GReIC

Services for Heavy User Communities

EGI Technical Forum 2011

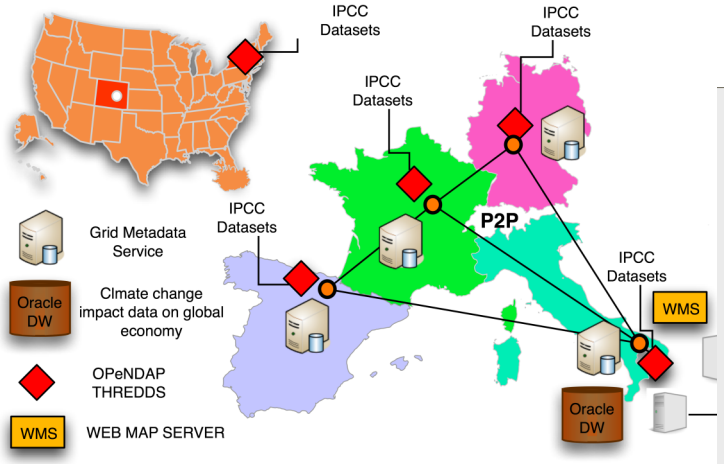
S. Fiore and G. Aloisio

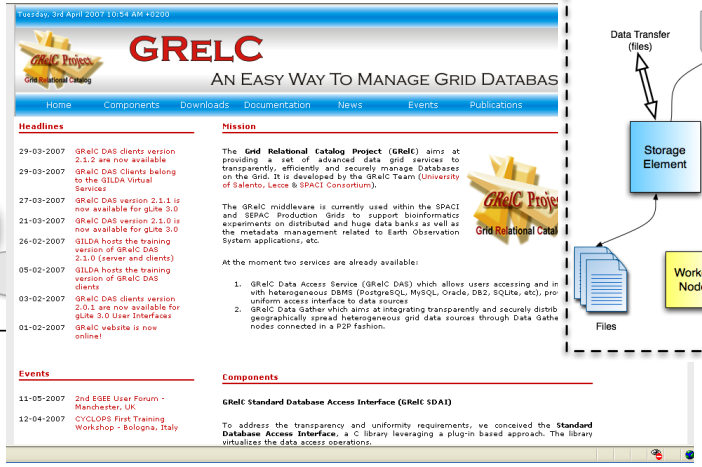
SPACI and University of Salento

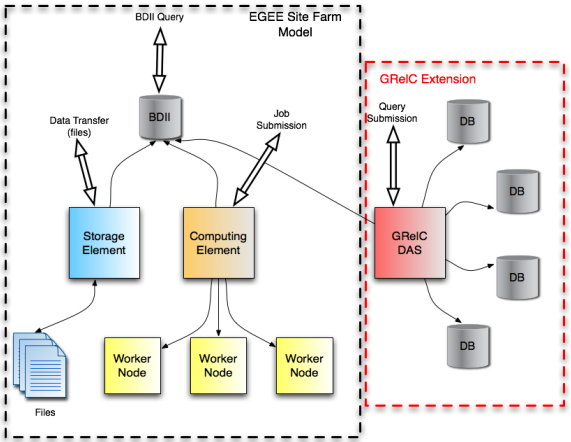
The GRelC Project: main goal and service

- Grid Relational Catalog is a project which aims at designing and developing a set of efficient, secure and transparent Data Grid Services (Starting date, January 2001).

- **GRelC Service** aims at providing a large set of functionalities to access to both relational and non relational Databases in a grid environment.

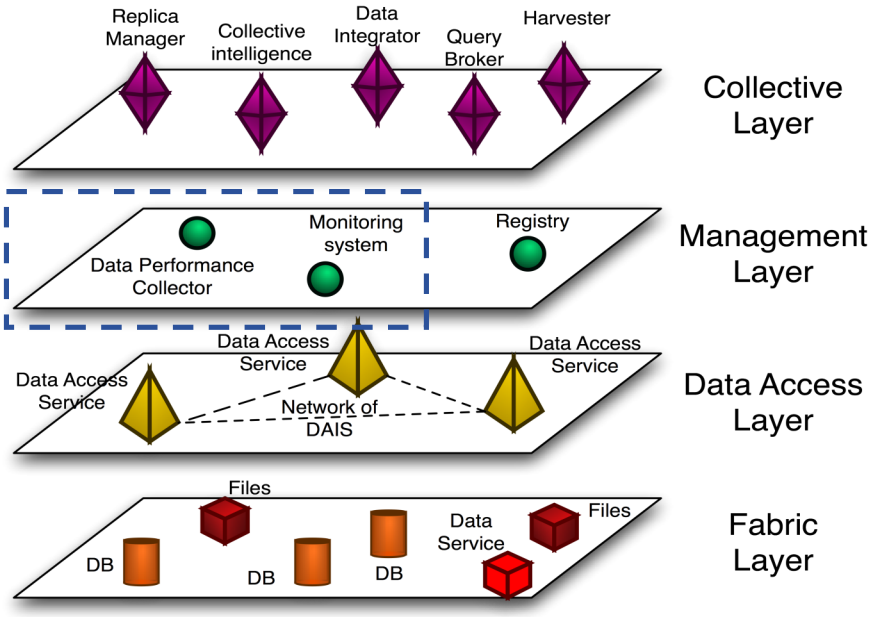


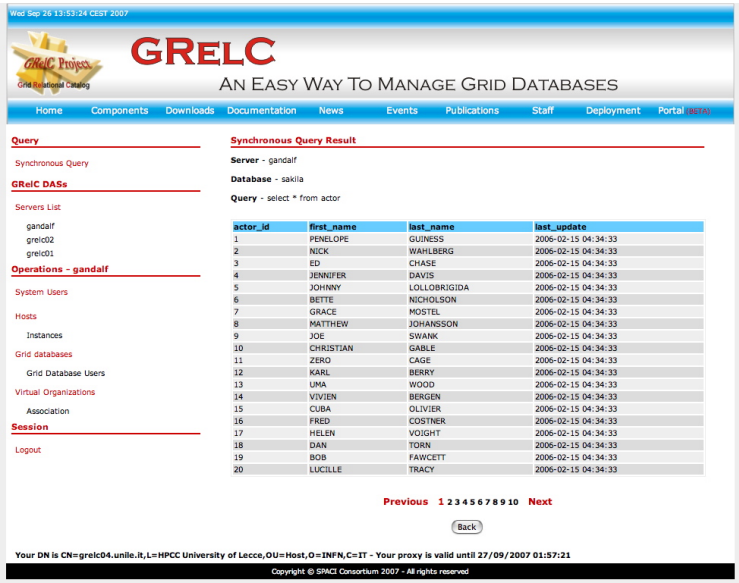




DoW (I) - “Management Layer”

An important task will be related to the monitoring and control functionalities connected with the underlying P2P infrastructure of the GRelC system. Such a management framework will be managed through the GRelC Portal by means of a new set of web pages exploiting the dashboard approach (charts, reports, table, diagrams able to provide global and local views about the status of the system). Users will be able to configure, manage and query their own GRelC services, exploiting a wide set of management functionalities embedded into the GRelC Portal. A key point will be to make easier (few steps in a web-based wizard) the *gridification process* (bringing into the grid) of a database resource.

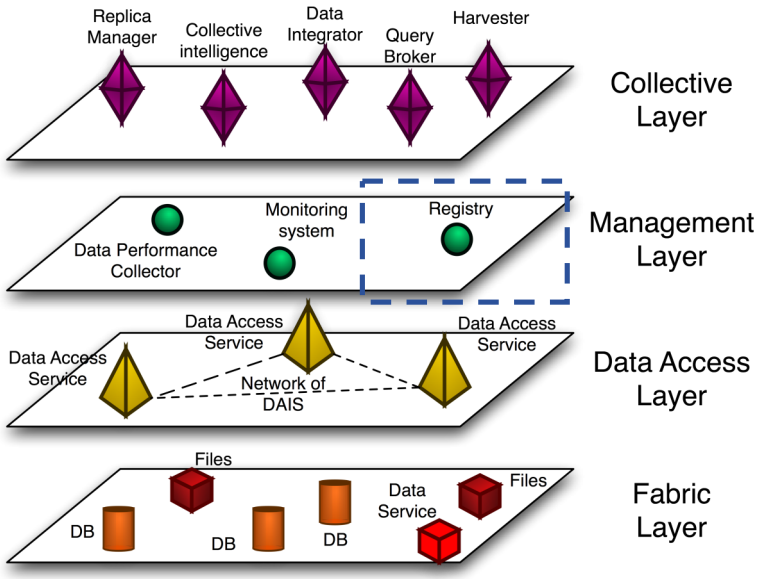


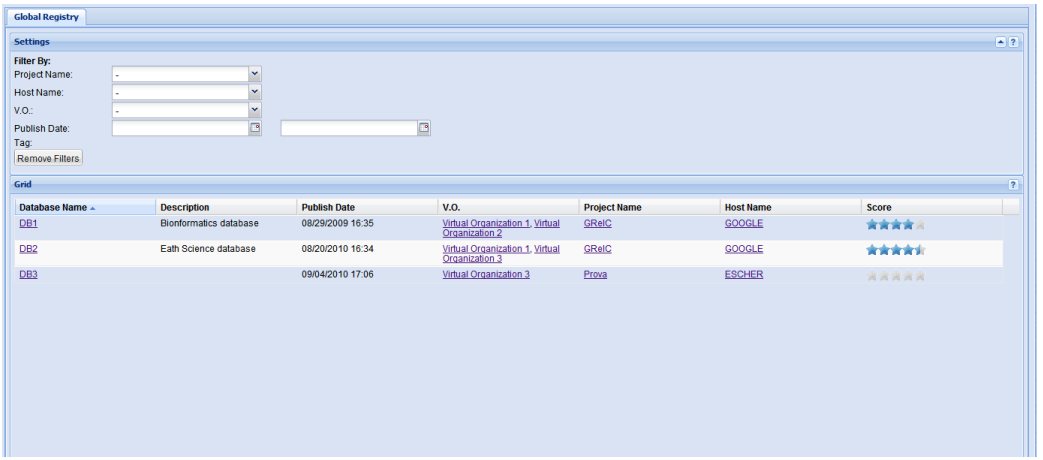


DoW (II) - “Management Layer”

A key task will be the creation of the EGI Database of Databases, a registry service accessible through a specific GRelC Portal web page that will contain all of the information about the grid-databases available in the GRelC System. Users will be able to:

- *query the registry* (exploiting a keyword-based approach) asking for specific databases, filtering by VO, keywords, domain, etc. This will help people working in a specific domain to quickly identify available and related resources, identify key people working on specific subjects, easily contact them to establish collaborations, etc.
- *join a specific grid-database*, submitting via the web a request to the grid-database administrator to know more about the supported VOs, etc.;
- *add comments* on the available data and the related data sources, being part of a community exploiting a collaborative and Web2.0 oriented approach. All of this data will be available for future users, creating a knowledge base centred around community-oriented topics.





DashboardDB and EGI (I)

- A new system (**DashboardDB**) more targeted on the GRelC service has been designed during Y1
  - It represents a unified environment (web based) joining social, management and monitorings aspects
  - Key aspect: focus on “grid-databases”

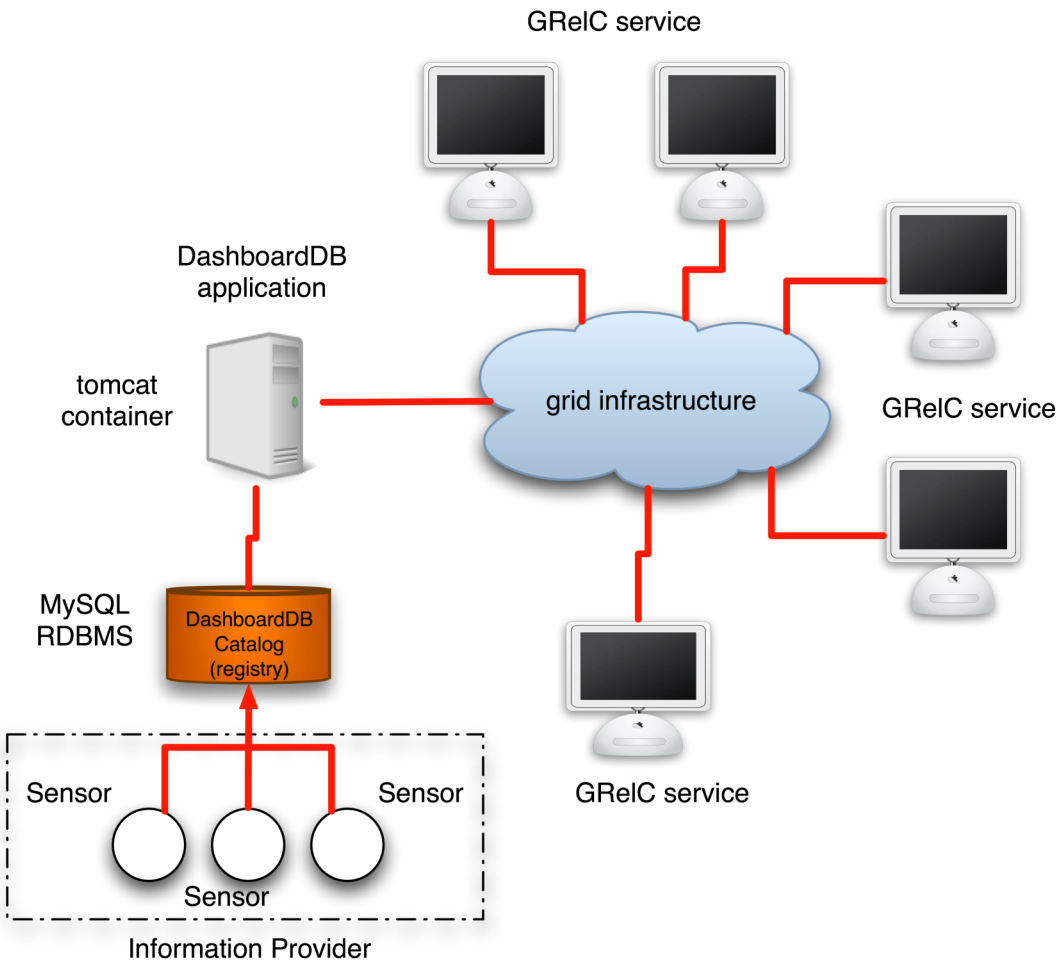
- Non functional requirements:
  - Pervasivity, user-friendliness and transparency
    - A web based solution is a good candidate
  - Security
    - …taking into account the security implementation must not be a barrier for new users
  - Look and Feel
    - Technolgical impacts on the adopted software libraries

DashboardDB and EGI (II)

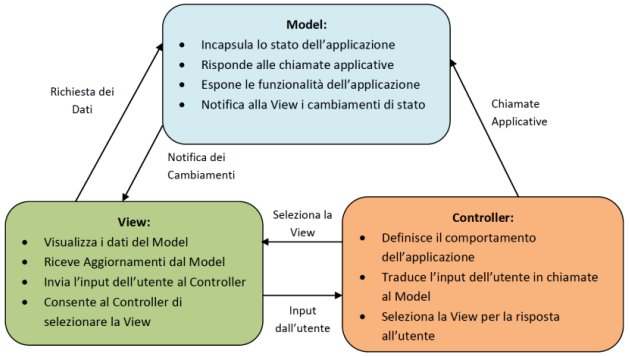
- Functional requirements:
  - Monitoring of GReIC service instances
  - Provision of specialized views related to the “network” of GReIC services
    - Database/VO association
    - Database distribution
    - etc.
  - Creation of a community oriented registry of grid-database resources
    - Discussion groups
    - Tagging capabilities
    - etc.
  - Important Features
    - Permalink support
    - Support for multiple views
      - *Based on countries, goal, etc.*

DashboardDB: Architecture

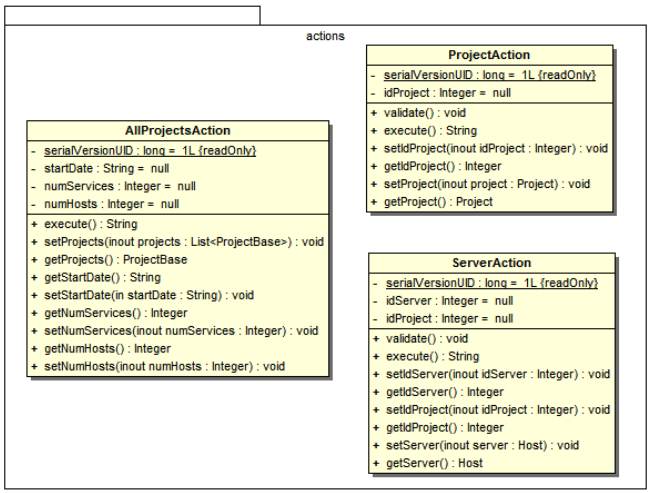
System Architecture



Model View Controller Pattern

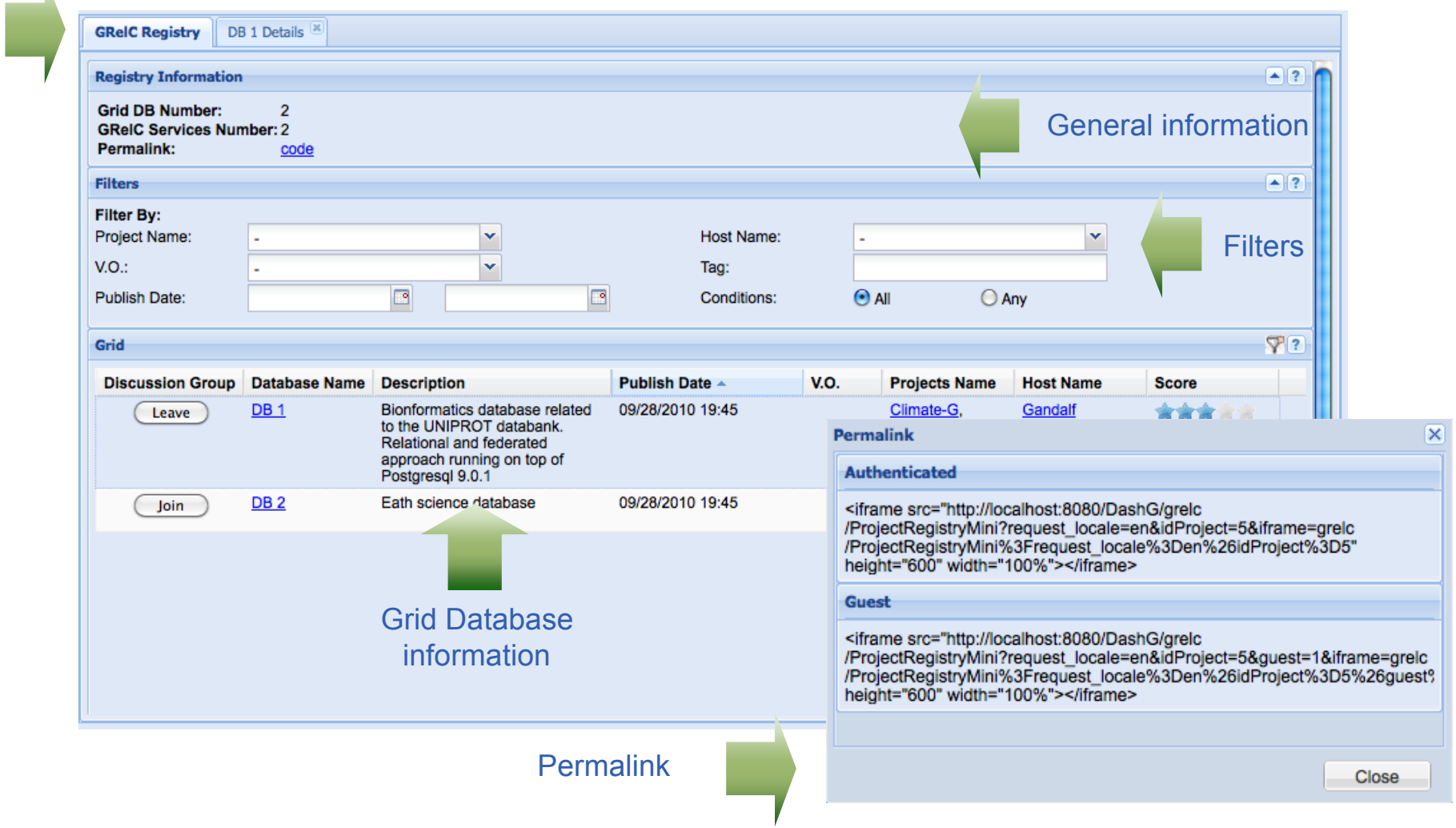


Main actions



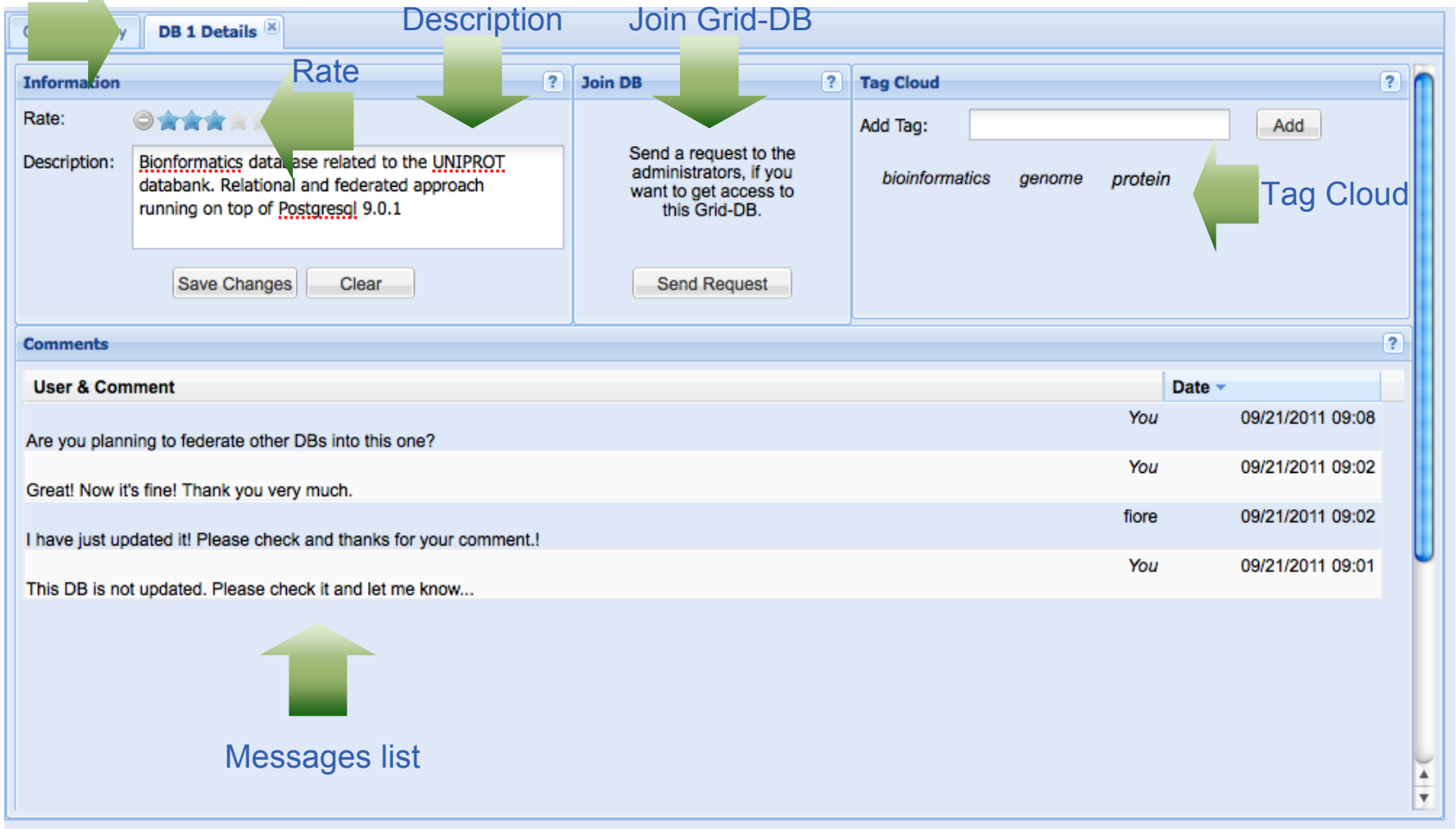
The DashboardDB Registry: main view

GRelC Registry

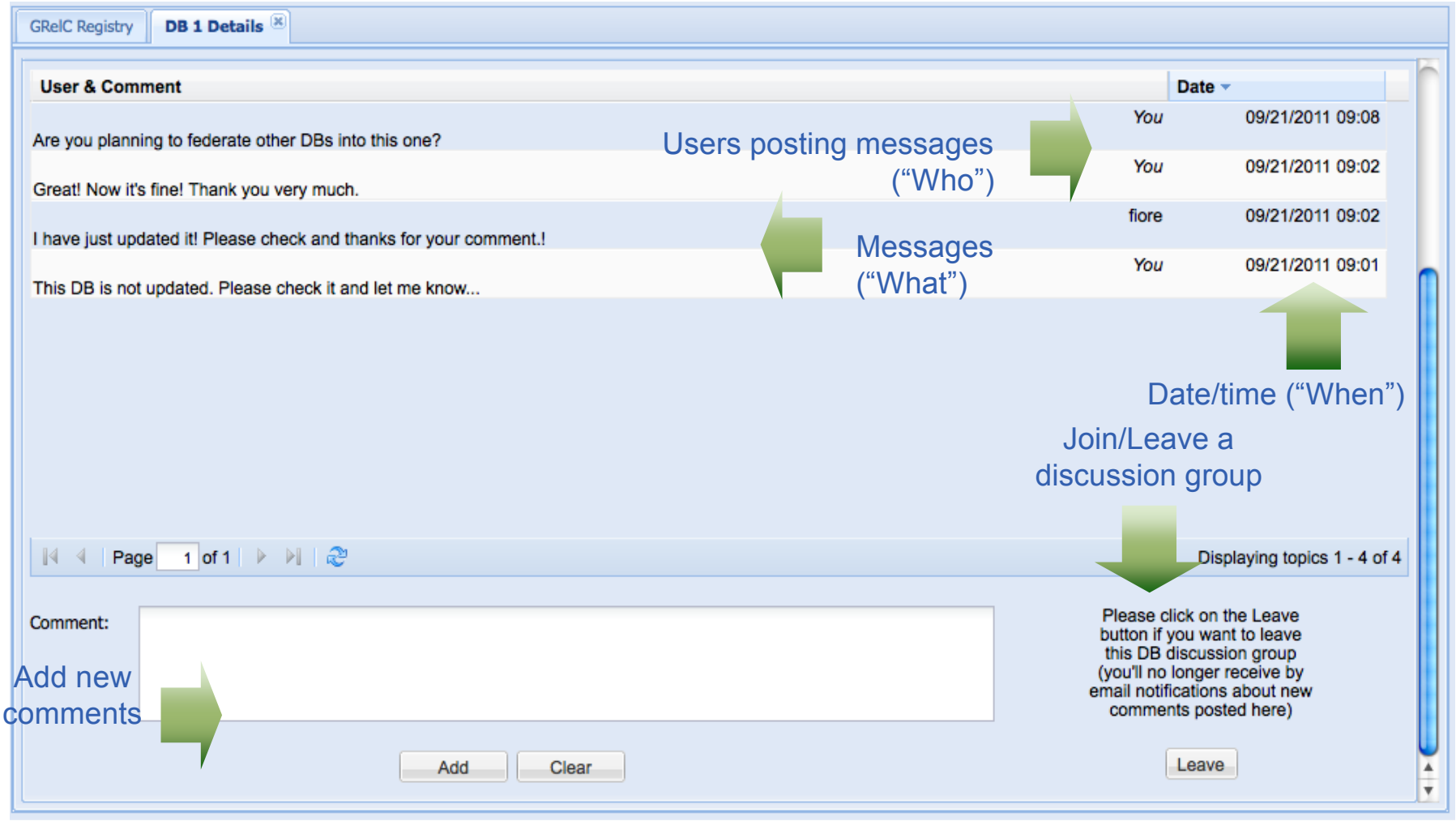


The DashboardDB Registry: grid-DB view

Grid-DB Details



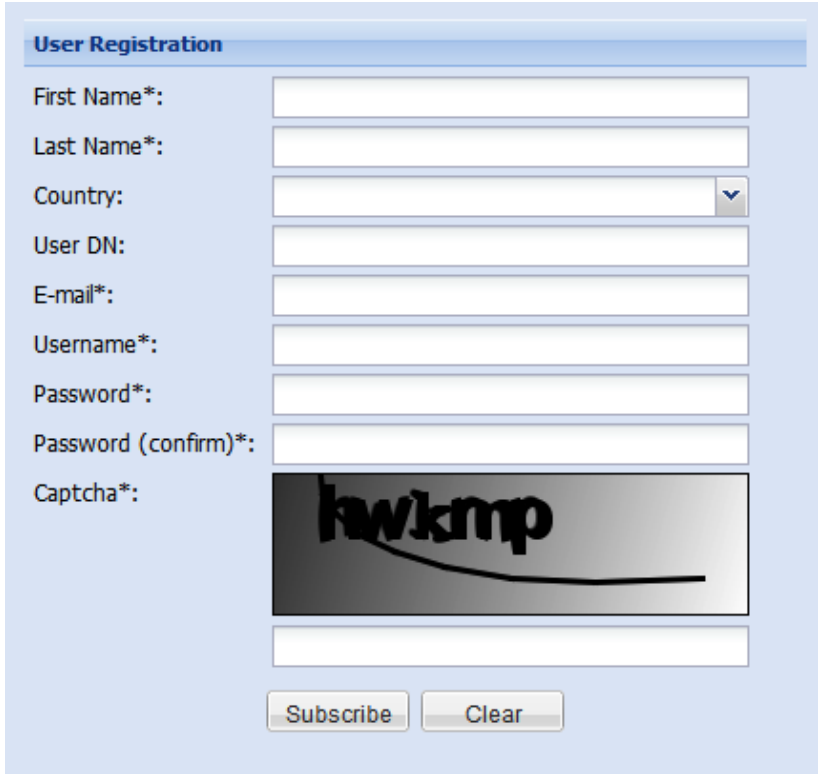
DashboardDB: the Registry

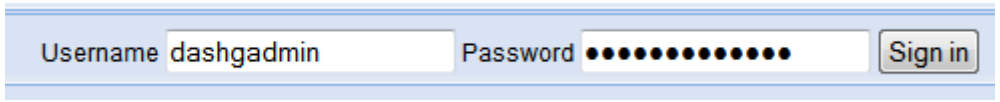


DashboardDB: Security aspects

Security Management:

- User Registration
- User Authentication
- User Profile Management
- User Authorization
- Guest users (access to public projects)





DashboardDB: Permalinks and Mashup

By including into a target web page a simple line of code like:

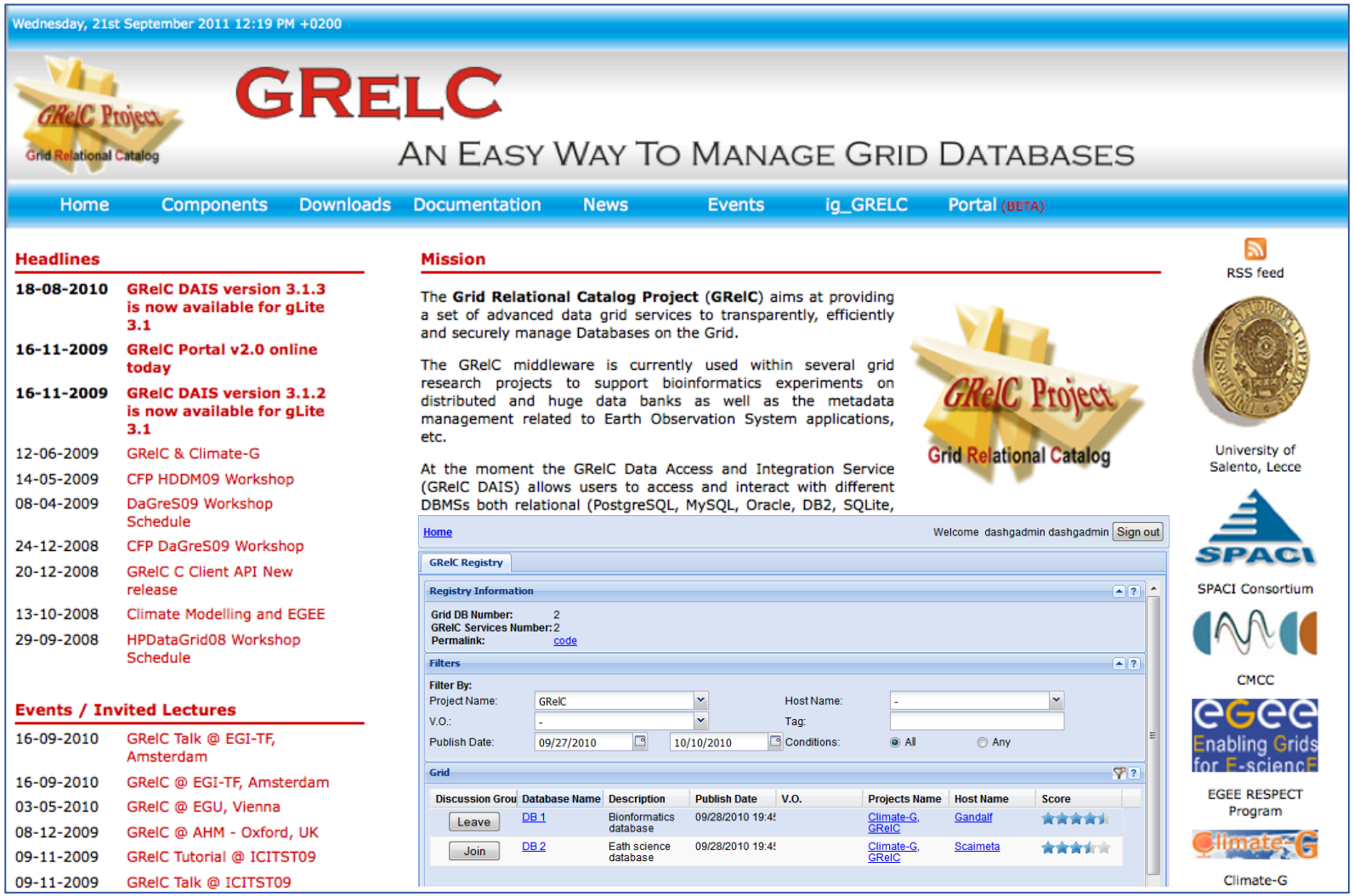
```
…
<iframe
src="http://host:8080/dashboardDB/…./ProjectRegistry….?request_locale=en&idProject=…&frame=…/ProjectRegistry…%3Frequest_locale%3Den%26idProject%3D5" height="600" width="100%"></iframe>
…
```

you can embed the DashboardDB registry into your web application in a straightforward manner like a YouTube video.

Authorization can be turned on/off into the target web page

Reusability can strongly be addressed by exploiting permalink capabilities (key issue for software sustainability)

DashboardDB: “embedding” the registry



Ongoing activities and new ones planned for Y2

Ongoing activities and new ones planned for Y2:

- **Porting** of the GRelC software on:
  - **gLite 3.2** (SL5.x) very soon (some problems with 64bits SSL libraries on SL5, prevented the team to release the software at the end of Y1)
  - … and on **EMI** soon after that

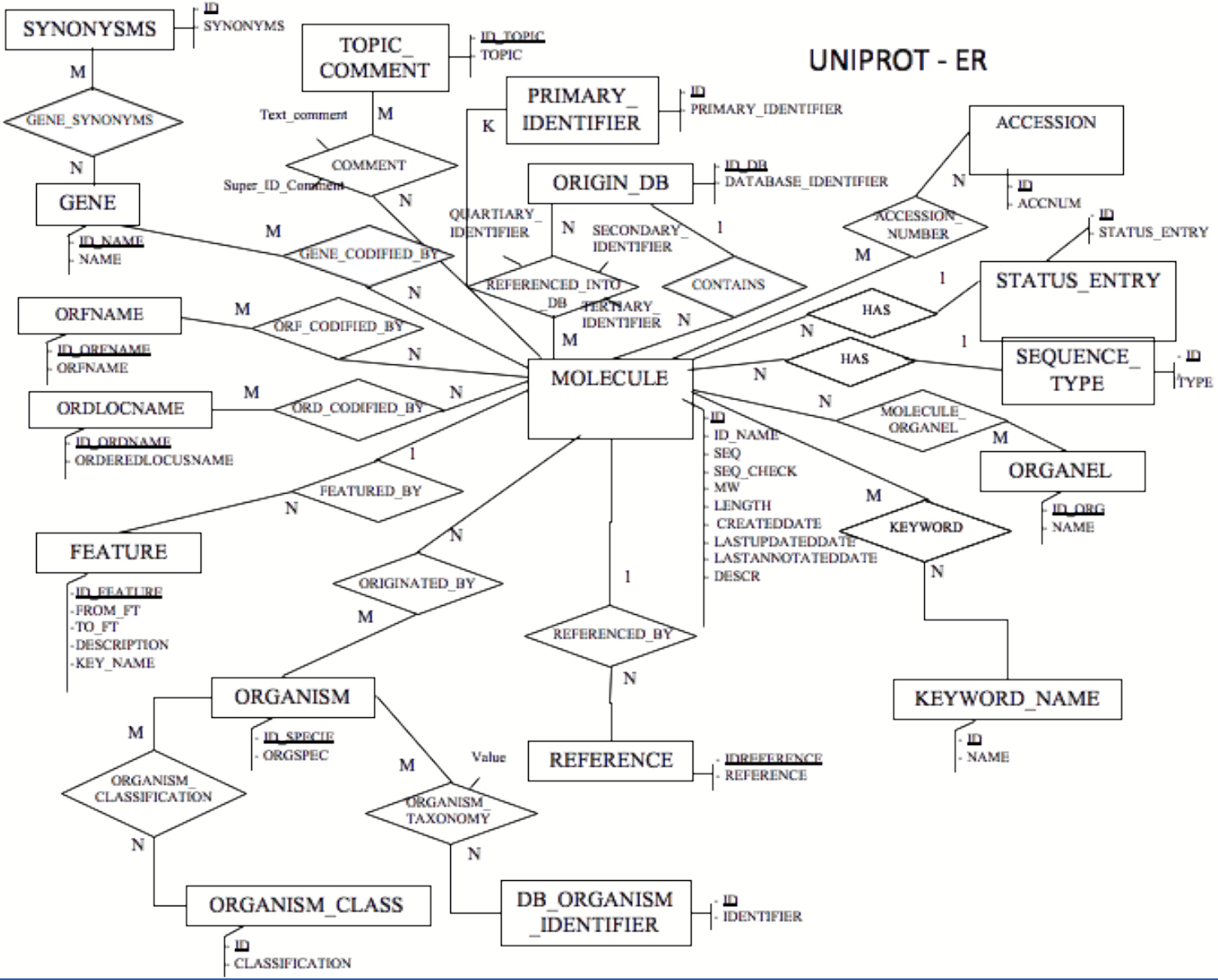
- HUC support activities:
  - **LS**: A GRelC service has been deployed in our site to support LS database management (user support activity). In particular a use case regarding the **UNIPROT data bank** has been implemented
  - **ES**: the **Climate-G Portal** will integrate the **DashboardDB monitoring** facility.
  - **Tutorial** and **training** events (next event scheduled in December at the PDCS2011 conference, Dallas, Texas)
  - Participation in "**user community oriented**" activities (i.e. ES, LS), initiatives and conferences (AGU2011, EGU2011 and EGU2012, etc.)
  - **Project website** and **GILDA tutorials**

HUC Life Sciences Support: the UNIPROT use case

- In Q4-Q5 a **new use case for LS** has been jointly defined with bioinformatics people at the University of Salento. The main goal of this use case was to make the Uniprot database available to the LS community through a GRelC service interface.
- A relational-based schema of the Uniprot database has been designed and implemented.
- An ETL (Extraction-Transformation-Loading) tool to move the data from the Uniprot/Swiss-Prot flat file into a relational DB has been implemented and tested jointly with the bioinformatics group.
- The database schema includes **30 relational tables (13GBs of data)**.
- The relational version of the Uniprot DB has been deployed on the machines provided by SPACI to support these use cases.
- The database allows submitting queries like:
    - **Query 1**: *Given a protein, select the OG (OrGanelle) that indicates if the gene coding for a protein originates from mitochondria, a plastid, a nucleomorph or a plasmid.*
    - **Query 2**: *Given a protein, select the specie, its classification and taxonomy.*

Contact point for this activity: maria.mirto@unisalento.it

HUC Life Sciences Support: the UNIPROT use case



UniProtKB/Swiss-Prot Release 2011_05 of 03-May-2011 30 Tables

Table Name	Num_entry
OriginDB	129
Gene	84260
Organism	13008
ordlocname	381179
orfname	73583
organel	845
topic_comment	40
db_organism_identifier	1
organism_class	8347
synonyms	52157
primary_identifier	3937759
keyword_name	1052
sequence_type	1
status_entry	1
Molecule	526969

Table Name	Num_entry
organism_taxonomy	12463
organism_classification	122209
originated_by	537622
gene_synonyms	56117
accession	694964
accession_number	711407
gene_codified_by	458981
orf_codified_by	76407
ord_codified_by	381473
molecule_organel	20223
comment	2206025
feature	3358177
referenced_into_db	8711973
keyword	3250350
reference	931428

HUC Life Sciences Support: the UNIPROT use case

Advantages

- Reducing the redundancy present into the flat file;
- Reducing the inconsistency of data that could have different values in the flat file;
- More performing searches querying the relational database, by using the GRelC service;
- Complex queries by using a standard language such as SQL.

Next steps

- Taking into account the user requirements, in the near months it is expected to increase the number of biological data banks accessible via the GRelC interface
- The UNIPROT data bank will be published on the DashboardDB registry

User support: the GRelC WebSite

GRelC Website

URL: http://grelc.unile.it/

Mailing List

mail: grelc-user@sara.unile.it

Main sections:

- Download **(rpms available)**
- News
- Publications
- Events
- Deployment
- Documentation
- Components
- …..

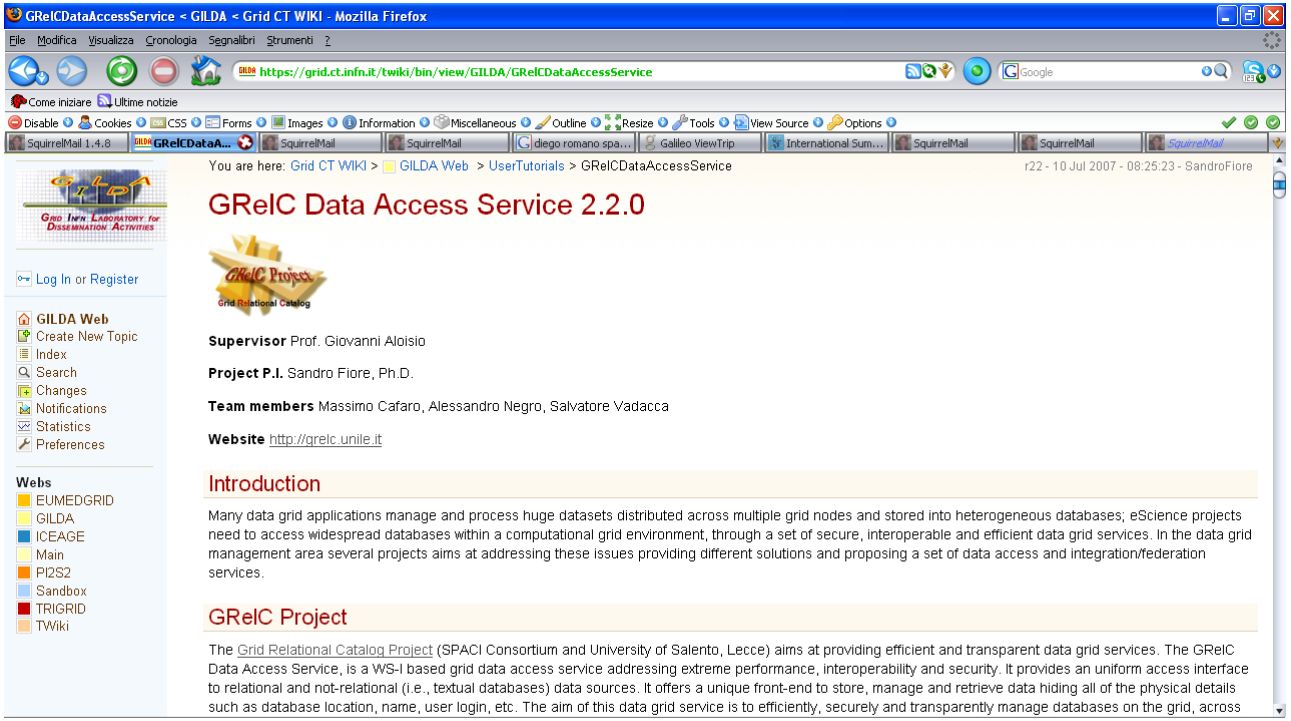
User support: tutorials on GILDA

GRelC DAS User Tutorial

on GILDA Grid CT Wiki Website

Info about:

- Log in to the grid
- Query Submission



For any information about GILDA t-Infrastructure please contact roberto.barbera@ct.infn.it & grid-prod@ct.infn.it

GRelC DAS Tutorial link:

https://grid.ct.infn.it/twiki/bin/view/GILDA/GRelCProject

Special thanks to the GILDA Staff for their support

Some useful information

Fon any information

Project P.I.: S. Fiore (sandro.fiore@unisalento.it)
GRelC WebSite: http://grelc.unile.it
GILDA support: https://grid.ct.infn.it/twiki/bin/view/GILDA/GRelCProject
Mailing lists: grelc-user@sara.unisalento.it

Some useful references

[1] S. Fiore, et al., The Climate-G Portal: The context, key features and a multi-dimensional analysis,, Future Generation Computer System, Vol 28, pp.1-8 (2012), doi:10.1016/j.future.2011.05.015.

[2] S. Fiore, G. Aloisio, Special section: Data management for eScience. Future Generation Computer System 27(3): 290-291 (2011)

[3] S. Fiore, et al., The Data Access Layer in the GRelC System Architecture, Future Generation Computer System, 27(3): 334-340 (2011), http://dx.doi.org/10.1016/j.future.2010.07.006

[4] S. Fiore, et al., The GRelC Project: from 2001 to 2011, ten years working on Grid-DBMSs, in Grid and Cloud Database Management, Springer. Edited by S. Fiore and G. Aloisio.

[5] S. Fiore, G. Aloisio, P. Fox, M. Petitdidier, H. Schwichtenberg, S. Denvil, J. D. Blower, A. Cofino, The Climate-G testbed: towards large scale distributed data management for climate change, Proceedings of the International Conference on Computational Science ICCS 2011, June 1 - June 3, 2011, Nanyang Technological University, Singapore, Procedia Computer Science, Elsevier, pp. 567-576.

[6] S. Fiore and G. Aloisio, "Grid and Cloud Database Management", 2011. Springer, ISBN 978-3-642-20044-1