Oncology over Internet:

integrating data and analysis of oncology interest on the net by means of workflows

Paolo Romano (paolo.romano@istge.it)
National Cancer Research Institute,
Genoa, Italy

Outline

- Motivations for the system
- Goals of the system
- The system
  - architecture
  - ERA schema
  - users profiling
  - workflow annotation

Information in biology: well known facts

- Biomedical research produces an increasing quantity of new data and new data types
  - EMBL size: ver 82 7,3% more vs ver 81 (3 months), 112,9% vs ver 74 (24 months)
- Emerging domains, like mutation and variation analysy, polymorphisms, metabolism, as well as new technologies, e.g., microarrays, will contribute with even huger amounts of data
- Analysis softwares must interoperate with databases
  - Databases as input for softwares
  - Results as new data to record and analyze

Heterogeneicity of databanks

- A few dbs are managed in an almost homogenous way (e.g., sequences at EBI, NCBI, DDBJ)
- Secondary databases are of high quality (good and extended annotation, quality control)
- Many databases are highly specialized, e.g. by gene, organism, disease, mutation, etc…
- Many databanks are created by small groups or by single researchers
- Databanks are distributed:
  - Different DBMS, data structures, query methods
  - Different information, semantics

Goals of the integration

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

- Carry out analysis and/or searches involving more databases and softwares automatically
- Perform analysis involving large data sets effectively
- Achieve a better and wider view of all available information
- Carry out a real data mining

Data integration longevity

- Integration needs stability
  - Standardization……
  - Good domain knowledge
  - Well defined data
  - Well defined goals
- Integration fears:
  - Heterogeneicity 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 the data is very difficult, because data and related knowledge change very quickly
- Complexity of information makes it difficult to design data models which can be 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 to extend

Workflows management

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

Main goal is:

- the implementation of data analysis processes in standardized environments

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 management software

Workflow management softwares for bioinformatics applications:

- Biopipe, an add-on to bioperl
- GPipe, an extension of the Pise interface
- Taverna (EBI), a component of the myGrid platform
- Wildfire (Bioinformatics Institute, Singapore)
- Pipeline Pilot (SciTegic)
- BioWBI, Bioinformatic Workflow Builder Interface, from IBM

They all require knowledge of the systems and skills and time for development of the workflows.

Oncology over Internet (O_2I)

We designed a web system that:

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

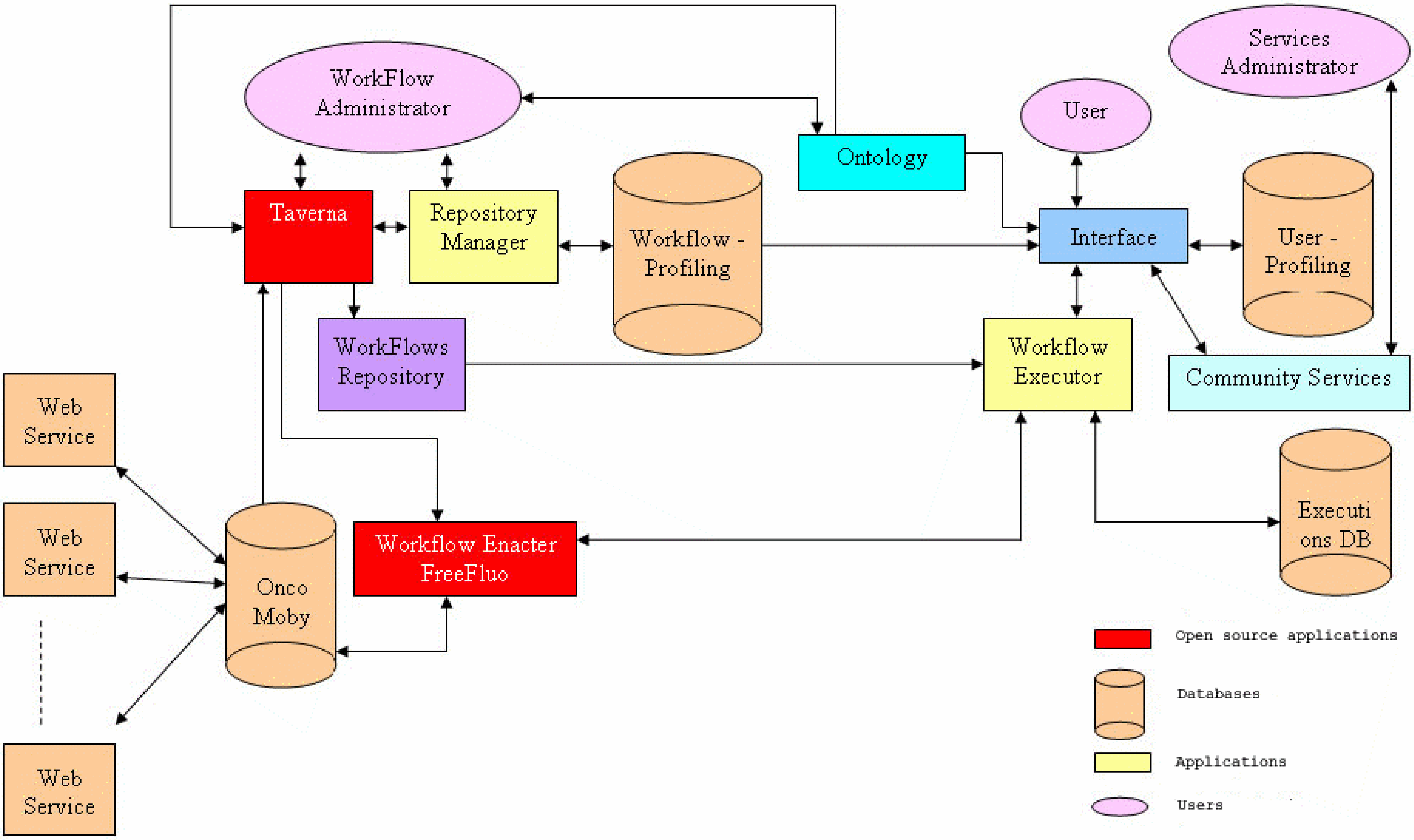
Oncology over Internet (O_2I)

We designed a web system that:

- makes access to and retrieves data from Web Services and registries of Web Services
- stores workflows using the Simple conceptual unified flow language (Scufl) format
- is partially based on open source tools (Taverna WB, FreeFluo and mySQL)

Prototype available on-line by end of 2005: http://www.o2i.it:8080/portal/

O$_2$I architecture



Predefined workflows

Workflows are:

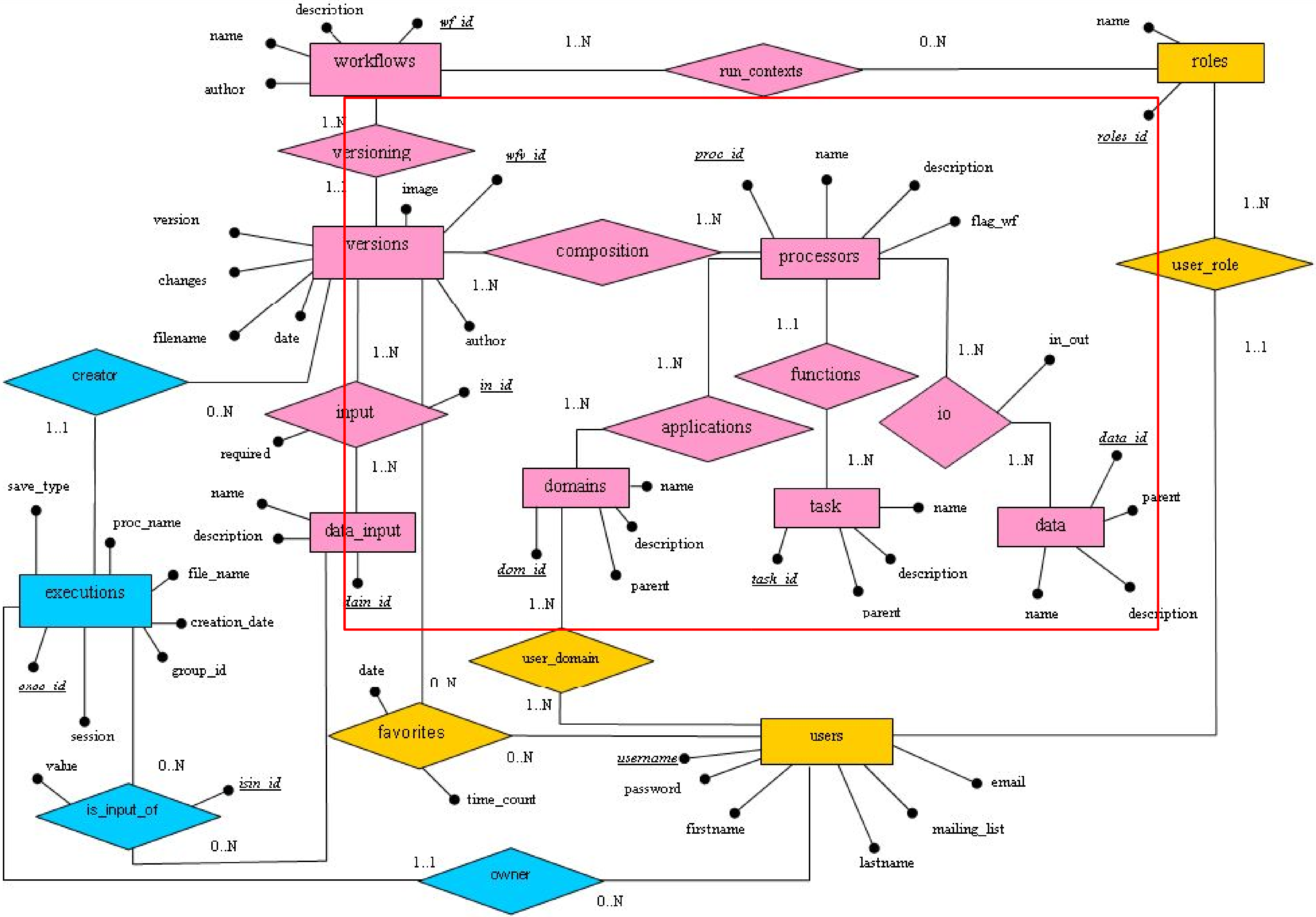
- created by internal staff using Taverna
- stored in Scufl format
- maintained (workflow vs version)
- submitted by:
  - users
  - service providers

Annotation of workflows

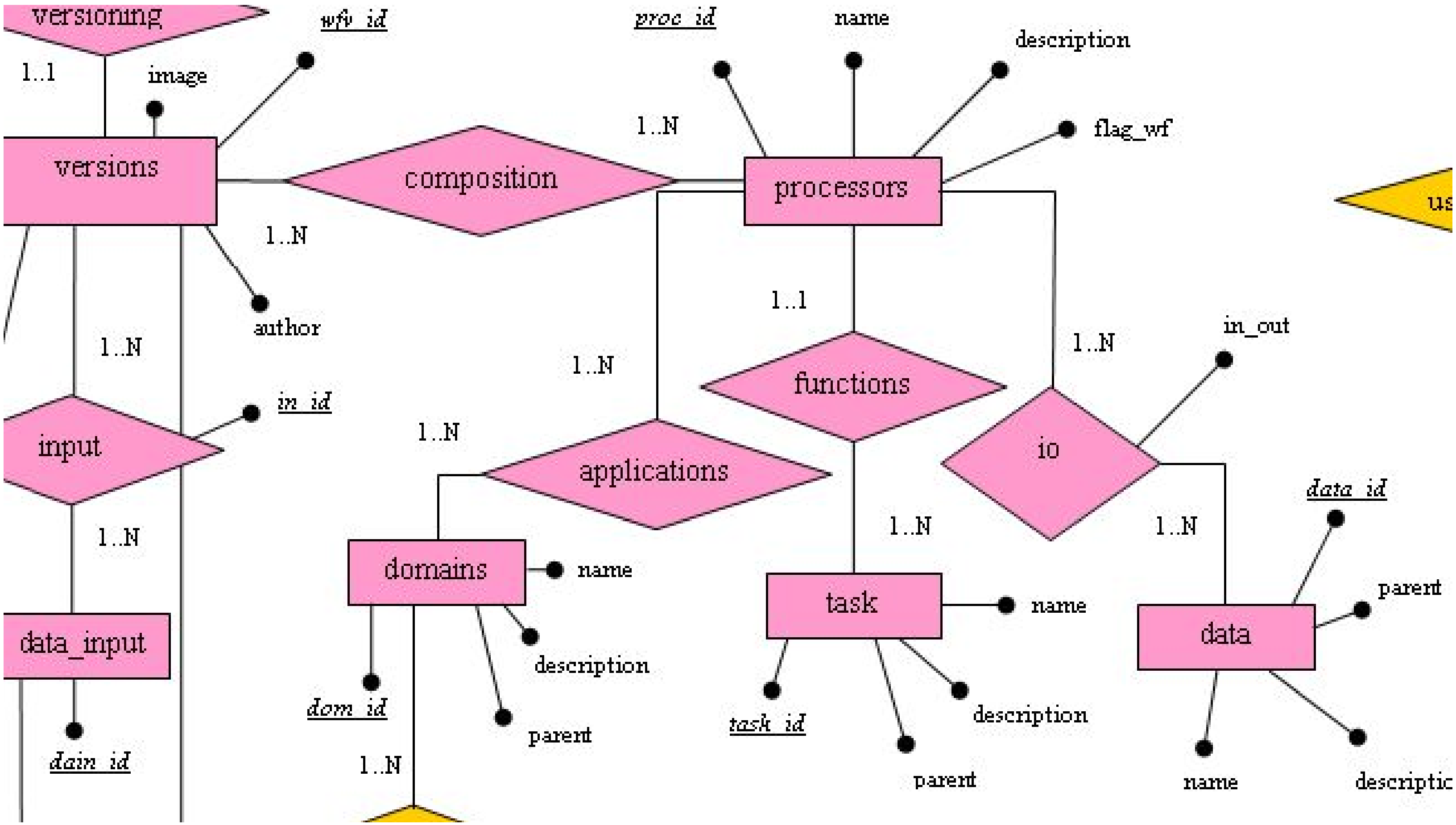
Workflows are annotated on the basis of:

- a simple ontology for bioinformatics processors:
  - application domains
  - task
  - inputs/outputs
- ontology derived from Taverna:
  - new structure
  - some additions (biological resources, images, …)
  - under further development

O$_2$I ERA schema



O_2I workflows annotation

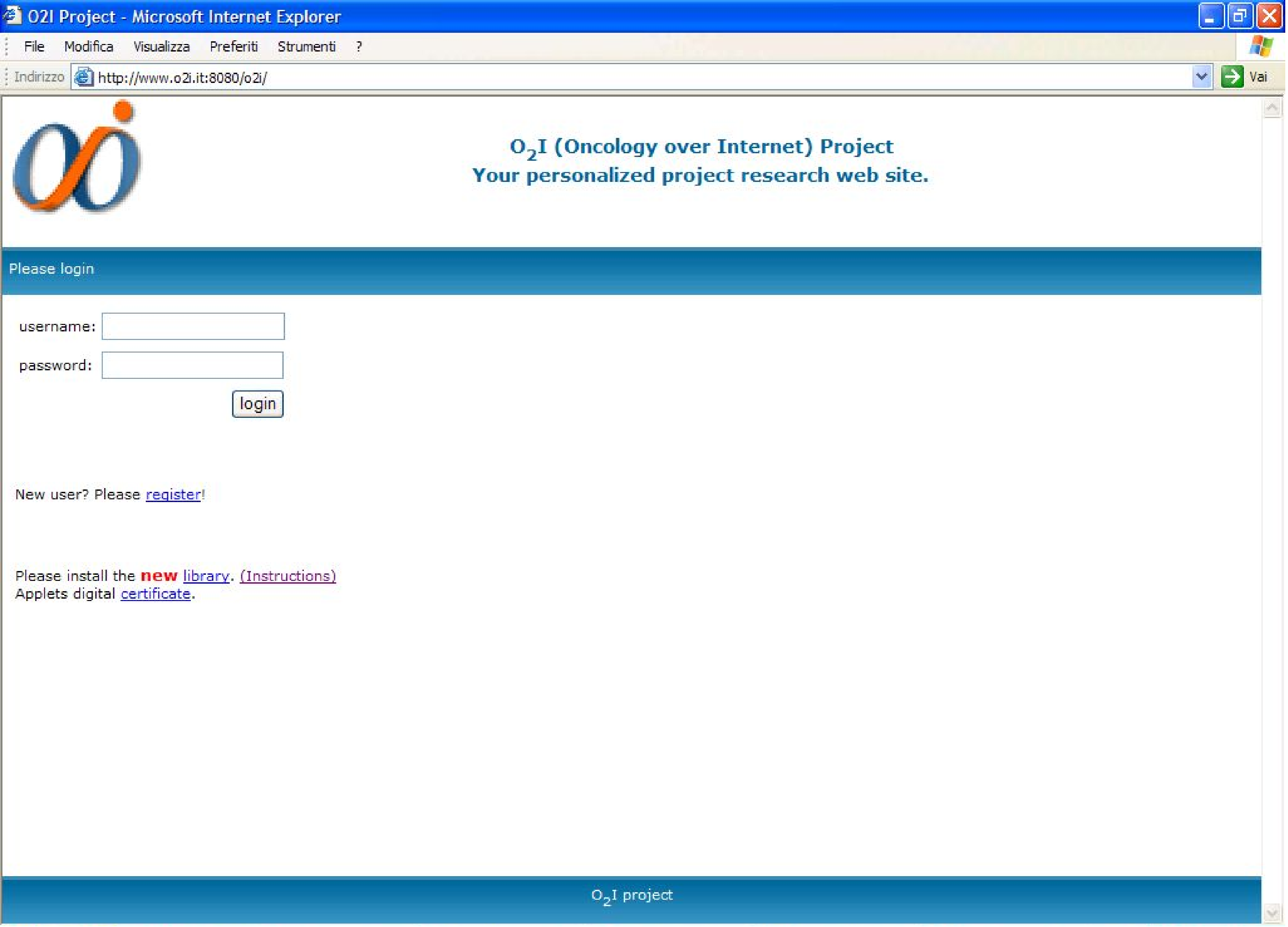


P. Romano, Naples, October 6, 2005

Users’ registration and profiling

Users are profiled on the basis of:

- role in their organization
  - computer scientist / physician / researcher / patient / journalist / …
- domains of interest
- past workflows’ executions



O2I Project - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indirizzo
Vai

http://www.o2i.it:8080/o2i/

O_2I (Oncology over Internet) Project
Your personalized project research web site.

Please login

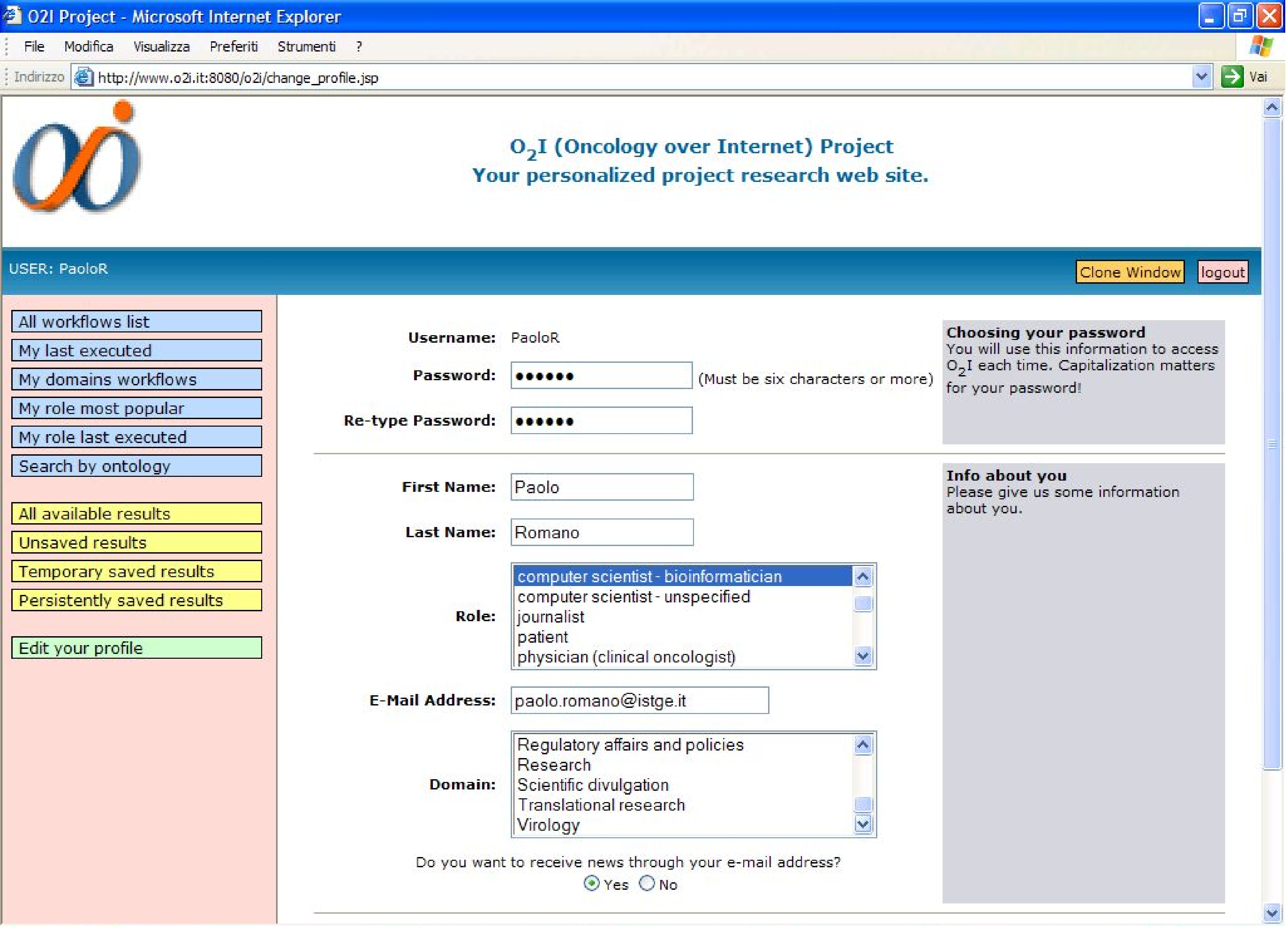
username:

password:

New user? Please register!

Please install the **new** library. (Instructions)
Applets digital certificate.

O_2I project



O_2I (Oncology over Internet) Project
Your personalized project research web site.

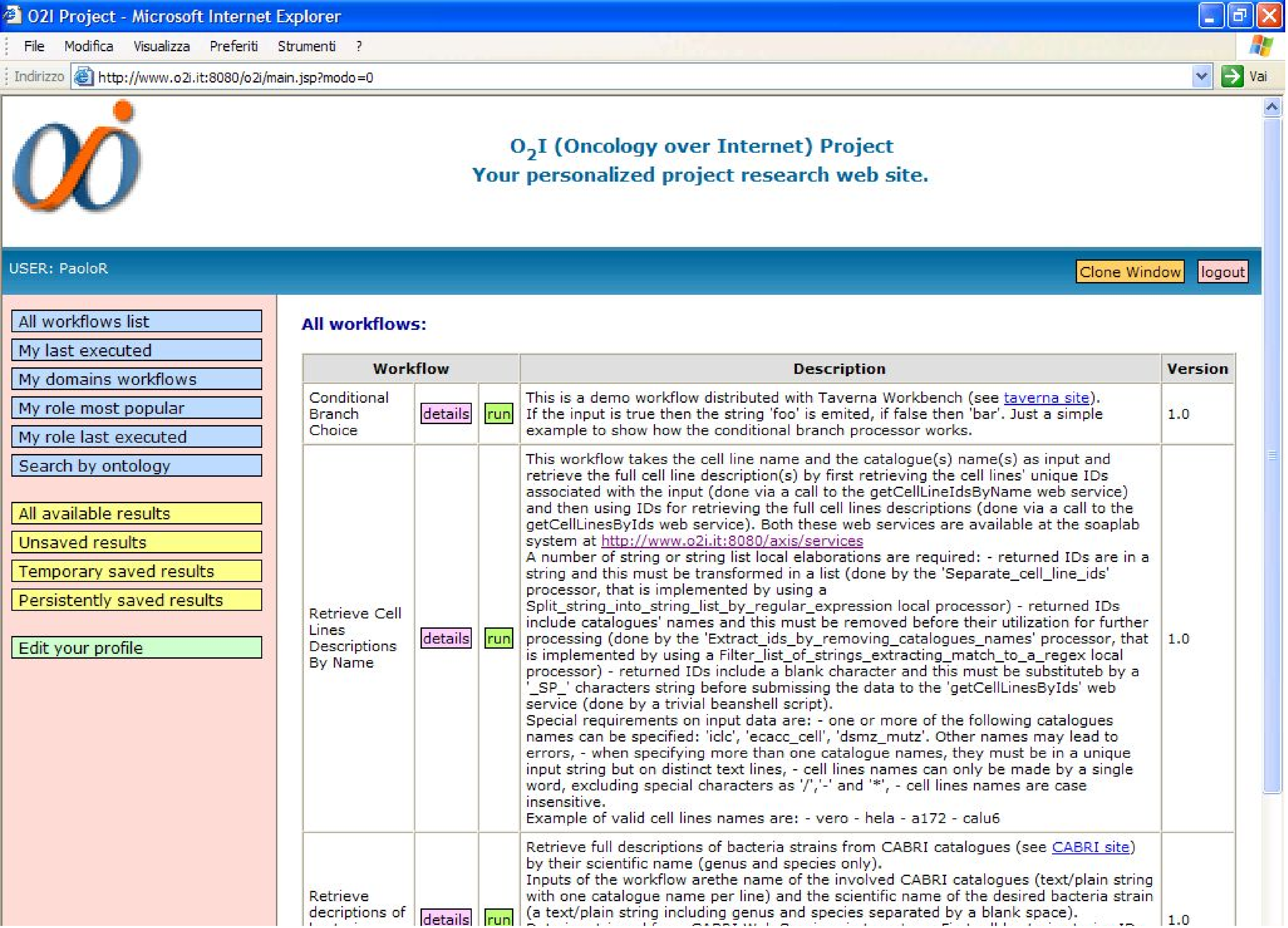
USER: PaoloR

Username:	PaoloR
Password:	•••••• (Must be six characters or more)
Re-type Password:	••••••
First Name:	Paolo
Last Name:	Romano
Role:	computer scientist - bioinformatician computer scientist - unspecified journalist patient physician (clinical oncologist)
E-Mail Address:	paolo.romano@istge.it
Domain:	Regulatory affairs and policies Research Scientific divulgation Translational research Virology

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

Choosing your password
You will use this information to access O_2I each time. Capitalization matters for your password!

Info about you
Please give us some information about you.



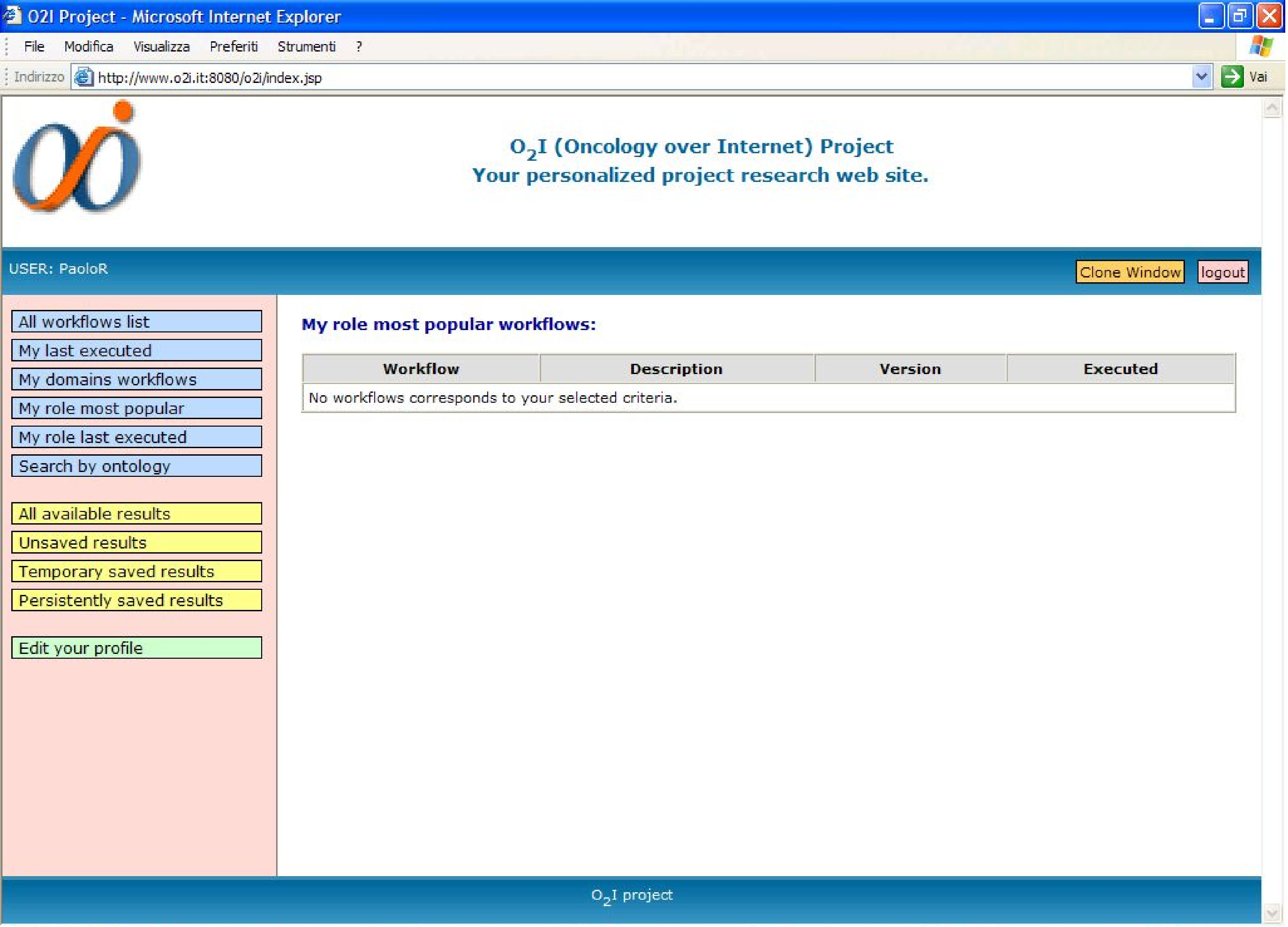
O_2I (Oncology over Internet) Project
Your personalized project research web site.

USER: PaoloR

All workflows:

Workflow			Description	Version
Conditional Branch Choice	details	run	This is a demo workflow distributed with Taverna Workbench (see taverna site). If the input is true then the string 'foo' is emited, if false then 'bar'. Just a simple example to show how the conditional branch processor works.	1.0
Retrieve Cell Lines Descriptions By Name	details	run	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	1.0
Retrieve decriptions of	details	run	Retrieve full descriptions of bacteria strains from CABRI catalogues (see CABRI site) by their scientific name (genus and species only). Inputs of the workflow arethe name of the involved CABRI catalogues (text/plain string with one catalogue name per line) and the scientific name of the desired bacteria strain (a text/plain string including genus and species separated by a blank space).	1.0

P. Romano, Naples, October 6, 2005

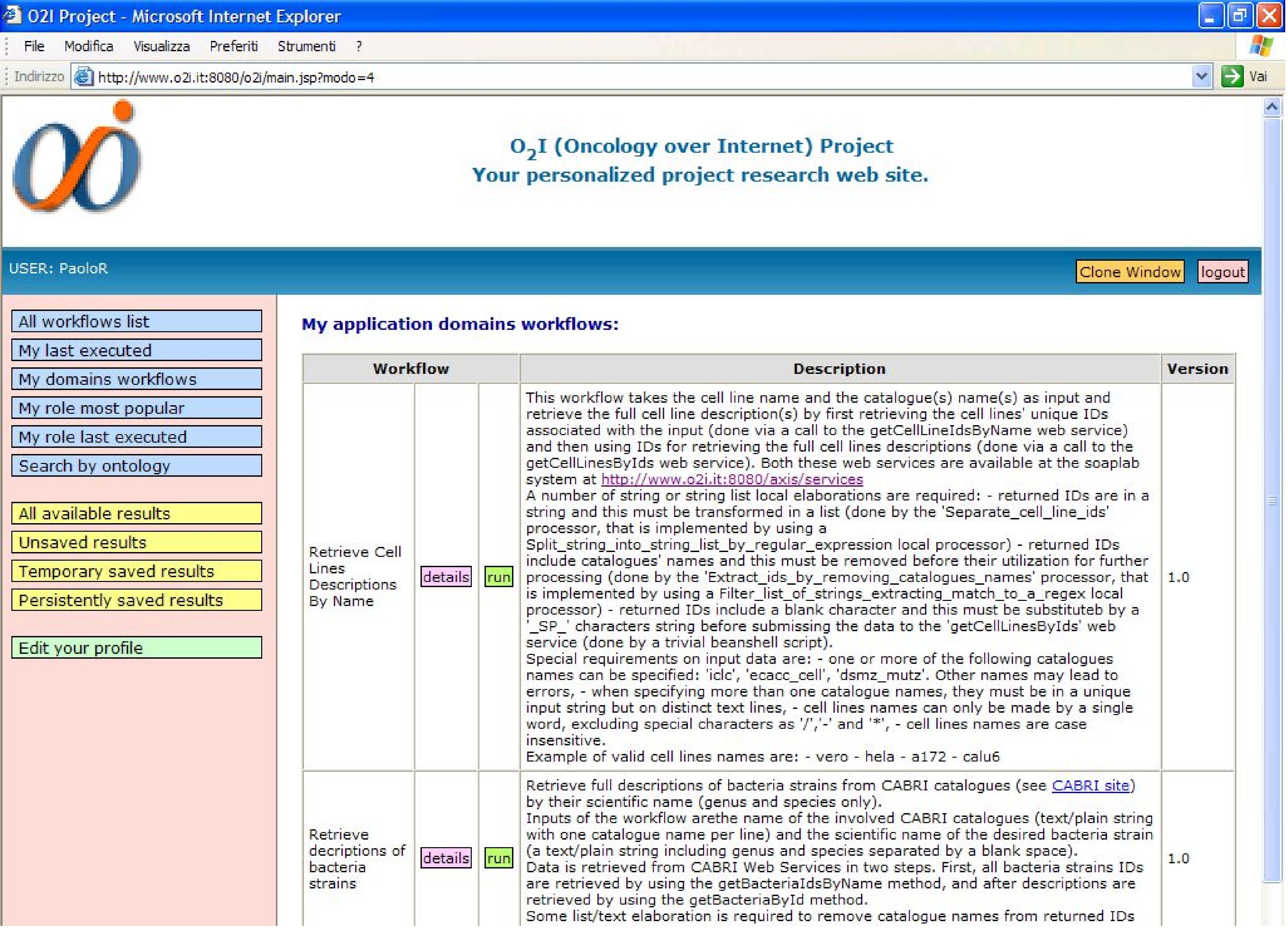


O_2I (Oncology over Internet) Project
Your personalized project research web site.

My role most popular workflows:

Workflow	Description	Version	Executed
No workflows corresponds to your selected criteria.			

O_2I project


O2I Project - Microsoft Internet Explorer
File
Modifica
Visualizza
Preferiti
Strumenti
?
Indirizzo
http://www.o2i.it:8080/o2i/main.jsp?modo=4
Vai

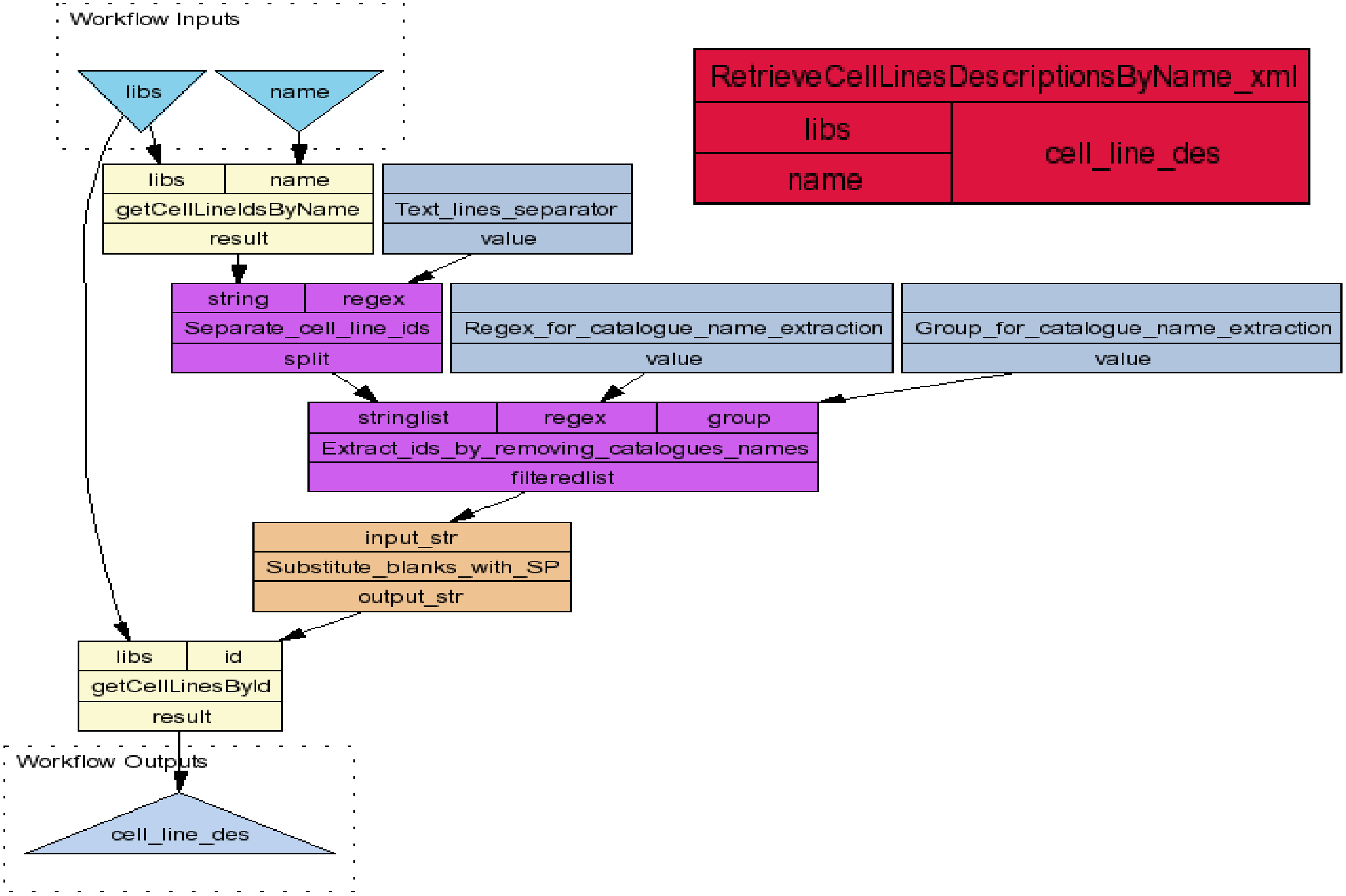
O$_2$I (Oncology over Internet) Project
Your personalized project research web site.

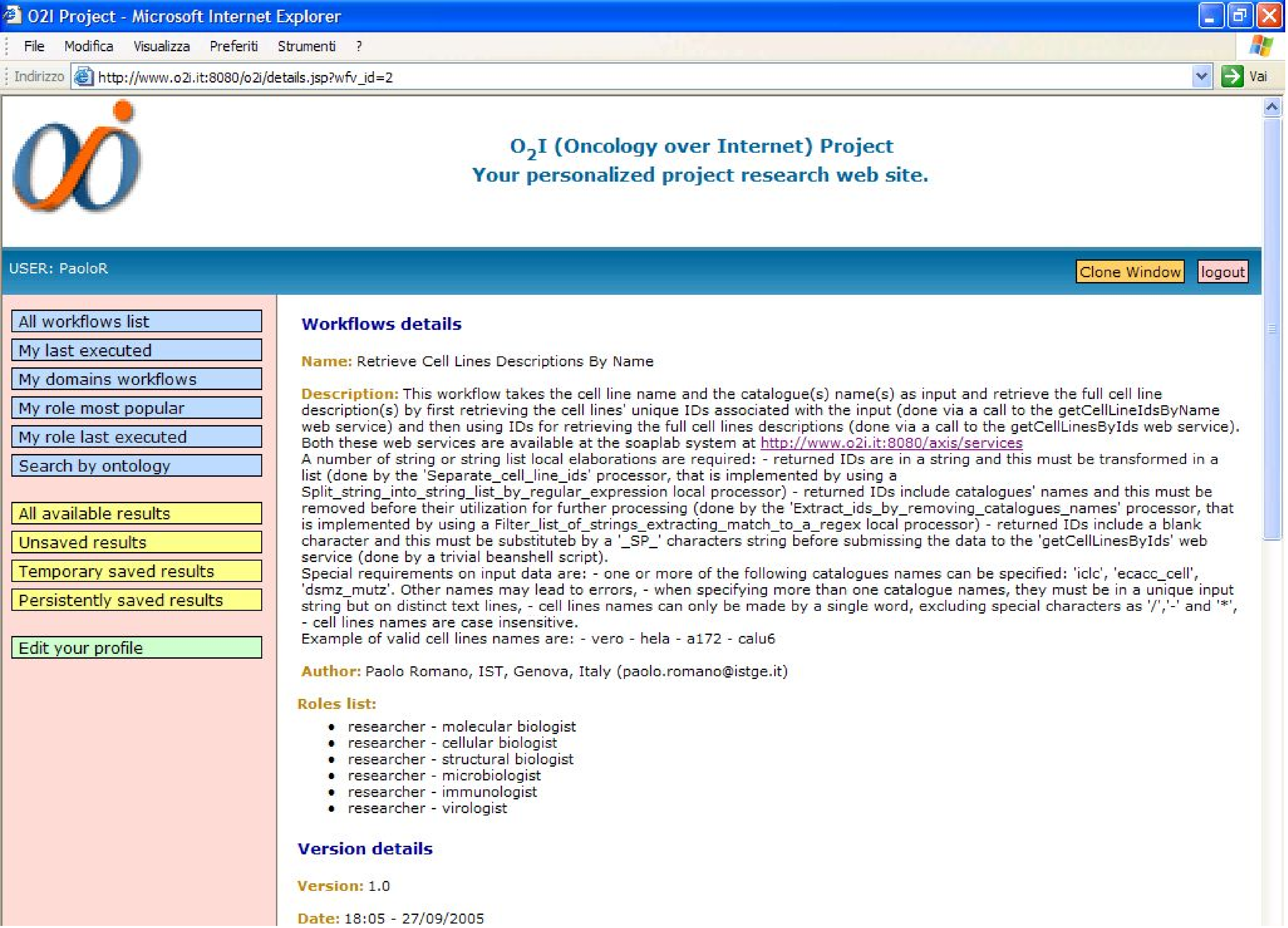
USER: PaoloR

My application domains workflows:

Workflow			Description	Version
Retrieve Cell Lines Descriptions By Name	details	run	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	1.0
Retrieve decriptions of bacteria strains	details	run	Retrieve full descriptions of bacteria strains from CABRI catalogues (see CABRI site) by their scientific name (genus and species only). Inputs of the workflow arethe name of the involved CABRI catalogues (text/plain string with one catalogue name per line) and the scientific name of the desired bacteria strain (a text/plain string including genus and species separated by a blank space). Data is retrieved from CABRI Web Services in two steps. First, all bacteria strains IDs are retrieved by using the getBacteriaIdsByName method, and after descriptions are retrieved by using the getBacteriaById method. Some list/text elaboration is required to remove catalogue names from returned IDs	1.0

Simple demo workflow





O_2I (Oncology over Internet) Project
Your personalized project research web site.

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)

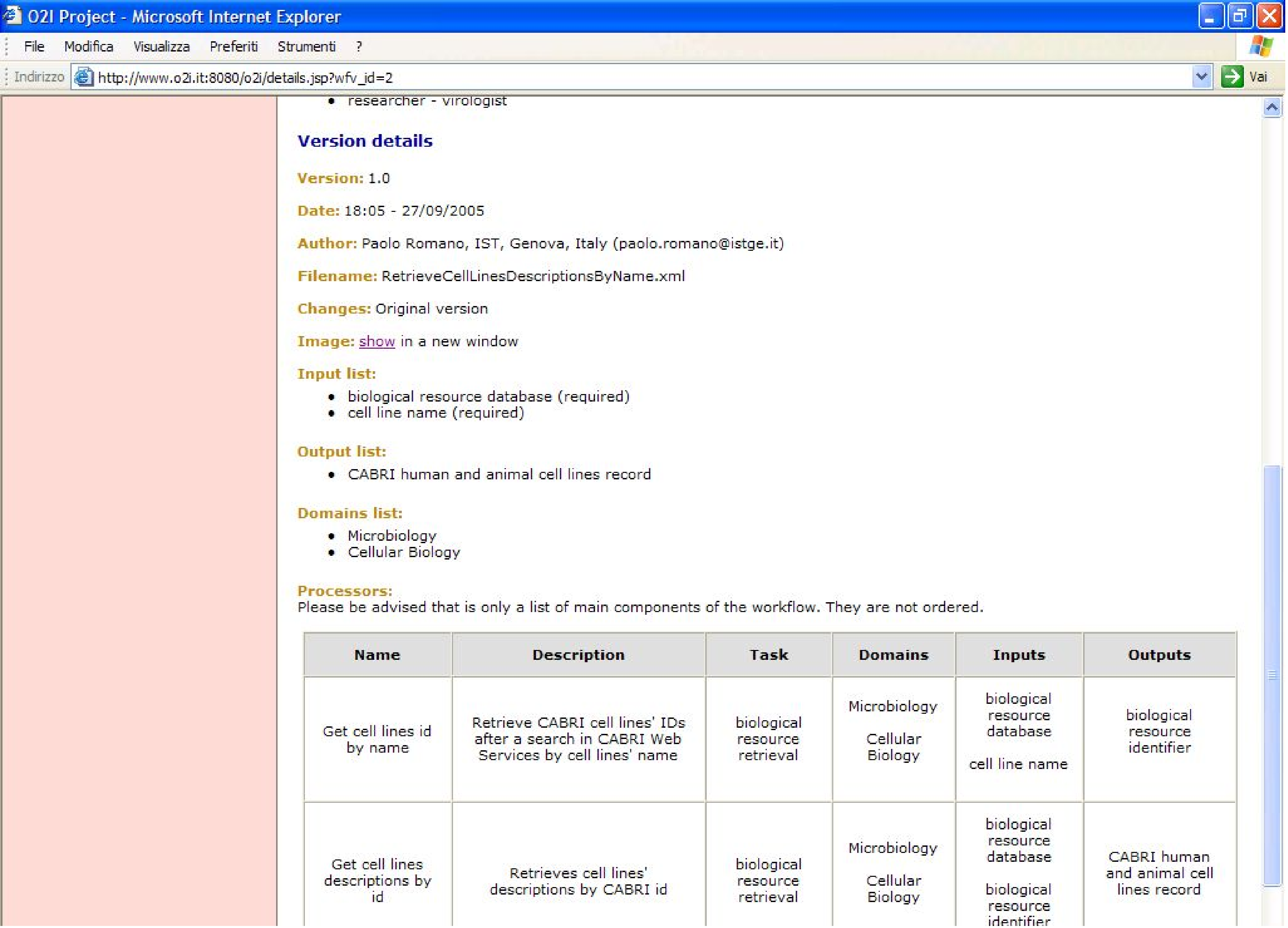
Roles list:

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

Version details

Version: 1.0

Date: 18:05 - 27/09/2005



http://www.o2i.it:8080/o2i/details.jsp?wfv_id=2

- researcher - virologist

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:

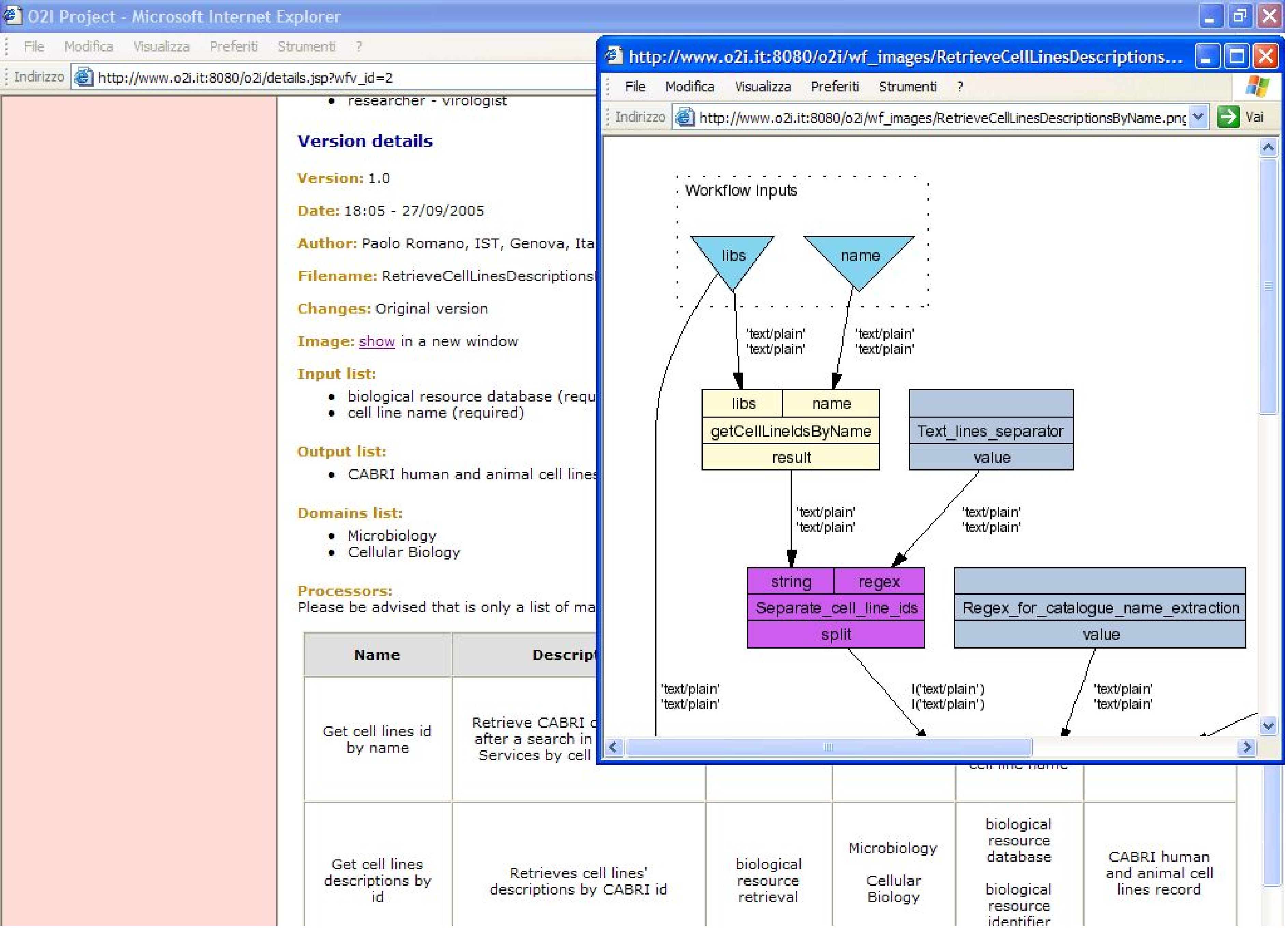
- CABRI human and animal cell lines record

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



O2I Project - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indirizzo http://www.o2i.it:8080/o2i/details.jsp?wfv_id=2
http://www.o2i.it:8080/o2i/wf_images/RetrieveCellLinesDescriptions...
File Modifica Visualizza Preferiti Strumenti ?
Indirizzo http://www.o2i.it:8080/o2i/wf_images/RetrieveCellLinesDescriptionsByName.pnc
Vai

- researcher - virologist

Version details

Version: 1.0

Date: 18:05 - 27/09/2005

Author: Paolo Romano, IST, Genova, Ita

Filename: RetrieveCellLinesDescriptions

Changes: Original version

Image: show in a new window

Input list:

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

Output list:

- CABRI human and animal cell lines

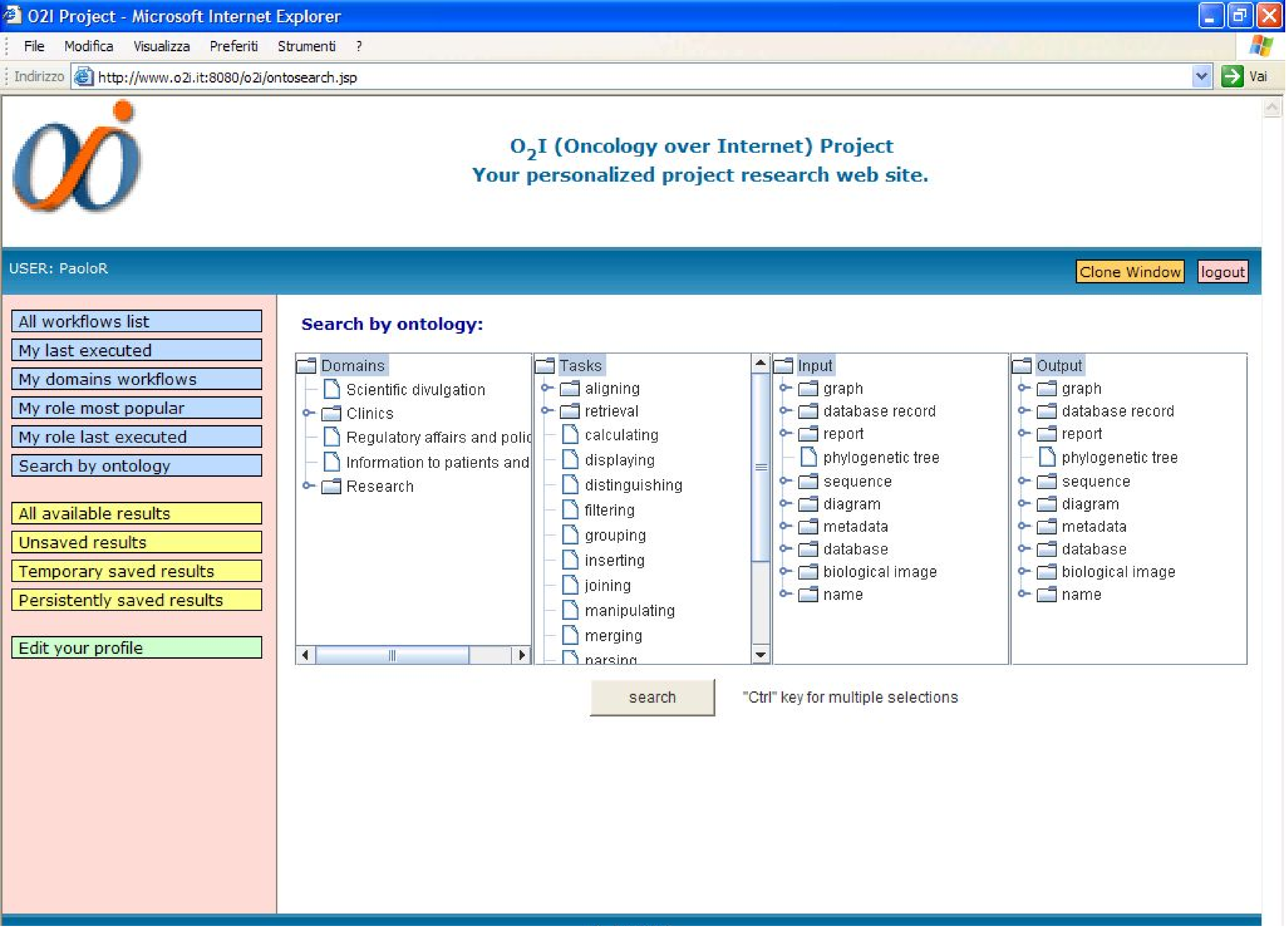
Domains list:

- Microbiology
- Cellular Biology

Processors:
Please be advised that is only a list of ma

Name	Descript				
Get cell lines id by name	Retrieve CABRI c after a search in Services by cell			cell line name	
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

P. Romano, Naples, October 6, 2005


O2I Project - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indirizzo http://www.o2i.it:8080/o2i/ontosearch.jsp
Vai

O$_2$I (Oncology over Internet) Project
Your personalized project research web site.

USER: PaoloR

- All workflows list
- My last executed
- My domains workflows
- My role most popular
- My role last executed
- Search by ontology

- All available results
- Unsaved results
- Temporary saved results
- Persistently saved results

- Edit your profile

Search by ontology:

Domains
- Scientific divulgation
- Clinics
- Regulatory affairs and polic
- Information to patients and
- Research

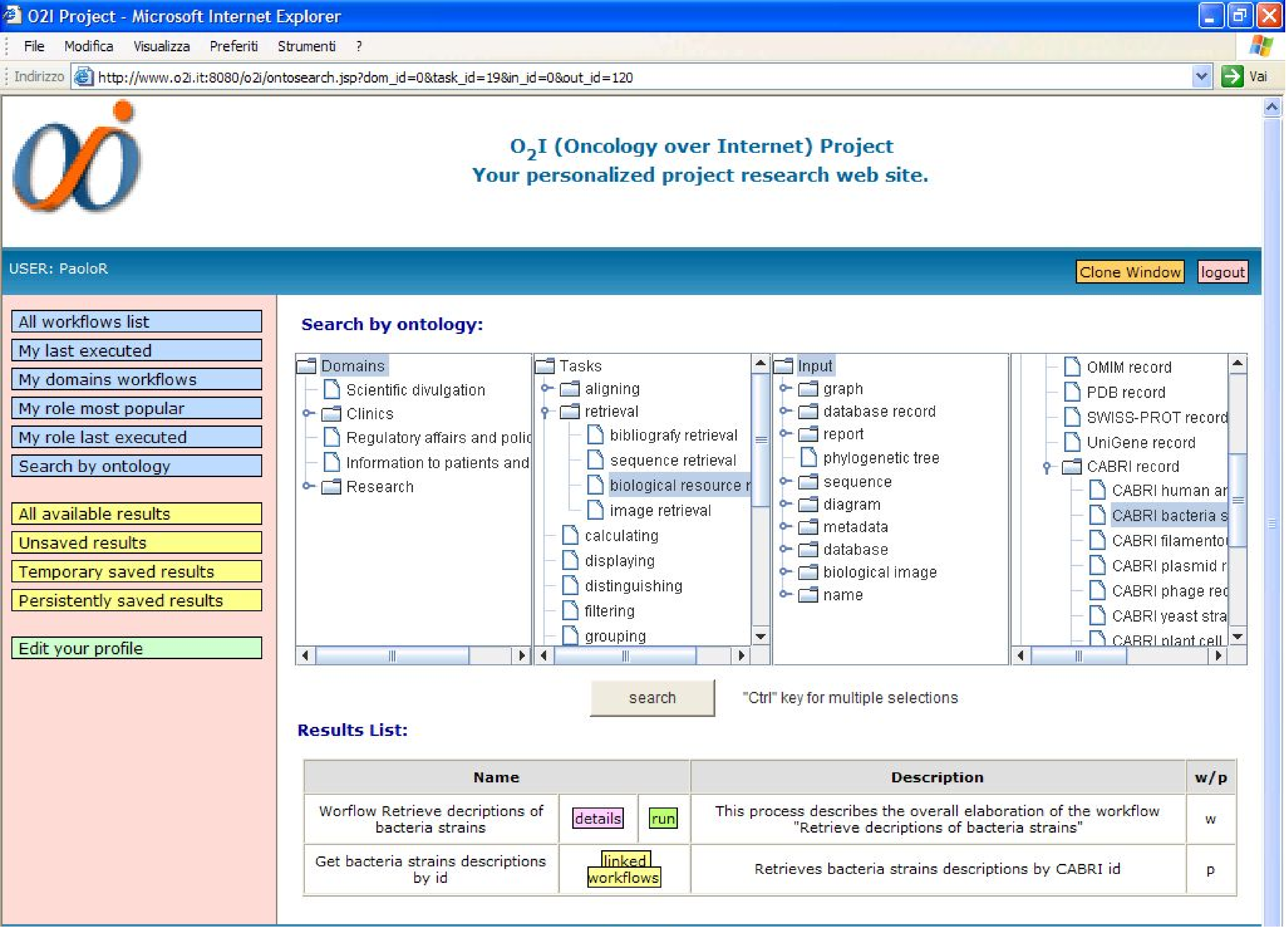
Tasks
- aligning
- retrieval
- calculating
- displaying
- distinguishing
- filtering
- grouping
- inserting
- joining
- manipulating
- merging
- parsing

Input
- graph
- database record
- report
- phylogenetic tree
- sequence
- diagram
- metadata
- database
- biological image
- name

Output
- graph
- database record
- report
- phylogenetic tree
- sequence
- diagram
- metadata
- database
- biological image
- name

search

"Ctrl" key for multiple selections



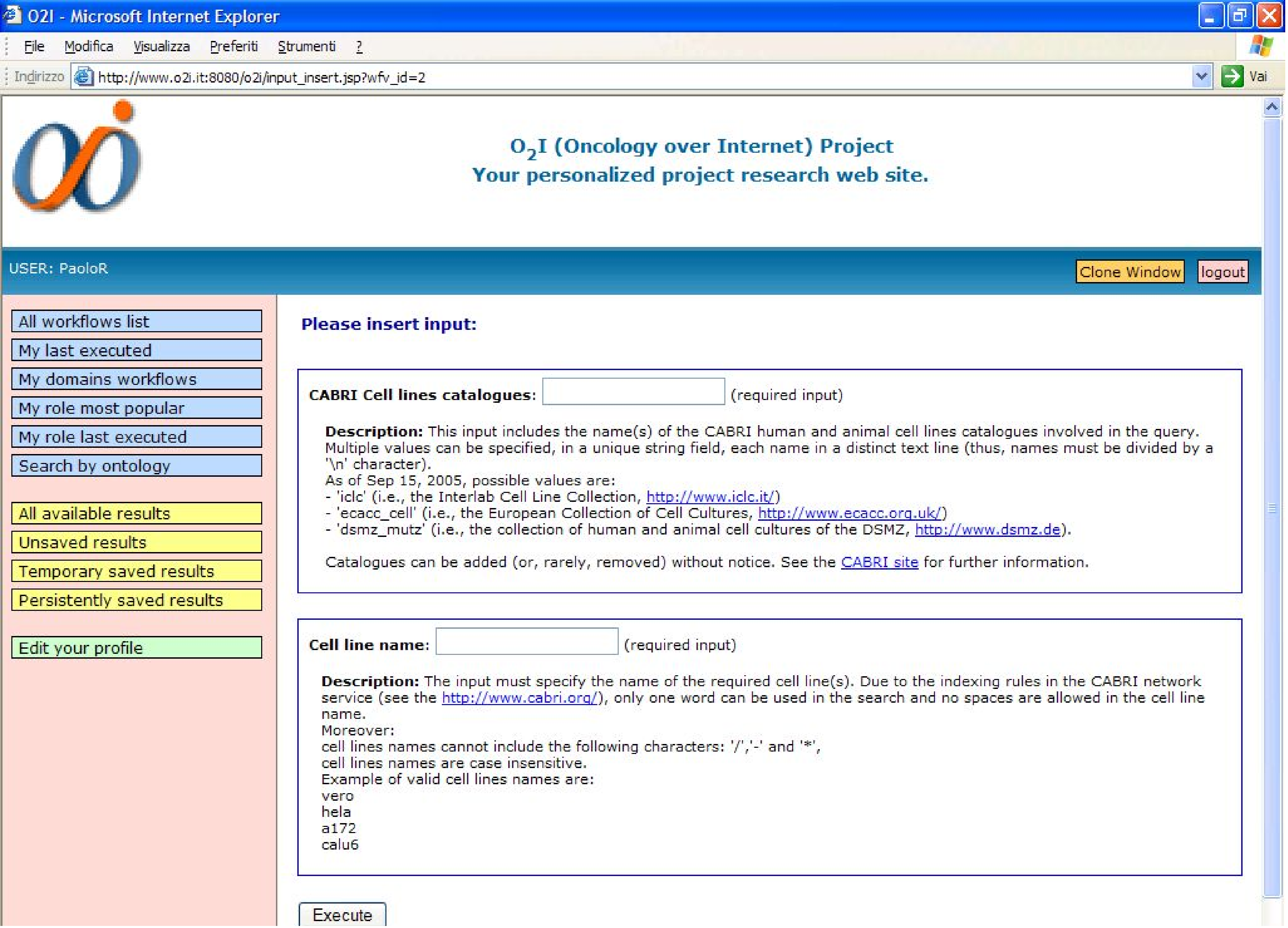
O_2I (Oncology over Internet) Project
Your personalized project research web site.

Search by ontology:

"Ctrl" key for multiple selections

Results List:

Name			Description	w/p
Worflow Retrieve decriptions of bacteria strains	details	run	This process describes the overall elaboration of the workflow "Retrieve decriptions of bacteria strains"	w
Get bacteria strains descriptions by id	linked workflows		Retrieves bacteria strains descriptions by CABRI id	p



O_2I (Oncology over Internet) Project
Your personalized project research web site.

USER: PaoloR

Please insert input:

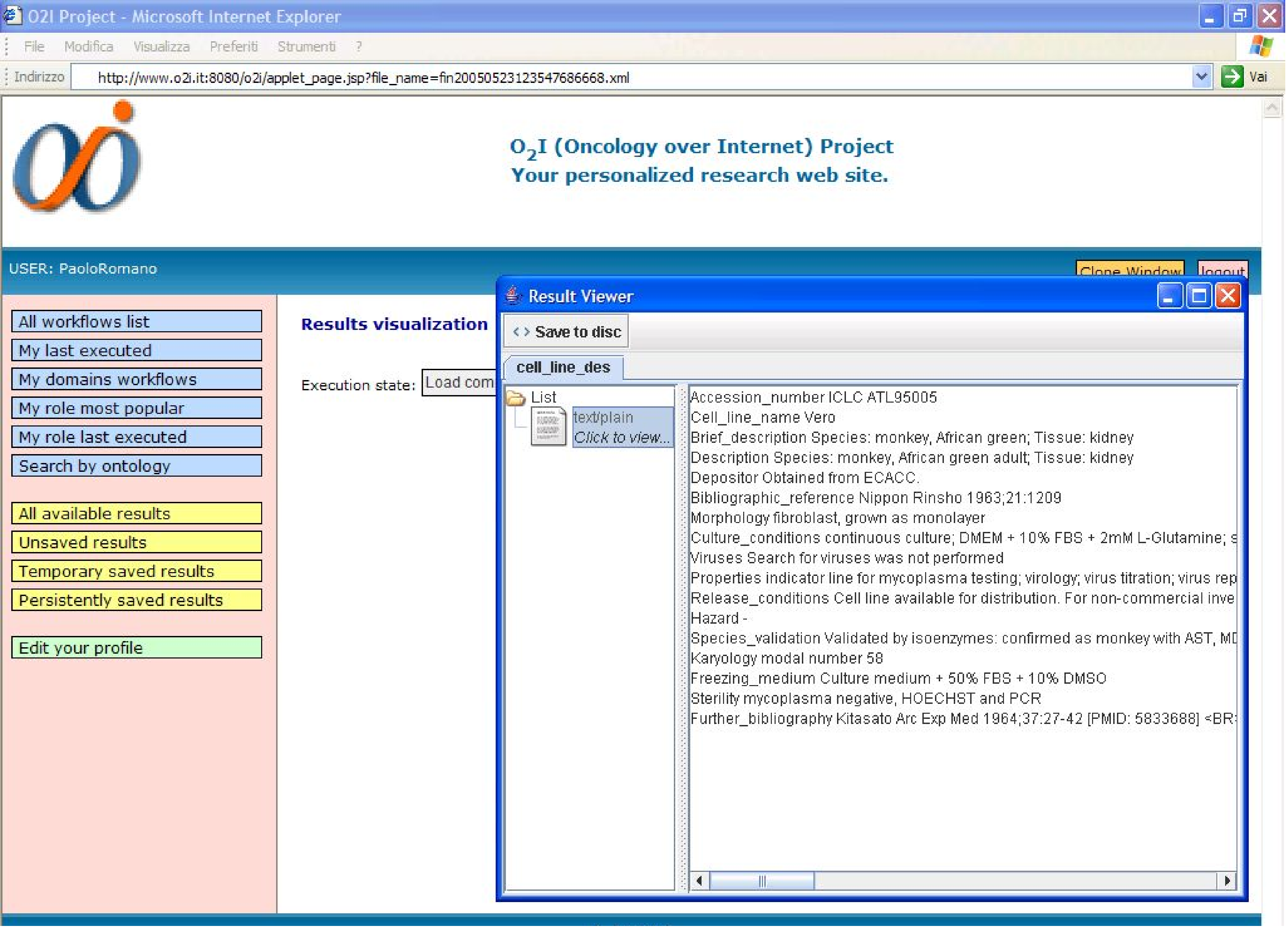
CABRI Cell lines catalogues: (required input)

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


O2I Project - Microsoft Internet Explorer
File
Modifica
Visualizza
Preferiti
Strumenti
?
Indirizzo
http://www.o2i.it:8080/o2i/applet_page.jsp?file_name=fin20050523123547686668.xml
Vai

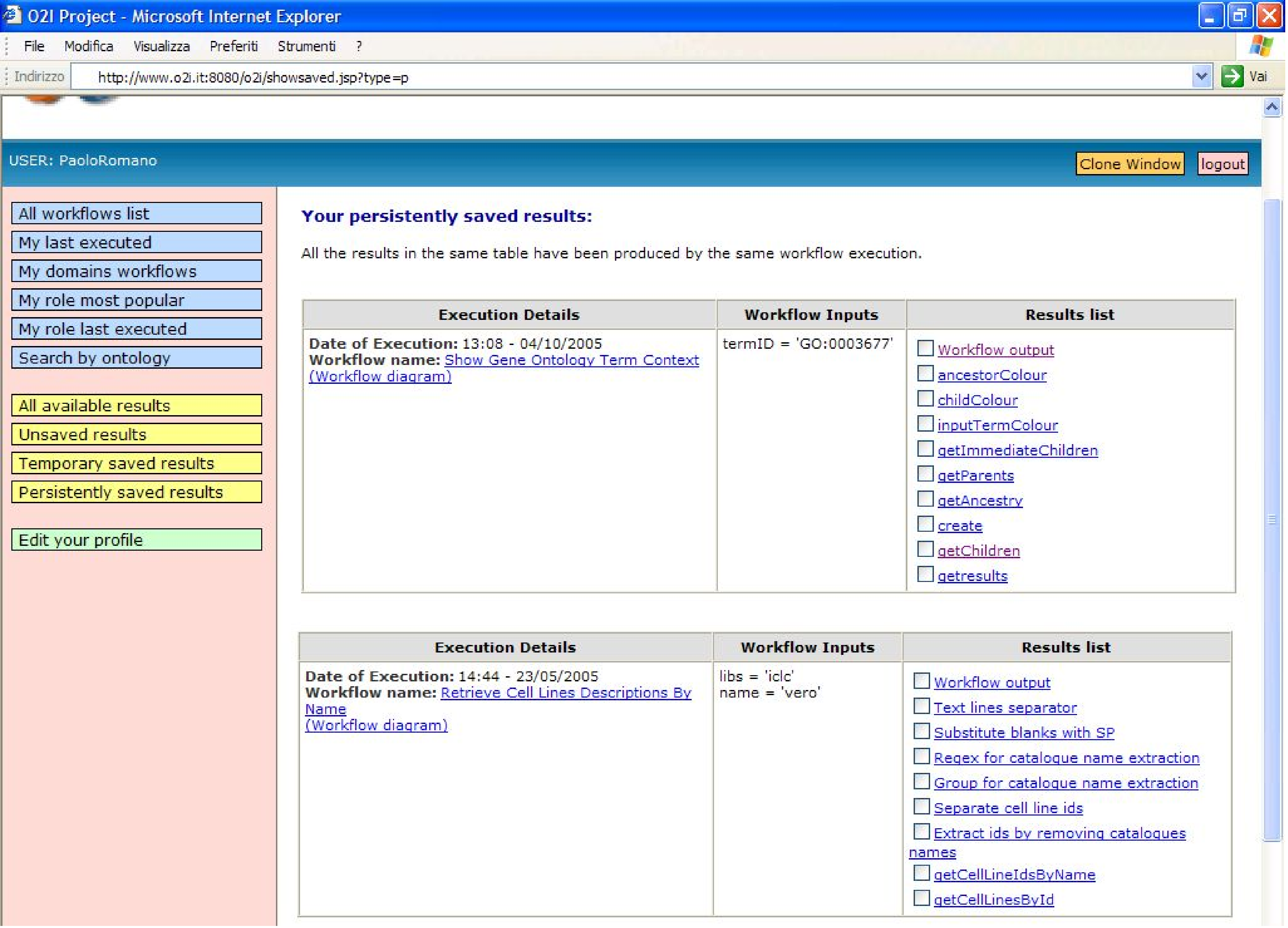
O$_2$I (Oncology over Internet) Project
Your personalized research web site.

USER: PaoloRomano

Results visualization

Execution state: Load com

```
Accession_number ICLC ATL95005
Cell_line_name Vero
Brief_description Species: monkey, African green; Tissue: kidney
Description Species: monkey, African green adult; Tissue: kidney
Depositor Obtained from ECACC.
Bibliographic_reference Nippon Rinsho 1963;21:1209
Morphology fibroblast, grown as monolayer
Culture_conditions continuous culture; DMEM + 10% FBS + 2mM L-Glutamine; s
Viruses Search for viruses was not performed
Properties indicator line for mycoplasma testing; virology; virus titration; virus rep
Release_conditions Cell line available for distribution. For non-commercial inve
Hazard -
Species_validation Validated by isoenzymes: confirmed as monkey with AST, MD
Karyology modal number 58
Freezing_medium Culture medium + 50% FBS + 10% DMSO
Sterility mycoplasma negative, HOECHST and PCR
Further_bibliography Kitasato Arc Exp Med 1964;37:27-42 [PMID: 5833688] <BR>
```



USER: PaoloRomano

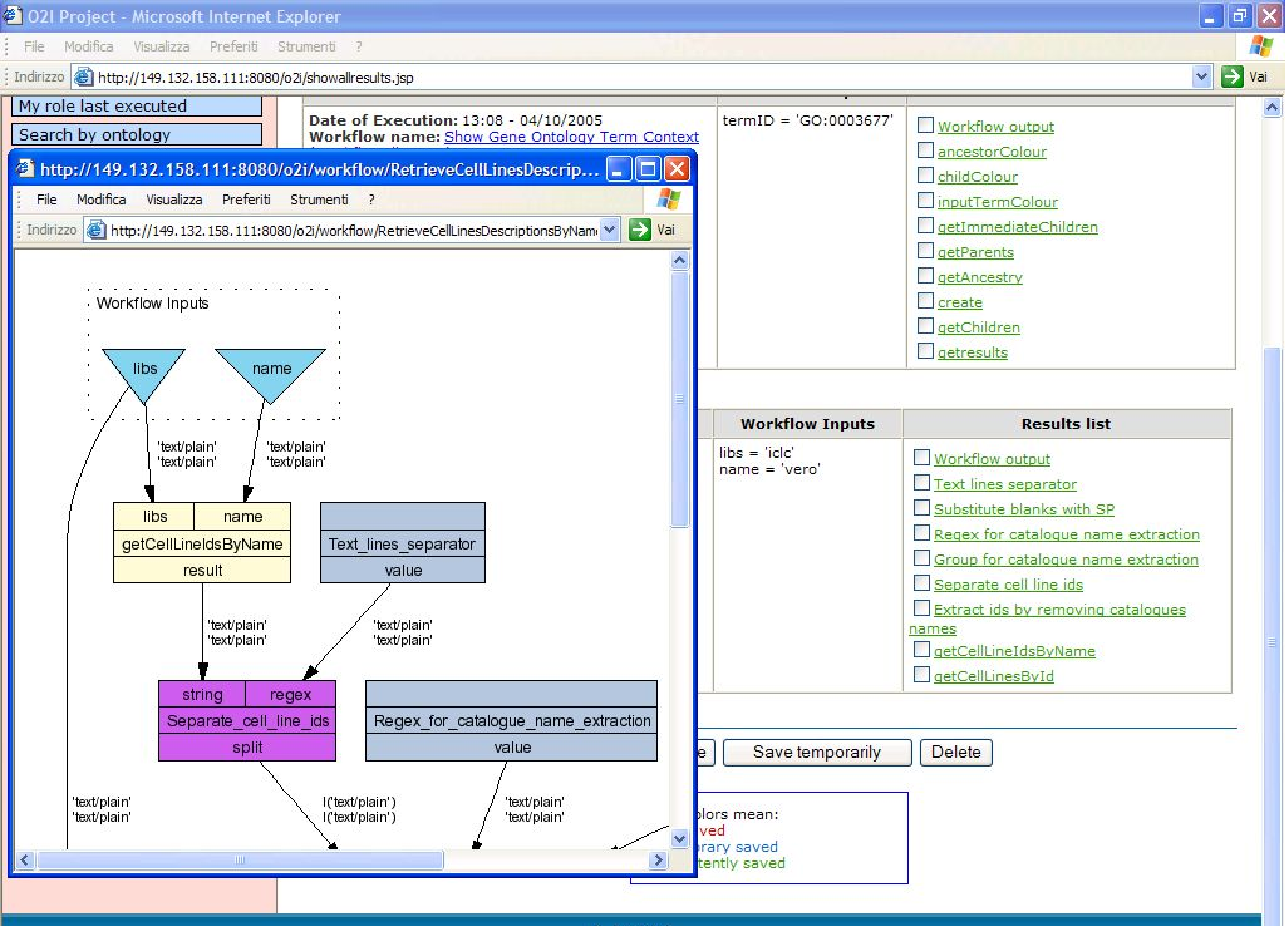
Your persistently saved results:

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

Execution Details	Workflow Inputs	Results list
Date of Execution: 13:08 - 04/10/2005 **Workflow name:** Show Gene Ontology Term Context (Workflow diagram)	termID = 'GO:0003677'	□ Workflow output □ ancestorColour □ childColour □ inputTermColour □ getImmediateChildren □ getParents □ getAncestry □ create □ getChildren □ getresults

Execution Details	Workflow Inputs	Results list

Date of Execution: 14:44 - 23/05/2005
Workflow name: Retrieve Cell Lines Descriptions By Name
(Workflow diagram)
	libs = 'iclc'
	name = 'vero'
		□ Workflow output
		□ Text lines separator
		□ Substitute blanks with SP
		□ Regex for catalogue name extraction
		□ Group for catalogue name extraction
		□ Separate cell line ids
		□ Extract ids by removing catalogues names
		□ getCellLineIdsByName
		□ getCellLinesById



http://149.132.158.111:8080/o2i/showallresults.jsp

My role last executed

Search by ontology

Date of Execution: 13:08 - 04/10/2005
Workflow name: Show Gene Ontology Term Context

termID = 'GO:0003677'

- □ Workflow output
- □ ancestorColour
- □ childColour
- □ inputTermColour
- □ getImmediateChildren
- □ getParents
- □ getAncestry
- □ create
- □ getChildren
- □ getresults

http://149.132.158.111:8080/o2i/workflow/RetrieveCellLinesDescriptionsByNam

Workflow Inputs	Results list

libs = 'iclc'
name = 'vero'
	□ Workflow output
	□ Text lines separator
	□ Substitute blanks with SP
	□ Regex for catalogue name extraction
	□ Group for catalogue name extraction
	□ Separate cell line ids
	□ Extract ids by removing catalogues names
	□ getCellLineIdsByName
	□ getCellLinesById

olors mean:
ved
rary saved
tently saved

Some acknowledgements

IST, Genoa
Paolo Romano,
Ulrich Pfeffer,
Domenico Marra,
Valentina Mirisola,
M. Assunta Manniello

ISMAC, CNR, Genoa
Patrizio Arrigo,
Matteo Fattore

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 (MC)
Emanuela Merelli

This work has partially been supported by the Italian Ministry for Education, University and Research (MIUR), project "Oncology over Internet" (2002 – 2005).