Project Acronym:	INDICATE
Project Title:	International Network for a Digital Cultural Heritage e-Infrastructure
Contract Number:	261324
Starting date:	1 September 2010 **Ending date:** 31 August 2012

Deliverable Number:	D 4.2
Title of the Deliverable:	Pilot on e-Collaborative digital archives
Task/WP related to the Deliverable:	Task 4.2 e-Collaborative Digital Archives protected by access control and rights management
Type (Internal or Restricted or Public):	Public
Author(s):	Roberto Barbera, Antonio Calanducc
Partner(s) Contributing:	All

Contractual Date of Delivery to the CEC:	29 February 2012
Actual Date of Delivery to the CEC:	26 March 2012
Date of Delivery of the addendum:	26 July 2012

Project Co-ordinator

Company name :	Istituto Centrale per il Catalogo Unico (ICCU)
Name of representative :	Rosa Caffo
Address :	Viale Castro Pretorio 105, I-00185 Roma
Phone number :	+39.06.49210427
Fax number :	+39.06.4959302
E-mail :	rcaffo@beniculturali.it
Project WEB site address :	http://www.indicate-project.eu

Context

WP 4	Research Pilots
WP Leader	NTUA
Task 4.2	Task 4.2 e-Collaborative Digital Archives protected by access control and rights management
Task Leaders	GARR and COMETA
Dependencies	
Starting date	September 2010
Release date	February 2012

Author(s)	Roberto Barbera, Antonio Calanducci, Maria Laura Mantovani, Sabrina Tomassini, Federica Tanlongo, Gabriella Paolini
Contributor(s)	All
Reviewers	Sergi Fernandez, Franc J. Zakrajsek, Mercè López
Approved by:	Rossella Caffo (ICCU)

HISTORY

Version	Date	Author	Comments
0.1	14 February 2012	Roberto Barbera	First version, index and structure
0.2	25 February 2012	Antonio Calanducci	Content creation and summary
0.3	25 February 2012	Roberto Barbera	Conclusions added and whole document checked and revised
0.4	06 March 2012	Roberto Barbera	Names of the internal reviewers added
0.5	21 March 2012	Sergi Fernandez, Mercè López	Internal review
0.6	24 March 2012	Roberto Barbera	Internal review comments and corrections applied
1.1	26 July 2012	Maria Laura Mantovani	Addendum about Identity Federation

Table of Contents

1 Executive summary

This document shows the architecture and implementation of the INDICATE e-Culture Science Gateway that is the platform used to create the e-Collaborative Digital Archives deployed on the COMETA e-Infrastructure. Particular attention has been paid to simplify the access to non-expert users and to define a fine-grained authentication and authorization system that could protect digital cultural contents in an effective way.

We start presenting the reasons and the benefits of adopting the Science Gateway paradigm as a means to provide a simplified access for users to the contents of the digital archives.

In particular, we will describe how we achieved user authentication and authorization integrating Science Gateways single-sign-on mechanism with Identity Federations and user tracking and logging for any Grid transaction.

Then, we present the gLibrary platform that have been used to create the repositories on the storage resources and metadata service of the Grid infrastructure used, and the gLibrary APIs to create a set of portlets, deployed into the e-Culture Science Gateway, to provide an easy-to-use front-end for discovering, finding and retrieving assets of the three digital archives implemented.

2 Introduction

The goal of this pilot application is to demonstrate how to implement e-Collaborative Digital Archives on e-Infrastructures and protect them with access control and rights management, giving birth to a real e-Culture Science Gateway.

Three repositories have been successfully created and deployed on the e-Infrastructure platform provided by COMETA and insisting on the GARR research network:

- The **Federico De Roberto literary works archive** (De Roberto DR);
- The **Architectural and Archaeological Heritage present in Mediterranean Area** (MED Repo);
- A **China Relics Digital Repository** (China Relics).

The Indicate e-Culture Science Gateway (e-CSG) aims at proposing a model to enable transparent access to the Digital Cultural Heritage for millions researchers all around the world. However, the management of authorisation procedures, if implemented in a traditional way, i.e. assigning credentials to each new user and maintaining them during their lifecycle, would imply a significant overhead for the e-CSG manager. In the meantime, end users would also get an additional set of credentials to be remembered and kept private, with the usual drawbacks: usage of weak password, re-use of the same password (thus weakening security levels) and risk of identity theft.

For these reasons, we decided to implement the Federated Access to the eSCG.

This approach offers a number of advantages:

- The pool of potential users dramatically increases and it is immediately extended to all end users belonging to existing identity federations supported by the e-CSG;
- The e-CSG manager is exonerated from creating and keeping on its servers the users' credentials, as they are managed by Identity Providers at single Federated organizations which connect to the e-CSG;
- End users don't need to obtain, manage and remember a new set of credentials, and use the usual credentials provided by their home organization.

For this pilot, we started with the integration of IDEM[1], the Italian AAI Federation dedicated to Research, Education and Culture, managed by GARR[2]. However, this approach is easily extensible to any other Federation basing on the SAML OASIS standard[3] and indeed, as of today, the e-CSG is a Service Provider of the following Identity Federations: CARSI[4] (China), GRNET-AAI[5] (Greece), SIR[6] (Spain) as well as of the eduGAIN[7] international inter-federation.

[1] http://www.idem.garr.it

[2] http://www.garr.it

[3] http://saml.xml.org/saml-specifications

[4] http://www.carsi.edu.cn/

[5] http://aai.grnet.gr/

[6] http://www.rediris.es/sir/

[7] http://www.edugain.org

3 Enabling transparent access to e-Collaborative Digital Archives through Science Gateways and Identity Federations

One of the main obstacles for non-IT-expert users to exploit e-Infrastructures, such as Grids, is the fact that they are based on complex security mechanisms such as Public Key Infrastructures (PKI) and accessed through low level (command-line based, I.e. non-graphical) user interfaces. The approach used to solve both the previous problems and make available the previous listed digital repositories to the largest possible number of users, was to deploy them into a "Science Gateway" whose access is regulated by "Identity Federations".

In the recent past, interesting developments have been independently carried out by the Grid community with the Science Gateways and by the National Research and Education Networks with the Identity Federations to ease, from one side, the access and use of Grid infrastructures and, from the other side, to increase the number of users authorised to access network-based services.

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 (US Teragrid project).*"

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 $O(10^7)$.

To address the issue of the use of e-Infrastructures, The Italian National Institute of Nuclear Physics and the Consorzio COMETA are developing since almost two years, a new type of Science Gateway that implements an authentication schema based on Identity Federations, that has been adopted to deploy the e-collaborative Digital Archives discussed in this document, referred since now on as an e-Culture Science Gateway (e-CSG).

3.1 Linking the e-CSG to Identity Federations

The IDEM GARR Federation brings together organizations in the field of Education, Research and Culture, namely Universities, Research Institutes such as the National Research Council, the Italian Institute of Nuclear Physics and the Italian Institute of Astrophysics, supercomputing centres, medical research centres, National Libraries and Museums, and other cultural institutions.

Organizations subscribing to IDEM link their Identity Provider (IdP) to the Federation. An Identity Provider is a service which enables end users belonging to the organization to use their usual credentials, and more generally their Digital Identity, in order to connect not only to resources provided by their own organization, but also to those offered by other federated organizations. Thanks to the federated approach, once the e-CSG links to a specific Federation, all end users belonging to that federation are immediately enabled to ***authenticate*** into the e-CSG. This does not imply that they are automatically ***authorised*** to do so: as a matter of fact, this approach decouples the authentication and authorization steps, the first one being demanded to the Federation's IdPs, and the second remaining with resource owners and implementing their own access policies.

For example, supporting the IDEM GARR Federation enabled a pool of about 3 million *potential* users belonging to all federated organizations (currently amounting to 35, but steadily growing) to connect to the e-CSG. Each of these users will need to be authorised to access a specific resource within the e-CSG

according to its owner's policies. So, different user groups will access different subsets of resources, and have different rights on them.

3.2 User authentication on the e-CSG with Identity Federations and Identity Providers

One of the strengths of this e-CSG, and of Catania Science Gateways in general, is the decoupling of the authentication (AuthN) phase from the authorization (AuthZ) one. In order to access the e-CSG, a user must be both authenticated and authorized but we treat the two steps separately and with different technologies. The schema for authentication and authorization is depicted in fig. 1. User authentication relies on Identity Providers (IdPs) that are members of one or more Identity Federations. We only support federations based on the SAML 2.0 standard specifications and on its implementation done by Shibboleth and SimpleSAMLphp.

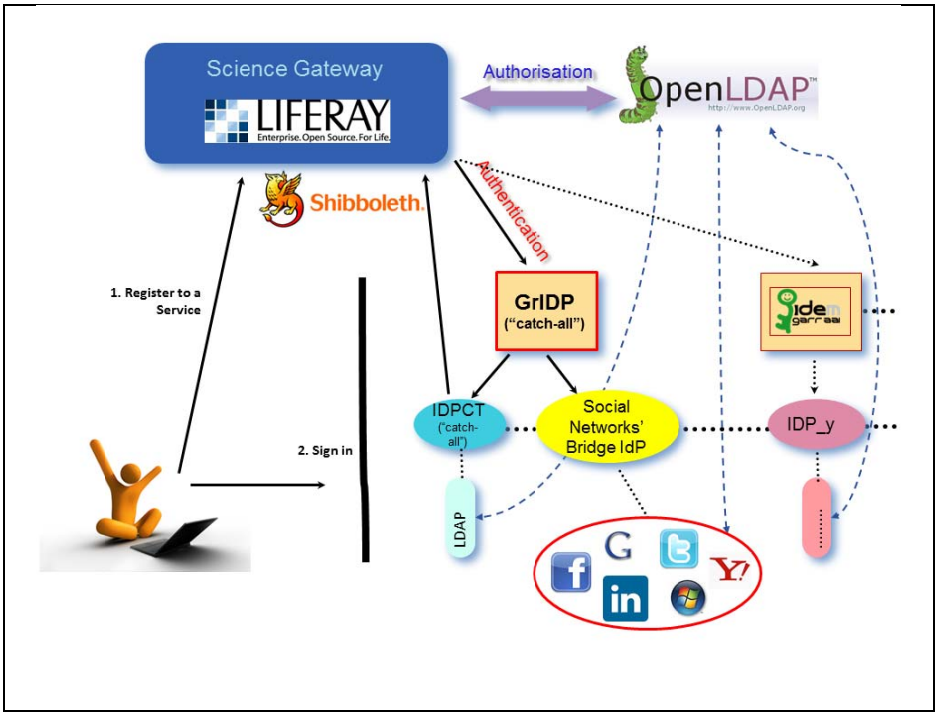

Figure 1. Authentication and authorization schema of the Science Gateway.

We currently support several official Identity Federations and the INDICATE e-CSG is already registered as Service Provider of the eduGAIN inter-federation service within the GÉANT project.

We also support all the Identity Providers of the Grid IDentity Pool (GrIDP), a "catch-all" Identity Federation that we have expressly created to gather all the IdPs that do not already belong to any official federations and all the users of the e-CSG 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, we have also created a special IdP, the"Social Networks' Bridge Identity Provider", that allows people to get authenticated with the same credentials they already have with the most known and populated social networks.

3.2.1 Supporting Identity Federations: the technical steps

The following are the technical steps implemented in the pilot in order to link the e-CSG with the IDEM GARR federation.

3.2.1.1 Integrating the Service Provider funcionality into the e-CSG

Being basically a web portal, the eCGS can be easily integrated with the Relying-party functionality (also known as Service Provider) from the SAML Web Single-Sign-On profile. This can be done by exploiting several software frameworks. In the pilot we chose the Shibboleth Service Provider solution. Once implemented Service Provider functionalities on the e-CSG, the service can be registered as a resource in the Federation.

3.2.1.2 Installing a Discovery Service

A discovery service provides a browser-based interface where a user selects his/her Identity Provider. The service provider uses this information to initiate SAML Web Browser SSO.

By activating a dedicated Discovery Service, it is possible to support other federations and single IdPs that are not integrated in any federation. In the pilot, the selected implementation for this component is Shibboleth's Centralized Discovery Service.

3.2.1.3 Registering the e-CSG INDICATE as a federated resource

In order to become part of a Federation, an Organization needs to subscribe an agreement with the Federation itself. The agreements imply accepting shared security policies and relating to the offered services and management of identity data shared by federation members. By accepting these policies and security standards, federation participants build a network of trust, which allows Service Providers to accept Digital Identities that are guaranteed by Identity Providers without any need to verify them. This trust also allows Identity Providers to share the users' attributes with Service Providers, who are bound to use them according to the Federation's rules.

In this case, the organization responsible for the e-CSG, i.e. Consorzio COMETA, had to subscribe to IDEM Federation. Once registered the e-CSG as an IDEM resource, all IDEM end users are immediately enabled to authenticate in it.

Likewise, it was possible to register the e-CSG in other European Federations such as the Chinese (CARSI), the Greek (GRNET-AAI), and the Spanish (SIR) ones. The e-CSG was also registered as a resource of the eduGAIN interfederation.

eduGAIN is intended to enable the trustworthy exchange of information related to identity, authentication and authorisation between the GÉANT[8] Partners' federations. To this end, eduGAIN co-ordinates elements of the federations' technical infrastructure and offers a common policy framework controlling the exchange of this information. Its initial goal is to enable Pan-European Web Single Sign On (Web SSO) to both GÉANT services and to those provided by other communities represented by, or associated with, the GN3 partners. Currently, eduGAIN includes 14 Federations: Spain, Italy, Greece, Croatia, Hungary, Switzerland, Belgium, Germany, Czech republic, Norway, Sweden, Finland, Brazil and the Netherlands. In addition, France, Latvia and Canada are in process to enter the inter-federation and Turkey and Poland are in a pilot phase. Thanks to the participation in eduGAIN, users from the member Federations can be automatically enabled to access the e-CSG.

Identity Federations in the Field of Education and Research and similar to those already integrated are already operating in Countries involved in the INDICATE project: FER (France)[9], ARNES-AAI (Slovenia)[10] and ULAKBIM (Turkey) could be easily linked to the e-CSG.

[8] http://www.geant.net

[9] https://services.renater.fr/federation/index

3.2.2 The login procedure

1. The user opens the URL https://indicate-gw.consorzio-cometa.it and chooses the “Sign In” option (on the upper-right part of the webpage).

Figure 1-a. Home page of the INDICATE e-Culture Science Gateway

2. The user is redirected to the Discovery Service, i.e. a webpage where s/he selects his/her federation and home organization.

[10] http://www.arnes.si/en/services/arnesaai.html

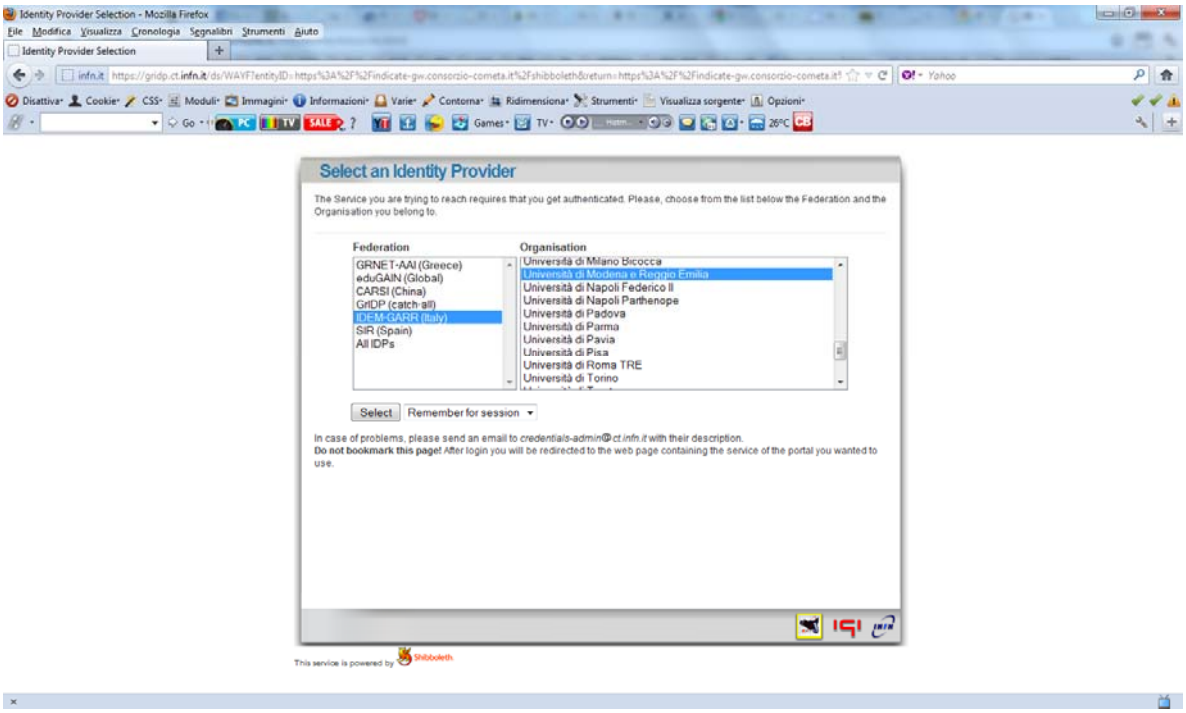

Figure 1-b. Selection of the Identity Federation and Identity Provider from the discovery service

3. The user is now redirected to his/her usual IdP login webpage, where s/he logs in with his/her credentials.

Figure 1-c. Login page of the selected Identity Provider

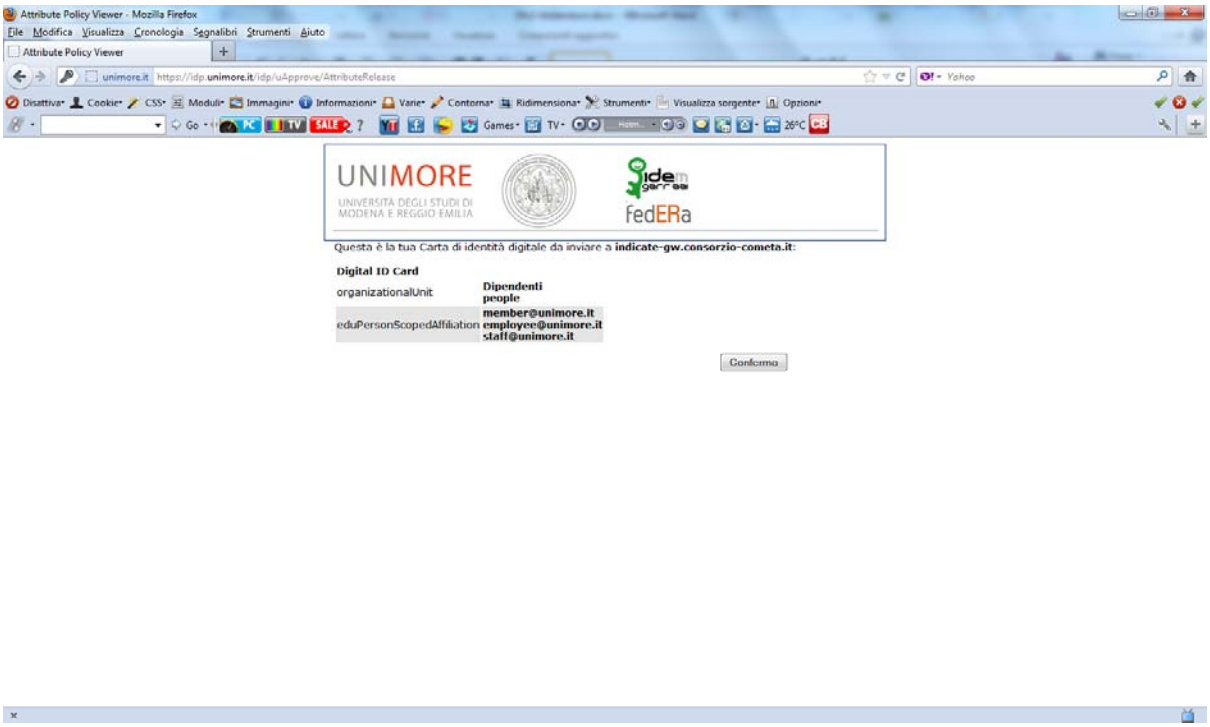

Figure 1-d. Validation of user credentials by the Identity Provider

4. The user is now logged-in and redirected to the e-CSG, where s/he will access only those resources for which s/he got an authorisation.

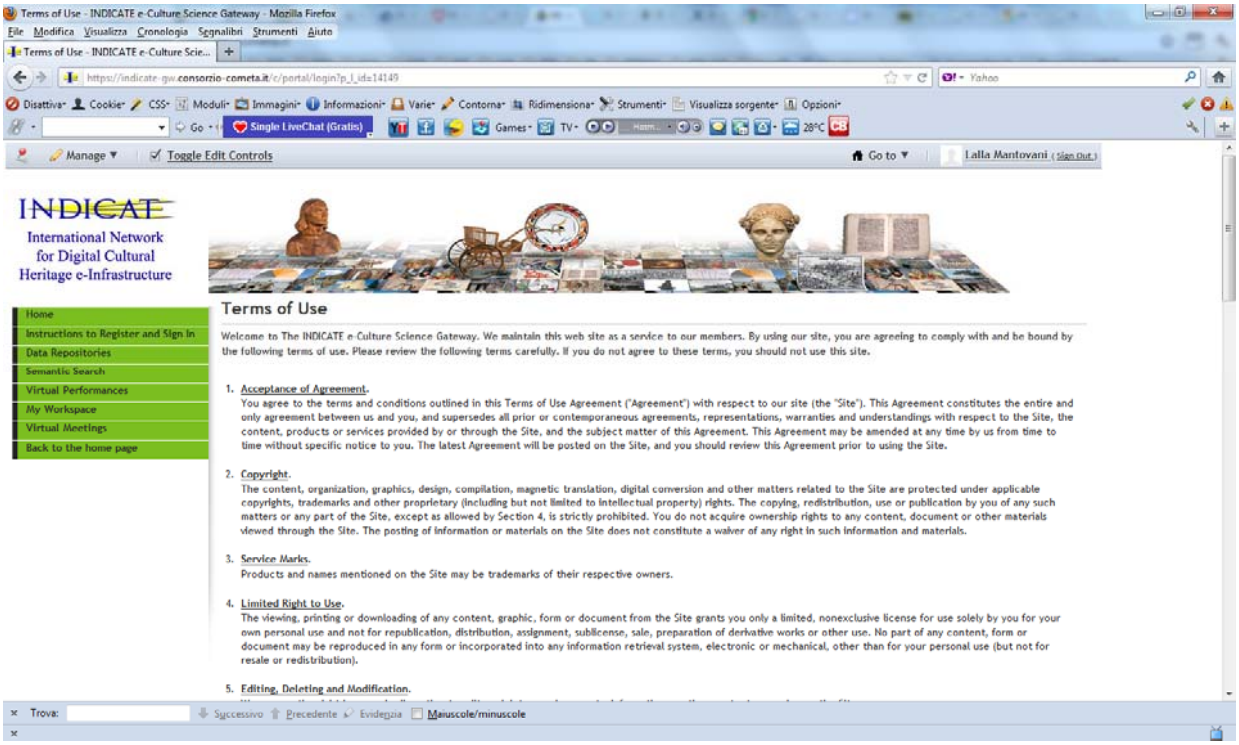

Figure 1-e. Redirection to the e-CSG home page with authentication confirmed

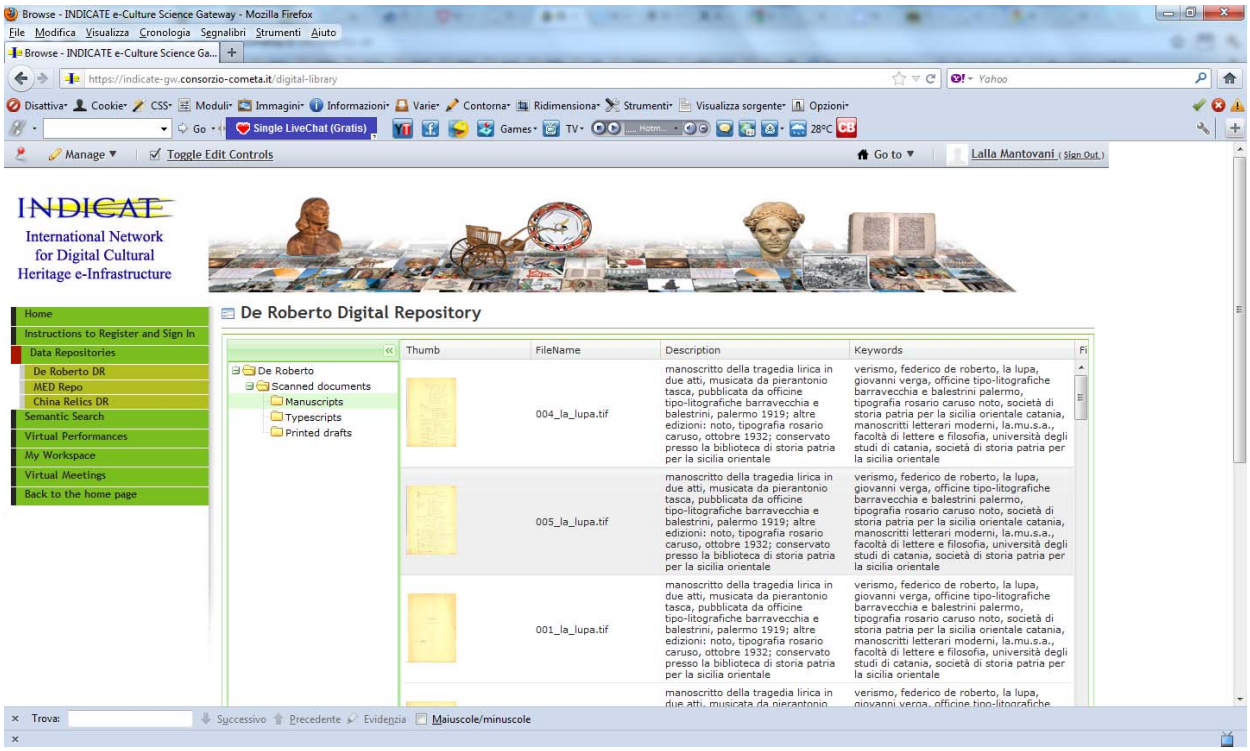

Figure 1-f. Browsing of the contents of the e-CSG

3.3 User authorization on the e-CSG

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

The workflow by which users can register to the e-CSG is shown in fig. 2. A user points her browser to the e-CSG website and asks to be registered by filling a dedicated form. Here, she can specify the Identity Federation she belongs or ask to be enrolled as a member of GrIDP. The request of registration, once it is confirmed by the user via email, is then forwarded to the administrators of the portal. If it is accepted, user information is stored on the LDAP registry and the user is notified that she can sign in. Otherwise, she is notified that her request has been denied. This procedure has been put in place in order to ensure that authorizations are not provided automatically to everybody and that a check be done on the requests by a human being.

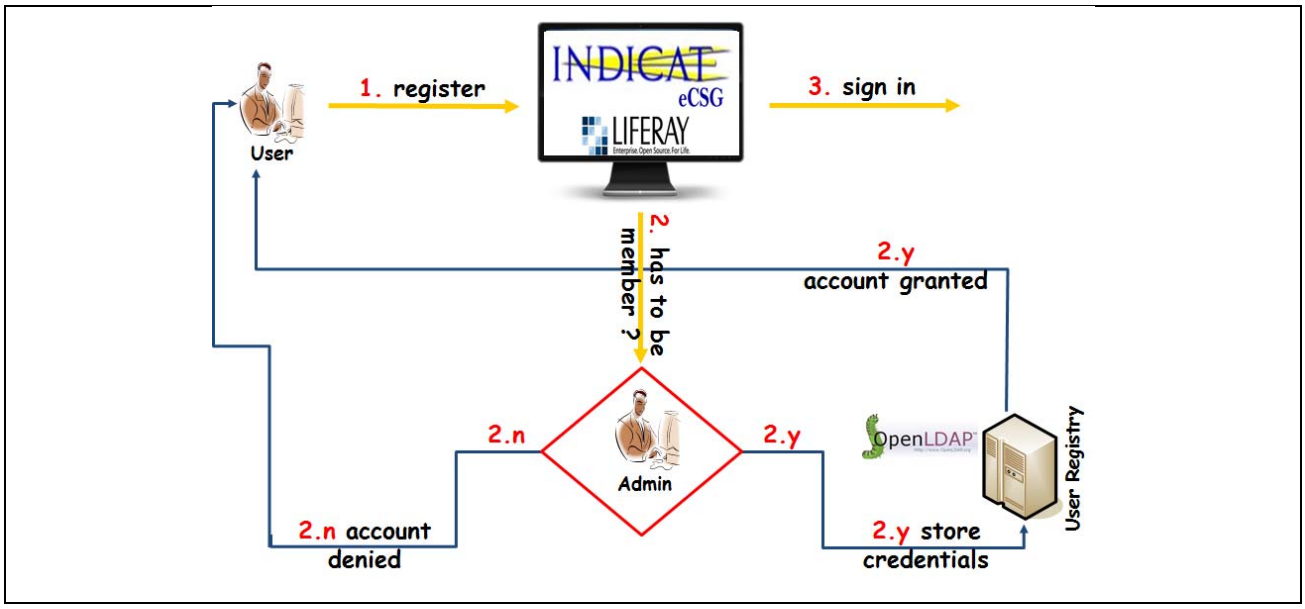

Figure 2. Workflow of the registration procedure

Once a user has been authorized to access the e-CSG, he/she can then sign in and run the applications he/she is allowed to from within the portal. The workflow of this phase is depicted in fig. 3. When the user signs in, he/she is asked to select in a web page the Identity Federation and the Identity Provider he/she belongs to.

Then, he/she is redirected to the login page of his/her Identity Provider where he/she can insert his/her credentials. If they are correctly verified, the control returns to the e-CSG that checks if the user is inserted in the LDAP registry. If he/she is, the user is then presented with the digital archives he/she has the privilege to access to. At this time, the portal contacts an eToken server that returns a valid "proxy" certificates (see next section) to be used to perform any Grid transaction (upload/download/metadata editing).

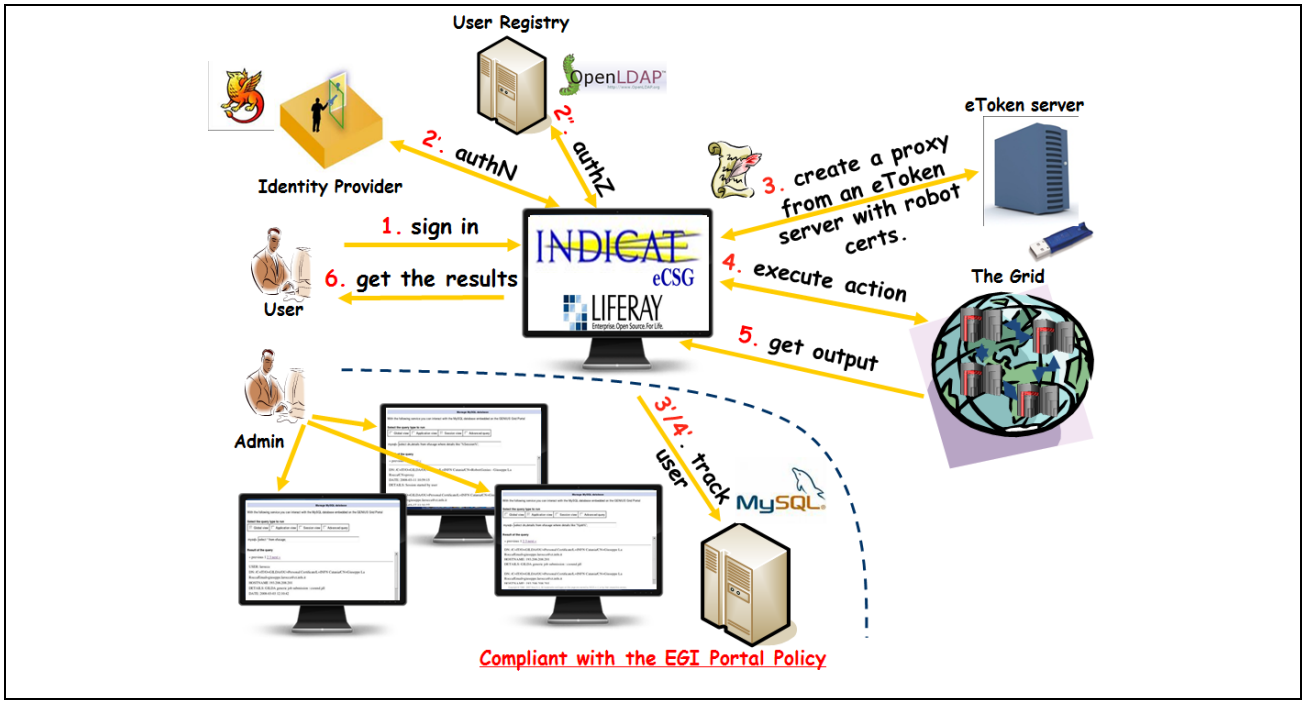

Figure 3. Workflow of the sign in procedure

3.4 User tracking and logging on the e-CSG

It is worth mentioning that, in order to be compliant with the strict rules of the European Grid Infrastructure VO Portal and Grid Security Traceability and Logging policies, each operation done by the user inside the e-CSG Gateway is stored on a User Tracking Database that can be inspected at any time by the administrator of the portal. This ensures the non-repudiability of Grid transactions which is one of the most important requirements of the Grid Security Infrastructure.
The authentication and authorization mechanism described above has the big advantage of being based on standards and greatly simplifies the access to e-Infrastructure by non-IT-expert users avoiding the need for them to get and manage personal digital certificates. However, all Grid transactions must be signed with proxies generated by standard X.509 digital certificates so we have implemented in the e-CSG a mechanism that creates proxies on the fly and on user request. This is done by a service called 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 **digital repositories** and allow people to access them without any personal credentials. According to this schema, when a user is authorized to access the e-CSG and wants to access one of the digital archives she is allowed to, the portal retrieves on her behalf a valid proxy for the eToken server. The proxy generated on the fly contains the extensions that specify the role and privileges of the robot certificate inside the VO supported by the Science Gateway, so different proxies can be created according to the different roles and privileges of the user in the LDAP registry. This ensures a fine grained authorization and provides the portal manager with the complete control of deciding what a given user can see and do.

4 Implementation of the e-Collaborative Digital Archives with gLibrary

As stated at the beginning of this document, during the first 6 months of the project, COMETA, after the creation and setup of the e-Culture Science Gateway, had the responsibility to implement two Grid-enabled digital repositories, as foreseen in the Description of Work:

- The Federico De Roberto literary works archive (De Roberto DR);
- The Architectural and Archaeological Heritage present in Mediterranean Area (MED Repo).

A third digital archive, a China Relics Digital Repository, has been implemented as the result of a Grid training organized in the context of the joint CHAIN/EPIKH School for Application Porting, held in Beijing on 16-27 May 2011, whose results have been presented at the Joint CHAIN/EPIKH Workshop[11].

The main software package behind the INDICATE e-Culture Science Gateway is gLibrary, a framework developed by COMETA and INFN Catania that allows the creation, organization, browsing and retrieving of digital assets on Grid-enables digital repositories, hiding the underlying technical details to the end users.

In the first phase, we defined with the help of the content's providers, which metadata had to be used to describe each repository, in which ways the users should access, browse and query the contents.

In the second phase, we uploaded all the digital objects to the storage resources of the COMETA Grid infrastructure. In order to achieve fault tolerance, most of the contents have been replicated in several storage elements, located in Catania, Messina, Palermo and Beijing, providing load balancing in case of high traffic, hiding network latencies and allowing users to choose a storage closer to them.

[11] http://agenda.ct.infn.it/conferenceOtherViews.py?view=standard&confId=473

In the third phase, we registered the metadata for each digital object that has been previously uploaded and set up the three repositories in the gLibrary backend services.

In the fourth and last phase, we had to deal with the creation of a simple front-end to access the repositories from within the e-Culture Science Gateway. The natural choice was to make use of portlets that interact with the gLibrary RESTful APIs. Two portlets (shown in figures 4-5) have been developed and deployed into the e-CSG to access, browse and retrieve the repositories' contents.

The browsing system has been designed to quickly retrieve the desired content among thousands of items using an intuitive filtering system on metadata, accessed on the header of each column. Once user has found the object she needs, a 3D geo-map (based on Google Earth) of the storage resources that have an available replica of the selected item is shown letting her choose the storage element to download the desired content from (see figures 6 and 7).

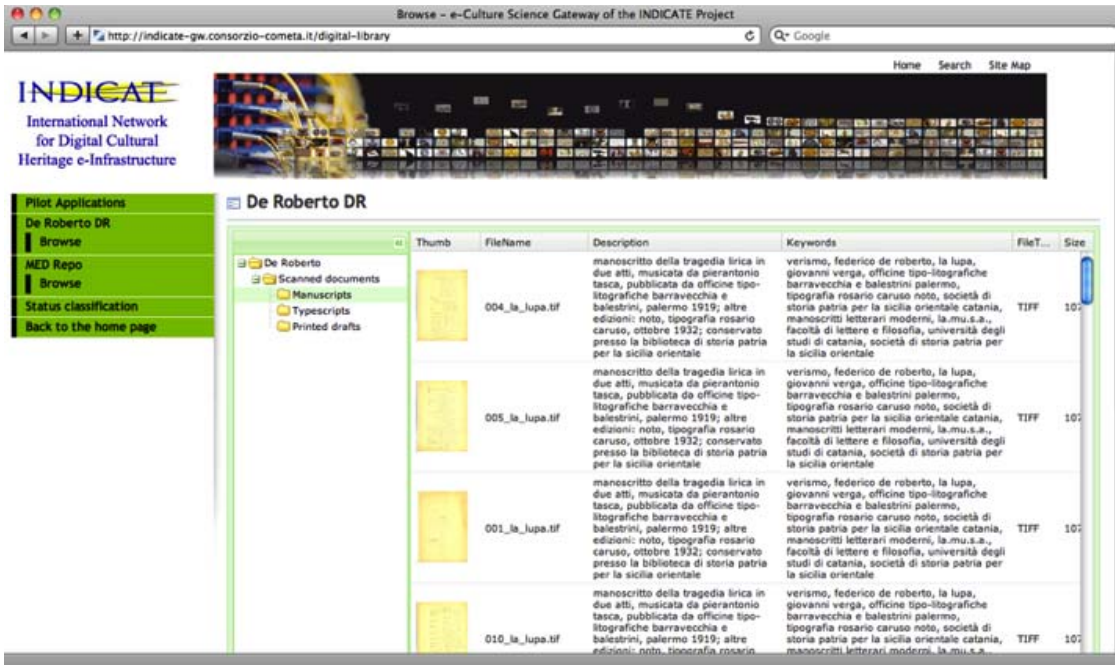

Figure 4: De Roberto DR browser portlet

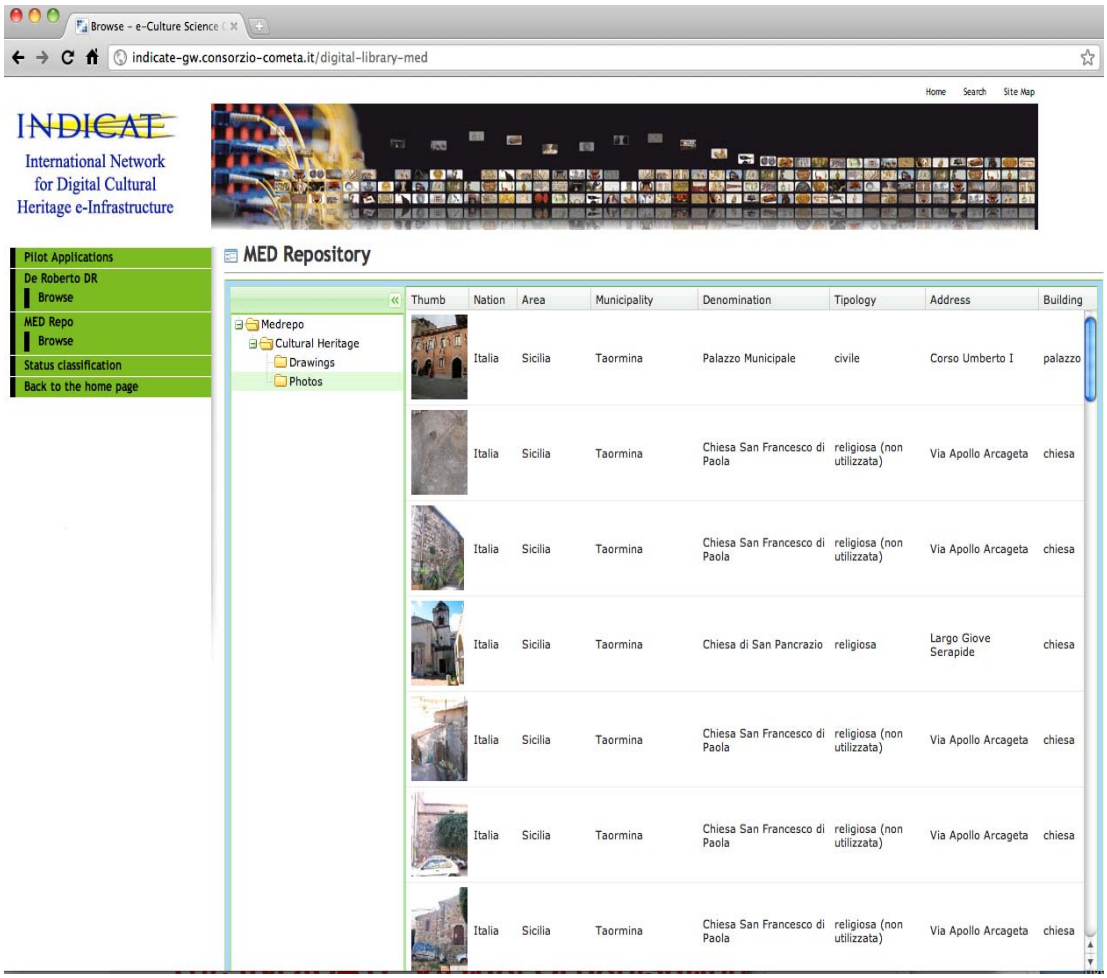

Figure 5: Mediterranean Repository (MED Repo) browser portlet

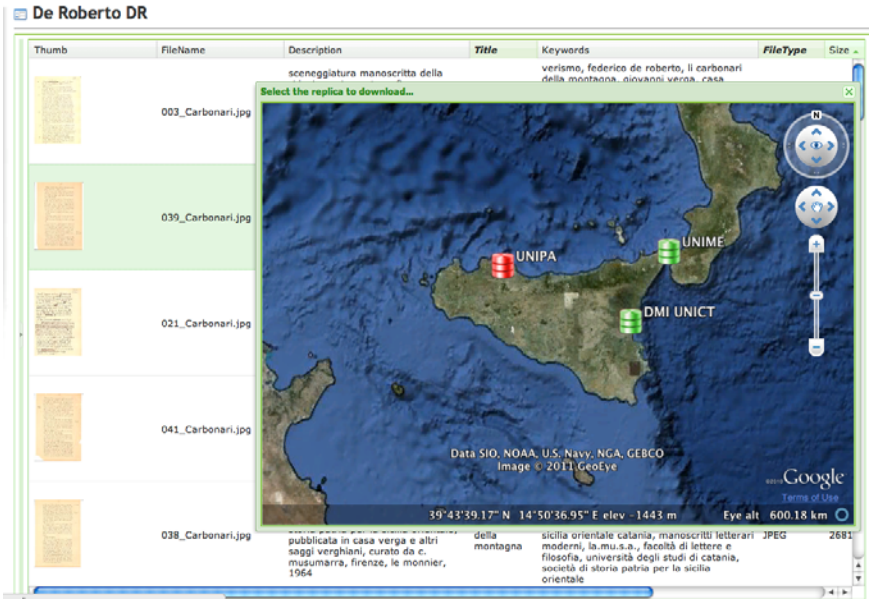

Figure 6: 3D geomap of storage elements based on Google Earth

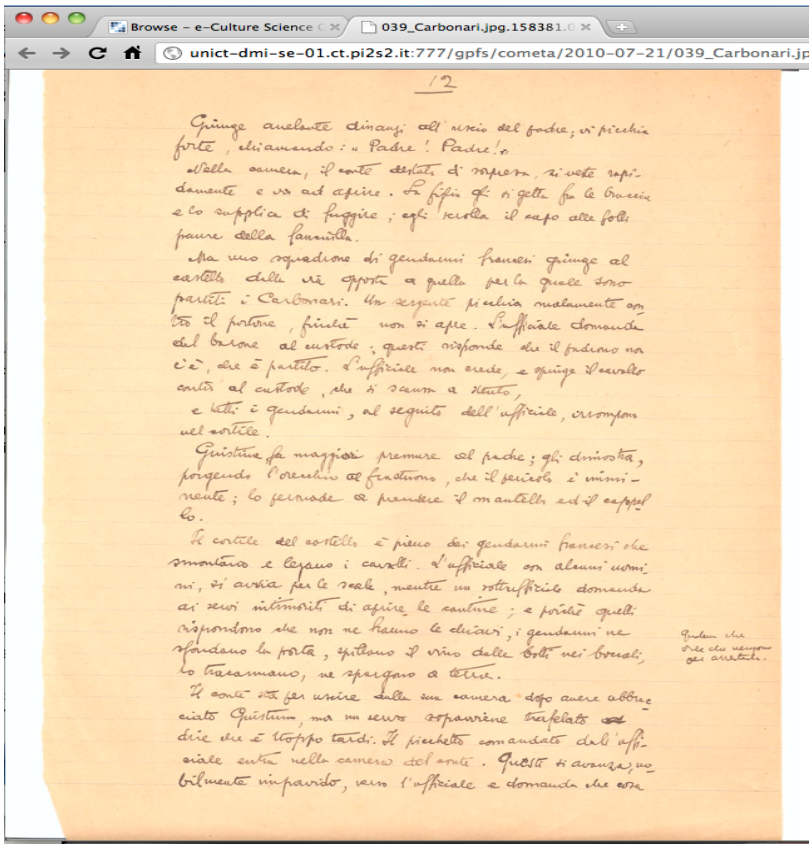

Figure 7: A sample of a digital object (page 39 of the manuscript "I Carbonari" from Federico De Roberto repository) downloaded from one of the e-Infrastructure storage to the user's browser

The following picture (figure 8) describes the architecture of the e-Cultural Science Gateway (e-CSG) integrated with gLibrary and the storage resources and metadata services.

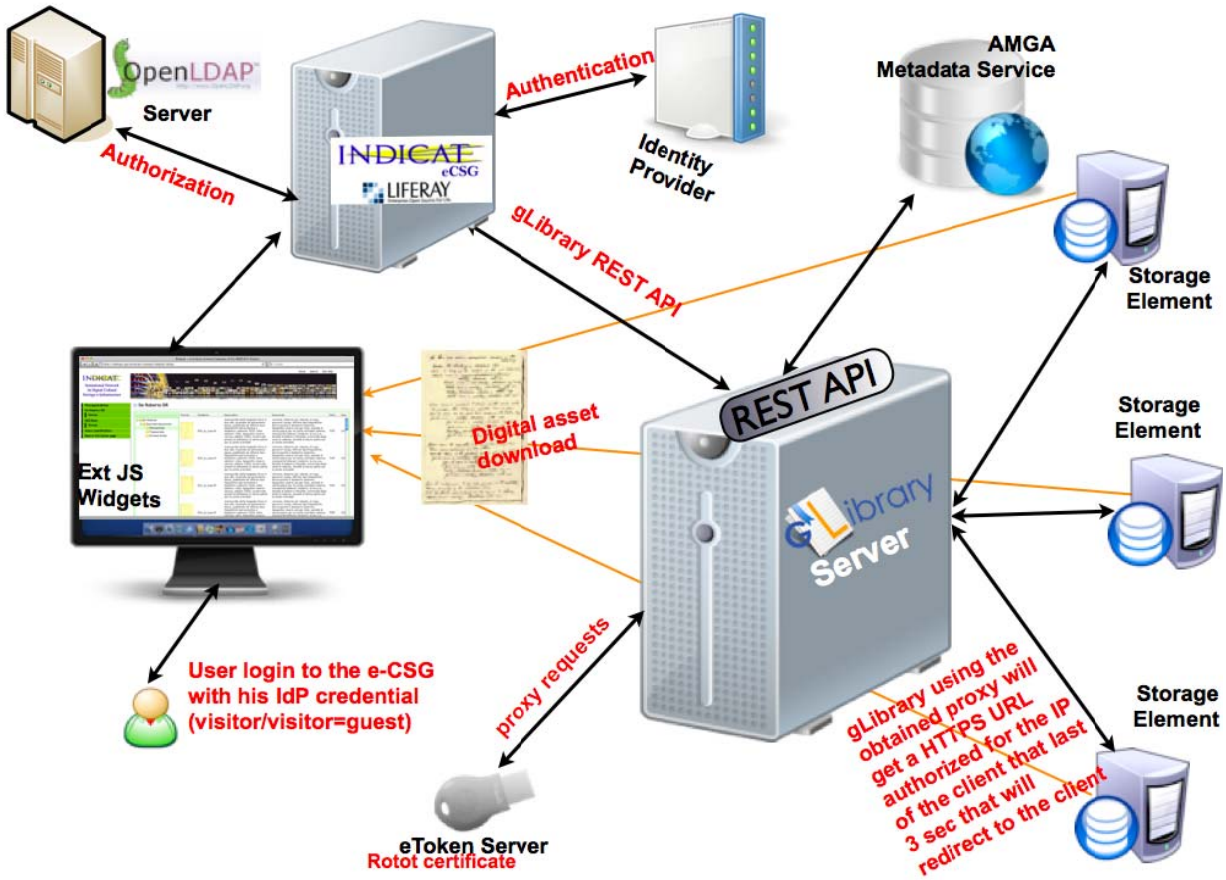

Figure 8: Architecture of the integration among the e-CSG, gLibrary repository services and the e-Infrastructure resources used

After a successful login to the e-CSG Science Gateway, the user can select one of the portlets (three portlets are currently deployed, one per each digital archive). The browsing portlet and its cascading filtering system builds and run "behind the scenes" complex queries in real time, according to the section of attributes and values, to the metadata service of the e-Infrastructures using the gLibrary services REST APIs, showing back the results through a nice AJAX and desktop-like GUI.

The authentication and authorization information, handled by the Identity Federation and OpenLDAP services is passed to and used by the gLibrary backend to enforce the access control on metadata and data stored once it needs to interact with the storage and metadata resources. At every interaction of the user with the portlet, a call to a given gLibrary RESTful API is augmented with authentication/authorisation information, so that the proper proxy certificate, generated by the eTokenServer (as described in the first section), can be created/reused to access grid services that require X.509 credentials. At the same time, every operation is logged in the user tracking DB, as described above. Once user chooses a link to a digital object she is interested in, the gLibrary backend asks the proper storage to generate a one-time short-lived (3 sec) URL, authorized for the IP of the requesting client and forwards it back to the user's browser that follows it and immediately download the digital file, directly from the storage resource to his/her desktop, without any streaming into gLibrary server or the e-CSG server. This mechanism allows offloading the e-CSG from the heavy duty of serving numerous and big downloads, that could easily collapse the hosting server, so that the gateway will be free to handle just the browsing and search functionalities, forwarding the charge of handling downloads and potentially huge data transfers directly to the storage resources of the e-Infrastructure.

5 Conclusions

E-Infrastructures can be very beneficial platforms for the Digital Cultural Heritage (DCH) community, provided they are «easy to use».

The INDICATE e-Culture Science Gateway is a major step forward towards the uptake of Grid technology by the DCH community. The adopted Science Gateway model, supporting Identity Federations and Social Networks, can revolutionize the way Grid infrastructures are used, hugely widening their potential user base, especially non-IT experts and the “citizen scientist”. The adoption of standards, in particular, represents a concrete investment towards sustainability.

By design, the components (the “portlets” – our “Lego bricks”) of the e-Culture Science Gateway have maximum re-usability and, indeed, they have already been adopted by other projects (e.g., agINFRA, CHAIN, DECIDE, EarthServer, EUMEDGRID-Support, and GISELA).