

#### Consorzio COMETA



# The COMETA Grid Infrastructure and its Services for Digital Cultural Heritage Applications

Roberto Barbera (<u>roberto.barbera@ct.infn.it</u>)
COMETA Chief Technical Officer
DC-NET Concertation Seminar
Tallinn, 13-14.01.2011





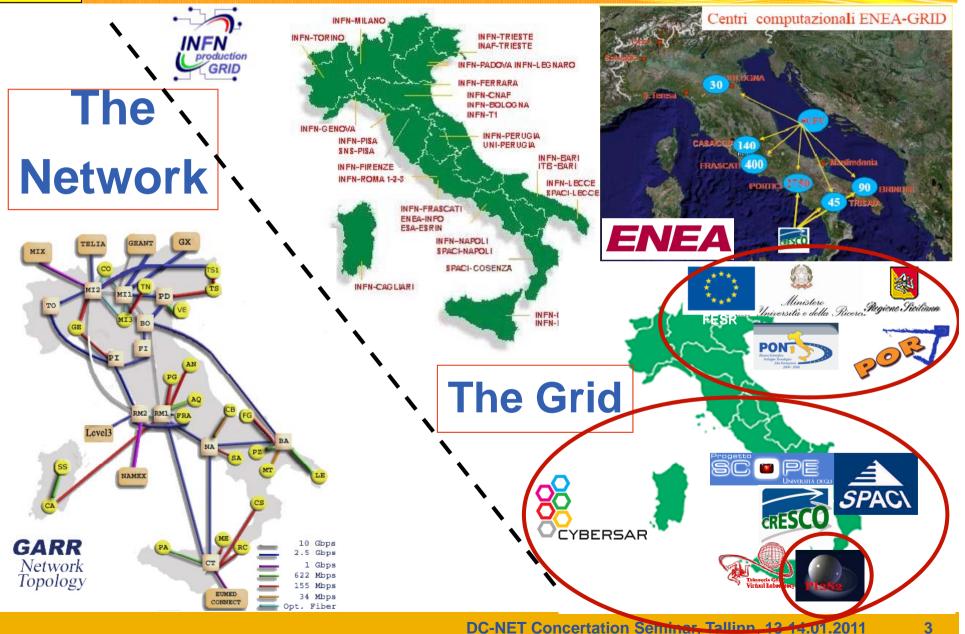


- The Italian e-Infrastructure
- The COMETA Grid and Applications
- The COMETA Services for Cultural Heritage
  - gLibrary
- COMETA, DC-NET and INDICATE
- The INDICATE Virtual Organisation
- The INDICATE e-Culture Science Gateway
- Summary and Conclusions



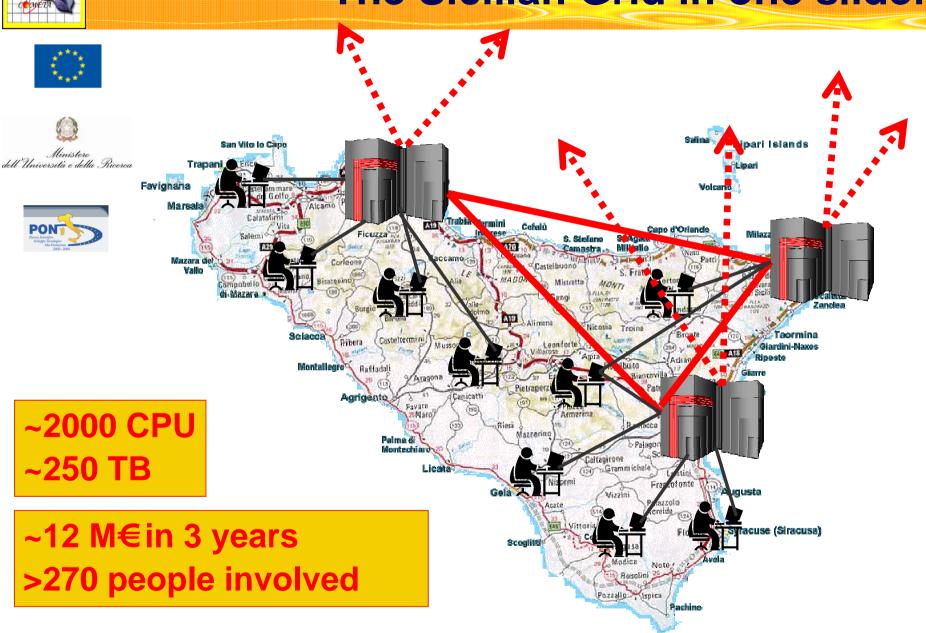
#### The Italian e-Infrastructure

(interoperable with gLite as common communication protocol)





#### The Sicilian Grid in one slide...





#### **The COMETA Consortium**

(www.consorzio-cometa.it)

















## The Sicilian e-Infrastructure (1/2)





## The Sicilian e-Infrastructure (2/2)









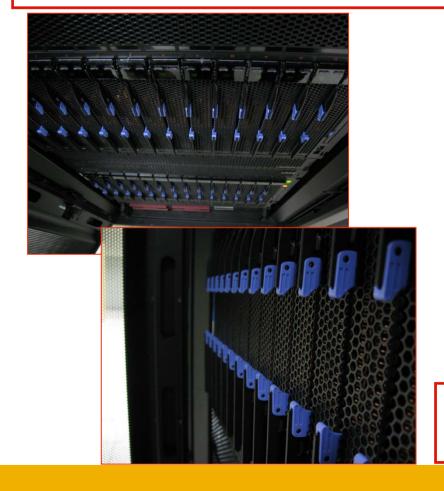


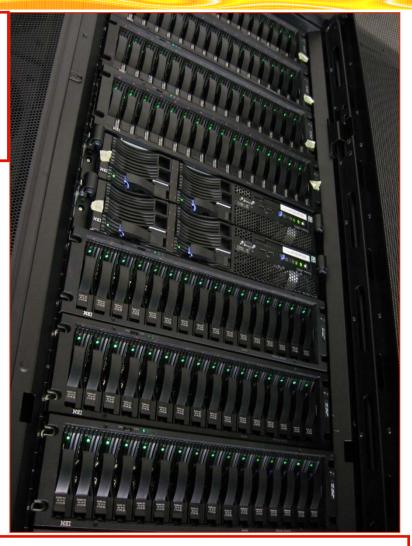




#### Computing and Storage resources available

- 1. <u>~2000 cores</u> AMD Opteron 2218 rev. F
- 2. 2 GB of RAM per core
- 3. Commercial LRMS (LSF)
- 4. Infiniband-4X (for MPI applications)





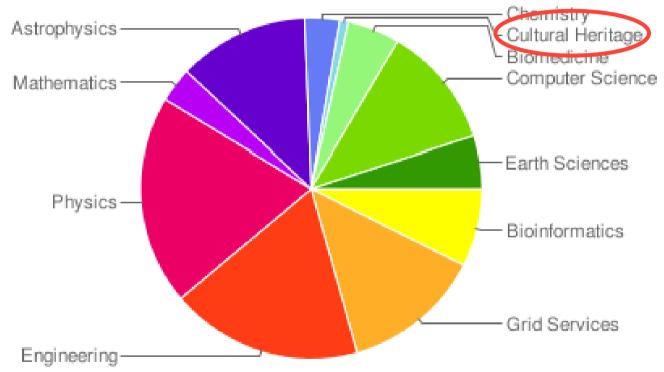
- 1. ~250+ TB of storage
- 2. Distributed parallel filesystem (GPFS)



## **COMETA Applications**

- A comprehensive database is available at:
  - www.pi2s2.it/applications (122 entries)

Distribution by domain



- To know more, look at:
  - agenda.ct.infn.it/materialDisplay.py?materialId=12&confld=24
  - agenda.ct.infn.it/materialDisplay.py?materialId=20&confld=87



## Why Grids for Digital Cultural Heritage?

- High performance storage systems:
  - Geographically distributed replicas of files
    - Fault-tolerant digital preservation;
- Simplified authentication systems:
  - Single sign-on;
- Fine-grained authorization systems:
  - Allow to define exactly who:
    - Individual user(s), group(s), organization(s), the whole world;
  - And what:
    - Read/edit/delete data and metadata;
    - Search, browse;
    - Creation of new repositories;
    - Role assignments;
    - Etc.



ibrary

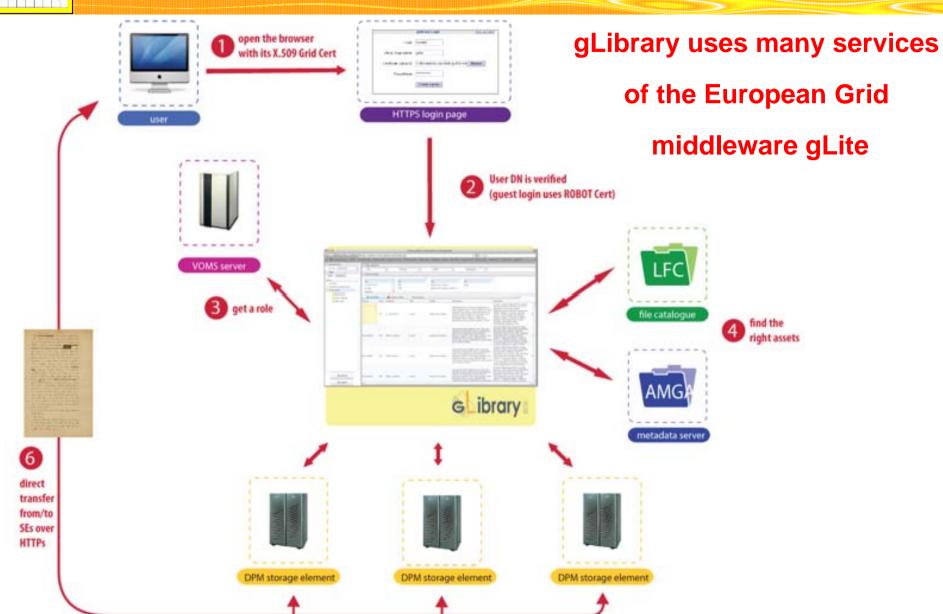


- gLibrary is the INFN/COMETA platform that provides a simple yet powerful system to store, organize, search and retrieve "digital assets" in repositories built on e-Infrastructures
  - hides the underlying technical details to the users
- What we mean by "digital asset":
  - data + metadata





## gLibrary architecture





## **Asset organization**

- Digital Asset of different kind (or type) are organized into repositories:
  - "types" e "collections" definition by the administrator(s) of a given

repository. Assets organized by:

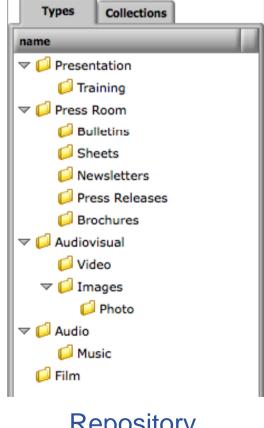
type:

a list of attributes to describe a given kind of asset

- hierarchical (a child type shares and extend parent's attributes)
- filters: subset of attributes queried during browsing

EXAMPLE OF TYPES AND ATTRIBUTES' LIST

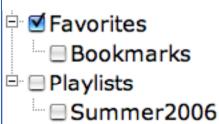
Туре	Attricutes' list
Audic	Formut, Bitrate, Samplerate, Time
Music	(Format, Bitrate, Samplerate, Time), Name. Artist, Album, Genre, Tracknumber, Year, Artwork, Lyric, Rating
Presentation	Format, NumOf?ages
Training	(Fermat, NumOPages), Title, Runtime, Speaker, Author, Subject, Event, Date, Type
(Root,	FileSome, Sui missionDate, L'escription, Keywords, LastModificationDate, Size



Repository

- and/or collection:
- group together related assets belonging to different types
- useful to define subsets of assets of the same type (i.e. favorites, playlists)
- the same asset can belong to multiple collection (sort-of tagging)

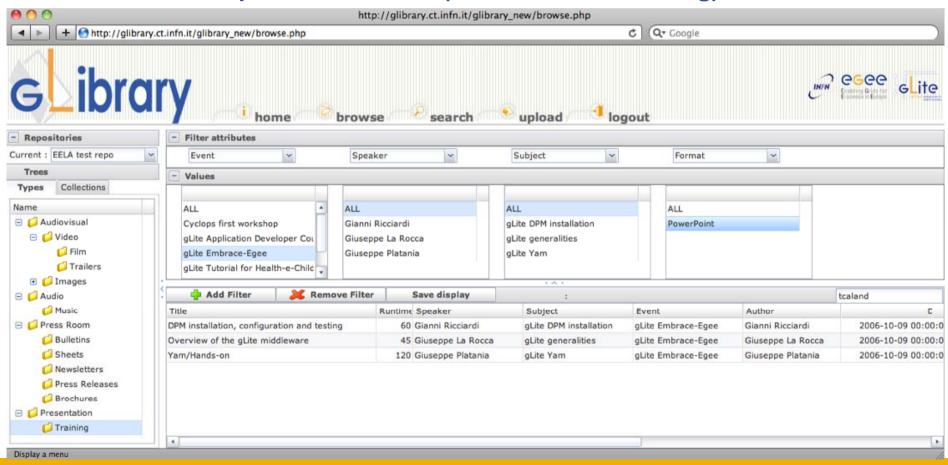
#### Collections





# **qLibrary** browser

- "à la iTunes"
- Assets are browsed selecting a type (or a collection) and then one of its filter:
  - attributes of the selected types/collection, chosen from a defined list, used to narrow the result set
- Filter application is cascading: the selection of a filter value dynamically influences subsequent filter values ("à la iTunes" browsing)





#### Asset upload and retrieval

- Users can upload their local assets to one or more (i.e., creating replicas) Storage Elements of the Grid;
- Download from Grid storage to user PC/laptop with a click:
  - choosing a replica from a list:



- future functionalities: storage maps, closer storage;
- Transfers (handled over HTTPS) are authorized by the digital certificate imported into the user's browser:
  - only authorized users can download a given assets;
  - guest access is supported through robot certificates.







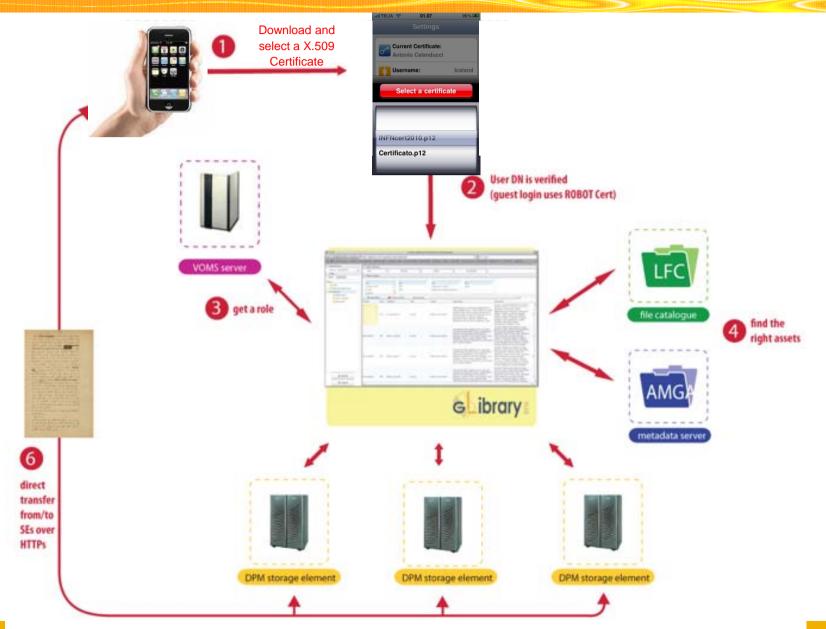


- Mobile interface to access Grid Digital Libraries:
  - iPhone, iPod Touch, iPad;
- Advantages:
  - provides an extremely intuitive and touch-based user interface to Grid storage elements and metadata, especially for non-expert users;
  - (to do) automatic selection of the closest replica, according to the user physical location retrieved by the integrated GPS;
  - offline access to the assets already downloaded.

In collaboration with Dr. C. Pistagna

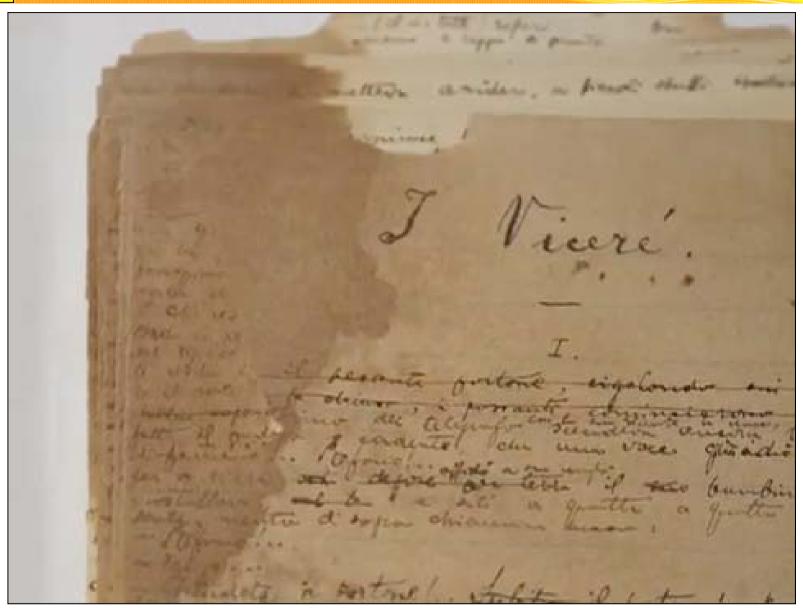


## gLibrary mobile architecture





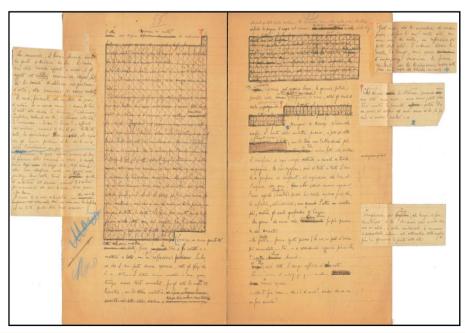
#### A use case: Federico De Roberto works





# Digitise to preserve

- Some sheets are damaged (mold, crumbed pieces) and need physical restoration;
- Digitisation is mandatory to avoid the loss of this heritage; some works are still unpublished and relevant for the humanistic community.







## Digitisation phase @ University of Catania





# **Acquisition stage**

- Digitalization of manuscripts, typescripts, printed works:
  - TIFF Files, one per page, 600 dpi, about 100 MB for A3
    - High resolution scans for in-depth examination;
  - PDF, one per work, 300 dpi, varying file sizes 40-400MB
    - Overall examination of works;
  - 8000 sheets/scans, 3 Terabyte of disk space;
  - Different physical formats, A3/A4/custom size;
- Embedded Metadata:
  - TIFF with embedded metadata to provide scan physical features and information about the content:
    - ImageWidth, ImageHeight, XResolution, FileSize, CreationDate, ModifyDate
    - Description, Keywords, CaptionWriter, Title, Author, Copyright Status, Copyright Notice;
  - Added with Photoshop after the digitisation phase (Adobe XMP format).



## Goals and requirements

- Make manuscripts accessible to the humanistic community
  - Always on-line: 24 x 365
  - Available from everywhere
  - Simple and easy-to-use interface for non-expert people
- Quickly find the desired document
  - Document organization according the physical and semantic metadata
    - Organization by type/collections
    - Dynamic filtering of search result sets according the selection of one or more document metadata
- Long-term preservation (digital preservation)
  - Multiple copies (replicas) spread over different geographical sites
  - Reliability of storage systems and replica redundancy to achieve real preservation



#### De Roberto DR on Grid with gLibrary

#### Goals:

- store the 8000 scans of De Roberto Heritage
   Data Grid Storage Elements
- enable an ubiquitous and 24/24h access to scientists ---> Web Application
- document organization for a quick search
   Metadata Services
- long-term digital preservation of data
   redundancy through Replicas of files on several Storage Elements
- simple and easy-to-use system for searches, organization, upload and download of digitalized documents on the Grid ---->



## Technologies used

#### Web 2.0 standards:

- Javascript/AJAX/JSON on the client side;
- PHP5 classes to implement business logic on the server side;

#### Grid technologies:

- Storage Element SRM interface to get the TURLs (Transfer URLs);
- Transfers handled with GridFTP and X.509 cert auth HTTPS;
- X.509 based Globus Security Infrastructure with the VOMS extensions to handle authentication and authorization (ACL based) on Metadata and Storage Elements;
- All grid services implemented with the EGEE gLite middleware (DPM Storage Elements, AMGA Metadata Catalogue, LFC File Catalogue, VOMS Services);

#### Other standards:

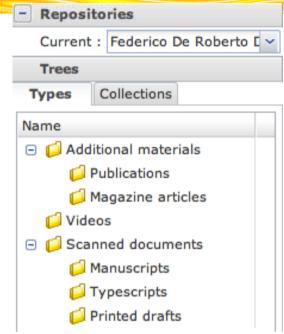
- Subset of XMP Metadata:
  - Dublin Core Metadata set will be included in the next release:



#### Metadata used in the DR digital library

- Type tree definition for the DR repository
- Attributes definition per type (e.g., manuscripts):

Attribute	Value
Title	la lupa
Author	federico de roberto, giovanni verga
Description	manoscritto della tragedia lirica
Keywords	verismo, federico de roberto, la lupa,
CaptionWriter	stefania iannizzotto, alessandro
CopyrightStatus	copyrighted
PageNum	5
TotalPages	34
DocumentGenre	tragedia lirica
PublicationYear	1916
Publsher	officine tipo-litografiche barravecchia e balestrini
FileType	PDF
Resolution	300
ScanQuality	good



#### Filter defined per type. Es:

-DocumentGenre

-Title

-FileType

-ScanQuality

-DocumentType

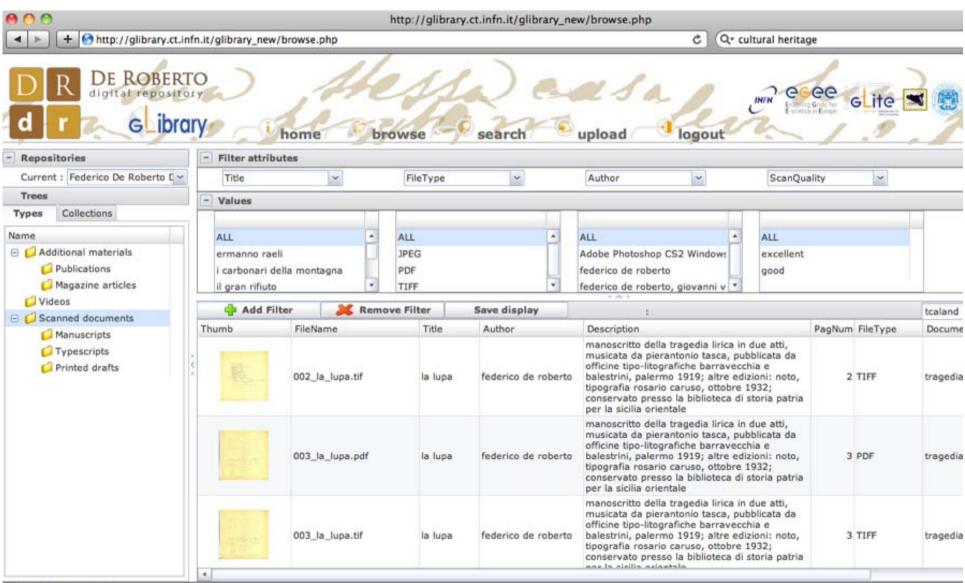
-PublicationYear

-Publisher

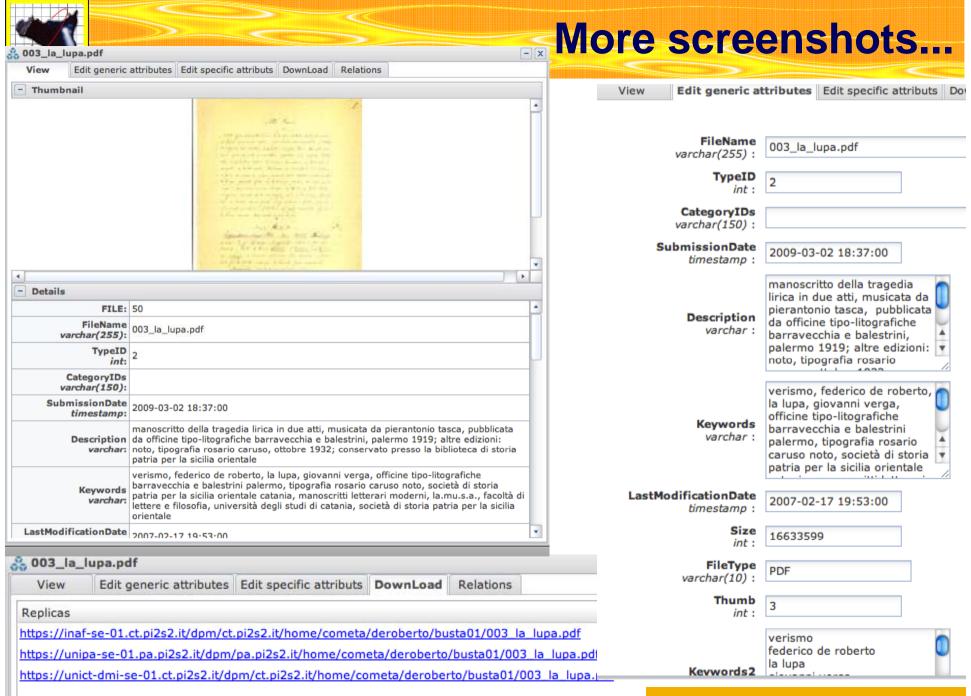
-Location



#### Some screenshoots...

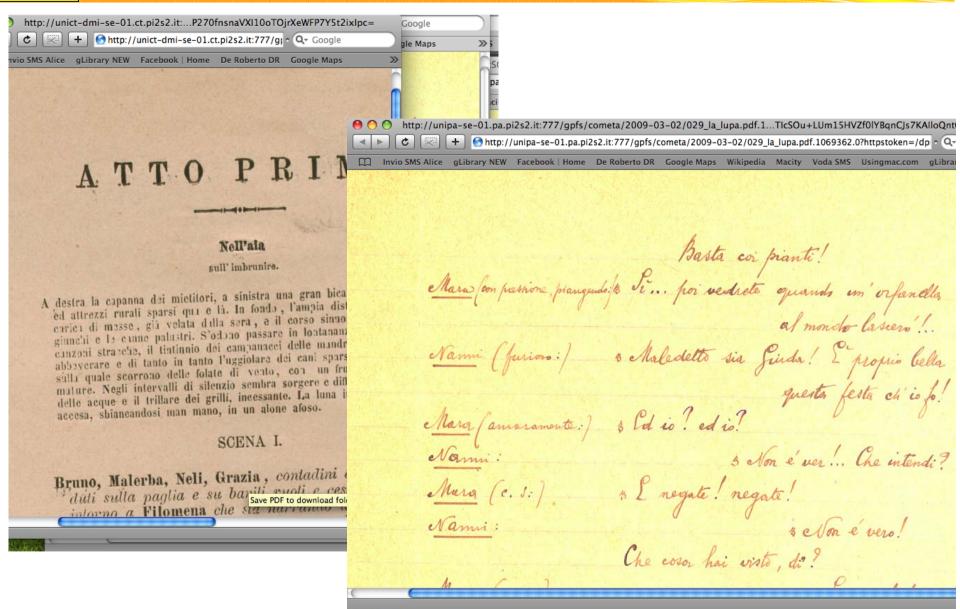


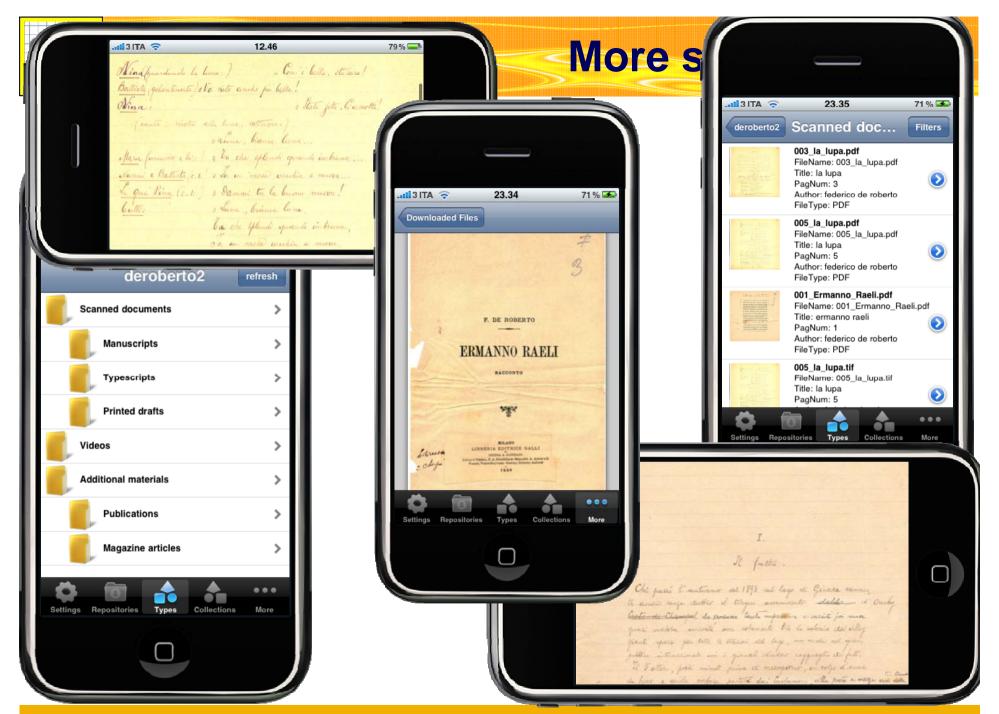
Cancalad opening the gage





#### More screenshots...







#### **COMETA, DC-NET and INDICATE**

#### COMETA and DC-NET:

COMETA has signed a Cooperation Agreement with DC-NET;

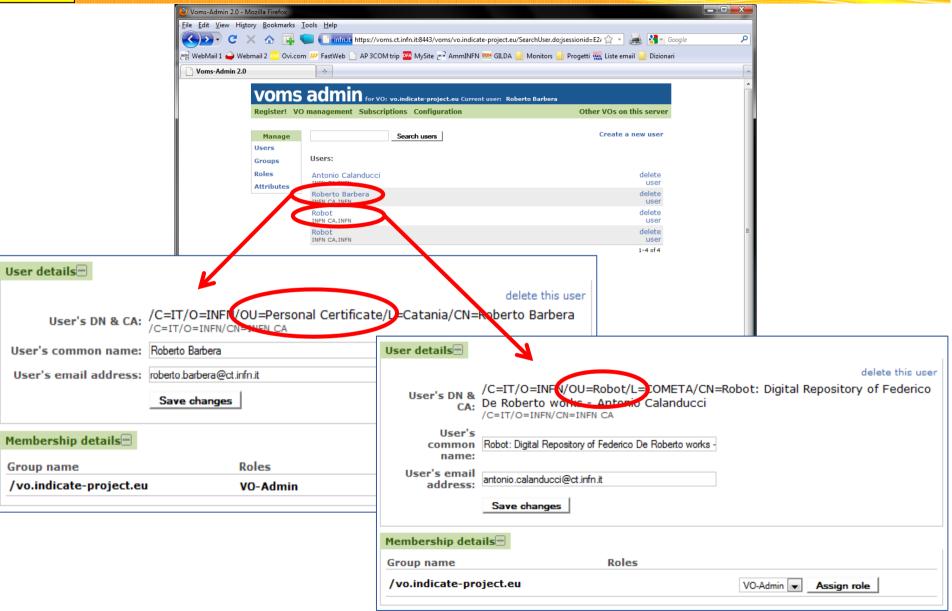


#### COMETA and INDICATE:

- Develop an e-Culture Science Gateway (eCSG) with general user authentication and authorisation schemas;
- Make two Digital Cultural Heritage pilot use cases available through the eCSG:
  - Digital Repository of Federico De Roberto works;
  - Digital Repository of the Architectural and Archaeological Heritage in the Mediterranean Area.



## The INDICATE Virtual Organisation



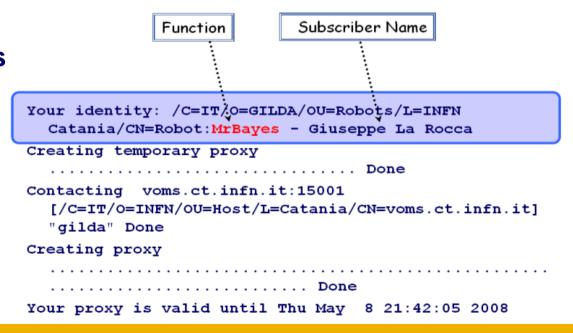


#### Robot certificates "in a nutshell"

Robot certificates have been introduced to allow users, who do not have/want personal certificates and do not belong to any Virtual Organisation, to use the Grid; they are usually issued on <a href="mailto:smart-cards">smart-cards</a> to be plugged on the machine where the service(s) is(are) executed.

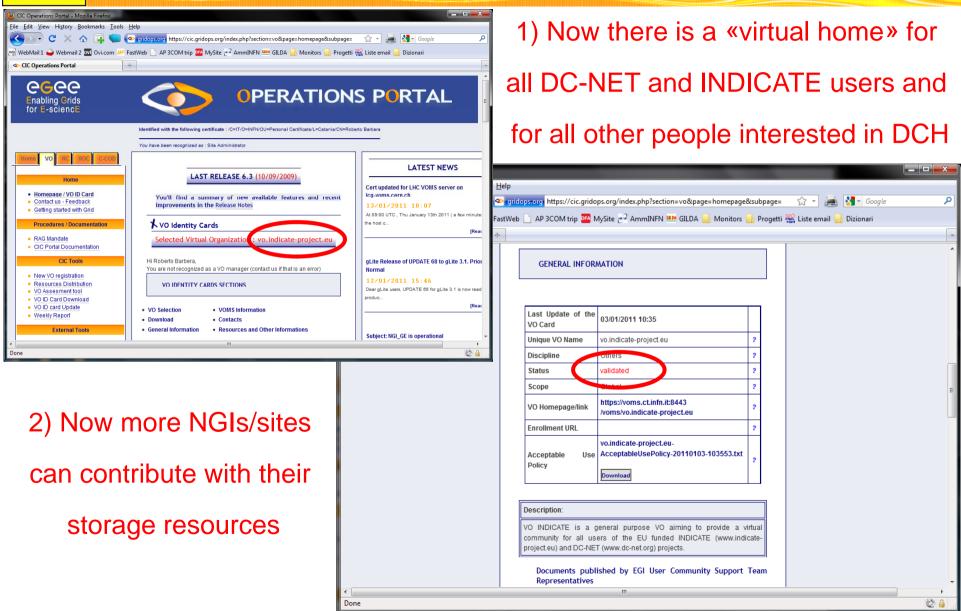


Basically, these certificates can be used to identify a person responsible for an unattended service or process acting as client and/or server for a Virtual Research Community.



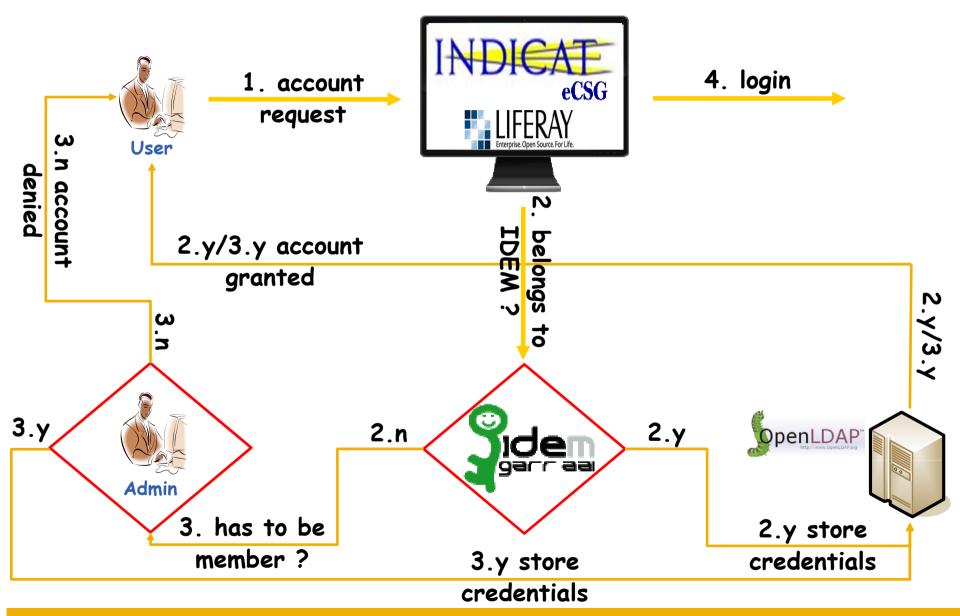


## The INDICATE VO registered on the CIC



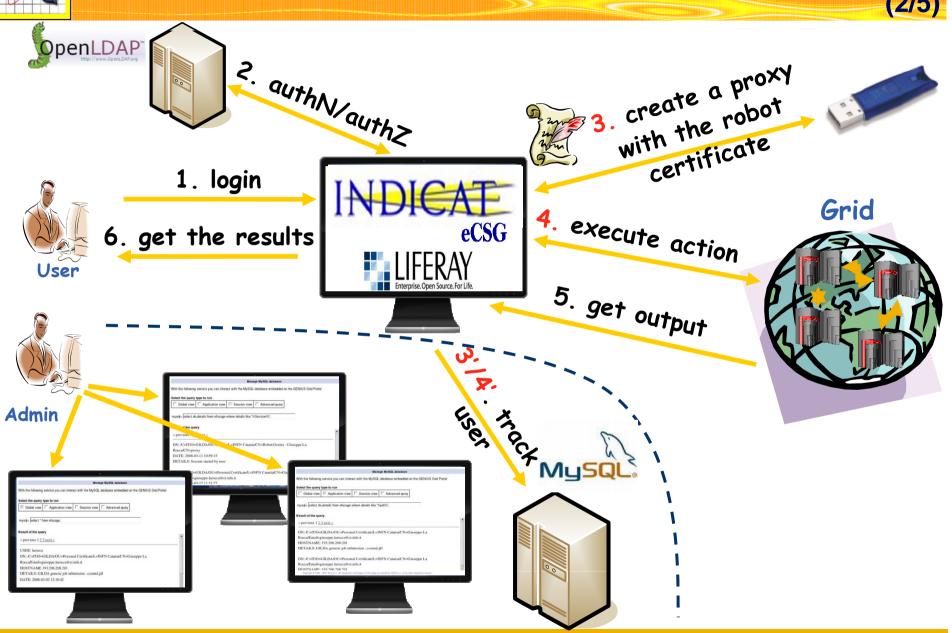


(1/5)



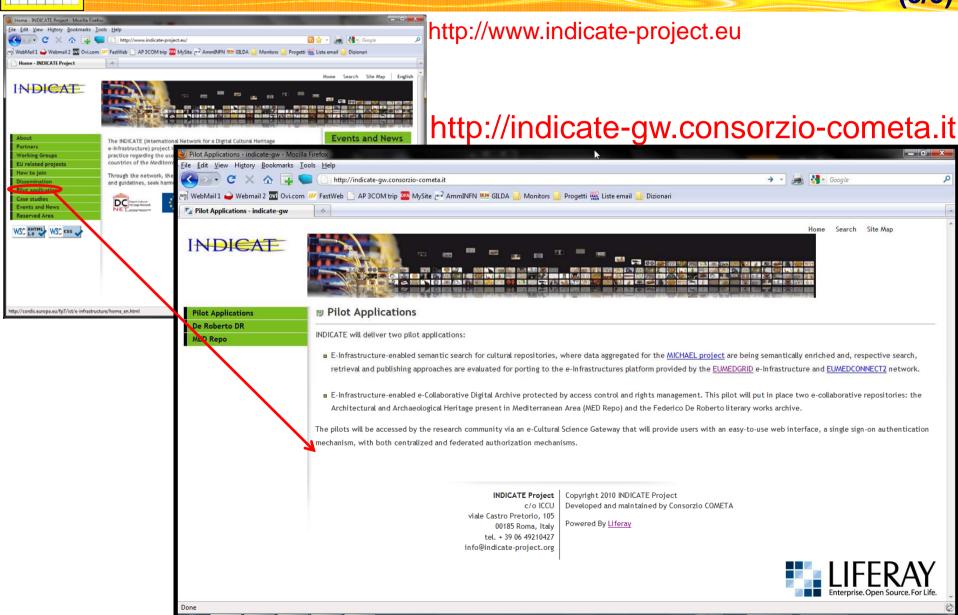


(2/5)



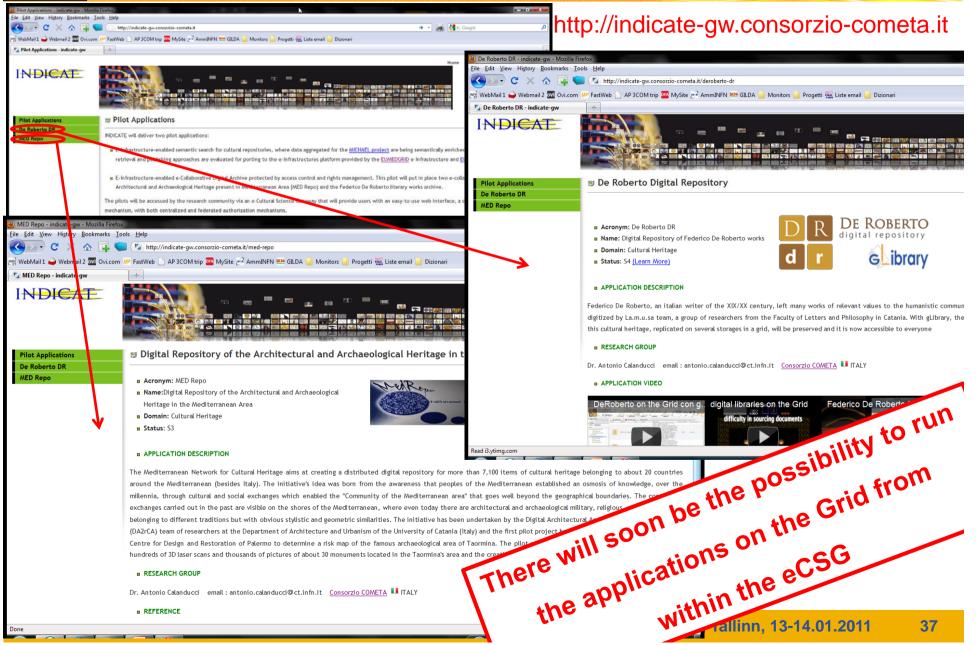


(3/5)



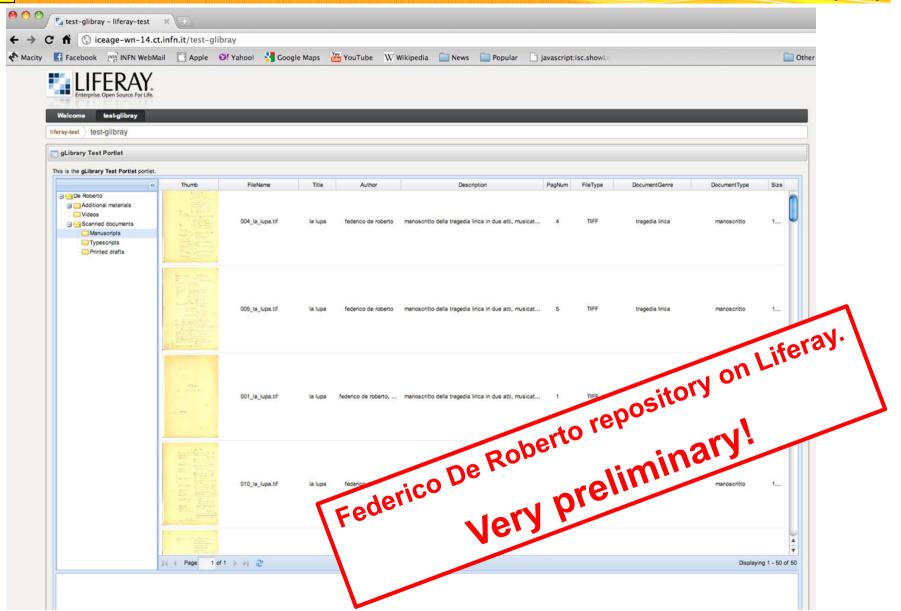


(4/5)





5/5)





## **Summary and Conclusions**

- COMETA is one of the Grid resource providers of the Italian e-Infrastructure and, through it, of the European Grid Infrastructure;
- COMETA has a long experience in developing services for Virtual Research Communities, also in the field of Digital Cultural Heritage;
- gLibrary is entirely based on gLite, one the most adopted middleware in the world;
- In the context of DC-NET and, mainly, INDICATE projects, COMETA will concretely show, through exemplar use cases, how Grid infrastructures can effectively be used by the Digital Cultural Heritage community in a simple and intuitive way;
- I'm at your disposal to discuss all possible ways of collaboration.



## Thank you for your kind attention!

# Any questions?



#### References

#### gLibrary contacts:

antonio.calanducci@ct.infn.it, glibrary@ct.infn.it

#### Federico De Roberto repository:

https://glibrary.ct.infn.it/

#### Videos:

- http://www.youtube.com/watch?v=VNN3OnpmUUU
- http://www.youtube.com/watch?v=IhFFjHD8IsI

#### • Publications:

- A.Calanducci, R.Barbera, J.Sevilla, A. De Filippo, M.Saso, S. Iannizzotto, F. De Mattia, F.Vicinanza.
   "Data Grids for Conservation of Cultural Inheritance", 1st International Workshop on Data Grids for e-Science (DaGreS09) at ACM International Conference on Computing Frontiers, May 18-20, 2009 (<a href="http://www.computingfrontiers.org/2009/">http://www.computingfrontiers.org/2009/</a>)
  - https://glibrary.ct.infn.it/m/DaGRes-editor.pdf
- A. Calanducci, C. Cherubino, L. N. Ciuffo, D. Scardaci, "A Digital Library Management System for the Grid", Fourth International Workshop on Emerging Technologies for Next-generation GRID (ETNGRID 2007) at 16th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE-2007), GET/INT Paris, France, June 18-20, 2007 (<a href="http://etngrid.diit.unict.it/2007/index.html">http://etngrid.diit.unict.it/2007/index.html</a>).
  - https://glibrary.ct.infn.it/glibrary/downloads/gLibrary\_paper\_v2.pdf