Supporting the Bioinformatics Community in the OpenAdap.net framework

F. Masulli[1] A.E.P. Villa[2] J. Iglesias[2]

[1] **CNISM, Consorzio Nazionale Interuniversitario per le Scienze Fisiche della Materia, Genova, Italy** and
DISI, University of Genova, Italy
masulli@disi.unige.it

(2) GIN, University J. Fourier, Grenoble, France and
ISD, University of Lausanne, Switzerland

OpenAdap.net

(A. Villa and J. Iglesias)

Open Source project aimed at *breaking the barriers* existing in the flow of *information access* and *information processing.*

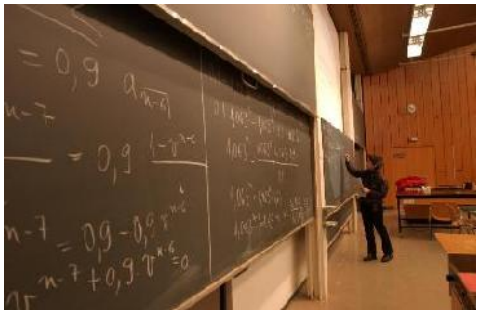
The infrastructure makes it possible the *dissemination* of resources such as knowledge, tools or data, their exposure to *evaluation* in ways that might be unanticipated and hence support the *evolution of communities* of users around a specific domain.

The architecture is designed by analogy with a *virtual distributed operating system* in which the *dynamic resources* are presented as *files* in a structured virtual file system featuring ownership and access permissions.

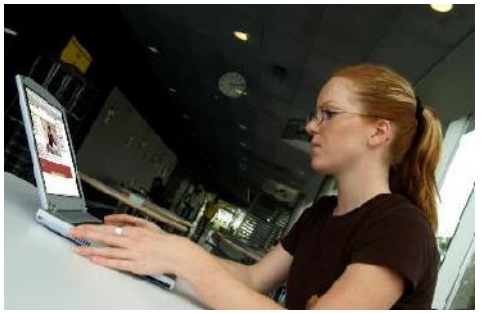
Target Audience

- **COMMUNITIES**: People who share a knowledge representation (common data format) driven by common interests.

- **CONTRIBUTOR**: people who would like to share their knowledge with the Community. They maintain the authorship and keep control and responsibility over their contribution.



- **USER**: people interested to process their own information or access knowledge stored elsewhere (e.g., in a public database) and extract the results of their processing. They exploit Contributors' applications in a trusted way.



Community members who provide new knowledge become able to share their contribution and members who have information to be processed can access these services

Components

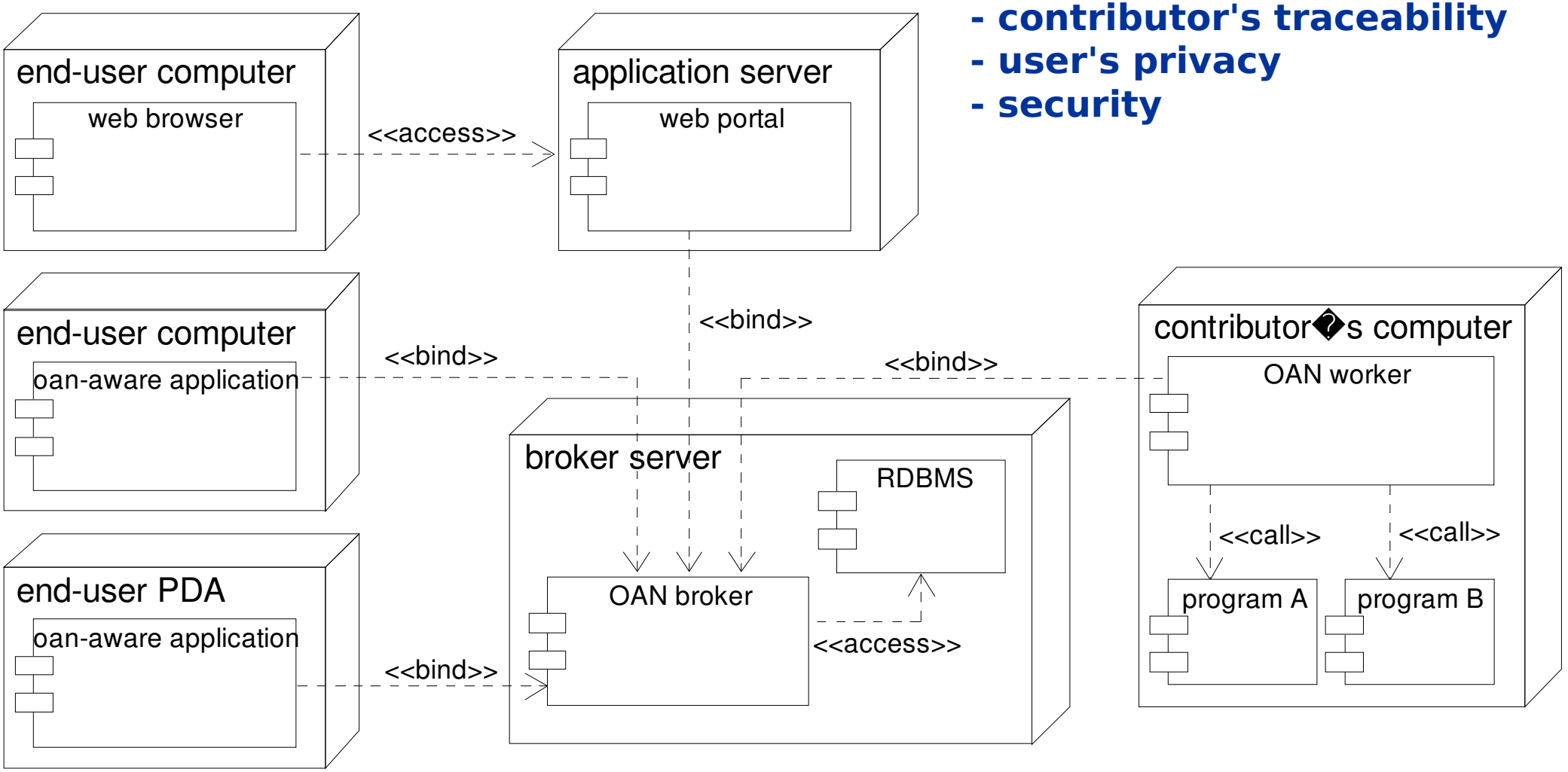
- **Brokers**: processes permanently running on a server, each of them in charge of managing a community of users and dispatching tasks and results on their behalf.

- **Workers**: processes shared by community members in charge of giving secured distant access to contributed resources like programs or data.

- **OAN-aware applications**: pieces of software (standalone applications, web portals, command line tools, etc.) providing access for an end-user to the community shared resources through identified connections to a broker.

Methodology

- **contributor's traceability**
- **user's privacy**
- **security**



UML DEPLOYMENT SCHEME

Software Architecture

	Data treatment istribution	Hardware resource allocation	Hidden execution hosts	Application sharing	Published application interface	Data sharing	Highly dynamic system	Transparent user/resource connection
Grid	×	×	×					
WS		×	×	×	×			
P2P					×	×	×	
OAN	×	×	×	×	×	×	×	×

- **Grid**: Each user has a large dataset to manipulate with one application distributed on a set of computers.
- **Web-services**: Many users exploit the same services permanently provided trough a centralized authority.
- **Peer-to-Peer**: Many users exchanging pieces of data in an unsupervised way

OAN Aims

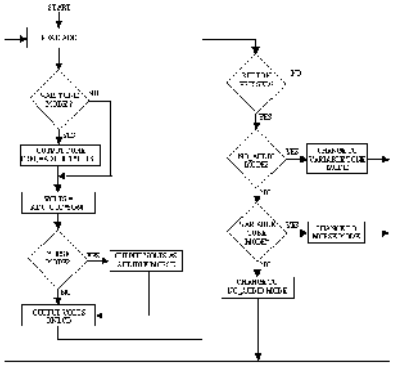
OpenAdap.net falls somewhere among these three architectures but with the intention to address a double-face problem:

- to provide to a community of users in the same domain a means to *interchange their resources* in an open, dynamic and secured way, and
- to provide to a community of users the access to the *exploitation* of information processing solutions contributed by users belonging to *other communities*.

contributor: joining the Community

CONTRIBUTOR

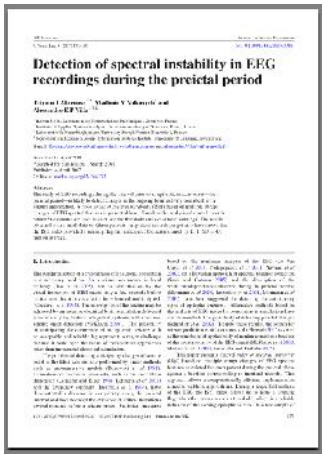
PROGRAMMING FLOWCHART



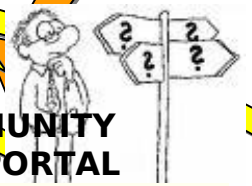
EDITING SOURCE CODE

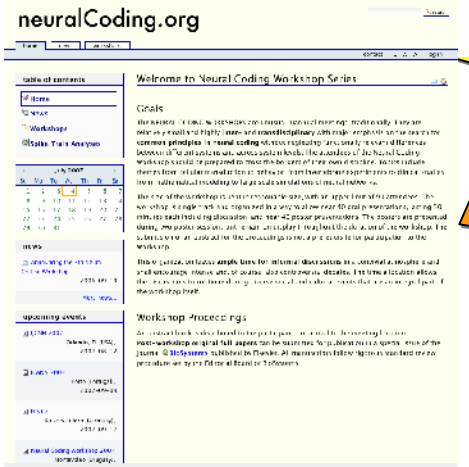
RUNNING THE APPLICATION

PUBLISHING

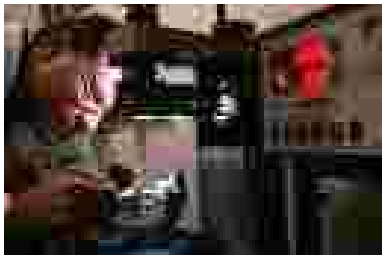


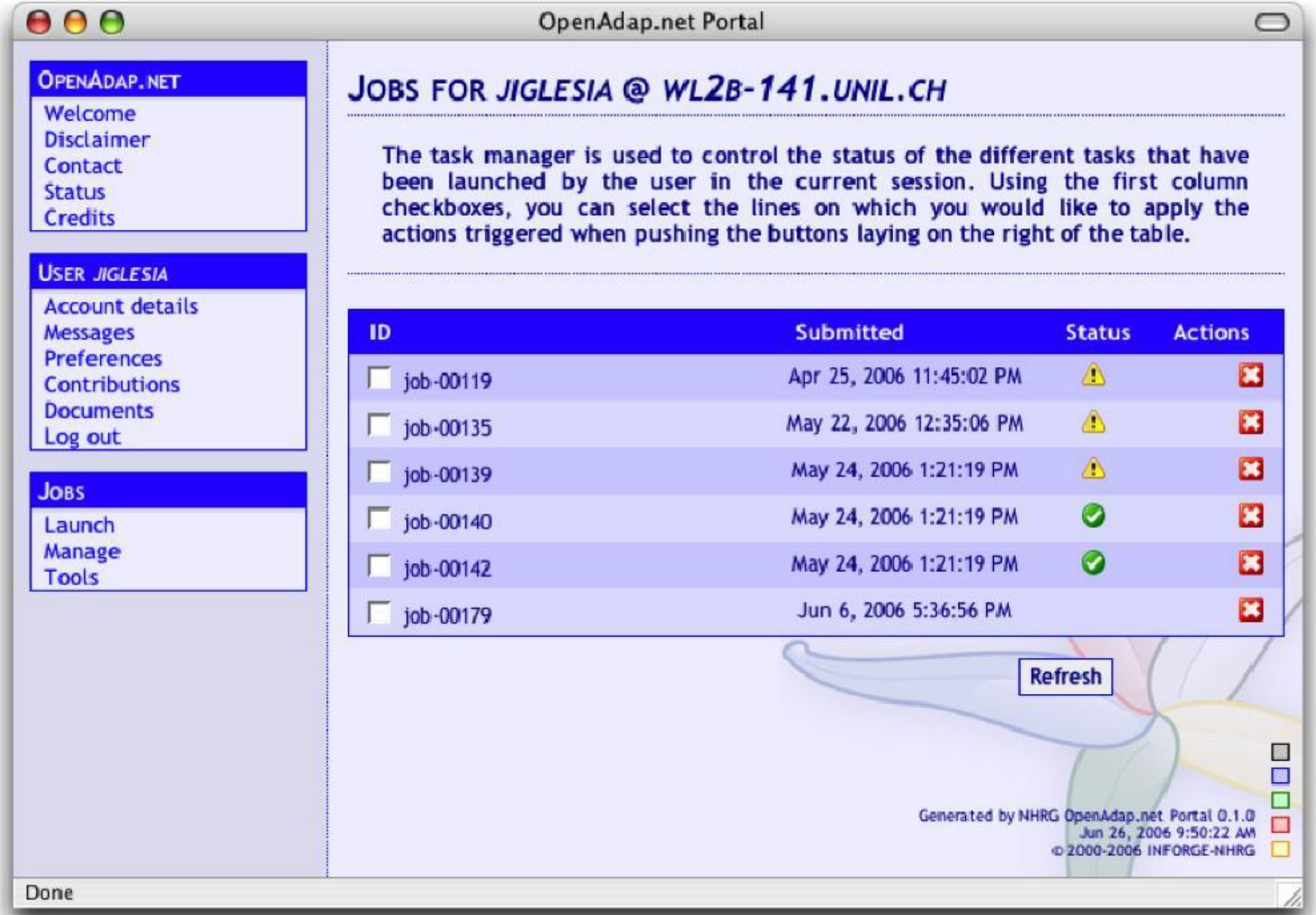
COMMUNITY WEB PORTAL





USER INTERESTED IN THE APPLICATION




OpenAdap.net Portal
OpenAdap.net
Welcome
Disclaimer
Contact
Status
Credits
User jiglesia
Account details
Messages
Preferences
Contributions
Documents
Log out
Jobs
Launch
Manage
Tools
Jobs for jiglesia @ wl2b-141.unil.ch
The task manager is used to control the status of the different tasks that have been launched by the user in the current session. Using the first column checkboxes, you can select the lines on which you would like to apply the actions triggered when pushing the buttons laying on the right of the table.
ID
Submitted
Status
Actions
job-00119
Apr 25, 2006 11:45:02 PM
job-00135
May 22, 2006 12:35:06 PM
job-00139
May 24, 2006 1:21:19 PM
job-00140
May 24, 2006 1:21:19 PM
job-00142
May 24, 2006 1:21:19 PM
job-00179
Jun 6, 2006 5:36:56 PM
Refresh
Generated by NHRG OpenAdap.net Portal 0.1.0
Jun 26, 2006 9:50:22 AM
© 2000-2006 INFORGE-NHRG
Done

- Inter-Broker Communication

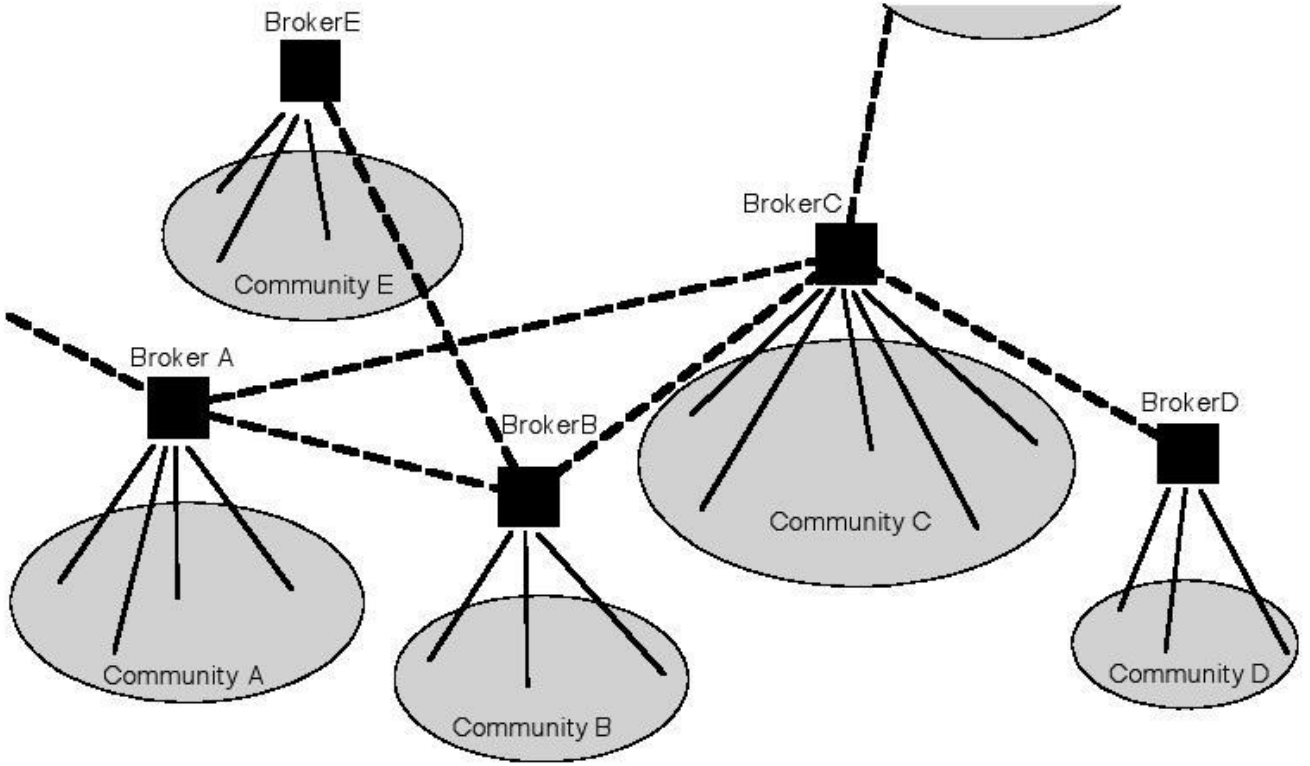


Figure 1: OpenAdap.net communities (grey ellipses) virtually laid down in an ontological space organized around their domain-oriented brokers (black squares). Intelligent inter-broker communications (interrupted lines) allow the contextualized sharing of community resources across the semantic barriers.

- Ontologies

- pervasive access
- trust of resources: dynamic resoources characterization
- user profile: dynamic users characterization

Initially composed by a default set of resources the system will learn and adapt user's preferences.

With an increase in user's activity within the Community, the OpenAdap.net broker will be able to autonomously evolve and suggest the most relevant resources, on the basis of user's profile and the semantic content of the different Interconnected brokers.

Discussion/Perspectives

- To avoid the "re-invention" and "re-discovery" of existing knowledge so to save time and expenses by the whole society and prevent incorrect applications.

- To promote the development of Communities with semantically related interests.

- To provide Communities the possibility to transparently compose meta-resources, based on the available resources, by means of a network-centric operating system driven by the activity of intelligent adaptable brokers.

- To develop a new OpenAdap.net protocol (oan://) for field instrumentation and wireless communication over Internet aimed at easy ubiquitous knowledge sharing and access.

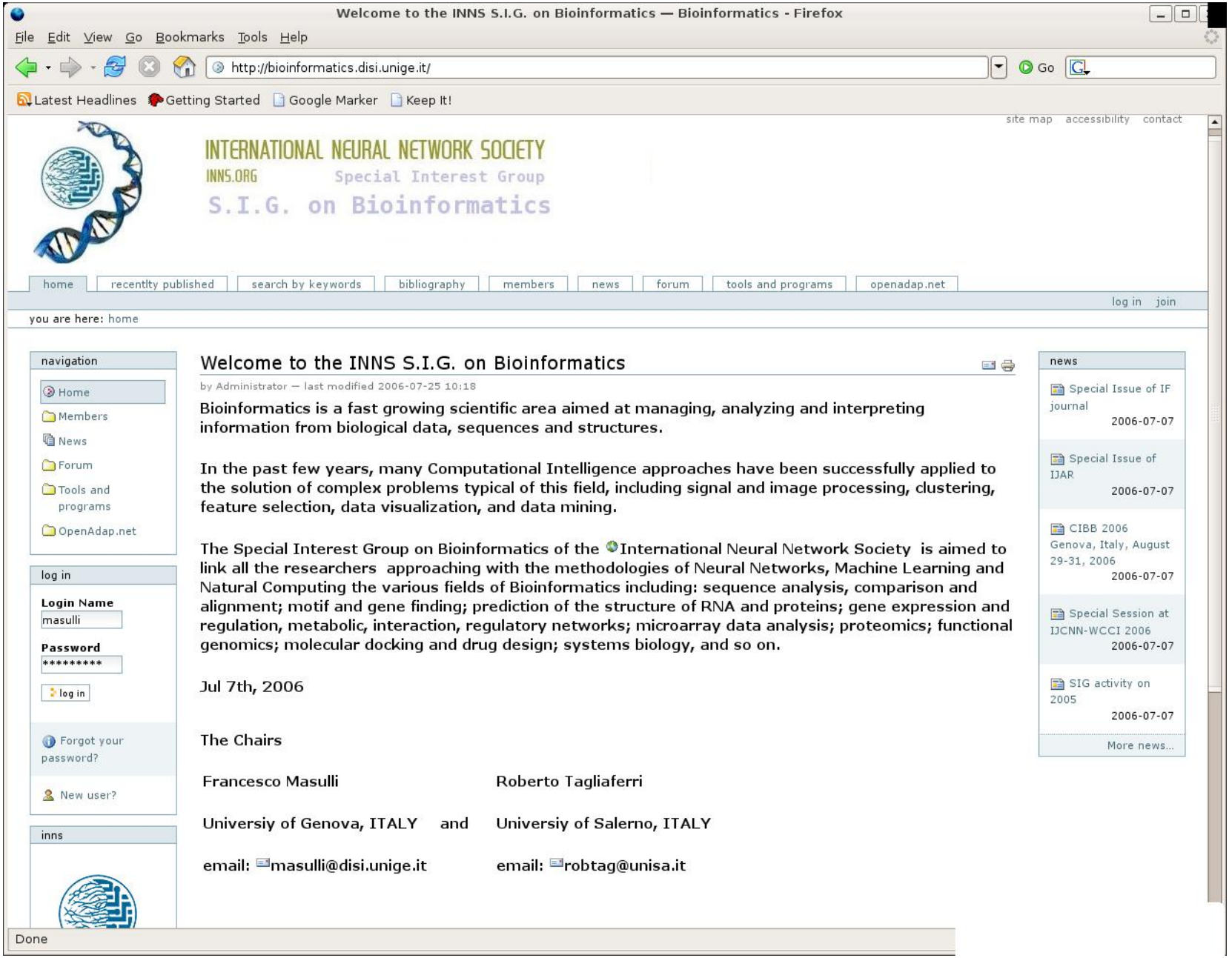
- To develop new business opportunities for Third Parties from all over the world, SMEs in particular, aimed at added value services (educational, commercial, governmental, ...).

Available Resources for Bioinformatics Community

Nowadays, in the Post-Genomic era, we have **many Bioinformatic data sets available** (most of them released in public domain on the Internet) and **large computing infrastractures.**

The information embedded in many Bioinformatic data sets has no yet completely exploited, due to the lack of accurate **computing tools** (including machine learning tools) and/or of their **diffusion** in the Bioinformatics community.

A portal for the Bioinformatics Community



A portal for the Bioinformatics Community

- http://bioinformatics.disi.unige.it/
- to be released on March, 2008
- contributions by
  - Stefano Rovetta
  - Davide Chicco (student)
  - Christian Fornaciari (student)
  - ...

A portal for the Bioinformatics Community

- Broker on a Virtual WEB server
- Worker on a small Cluster of 4 AMD (2 GHz, 64 bits) under Rocks cluster distribution http://www.rocksclusters.org
  - Rocks is an open-source Linux cluster distribution intended for high-performance computing clusters.
  - Rocks is now based on CentOS and includes many tools (such as MPI) which are not part of CentOS but are integral components that make a group of computers into a cluster.
  - Installations can be customized with additional software packages at install-time by using special user-supplied CDs (called "Roll CDs")

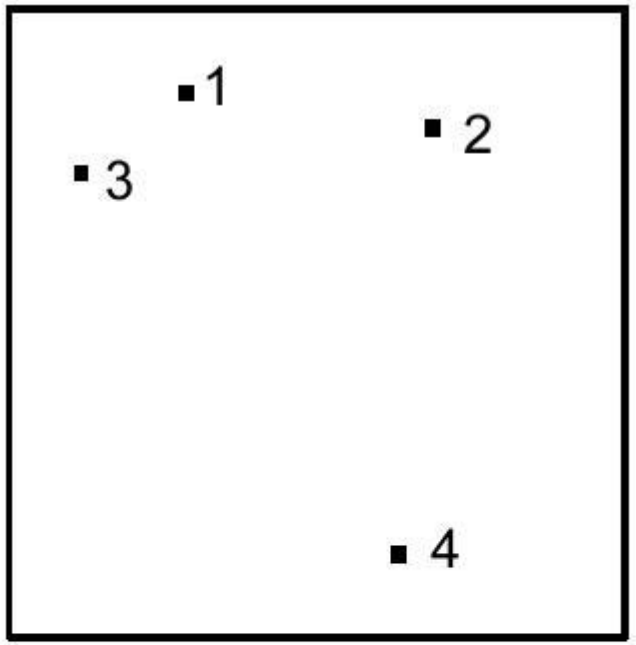
Proof of concepts

Porting on OpenAdap.net of the **Shared Farthest Neighbor clustering** (SFN) **algorithm**

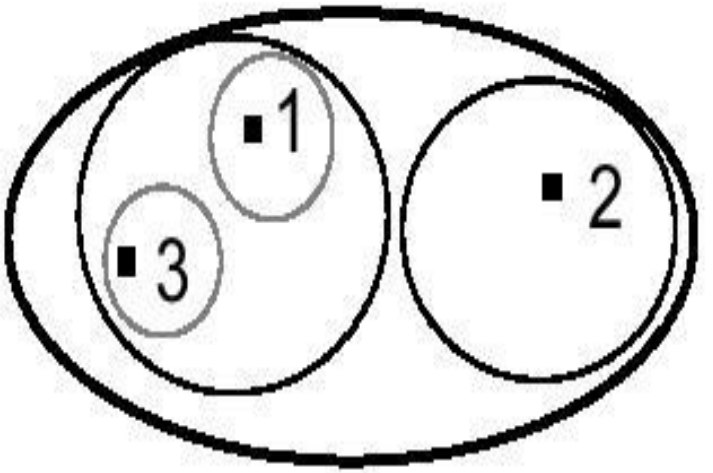
Rovetta, F. Masulli, "Shared farthest neighbor approach to clustering of high dimensionality, low cardinality data", Pattern Recognition, 39, pp. 2415-2425, 2006

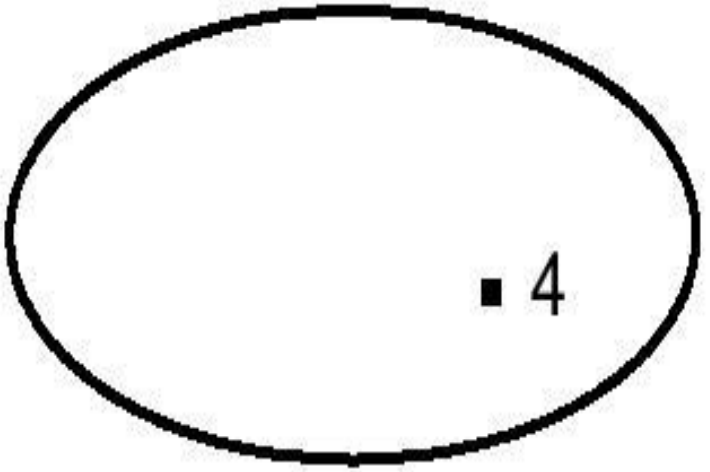
Available as Open source at http://mlsc.disi.unige.it/downloads/C/sfn/

Shared Farthest Neighbor clustering algorithm



Data points	1	2	3	4
I Neighbor	3	1	1	3
II Neighbor	2	3	2	2
III Neighbor	4	4	4	1





"*Points in perspective*" principle

Actual development

- Porting in OpenAdap.net of the Rocks Bioinformatics roll called **Bio Roll** containing a suite of Bio-informatics applications, most commonly in use by the bio-informatics community.
- The list of applications is as follows:
  - **HMMER** - From Washington University at St. Louis - http://hmmer.wustl.edu/
  - **NCBI BLAST** - From National Center for Biotechnology Information www.ncbi.nlm.nih.gov/BLAST/
  - **MpiBLAST** - From Los Alamos National Laboratory http://mpiblast.lanl.gov/
  - **biopython** - www.biopython.org

- **ClustalW** - From the European BioInformatics Institute - http://www.ebi.ac.uk/clustalw/
- **MrBayes** - From School of Computational Science at the Florida State University - http://mrbayes.csit.fsu.edu/
- **T_Coffee** - From Information Genomique et Structurale at Centre National de la Recherche Scientifique
- **Emboss** - From European Molecular Biology Institute - http://emboss.sourceforge.net/
- **Phylip** - From the Dept. of Biology at the University of Washington - http://evolution.genetics.washington.edu/phylip.html
- **fasta** - From the University of Virginia - http://fasta.bioch.virginia.edu/
- **Glimmer** - From Center for Bioinformatics and Computational Biology at the University of Maryland http://www.cbcb.umd.edu/software/glimmer/
- **perl-bioperl**, **perl-bioperl-run**, **perl-bioperl-gui** - From CPAN

Conclusions / Future Work

- Starting the Bioinformatics Community in the in the OpenAdap.net framework
- We plan to port on OpenAdap.net other algorithms developend by our group
- GMU Univ. Fairfax, Virginia USA is starting to develop another broker for the Bioinformatics community giving access to their applications and to NIH Bioinformatic data sets
- Extending the collaboration to other Contributors