creation of new Science Gateways providing their developers with a simple interface and avoiding worry about middleware specificities. This is possible thank to the adoption of the SAGA [12] standard. SAGA stands for Simple API for Grid Applications. SAGA is a family of related standards specified by the Open Grid Forum (OGF) to define an application-programming interface (API) for common distributed computing functionality. The specification of services, and the protocols to interact with them, is out of the scope of SAGA. Rather, the API seeks to hide the detail of any service infrastructures that may or may not be used to implement the functionality that the application developer needs. The API aligns, however, with all middleware standards within Open Grid Forum (OGF).

The Catania Grid Engine makes use of the JSAGA [13] implementation of the SAGA standard. JSAGA is a project developed and maintained by the IN2P3 institute in France. JSAGA implements the SAGA specification providing a lightweight, modular and pluggable set of java libraries. The Catania Science Gateway Framework well suites the integration of java libraries offering a homogeneous architecture to the whole system.

JSAGA library components are mainly three respectively dedicated to: Security Context, File management and Job management. The APIs are Object Oriented in order to facilitate the development of distributed applications.

JSAGA offers a set of command lines written on top of the low-level APIs capable to access to the distributed infrastructures exactly like their related native User Interfaces in order to demonstrate the powerful approach of the SAGA standard.

PoS(ISGC 2013)034

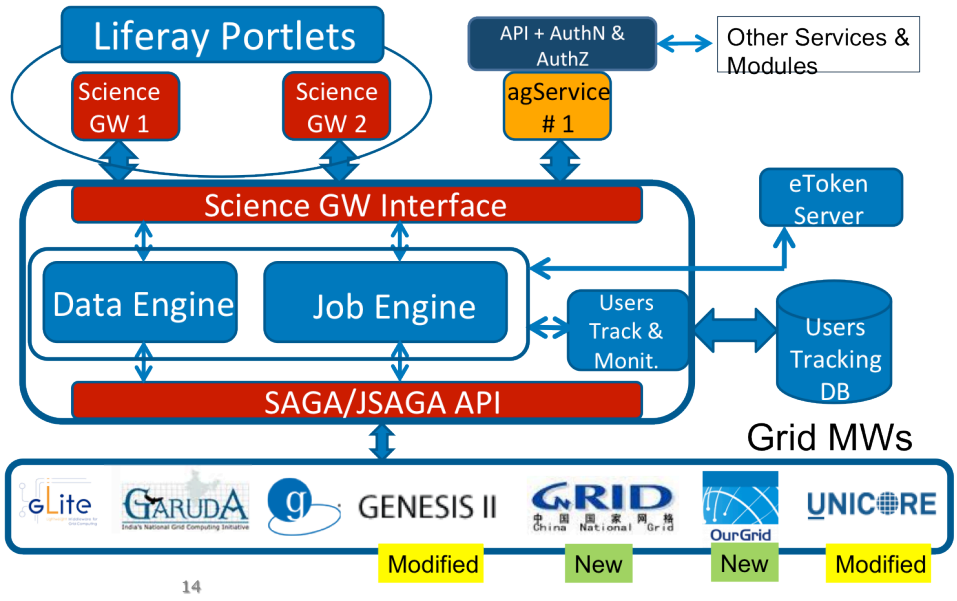


Figure 2-4: The Grid Engine

The Catania Grid Engine software layer has been developed designing a "job engine" and a "data engine" which, in turn, call the JSAGA API for respectively job and data management. The Science GW Interface also contains the functions to interact with the User Tracking DB which provides the right compliancy with the strict rules of the European Grid infrastructure VO Portal and Grid Security Traceability and Logging policies[4], each operation done by the user inside the Science Gateway is stored on a User Tracking DB that can be inspected at any time by the administrator of the portal. This ensures the non-repudiability of any Grid transaction, which is one of the most important requirements of the Grid Security Infrastructure.

Beside the GridEngine core components related to the pure Data and Job management, another important component takes care of the distributed infrastructure' authentication and authorization; in particular for all those distributed environments based on the GSI security infrastructure [15]. In general all Grid transactions must be signed with proxies generated by standard X.509 digital certificates. For this reason has been implemented in the Science Gateway a mechanism that creates proxies on the fly and on user request. This action is performed by a service called eToken server.

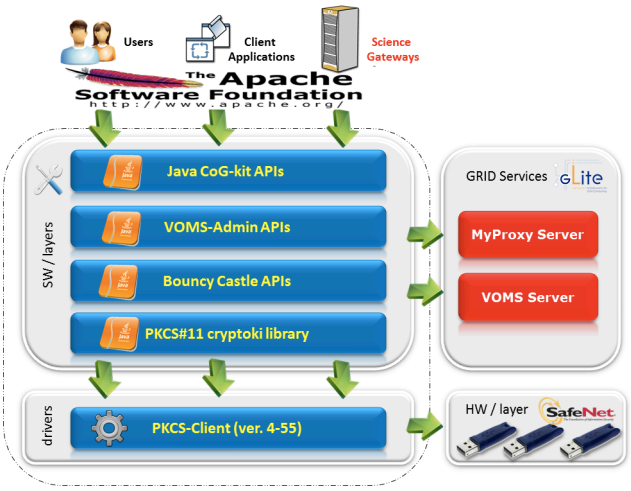


Figure 2-5 eToken Server

The eToken server generates proxies starting from robot certificates. Robot certificates are special, yet standard, digital certificates stored in USB Smart Card, referred to as etokens. It is possible to bind robot certificates with applications and allow people to run them without any personal credentials. According to this schema, when a user is authorised to access the Science Gateway and wants to run one of the authorized applications, the portal retrieves valid proxy for the eToken server on the user behalf. In case of gLite-based infrastructures, such as agINFRA the generated proxy on the fly contains the extensions specifying the VOMS role and privileges of the robot certificate inside the Science Gateway supported VOs. So that, different proxies can be created according to the different roles and privileges stored in the LDAP registry. This ensures a fine-grained authorisation and provides the portal manager with the complete control of deciding what a given user can access and do.

The core of the eToken server is a “lightweight” grid crypto library implemented according to the Service Oriented Architecture. The multi-threaded eToken server holds the web services to access the smart cards and interacts both with the Virtual Organisation and the automatic proxy renewal (MyProxy) server. A Java multi-platform client configured for inter-service communication via HTTPS completes the architecture. In order to improve the performances, the server is built on top of the Apache Tomcat Application Server and configured to accept requests only from a set of authorized “clients” (e.g., the Science Gateway). The adoption of the Apache Tomcat as Application Server ensures scalability and high performances especially when the server has to deal with huge numbers of requests. To further improve its performances and reduce the waiting time to get a proxy, the eToken server implements a mechanism for caching the proxies.

2.3.1 The Job Engine

The Job Engine is the most relevant part of the Catania Grid Engine for what concerns the access to computational Grids from the Catania Science Gateway Framework. The Job Engine manages the whole cycle of the job execution starting from its submission until the retrieval of the output. It receives requests to submit jobs from the Science Gateway interface and takes care of jobs until they are properly executed. Job operations such as submissions, statuses check and output retrieval, are mapped onto the functions of JSAGA. Moreover, the Job Engine deals with all the preliminary operations needed to execute a job such as associating a proxy to the job,

PoS(ISGC 2013)034

looking for a working resource manager, satisfying special user requirements, and so on. The Job Engine has an interface to the eToken Server to create proxies from robot certificates and associate them to a job submission. It also offers a fault tolerance capability by submitting a job to the Grid infrastructure until it will be properly executed, shielding users from possible infrastructure failures. Finally, it updates the User Tracking and Monitoring module providing the necessary input to account user operations and to control the Grid interactions rate. In order to scale to large numbers of concurrent jobs and users, the Job Engine internally makes use of the Java Thread Pools made available by the underlying portal application server (Glassfish, in our case) where Liferay is running on (see).

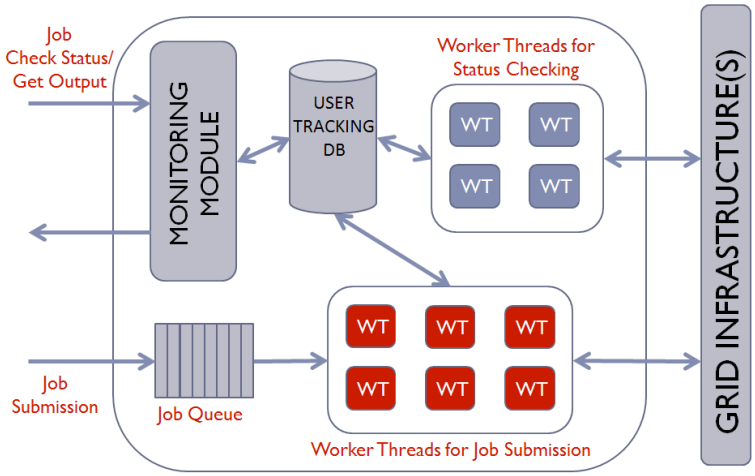


Figure 2-6: The Grid Engine' Job Module

Submitted jobs are queued and dispatched at constant rate as parallel threads of a pool so eliminating the risk that a large number of jobs can block the Science Gateway and induce a Denial of Service. The same concept of threads is used for job monitoring avoiding the Science Gateway to slow down when for instance; the user often checks the status of many jobs.

2.3.2 The Data Engine

The Data Engine realises a direct transfer between the Science Gateway and the Grid or other less specific kind of distributed Storage Elements, simulating a file system like access to the Science Gateway users. A good data service is the one that provides users with the possibility to arrange files in folders ordered in a tree in the same way a real file system does on a physical disk. A database consisting of three main tables, represented in Figure 2-6, achieves this task.

PoS(ISGC 2013)034

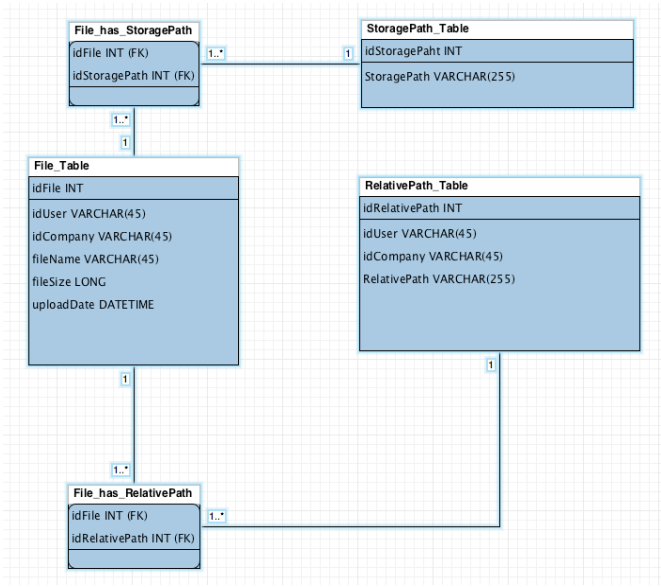


Figure 2-7: The Data Engine' Tables schema

Every operation on a file such as upload, copy, move or delete implies adding or editing some entries in the correct table. When users simply need to browse the file tree, in order to obtain the list of files, all the information stored in the database is used to create the user view of the file hierarchy. The Data Engine internally makes use of HIBERNATE an object-relational mapping (ORM) library providing a framework able to realise a mapping between the specific database instance (MySQL in our case) tables and Java objects. Database tables related to the virtual file system are not the only ones used. The European Grid Infrastructure Portal and User Traceability policies require to track every single Grid transaction performed by the Science Gateway, for this reason also the Data Engine is integrated with the User Tracking Database module which stores any user transaction done by the Science Gateway users over the distributed files.

The DataEngine provides client APIs that satisfy the SRM v2.2 standard protocol to interact with the storage elements of SToRM and DPM kind. The Scientific Data Management Research Group Lawrence Berkeley National Laboratory, University of California, US[2], has developed a java implementation of the SRM v2.2 protocol.

2.3.3 Web Applications

The remaining module of Catania Science Gateway Framework as shown in Figure 1-2 consists of the user interface made of a set of web applications. For the development of these basic elements of the Science Gateway, the JSR 286 standard (also known as "portlet 2.0") was adopted. These applications, normally in the form of portlets are used and they can be exchanged among different portals or many of them arranged together to build up complex and sophisticated scientific environments. As portlet container, the award winning Liferay [14] portal framework has been chosen which offers a rich, easy-to-use "web 2.0" interface using AJAX and other presentation layer technologies. It features effortless GUI-based

[2] https://sdm.lbl.gov/bestman/

PoS(ISGC 2013)034

personalization, drag-and-drop portlets, dynamic navigation, and an instant-add portlet library. The portal platform also integrates with most used packages such as YUI3 and jQuery and with the JavaScript library of the portal developers. Liferay is currently the most used framework to build Science Gateways in the Grid world.

3. Use Cases

In this chapter will be examined in detail the Catania Science Gateway Framework implementation build up for the agINFRA project. As already described in the introduction, the main capabilities offered by the framework are the possibility to execute software into a distributed environment and access to a remote hosted service protecting or not its access through the portal AAI module.

All these possibilities will be examined in detail by the next chapters.

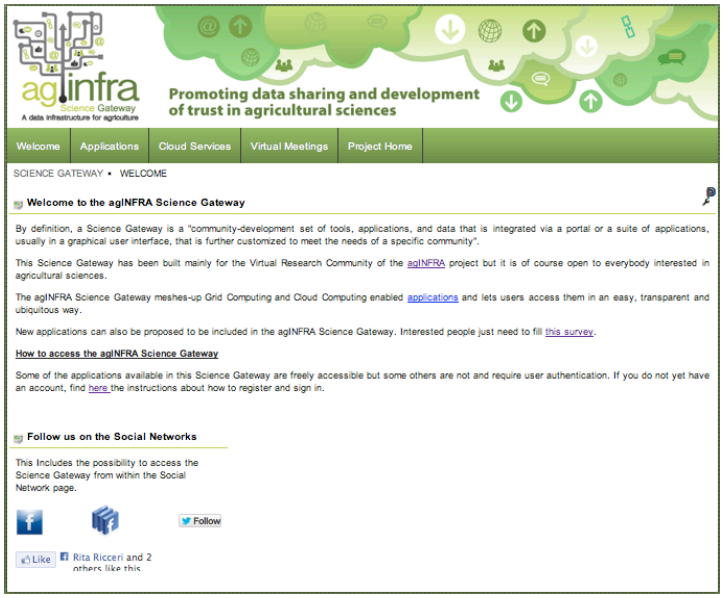


Figure 3-1 agINFRA science gateway home page

The overall portal look and feel recalls the main project web site and the only change consists of the different menu voices briefly described below:

- **Welcome**, points to the agINFRA science gateway main page
- **Applications**, points to a special portlet, which shows the list of all applications installed in the portal.
- **Applications/<App Name>;** under the Applications menu, a list of installed applications is available. If the user clicks on one of these voices, the specific application description page will be presented. In this page there will be a 'Run' link or button that points to the application input form. Once the user fills the input form, the application can be executed into any supported distributed infrastructure. The application 'Run' link does not appear in case the user is not authorised or not yet logged in.
- **Cloud Services;** this menu voice points to the list of available services hosted by Cloud nodes. Before to access any service from this menu the user must be logged-

PoS(ISGC 2013)034

in. At the moment this page reports all those services hosted by the CLEVER cloud management system.

- **Virtual Meetings**; this voice points to the agINFRA project' online web conferencing system.
- **My Workspace**; points to a special portlet that monitors the user job submissions and manages the job outputs.
- **My Cloud;** points to a special portlet that allows manage cloud nodes. This application can be seen and accessed by a particular kind of allowed users: 'Cloud Managers'.
- **Project Home**; points to the agINFRA project main web site.

3.1 The Applications

The next chapters will introduce a list of applications that have been specifically developed for the agINFRA science gateway. Each of these applications will be briefly examined below.

3.1.1 Agris XML AP 2 RDF

This application has been developed during the first project-training event targeted for high-level application developers. This application was at the beginning just a demonstration of the Catania Science Gateway Framework capabilities but later on it has been improved and installed in the production portal since its usage simplicity. This application converts the AGRIS AP XML files into the corresponding RDF format, splitting the given input file in several pieces and then each input file processed by a dedicated Grid job. This application is very useful when lot of big files are given as input. The FAO organisation, which is one of the partners of the agINFRA project, is using this application to convert XML files into the corresponding RDFs records. The produced RDFs are then used to populate a new introduced semantic database.

3.1.2 Agrovoc Tagging

The Agrovoc Tagging application takes as input a big text file coming from a web crawling process and its purpose is to index documents with the Agrovoc Thesaurus (English version). Starting from a list of URLs (the web Crawler output), the application generates a TXT file containing the mapping: DOCUMENT_URL=LIST_OF_EXTRACTED_KEYWORDS. In the future this application will be modified in order to store the output directly in a triple-store. This application is still targeted to the FAO project partner facilitating the Tagging process which requires lot of computational power.

3.1.3 Weka

WEKA consists of various machine-learning algorithms that can be applied to Data Mining problems. WEKA was developed at the University of Waikato, New Zealand. It is open source and issued under the GNU General Public License. All WEKA algorithms are developed in Java and callable by a given set of libraries in the form of jar files.

The application input interface requests a file and then a list of possible arguments in order to process the same input file with different WEKA algorithms with a dedicated Grid job. This

PoS(ISGC 2013)034

application has been developed by a student of the Indian Statistical Institute and used as working tool for its study case.

3.1.4 Rice Info

RiceInfo is a search interface providing results from an ontological database about pests (like viruses, worms, weeds, insects etc.), nutritional deficiency, toxicity, control mechanisms (for pests, nutritional deficiency, toxicity, diseases etc.) and related information. The result displayed for the RiceInfo ontology is in N3 triple format. This interface also integrates search results from DBpedia and GoogleScholar. Also this application is a study case for a student of the Indian Statistical Institute.

3.1.5 R Statistical Analysis tool

R is a language and environment for statistical computing and graphics.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

The application allows any user to upload a R macro file or directly type the macro into a dedicated input form and then execute the R macro into the Grid. R application is a generic tool and for this reason introduced in the agINFRA portal for disseminative purposes.

3.2 Hosted services

Together with the classic interfaces able to execute jobs into a distributed infrastructure, the Catania Science Gateway framework allow to access remote hosted services from its portal providing both secured and insecure access to its contents. The figure below shows the architecture of this solution that has started as a prototype with just one agINFRA service and then replicated by other services especially for those hosted by Cloud systems.

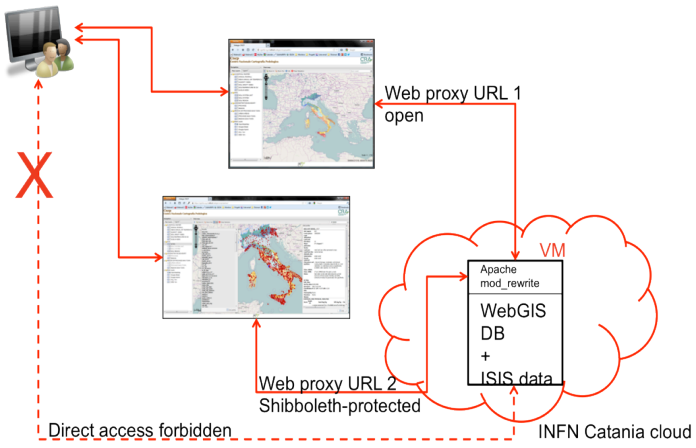


Figure 3-2 Remote service hosting architecture

PoS(ISGC 2013)034

As show by the Figure 3-2 any user direct connection is not allowed. The only way to access the remote service is from the Science Gateway, which maps different possible URLs into protected and public accessible pages. As soon the user request a protected URL, the portal provides the IdF/IdP selection and login window and, in case the user is allowed to access the service checking the authorization LDAP records, the service URL will be shown.

The following chapters will examine this kind of services.

3.2.1 Italian Soil Information System (SISI)

This is the first service integrated in the agINFRA Science Gateway portal and it is running for the online Italian Soil Data Consultation (CRA) [17] who developed the software. SISI is made up of a hierarchy of geo-databases, which include soil regions and aim at correlating the soils of Italy with those of other European countries with respect to soil typological units (STUs), at national level, and soil sub-systems, at regional level.

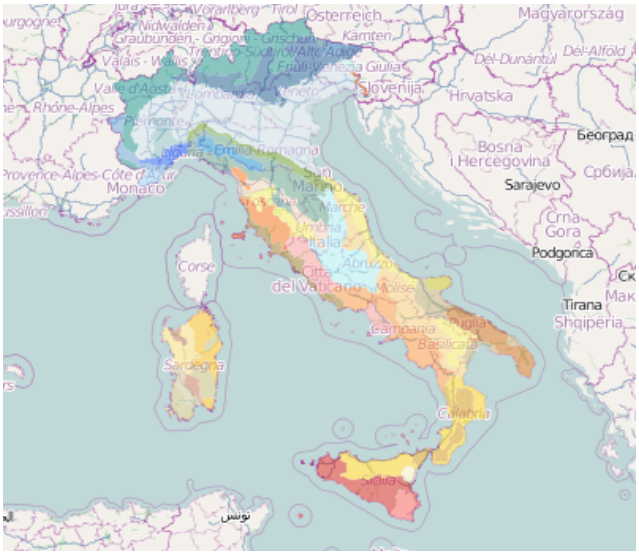


Figure 3-3 ISIS Italian Soil Information System

SISI provides an inventory of Italian soilscapes at two reference scale, Soil region (1:5.000.000) and Soil systems (1:500.000). Soil regions are a regionally restricted part of the soil cover characterized by a typical climate and parent material association. Soil systems illustrate main Italian soilscapes and are composed of homogeneous areas as for physiography, lithology, river drainage network, and land cover.

The portal allows two different kinds of user access. A public accessible mode, which provides a limited set of features, and a restricted mode accessible only by allowed and registered users. The restricted mode allows the users to access to all the features provided by this service.

3.2.2 Annotate and Browse Soil Maps

This application is in close relationship with the ISIS service and provides access to the metadata associated to the ISIS soil map files. In particular it is possible to add to the existing metadata one or more user annotations. On top of the existing map metadata and even user annotations it is possible to make queries in order to discover a particular map. The annotation

PoS(ISGC 2013)034

feature is well appreciated in all those environments that require the sharing of data among big or more in general virtual communities.

3.3 CLEVER Cloud

The Catania Science Gateway Framework allows managing services hosted by Clouds and in particular by cloud nodes managed by the CLEVER Cloud management system.

CLEVER [18] is an innovative Cloud middleware designed and developed at the University of Messina for managing virtual appliances supplying an abstraction in the management of virtual resources providing a useful and easy management of private/hybrid clouds: by handling simple and easily accessible interfaces, it allows the possibility to interact with different "interconnected" computing infrastructures.

CLEVER aims to provide Virtual Infrastructure Management services and suitable interfaces at the High-level Management layer to enable the integration of high-level features such as Public Cloud Interfaces, Contextualization, Security and Dynamic Resources provisioning.

The middleware is based on a distributed clustered architecture, where each cluster is organized in two hierarchical layers, as depicted in Figure 3-4. CLEVER nodes contain a host level management module, called Host Manager (HM). A single node may also include a cluster level management module, called Cluster Manager (CM). The CM contains the intelligence for treating and analysing all incoming data whereas the HM has simple characteristics at lower level. Indeed it represents the remote agent of the CM. Thus, we have in the cluster at least one active CM at higher layer and, at lower layer, many HMs depending on it. A CM acts as an interface between the clients (software entities, which can exploit the Cloud) and the software running on the HMs. The CM receives commands from the clients, gives instructions to the HMs, elaborates information and finally sends results to the clients. It also performs the management of resources (uploading, discovering, etc.) and the monitoring of the overall state of the cluster (workload, availability, etc.).

Both CMs and HMs are composed by several sub-components, called agents, which are designed to perform specific tasks. Since agents are separated processes running on the same host, (due to fault tolerance purposes) internal communication is based on Inter Process Communications (IPCs), such as exploiting D-Bus interfaces. We call this Internal Remote Method Invoker (IRMI) communication, which refers to the message exchanging protocol among agents within the same manager (both in CMs and HMs).

On the other hand, External Remote Method Invoker (ERMI) communications allow to CM agents to exchange messages with the HMs ones. It is based on the XMPP protocol [19], which was born to drive the communications in the heterogeneous instant messaging systems, where it is possible to convey any type of data. In particular, the protocol is able to guarantee the connectivity among different users even with restrictive network security policies (NAT transversal, firewalling policies, etc.). The XMPP protocol is also able to offer a decentralized service, scalability in terms of number of hosts, flexibility in the system interoperability and native security features based on the use of channel encryption and/or XML encryption. The current implementation of CLEVER is based on the employment of an Ejabberd XMPP server [20].

About storing persistent information of the infrastructure, the current implementation of CLEVER is based on a specific plugin able to interact with the Sedna native XML database (see [21]). Sedna allows the possibility of creating incremental hot backup copies of the databases

PoS(ISGC 2013)034

and supports ACID transactions. It has been also chosen because it natively supports the XML data containers.

Another important thing to report, is that CLEVER has been designed with an eye toward horizontal federation, thanks to the choice of using XMPP for module communication, made thinking about the possibility to support in the future also inter-domain communication between different CLEVER administrative domains.

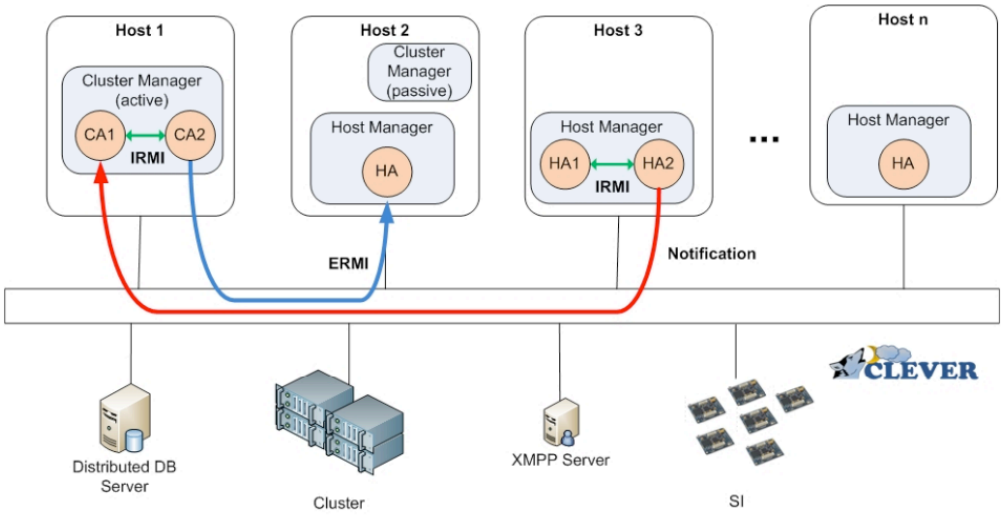
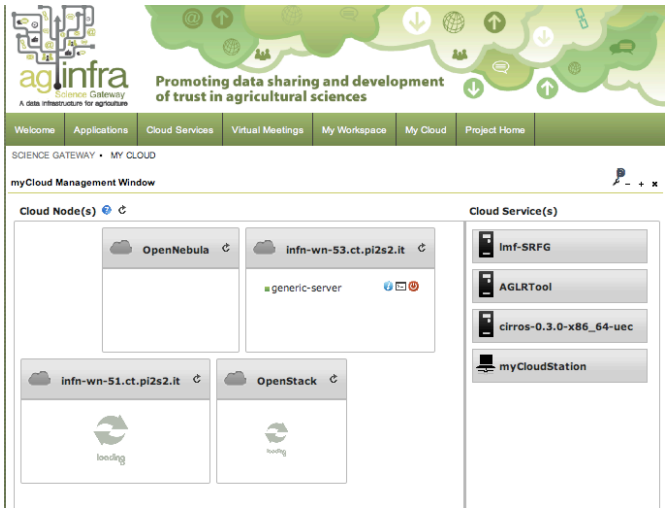


Figure 3-4: CLEVER middleware

3.3.1 MyCloud portlet

The Catania Science Gateway Framework provides a special portlet able to manage Cloud node instances with a high-level user interface. With the MyCloud portlet it is possible to instantiate new virtual machines, move virtual machines among different nodes and many other operations just with simple mouse clicks and dragging operations, normally accomplished by Cloud managers by command line utilities.



PoS(ISGC 2013)034

Figure 3-5 MyCloud portlet

One of the most recent changes in CLEVER and MyCloud portlet is the possibility to define a special Node able to instantiate and manage virtual machines into Nebula or OpenStack Cloud Management System, seamlessly.

4. Conclusions

Services and tools explained by the previous chapters offer to the agINFRA user community a very powerful environment to easily, securely and seamlessly access a huge amount of distributed data and computation power. The adoption of the Catania Science Gateway paradigm has a key role for the dissemination and achievement of the very challenging project objective, to build up a reliable and freely accessible agricultural data infrastructure. The agINFRA portal is one of the main entry points for the agricultural user community and even from other services accessing data from other Web portals. All these operations are performed hiding the complexities behind the exploitation of the distributed infrastructures such as Grids, Clouds, and any other kind of distributed data storage. The adoption of standard components in developing the Catania Science Gateway Framework is key element to reach the technological sustainability of the whole architecture.

PoS(ISGC 2013)034

References

[1] Valeria Ardizzone, Roberto Barbera, Antonio Calanducci, Marco Fargetta, E. Ingrà, Ivan Porro, Giuseppe La Rocca, Salvatore Monforte, R. Ricceri, R. Rotondo, Diego Scardaci, Andrea Schenone: "The DECIDE Science Gateway". Journal of Grid Computing Vol. 10(4), pages 689-707, Editor Springer 2012.

[2] www.fao.org

[3] Teragrid/Xede, https://www.xsede.org

[4] EGI Portal traceability policy - https://documents.egi.eu/public/ShowDocument?docid=80

[5] saml.xml.org

[6] shibboleth.internet2.edu

[7] simplesamlphp.org

[8] www.edugain.org

[9] www.geant.net

[10] gridp.ct.infn.it

[11] idpsocial.ct.infn.it

[12] saga-project.github.io

[13] http://grid.in2p3.fr/jsaga/

[14] www.liferay.com

[15] www.globus.org/security/overview.html

[16] http://clever.unime.it/&#8206

[17] www.entecra.it

[18] F. Tusa, M. Paone, M. Villari, and A. Puliafito., "CLEVER: A CLoud-Enabled Virtual EnviRonment," in 15th IEEE Symposium on Computers and CommunicationsS Computing and Communications, 2010. ISCC '10. Riccione, June 2010

[19] The Extensible Messaging and Presence Protocol (XMPP) protocol. Retrieved June 10, 2012, from http://tools.ietf.org/html/rfc3920/

[20] Ejabberd, the Erlang Jabber/XMPP. Retrieved June 10, 2012, from http://www.ejabberd.im/

[21] Sedna, Native XML Database System. Retrieved June 10, 2012, from http://http://www.sedna.org/

PoS(ISGC 2013)034