Automation of bioinformatics processes through workflow management systems

Paolo Romano

Bioinformatics
National Cancer Research Institute of Genoa, Italy
paolo.romano@istge.it

Summary

- Information and data integration in biology
- A methodology for automation of processes
- Features and limitations of workflow management systems
- Biowep: a workflow enactment portal

Information in biology: well known facts

Biomedical research produces an increasing quantity of **new data** and **new data types**

- Genomics is producing an immense quantity of data
- Emerging domains, like mutation and variation analysis, polymorphisms, metabolism, as well as new high-throughput technologies, e.g., microarrays, will also contribute with huge amounts of data
- Analysis software must interoperate with databases
  - Databases as input for software
  - Results as new data to store and analyze

Information in biology: some figures

EMBL Data Library 88 (Sep 2006)

Sequences: 80,591,891 (Bases: 146,595,277,574)
Increase: +8,86% (+8,91%) since previous release
Increase: +37,16% (+36,297%) since Sep 2005
(http://www3.ebi.ac.uk/Services/DBStats/)

ArrayExpress (31/10/2006)

Experiments 1,739 (ca 140 Gb)
Increase +100% from October 2005 to October 2006
(http://www.ebi.ac.uk/arrayexpress/Help/stats/index.html)

Nucleic Acids Research Supplement (2006)

858 molecular biology databases
(http://nar.oxfordjournals.org/cgi/content/full/33/suppl_1/D5/DC1)

SRS public sites (17/11/2006)

1,300 libraries
(http://downloads.lionbio.co.uk/publisrs.html)

Heterogeneicity of databanks

- A few dbs are managed in a homogenous way (nucleotide sequences at EBI, NCBI, DDBJ)
- Secondary databases are of the highest quality (good and extended annotation, quality control)
- Many databases are highly specialized, e.g. by gene, organism, disease, mutation
- Many databases are created by small groups or even by single researchers

- Databanks are distributed:
  - Different DBMS, data structures, query methods
  - Same information, different syntax and semantics

Goals of the integration

In this context, data integration and process automation are needed to:

- Automatically carry out analyses and/or searches involving more databases and software
- Effectively perform analyses involving large data sets
- Achieve a better and wider view of all available information
- Carry out a real data mining and discover new data

This can be done by computers, but…

Data integration longevity

- Integration needs stability
  - Standardization……
  - Good domain knowledge
  - Clearly defined data
  - Clearly identified goals

- Integration fears:
  - Heterogeneity of data and systems
  - Uncertain domain knowledge
  - Fast evolution of data
  - Highly specialized data
  - Lacking of predefined, clear goals
  - Originality, experimentalism (“let me see if this works”)

Integration of biological information

In biology:

- A pre-analysis and reorganization of information is very difficult, because knowledge and related data change very quickly
- Complexity of information makes it difficult to design data models which are valid for different domains and over time
- Goals and needs of researchers evolve very quickly according to new theories and discoveries

Integration must therefore be carried out by using flexible systems that are easy to adapt and extend

A possible methodology

A methodology based on data standardization:

- XML schemas for the creation of the models of the information
- XML based languages for data representation and storage
- Web Services for data exchange and interoperation between software
- Computerised workflows for the definition and execution of analysis processes
- “User friendly” portals for accessibility and usability of workflows by all scientists

XML and Web Services in bioinformatics

Models are lacking, XML and Web Services are increasingly being used

XML dialects

Sequences (BSML, Agave)

Proteins (SPML)

NCBI outputs (BlastXML)

Microarray (MAGE-ML)

Systems Biology Markup Language (SBML)

Biological Variation Markup Language (BVML)

Web Services

EMBOSS, XEMBL, Interpro (EBI)
eUtils (NCBI)
caBIO (NCICB)
KEGG API

GeneCruiser, Biosphere (microarray)
SIMAP (proteins)

CABRI (biological resources)
TP53 mutations (gene mutations)

bioMOBY (directory)
Soaplab (tools)

Workflow management

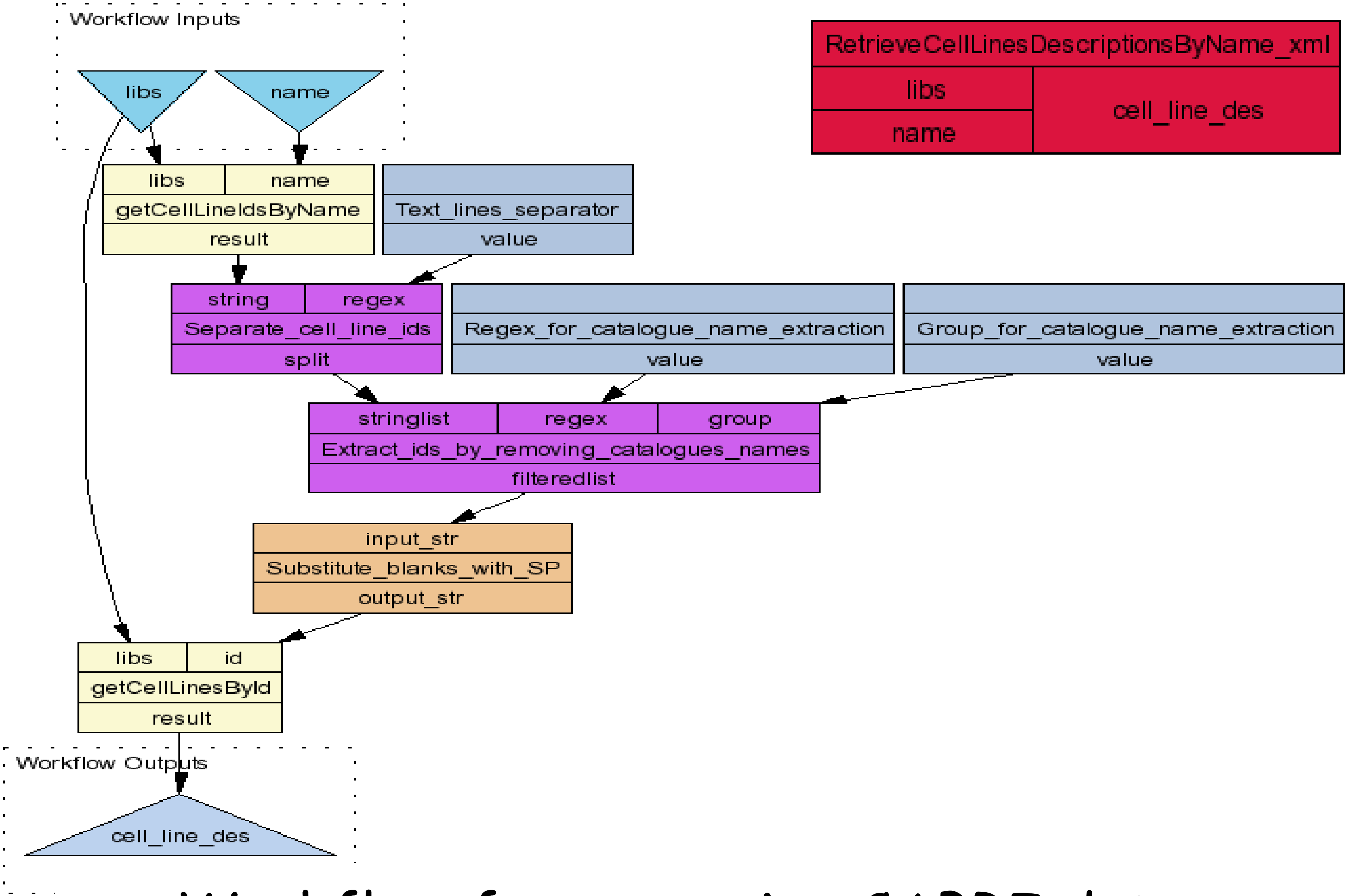
“**A computerized facilitation or automation of a business process, in whole or part**". (Workflow Management Coalition)

Its main goal is:

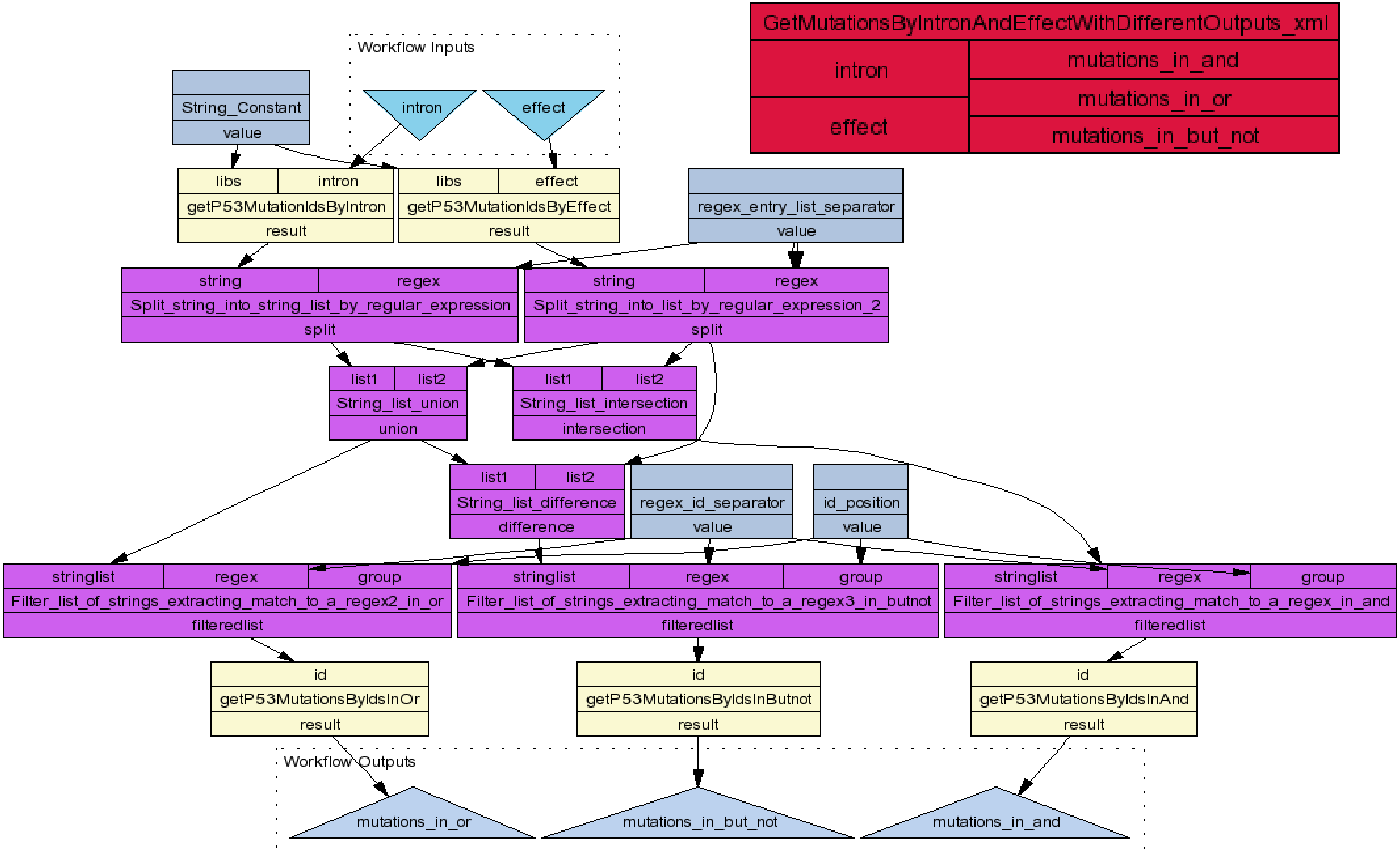
- the implementation of data analysis processes in standardized environments

Its main advantages relate to:

- **effectiveness**: being an automatic procedure, it frees bio-scientists from repetitive interactions with the web and it supports good practice,
- **reproducibility**: analysis can be replicated over time,
- **reusability**: intermediate results can be reused,
- **traceability**: the workflow is carried out in a transparent analysis environment where data provenance can be checked and/or controlled.



Workflow for accessing CABRI data



GetMutationsByIntronAndEffectWithDifferentOutputs_xml

intron

effect

mutations_in_and

mutations_in_or

mutations_in_but_not

Workflow Inputs

intron

effect

String_Constant

value

libs

intron

getP53MutationIdsByIntron

result

libs

effect

getP53MutationIdsByEffect

result

regex_entry_list_separator

value

string

regex

Split_string_into_string_list_by_regular_expression

split

string

regex

Split_string_into_list_by_regular_expression_2

split

list1

list2

String_list_union

union

list1

list2

String_list_intersection

intersection

list1

list2

String_list_difference

difference

regex_id_separator

value

id_position

value

stringlist

regex

group

Filter_list_of_strings_extracting_match_to_a_regex2_in_or

filteredlist

stringlist

regex

group

Filter_list_of_strings_extracting_match_to_a_regex3_in_butnot

filteredlist

stringlist

regex

group

Filter_list_of_strings_extracting_match_to_a_regex_in_and

filteredlist

id

getP53MutationsByIdsInOr

result

id

getP53MutationsByIdsInButnot

result

id

getP53MutationsByIdsInAnd

result

Workflow Outputs

mutations_in_or

mutations_in_but_not

mutations_in_and

Workflow for retrieving TP53 mutations

Workflow management systems

Different kinds of WMSs

- Software libraries

  Add-on to development tools, need programming efforts.

- Standalone systems

  Normally implemented on personal computers for accessing distributed services.

- Web interfaces

  Implement remote access to the functions of the WMS.

  Services can be local to the server or distributed.

  Maybe Grid enabled.

Workflow management systems

Different languages for workflows

- Proprietary
  Developed for a WMS, optimized to its goals
  Commercial software often not standardized
- Standards
  Standard can be very general, not really goal-oriented and specific.
  Different organizations, different standards
  WfMC, W3C

Different availability schemes

Commercial / Public domain or use / Open source

Workflow management systems

Software	Type	XML	Avail	URL
Taverna Workbench	**Stand-alone**	**XScufl**	**Open source**	**http://taverna.sourceforge.net/**
Biopipe	Library	Pipeline XML	Open source	http://www.gmod.org/biopipe/
ProGenGrid	Stand-alone	NA	NA	http://datadog.unile.it/progen
DiscoveryNet	Stand-alone	DPML	Commercial	http://www.discovery-on-the.net/
Kepler	Stand-alone	MoML	Open source	http://kepler-project.org/
GPipe	Web Interface , local services	GPipe XML	Open source	http://if-web1.imb.uq.edu.au/Pise/5.a/gpipe.html
EGene	Stand-alone	NA	Open source	http://www.lbm.fmvz.usp.br/egene/
BioWMS	Web Interface, remote services	XPDL	Public use	http://litbio.unicam.it:8080/biowms/
Biowep	**Portal**	**XScufl** **XPDL**	**Open source**	**http://bioinformatics.istge.it/biowep/**
BioWBI	Web Interface, local services	Proprietary	Commercial	http://www.alphaworks.ibm.com/tech/biowbi
Pegasys	Stand-alone	Pegasys DAG	Open source	http://bioinformatics.ubc.ca/pegasys/
Wildfire	Stand-alone	GEL	Open source	http://wildfire.bii.a-star.edu.sg/wildfire/
Triana	Stand-alone	Triana WL	Open source	http://www.trianacode.org/
Pipeline Pilot	Stand-alone	Proprietary	Commercial	http://www.scitegic.com/
FreeFluo	Library	WSFL & XScufl	Open source	http://freefluo.sourceforge.net/
Biomake	Library	NA	Open source	http://skam.sourceforge.net/

Bartocci et al, Proc. BITS Conf., 2006

Workflow management systems: issues

Network issues

Quality of Service

Availability / Access restrictions

Speed / Timeouts

Practical issues

Long running jobs -> timeouts / time limits

Huge data I/O -> timeouts / time limits

Access to Grid networks & services

Human interactions

Data reuse / data caching

Heterogeneity of WS, complexity of WS I/O

Heterogeneity of data -> shims, adapters (format conversions), data manipulation

Some solutions?

Scheduling: job IDs, monitoring execution

Reference data models

Semantic Web Services

Summarizing….

Moving from an interactive to an automated approach for data integration requires new technologies and tools.

Some starting assumptions

- XML schemas for the creation of the models of the information,
- XML based languages for data representation and exchange,
- Web Services for the interoperability of software
- Computerised workflows for the definition and execution of analysis processes

All proposed workflow management software for bioinformatics require knowledge of the WSs and skills and time for development and maintenance of workflows.

PORTALS CAN BE THE ANSWER

Portals: list of requirements

Workflow repository

Workflow edit, upload & download

Semi-automatic workflow editing

Reference repository

Workflow search

- By type
- By authors
- By linked publications
- By services (ontology)

Workflow description

- Annotation
- Tagging
- Peer reviewing, ratings

Workflow run time environment

Workflow execution

Pre-run workflow diagnosis

Automated process logging

Data management

Interactive workflows

Taverna Portal Party,
Manchester, September 28-29, 2006
(partial list)

Biowep: features (i)

We designed biowep, a workflow enactment portal for bioinformatics that:

- allows for the carrying out of predefined workflows
- supports workflows annotation through a simple ontology for bioinformatics processors (domain, task, i/o)
- implements search and selection of workflows on the basis of their annotation
- supports retrieval of workflows on the basis of users' profiling
- allows storing and retrieval of workflows' executions and related results

Biowep: features (ii)

The system:

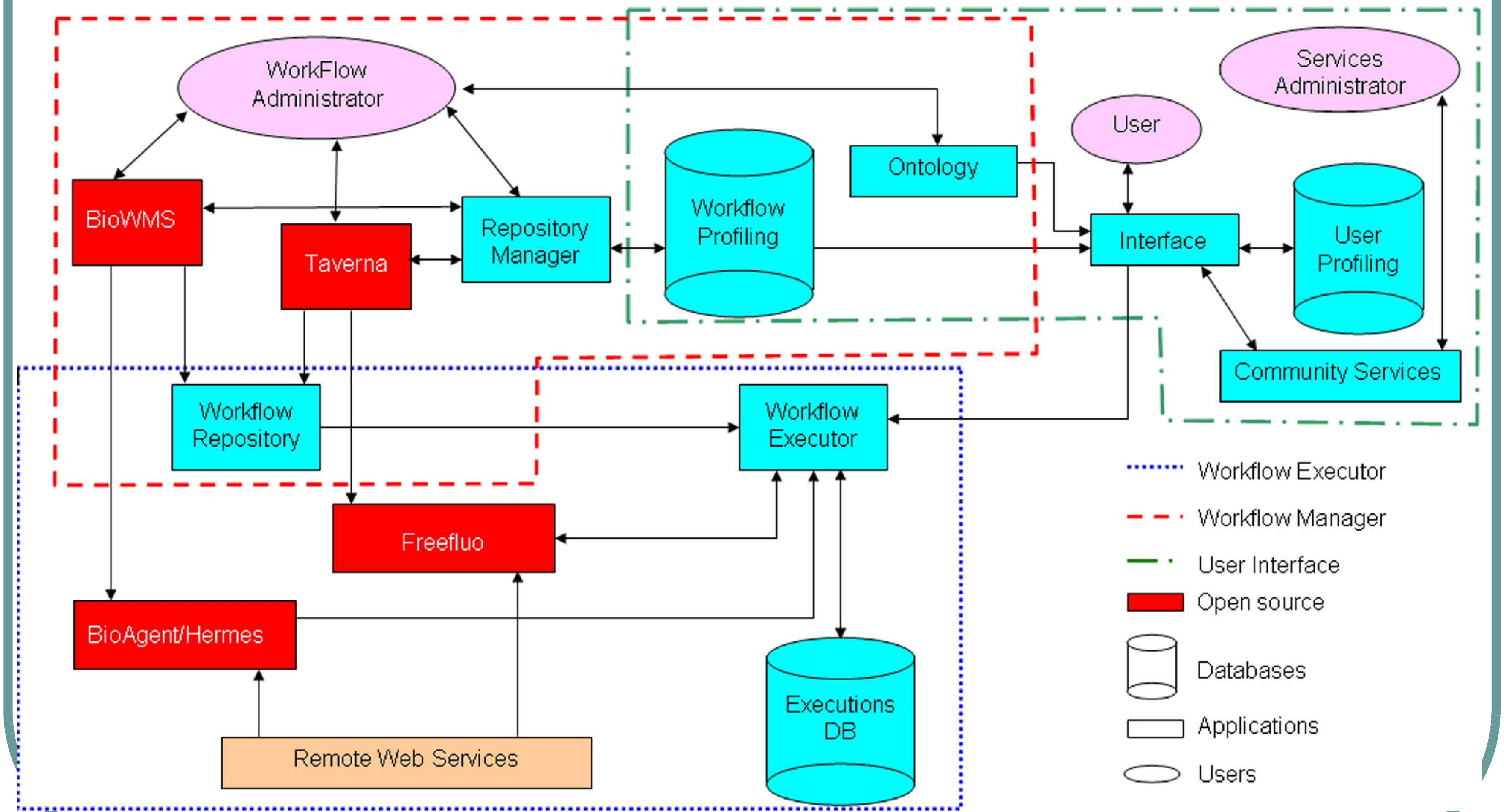
- uses open source sw (Taverna WB and mySQL)
- makes access to all services supported by Taverna
- stores workflows in Scufl and/or XPDL formats
- is available under LGPL license
- requirements: Java SDK + Tomcat with Axis

Available on-line:

http://bioinformatics.istge.it/biowep/ (reference site)
http://bioinformatics.istge.it:8080/biowep/ (portal)

Romano P et al, Biowep: a workflow enactment portal for bioinformatics applications, BMC Bioinformatics (accepted)

Biowep: architecture

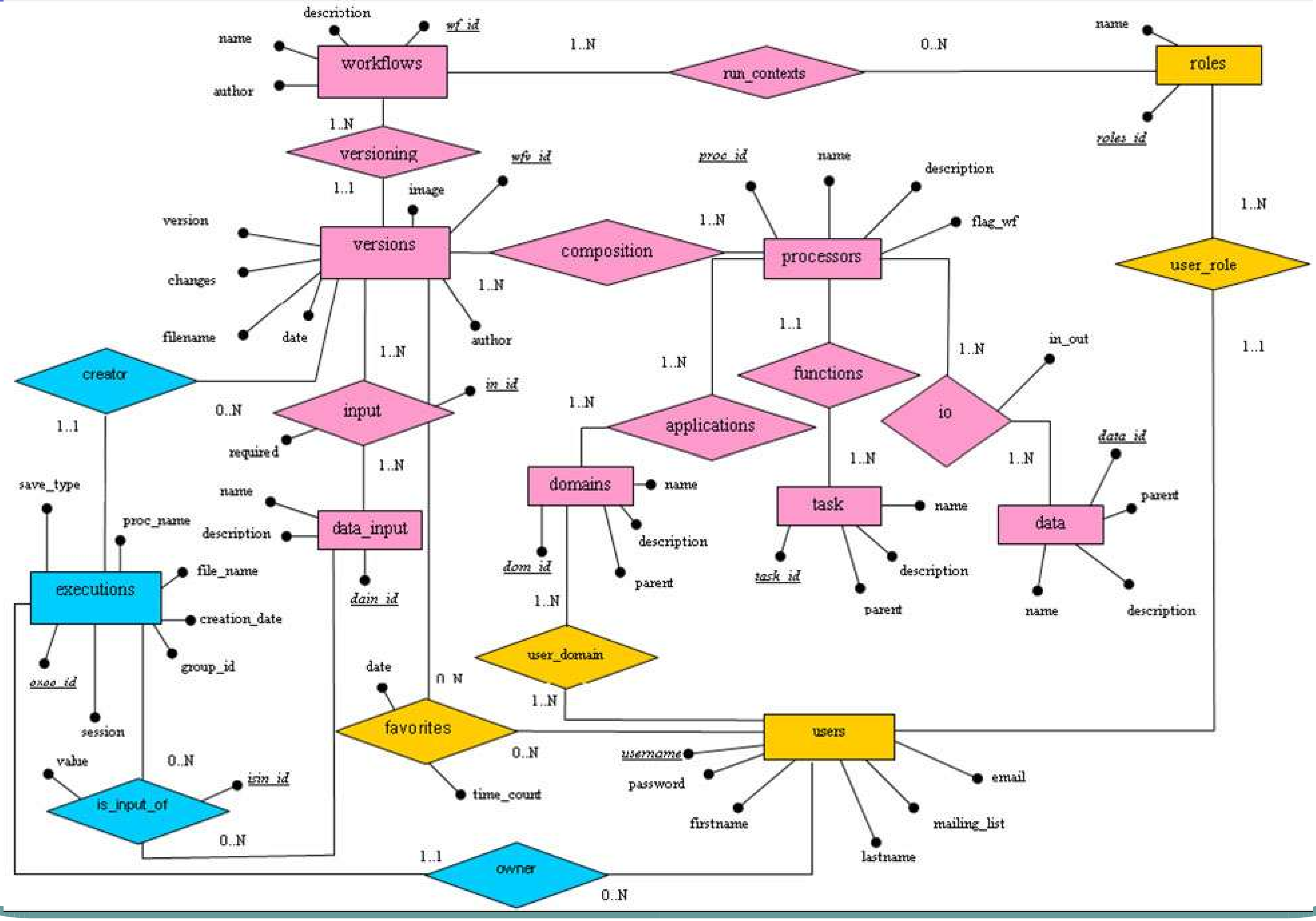


Biowep: workflows

Workflows are:

- created by the administrator by using Taverna or BioWMS
- archived in the related format in the repository
- annotated by using a simple ontology of bioinformatics data and tasks
- verified and updated when needed (workflow vs version)
- can be submitted by users or service providers through the reference site

Biowep: ERA schema



P. Romano, JdB06, 22/11/2006

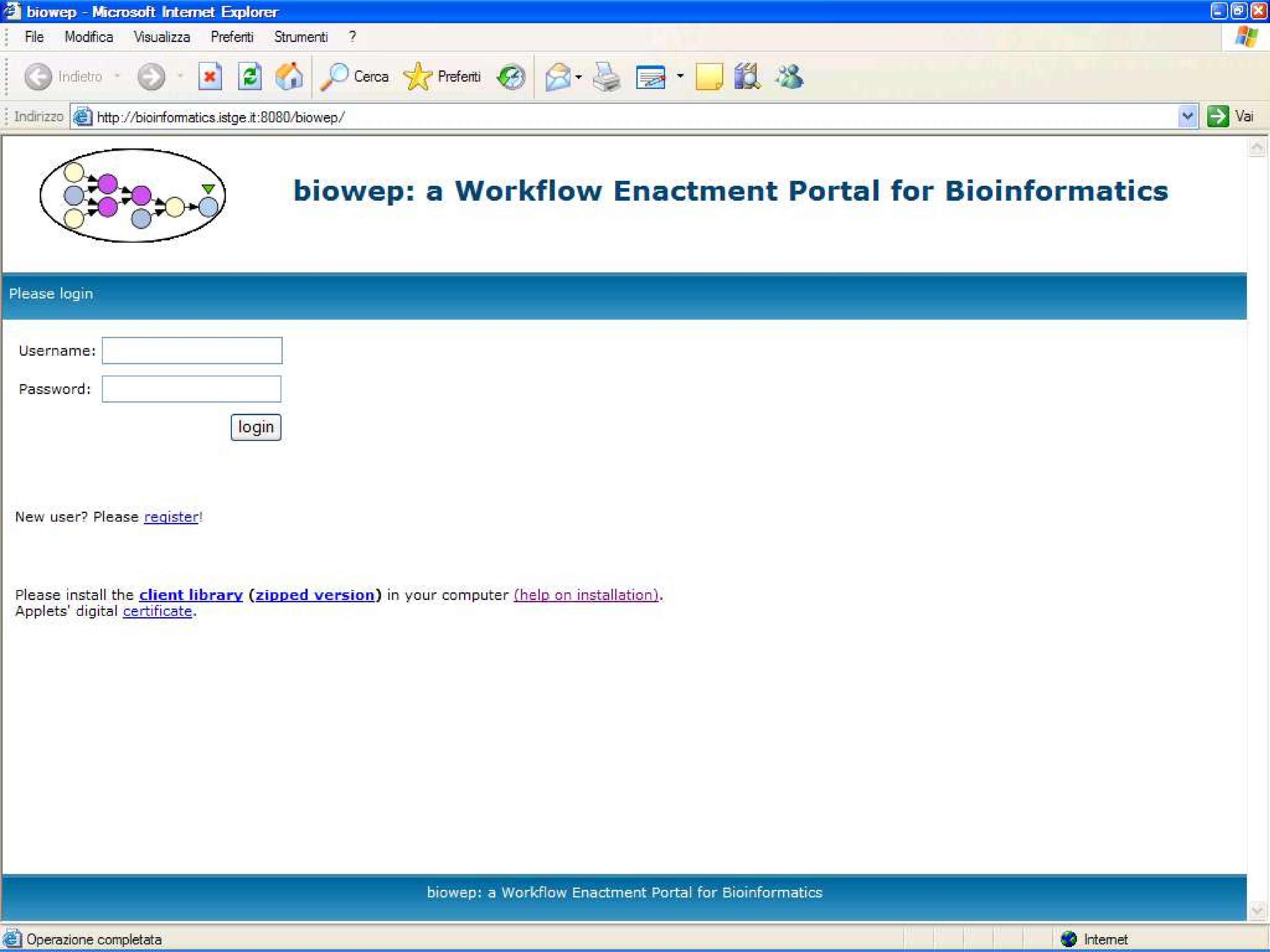
Biowep: users' profiling

Users are registered on the basis of:

- role in the organization / life
  - computer scientist / physician / researcher / patient / …
- scientific domains of interest

Users do not own workflows, they only own results of their executions. Results can be saved for later use and analysis

Workflows are listed according to user's profiling, frequency of use by other users. Last executed worflows are listed first.



biowep - Microsoft Internet Explorer
File
Modifica
Visualizza
Preferiti
Strumenti
?
Indietro
Cerca
Preferiti
Indirizzo
http://bioinformatics.istge.it:8080/biowep/
Vai

biowep: a Workflow Enactment Portal for Bioinformatics

Please login

Username:

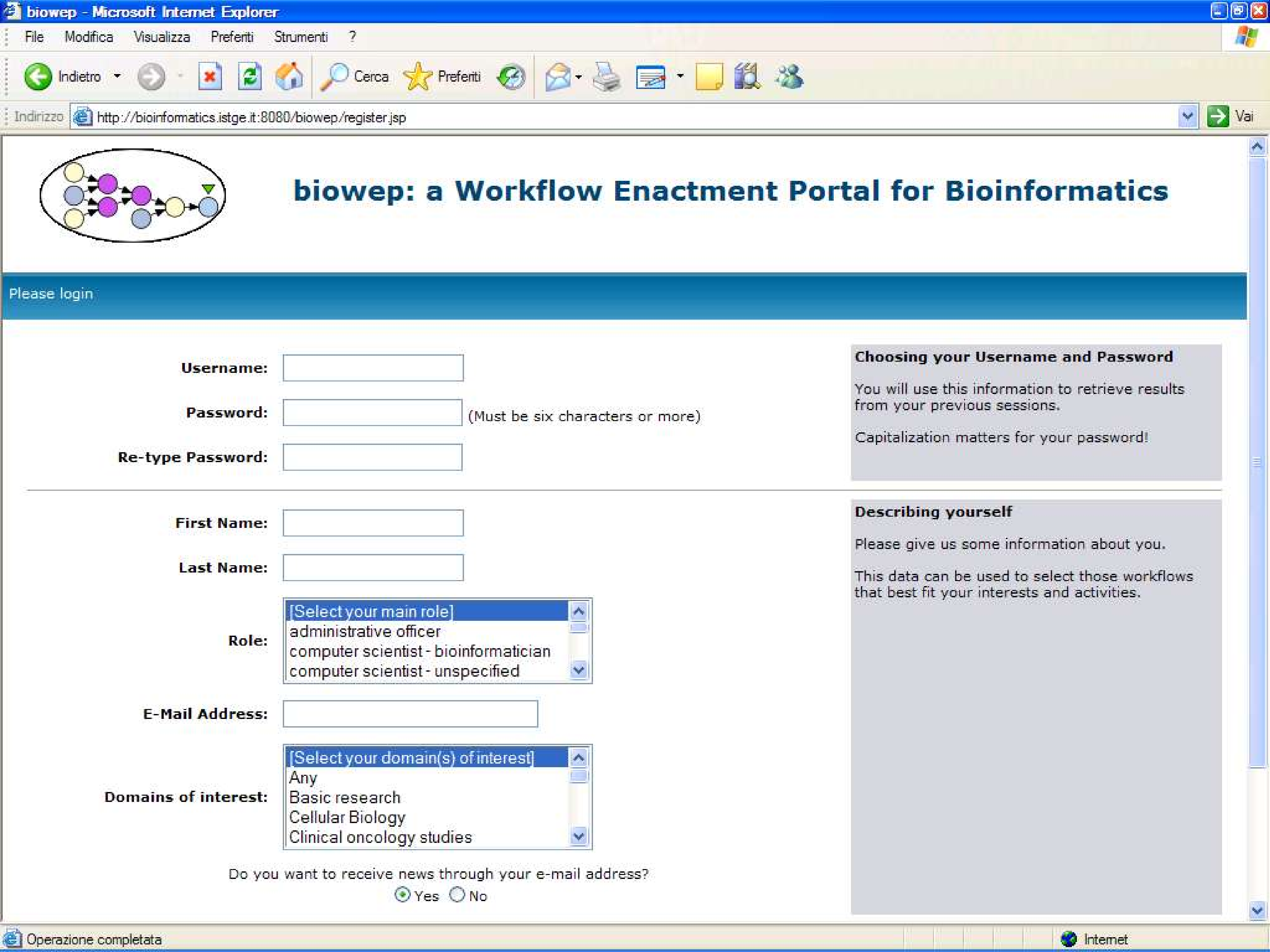
Password:

login

New user? Please register!

Please install the **client library** (**zipped version**) in your computer (help on installation).
Applets' digital certificate.

biowep: a Workflow Enactment Portal for Bioinformatics



biowep: a Workflow Enactment Portal for Bioinformatics

Please login

Username:

Password: (Must be six characters or more)

Re-type Password:

First Name:

Last Name:

Role: [Select your main role]
administrative officer
computer scientist - bioinformatician
computer scientist - unspecified

E-Mail Address:

Domains of interest: [Select your domain(s) of interest]
Any
Basic research
Cellular Biology
Clinical oncology studies

Do you want to receive news through your e-mail address?
● Yes ○ No

Choosing your Username and Password

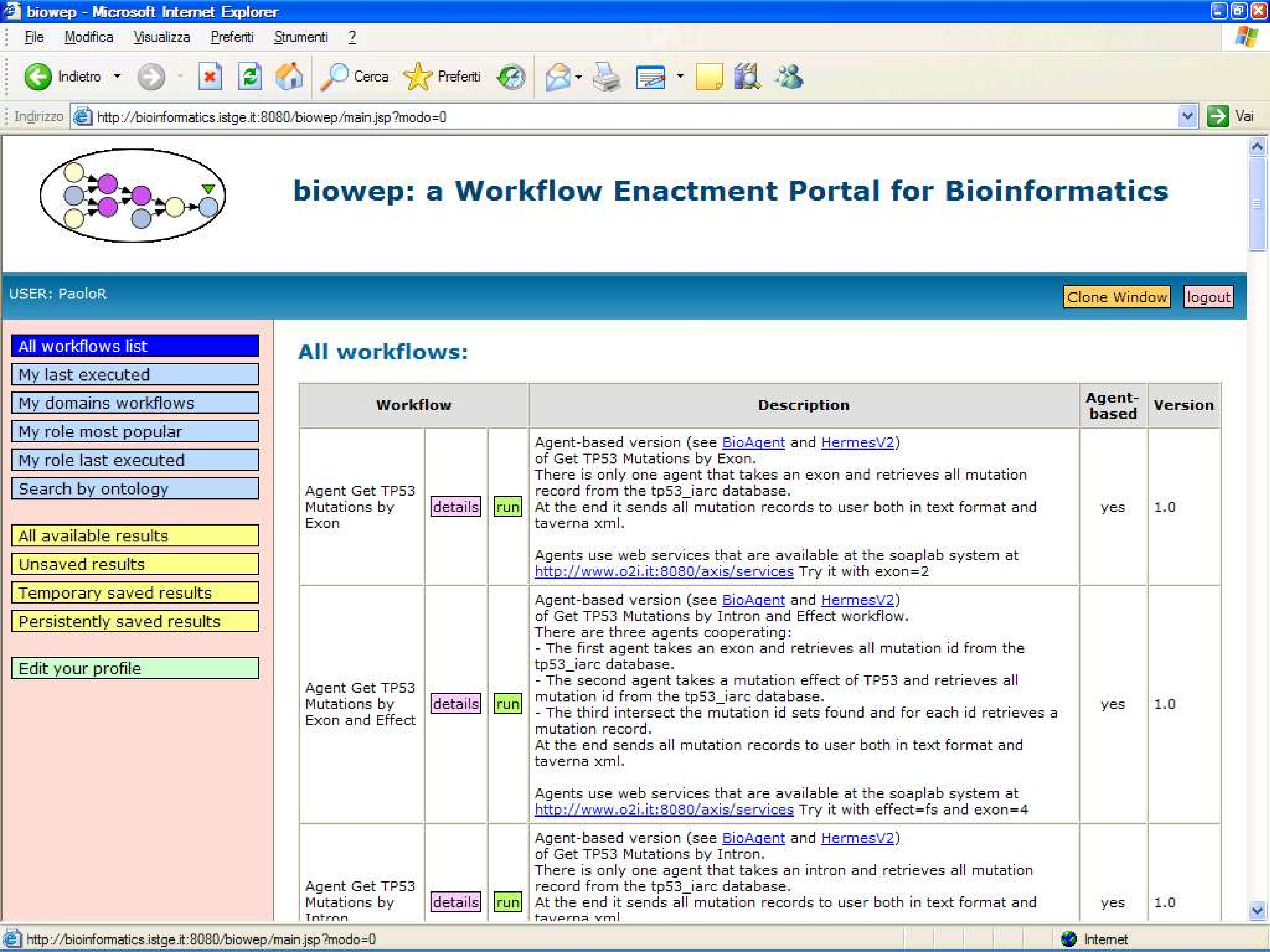
You will use this information to retrieve results from your previous sessions.

Capitalization matters for your password!

Describing yourself

Please give us some information about you.

This data can be used to select those workflows that best fit your interests and activities.

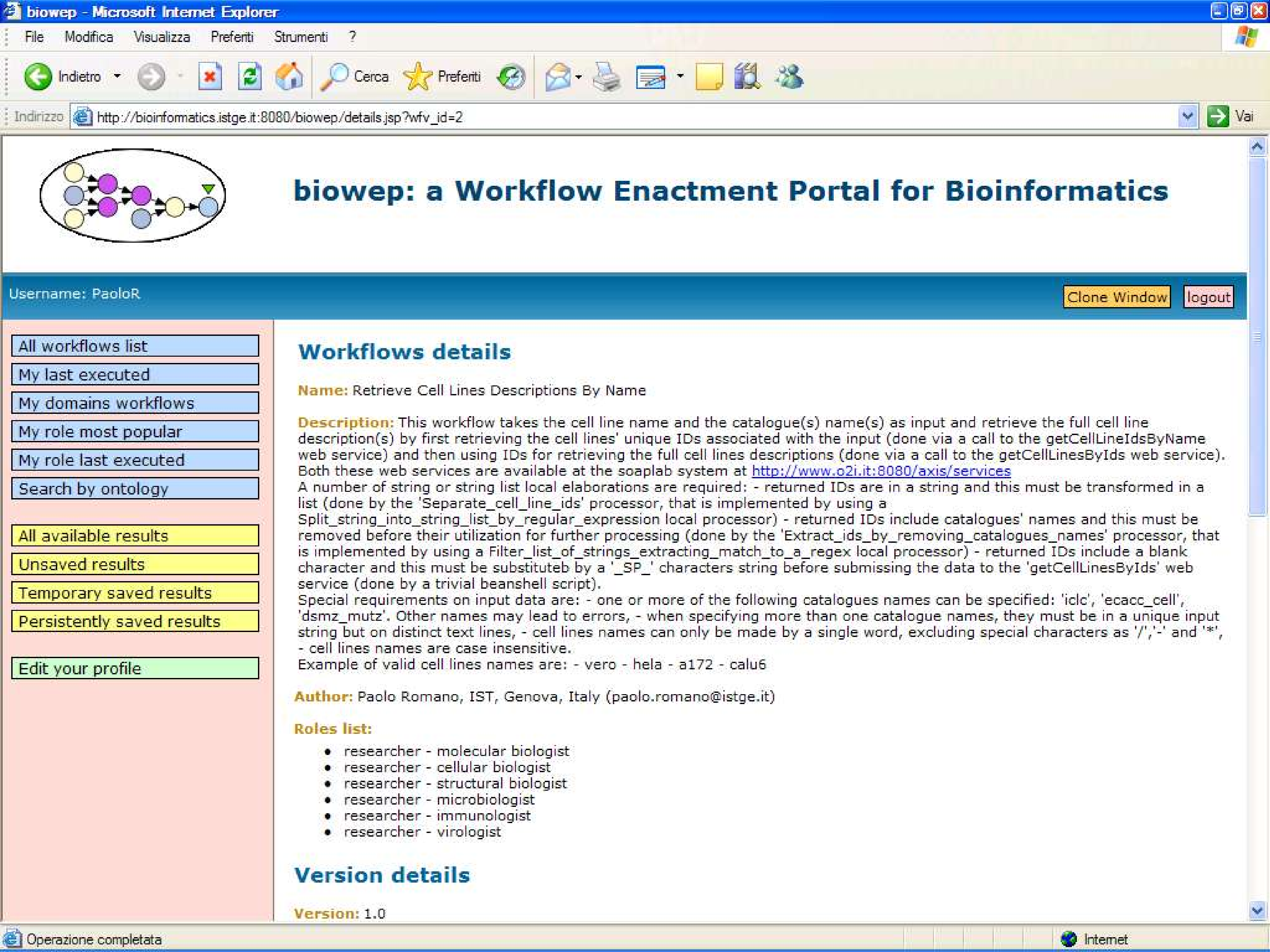


biowep: a Workflow Enactment Portal for Bioinformatics

USER: PaoloR

All workflows:

Workflow			Description	Agent-based	Version
Agent Get TP53 Mutations by Exon	details	run	Agent-based version (see BioAgent and HermesV2) of Get TP53 Mutations by Exon. There is only one agent that takes an exon and retrieves all mutation record from the tp53_iarc database. At the end it sends all mutation records to user both in text format and taverna xml. Agents use web services that are available at the soaplab system at http://www.o2i.it:8080/axis/services Try it with exon=2	yes	1.0
Agent Get TP53 Mutations by Exon and Effect	details	run	Agent-based version (see BioAgent and HermesV2) of Get TP53 Mutations by Intron and Effect workflow. There are three agents cooperating: - The first agent takes an exon and retrieves all mutation id from the tp53_iarc database. - The second agent takes a mutation effect of TP53 and retrieves all mutation id from the tp53_iarc database. - The third intersect the mutation id sets found and for each id retrieves a mutation record. At the end sends all mutation records to user both in text format and taverna xml. Agents use web services that are available at the soaplab system at http://www.o2i.it:8080/axis/services Try it with effect=fs and exon=4	yes	1.0
Agent Get TP53 Mutations by Intron	details	run	Agent-based version (see BioAgent and HermesV2) of Get TP53 Mutations by Intron. There is only one agent that takes an intron and retrieves all mutation record from the tp53_iarc database. At the end it sends all mutation records to user both in text format and taverna xml	yes	1.0



biowep: a Workflow Enactment Portal for Bioinformatics

Username: PaoloR

Workflows details

Name: Retrieve Cell Lines Descriptions By Name

Description: This workflow takes the cell line name and the catalogue(s) name(s) as input and retrieve the full cell line description(s) by first retrieving the cell lines' unique IDs associated with the input (done via a call to the getCellLineIdsByName web service) and then using IDs for retrieving the full cell lines descriptions (done via a call to the getCellLinesByIds web service). Both these web services are available at the soaplab system at http://www.o2i.it:8080/axis/services
A number of string or string list local elaborations are required: - returned IDs are in a string and this must be transformed in a list (done by the 'Separate_cell_line_ids' processor, that is implemented by using a Split_string_into_string_list_by_regular_expression local processor) - returned IDs include catalogues' names and this must be removed before their utilization for further processing (done by the 'Extract_ids_by_removing_catalogues_names' processor, that is implemented by using a Filter_list_of_strings_extracting_match_to_a_regex local processor) - returned IDs include a blank character and this must be substituteb by a '_SP_' characters string before submissing the data to the 'getCellLinesByIds' web service (done by a trivial beanshell script).
Special requirements on input data are: - one or more of the following catalogues names can be specified: 'iclc', 'ecacc_cell', 'dsmz_mutz'. Other names may lead to errors, - when specifying more than one catalogue names, they must be in a unique input string but on distinct text lines, - cell lines names can only be made by a single word, excluding special characters as '/','-' and '*', - cell lines names are case insensitive.
Example of valid cell lines names are: - vero - hela - a172 - calu6

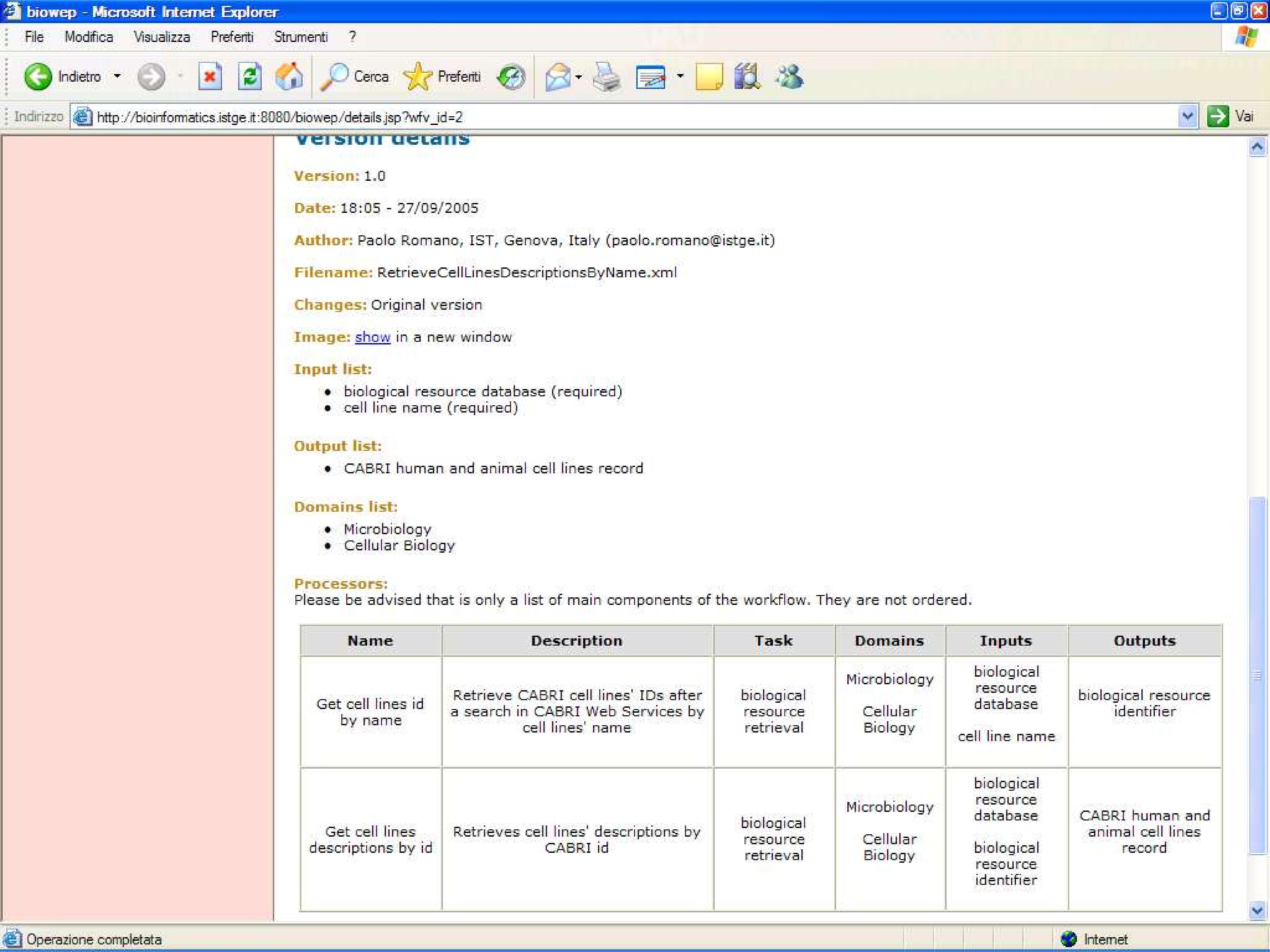
Author: Paolo Romano, IST, Genova, Italy (paolo.romano@istge.it)

Roles list:

- researcher - molecular biologist
- researcher - cellular biologist
- researcher - structural biologist
- researcher - microbiologist
- researcher - immunologist
- researcher - virologist

Version details

Version: 1.0



Version details

Version: 1.0

Date: 18:05 - 27/09/2005

Author: Paolo Romano, IST, Genova, Italy (paolo.romano@istge.it)

Filename: RetrieveCellLinesDescriptionsByName.xml

Changes: Original version

Image: show in a new window

Input list:

- biological resource database (required)
- cell line name (required)

Output list:

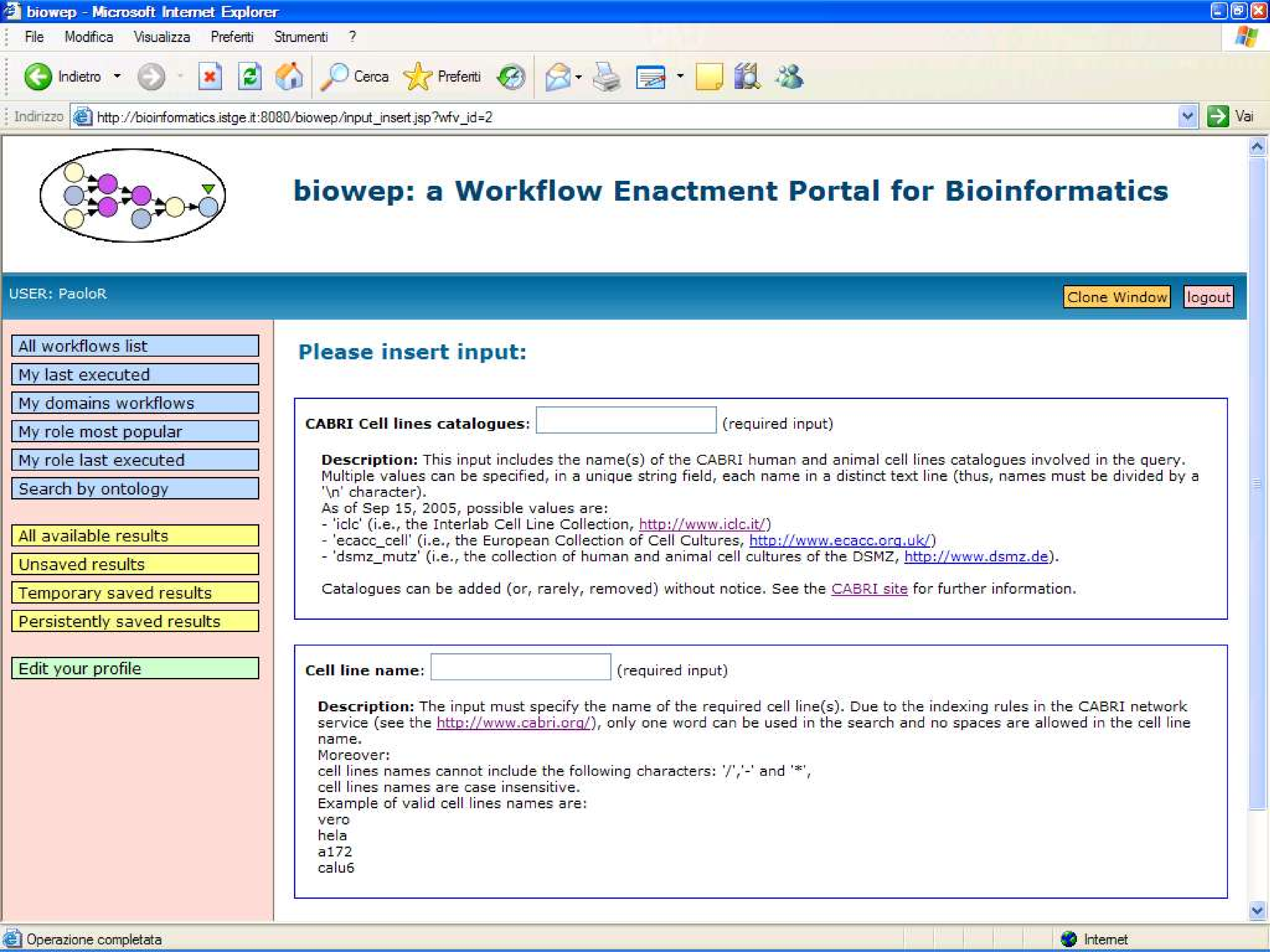
- CABRI human and animal cell lines record

Domains list:

- Microbiology
- Cellular Biology

Processors:
Please be advised that is only a list of main components of the workflow. They are not ordered.

Name	Description	Task	Domains	Inputs	Outputs
Get cell lines id by name	Retrieve CABRI cell lines' IDs after a search in CABRI Web Services by cell lines' name	biological resource retrieval	Microbiology Cellular Biology	biological resource database cell line name	biological resource identifier
Get cell lines descriptions by id	Retrieves cell lines' descriptions by CABRI id	biological resource retrieval	Microbiology Cellular Biology	biological resource database biological resource identifier	CABRI human and animal cell lines record



biowep: a Workflow Enactment Portal for Bioinformatics

Please insert input:

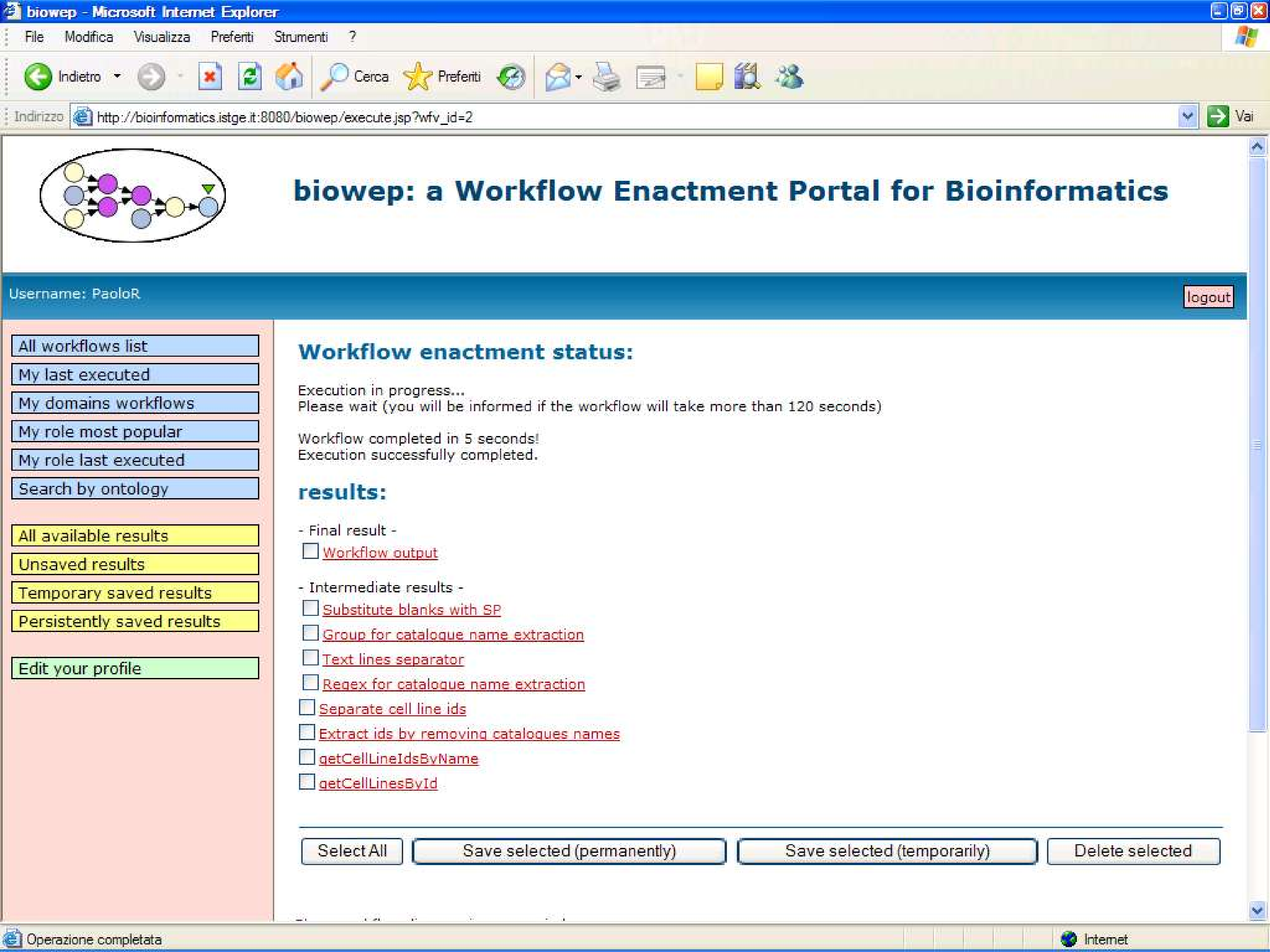
CABRI Cell lines catalogues: (required input)

Description: This input includes the name(s) of the CABRI human and animal cell lines catalogues involved in the query. Multiple values can be specified, in a unique string field, each name in a distinct text line (thus, names must be divided by a '\n' character).
As of Sep 15, 2005, possible values are:
- 'iclc' (i.e., the Interlab Cell Line Collection, http://www.iclc.it/)
- 'ecacc_cell' (i.e., the European Collection of Cell Cultures, http://www.ecacc.org.uk/)
- 'dsmz_mutz' (i.e., the collection of human and animal cell cultures of the DSMZ, http://www.dsmz.de).

Catalogues can be added (or, rarely, removed) without notice. See the CABRI site for further information.

Cell line name: (required input)

Description: The input must specify the name of the required cell line(s). Due to the indexing rules in the CABRI network service (see the http://www.cabri.org/), only one word can be used in the search and no spaces are allowed in the cell line name.
Moreover:
cell lines names cannot include the following characters: '/','-' and '*',
cell lines names are case insensitive.
Example of valid cell lines names are:
vero
hela
a172
calu6



biowep: a Workflow Enactment Portal for Bioinformatics

Username: PaoloR

Workflow enactment status:

Execution in progress...
Please wait (you will be informed if the workflow will take more than 120 seconds)

Workflow completed in 5 seconds!
Execution successfully completed.

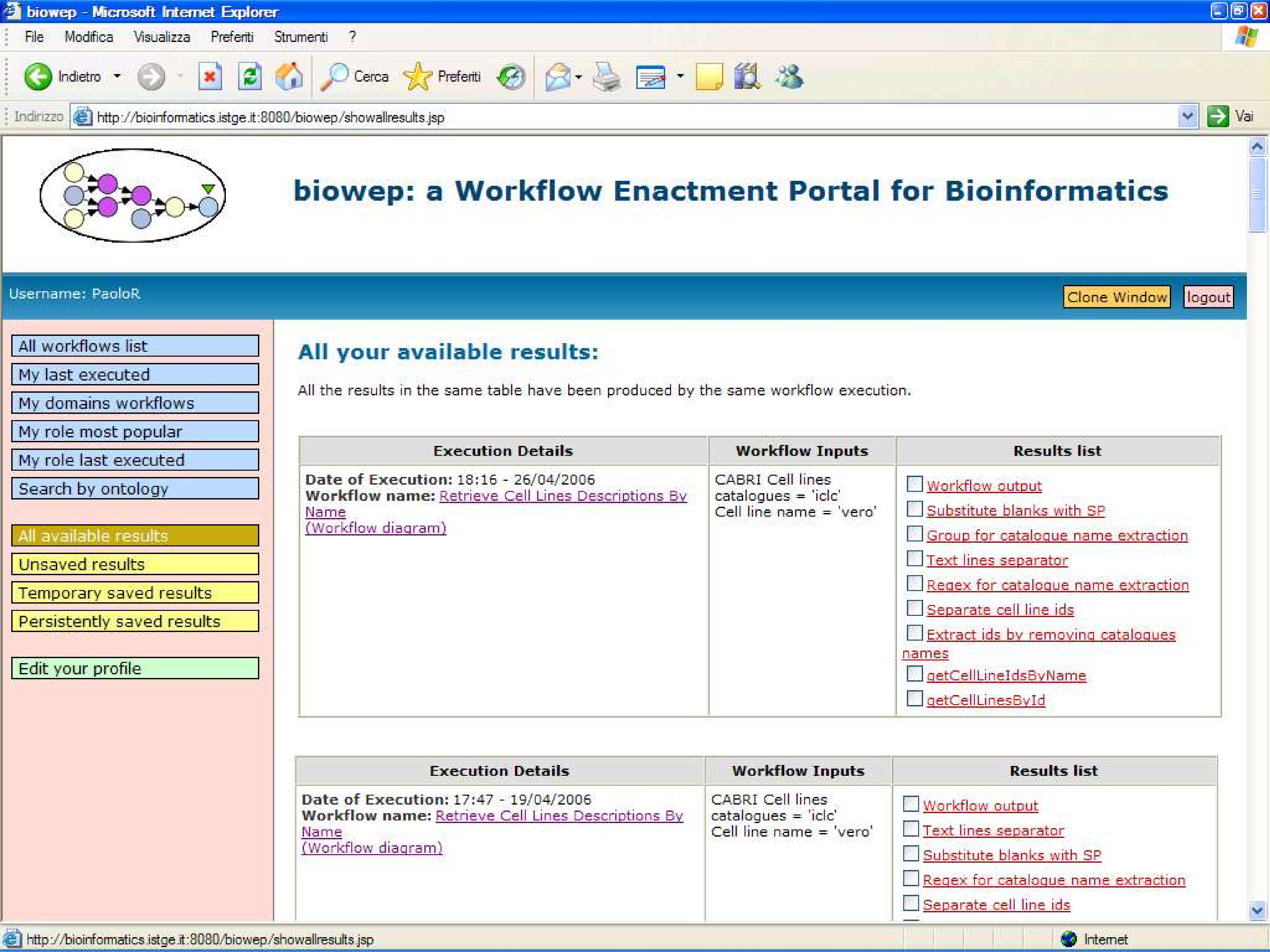
results:

- Final result -

□ Workflow output

- Intermediate results -

□ Substitute blanks with SP
□ Group for catalogue name extraction
□ Text lines separator
□ Regex for catalogue name extraction
□ Separate cell line ids
□ Extract ids by removing catalogues names
□ getCellLineIdsByName
□ getCellLinesById



biowep: a Workflow Enactment Portal for Bioinformatics

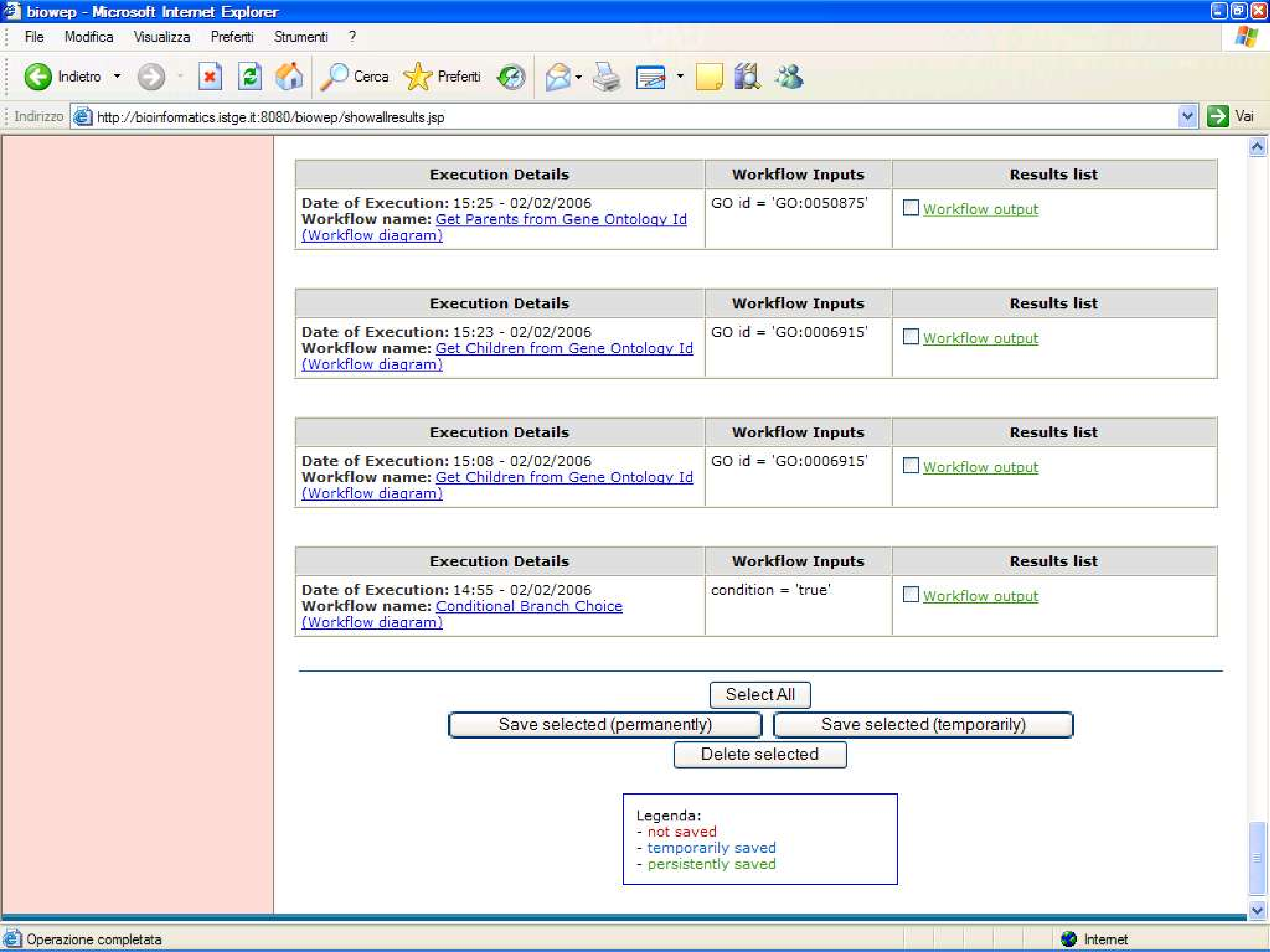
Username: PaoloR

All your available results:

All the results in the same table have been produced by the same workflow execution.

Execution Details	Workflow Inputs	Results list
Date of Execution: 18:16 - 26/04/2006 **Workflow name:** Retrieve Cell Lines Descriptions By Name (Workflow diagram)	CABRI Cell lines catalogues = 'iclc' Cell line name = 'vero'	□ Workflow output □ Substitute blanks with SP □ Group for catalogue name extraction □ Text lines separator □ Regex for catalogue name extraction □ Separate cell line ids □ Extract ids by removing catalogues names □ getCellLineIdsByName □ getCellLinesById

Execution Details	Workflow Inputs	Results list
Date of Execution: 17:47 - 19/04/2006 **Workflow name:** Retrieve Cell Lines Descriptions By Name (Workflow diagram)	CABRI Cell lines catalogues = 'iclc' Cell line name = 'vero'	□ Workflow output □ Text lines separator □ Substitute blanks with SP □ Regex for catalogue name extraction □ Separate cell line ids


biowep - Microsoft Internet Explorer
File
Modifica
Visualizza
Preferiti
Strumenti
?
Indietro
Cerca
Preferiti
Indirizzo
http://bioinformatics.istge.it:8080/biowep/showallresults.jsp
Vai
Select All
Save selected (permanently)
Save selected (temporarily)
Delete selected
Operazione completata
Internet

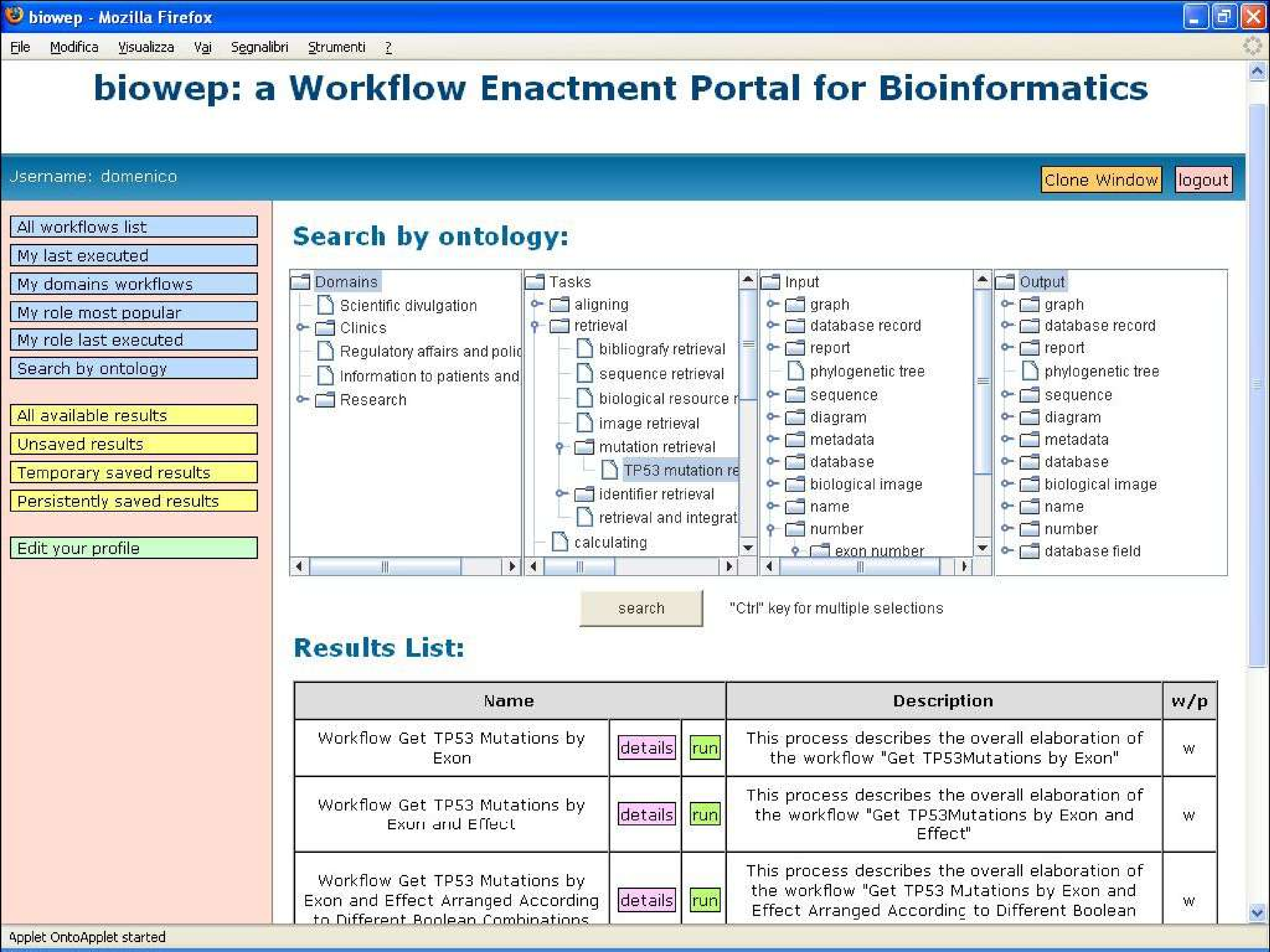
Execution Details	Workflow Inputs	Results list
Date of Execution: 15:25 - 02/02/2006 **Workflow name:** Get Parents from Gene Ontology Id (Workflow diagram)	GO id = 'GO:0050875'	□ Workflow output

Execution Details	Workflow Inputs	Results list
Date of Execution: 15:23 - 02/02/2006 **Workflow name:** Get Children from Gene Ontology Id (Workflow diagram)	GO id = 'GO:0006915'	□ Workflow output

Execution Details	Workflow Inputs	Results list
Date of Execution: 15:08 - 02/02/2006 **Workflow name:** Get Children from Gene Ontology Id (Workflow diagram)	GO id = 'GO:0006915'	□ Workflow output

Execution Details	Workflow Inputs	Results list
Date of Execution: 14:55 - 02/02/2006 **Workflow name:** Conditional Branch Choice (Workflow diagram)	condition = 'true'	□ Workflow output

Legenda:
- not saved
- temporarily saved
- persistently saved



biowep: a Workflow Enactment Portal for Bioinformatics

Search by ontology:

"Ctrl" key for multiple selections

Results List:

Name			Description	w/p
Workflow Get TP53 Mutations by Exon	details	run	This process describes the overall elaboration of the workflow "Get TP53Mutations by Exon"	w
Workflow Get TP53 Mutations by Exon and Effect	details	run	This process describes the overall elaboration of the workflow "Get TP53Mutations by Exon and Effect"	w
Workflow Get TP53 Mutations by Exon and Effect Arranged According to Different Boolean Combinations	details	run	This process describes the overall elaboration of the workflow "Get TP53 Mutations by Exon and Effect Arranged According to Different Boolean	w

Biowep: planned activities

- Improvement of end users support (two implementations available)
  - manuals, FAQs
  - support through biowep@istge.it , mailing lists
- Bug fixing (curation of children diseases)
- Revision of applications:
  - administration (end of 2006)
  - visualization of results (end of march 2007)
- Inclusion of new workflows
- Inclusion of further WMS
  - analysis and selection undergoing

Some acknowledgements…

IST, Genoa

Paolo Romano,

Domenico Marra,

Chiara Rasi,

Ulrich Pfeffer,

Valentina Mirisola,

M. Assunta Manniello

Gilberto Fronza

ITB, CNR, Milan

Luciano Milanesi

DISCo, University of Milan Bicocca

Guglielmo Bertolini,

Flavio De Paoli,

Giancarlo Mauri

DIST, University of Genoa

Ivan Porro,

Silvia Scaglione

DMI, University of Camerino

Emanuela Merelli,

Ezio Bartocci

This work has partially been supported by the Italian Ministry for Education, University and Research (MIUR), projects "Oncology over Internet– O_2I" and "Laboratory for Interdisciplinary Technologies in Bioinformatics - LITBIO"