



Enabling Grids for E-science

gLibrary: a Multimedia Contents Management System on the grid

Tony Calanducci

INFN Catania, NA3 & NA4

EGEE User Forum

01-03 March 2006, CERN, Geneva

www.eu-egee.org



Information Society
and Media



- **Motivations**
- **gLibrary features**
- **Implementation details**
- **Security features**
- **Future planned improvements**
- **Conclusions**

- Huge amounts of data can be saved on SEs (did we forget about the existence of Data Grids?)
- But how can we easily find later a file that we need?
 - (if you have good memory, its GUID could be a solution ☺)
 - File Catalogues just let us to arrange files in folders and subfolders, no way to *query on their contents*
 - Metadata Catalogues are a possible solution, but not always “affordable” especially for non expert users (powerful but complex to use)
- Our solution: a higher level application built on top of many gLite grid services: a Metadata Catalogue + File Catalogues + Storage Elements → **gLibrary**
- Requirements: **easy to use, fast, secure, extensible**

- **Attempt to create a Multimedia Management System on the Grid**
 - Examples of Multimedia Contents handled by gLibrary:
 - Images
 - Movies
 - Audio Files
 - Office Documents (Powerpoint, Word, Excel, OpenOffice)
 - E-Mails, PDFs, HTMLs
 - Customized versions of well-know document type (ex. EGEE PPTs)
 -
- **Keep track and organize in a uniform way all the additional details (metadata) of files saved in Storage Elements and registered in File Catalogues**
- **Provide users an easy way to locate and retrieve files based on their contents**

- **Example 1:**
 - Locate all theoretical (PPTType) PowerPoint (Type) presentations about FireMan (Keywords) given in 2005 (Date) by Uncle Sam (Speaker);
 - Find all the movies (Type) in which Julia Roberts (Cast) performed together with Hugh Grant (Cast) produced in USA (Country) in 2004 (ReleaseDate); or all the acoustic (Genre) mp3 (Format) audio files (Type) of Alanis Morissette (Singer) that last more than 3 minutes (Runtime).
- **Example 2:**
 - A doctor is looking for brain (keyword) DICOM (Type) images of male (Gender) patients older than 65 (Age).
- **Example 3:**
 - A job can behave as a storage crawler: it scans pre-existing files in Storage Elements to extract relevant metadata that will be published on gLibrary for further data mining.

- Files are saved on SEs and registered into file catalogues (LFC and/or FiReMan)
- **The AMGA Metadata Catalogue is** used to archive and organize metadata and to answer users' queries.
- gLibrary is built using the following AMGA collections:
 - /gLibrary contains generic metadata for each entry
 - /gLAudio, /gLImage, /GLVideo, /gLPPT, /EGEEPPT, /gLDoc, ... are examples of collections of “additional features” (shown later)
 - /gLTypes
 - keeps the associations between document types and the names of the collection that contains the “additional features”
 - is used by gLibrary to find out where it has to look when new document types are added into the system (extensibility)
 - /gLKeys is used to store Decryption Keys

Collection		/gLibrary		
Entry Names	Attributes			
	FileName	PathName	Type	Submitter
	4ffaafc8-26e7-4826-b460-3d5bf08081a4	DedicatoAte.mp3	/grid/gilda/calanducci	Audio Tony Calanducci
00454dca-a269-4b93-8a45-c4012af05600	ardizzonelarocca_is_231005.ppt.gpg	/grid/gilda/calanducci/ EGEE	EGEEDOC Tony Calanducci	

/gLibrary (continuum)

Attributes				
SubmissionDate	Encryption	Description	Keywords	CreationDate
2006-01-05 00:00:00	false	Canzone delle vibrazioni che ha ricevuto un enorme successo tra i teenagers nel 2003	Vibrazioni	2004-02-05 00:00:00
2005-01-05 16:44:22	true	gLite Information System	R-GMA, RGMA, BDII, IS	2005-10-05 23:40

Example of gLibrary collections

Collection	/gLTypes
Entry names	Attributes
	Path (<i>refers to a collection</i>)
Audio	/gLAudio
Image	/gLImage
Video	/gLVideo
Documents	/gLDOC
PowerPoint	/gLppt
EGEEDOC	/EGEEPPT

Collection	/gLKeys
Entry names	Attributes
	Passphrase
00454dca-a269-4b93-8a45-c4012af05600	ardizzo

“additional features”

Collection	/EGEEPPT							
Entry names	Attributes							
	Title	Runtime	Author	Type	Date	Event	Speaker	Topic
00454dca-a269-4b93-8a45-c4012af05600	Information Systems	00:30:00	Valeria Ardizzone, Giuseppe La Rocca	Theoretical	2005-10-23	4 th EGEE Conference	Giuseppe La Rocca, Valeria Ardizzone	R-GMA, BDII

Collection	/gLAudio					
Entry names	Attributes					
	SongTitle	Duration	Album	Genre	Singer	Format
4ffa9fc8-26e7-4826-b460-3d5bf08081a4	Dedicato A Te	00:03:27	Dedicato A Te	Pop	Le Vibrazioni	MP3


```
Query> selectattr /gLibrary:FILE /gLibrary:FileName /gLibrary:Description
/EGEEPPT:Author /EGEEPPT:Title /EGEEPPT:Event '/gLibrary:FILE=/EGEEPPT:FILE and
like(/gLibrary:Keywords, "%VOMS%") `
>> 1f6e9ac6-5c86-4599-b03b-560e0e7ea38a
>> VOMS_server_Installation.ppt.gpg
>> VOMS Server installation tutorial done in Venezuela
>> ziggy, Giorgio
>> Installing a gLite VOMS Server
>> First Latin American Workshop for Grid Administrators
```

```
Query> selectattr /gLibrary:FileName SubmissionDate Submitter
/gLAudio:SongTitle Singer Duration Genre '/gLibrary:FILE=/gLAudio:FILE and
/gLAudio:Format="MP3" '
>> DedicatoAte.mp3
>> 2006-01-05 00:00:00
>> Tony Calanducci
>> Dedicato A Te
>> Le Vibrazioni
>> 00:03:27
>> Pop
```

- **User Requirements:**
 - a valid proxy with VOMS extensions
 - VOMS Role and Group needed to be recognized by gLibrary as a contents manager.
- **3 kinds of users:**
 - **gLibraryManager:** (s)he can create new content type and allows a generic VO user to become gLibrarySubmitter
 - **gLibrarySubmitters:** they can add new entries and define access rights on the entries they create.
 - Fine-grained permission (reading, writing, listing, decrypting) settings on each entry: whole VO members, VO groups, list of DNs
 - **generic VO users:** browse and make queries (on entries they have access to)
- **Basic level of cryptography:**
 - New files saved on SEs can be encrypted beforehand with a symmetric passphrase that will be saved in /gLKeys. Only selected users (that have a specific DN in the subject of their VOMS proxy) can access the passphrase and decrypt the file.

```

Connecting to amga.ct.infn.it:8822...
ARDA Metadata Server 1.1.0
Query> whoami
>> tony
Query> user_listcred tony
>> 'C = IT, O = GILDA, OU = Personal Certificate, L = INFN Catania,
    CN = Tony Calanducci, emailAddress = tony.calanducci@ct.infn.it'
Query> grp_member
>> gilda:users
>> gLibraryManager:glibrarysubmitters
Query> addentry /gLibrary/1f6e9ac6-5c86-4599-b03b-560e0e7ea38a
  FileName VOMS_server_Installation.ppt.gpg PathName
  /grid/gilda/calanducci/EGEE Type EGEEDOC Submitter 'Tony
  Calanducci' SubmissionDate '2006-01-07 18:44' DecryptKeyDir
  '/DLKeys/gildateam' Description 'VOMS Server installation
  tutorial done in Venezuela' Keywords 'VOMS Server' CreationDate
  '2005-10-08 18:28'
Query> acl_show /gLibrary/1f6e9ac6-5c86-4599-b03b-560e0e7ea38a
>> tony rwxr-x
>> gLibraryManager:glibrarysubmitters rwx
  
```

```
Query> dir /gLibrary
>> /gLibrary/00454dca-a269-4b93-8a45-c4012af05600
>> entry
>> /gLibrary/abd52d35-1bee-4de9-b234-a9abd920307e
>> entry
>> /gLibrary/1f6e9ac6-5c86-4599-b03b-560e0e7ea38a
>> entry
```

Let's logout and login again using a VOMS proxy with just VO Gilda membership (No Role or group)

ARDA Metadata Server 1.1.0

```
Query> whoami
```

```
>> gilda
```

```
Query> grp_member
```

```
>> gilda:users
```

```
Query> dir /gLibrary
```

```
>> /gLibrary/00454dca-a269-4b93-8a45-c4012af05600
```

```
>> entry
```

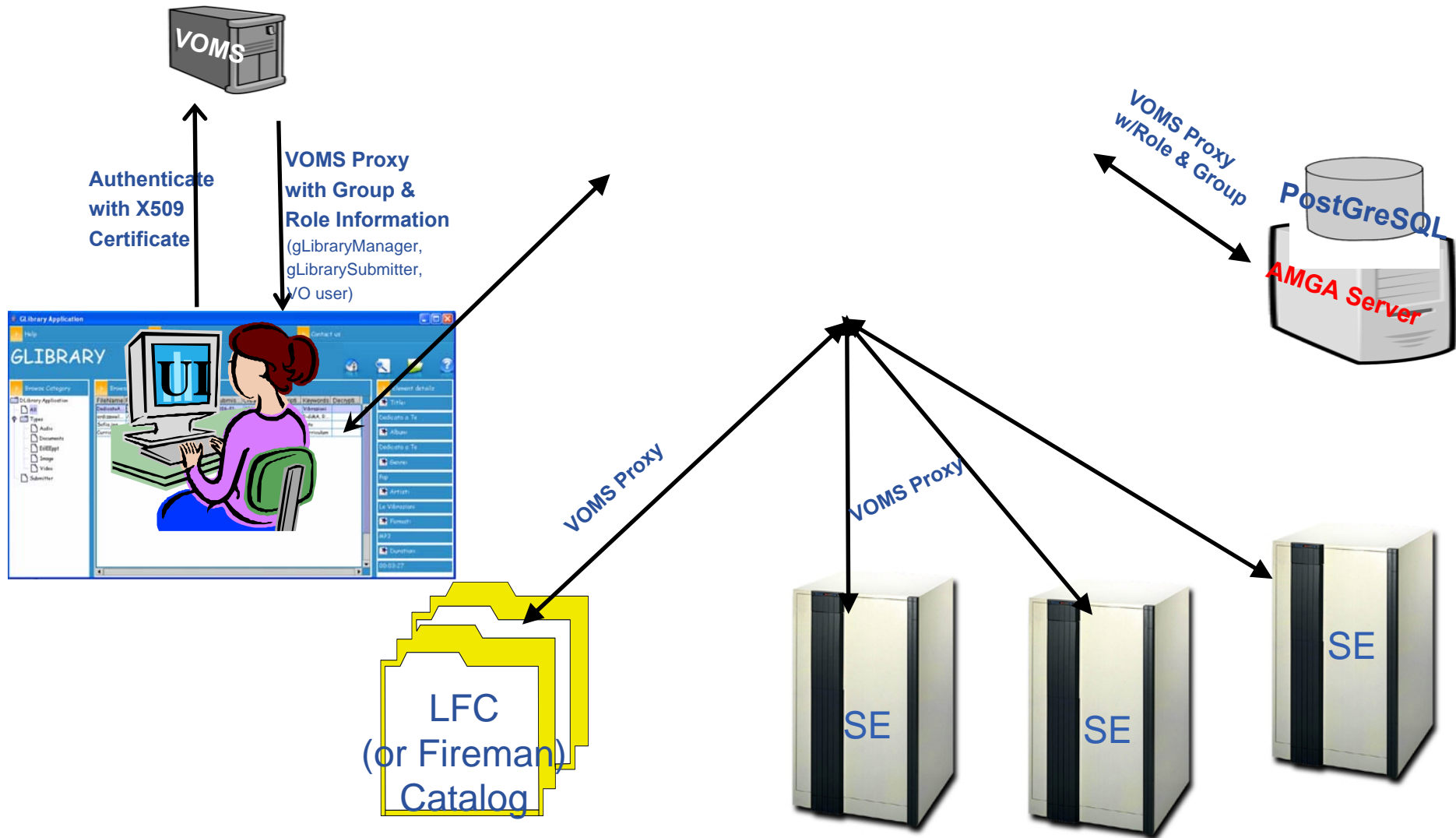
```
Query> acl_show /gLibrary/00454dca-a269-4b93-8a45-c4012af05600
```

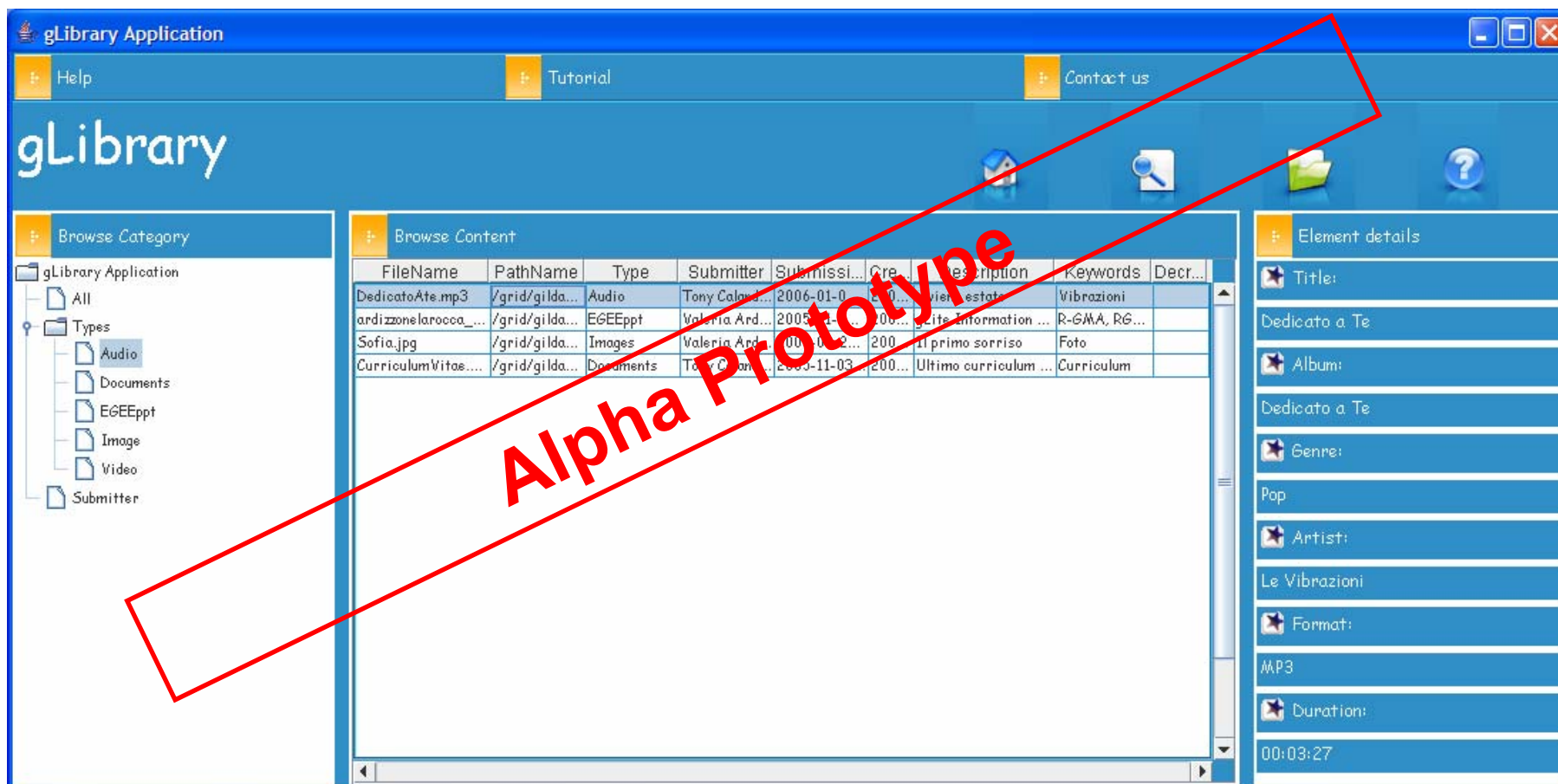
```
>> gLibraryManager rwxr-x
```

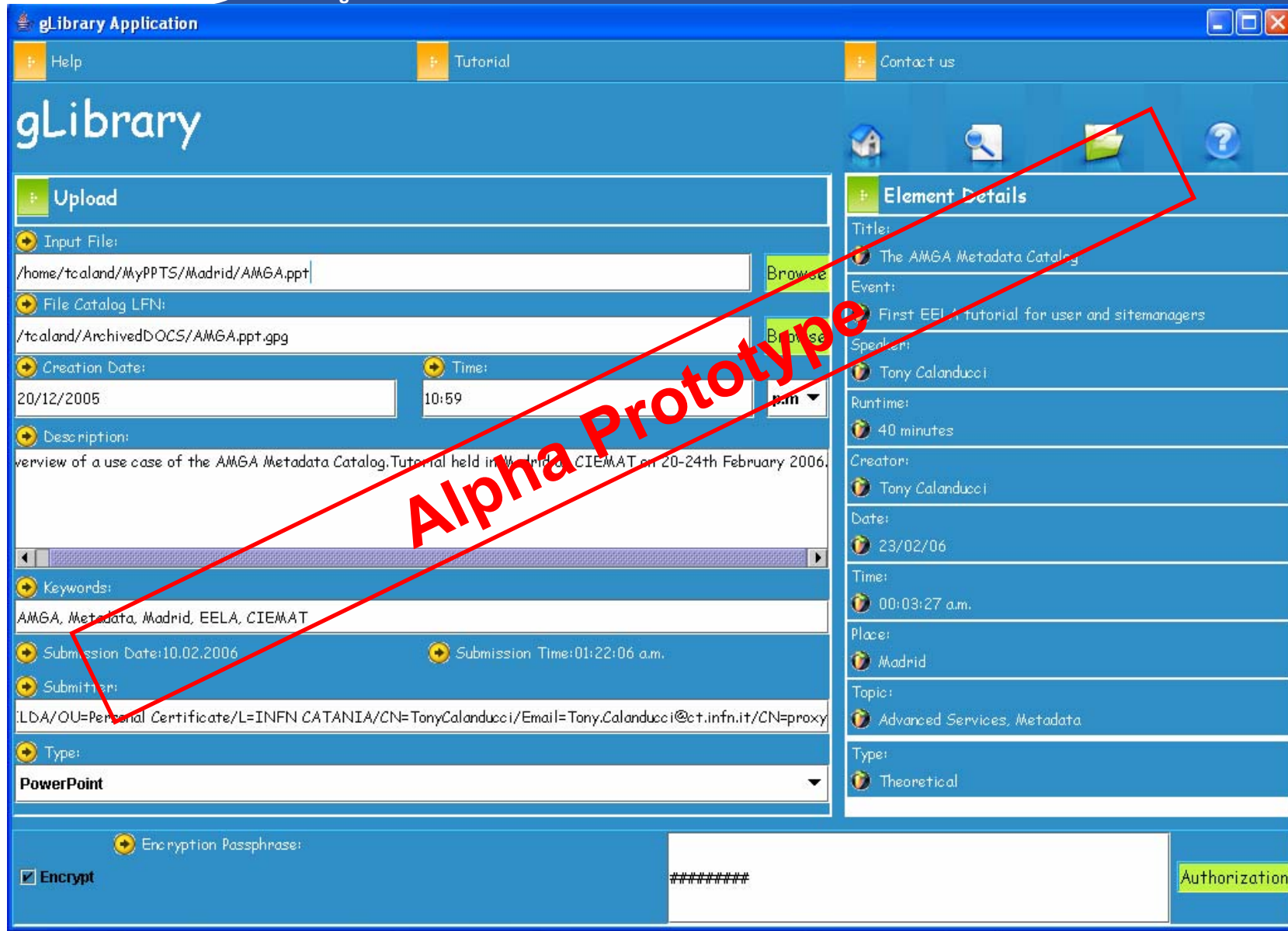
```
>> gilda:users rx
```

The entry previously created does not even appear to non authorized users

- **Heavy exploitation of AMGA features**
 - support for VOMS proxy authentication
 - fine-grained authorization capabilities to set ACLs per entry basis to restrict access to the decryption keys.
 - Allow gLibrarySubmitters to control which users (based on DNs, VOMS Roles and Groups) can list and get the attributes' value for the submitted entries
- **GUI Front-ends (to achieve the “easy of use” promise):**
 - Java SWING GUI to be run on a Grid UserInterface (JVM required) -- prototype is under way
 - Portlet based front-end will be deployed in GENIUSPHERE and made available for any other JSR168 compliant portlets container
 - Both use AMGA Java APIs







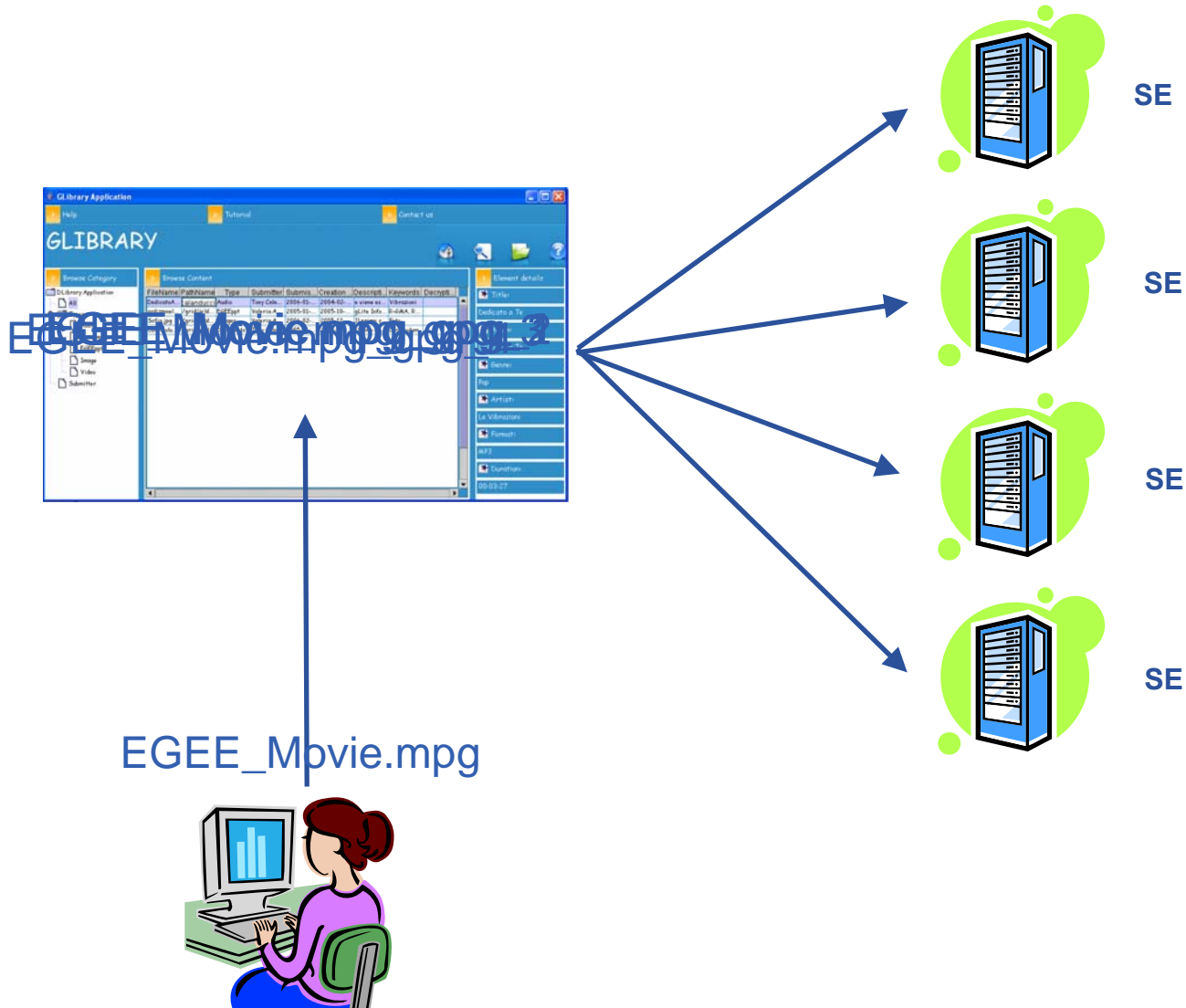
The screenshot shows the gLibrary application window with a blue header and navigation tabs for Help, Tutorial, and Contact us. The main content is divided into two columns:

- Upload Section:**
 - Input File:** /home/tcaland/MyPPTS/Madrid/AMGA.ppt
 - File Catalog LFN:** /tcaland/ArchivedDOCS/AMGA.ppt.gpg
 - Creation Date:** 20/12/2005
 - Time:** 10:59
 - Description:** Overview of a use case of the AMGA Metadata Catalog. Tutorial held in Madrid, CIEMAT on 20-24th February 2006.
 - Keywords:** AMGA, Metadata, Madrid, EELA, CIEMAT
 - Submission Date:** 10.02.2006
 - Submission Time:** 01:22:06 a.m.
 - Submitter:** LDA/OU=Personal Certificate/L=INFN CATANIA/CN=TonyCalanducci/Email=Tony.Calanducci@ct.infn.it/CN=proxy
 - Type:** PowerPoint
 - Encryption Passphrase:** [Redacted]
 - Encrypt**
- Element Details Section:**
 - Title:** The AMGA Metadata Catalog
 - Event:** First EELA tutorial for user and sitemanagers
 - Speaker:** Tony Calanducci
 - Runtime:** 40 minutes
 - Creator:** Tony Calanducci
 - Date:** 23/02/06
 - Time:** 00:03:27 a.m.
 - Place:** Madrid
 - Topic:** Advanced Services, Metadata
 - Type:** Theoretical

A red diagonal watermark reading "Alpha Prototype" is overlaid across the center of the interface. At the bottom right, there is an "Authorization" button.

- **Splitting of big files among several SEs (different chunks stored in different SEs):**
 - Enforce security of data: even if a chunk is intercepted it has no meaning alone.
 - Increase upload/download bandwidth
 - Possible implementation:
 - one more NumberOfChunks attribute in /gLibrary collection.
 - /gLChunks collection keeps track of FirstChunkGUID-Chunk#-ChunkGUID
- **Automatic extraction and population of metadata for well known document types**
 - use of GNU libextractor to extract metadata from HTML, PDF, PS, OLE2 (DOC, XLS, PPT), OpenOffice (sxw), StarOffice (sdw), DVI, MAN, MP3 (ID3v1 and ID3v2), OGG, WAV, EXIV2, JPEG, GIF, PNG, TIFF, DEB, RPM, TAR(.GZ), ZIP, ELF, REAL, RIFF (AVI), MPEG, QT and ASF
 - use of Lucene algorithm for indexing document types containing text
- **Evaluation of gLite Hydra Key Store to save decryptions keys**

Splitting Implementation



- Born as an use case to demonstrate AMGA features
- Built on top of many gLite services
- Considering collaboration and integration with NA3 Document Digital Library System
- **Fast** → thanks to AMGA
- **Secure** → ACLs, encryption, and splitting
- **Easy to use** → User friendly Java GUI and portal soon available
- Easily **extensible** to support any document types (Medical Images and files, Invoices, Proceedings, Scientific Publications, Newspapers clips, ...)



Thanks for the attention