BioinfoGRID
Bioinformatics Grid Application for life science

The GENIUS Grid Portal : some success stories!

Giuseppe LA ROCCA
INFN Catania
giuseppe.larocca@ct.infn.it

Joint EGEE and SEE-GRID Summer School
on Grid Application Support
Budapest - Hungary, June 25-30, 2007

Outline

- Grid Portals technology
- EnginFrame framework
- The GENIUS Grid Portal
- Some Integrated Applications
- References

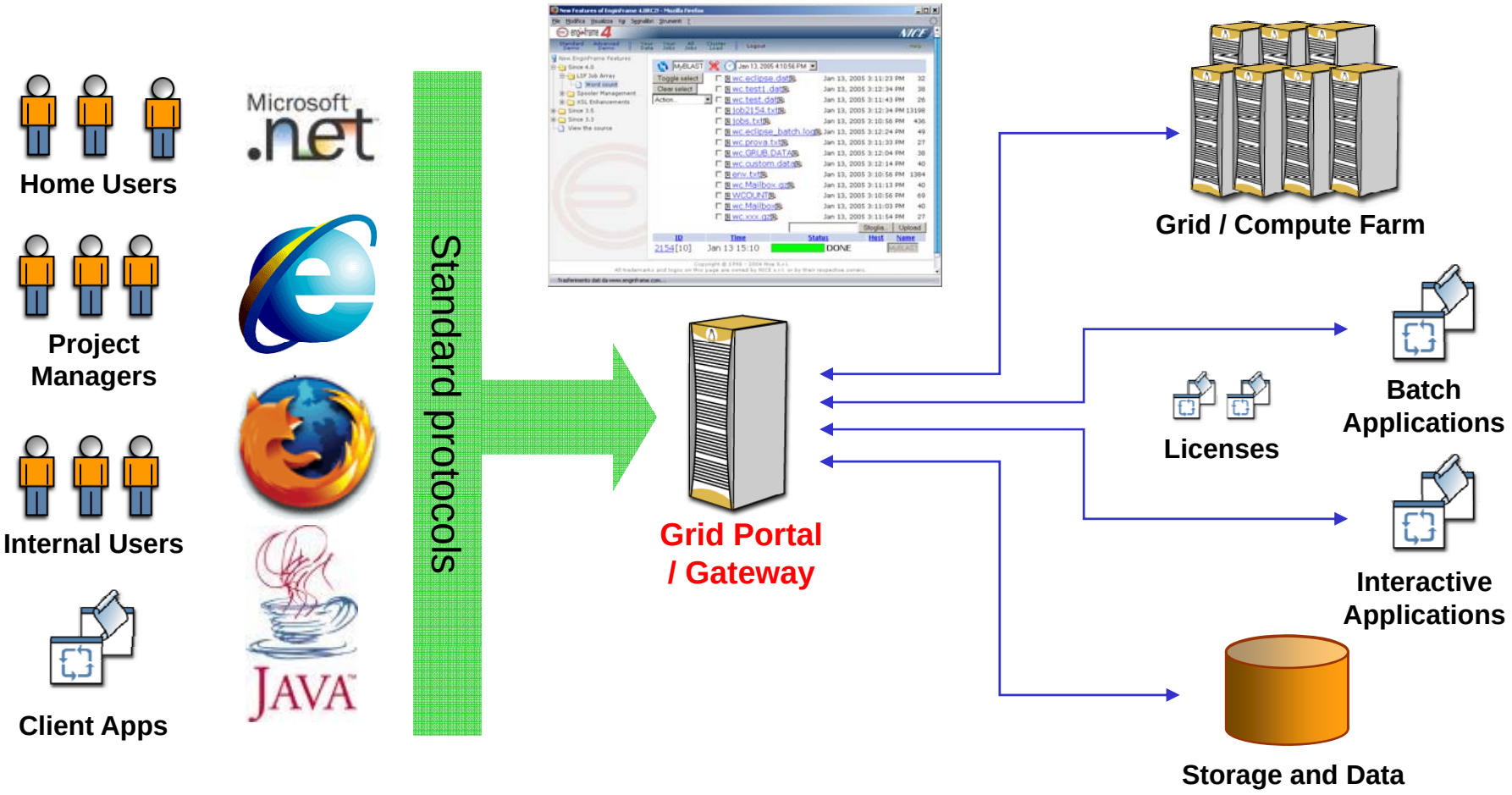
A grid portal: why and how

- It can be accessed from everywhere and by “everything” (desktop, laptop, PDA, cell phone).
- It can keep the same user interface to several back-ends.
- It must be redundantly “secure” at all levels:
  - 1) secure for web transactions,
  - 2) secure for user credentials,
  - 3) secure for user authentication,
  - 4) secure at VO/VOMS level.
- All available grid services must be incorporated in a logic way, just “one mouse click away”.
- Its layout must be easily understandable and user friendly.

Grid Portal benefits

- A Grid Portal improves usability of Grids
  - Lowering end-user requirements for accessing the Grid
  - Hiding the complexity of data and job services management in the Grid
- A Grid Portal improves utilization of Grids
  - Making the Grid (r)evolution transparent to the end-user
  - Providing an appealing user-friendly Web interface
  - Enforcing Grid utilization policies

The Grid Portal / Gateway



- Grid Portals technology
- EnginFrame Framework
- The GENIUS Grid Portal
- Some Integrated Applications
- References

What EnginFrame is ?

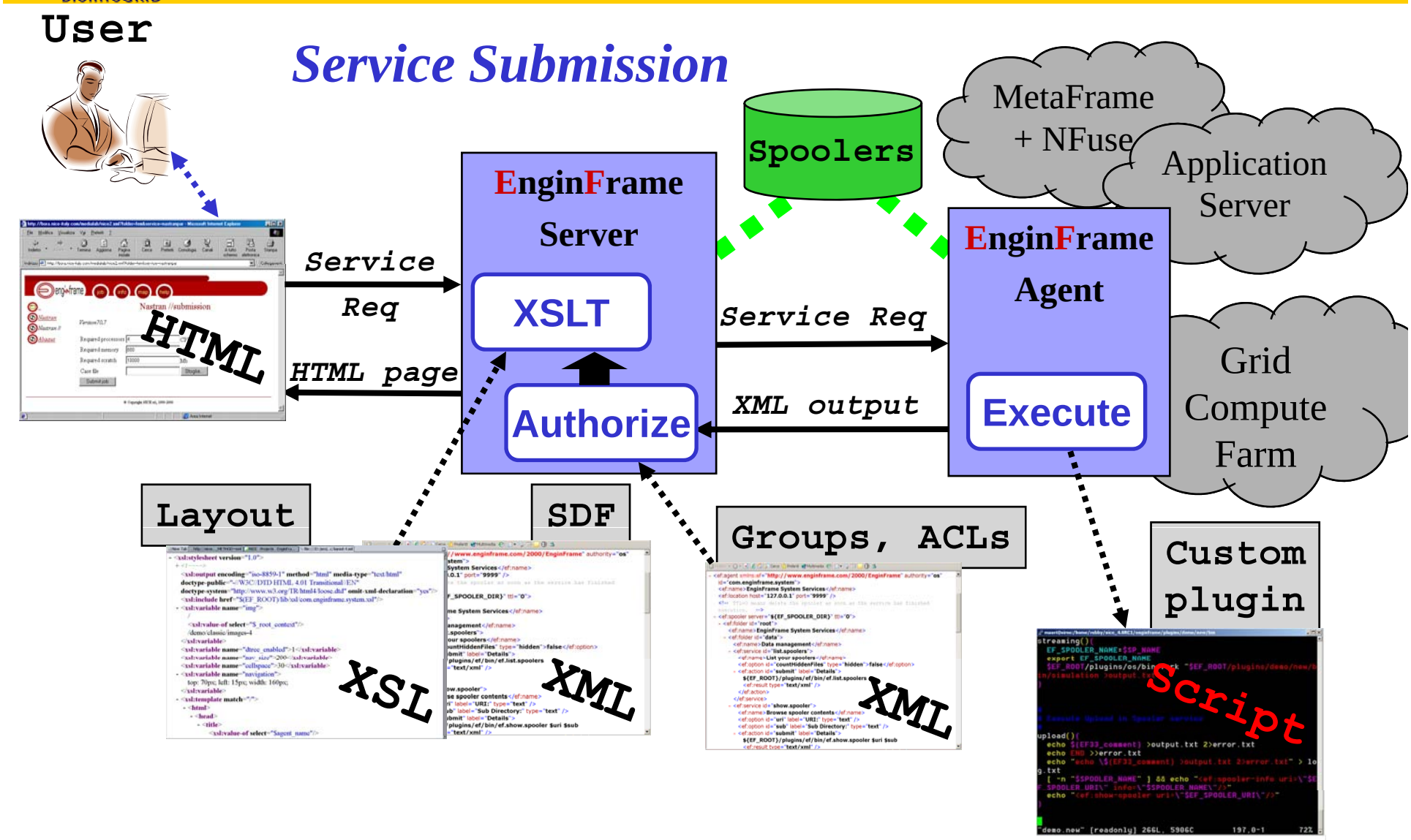
- It is a web-based technology able to expose Grid services running on Grid infrastructures
- It allows organizations to provide application-oriented computing and data services to both users (via Web browsers) and applications (via SOAP/WSDL and/or RSS)
- It's a Grid *gateway*
- It greatly simplifies the development of Web Portals exposing computing services that can run on a broad range of different computational Grid systems

Who uses EnginFrame?

- **Mechanical**
  - Ferrari, Audi, BMW, FIAT Auto, Elasis, Magneti Marelli, P+Z, Swagelok, Toyota, TRW
- **Manufacturing**
  - Bridgestone, Procter & Gamble, Galileo Avionica
- **Oil&Gas**
  - Slavneft, Schlumberger, TOTAL, VNIIGaz
- **Electronics**
  - STMicroelectronics, Accent, SensorDynamics, Motorola
- **Biotech**
  - ENEA, EGEE LS community
- **Telecom**
  - Telecom Italia
- **Research**
  - INFN, ASSC, CCLRC, CERN, CILEA, CINECA, CNR, CNRS/IN2P3, ENEA, FzU, ICI, IFAE, ITEP, JSC G.G.M., KU Leuven, SSC-Russia, SDSC
- **Education**
  - Dresda University, Ferrara University, ITU, Messina University, Politecnico of Milan, Technische Universität Dresden, Trinity College Dublin, Salerno University, S-PACI

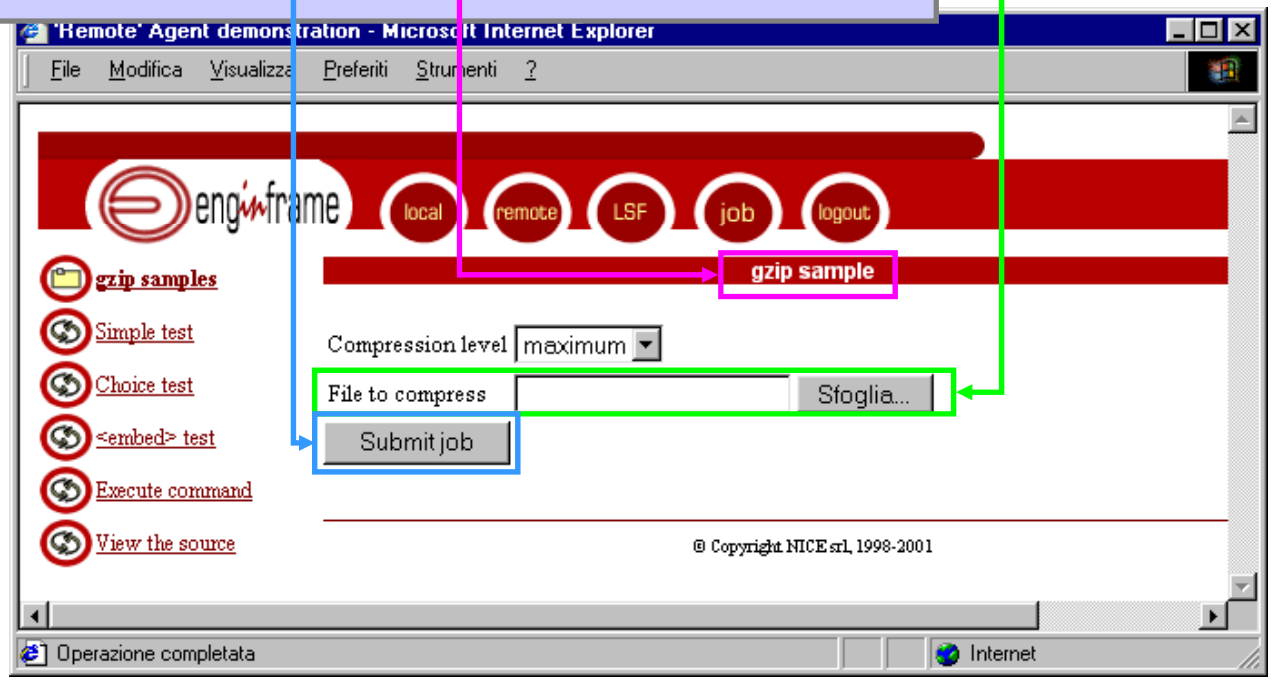
EnginFrame Working Environment

BioinfoGRID



Service example

```
<ef:service id="gzip">
  <ef:name>gzip sample</ef:name>
   <ef:option id="level" label="Compression level" type="list">
     <ef:option id="9">maximum</ef:option>
     <ef:option id="4">medium</ef:option>
     <ef:option id="0">none</ef:option>
   </ef:option>
   <ef:option id="FILE" label="File to compress" type="file"/>
   <ef:action id="submit" label="Submit job">
     EF_SPOOLER_NAME="gzip $file"
     export EF_SPOOLER_NAME
     ${EF_ROOT}/plugins/lsf/bin/bsub -o output.txt gzip -$level \"$FILE\"
     <ef:result type="text/xml"/></ef:action>
</ef:service>
```

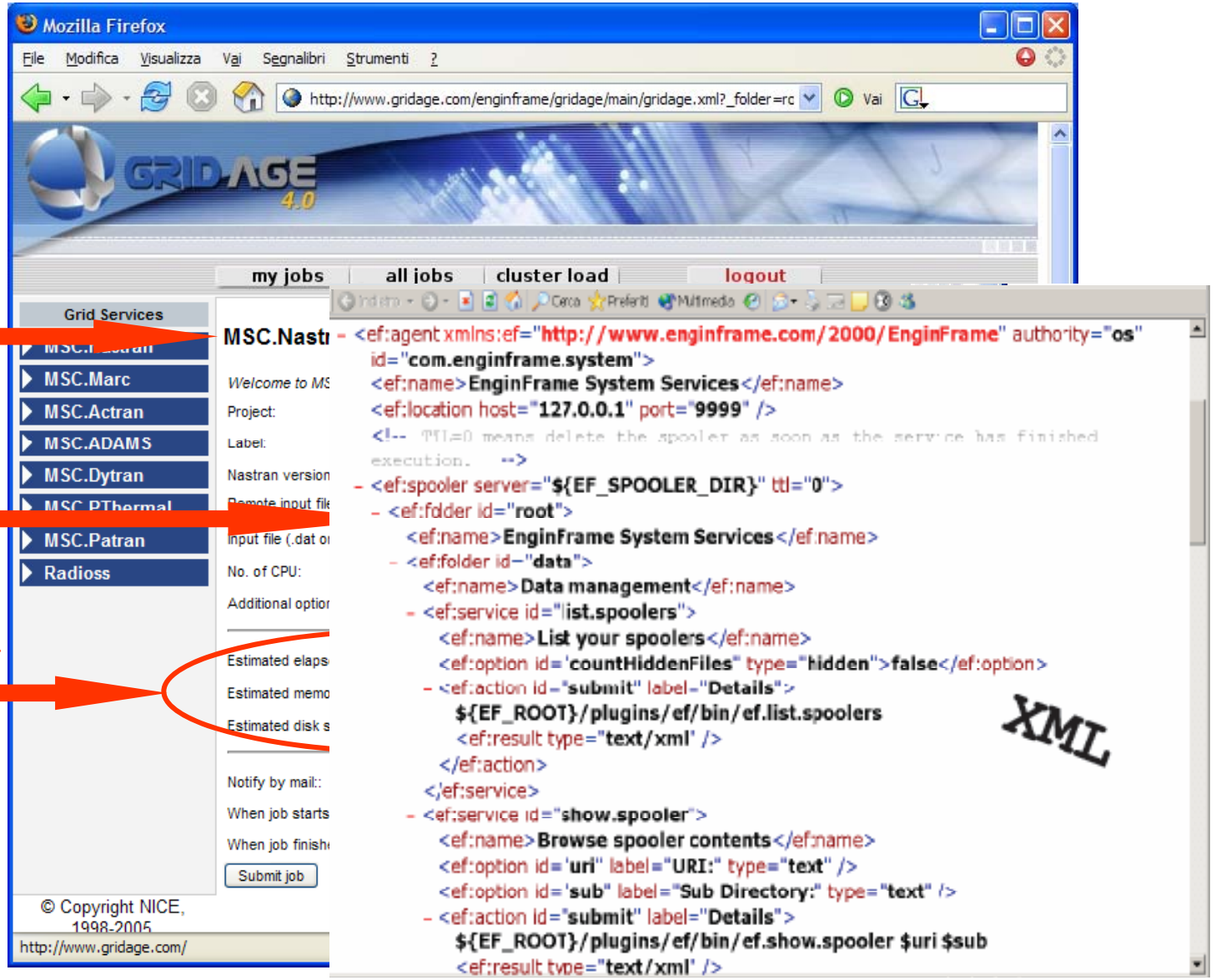


Usability & Input Management

User friendly, Application-oriented Job submission

Flexible and efficient Input file management

Hide complexity of underlying scheduler



- Grid Portals technology
- EnginFrame Framework
- The GENIUS Grid Portal
- Some Integrated Applications
- References

What GENIUS is ?

- GENIUS is a powerful Grid Portal that allows scientists to exploit Grid resources only using a conventional Web browser
- It's a gateway to European EGEE Project middle-ware
- It has been built on top of the EnginFrame framework
- It allows to expose gLite-enabled applications via Web-browser as well as Web Services

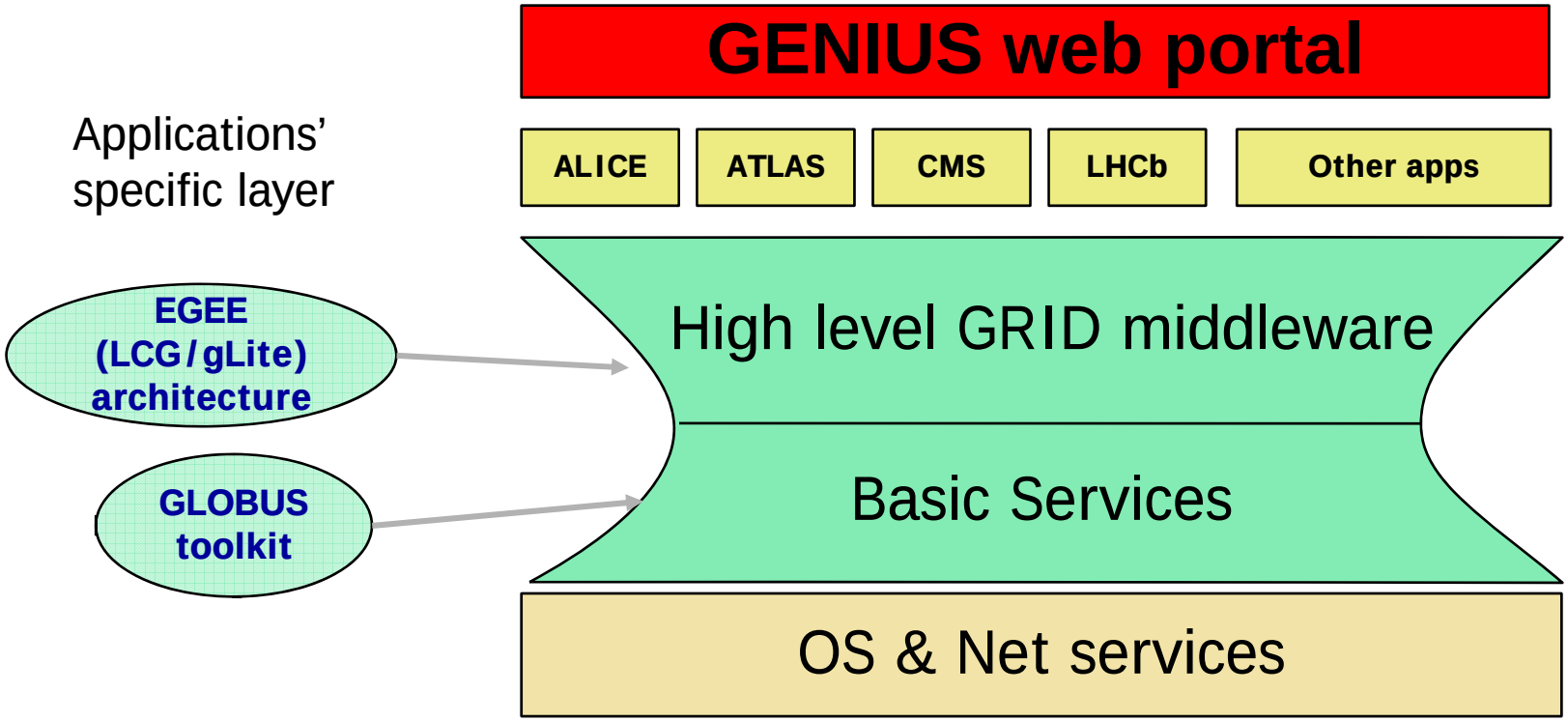
The GENIUS hourglass model

GENIUS

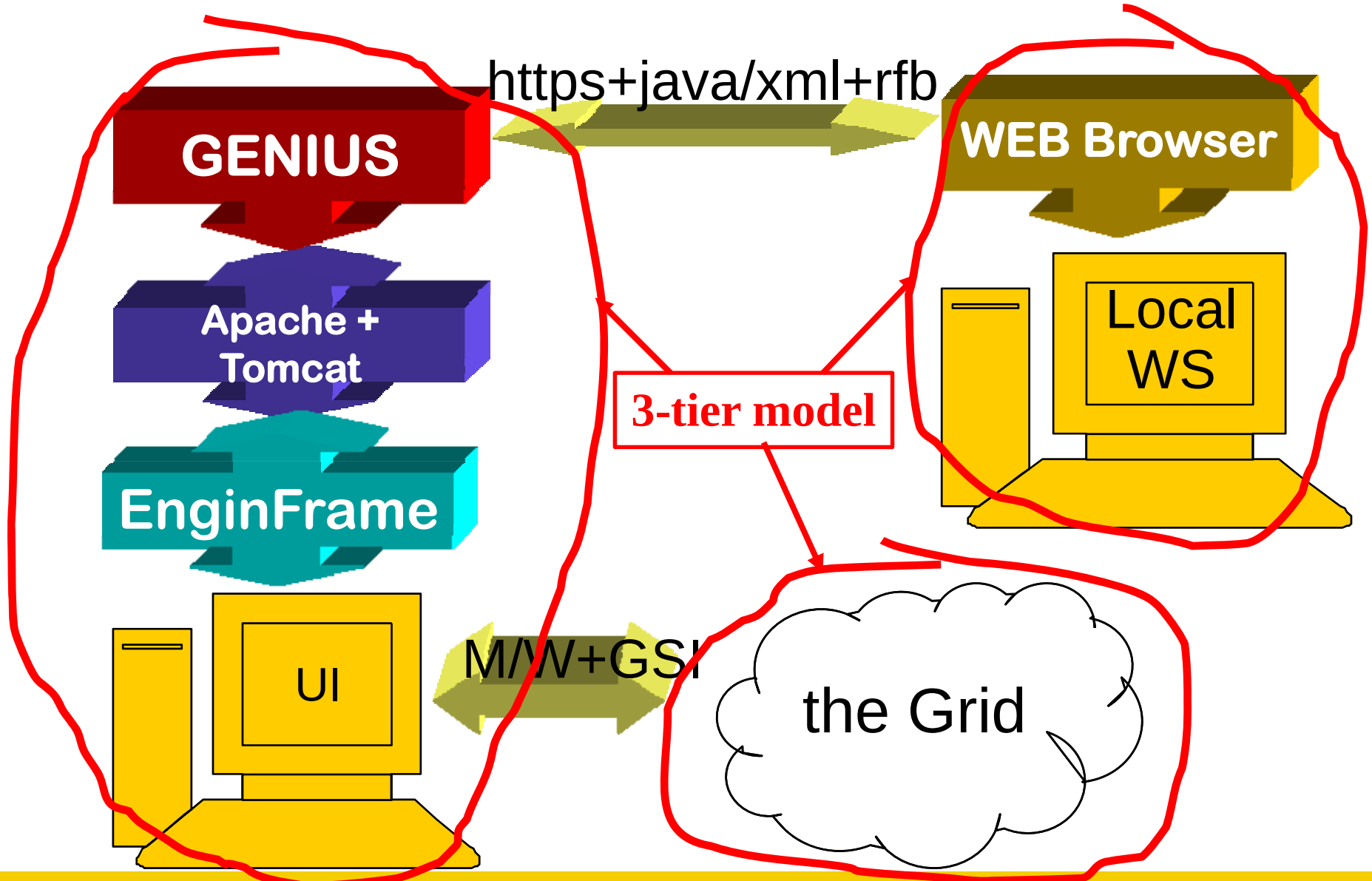
Grid Enabled web eNvironment for site Independent

User job Submission

INFN / NICE Collaboration



GENIUS Architecture



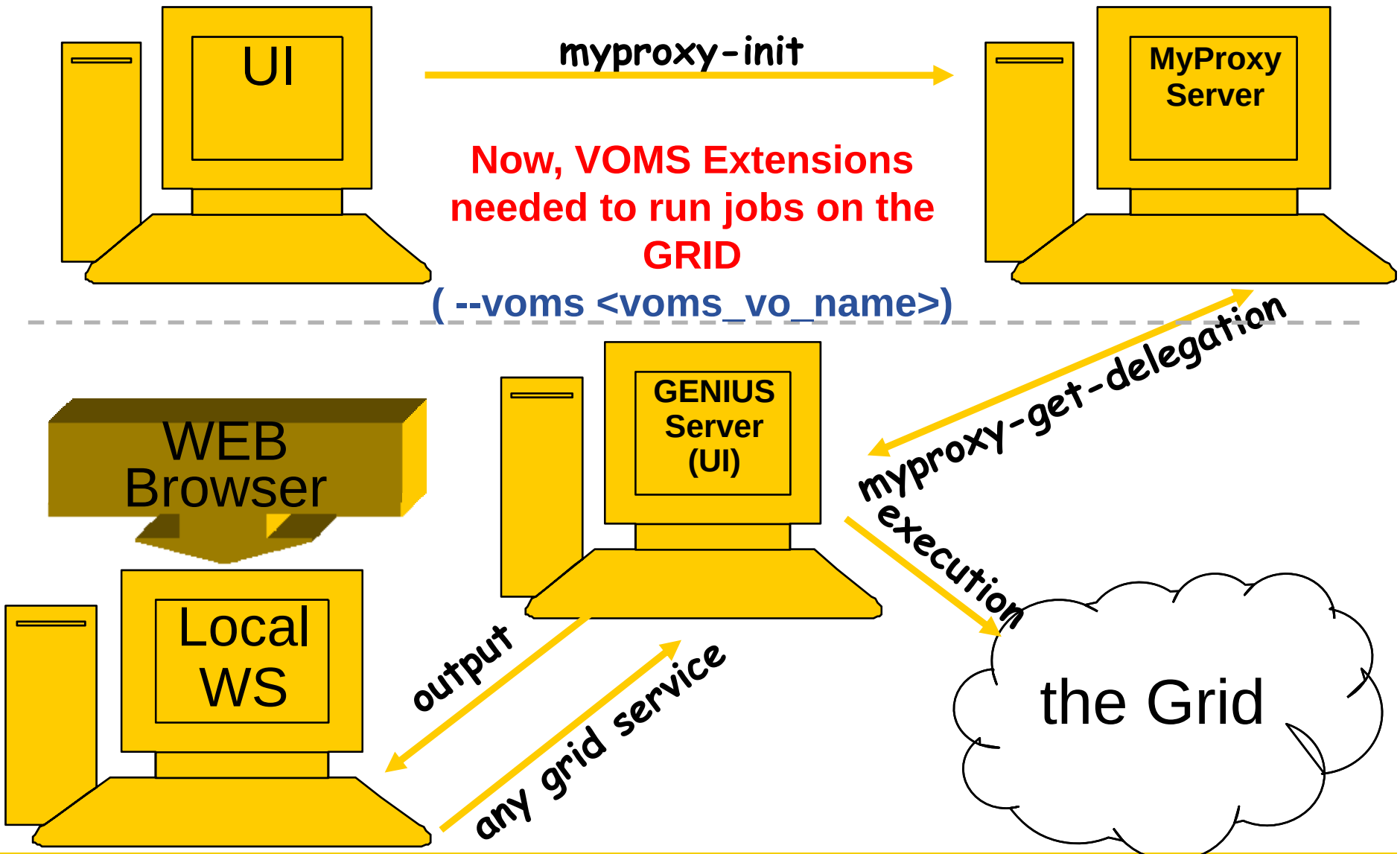
GENIUS Grid Portal

Reference Web Site: https://genius.ct.infn.it

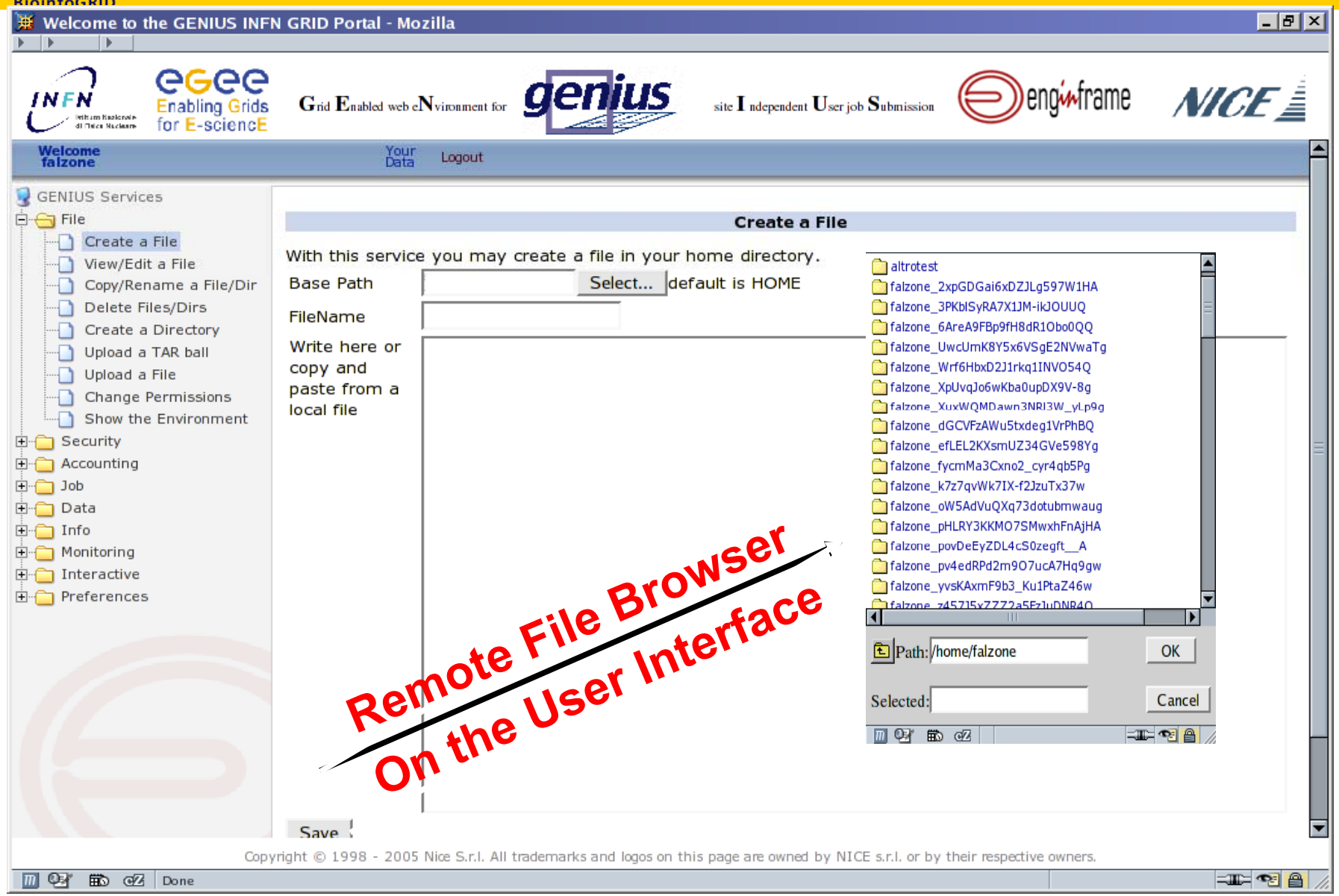
GENIUS : Security Infrastructure

- All web transactions are executed under the Secure Socket Layer (SSL) via HTTPS
- The user must have an account on the User Interface
- When the user wants to interact with the file-system of the UI, GENIUS will prompt for the username and password of user's account on that machine
- Grid authentication in GENIUS is currently based on the concept Proxy

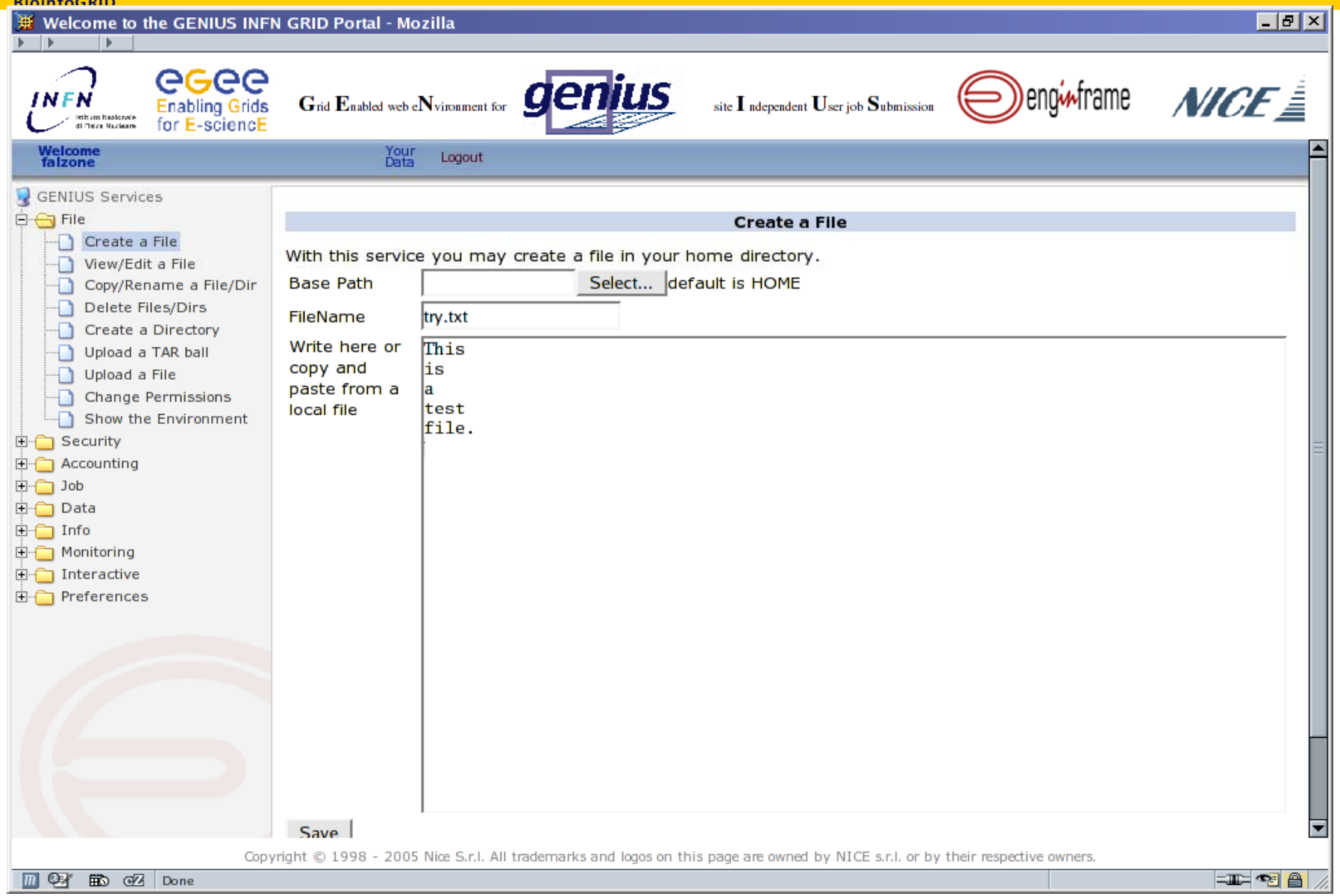
Grid authentication with MyProxy



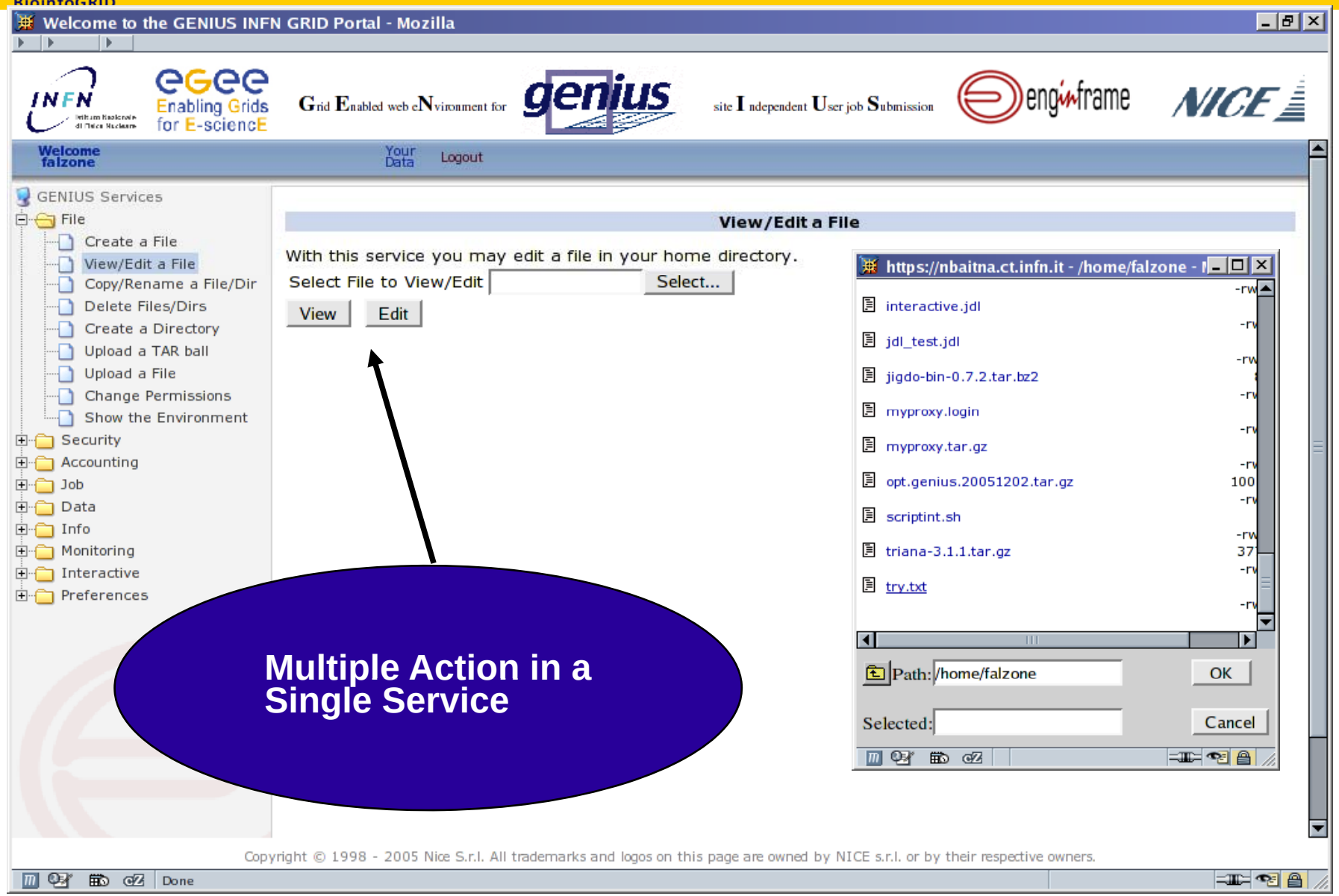
GENIUS: files management



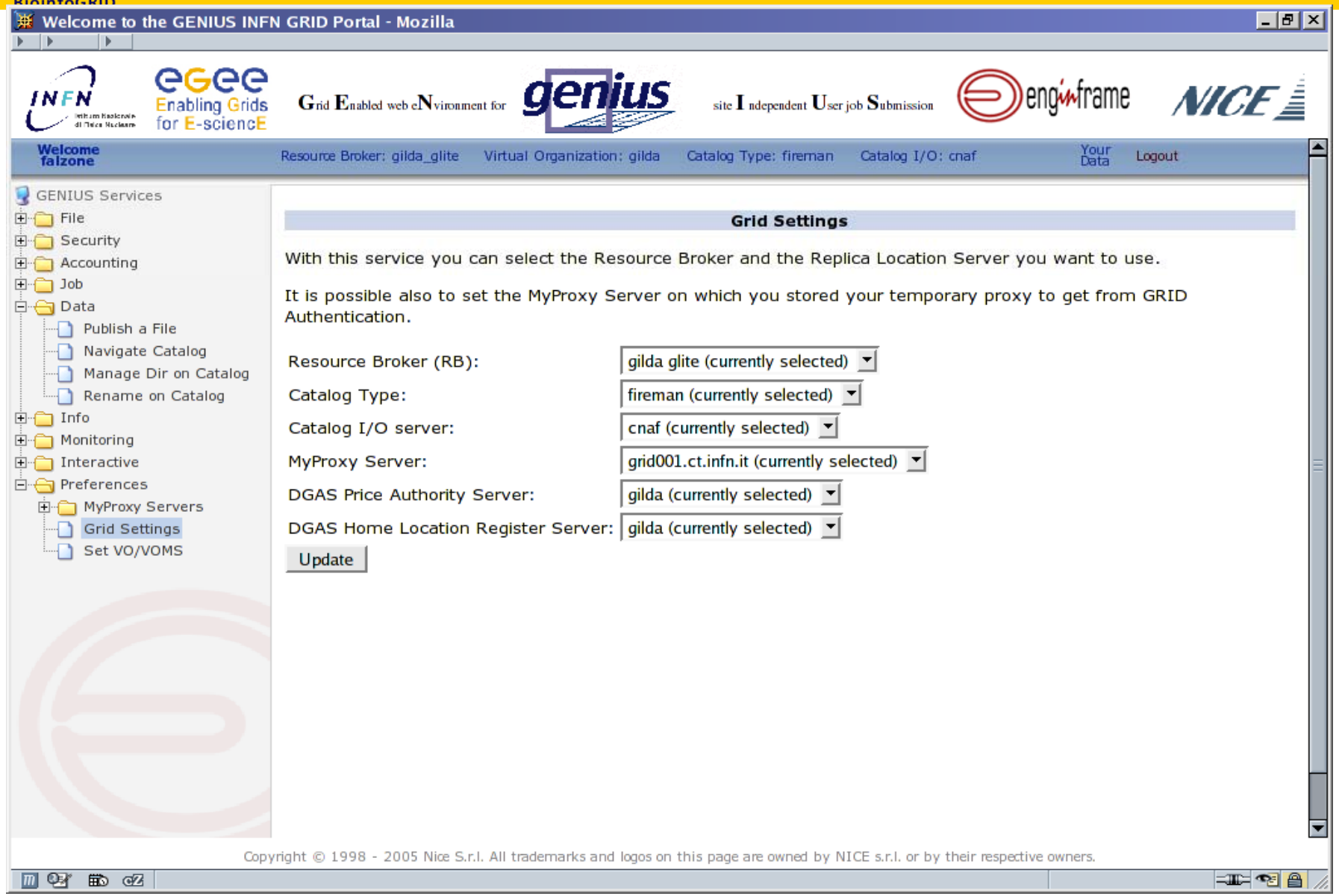
GENIUS: files management



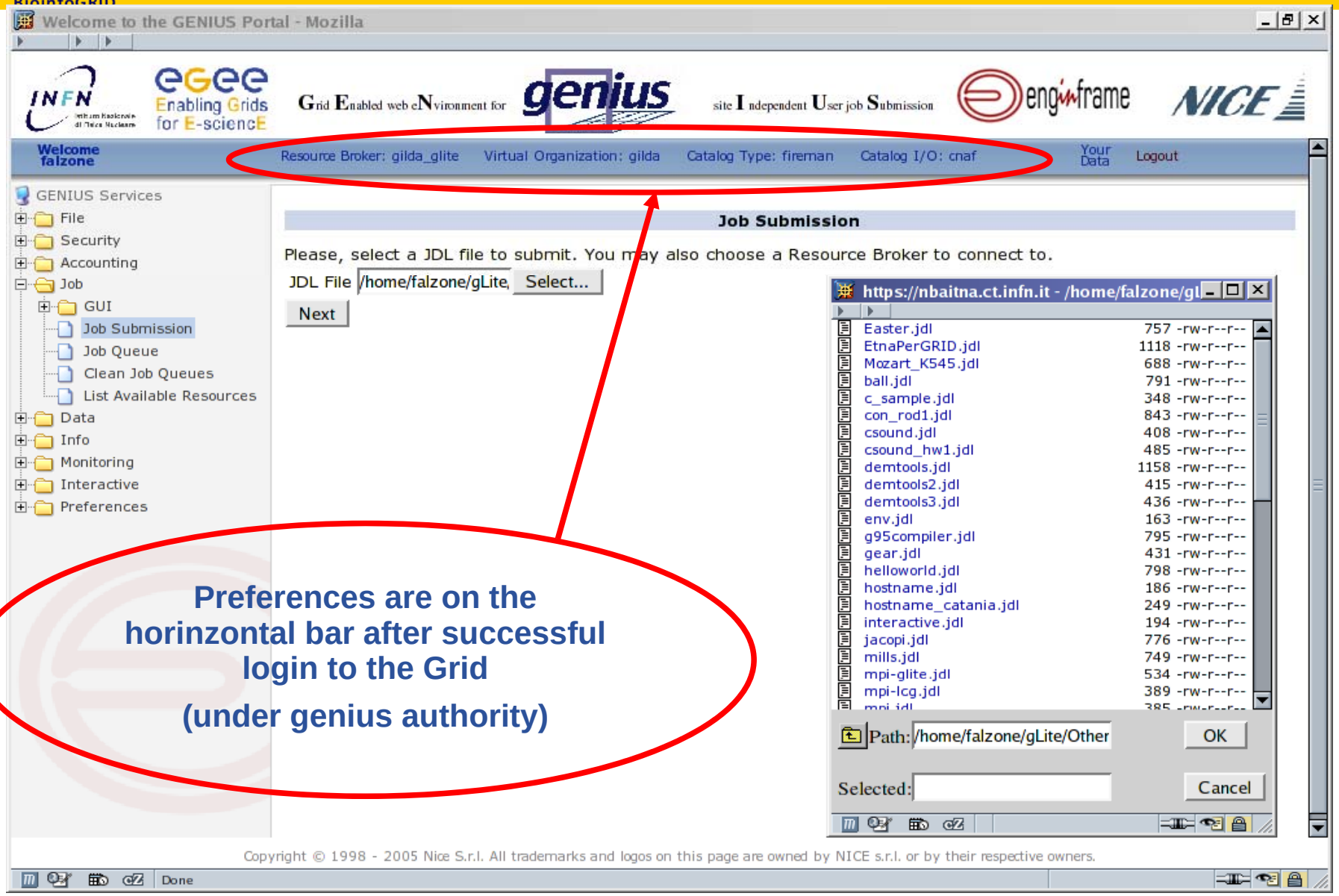
GENIUS: files management



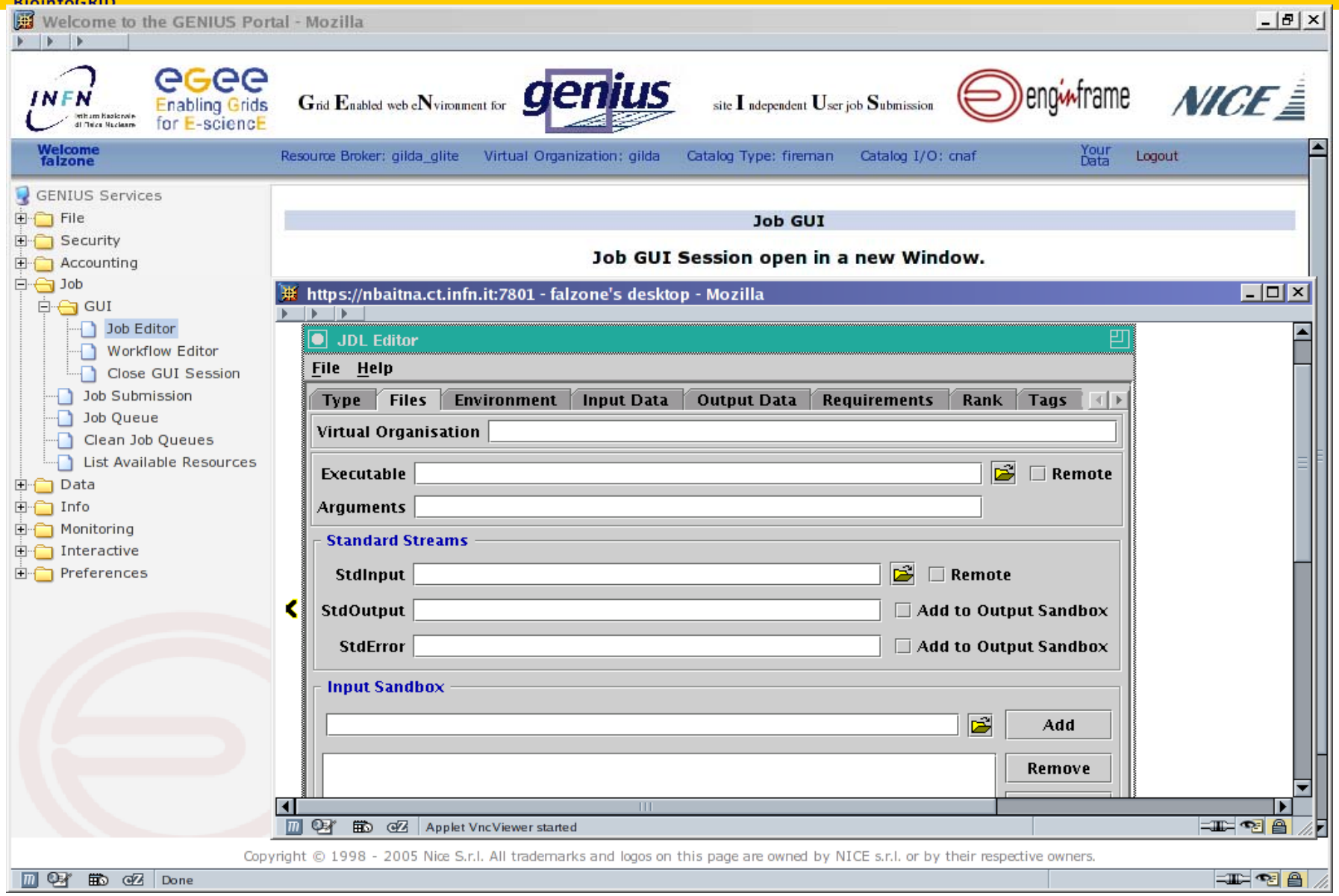
GENIUS: preferences



GENIUS: Job Submission

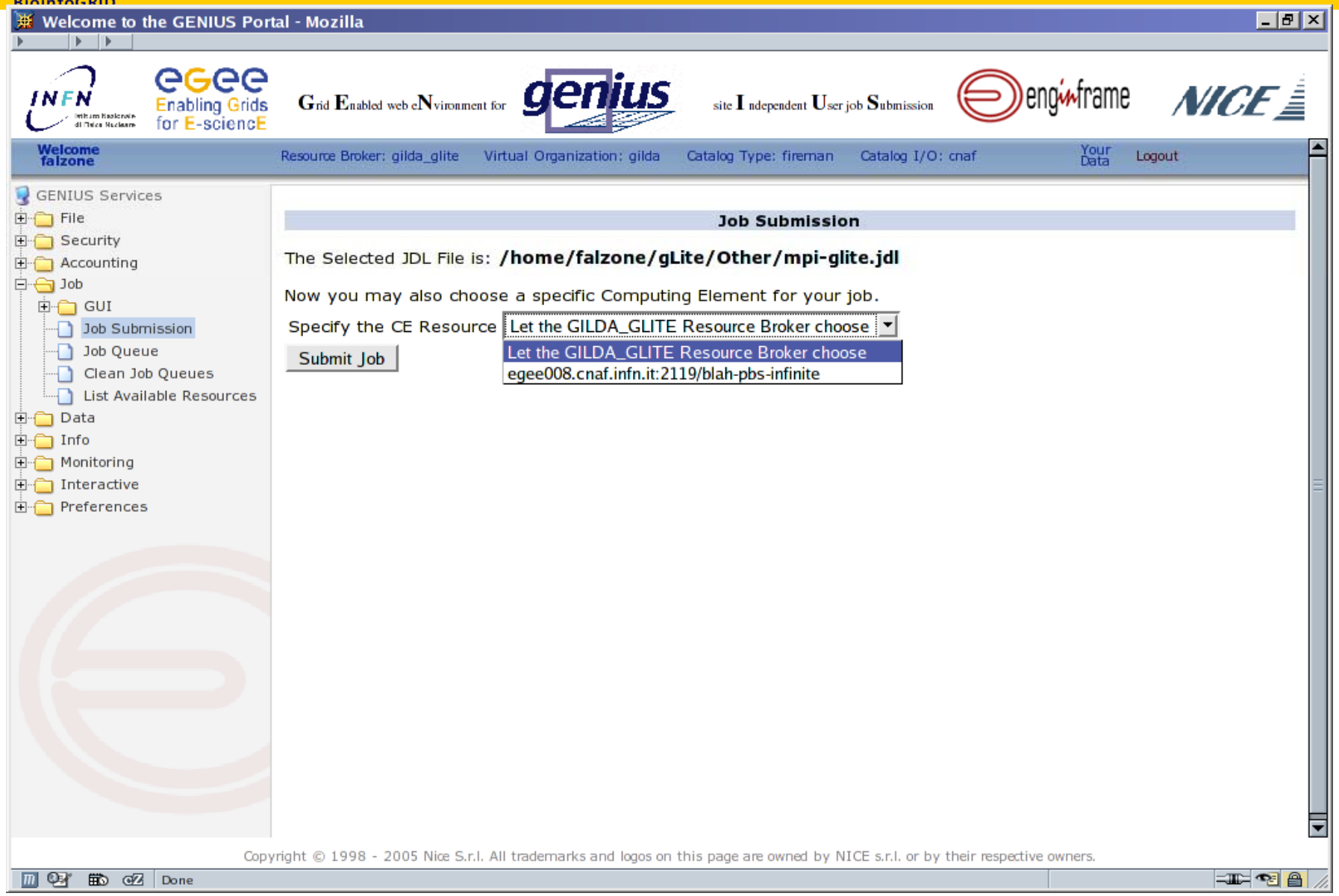


GENIUS: Job Editor for jdl files

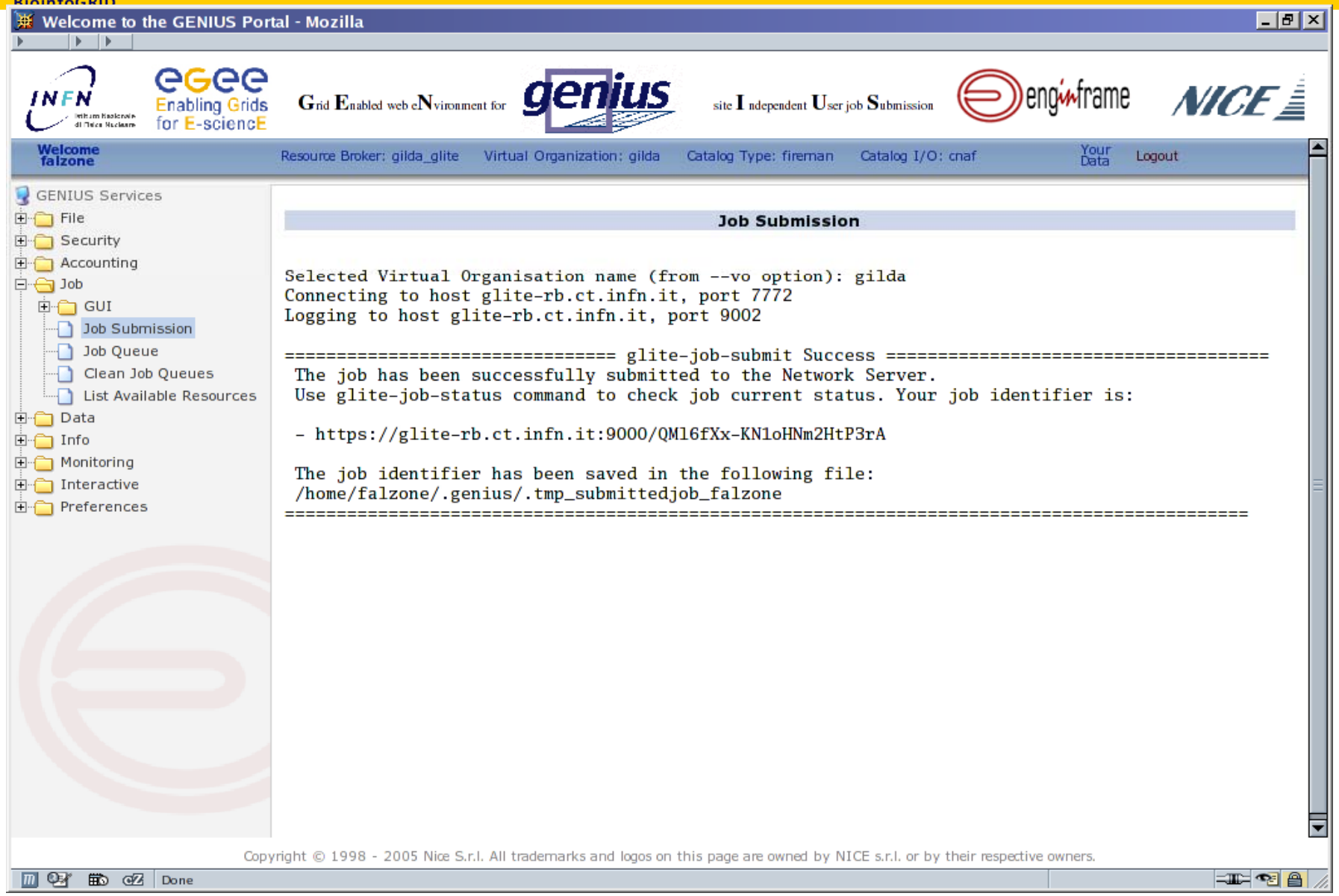


GENIUS: Job Submission

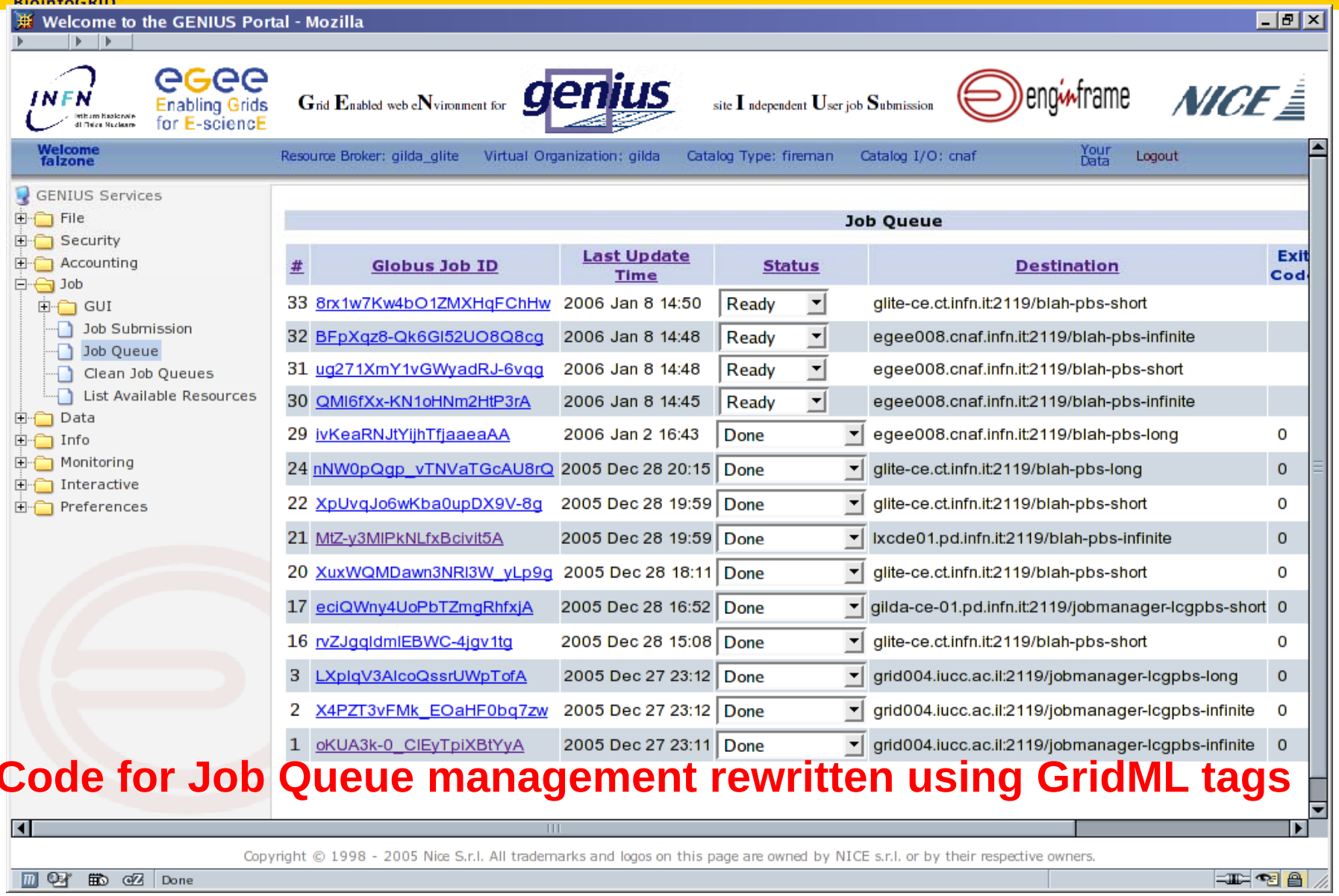
GENIUS: Job Submission



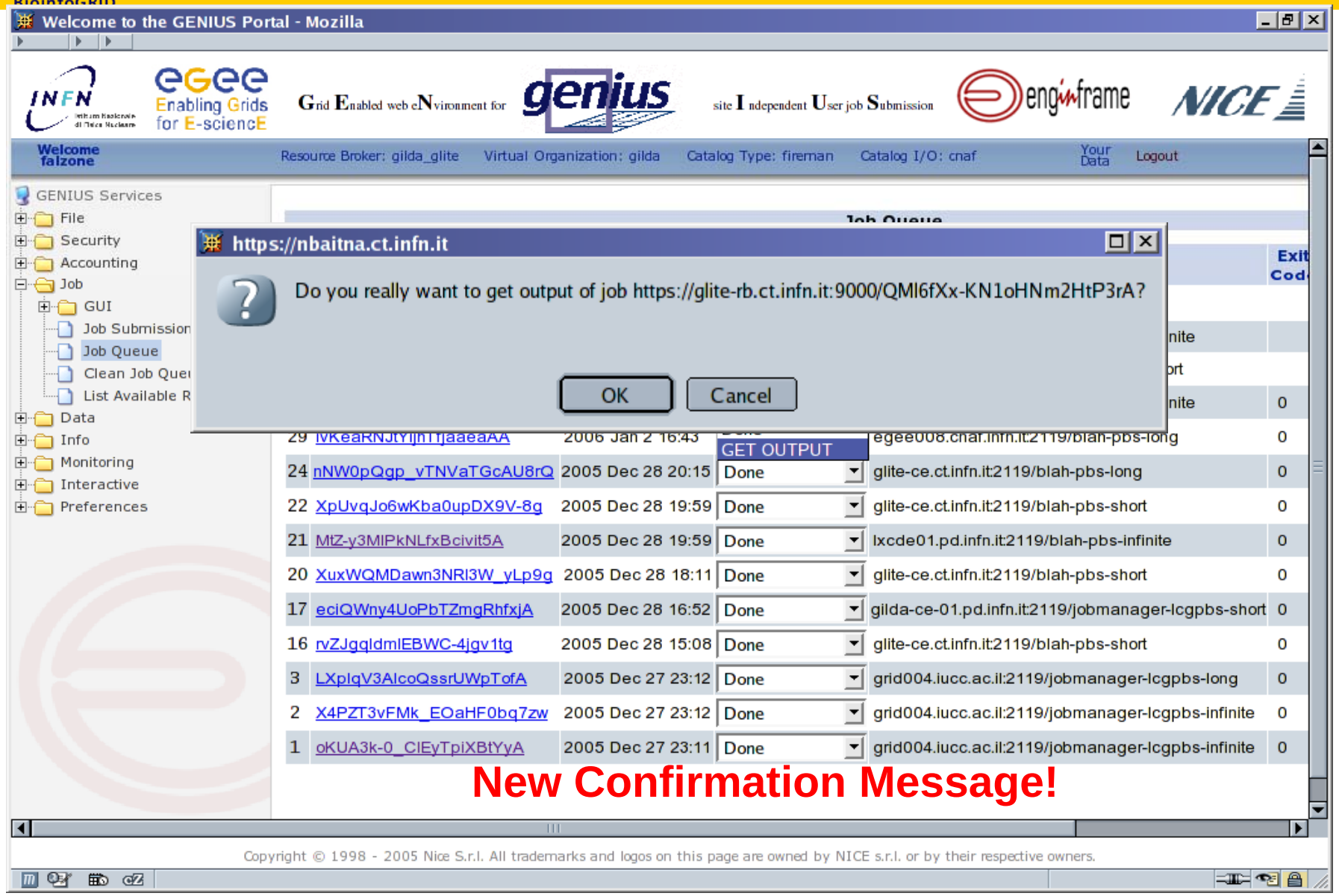
GENIUS: Job Submission



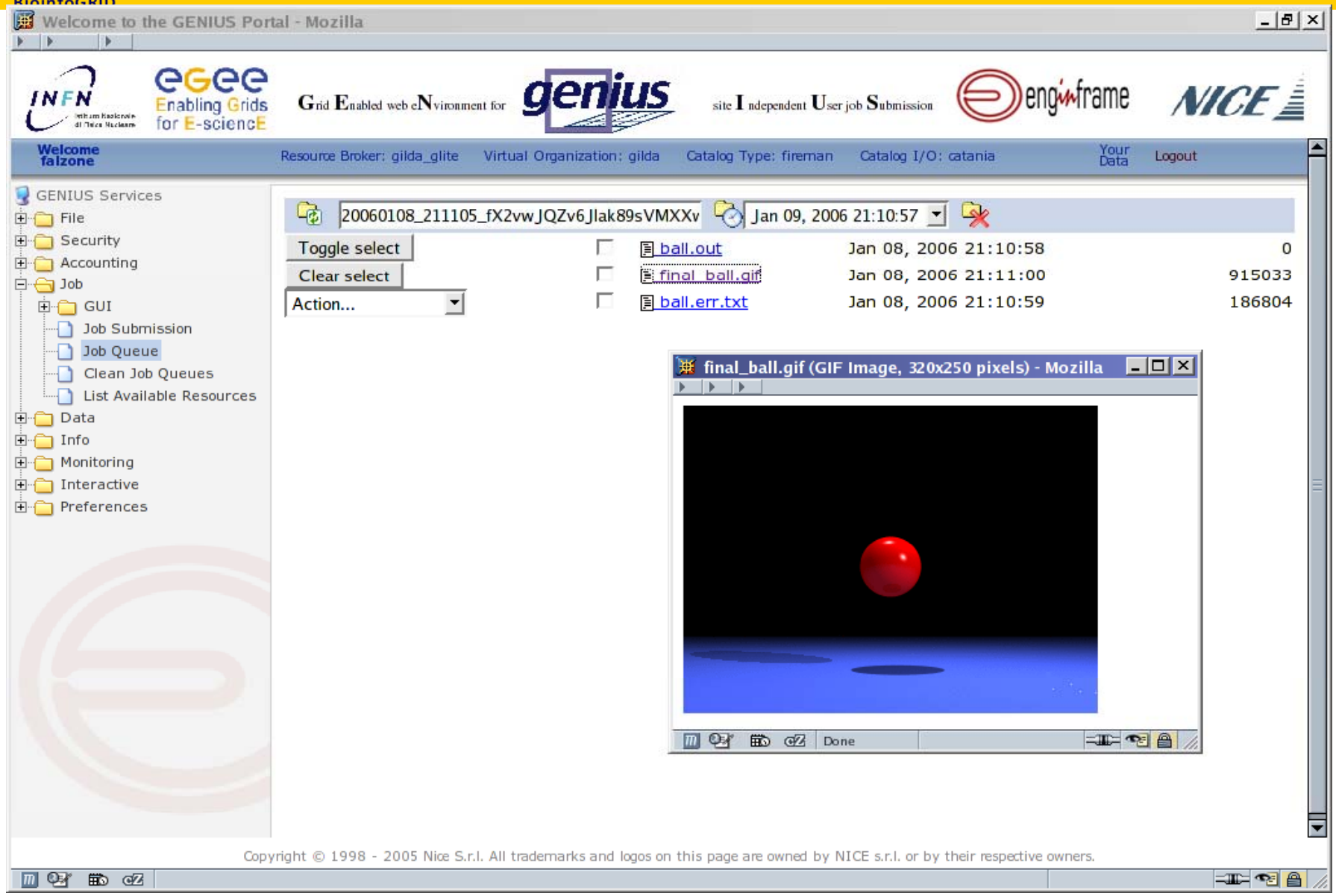
GENIUS: Job Queue



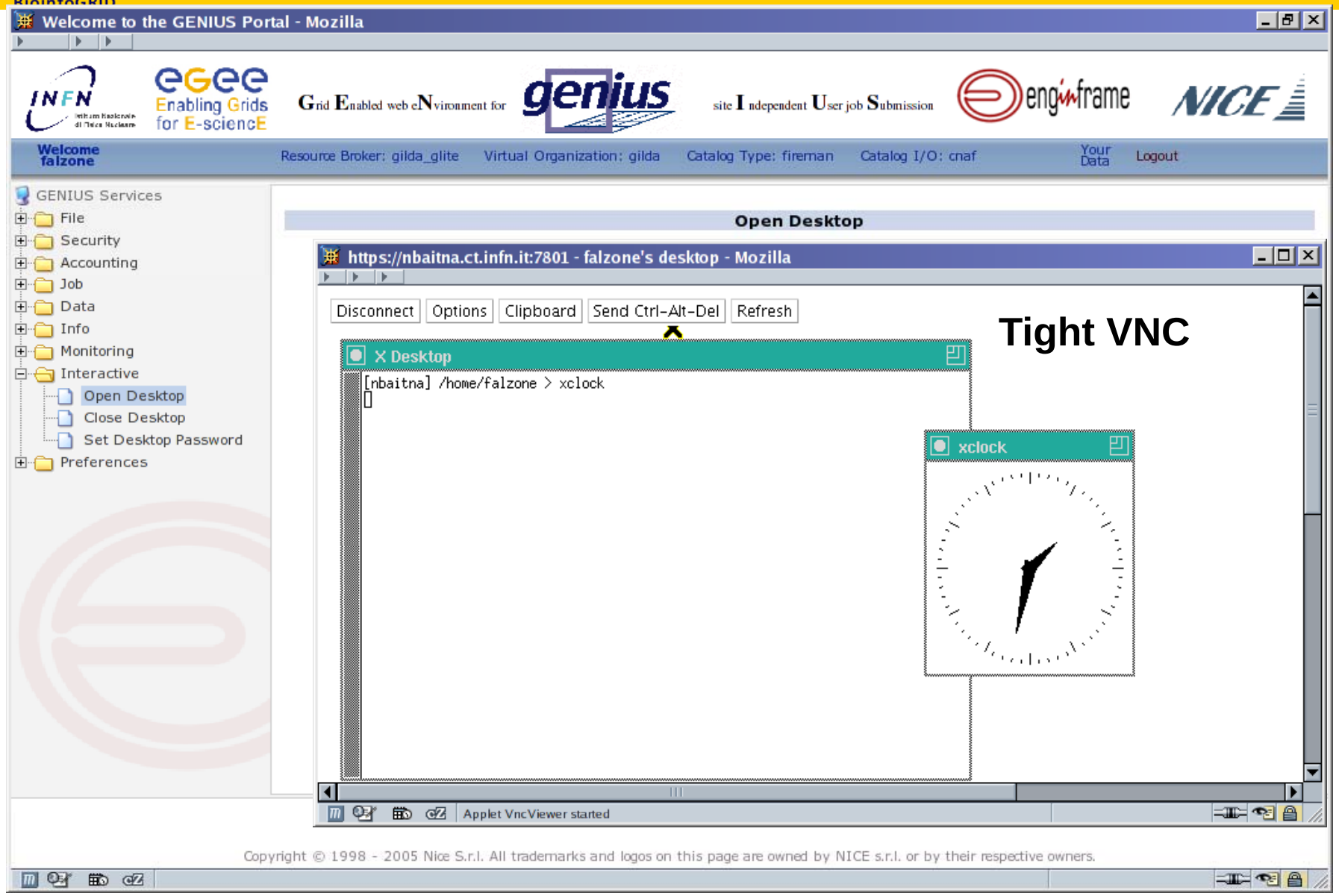
GENIUS: Get Output



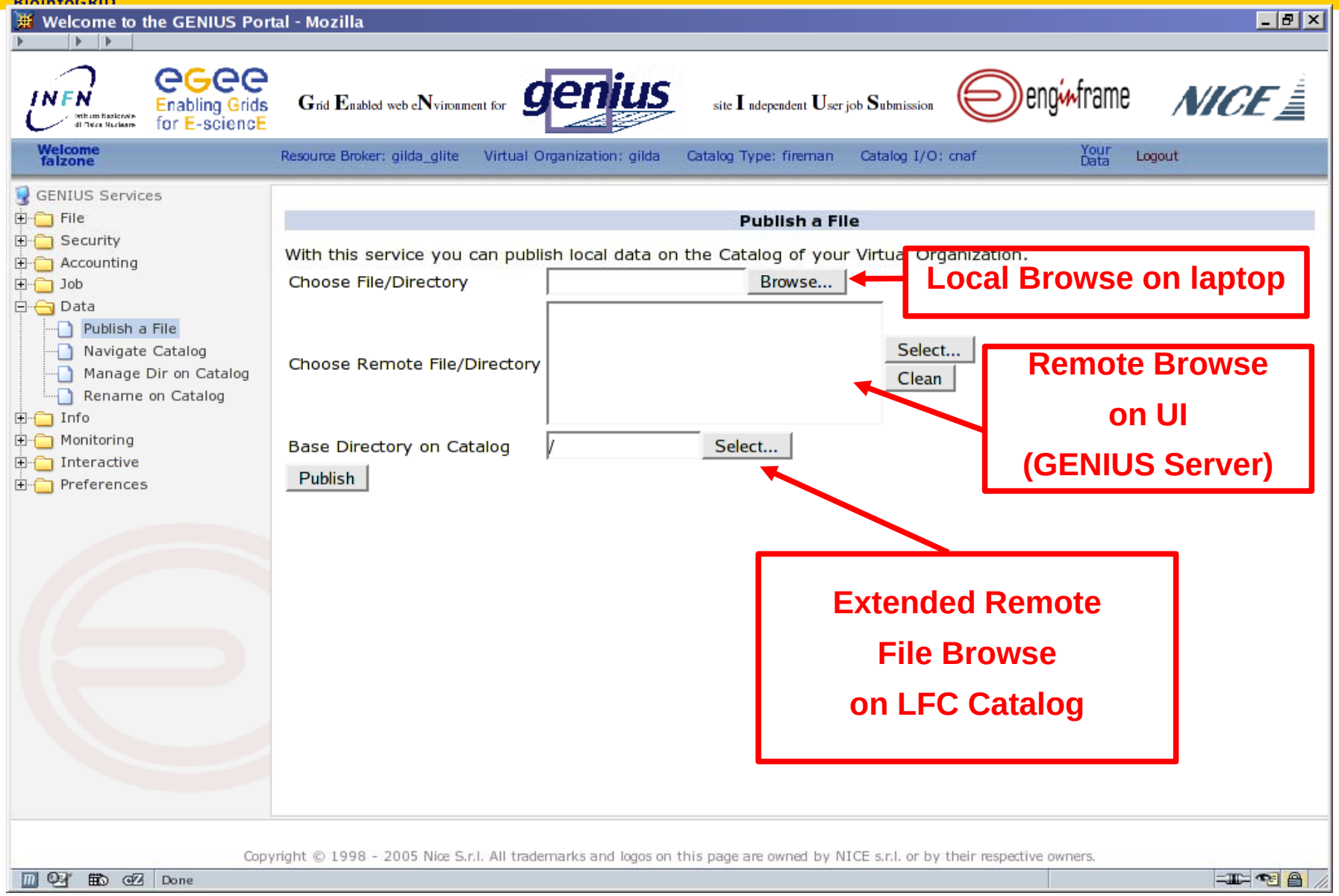
GENIUS: Job Queue - Spooler



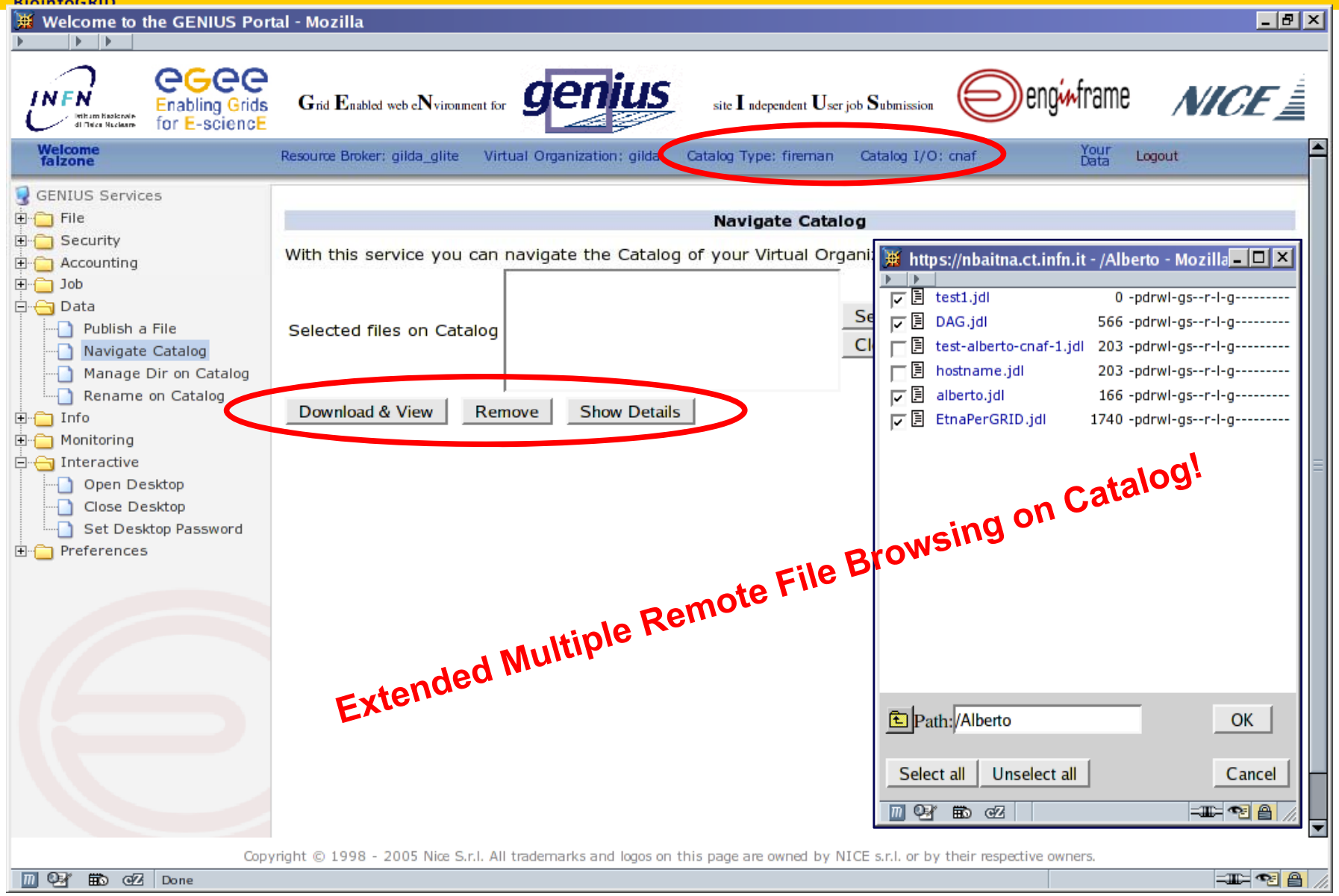
GENIUS: Interactive



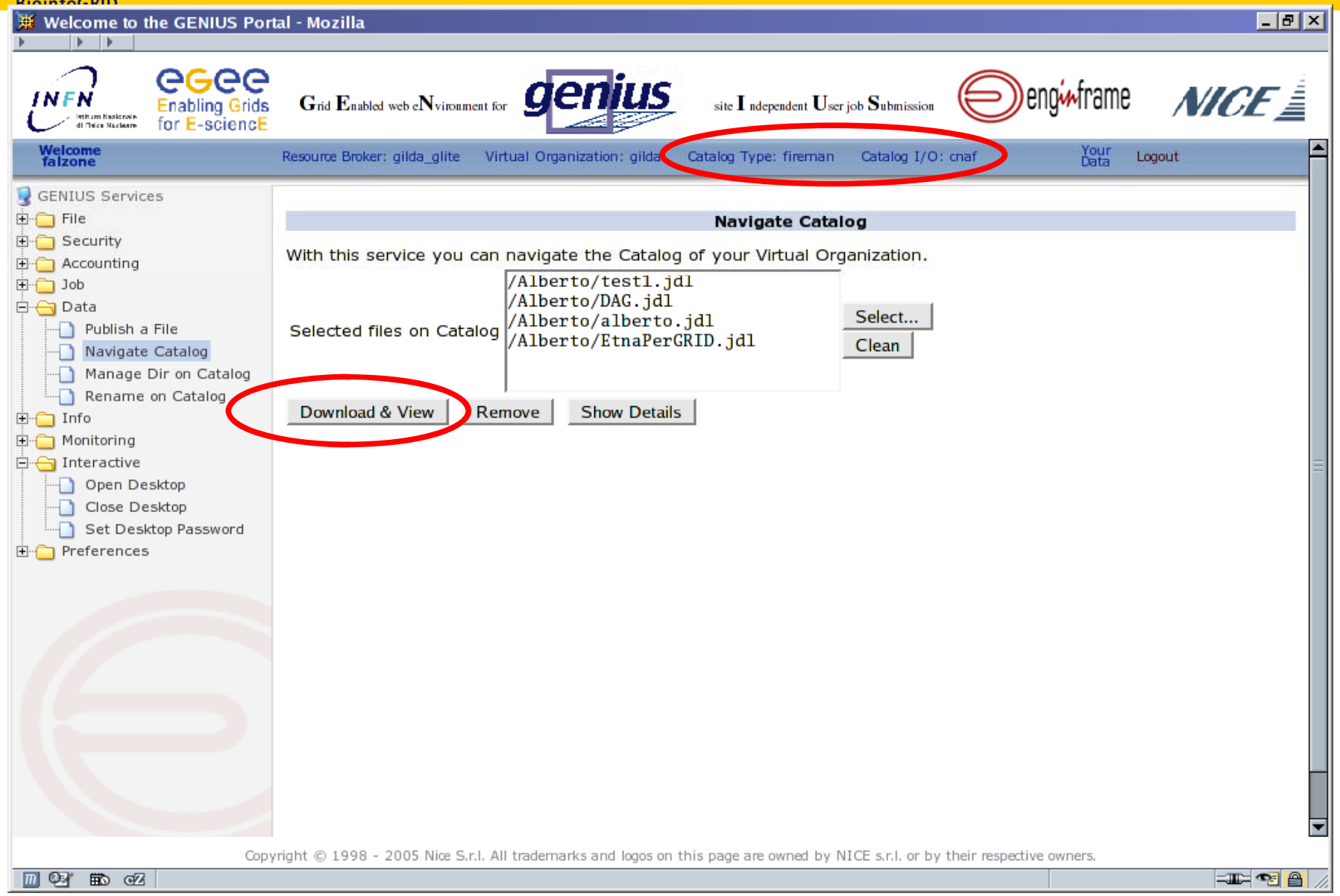
GENIUS: Data & Grid Catalogs



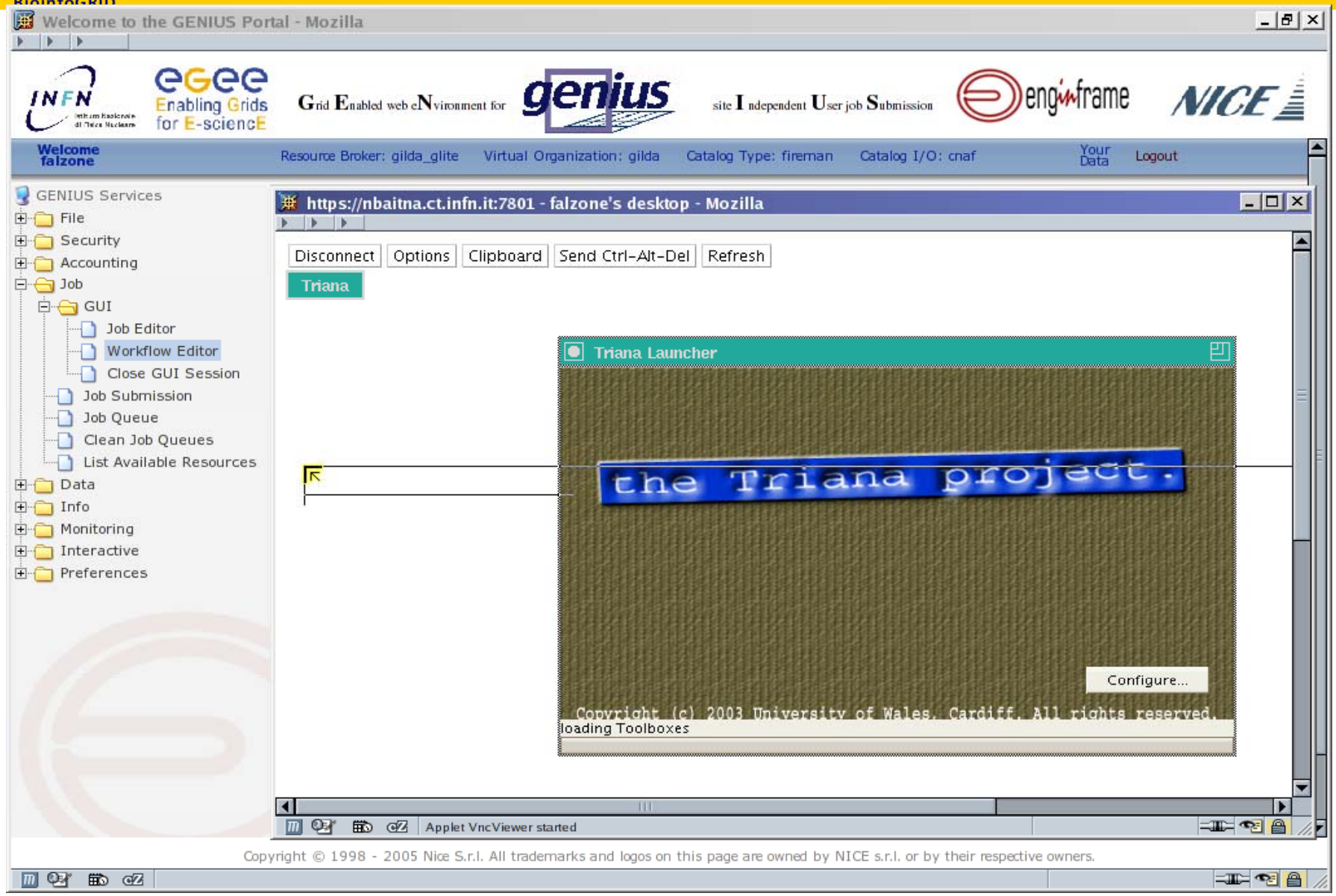
GENIUS: Data & Grid Catalogs



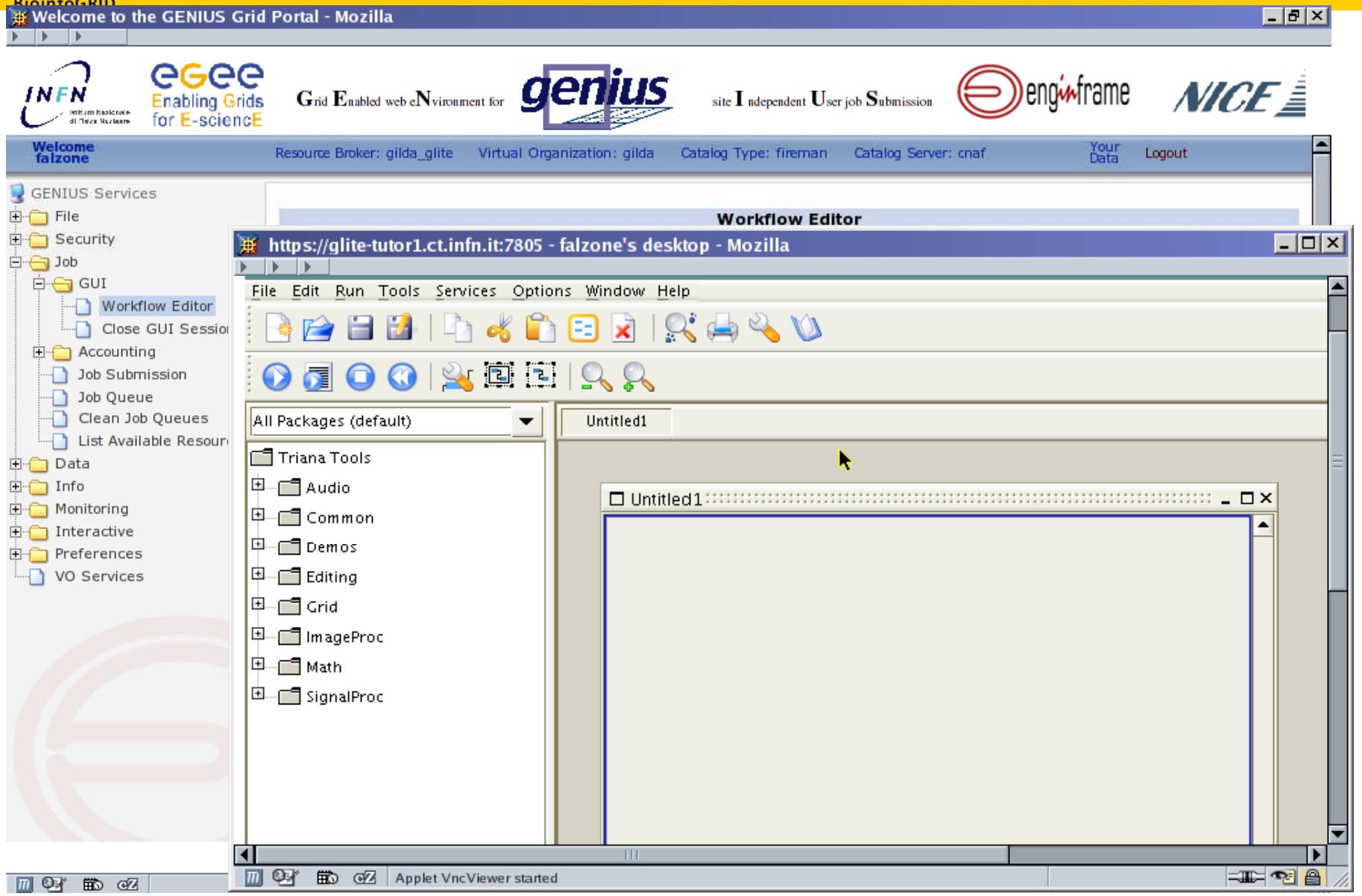
GENIUS: Data & Grid Catalogs



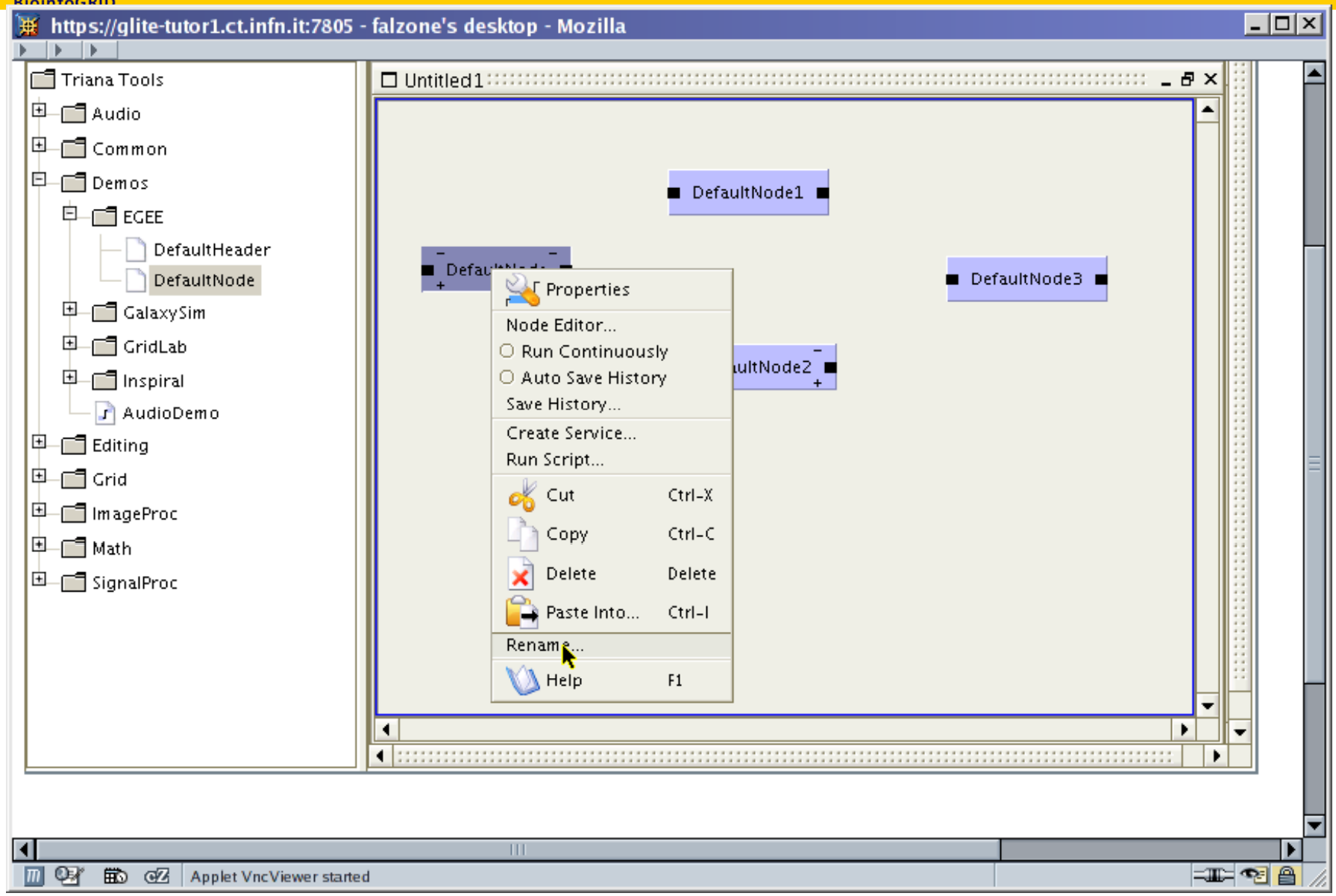
GENIUS & TRIANA /1



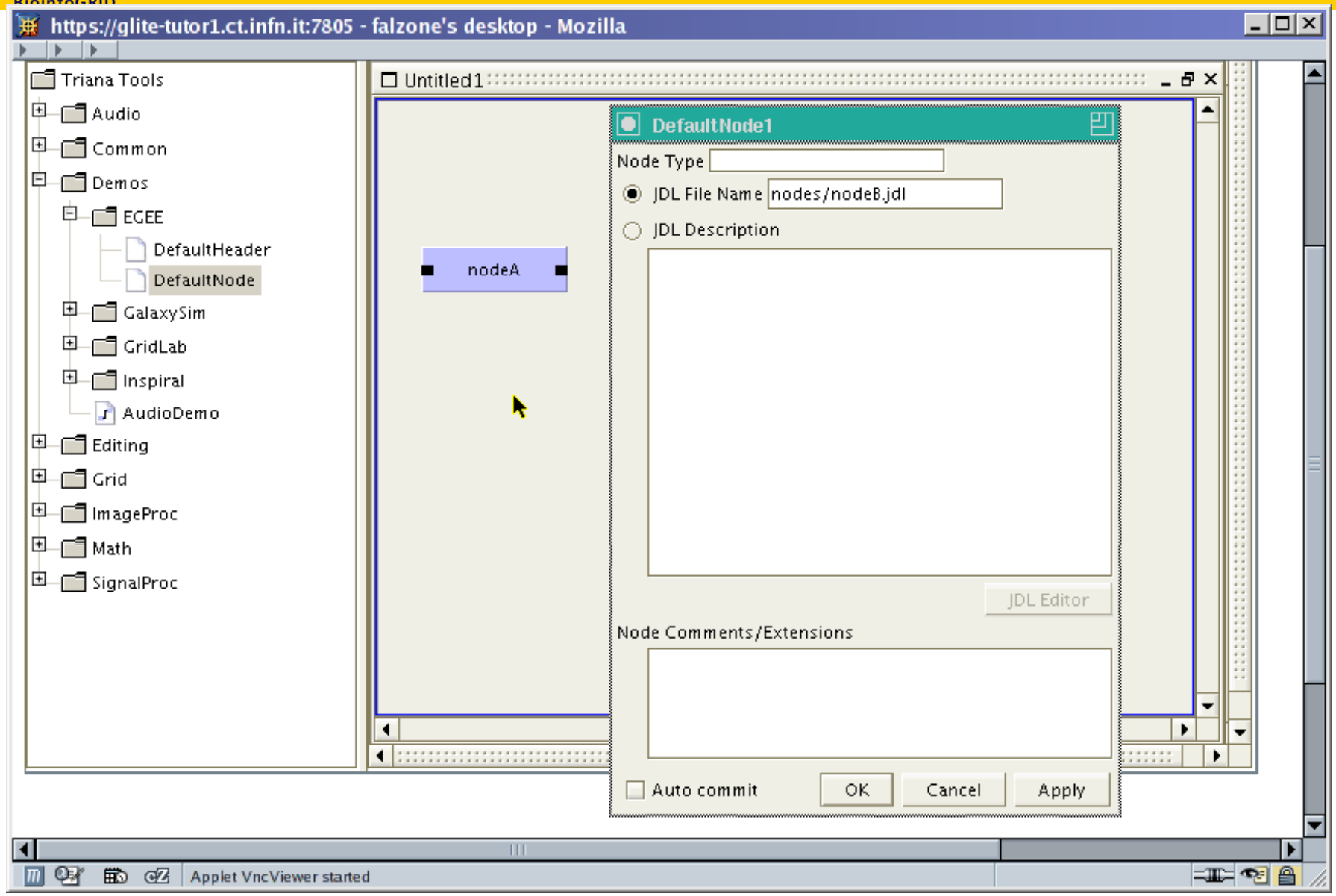
GENIUS & TRIANA /2



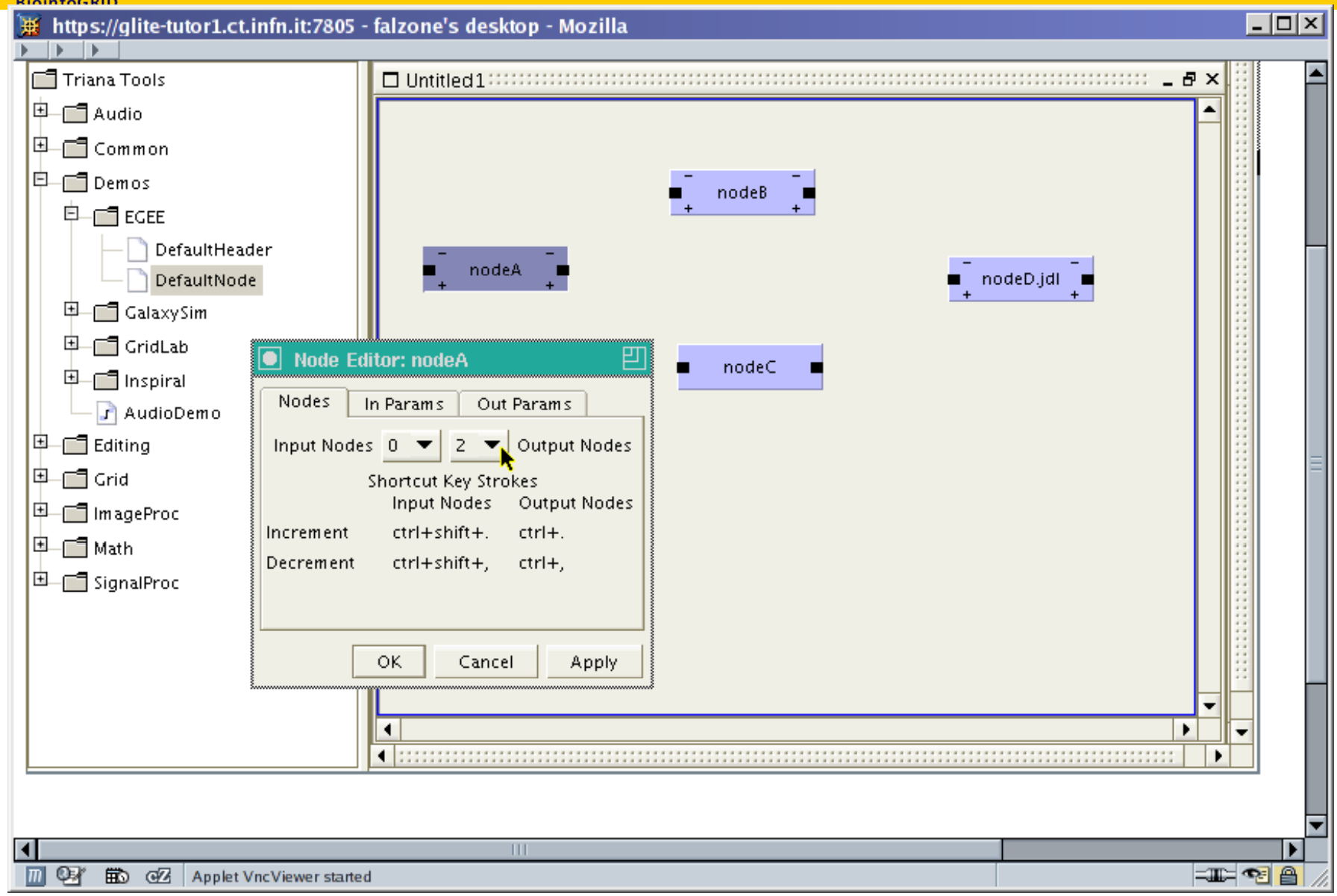
GENIUS & TRIANA /3



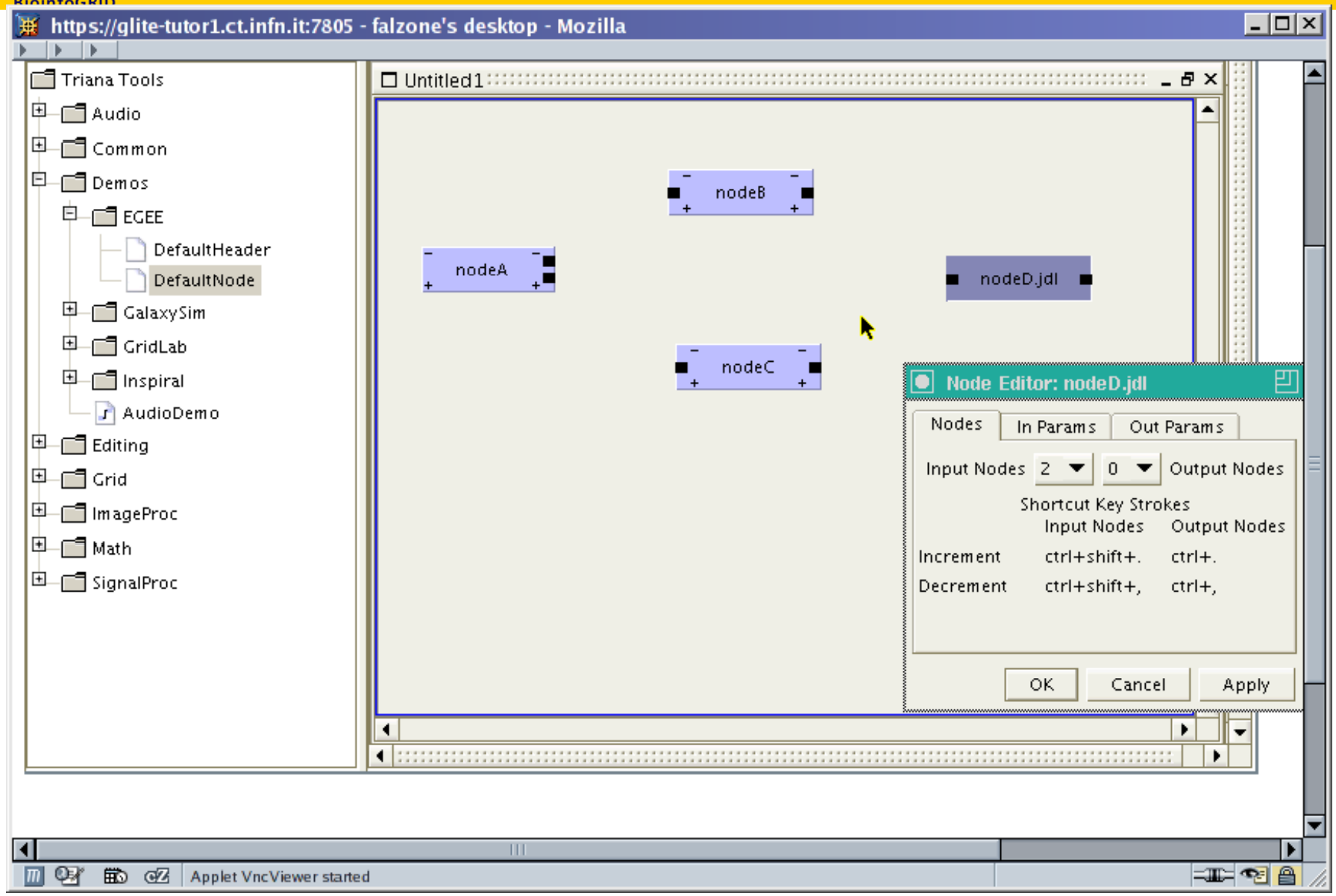
GENIUS & TRIANA /4



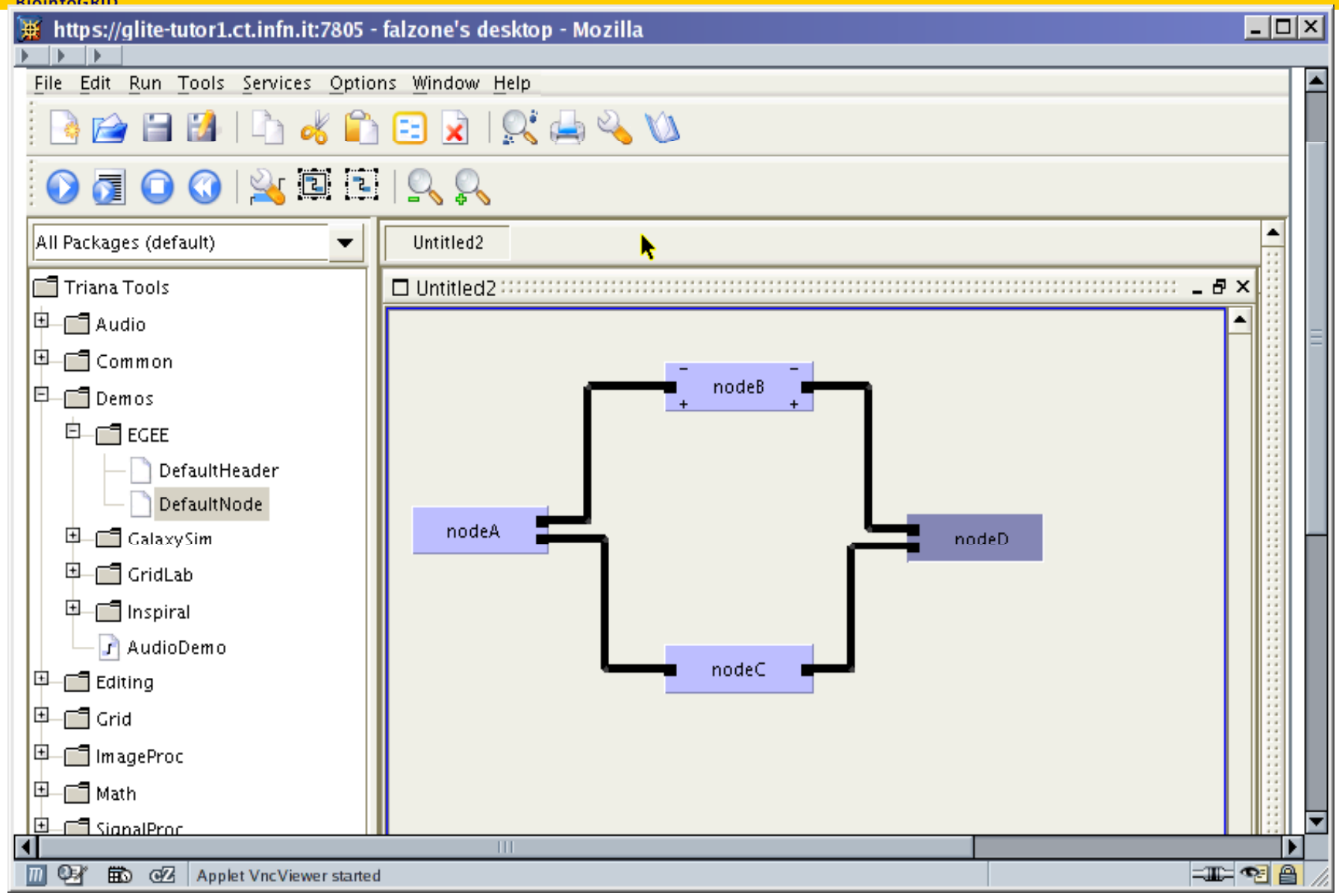
GENIUS & TRIANA /5



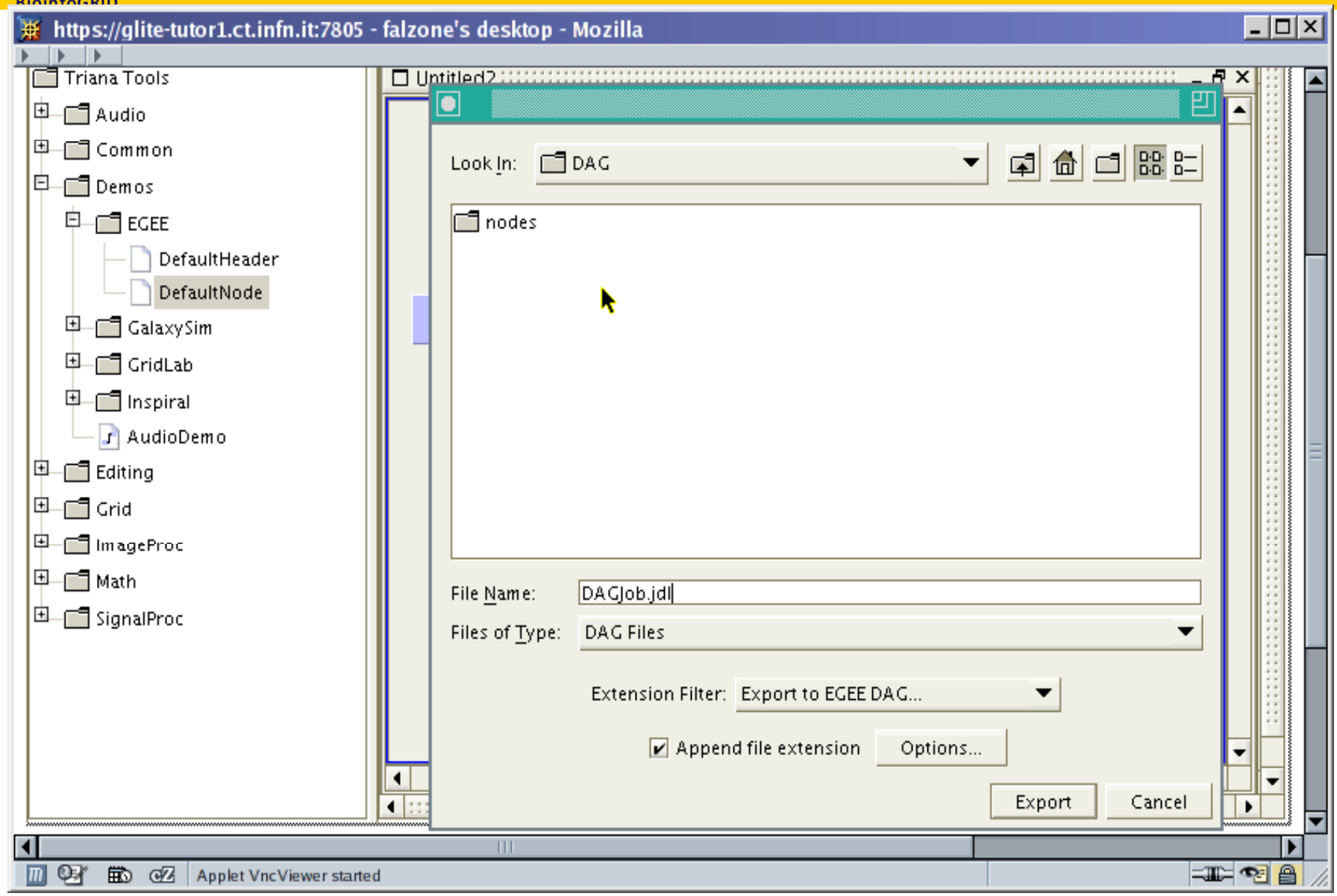
GENIUS & TRIANA /6



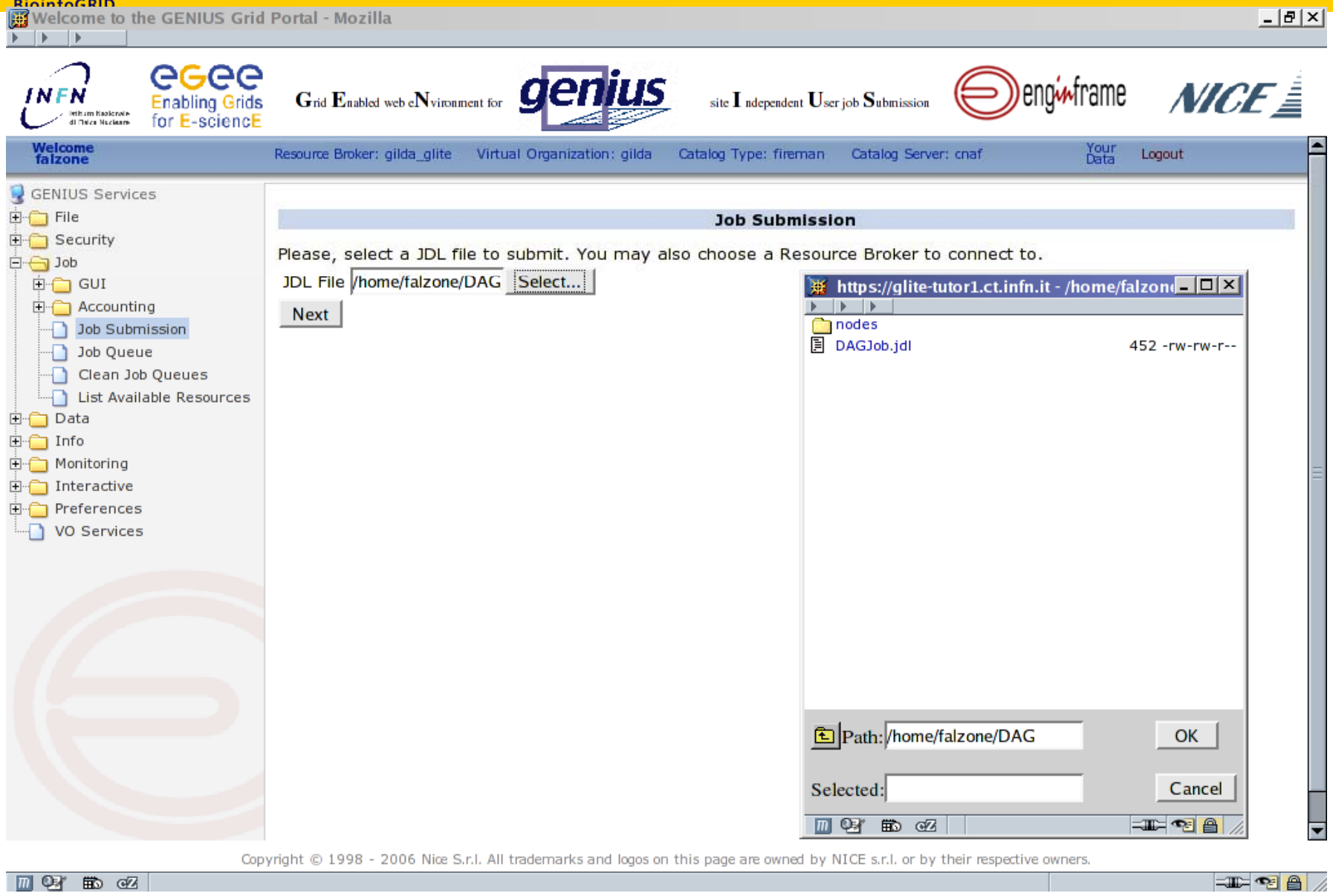
GENIUS & TRIANA /7



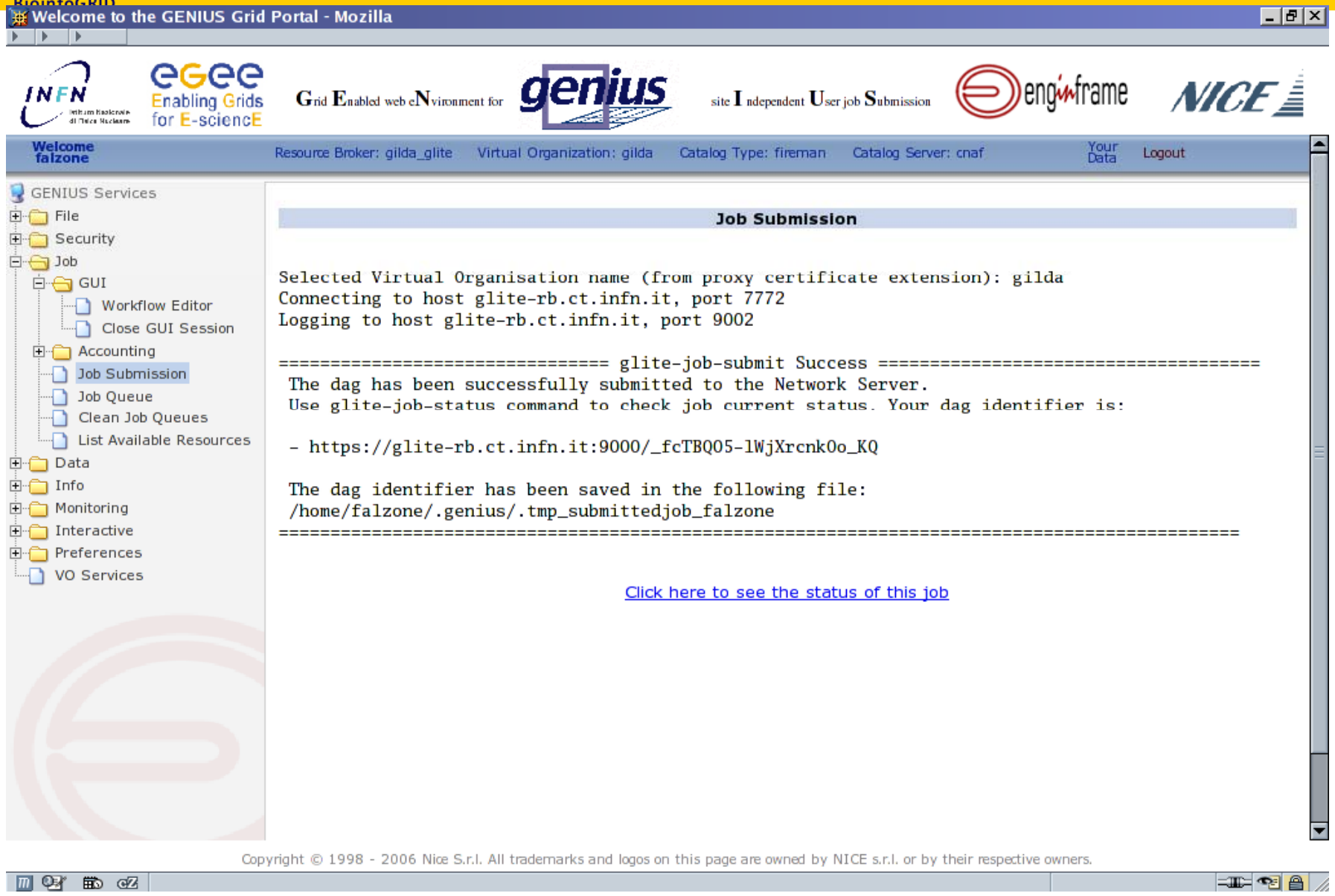
GENIUS & TRIANA /8



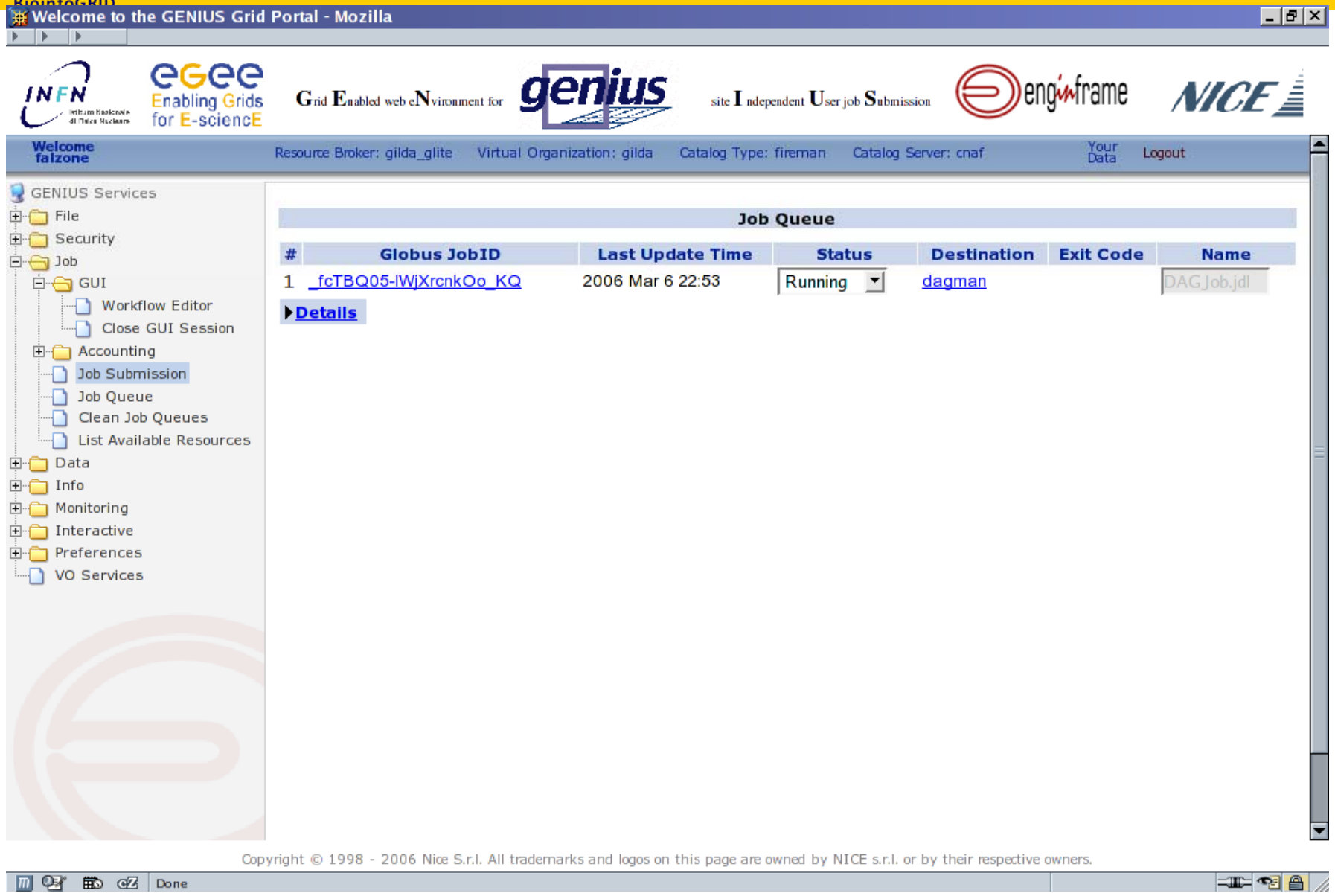
GENIUS & TRIANA /9



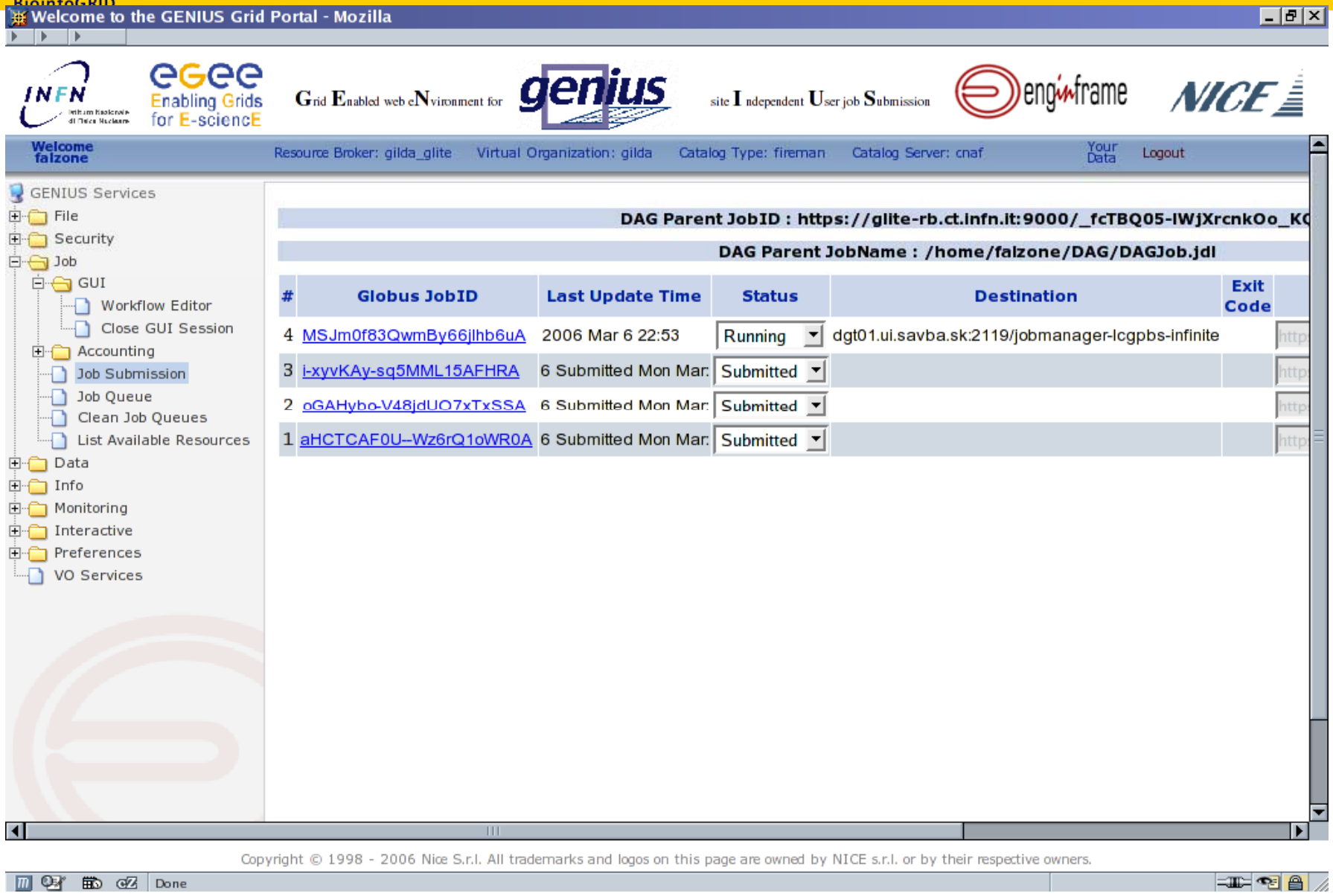
GENIUS & TRIANA /10



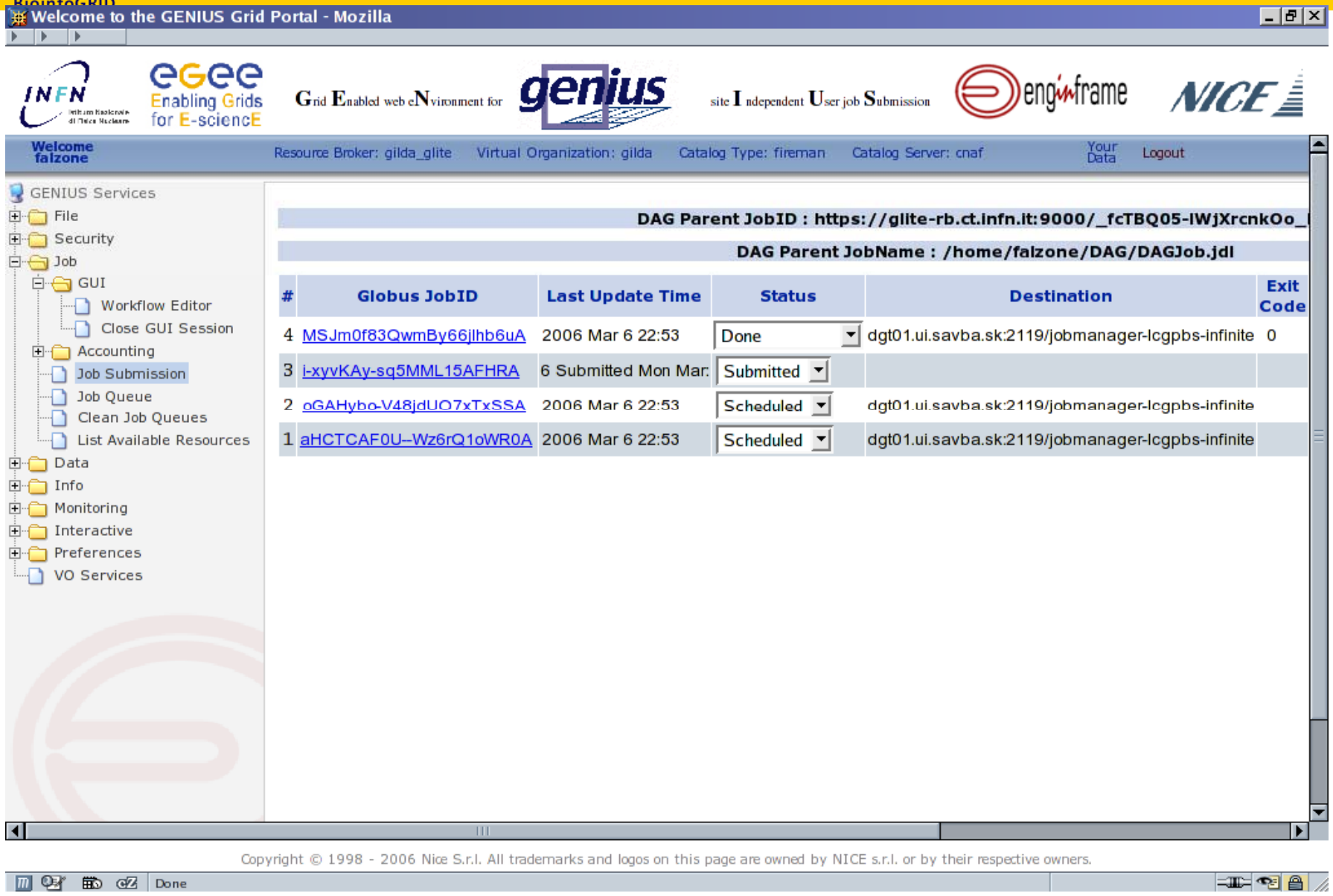
GENIUS & TRIANA /11



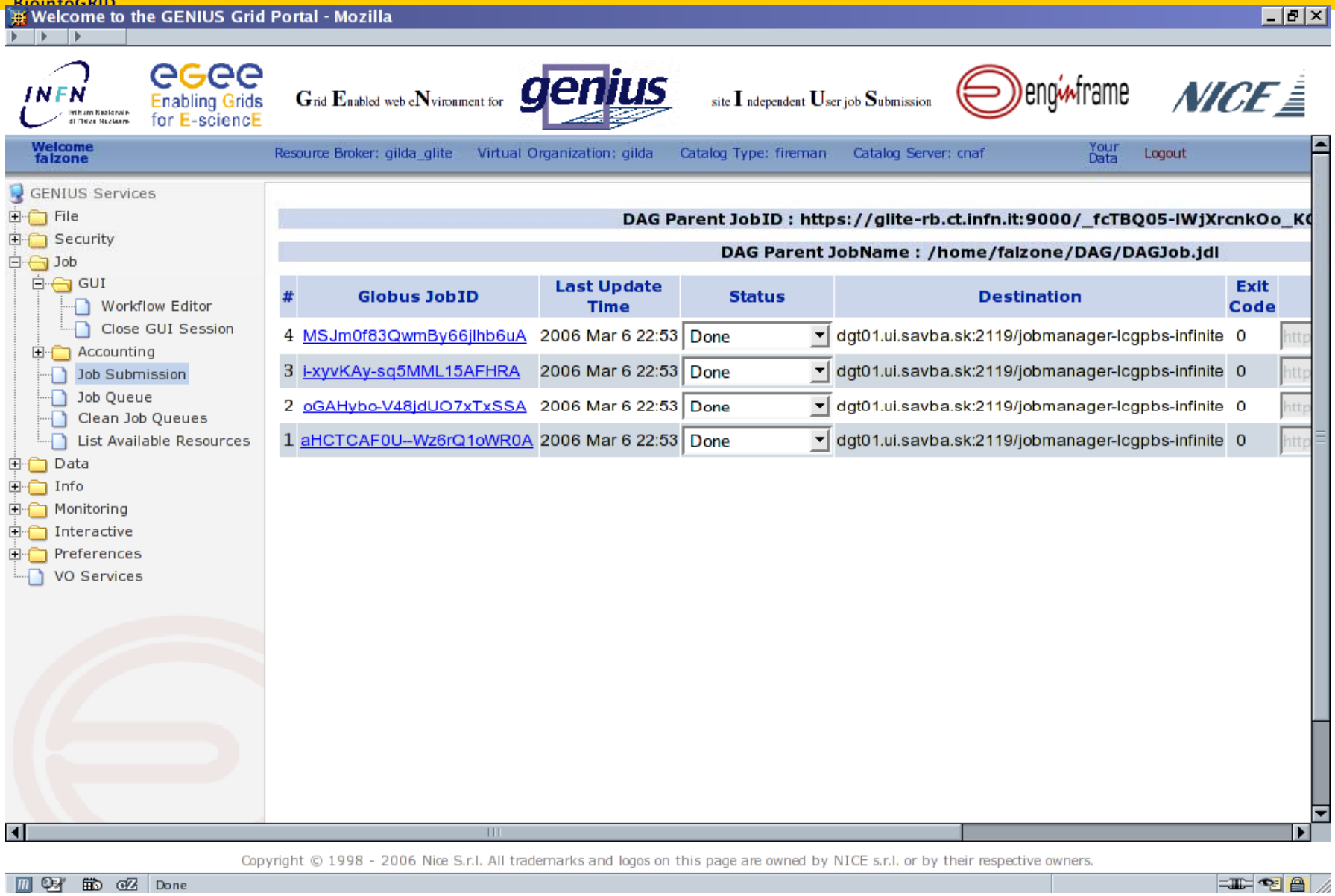
GENIUS & TRIANA /12



GENIUS & TRIANA /13



GENIUS & TRIANA /14

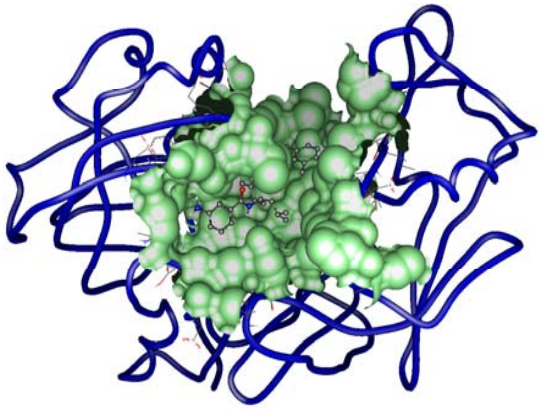


GENIUS & TRIANA /15

- Grid Portals technology
- EnginFrame Framework
- The GENIUS Grid Portal
- Some Integrated Applications
- References

Presented at CAPI 2006 Biocomputing - Milan

B L A S T



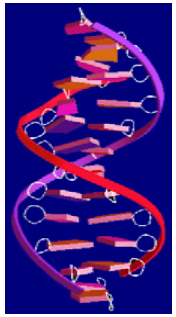
This is a work supported in part by the LIBI and the BIOINFOGRID project.

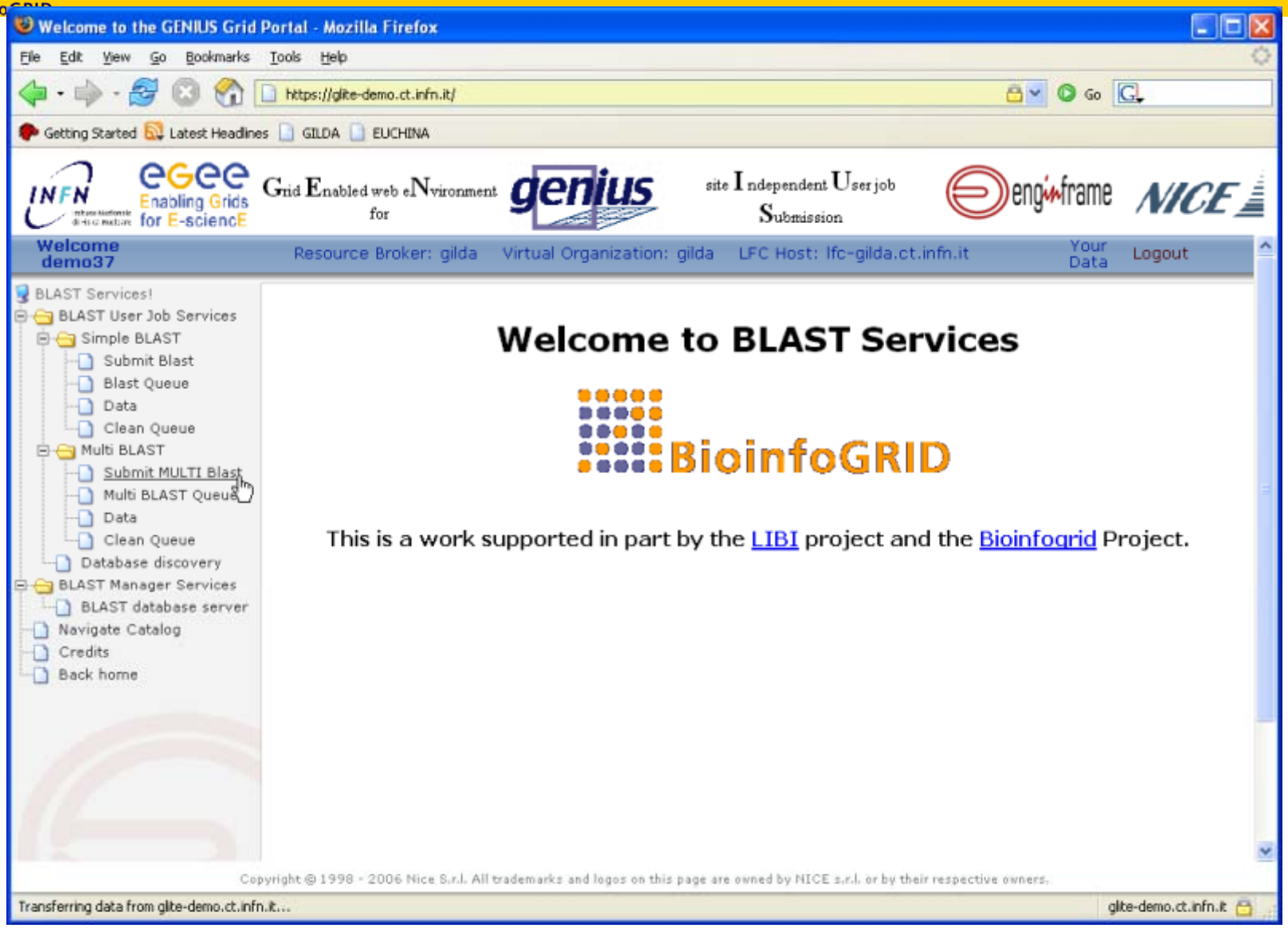
What is BLAST ?

BLAST (Basic Local Alignment Search Tool) is an algorithm used to compare biological sequences, such as the amino-acid of different proteins or the nucleotides of DNA sequences.

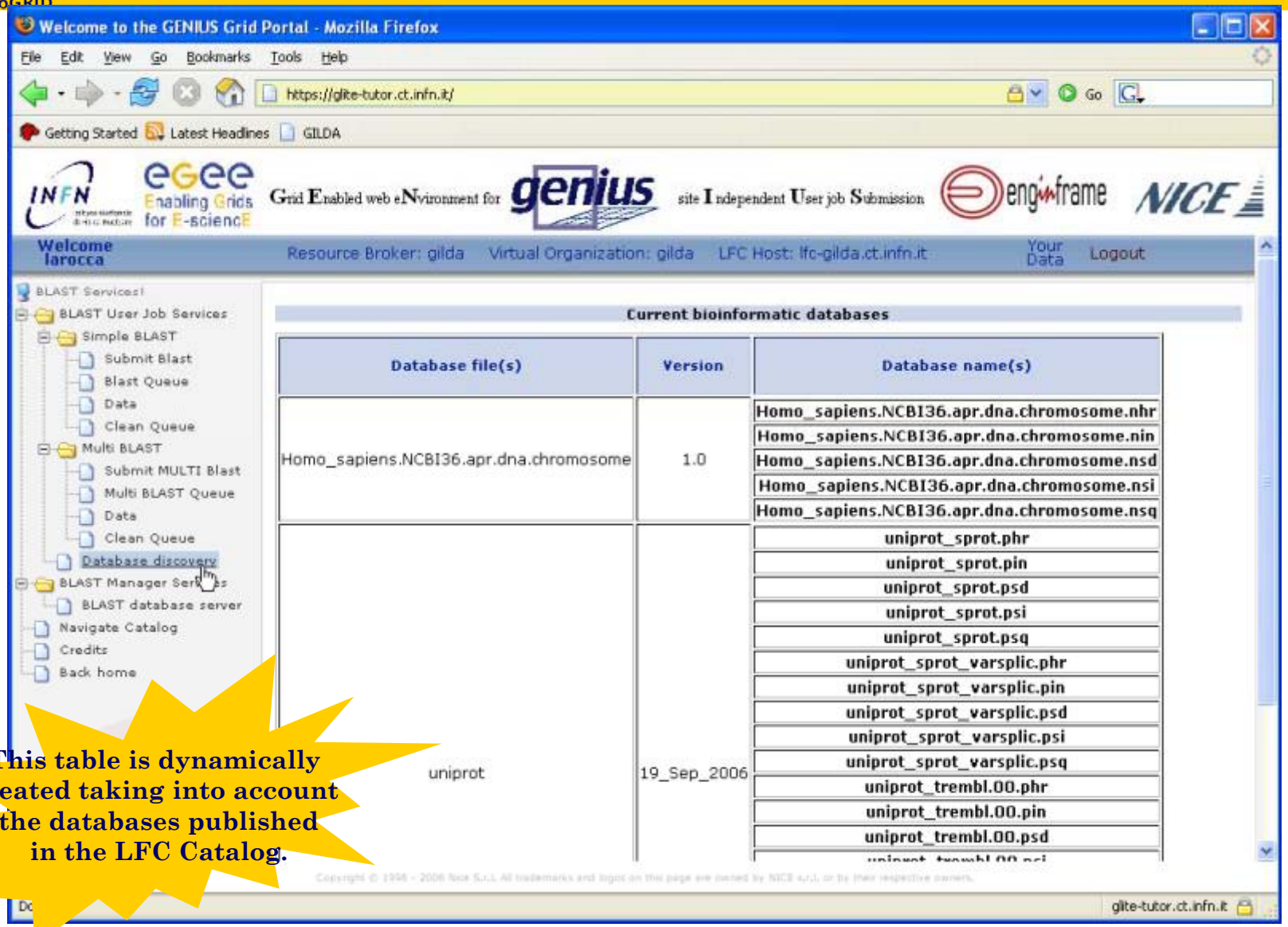
The program requires a query sequence (also called *target sequence*) and a database.

Its identifies all the sub-sequences of the target query that are similar to the sub-sequences in the database.



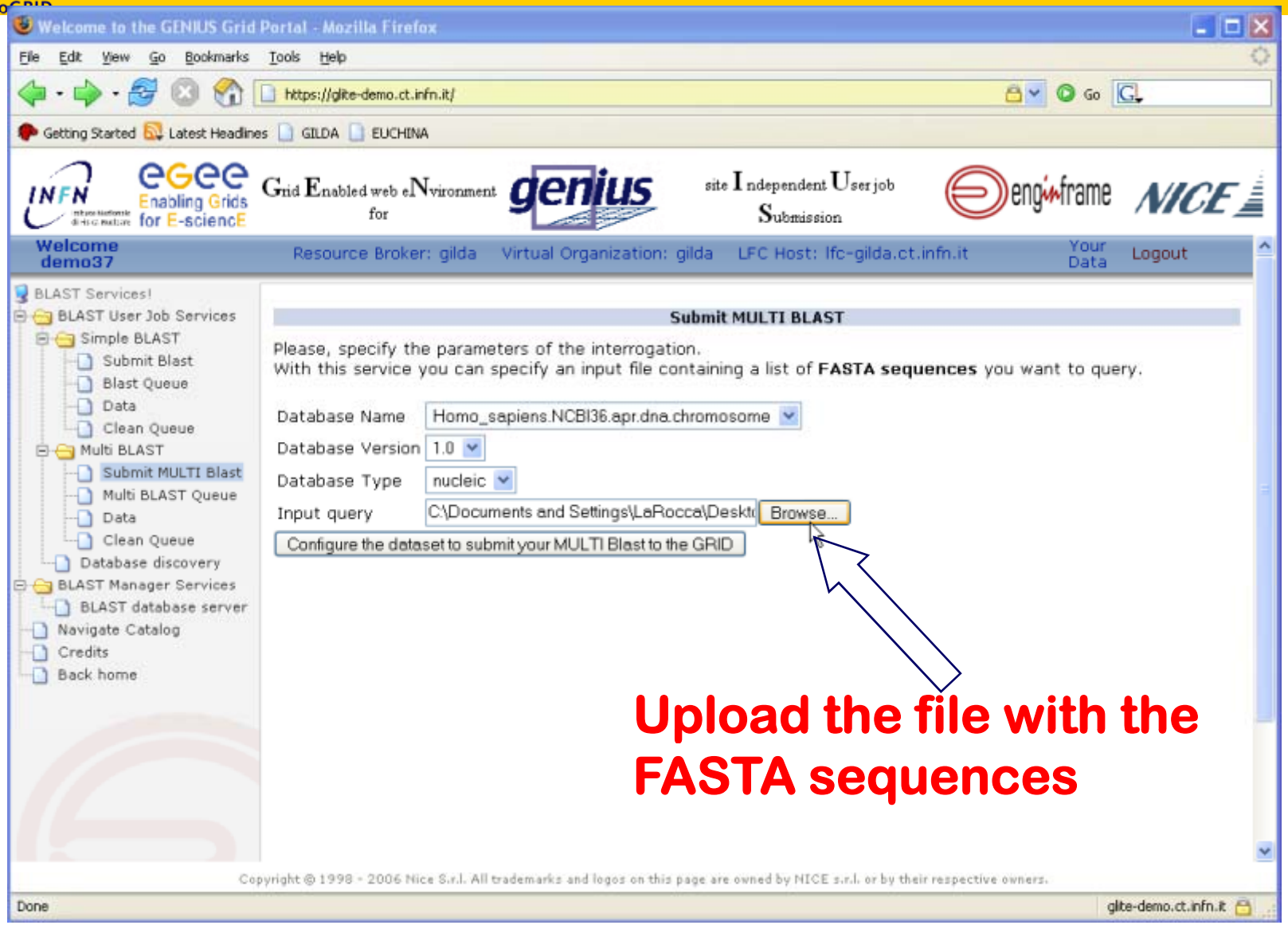

Welcome to the GENIUS Grid Portal - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
https://glite-demo.ct.infn.it/
Getting Started Latest Headlines GILDA EUCHINA
INFN
eGee Enabling Grids for E-sciencE
Grid Enabled web eNvironment for
genius
site Independent User job Submission
enginframe
NICE
Welcome demo37
Resource Broker: gilda
Virtual Organization: gilda
LFC Host: lfc-gilda.ct.infn.it
Your Data
Logout
BLAST Services!
BLAST User Job Services
Simple BLAST
Submit Blast
Blast Queue
Data
Clean Queue
Multi BLAST
Submit MULTI Blast
Multi BLAST Queue
Data
Clean Queue
Database discovery
BLAST Manager Services
BLAST database server
Navigate Catalog
Credits
Back home
Welcome to BLAST Services
BioinfoGRID
This is a work supported in part by the LIBI project and the Bioinfogrid Project.
Copyright © 1998 - 2006 Nice S.r.l. All trademarks and logos on this page are owned by NICE s.r.l. or by their respective owners.
Transferring data from glite-demo.ct.infn.it...
glite-demo.ct.infn.it

Bioinformatics databases

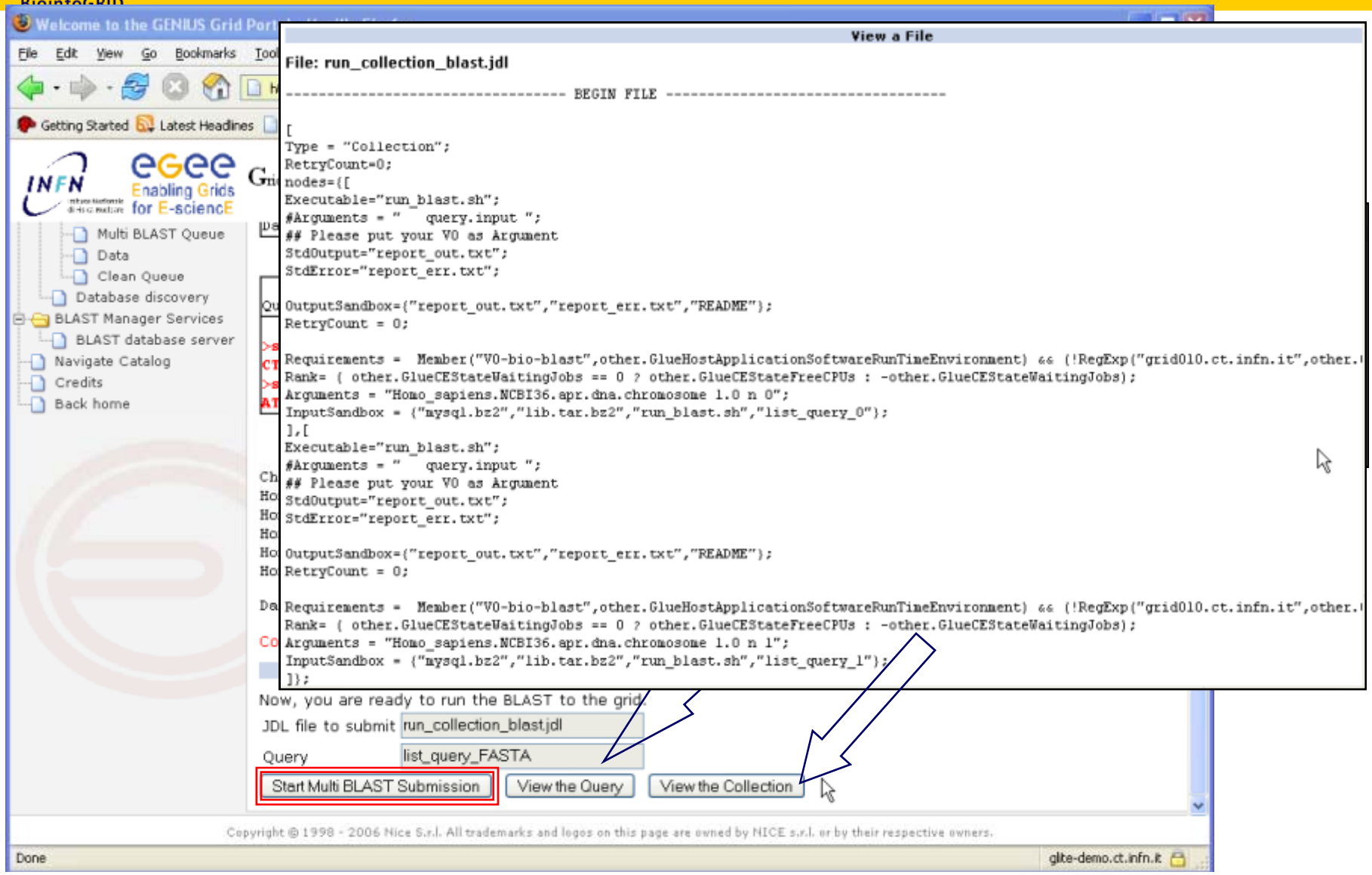


This table is dynamically created taking into account the databases published in the LFC Catalog.

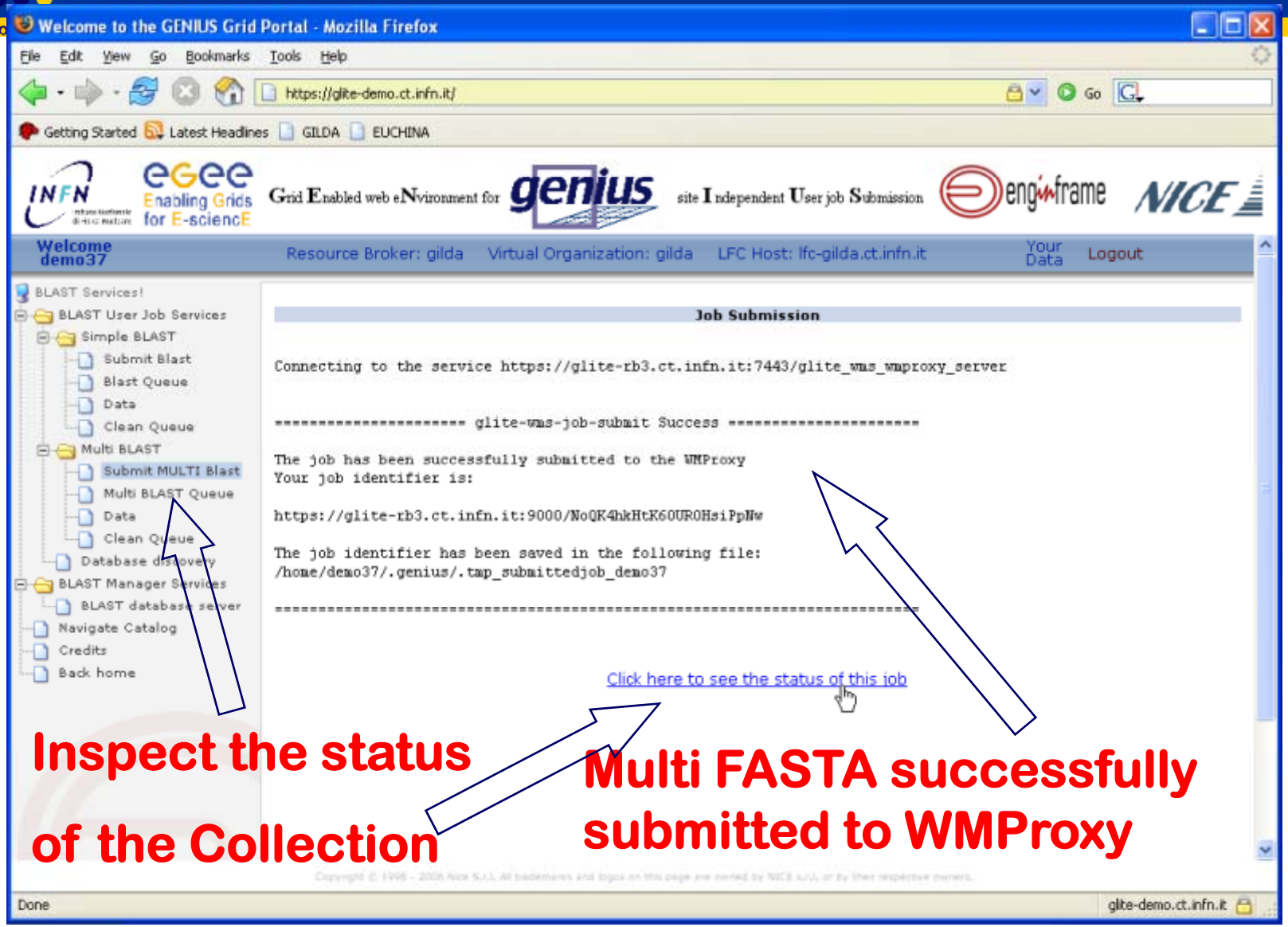
MULTI BLAST : input form /1



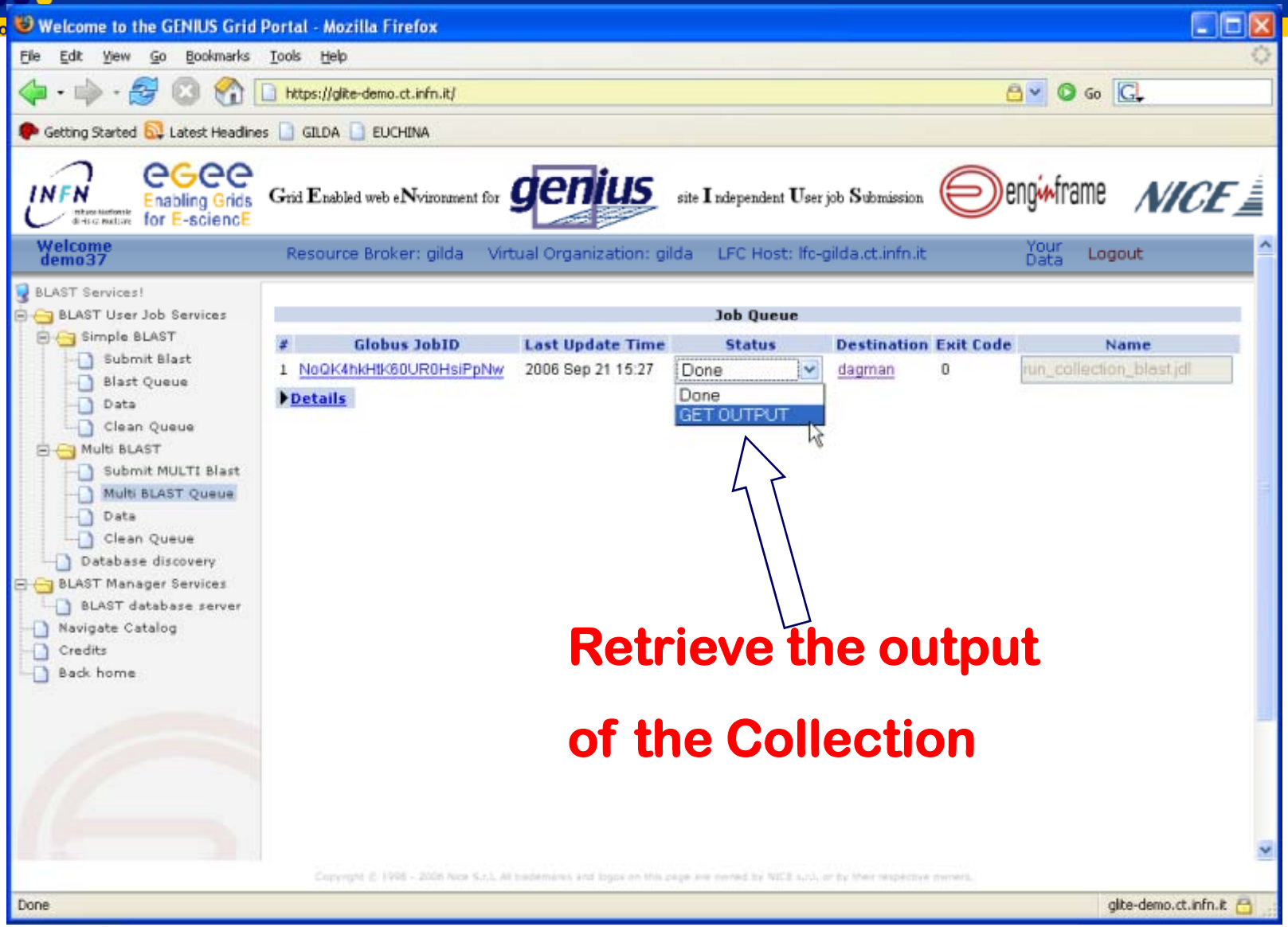
MULTI BLAST : input form /2



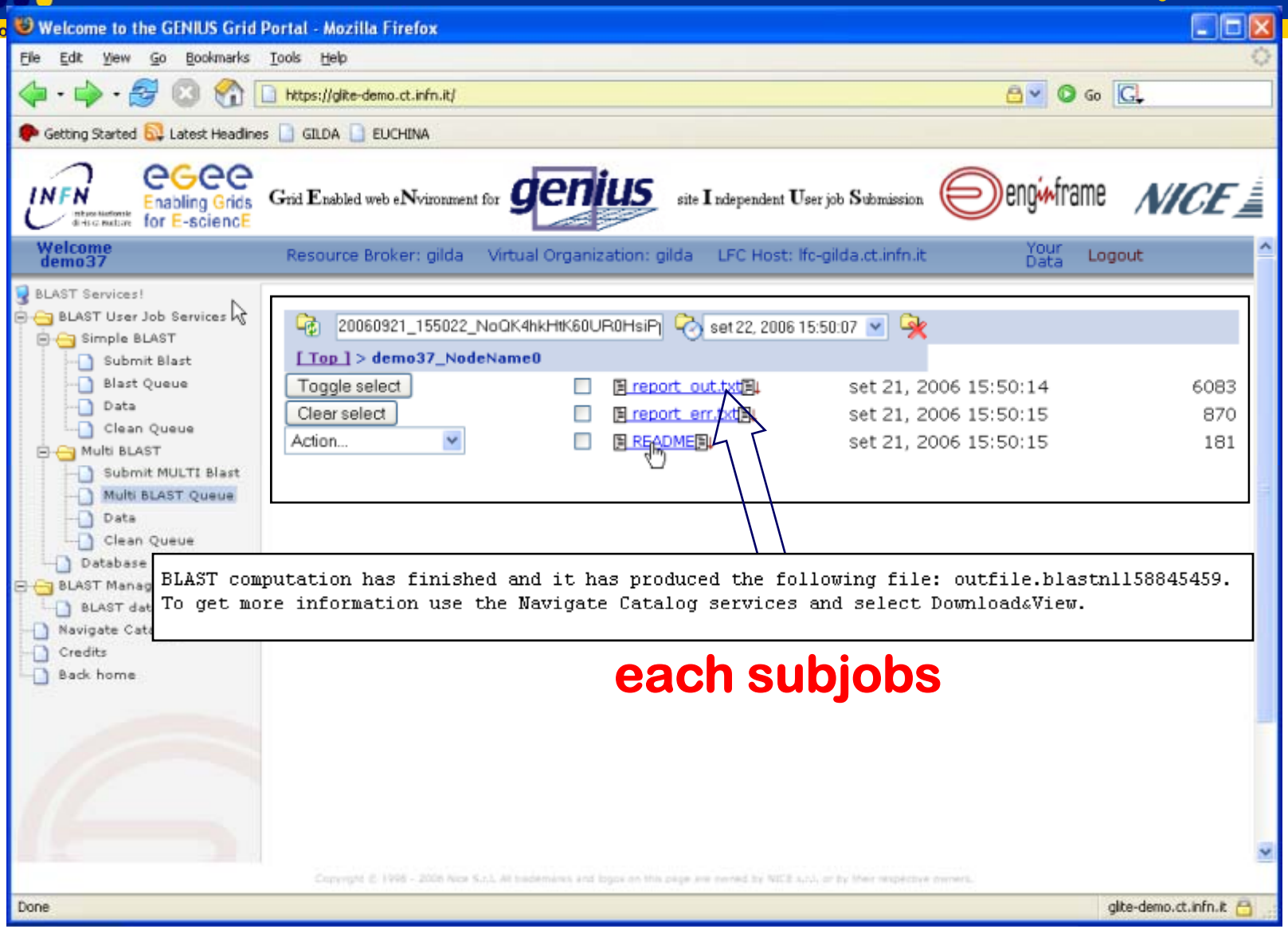
MULTI BLAST : submit



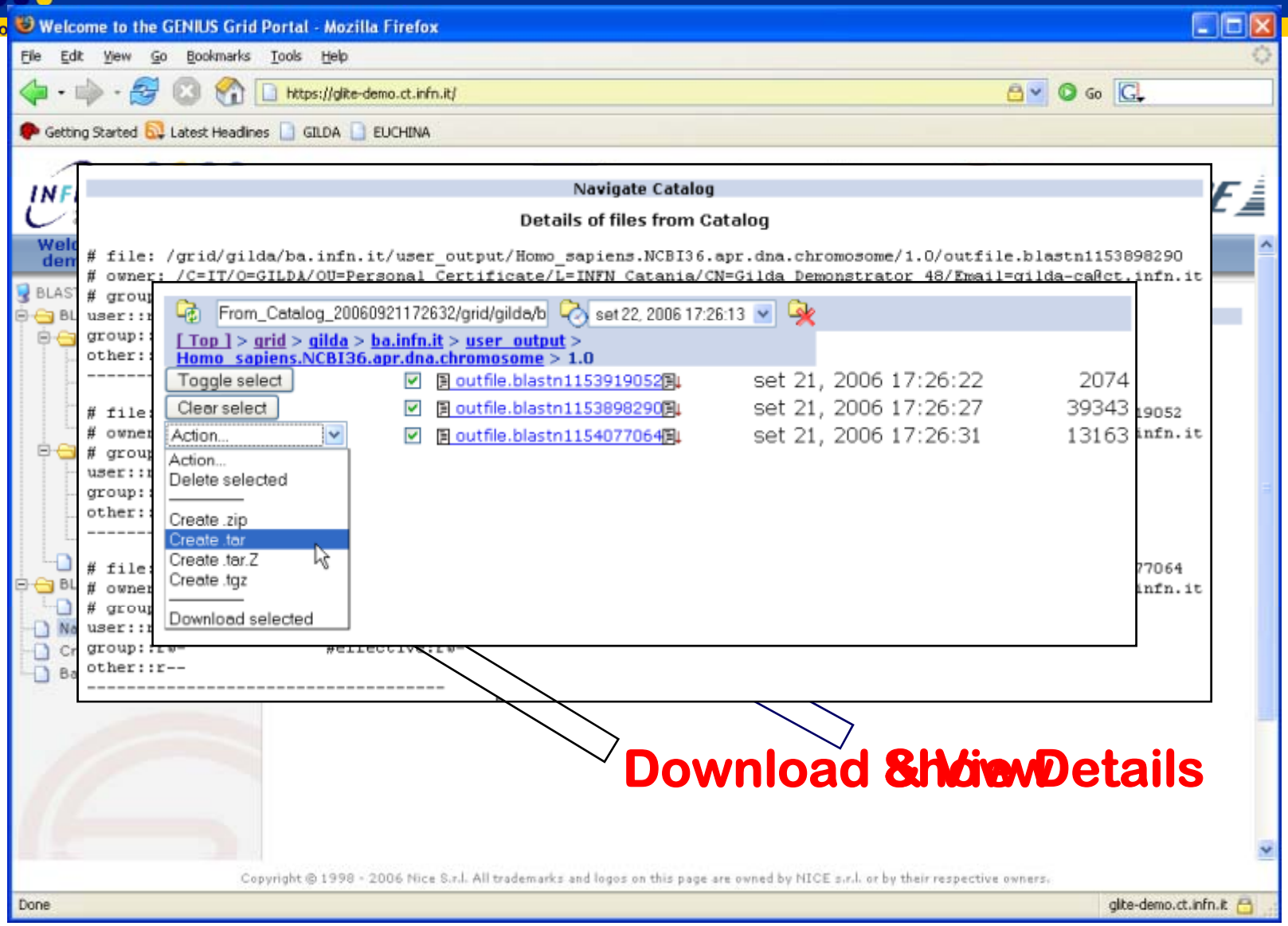
MULTI BLAST : queue



MULTI BLAST : Data Spooler



MULTI BLAST : LFC Catalog



The results

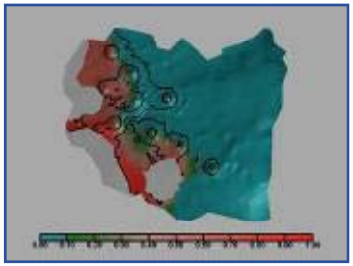
Click here to inspect a typical output file produced by BLAST during the computation

References

- "Muli BLAST on Demonstrator: an easy way to run BLAST in a GRID Infrastructure"
  - AIFTIMIEI, C. – CAROTA, L. – DE FILIPPIS, N. – DONVITO, G. – FALZONE, A – PIERRO, A. – TULIPANO, A, – LA ROCCA, G. – VERLATO, M. – Prof. BARBERA, R. – Prof. MAGGI, G.P. CAPI2006 BIOCOMPUTING – Milan 2006.

Presented at First EUMedGrid Project Review Madrid, 27-28 Feb. - SPAIN

CODESA–3D



This is a work supported in collaboration with

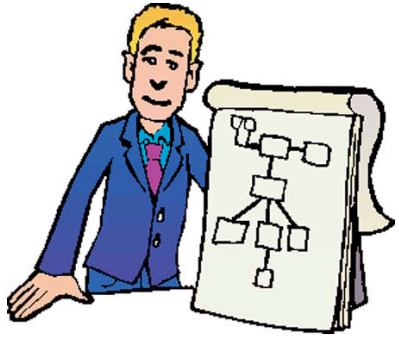
Giuditta Lecca[1], Jawher Kerrou[2], Philippe Renard[3]
Giuseppe La Rocca[4] and Roberto Barbera[4]

1 CRS4, Cagliari (Italy)
2 INAT, Tunis (Tunisia)
3 CHYN, University of Neuchatel (Switzerland)
4 INFN, Catania (Italy)

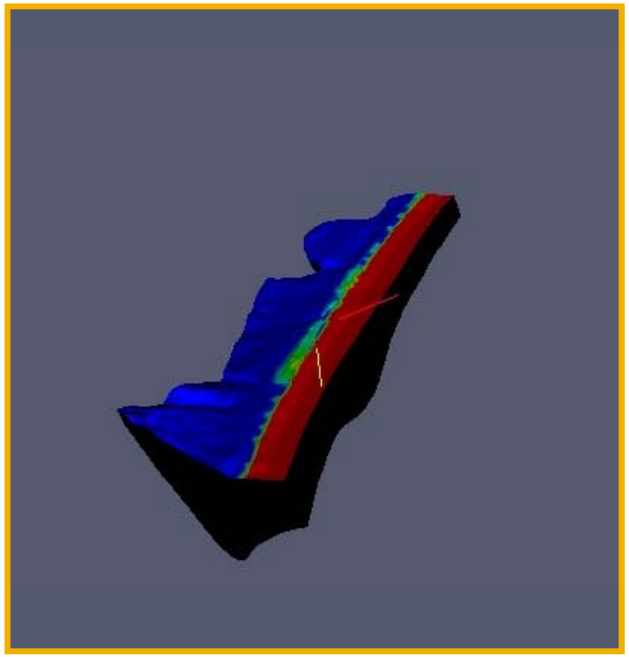
The CODESA-3D application

- **CODESA-3D** ***(COupled variable DEnsity and Saturation 3-Dimensional)*** **is a software developed at CRS4 (Center for Advanced Studies, Research and Development in Sardinia - http://www.crs4.it/).**

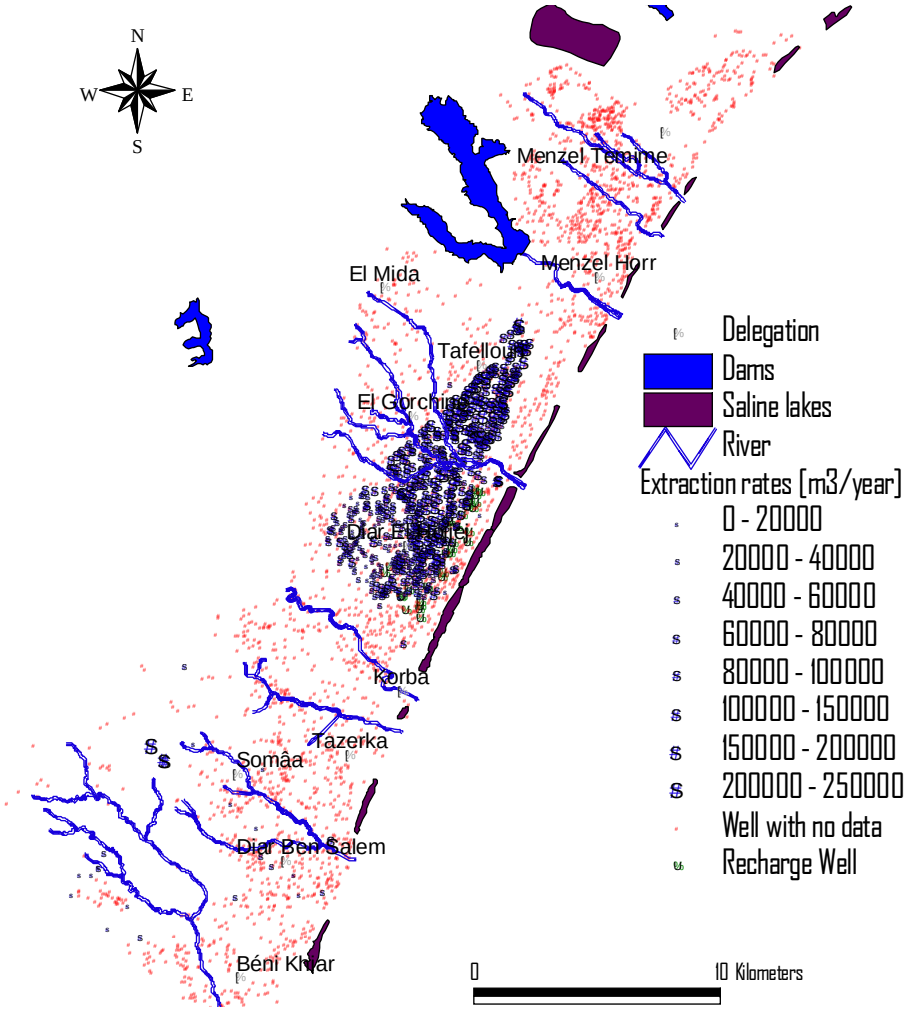
- **Its aims to study, taking into account detailed mathematical representations of physical, chemical and biological processes that govern water flow, saltwater intrusion in freshwater coastal aquifers.**



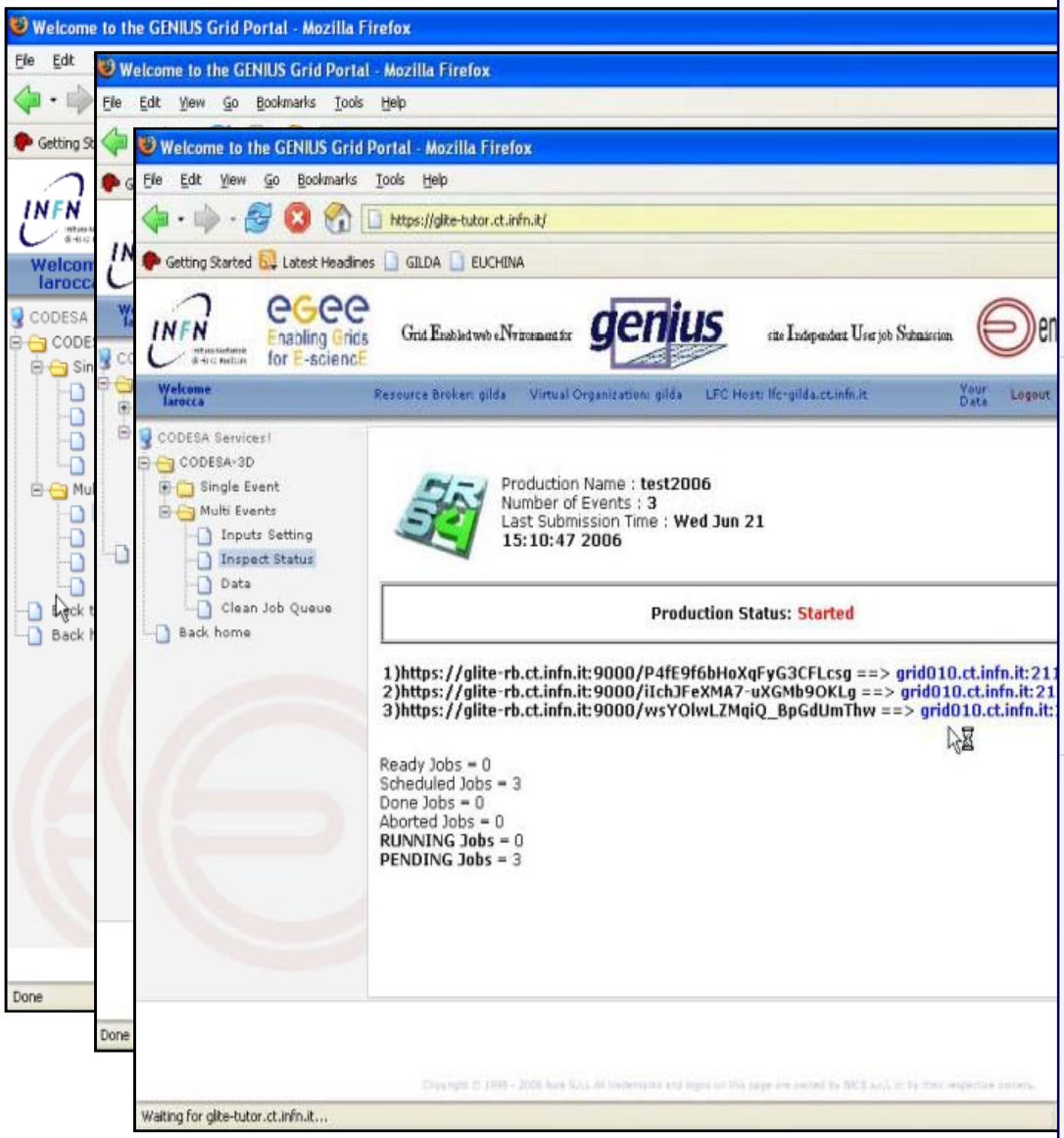
The Korba case-study (Tunisia)

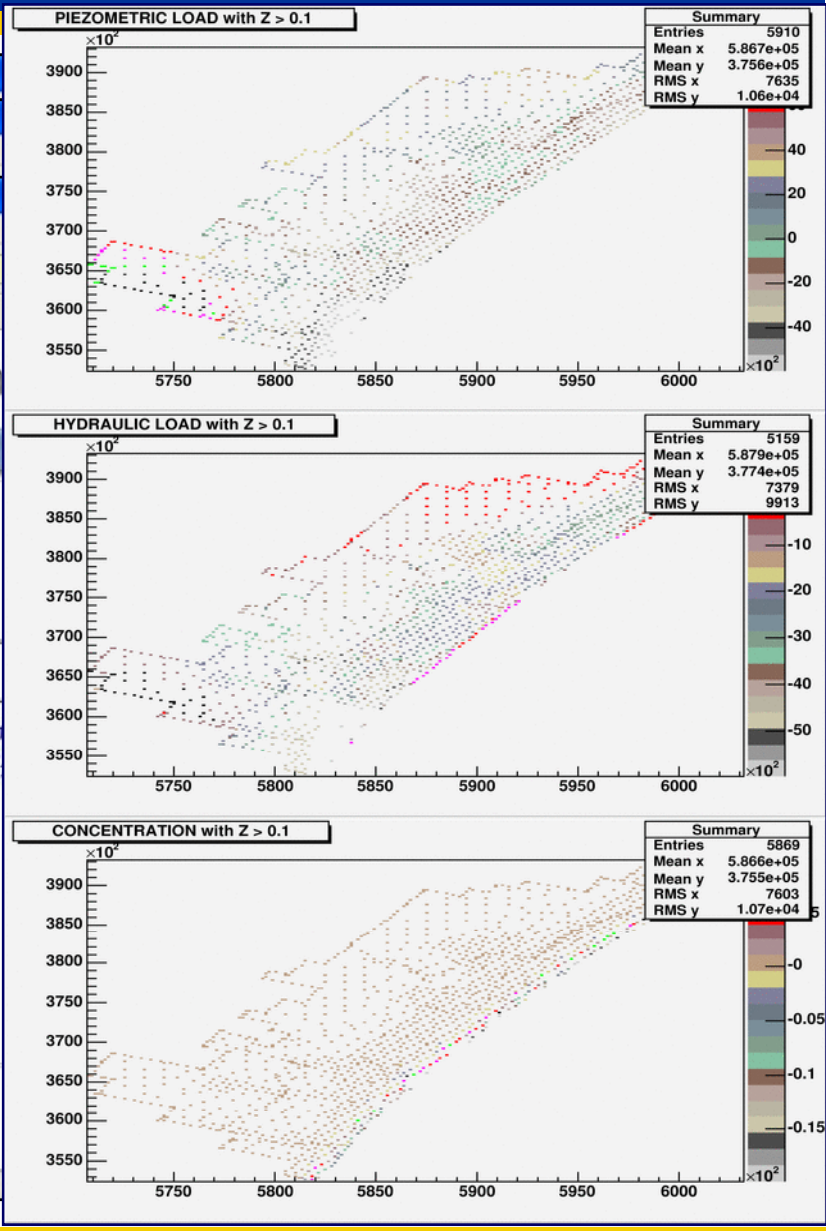


Extraction rates and Recharge Map



Integration & results





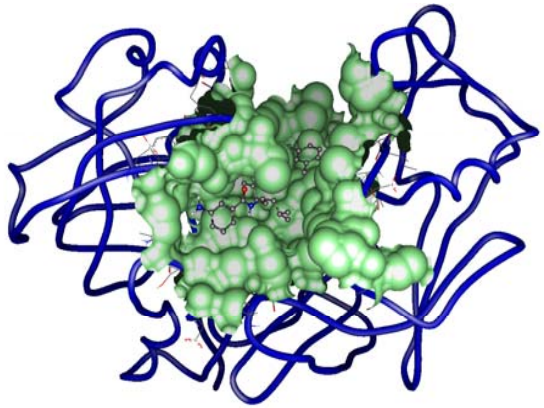
References

- "Sustainable management of groundwater exploitation using Monte Carlo simulation of seawater intrusion in Korba aquifer (Tunisia)"
  - Mr.KERROU, J. – Dr. LECCA, G. – Prof. RENARD, P. – Dr. LA ROCCA, G. – Prof. BARBERA, R. – EGEE User Forum 01-03 March 2006.

- Internation Science Grid This Week (ISGTW)
  - http://www.isgtw.org/?pid=1000341 - March, 2007

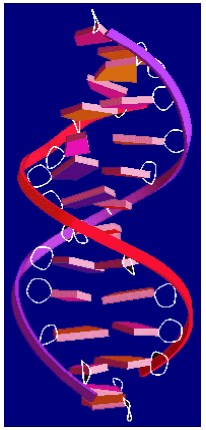
Presented at ISGC Symposium
27-29 March 2007
Taipei - TAIWAN

ROSETTA



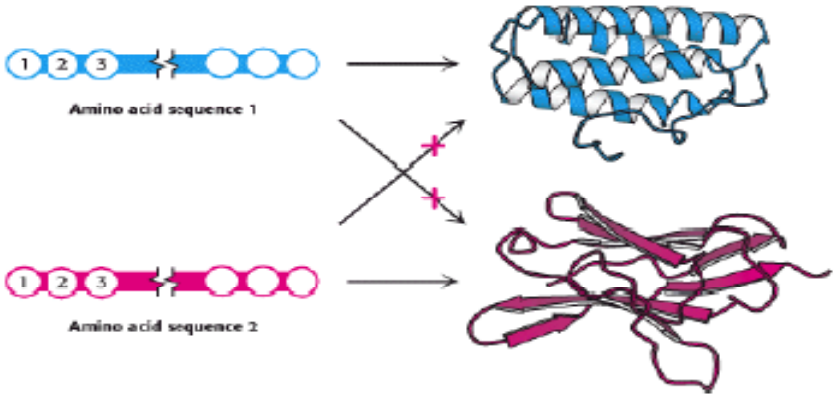
This is a work supported in collaboration with

Fabio POLTICELLI (polticel@uniroma3.it)
Pier Luigi LUISI (luisi@mat.ethz.ch)
Giovanni MINERVINI (gminervini@uniroma3.it)
Computational Biochemistry Lab.
Department of Biology – Roma 3



The Biological Applications

- **The protein folding "problem" and the structural genomics challenge**
  - The combination of the 20 natural amino acids in a specific sequence dictates the three-dimensional structure of the protein.
  - Protein function is linked to the specific three-dimensional arrangement of amino acids functional groups.
  - With the advancement of molecular biology techniques a huge amount of information on *protein sequences* has been made available but less information is available on structure and function of these proteins.
  - The "ab initio" prediction of protein structure is a key instrument to better understand the protein folding principles and successfully exploit the information provided by the "genomic revolution".



The protein sequences space

- The number of natural proteins, though apparently huge, represents just a tiny fraction of the theoretically possible protein sequences.
  - With 20 different co-monomers, a protein chain of just 60 amino acids can theoretically exist in 20^{60} chemically and structurally unique combinations.
- Estimates of the number of proteins present in nature vary from a minimum of 10^9 to a maximum of 10^{13}, thus the ratio between the number of existing proteins and those theoretically possible is very small.
  - A particularly suggestive example is that this ratio correspond to that between the volume of the hydrogen atom and that of the entire universe.

The "Never Born Proteins"

- **Rationale**
  - There exist a huge number of protein sequences that have never been exploited by biological systems, in other words enormous number of "never born proteins" (NBP).
  - The NBP pose a series of interesting questions for the biology and basic science in general:

- Which are the criteria with which the existing proteins have been selected?

- Do natural proteins have peculiar properties in terms for example of thermal stability, solubility in water or amino acid composition?

- Or else they represent just a subset of the possible protein sequences generated only by the contemporary action of contingency and physico-chemical forces?

What Rosetta does ?

- The Rosetta *ab initio* module (*developed by David Baker – University of Washington*) is a software application which *allows the prediction of the three-dimensional structure of an amino acid sequences* starting from a secondary structure of the sequence itself and a set of fragments extracted from the *Protein Data Bank (PDB).*

- The Protein Data Bank (http://www.wwpdb.org/) is a repository of proteins and nucleic acids that can be accessed for free by biologists and biochemists from around the world.

Rosetta: how it works !

- **Module I - Input generation**
  - The query sequence is divided in fragments of 3 and 9 amino acids
  - The software extracts from the data base of protein structures the distribution of three-dimensional structures adopted by these fragments based on their specific sequence
  - For each query sequence is derived a fragments data base which contains all the possible local structures adopted by each fragment of the entire sequence.

- **Module II - *Ab initio* protein structure prediction**
  - The sets of fragments are assembled in a high number of different combinations by a Monte Carlo procedure.
  - The resulting structures are subjected to a energy minimization procedure using a semi-empirical force field.
  - The principal non-local interactions considered are hydrophobic interactions, electrostatic interactions, main chain hydrogen bonds and excluded volume.
  - The compatible structures both with local biases and non-local interactions are ranked according to their total energy resulting from the minimization procedure.

Rosetta: module I

- The procedure for input generation is rather complex but computationally inexpensive (10 min of CPU time on a Pentium IV 3,2 GHz).
- Due to the many dependencies of module I (*Blast* and *psipred*), the input generation is carried out locally with a script that automatizes the procedure for a large dataset of sequences.
- Approximately 500 input datasets are currently being generated daily.

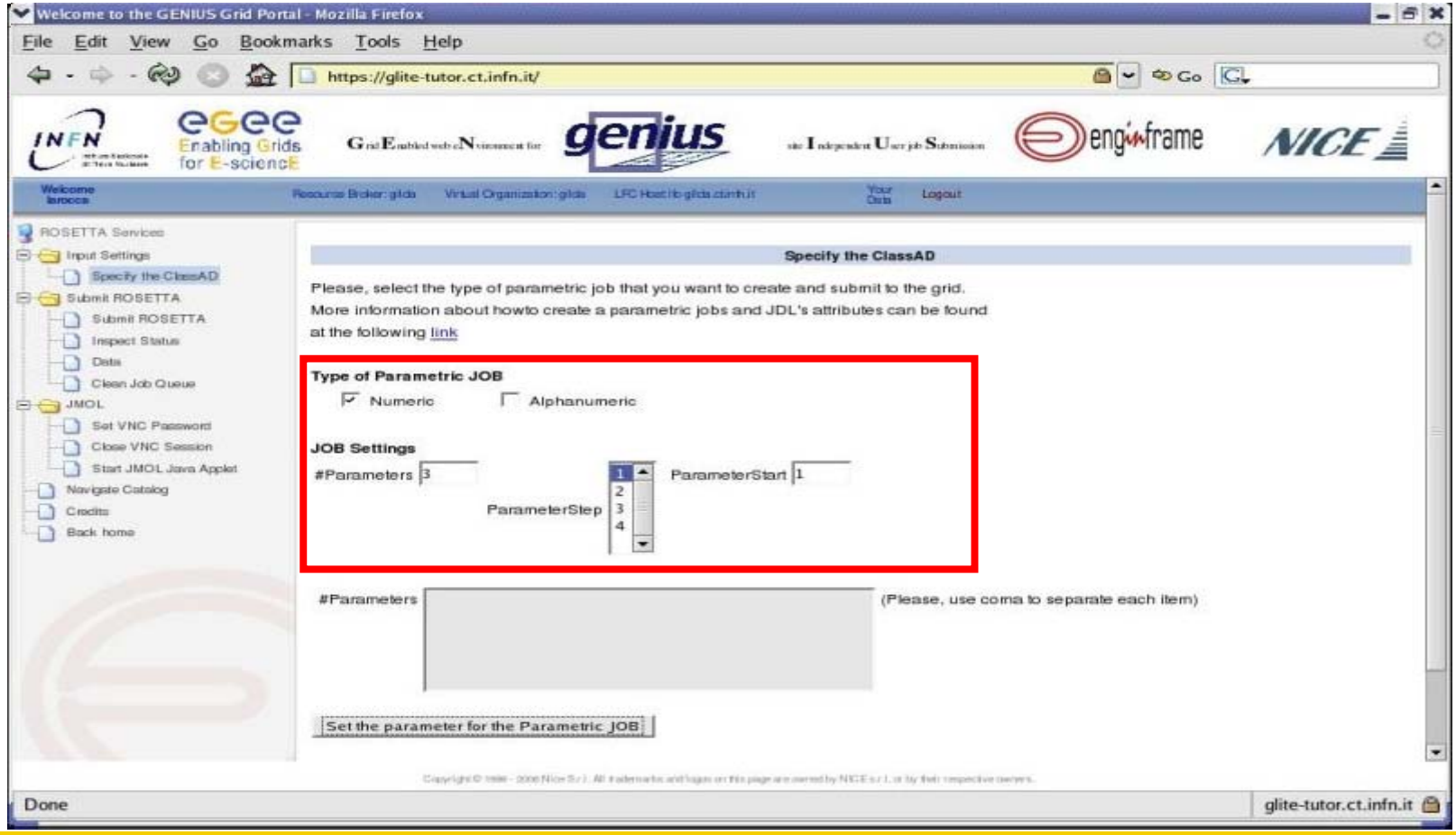
Rosetta: module II

- Input
  - fragment files generated by module 1
  - secondary structure prediction using psipred

- In output the user obtains a number of structural models of the query sequence ranked by total energy
- A single run with just the lowest energy structure as output takes approx. 10-40 min of CPU time depending on the degree of refinement of the structure
- The Module II has been implemented in GRID through the use of the GENIUS Grid Portal (https://glite-tutor.ct.infn.it)
  - From this portal, exploiting the last feature of the gLite middleware, (www.glite.web.cern.ch/glite) it's possible submitting parametric jobs and run, in one shot, a large number of jobs (structure predictions).

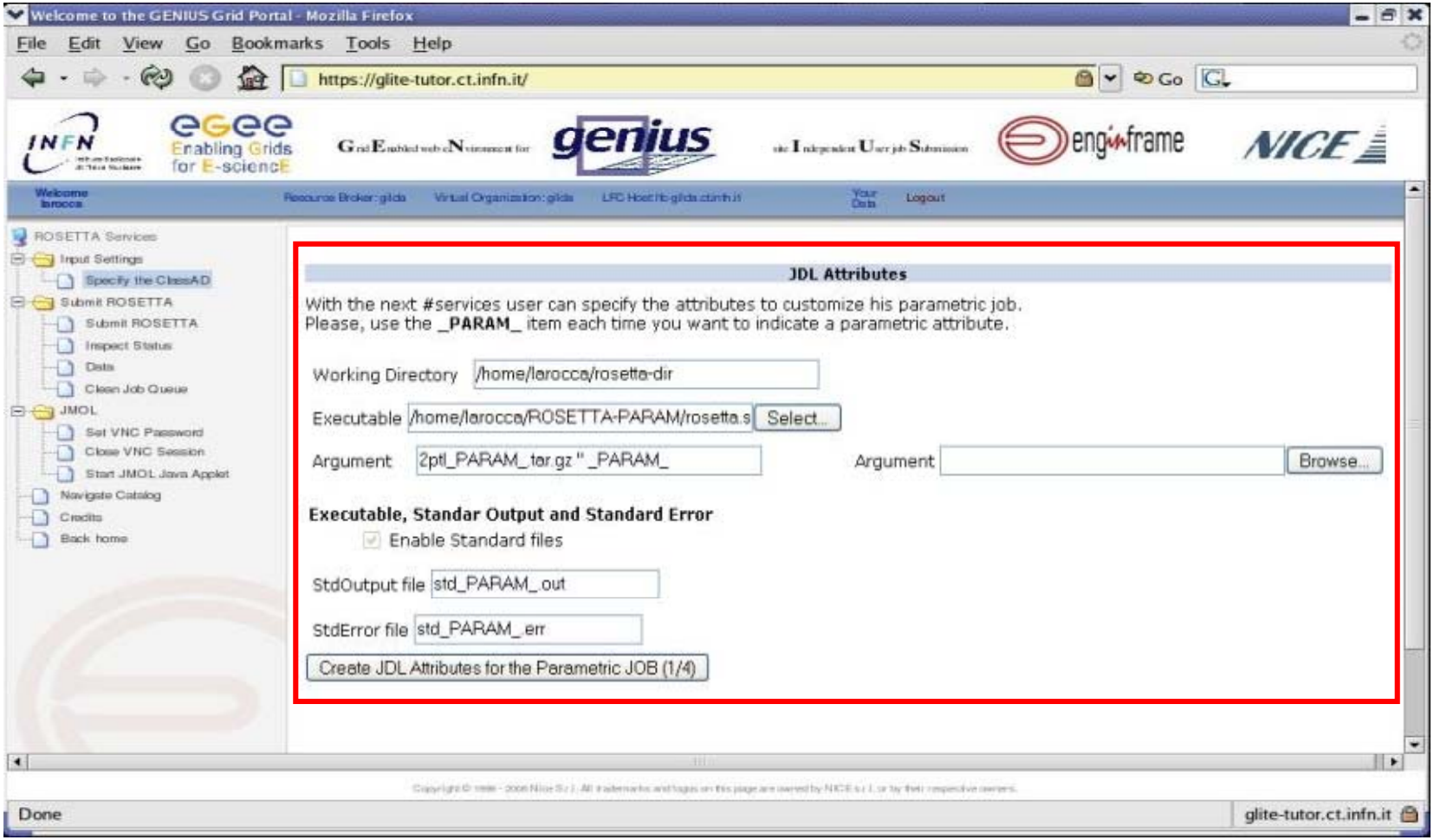
Create the dynamic ClassAD /1

- **Step 1.** After MyProxy initialization the user connects to the GENIUS portal to set up the parametric JDL, specifying the number of runs (equivalent to the number of amino acid sequences to be simulated) to be carried out.



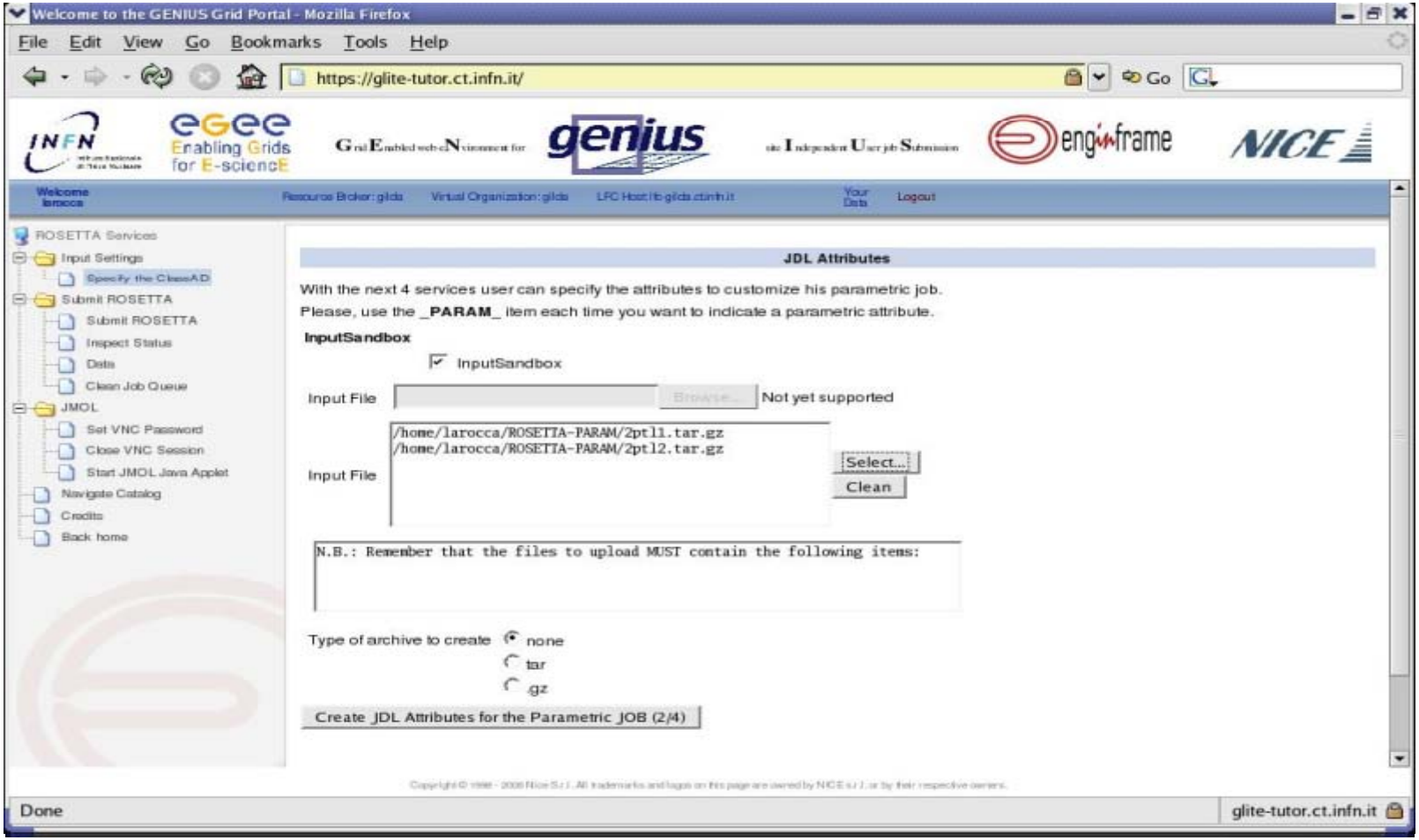
Create the dynamic ClassAD /2

- **Step 2.** The user specifies the working directory and the name of the shell script.



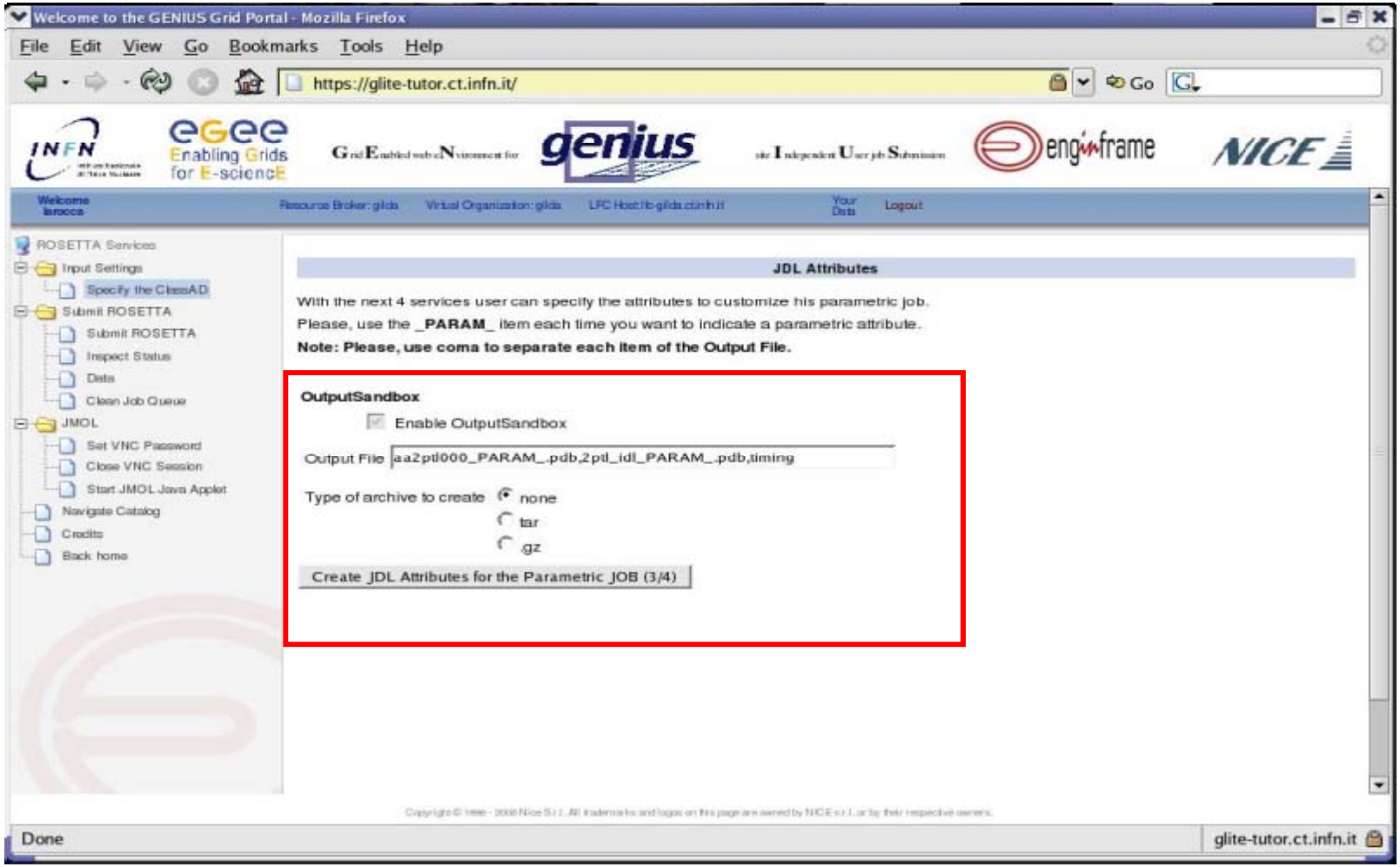
Create the dynamic ClassAD /3

- **Step 3.** Input files (fragment libraries) are loaded as a single .tar.gz folder per amino acid sequence.



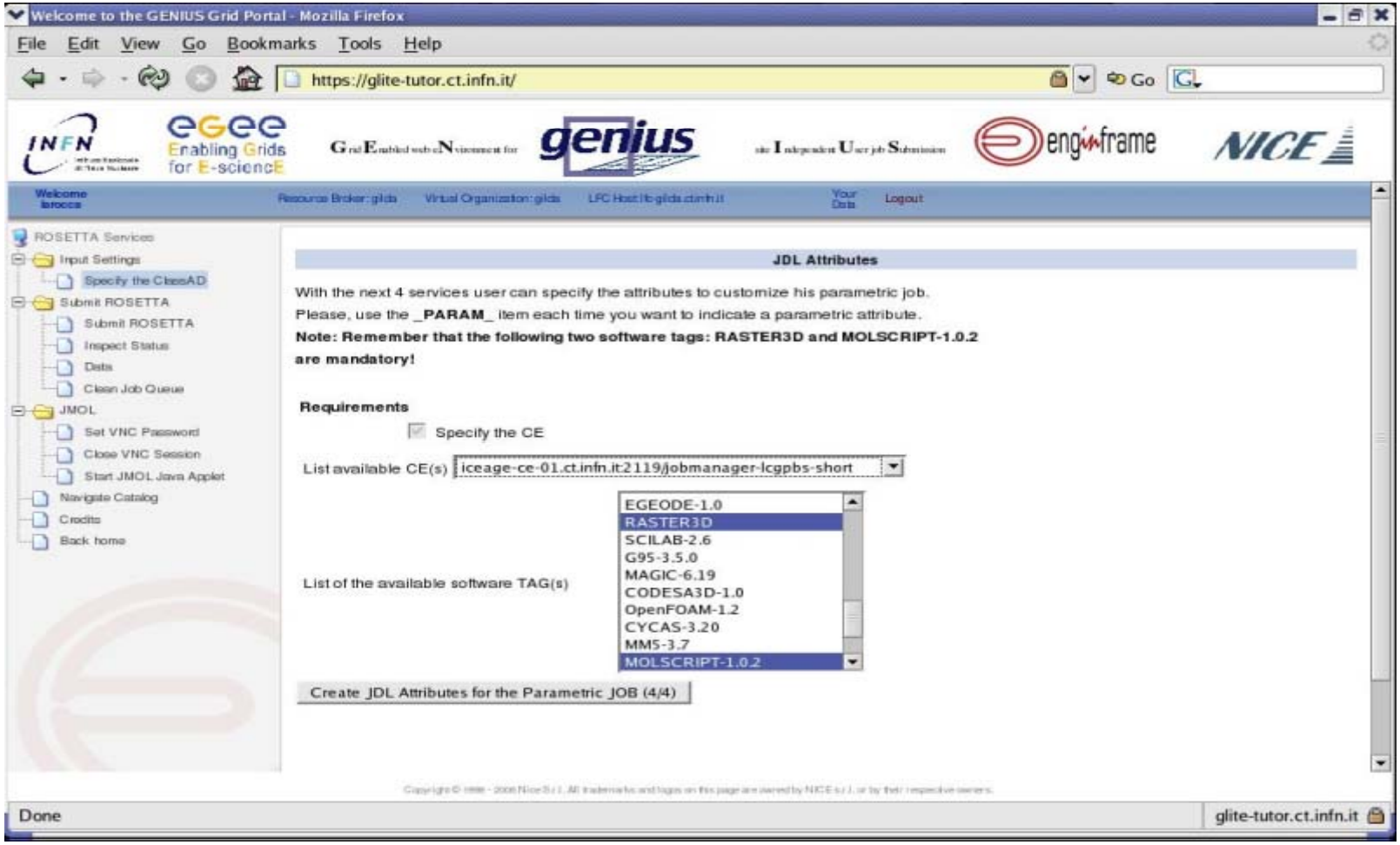
Create the dynamic ClassAD /4

- **Step 4.** Output files (initial and refined model coordinates) are specified in parametric form.



Create the dynamic ClassAD /5

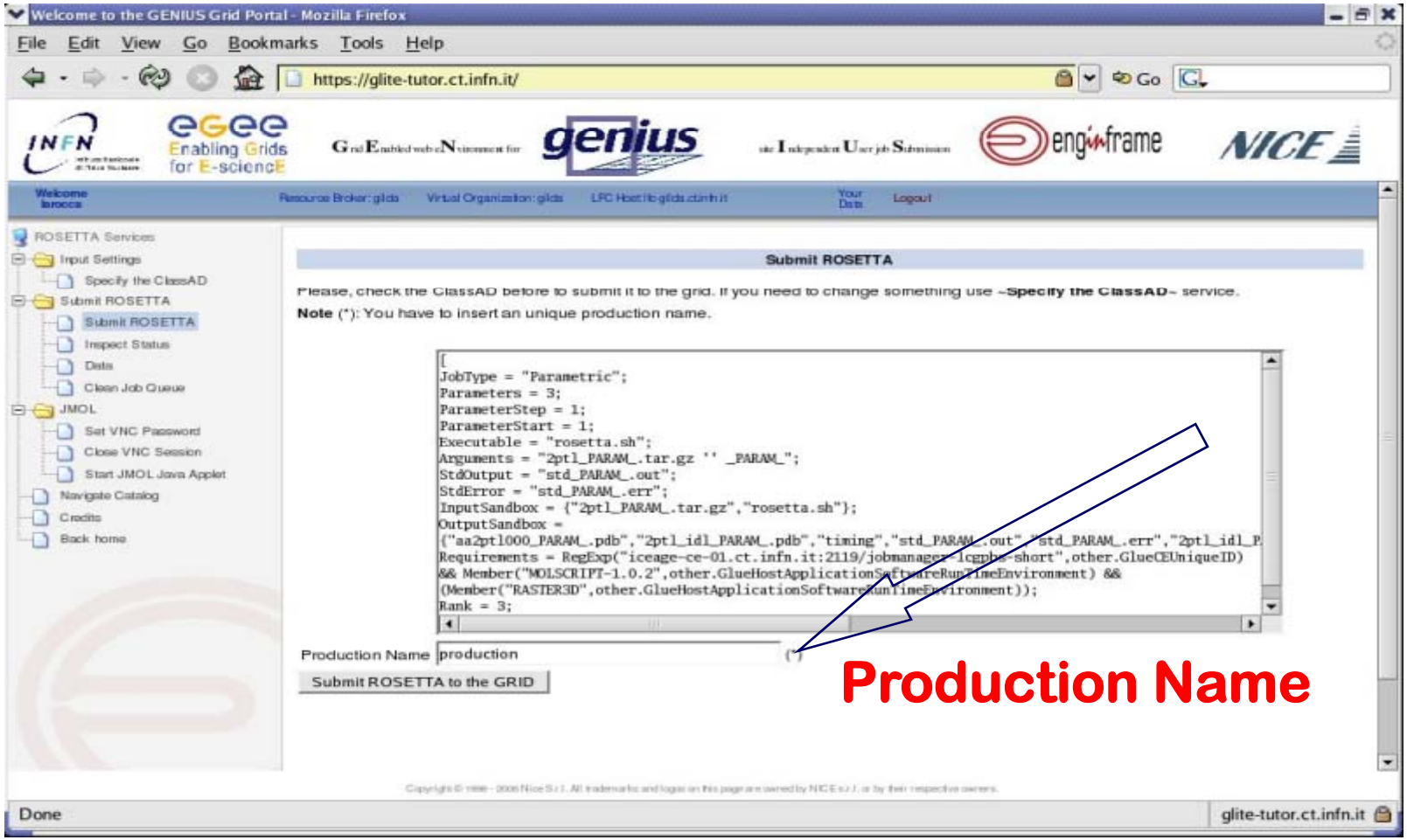
- **Step 5.** The software requirements are specified in order to properly run ROSETTA.



Submit ROSETTA to the Grid /1

BioinfoGRID

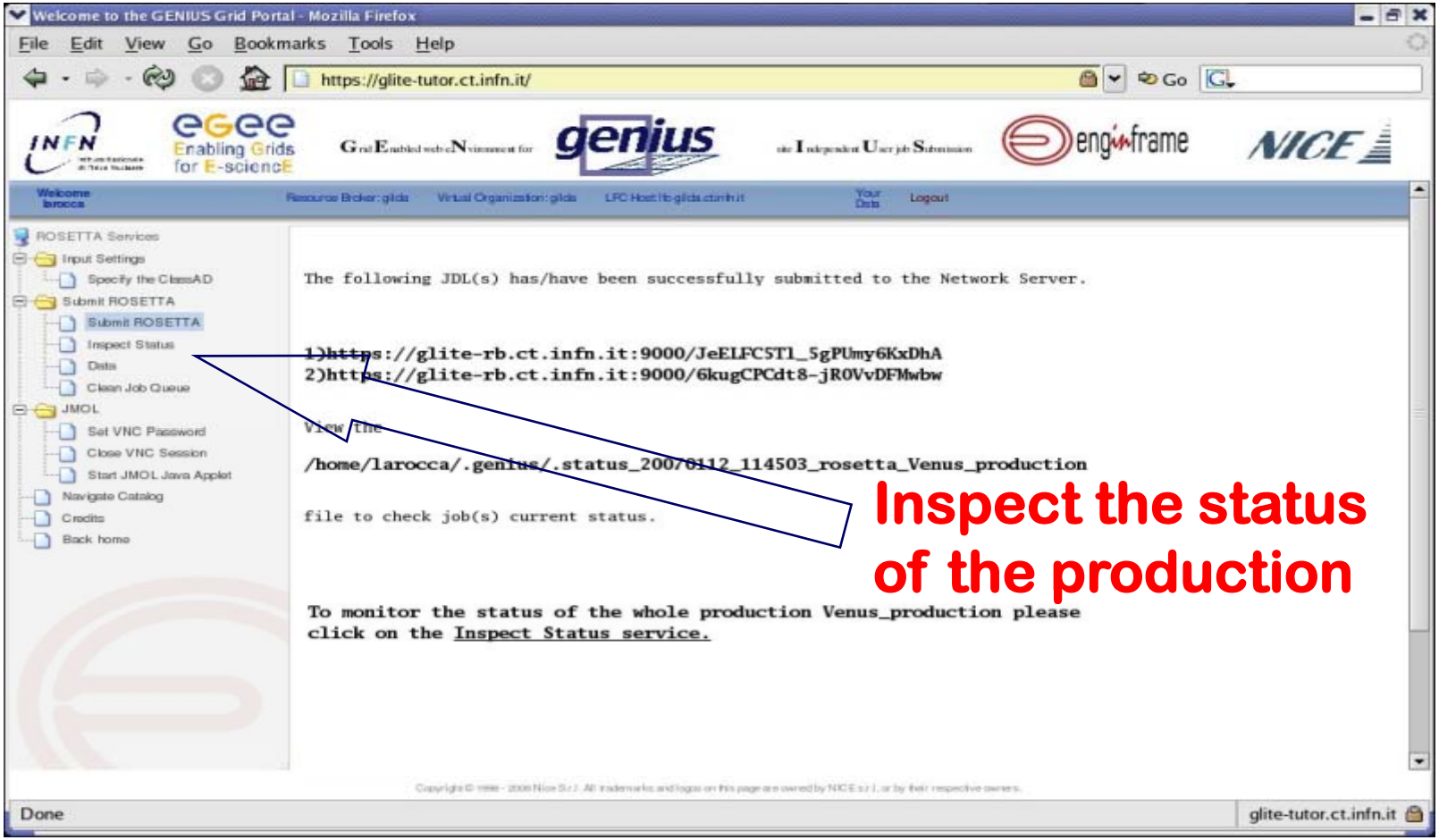
- **Step 6.** The parametric JDL file is generated and visualized to be inspected by the user.



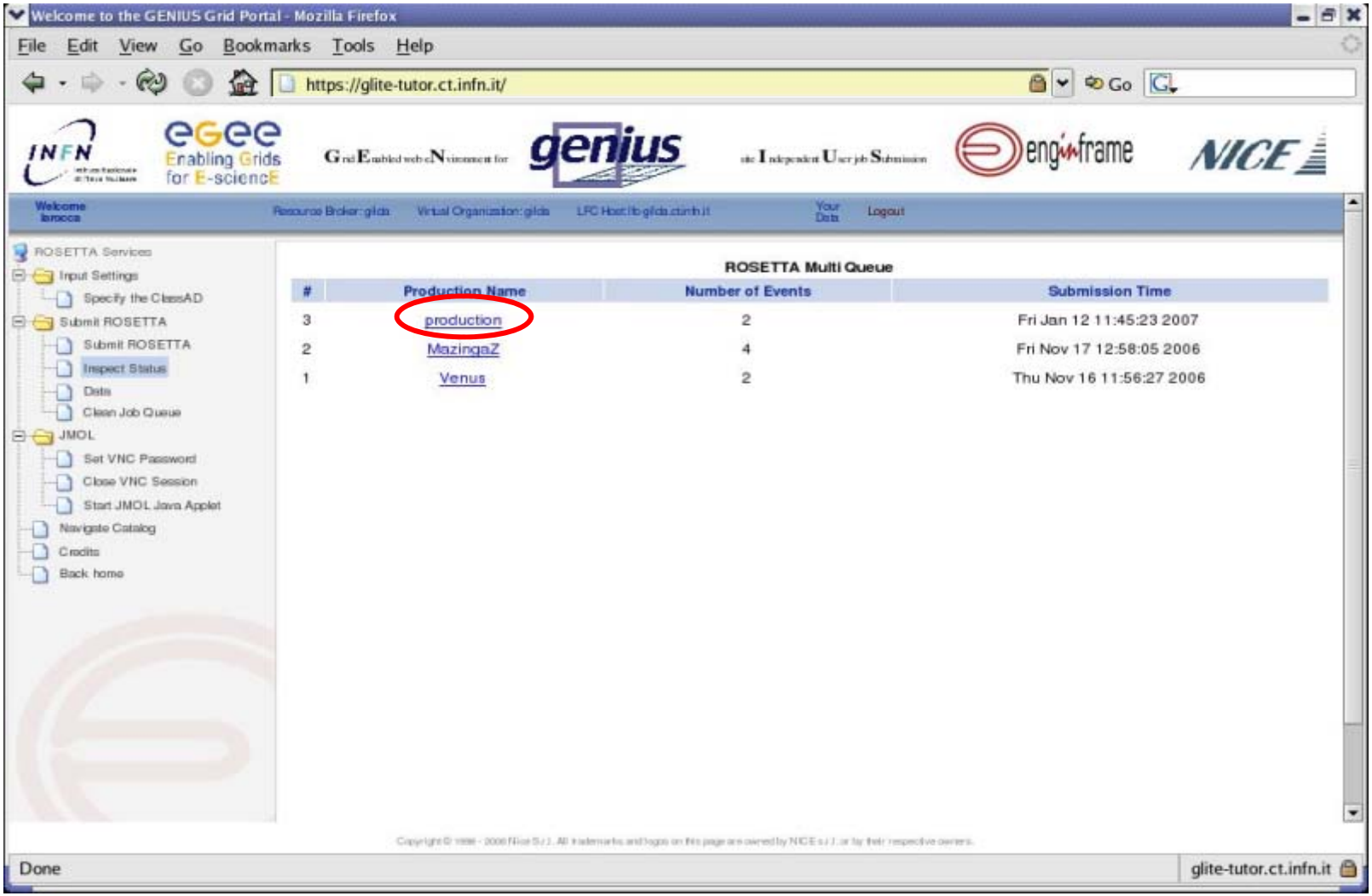
Submit ROSETTA to the Grid /2

BioinfoGRID

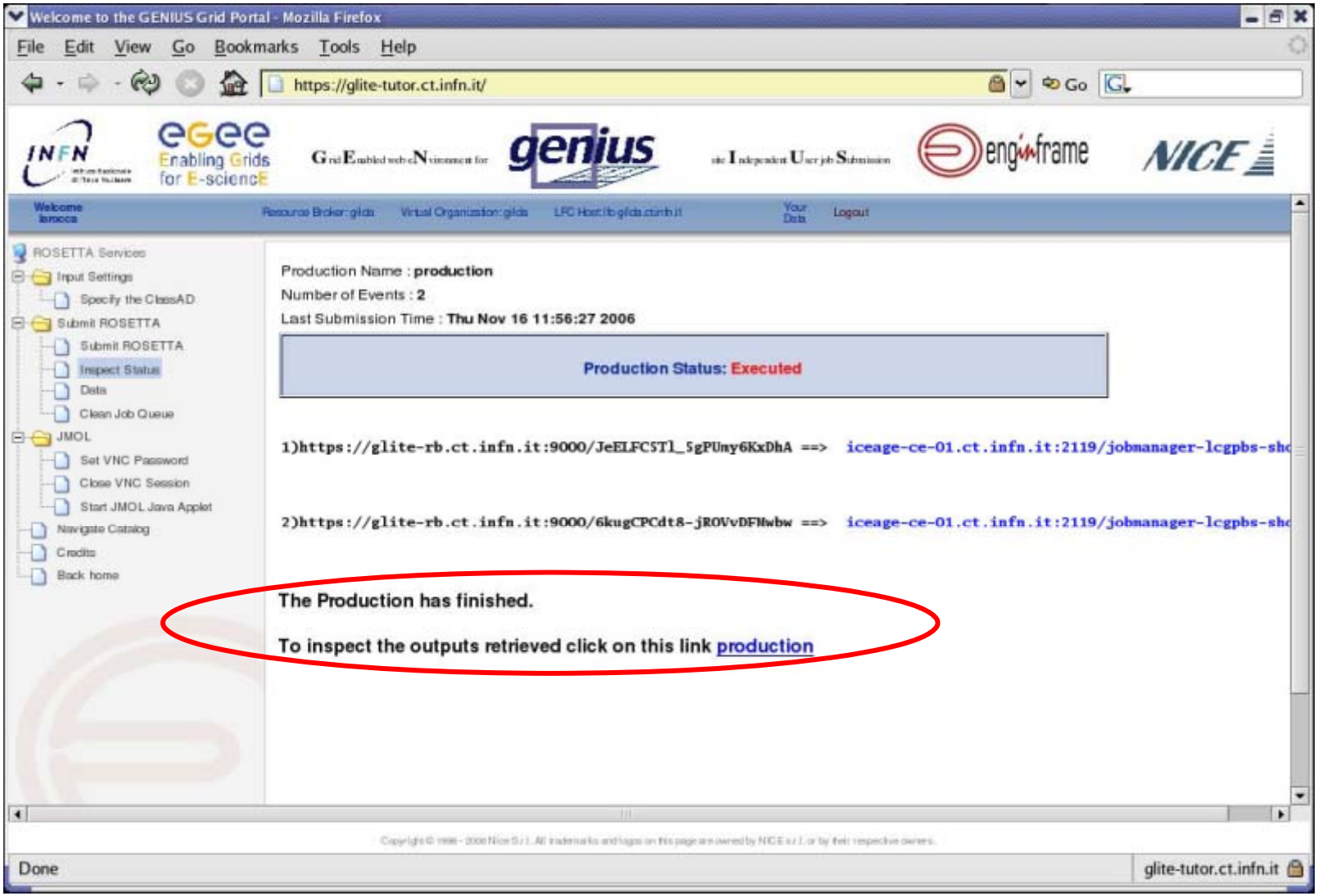
- **Step 7.** The parametric job is submitted and its status as well as the status of individual runs of the same job can be checked.



Inspect Status /1

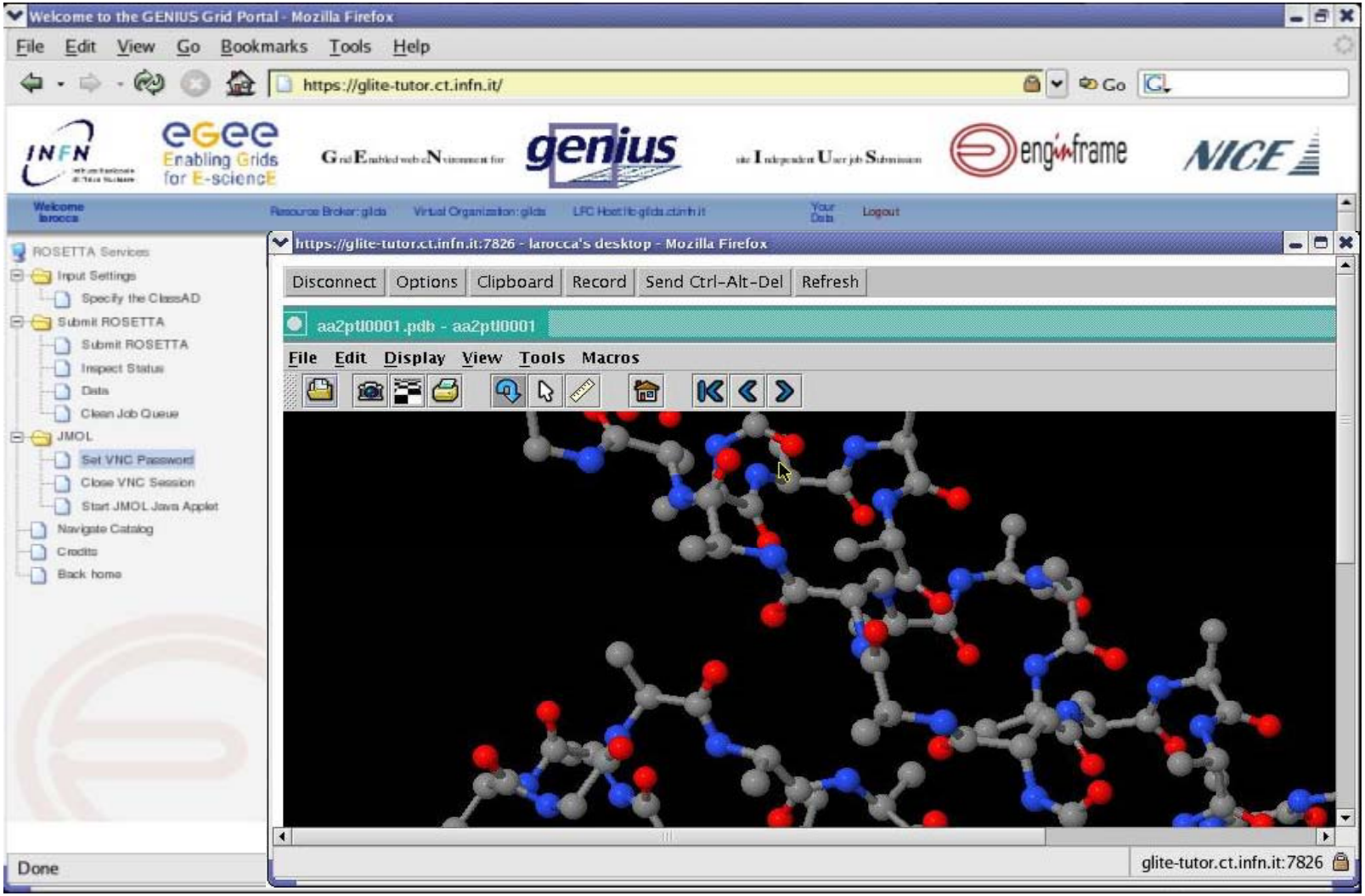


Inspect Status /2



Click here to inspect the typical output files produced by ROSETTA at the end of the prediction process

JMOL Applet Java



References

- "Porting Biological Applications in Grid : an experience within the EUChinaGRID framework"
  - Prof. POLTICELLI, Fabio – Dr. LA ROCCA, Giuseppe – Prof. LUISI, Pier Luigi, Dr. MINERVINI Giovanni – EGEE User Forum 27-29 March 2007.

- Protein structure prediction of "never born proteins". An experience within the EUChinaGRID framework.
  - Prof. POLTICELLI, Fabio – Dr. LA ROCCA, Giuseppe – Prof. LUISI, Pier Luigi, Dr. MINERVINI Giovanni – EGEE User Forum 09-11 May 2007.

Presented at First EUMedGrid Project Review Madrid, 27-28 Feb. - SPAIN

ARCHAEO CLIMATOLOGY

This is a work supported in collaboration with

Pier Giovanni PELFER (pelfer@fi.infn.it)

University of Florence - ITALY

The Mediterranean Region at the end of the urban revolution

BioinfoGRID

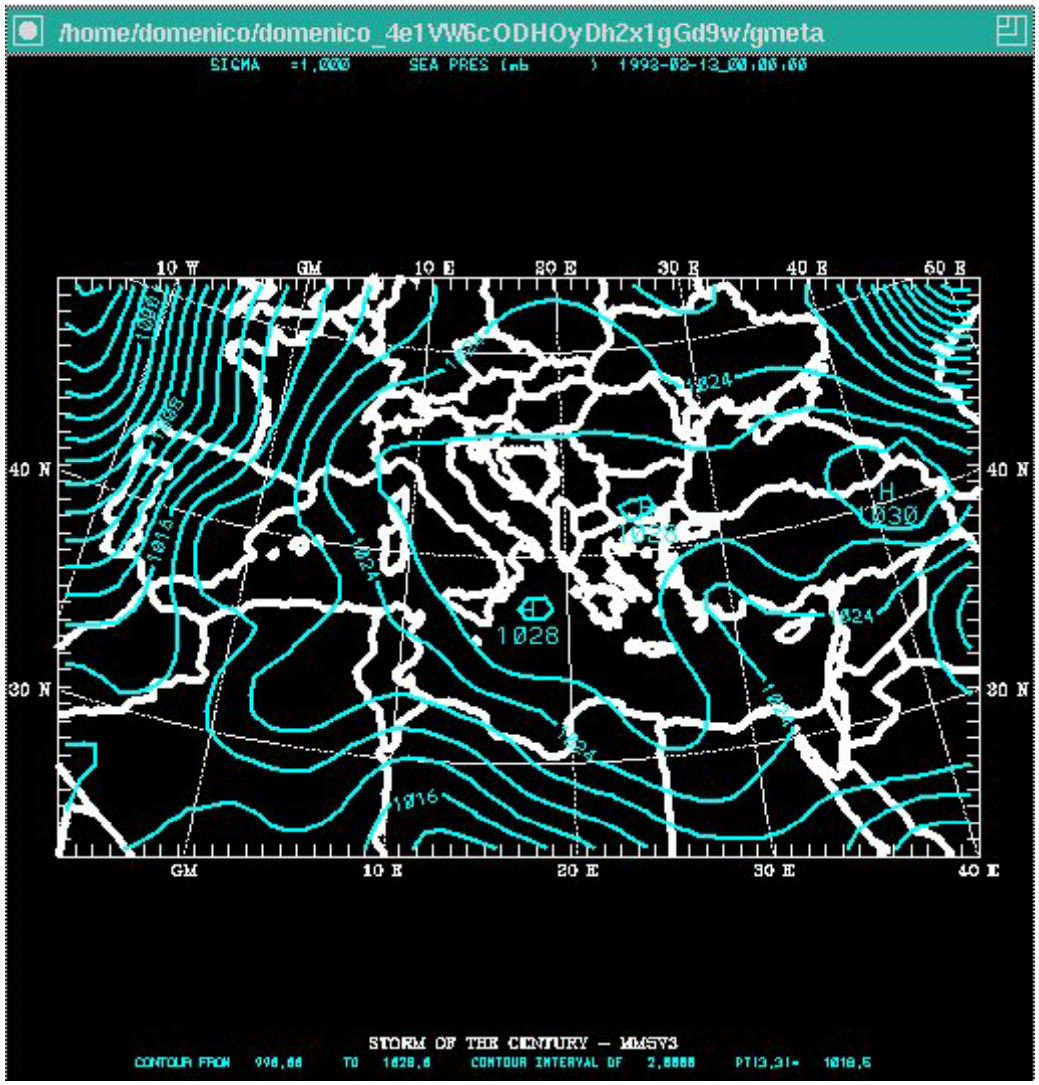
SHIPBUILDING and navigation were essential skills. This detail of a vase painting c.540 shows the god Dionysus at sea (having escaped capture by pirates).

ArchaeoGrid aims to investigate the paleoclimate situation in the Mediterranean Area between XVIII centuries B.C.

This kind of studies are made available, within EUMEDGRID, using a metereological modelling software called MM5. It allows to create complex atmospheric simulations and simulate weather evolution.

The result

The MM5 program, thanks to the description of the environment, elaborates the meteorological data and generate the forecast



References

- "Archaeological Applications on e-Infrastructures by ArchaeoGRID"
  - Dr. PELFER, Giuliano – Dr. CECCHINI, Roberto – Prof. PELFER, Pier Giovanni, Dr. POLITI, Antonio – EGEE User Forum 09-11 May 2007.

- Grid Portals technology
- EnginFrame Framework
- The GENIUS Grid Portal
- Some Integrated Applications
- References

References

- NICE web-site http://www.nice-italy.com
- EnginFrame Framework http://www.enginframe.com
- GENIUS Portal https://genius.ct.infn.it
- GENIUS Repository at https://geniuscvs.ct.infn.it

GENIUS Installation

- GENIUS Repository at https://geniuscvs.ct.infn.it
- Write an email message to alberto.falzone@nice-italy.com or nicola.venuti@nice-italy.com for an account request to download the GENIUS package

Thank you for your attention