PORTAL group "Introduction and usecase: the GPSA Bioinformatics portal"

Christophe Blanchet
Institut de Biologie et de Chimie des Protéines
CNRS IBCP
Lyon – Gerland, France
Christophe.Blanchet@ibcp.fr

PORTAL meeting 3
EGEE07, 4 oct., 2007, Budapest

www.eu-egee.org

What is a portal ?

- A portal is a Web-based application that provides personalization, single sign-on, and content aggregation from different sources, and hosts the presentation layer of information systems. Aggregation is the process of integrating content from different sources within a Webpage. A portal may have sophisticated personalization features to provide customized content to users. Portal pages may have different sets of portlets creating content for different users. *(javaworld.com)*

EGEE07 , Session "Uniform Grid Access"
GridSphere presentation

What the PORTAL group

- **is NOT !**
  - a study of the portal technology
    - Apache, Zope, Gridsphere, Websphere, EngineFrame, Tomcat, ...
  - a study of the CMS
    - Drupal, Joomla, Plone, Trac, ...
  - a study of the language
    - html, java, javascript, php, css, python, ajax, ...
  - a study of the performances of portal tools
    - time to get a page, number of simultaneous connected users, ...

What the PORTAL group IS ?

- **TCG working group**
  - chair: C. Blanchet
  - http://egee-intranet.web.cern.ch/egee-intranet/NA1/TCG/wgs/portal.htm
- **Topics:**
  - How a portal could connect to the Grid: authN
  - What could it be allow to do : authZ
  - In the name of who : what are the kind of portal users
  - Who would be responsible
  - Where could it be allowed : agreement of sites
  - What is the reliability of portal : intrinsic security against attack
  - Which technology : now and in the future

Goals

- **“état de l’art”**
  - requirements from application portals
  - requirement from sites
  - requirement from the grid platform in a whole
  - available technology
- **Define common diagram**
  - categorize users
  - categorize usecases
  - domains of application: data and tools
- **Best practices**
  - recommendations
  - policy agreement

Guidelines

- **Some questions to answer:**
  - How does one obtain service certificates from the EUGridPMA CAs?
  - How to register those certificates in a VO?
  - How to handle proxies of this certificate in a portal (e.g. wanting to add VOMS groups)?
  - Which software should be allow to be used ?
    - enabling users to put his own software
    - restrict to only portal ones
  - Do sites need to be advised of the use of these certificates (if so, how to inform them)?
  - What are the accounting/logging practices to follow on the portal?
  - Does the use of these need to appear in the VO AUP or similar documents?

Actions

- **May 2007 - May 2008**
- **Meeting 1, May, 2nd-3rd, Cern**
  - http://indico.cern.ch/conferenceDisplay.py?confId=15737
- **Meeting 2, June, 28th-29th, IBCP**
  - http://indico.cern.ch/conferenceDisplay.py?confId=17615
- **External presentation**
  - EGEE Bioinformatics meeting 3 (Valencia, March 6th, 2007)
    - great interest from the Bioinformatics community
  - EGEE JPSG meeting, March, 13rd-14th, Cern
    - good interest on the site operation and the policy aspects
- **Collaboration with other projects**
  - SwissBioGrid: portal requirement, technology
  - GridPP: portal requirement, technology
  - BioinfoGrid: portal requirement

Current status

- **“état de l’art”**
  - requirements from application portals
    - GPSA (LS), GridPP (HEP), BioinfoGrid (LS), SIMRI (LS)
  - requirement from sites
    - CNRS LAL
  - requirement from the grid platform in a whole
    - JSPG
  - available technology
    - VOMS / AMI integration (GridPP)
    - Shibboleth, SLCS and VOMS (SWITCH, CNRS UREC)
      - *Short Lived Credential Service (SLCS)*
      - *VOMS Attributes from Shibboleth (VASH)*

Common diagram

- **Categories of users:**
  - 1) End user is "anonymous" (i.e. neither grid credential nor registration).
  - 2) Pseudo-anonymous users (portal registration)
  - 3) Identified users w/o grid credentials (maybe other certificates).
  - 4) Identified users with grid credentials; portal credentials used.
  - 5) End user has his/her own grid credentials.
- **Usage of tools and data**
  - pre-defined software (avoid “unidentified” DOS)
  - let user upload their applications : only in case 5
  - data transfers through/by portal: only in case 5 (except user paramaters)
  - For example
    - Guest users should go through a mandatory page where they must acknowledge data privacy and tool licence respect.

DPK thoughts on Portals policy issues

- We need a matrix for the 5 use cases
- Per use case
  - Is Grid job well-known (canned job) or free?
  - Is Grid job executed under Portal ID or User ID?
  - Is Grid data well-known (canned data) or free?
  - Is Grid data owned by Portal ID or User ID?
- For User ID
  - Is user well identified (photo ID)?
  - Or just username/email-address and password?
  - Does user ID need to be transferred to the Grid
- Need to address accounting issues (as discussed earlier)
  - And data privacy issues
- Written policies are very important (to protect Grid and Portal)

PORTAL3 Agenda

- http://indico.cern.ch/sessionDisplay.py?sessionId=74&slotId=0&confId=18714#2007-10-04
-

Introduction and usecase: the GPSA Bioinformatics portal
by Dr. Christophe BLANCHET (CNRS IBCP)
(Rege II: 17:30 - 17:45)

[370] **Portals and Authentication**
by Mr. David GROEP (NIKHEF)
(Rege II: 17:45 - 18:00)

[371] **JSPG contribution**
by Dr. David KELSEY (RAL)
(Rege II: 18:00 - 18:15)

[372] **Web application security**
by Mr. Romain WARTEL (CERN)
(Rege II: 18:15 - 18:30)

[373] **Shibboleth and Grid Portals**
by Dr. Christoph WITZIG
(Rege II: 18:30 - 18:45)

[374] **Discussion**

Portal Authentication (D. Groep)

- **Issues**
  - assure approximate parity (~60 CAs with policy 6-7 pages)
  - namespaces for name overlaps
  - a forum to participate and raise issues
  - operation of a secure collection point about CAs which you accredit
  - common practices where possible
- **Users vs Hosts vs robots**
- **AuthN landscape for portal**
  - Service: infrastructure may not accept it
  - User: regularly provide the passphrase
  - Robots certs (example of NGS portals)
    - where available (UK, NL, soon CZ) these are the preferred choice
    - protects private key from abuse outside the portal system
  - alternative: MyProxy, Federation backed SLCS

JSPG policy issues

- Current Grid security audit requirement
  - Able to trace all actions to an well-identified individual person
  - Assurance of Identity is high (photo-id or equivalent)
  - Sharing of identities not allowed
    - Illegal/problem actions or data traceable to a person
- AUP
  - User accepts Grid AUP on registration with VO
  - VO AUP has to exist
    - And user also accepts this on registration

Portal and Shibboleth (V. Tschopp)

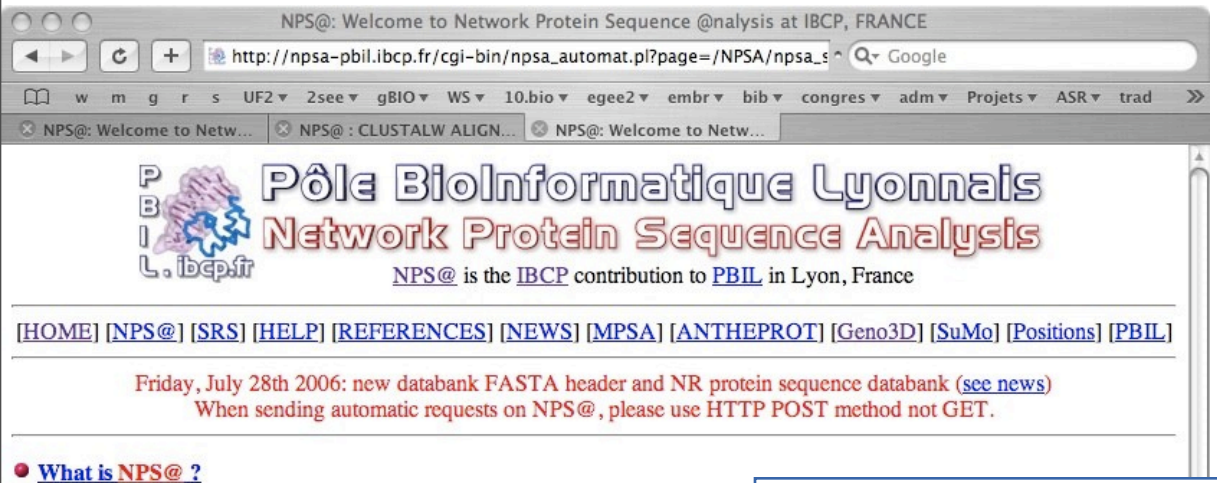
- **Shibboleth**
  - Federated Identity
    - Based on SAML (Security Assertion Markup Language)
  - Web resources SSO (Single Sign-On)
  - Open Source, Developed by Internet2
- **Advantages**
  - Integrate existing components in Portal
    - Reuse Shibboleth, SLCS and VOMS
  - Leverage existing Identity Management Systems
    - Semi-automated users management in Portal
  - User friendly
    - Same credential as usual
    - No certificate problem anymore
- **Growing coverage of Shibboleth based federations**
  - Finland, France, Switzerland, UK, US, ...

Map

- **Draft document**
  - “état de l’art”
  - Define common diagram
  - Best practices
  - + PORTAL policy in a separate document ?
- **Prototype**
  - Bioinformatics application: GPSA (CNRS IBCP)
  - portal: gridsphere (dev. team)
  - authN/authZ: Shibboleth and ... (Switch)
- **Roadmap**
  - next meeting in Feb 2008
  - place to be defined
  - draft documents and prototype available

Connecting portal to the Grid : Usecase of the GPSA Bioinformatics Portal

NPSA: Bioinformatics Web Portal

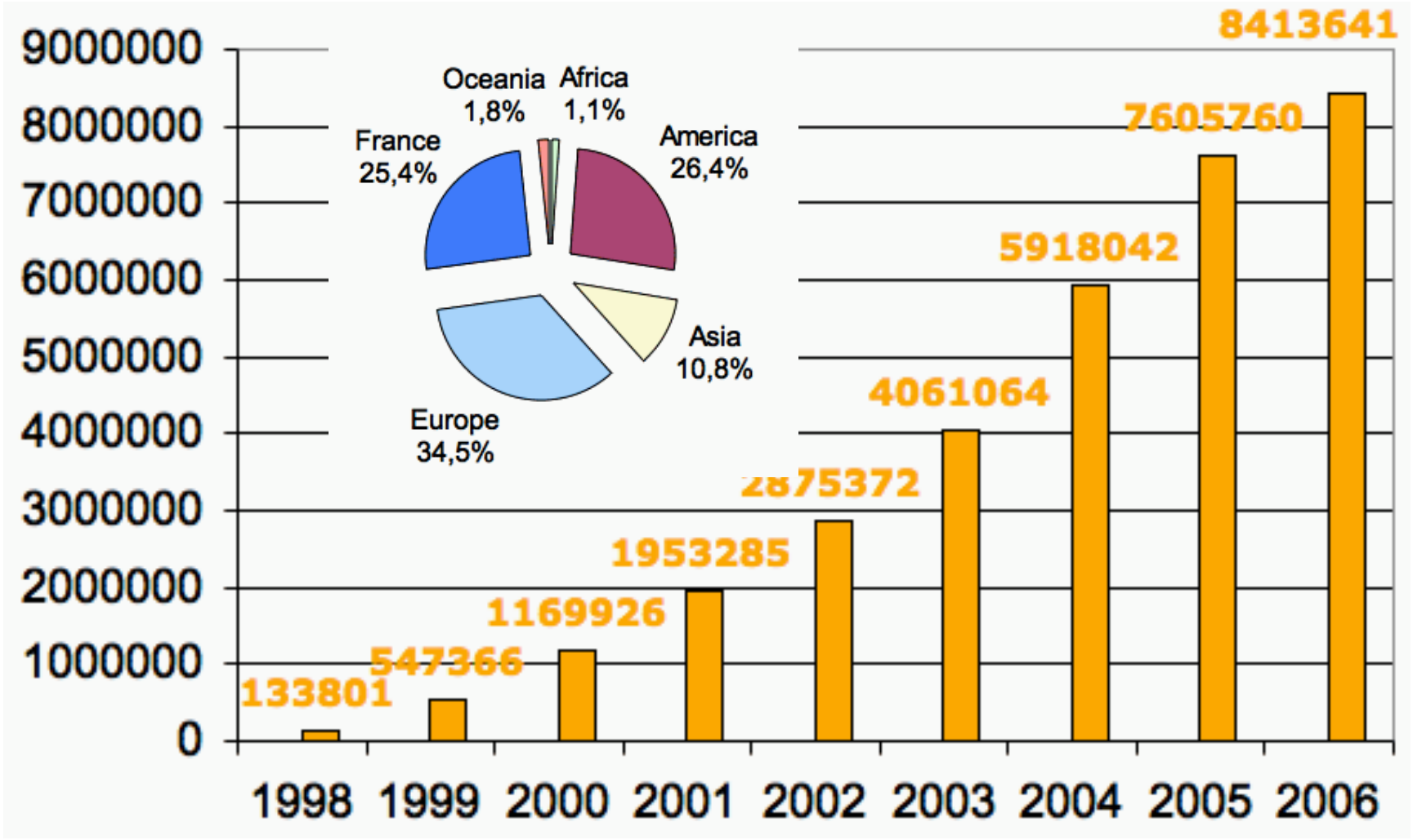


- Network Protein Sequence Analysis (NPS@ release 3)
  http://npsa-pbil.ibcp.fr
- Online since 1998
- 46 integrated methods for protein sequence analysis
- 12 Online up-to-date biological databanks

- *" NPS@: Network Protein Sequence Analysis", Combet C., Blanchet C., Geourjon C. et Deléage G. Tibs, 2000, 25, 147-150.*

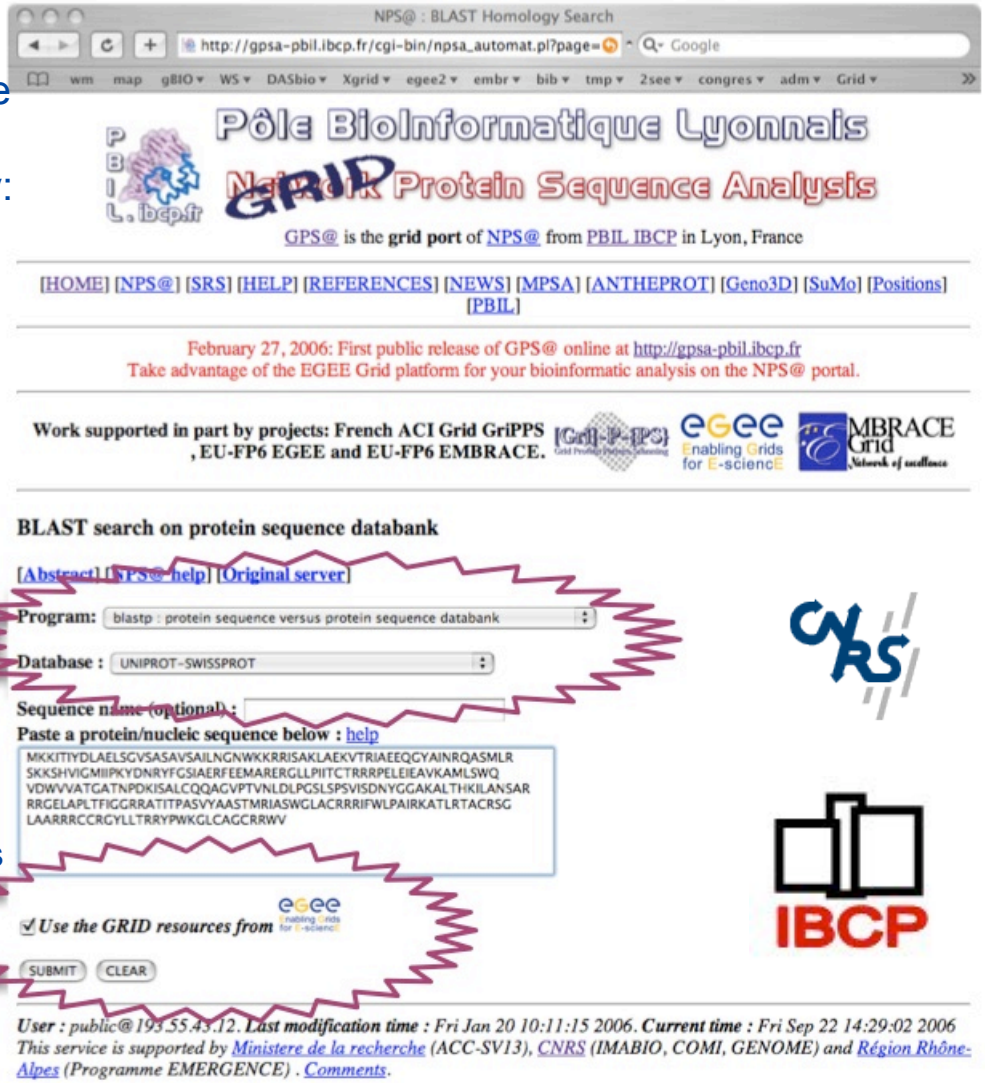
NPSA hits

- More than 10 millions analyses since 1998
- 5000 analyses/day



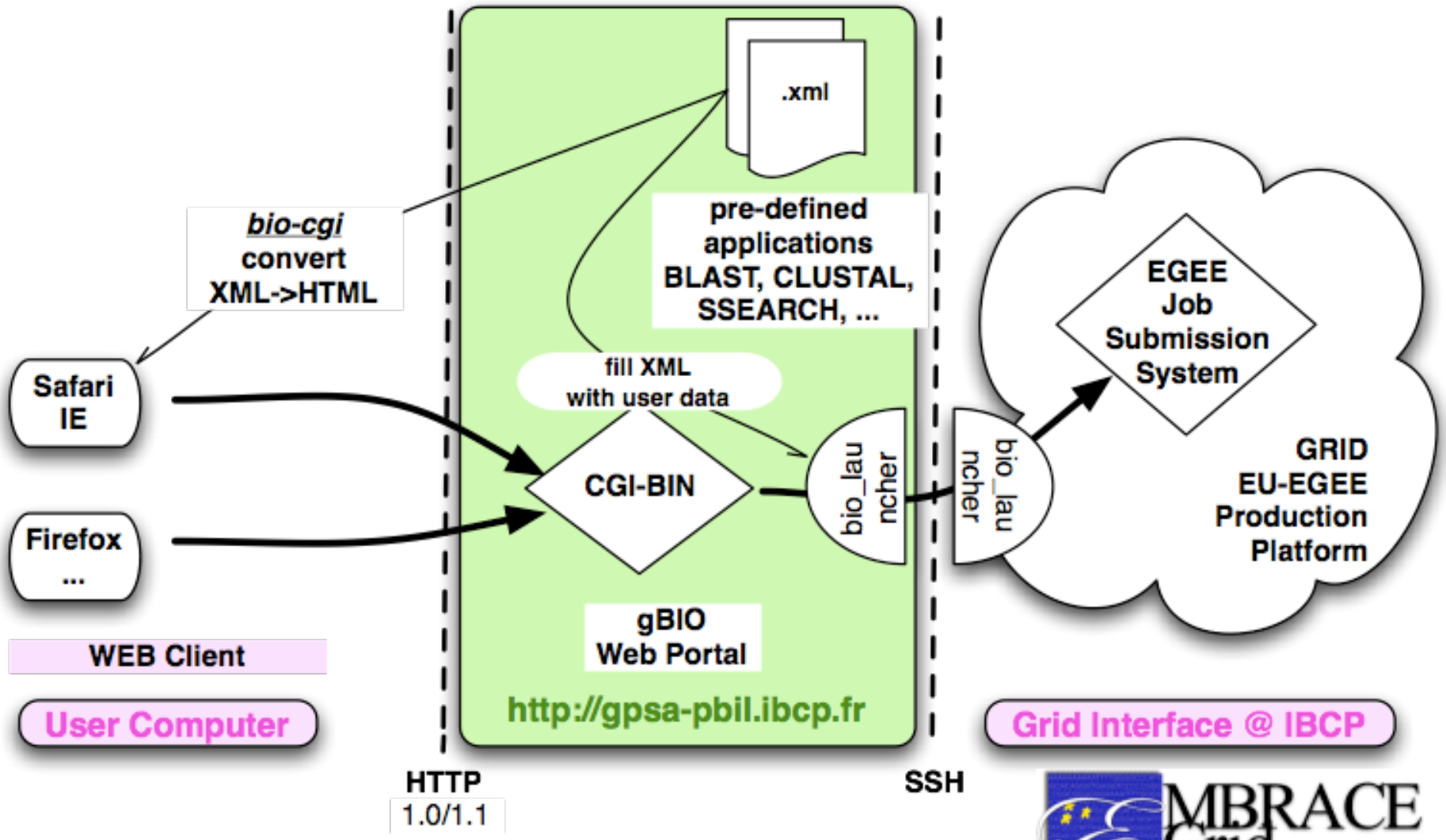
GPSA: Bioinformatics Grid Portal

- **Scientific objectives**
  - Molecular Bioinformatics: protein sequence analysis
  - Analyze data from high-throughput Biology: complete genome projects, EST, complete proteomes, structural biology, ….
  - Integration of biological data and tools
- **Method**
  - Provide Biologists with an usual Web interface: NPS@
    - § NPS@ Web portal online since 1998
    - § 46 tools & 12 updated databases
    - § + 9,000,000 jobs & 5,000 jobs/day
  - Ease the access to updated databases and algorithms.
    - § Protein databases are stored on the grid storage as flat files, encrypted if needed.
    - § Wrapping legacy bioinformatics applications
    - § Transparent remote access through local file-system accesses
  - Display results in graphical Web interface.
- **Status:** Prototype
- **Contact**: Christophe Blanchet@ibcp.fr



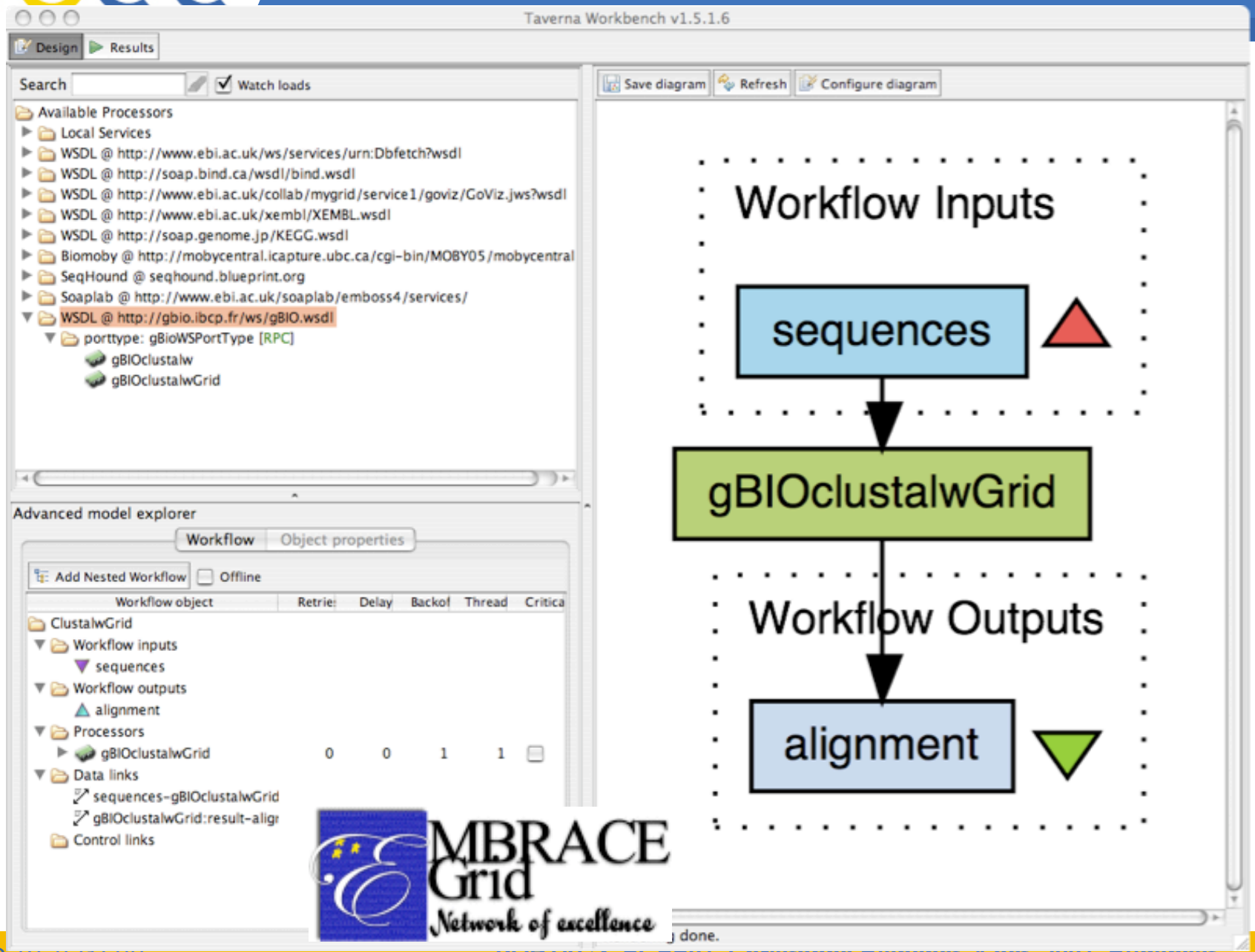
Christophe.Blanchet@ibcp.fr - http://gbio-pbil.ibcp.fr

Computation Virtualization: Web Portal

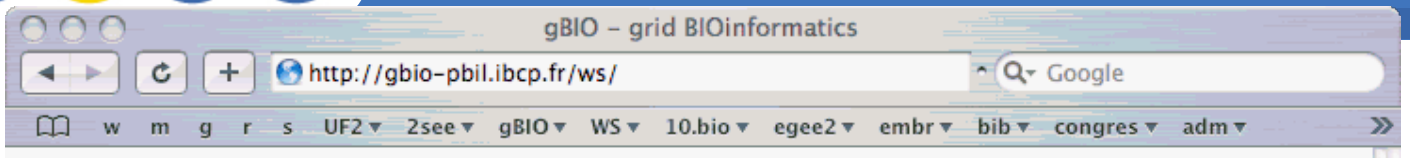


MBRACE Grid
Network of excellence

Grid WS of ClustalW



Bioinformatics WS on GRID



Grid activity of the PBIL-IBCP bioinformatics lab

Welcome at the Web Services of IBCP. We are providing Web services compatible to W3C standards (SOAP, WSDL, HTTP).

The available Web services are providing you with bioinformatics tools that you can call remotely by using a compatible client: Taverna, Triana, or your own SOAP client build with C/C++ gSOAP, perl SOAP::Lite or Java.

- List of Web Services available at IBCP:

  Global WSDL : http://gbio.ibcp.fr/ws/gBIO.wsdl

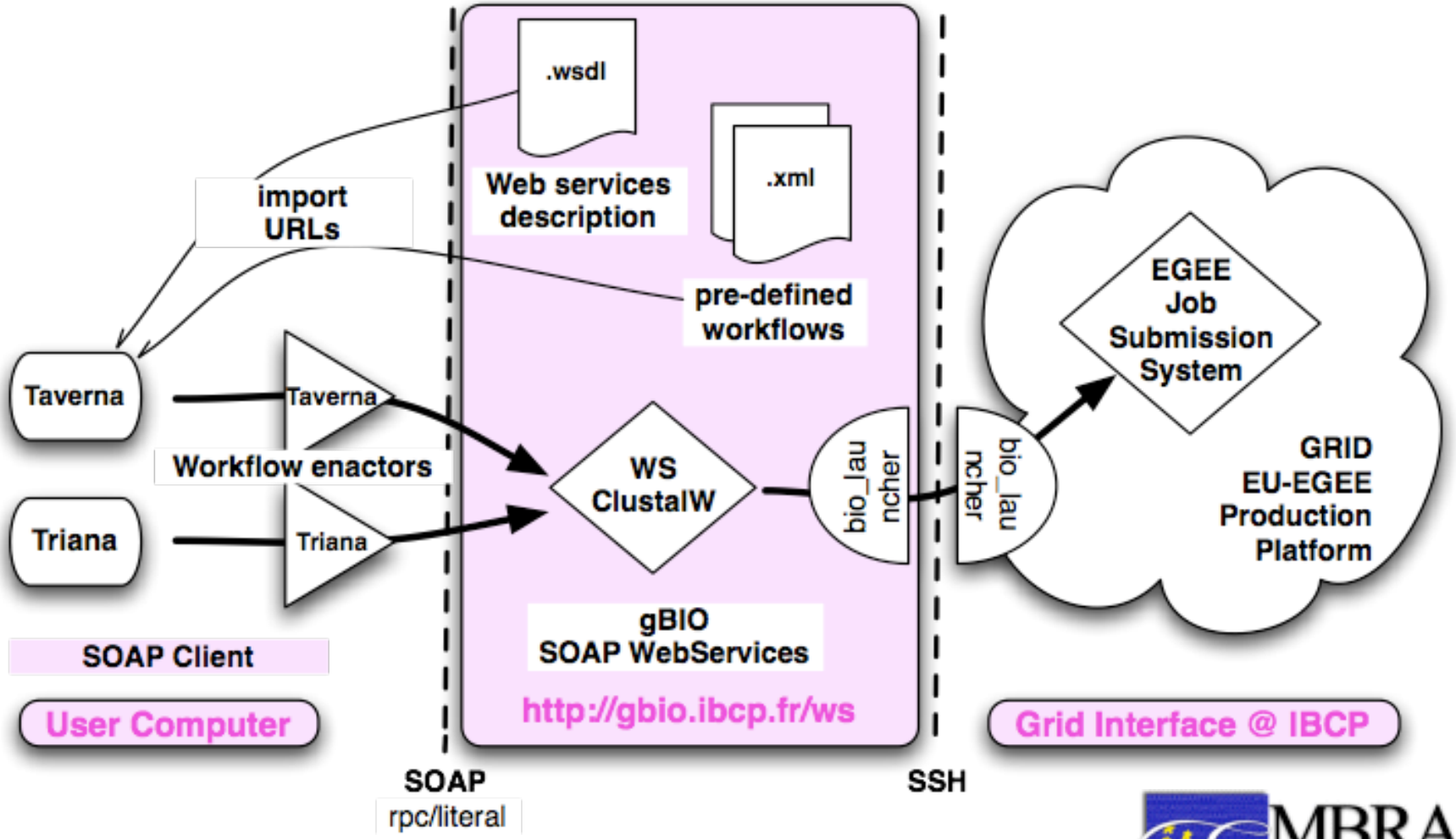
Service	Description
gBIOclustalw	ClustalW (run on local server resources)
gBIOclustalwGrid	ClustalW (run on GRID)

- List of pre-defined workflows available from IBCP:

  To use them, paste one URL below in the load field of your workflow enactor

Workflow	For Taverna
ClustalW	http://gbio.ibcp.fr/wf/Clustalw.xml
ClustalW on Grid	http://gbio.ibcp.fr/wf/ClustalwGrid.xml

Grid usage through Web Services



Christophe Blanchet, Christophe Combet, Vladimir Daric and Gilbert Deléage
Web Services Interface to run Protein Sequence Tools on Grid, Testcase of Protein Sequence Alignn
Lecture Notes in Computer Science : Biological and Medical Data Analysis (2006) 4345, 240-9

Conclusion

- **Usecase of the GPSA Bioinformatics Portal**
  - NPSA: a Bioinformatics Portal in action since 1998
  - Virtualize bioinformatics data and tools on Grid
  - “Ease of use” of Web interface
    - Portal: GPSA
    - Web Services : gBIO
- **Features**
  - non-authenticated access
  - tools can not be changed by users
  - data:
    - main ones are managed by portal admin
    - user data
  - Technology
    - Apache
    - HTML + CGI scripts: perl, python.