Cultural Heritage Digital Libraries on Data Grids

Antonio Calanducci[1], Jorge Sevilla[1], Roberto Barbera[1], Giuseppe Andronico[1], Monica Saso[2], Alessandro De Filippo[2], Stefania Iannizzotto[2], Domenico Vicinanza[3], and Francesco De Mattia[4]

[1] INFN Catania
Via S.Sofia 64, 95128 Catania, Italy
{antonio.calanducci,jorge.sevilla,roberto.barbera, giuseppe.andronico}@ct.infn.it

[2] Università di Catania, Facoltà di Lettere e Filosofia
Piazza Dante 32, Catania, Italy
lamusa@unict.it

[3] Conservatorio di Parma, ASTRA Project
Via del Conservatorio 27/A, 43100 Parma
francesco.demattia@conservatorio.pr.it

[4] Delivery Advanced Network Technology to Europe UK, ASTRA Project
City House 126-130 Hills Road, Cambridge CB2 1PQ
domenico.vicinanza@dante.co.uk

Abstract. Data Grids offer redundant and huge distributed storage capabilities, providing an ideal and secure place for the long-term preservation of digitized literary works and documents of artistic and historical relevance. In this demo, we are going to show how we deployed some digital repositories of ancient manuscripts making use of gLibrary, a grid-based system to host and manage digital libraries

Keywords: Digital Libraries, Grid Computing, Data Grid, Cultural Heritage, Digital Preservation.

1 Introduction

Data Grids offer redundant and huge distributed storage capabilities, providing an ideal and secure place for the long-term preservation of digitized literary works and documents of artistic and historical relevance.

In fact, digitization has been progressively used as a means for avoiding the loss of literary heritage on paper, caused by physical ageing and the environmental conditions in which documents are kept. Document consultation is another problem that leads to additional deterioration. Multiple copies of high resolutions scans stored in a distributed environment and made available for consultation with an easy to use interface is a means to guarantee conservation of cultural heritage. Grid authentication and authorization mechanisms allow a fine-grained access to archives by single users, groups or entire communities. Moreover, metadata services permit a structured organization of scanned files for quick searches.

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 469–472, 2009.

Two use cases have been considered to demonstrate how grid digital libraries can guarantee enduring preservation of literary heritage: the archives of the work of Italian writer Federico De Roberto, made up of almost 8000 scans, and the musical and the musical archives of the "Civiltà Musicale Napoletana" project, made up of more than 250,000 digitizations.

A working prototype of the De Roberto digital repository has been implemented on the gLibrary platform, a grid-based system to host and manage digital libraries developed by INFN Catania, on the Sicilian e-infrastructure of the COMETA consortium.

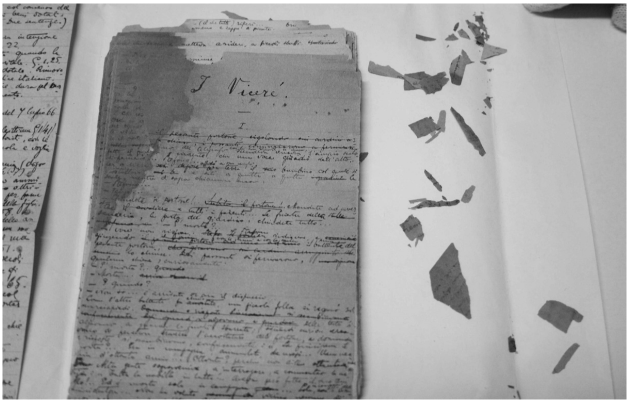


Fig. 1. First page of notes for the De Roberto "I Viceré" work

2 The gLibrary Platform

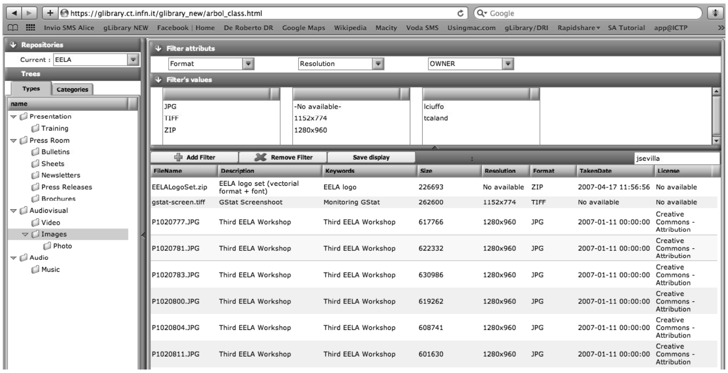
gLibrary challenge is to offer an extensible, robust, secure and easy-to-use system to handle digital assets widespread stored on a distributed Grid infrastructure. This goal can be achieved exploiting a series of Grid services together with a proper business logic, and providing an intuitive front-end, accessible from everywhere and anytime. Users do not have to care about the complexity of the underlying systems and the geographical location of their data and they can consider the available Grid storage as a huge virtual disk.

gLibrary can be used to store, organize, search and retrieve any kind of digital assets represented as files in a Grid environment. Consequently, it can be useful for different users that need a secure way to save and share their assets. Assets are saved on the grid storage servers and can be encrypted and replicated on several storage servers, assuring maximum security and high availability to the users' data.

All entries in a gLibrary repository are organized according to their type: a list of specific attributes to describe each kind of asset to be handled by the system. These are the same attributes that can be queried by users.

Each type can have multiple subtypes with additional attributes and all types share a common attribute list (root type), that is fixed by design (in the next release it will be the Dublin Core set). Before users can start uploading assets, a hierarchy of types has to be defined by the repository administrator.

A filtering system, similar to the ones used by the Apple iTunes application to organize iPod/iPhone multimedia collections, is available to browse each deployed repository: some of the attributes of each types can be selected as filters, and their cascading application narrow the result set dynamically, allowing the user to find the interested asset with few mouse clicks.



Current : EELA

Filter attributs: Format, Resolution, OWNER

Filter's values:

Format	Resolution	OWNER
JPG	-No available-	lciuffo
TIFF	1152x774	tcaland
ZIP	1280x960	

jsevilla

FileName	Description	Keywords	Size	Resolution	Format	TakenDate	License
EELALogoSet.zip	EELA logo set (vectorial format + font)	EELA logo	226693	No available	ZIP	2007-04-17 11:56:56	No available
gstat-screen.tiff	GStat Screenshoot	Monitoring GStat	262600	1152x774	TIFF	No available	No available
P1020777.JPG	Third EELA Workshop	Third EELA Workshop	617766	1280x960	JPG	2007-01-11 00:00:00	Creative Commons - Attribution
P1020781.JPG	Third EELA Workshop	Third EELA Workshop	622332	1280x960	JPG	2007-01-11 00:00:00	Creative Commons - Attribution
P1020783.JPG	Third EELA Workshop	Third EELA Workshop	630986	1280x960	JPG	2007-01-11 00:00:00	Creative Commons - Attribution
P1020800.JPG	Third EELA Workshop	Third EELA Workshop	619262	1280x960	JPG	2007-01-11 00:00:00	Creative Commons - Attribution
P1020804.JPG	Third EELA Workshop	Third EELA Workshop	608741	1280x960	JPG	2007-01-11 00:00:00	Creative Commons - Attribution
P1020811.JPG	Third EELA Workshop	Third EELA Workshop	601630	1280x960	JPG	2007-01-11 00:00:00	Creative Commons - Attribution

Fig. 2. gLibrary browsing front-end: type tree and filtering

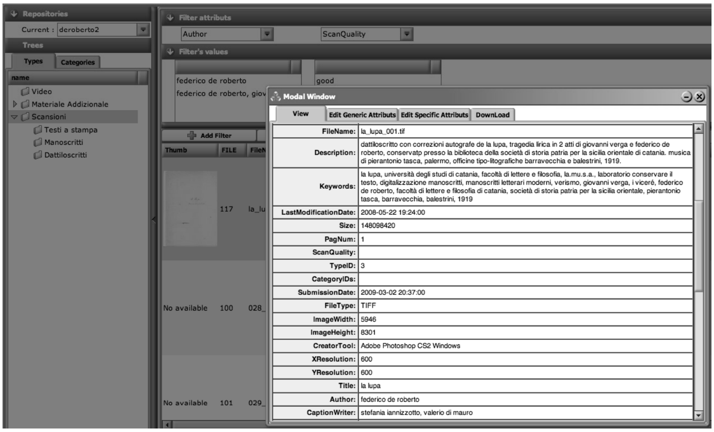
gLibrary servers can host multiple libraries, which can have their own hierarchy of types and can be accessed by different users. Permissions are set up using a fine-grained authorization mechanism. Each asset and/or type has a set of ACLs (Access Control Lists) based on X.509 certificate that restricts its usage, allowing asset owners to grant access to a whole organization, selected groups of users or just a single user. Those entries and types, on which users do not have permissions, are also not visible from the browsing interface.

Both uploads and downloads to/from the grid are carried out over HTTPS or GSIFTP using respectively the user X.509 certificate on the browser or a local grid proxy.

3 Demonstration Content

Two repositories has been actually deployed with gLibrary on the Grid: the archives of the work of Italian writer Federico De Roberto (1861-1927), made up of almost 8000 scans, and the musical archives of the “Civiltà Musicale Napoletana” project, made up of more than 250,000 documents.

During the demonstration, the gLibrary front-end will be used to browse through the repositories and to look for items with some given properties, exploiting the filtering system of gLibrary. For example, a scholar may need to look for all the De Roberto drafts, printed in the 1919. He will first select the *De Roberto* repository, then select *Scansioni* (scans) from the type tree, then *testi a stampa* (printed drafts). Choosing the *PublicationYear* as filter, and selecting *1919* as value among the available years, he will get back a result set of all the assets satisfying his request. The search can be further refined choosing *Publisher* as second filter, to group drafts by publisher.



FileName:	la_lupa_001.tif
Description:	dattiloscritto con correzioni autografe de la lupa, tragedia lirica in 2 atti di giovanni verga e federico de roberto, conservatp presso la biblioteca della società di storia patria per la sicilia orientale di catania. musica di pierantonio tasca, palermo, officine tipo-litografiche barravecchia e balestrini, 1919.
Keywords:	la lupa, università degli studi di catania, facoltà di lettere e filosofia, la.mu.s.a., laboratorio conservare il testo, digitalizzazione manoscritti, manoscritti letterari moderni, verismo, giovanni verga, i viceré, federico de roberto, facoltà di lettere e filosofia di catania, società di storia patria per la sicilia orientale, pierantonio tasca, barravecchia, balestrini, 1919
LastModificationDate:	2008-05-22 19:24:00
Size:	148098420
PagNum:	1
ScanQuality:	
TypeID:	3
CategoryIDs:	
SubmissionDate:	2009-03-02 20:37:00
FileType:	TIFF
ImageWidth:	5946
ImageHeight:	8301
CreatorTool:	Adobe Photoshop CS2 Windows
XResolution:	600
YResolution:	600
Title:	la lupa
Author:	federico de roberto
CaptionWriter:	stefania iannizzotto, valerio di mauro

Fig. 3. Metadata full set of a chosen manuscript

Once the sought asset has been found, the user is able to inspect the complete metadata set and finally is able to choose one of the replica links for downloading the file from the proper Storage Element to his desktop/laptop.