

Workflow management systems in supporto alla biologia: verso l'automazione dell'accesso ai dati e alla loro analisi in rete

Paolo Romano
Istituto Nazionale per la Ricerca sul Cancro
(paolo.romano@istge.it)

Sommario

- Integrazione dei dati in biologia
- XML e Web Services
- Soaplab ed esempi di implementazione
- Workflow management systems
- Taverna Workbench
- Portale biowep

L'informazione biologica

- La ricerca biomedica produce una quantità sempre crescente di dati
- Alcuni settori, quali la genomica e la proteomica, contribuiscono alla realizzazione di banche dati di rilevanti dimensioni
- Altri settori emergenti, legati all'analisi delle mutazioni, ai polimorfismi, al metabolismo, e derivati da nuovi strumenti quali i microarray, contribuiranno anch'essi con quantità di dati ancora superiori

Dimensione informazioni biologia

- EMBL Data Library 82 (Mar 2005):
 - Sequenze: 49,474,402 (Basi: 85,134,714,382)
- EMBL Data Library 85 (Dic 2005):
 - Sequenze: 64,739,883 (Basi: 116,106,677,726)
- EMBL Data Library 86 (Mar 2006):
 - Sequenze: 69,783,593 (Basi: 126,401,347,060)
 - Aumento: +7,79% (+8,87%) dalla precedente release
 - Aumento: +41,05% (+48,47%) in un anno

Si veda: <http://www3.ebi.ac.uk/Services/DBStats/>

Dimensione informazioni biologia

- ArrayExpress (21/12/2005)
esperimenti 1,187 (circa 800 Gb)
+100% da Ottobre 2004 a Ottobre 2005
(<http://www.ebi.ac.uk/arrayexpress/Help/stats/index.html>)
- Il numero speciale di Nucleic Acids Research del 2005
ha listato 724 database di biologia molecolare
(http://nar.oxfordjournals.org/cgi/content/full/33/suppl_1/D5/DC1
)
- La lista di siti SRS pubblici comprende 1,300 librerie
(<http://downloads.lionbio.co.uk/publisrs.html>)

Banche dati eterogenee

- Alcune banche dati sono gestite da pochi Centri (EBI, NCBI, DDBJ) in modo disomogeneo
- Le banche dati secondarie sono di ottima qualità (annotazione estesa, controllo accurato)
- Molte banche dati sono specializzate: gene/genoma, organismo, malattia, mutazione, etc...
- Molte banche dati sono sviluppate da singoli ricercatori o piccolo gruppi
- Le banche dati sono distribuite:
 - strutture dati, DBMS e metodi di distribuzione sono differenti
 - informazioni e significati (semantica) sono differenti

Software per analisi biologiche

- software specialistici sono essenziali per quasi tutte le analisi dati in biologia molecolare:
 - analisi di sequenza, predizione di struttura, predizione di siti di splicing, analisi dell'evoluzione molecolare, ricostruzione sequenze, etc...
- è necessario integrare questi software con le banche dati
 - banche dati come input dei software
 - risultati elaborazione/analisi come nuovi dati da memorizzare e analizzare

Obiettivi dell'integrazione

- L'integrazione dei dati e l'automazione dei processi sono necessari per:
 - Ottenere una visione complessiva e più precisa di tutte le informazioni disponibili
 - Eseguire automaticamente interrogazioni e/o analisi che coinvolgono più database e software
 - Eseguire con efficienza analisi che coinvolgono grosse quantità di dati
 - Realizzare un effettivo data mining

Integrazione: longevità

- L'integrazione necessita di stabilità:
 - Buona conoscenza e comprensione del dominio
 - Buona definizione dei dati
 - Standardizzazione
 - Obiettivi ben definiti
- L'integrazione teme:
 - Incertezza nella comprensione del dominio
 - Eterogeneità dei dati e dei sistemi
 - Specializzazione dei dati
 - Rapida evoluzione dei dati
 - Spontaneità, sperimentalismo
 - Mancanza di obiettivi predefiniti

Specificità dell'integrazione

- In ambito biologico:
 - Una pre-analisi e riaggregazione delle informazioni è impossibile, perché dati e conoscenze teoriche cambiano frequentemente e rapidamente
 - La complessità delle informazioni rende molto difficile creare modelli validi in diversi ambiti e che possano rimanere tali nel tempo
 - La disponibilità di strumenti assestati riduce le possibilità di implementare standard comuni
 - Le esigenze e gli obiettivi di ricerca evolvono rapidamente, seguendo le nuove acquisizioni e teorie

Specificità dell'integrazione

L'integrazione deve quindi essere sviluppata utilizzando sistemi flessibili che siano facilmente adattabili ed espandibili

- XML per strutturare i dati
- Web Services per consentire un'efficiente interazione tra computer
- Workflow management systems per automatizzare le procedure
- Portali “user friendly” per rendere i workflow accessibili a tutti i ricercatori

XML (eXtensible Markup Language)

- **Linguaggio Markup estensibile:**
 - Supera limiti HTML (orientato all'impaginazione)
 - Semplice definizione e implementazione nuovi documenti tramite Document Type Definitions (DTDs)
 - Modulare, nuovi DTD possono utilizzare i precedenti
- **Utilizzabile da applicazioni software:**
 - Corretto (Well formed, conforme allo standard XML)
 - Valido (conforme al DTD)
 - Definizioni standardizzate (namespaces)
 - Analisi sintattica ed estrazione dati automatiche

ID AA415057 standard; RNA; EST; 337 BP.
AC AA415057;
SV AA415057.1
DT 27-OCT-1997 (Rel. 53, Created)
DT 14-DEC-1999 (Rel. 62, Last updated, Version 2)
DE Mg0001 RCW Lambda Zap Express Library Pyricularia grisea cDNA clone RCW1
DE 5', mRNA sequence.
KW EST.
OS Magnaporthe grisea
OC Eukaryota; Fungi; Ascomycota; Pezizomycotina; Sordariomycetes;
OC Sordariomycetes incertae sedis; Magnaporthaceae; Magnaporthe.
RN [1]
RP 1-337
RA Wu S.-C., Bernstein B.D., Darvill A.G., Albersheim P.;
RT "Expressed sequence tags of the rice blast fungus grown on rice cell
RT walls";
RL Unpublished.
DR UNILIB; 863; 863.
CC Contact: Sheng-Cheng Wu
CC CCRC
CC University of Georgia
CC 220 Riverbend Road, Athens, GA 30602-4712, USA
CC Tel: 706 542 4446
CC Fax: 706 542 4412
CC Email: wusc@bscr.uga.edu
CC Seq primer: T3.

```
FH      Key      Location/Qualifiers
FH
FT      source    1..337
FT              /db_xref="taxon:148305"
FT              /db_xref="UNILIB:863"
FT              /note="Vector: Lambda Zap; Messenger RNAs prepared from
FT              Magnaporthe grisea grown at 23C in the dark with constant
FT              gyratory shaking (100 rpm) in Vogel's medium containing
FT              0.5% isolated rice cell walls as the sole carbon source"
FT              /organism="Magnaporthe grisea"
FT              /strain="CP987"
FT              /clone="RCW1"
FT              /clone_lib="RCW Lambda Zap Express Library"
FT              /tissue_type="Mycelium"
FT              /dev_stage="Day 5 post-inoculation"
XX
SQ      Sequence 337 BP; 56 A; 111 C; 74 G; 96 T; 0 other;
      ctttttcaat cagcccgaga actcctggtt ggggttttctg cctgttctga cagctacttg      60
      tcatcgcata gcccgttctt tggttccaga taccacaagc ctgggacatt gatttcccag      120
      caactctttc aaaatggtat tattagcctc ctcacgatcc ctcgcgcggt cgcttggtcc      180
      ccttgcttca cgctctcgag cttttcagag cagtgctgct tcccgttccc tctcgactgc      240
      caccgctcgc ggccagggca aatccacaac tctcctgagg cccgcggccg ccacgaggac      300
      aagcagggttg ttgtcgactg gttcgcctt tcgtatt      337
//
```

```
<!ELEMENT interpro (name, type, examplelist, memberlist, publist, parlist*, chlist*,
seclist*, abstract)>

<!ELEMENT name (#PCDATA)>
<!ELEMENT type (#PCDATA)>
<!ELEMENT abstract (#PCDATA|cite|dbxref|sub|sup|p|li|i|ol|reaction|pre)*>
<!ELEMENT examplelist (example*)>
<!ELEMENT example (#PCDATA|protein|dbxref|cite)*>
<!ELEMENT publist (publication*)>
<!ELEMENT memberlist (dbxref*)>
<!ELEMENT protein (#PCDATA|protein)*>

<!ATTLIST interpro
  id          ID          #REQUIRED>
<!ATTLIST dbxref
  db          CDATA #IMPLIED
  dbkey       CDATA #IMPLIED
  name        CDATA #IMPLIED>
<!ATTLIST protein
  sptr_ac     CDATA #REQUIRED
  status      (?|T|P|F|N) #IMPLIED
  start       CDATA #IMPLIED
  end         CDATA #IMPLIED>
```

DTD per Interpro

```
<interpro id="IPR000002">
  <name>FIZZY/CDC20 domain</name>
  <type>Domain</type>
  <abstract> This domain is found in proteins ...</abstract>
  <examplelist>
    <example>
      <protein sptr_ac="Q12834" />Mammalian protein, p55CDC
    </example>
    <example>
      <protein sptr_ac="Q09649" />
    </example>
  </examplelist>
  <publist>
    <publication pub_id="PUB00006167">
      <authorlist>Shirayama M., Toth A., Galova M., Nasmyth K.</authorlist>
      <title>APC(Cdc20) promotes exit from mitosis by .....</title>
      <dbxref db="MEDLINE" dbkey="20110935" />
      <journal>Nature</journal>
      <location firstpage="203" lastpage="207" volume="402" />
      <year>1999</year>
    </publication>
  </publist>
  <memberlist>
    <dbxref db="PREFILE" dbkey="PS50218" name="FIZZY_DOMAIN" />
    <dbxref db="PRODOM" dbkey="PD004563" name="PD004563" />
  </memberlist>
</interpro>
```


Linguaggi XML in biologia

- Sequenze
 - Bioinformatic Sequence Markup Language (BSML)
 - Agave
- Proteine (SPML)
- NCBI outputs (BlastXML)
- Microarray (MAGE-ML)
- Systems Biology Markup Language (SBML)
- Variabilità individuale
 - Biological Variation Markup Language (BVML)

Web Services

- Interfacce programmatiche (API) per l'accesso a servizi di rete basate su XML e protocolli standard di trasporto (HTTP, SOAP)
- Consentono alle applicazioni di accedere ai dati in maniera intelligente: individuazione dei contenuti e comprensione semantica
- Sono disponibili standard per la loro descrizione (WSDL), identificazione (UDDI) e composizione (WSFL)
- Utilizzano dati identificativi e descrittivi delle informazioni e dei servizi offerti (metadata)

WSDL: descrivere il WS

Web Services Description Language (WSDL)

- Standard per la descrizione dei Web Services
- Comprende dettagli per l'accesso concreti: localizzazione e modalità di accesso
- Comprende funzionalità astratte
- Implementazioni di WSDL per: SOAP, HTTP, MIME

UDDI: Registrare i servizi

Universal Description, Discovery and Integration (UDDI)

- Realizzazione di un framework per la descrizione dei Web Services, indipendente da HW e SW
- Consente la creazione di registri di Web Services
- Basato su standard World Wide Web Consortium (W3C) and Internet Engineering Task Force (IETF)
- Esistono alternative: bioMOBY

WSFL: Comporre servizi complessi

Web Services Flow Language (WSFL)

- Consente di descrivere insiemi di web services
 - Flow models: specifica come utilizzare un insieme di web services per raggiungere un certo obiettivo
 - Global models: descrive le interazioni tra più web services e il comportamento globale di un insieme
- Recursive composition: ogni model (flow o global) è un nuovo Web Service e può rientrare in altri modelli
- Esistono alternative:
 - WSCDL (Web Services Choreography Description Language)
 - Scufl (Simple Conceptual Unified Flow Language)
 - XPDL (XML Processes Description Language)

Web Services in bioinformatica

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

- GeneCruiser, Biosphere (microarray)
- SIMAP (proteine)

- Cataloghi CABRI (risorse biologiche)
- Mutazioni TP53

Common Access to Biological Resources and Information

Obiettivi

- Distribuzione di materiali biologici di qualità
- Linee Guida per la conservazione del materiale
- Cataloghi integrati tramite SRS
- Shopping cart

Partners

- 12 CRB europei, 28 collezioni + IST

Materiali

- Microrganismi (Batteri, lieviti, funghi filiformi)
- Linee cellulari animali e umane, ibridomi, linee B tip. HLA
- Plasmidi, fagi, virus, sonde DNA
- Complessivamente più di 116.000 risorse

IARC TP53 database

IARC TP53 Mutation Database

<http://www-p53.iarc.fr/>

- Release 9: 19,809 mutazioni somatiche, 1,769 articoli
- Informazioni: mutazione, materiale, stile di vita
- Vocabolari e annotazioni standard
- Ricerche on-line richiedono interazione

Implementazione SRS del database IARC TP53

<http://srs.o2i.it/srs71/>

- Basato su SRS
- Definizione di un DTD ad-hoc
- Trasferimento dati basato su XML
- Accesso programmatico semplificato

Web Services per CABRI e TP53

Implementare Web Services che consentano:

- L'accesso ai database CABRI e TP53 tramite i siti SRS
- La possibilità di includere questi task in workflow complessi

Riproducendo il comportamento attuale:

- Ricerca per nome, identificatore e a testo libero (CABRI)
- Ricerca per funzioni e proprietà (TP53)
- Combinare risultati
- Integrare i dati con altri sorgenti tramite ID o termini comuni

Due tipologie di Service:

- Ricerca per una specifica proprietà e restituzione degli ID
- Ricerca per ID e restituzione del record completo

Soaplab: SOAP-based Analysis Web Service

“Soaplab is a set of Web Services providing a programatic access to some applications on remote computers. It is often referred to as an **Analysis (Web) Service**” (Martin Senger, EBI).

Consente di implementare Web Services in grado di accedere a:

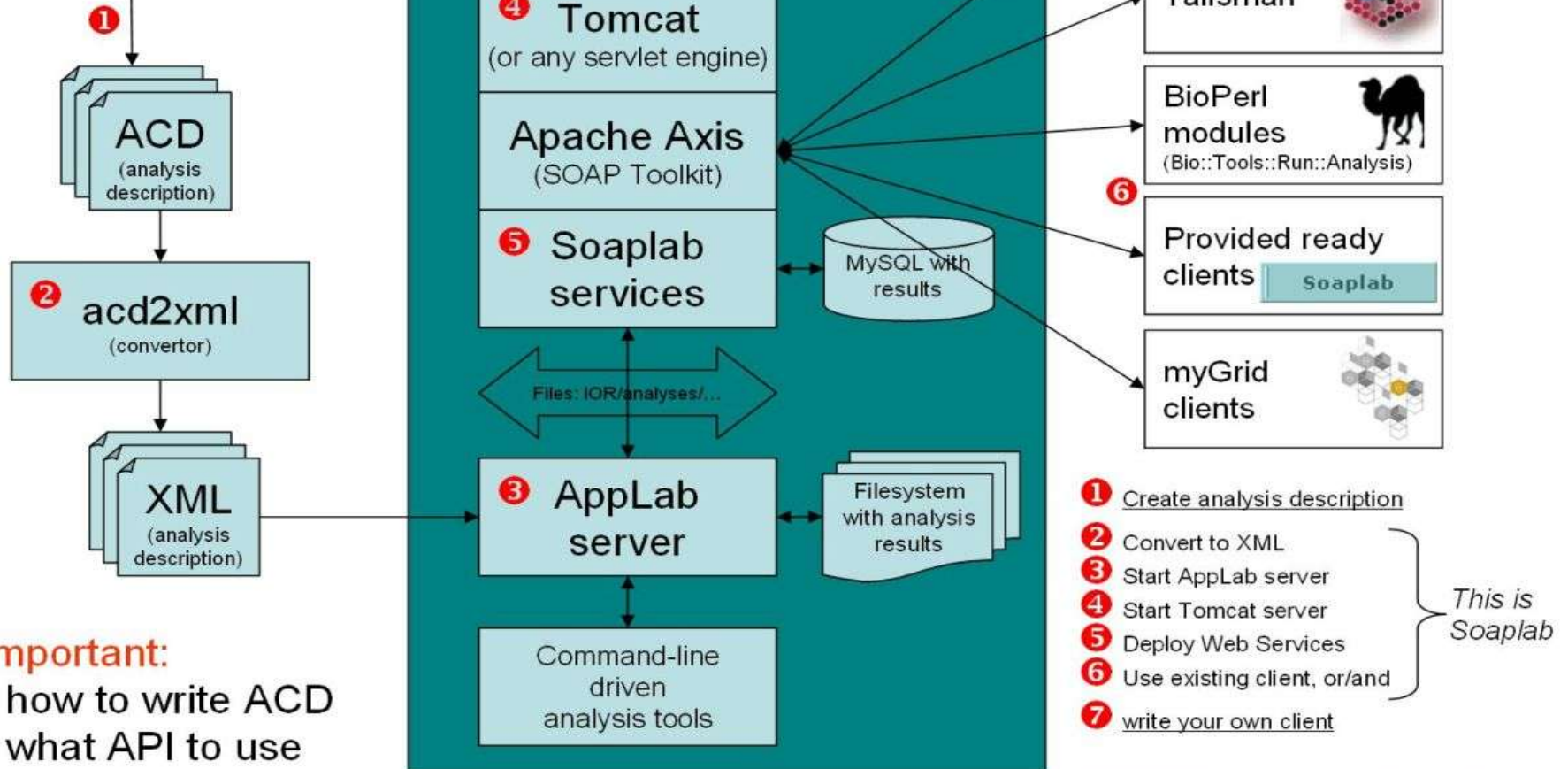
- Applicazioni locali eseguibili con “command-line”
- Applicazioni presenti in EMBOSS
- Contenuto di qualunque pagina Web (GowLab)

Requisiti

- Apache Tomcat servlet engine, Axis SOAP toolkit, Java
- perl, MySQL



Graphically...



Important:

- how to write ACD
- what API to use by the clients

Download: soaplab-clients-version

Download: applab-cembalo-soaplab-version (for EMBOSS)

Download: analysis-interfaces-version

<http://industry.ebi.ac.uk/soaplab/dist/>

Binary
distribution

Soaplab: getCellLineIdsByName

```
appl: getCellLineIdsByName [  
  documentation: "Get cell lines by name from CABRI human and animal  
    cell lines catalogues (see www.cabri.org)"  
  groups: "CABRI"  
  nonemboss: "Y"  
  comment: "launcher get"  
  supplier: "http://www.cabri.org/CABRI/srs-bin/wgetz"  
  comment: "method [{libs}-nam:'$name'] -ascii"  
]  
  
string: libs [ parameter: "Y" ]  
string: name [ parameter: "Y" ]  
  
outfile: result [ ]
```

Soaplab: getCellLinesById

```
appl: getCellLinesById [  
  documentation: "Get cell lines by Id from CABRI human and  
    animal cell lines catalogues (see www.cabri.org)"  
  groups: "CABRI"  
  nonemboss: "Y"  
  comment: "launcher get"  
  supplier: "http://www.cabri.org/CABRI/srs-bin/wgetz"  
  comment: "method -e [{ $libs }:'$id'] -ascii"  
]
```

```
string: libs [ parameter: "Y" ]  
string: id [ parameter: "Y" ]
```

```
outfile: result [ ]
```

Web Services CABRI

Web Service Name	Involved catalogues	Input	Output
getBacterialIdsByName	Bacteria strains	lib(s), name	id(s)
getBacterialIdsByProperty	Bacteria strains	lib(s), text	id(s)
getBacteriaById	Bacteria strains	id	full record
getFungiIdsByName	Filamentous fungistrains	lib(s), name	id(s)
getFungiIdsByProperty	Filamentous fungi strains	lib(s), text	id(s)
getFungiById	Filamentous fungi strains	id	full record
getYeastIdsByName	Yeasts strains	lib(s), name	id(s)
getYeastIdsByProperty	Yeasts strains	lib(s), text	id(s)
getYeastsById	Yeasts strains	id	full record
getPlasmidIdsByName	Plasmids	lib(s), name	id(s)
getPlasmidIdsByProperty	Plasmids	lib(s), text	id(s)
getPlasmidsById	Plasmids	id	full record
getPhageIdsByName	Phages	lib(s), name	id(s)
getPhageIdsByProperty	Phages	lib(s), text	id(s)
getPhagesById	Phages	id	full record
getCellLinesIdsByName	Human and animal cell lines	lib(s), name	id(s)
getCellLinesIdsByProperty	Human and animal cell lines	lib(s), text	id(s)
getCellLinesById	Human and animal cell lines	id	full record
getResourceIdsByName	All	lib(s), name	id(s)
getResourcesById	All	id	full record

Web Services TP53

Web Service Name	Input	Output
getP53MutationsByProperty	lib, text	Full record
getP53MutationsByIds	Id	Full record
getP53MutationIdsByType	lib, mutation type	id(s)
getP53MutationIdsByEffect	lib, effect	id(s)
getP53MutationIdsByExon	lib, exon number	id(s)
getP53MutationIdsByIntron	lib, intron number	id(s)
getP53MutationIdsByCodonNumber	lib, codon number	id(s)
getP53MutationIdsByCpgSite	lib, cpg site (true/false)	id(s)
getP53MutationIdsBySpliceSite	lib splice site (true/false)	id(s)
getP53MutationIdsByMetastasisLocalization	lib, metastasis localization (organ)	id(s)
getP53MutationIdsByTumorOrigin	lib, origin (primary, secondary, ...)	id(s)

Workflow

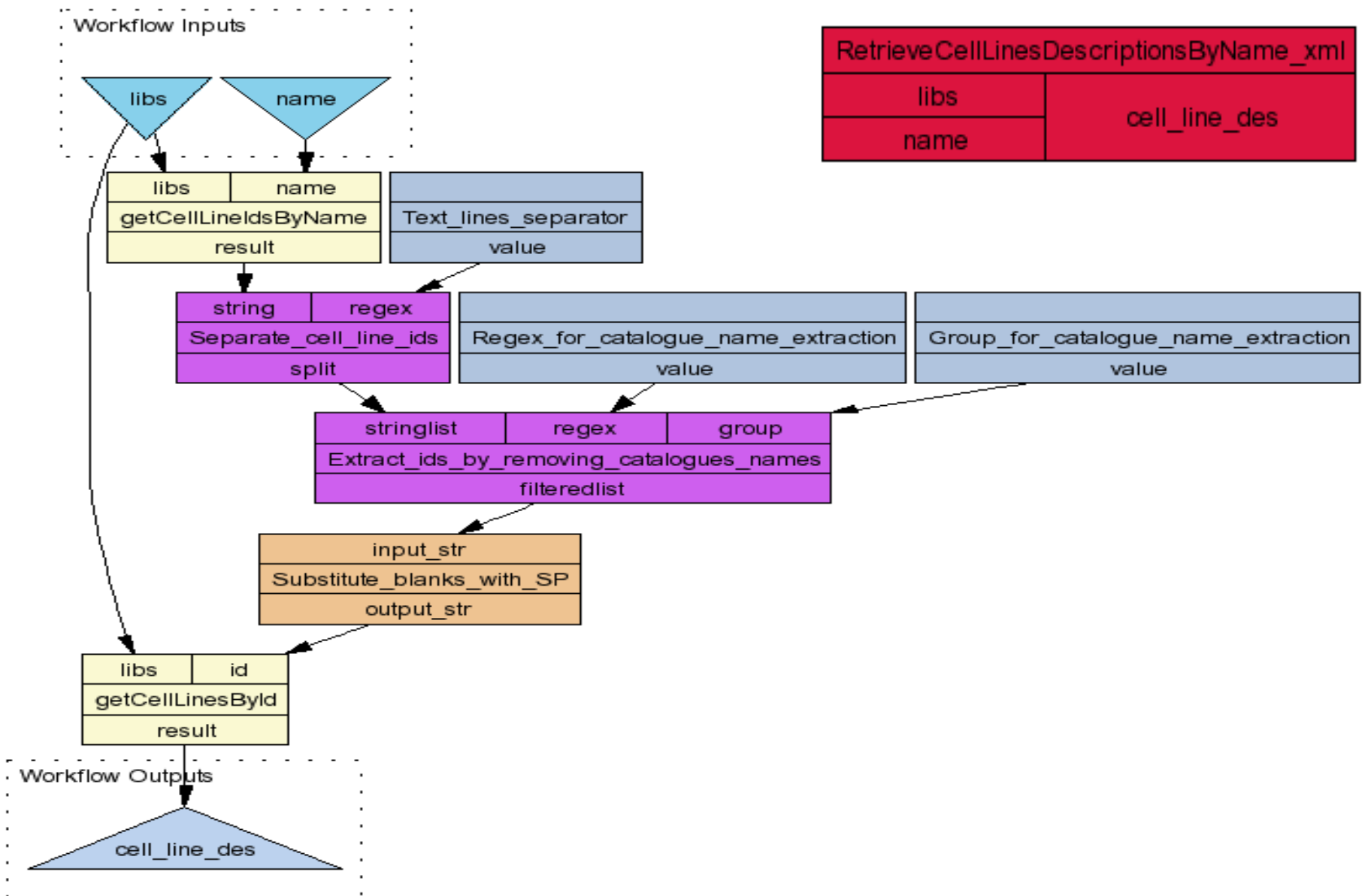
“A computerized facilitation or automation of a business process, in whole or part”. (Workflow Management Coalition)

Obiettivo:

- implementazione di processi di analisi dei dati in ambienti standardizzati

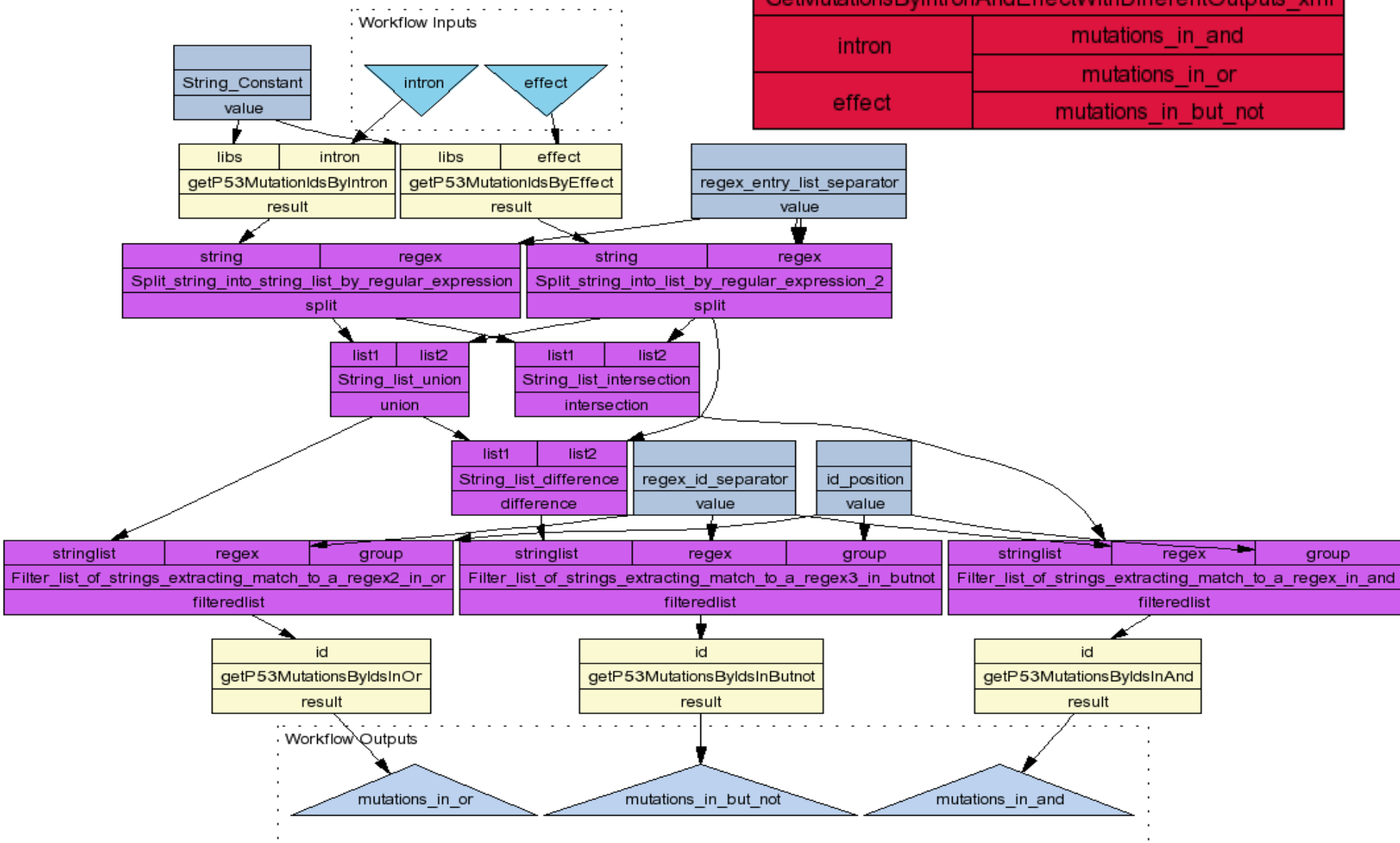
Vantaggi principali:

- **efficienza:** in quanto procedura automatica, libera il ricercatore dai compiti ripetitivi sul web e contribuisce a una “good practice”,
- **reproducibilità:** le analisi possono essere ripetute nel tempo,
- **riuso:** I risultati intermedi possono essere riutilizzati,
- **tracciabilità:** il workflow è eseguito in un ambiente trasparente nel quale la provenienza dei dati può essere verificata a posteriori.



Workflow per database CABRI

GetMutationsByIntronAndEffectWithDifferentOutputs_xml	
intron	mutations_in_and
	mutations_in_or
effect	mutations_in_but_not



Workflow per database TP53

Workflow management systems

Gestione di workflow per applicazioni bioinformatiche:

- Biopipe, un add-on per bioperl
- GPipe, una estensione dell'interfaccia Pise
- Taverna (EBI), una componente della piattaforma myGrid
- Pegasys (University of British Columbia)
- EGene (Universidade de São Paulo)
- Wildfire (Bioinformatics Institute, Singapore)
- Pipeline Pilot (SciTegic)
- BioWBI, Bioinformatic Workflow Builder Interface, di IBM

Richiedono una notevole conoscenza dei sistemi coinvolti e competenze e tempo per lo sviluppo dei workflow.

Workflow management systems

Software	Tipologia	Linguaggio XML	Disponibilità	URL
Taverna Workbench	Stand-alone	XScufl	Open source	http://taverna.sourceforge.net/
Biopipe	Libreria software	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	Interfaccia Web servizi locali	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	Interfaccia Web servizi remoti	XPDL	Public use	http://litbio.unicam.it:8080/biowms/
BioWEP	Portale	XScufl XPDL	Open source	http://bioinformatics.istge.it/biowep/
BioWBI	Interfaccia Web servizi locali	Proprietary	Commerciale	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 Workflow Language	Open source	http://www.trianacode.org/
Pipeline Pilot	Stand-alone	Proprietary	Commercial	http://www.scitegic.com/
FreeFluo	Libreria software	WSEL e XScufl	Open source	http://freefluo.sourceforge.net/
Biomake	Libreria software	NA	Open source	http://skam.sourceforge.net/

Presentano diverse tipologie e utilizzano diversi standard

Workflow management: Taverna

Taverna Workbench consente di

- costruire workflow per analisi complesse
- accedere a processori sia remoti che locali
- definire processori alternativi
- eseguire i workflow
- visualizzare i risultati in diversi formati
- descrivere i dati bioinformatici tramite un'ontologia

Requirements: java, Windows or Linux

Open source: <http://taverna.sourceforge.net/>

Current version: 1.4 (stable, next version 2.0)

Taverna: processori disponibili

Web Services descritti tramite WSDL

- Mette a disposizione i servizi descritti

Web Services accessibili tramite server Soaplab

- Mette a disposizione i servizi forniti da server Soaplab

Registri BioMOBY

- Interagisce con un “MOBY Central repository” per accedere ai servizi registrati

Workflow

- Incorpora interi workflow definiti con Scufi o aggiunge singoli processori

Biomart

- Interagisce con database Biomart per comporre query ed estrarre dati

SeqHound

- Interroga SeqHound (Sequence and Structure Database Management System), un insieme di Web Services basati sul modello dati NCBI, e accede a informazioni di sequenza e struttura

Processori locali

- Funzioni Java in grado di elaborare liste e stringhe, definire valori costanti, eseguire semplici elaborazioni, leggere e scrivere file. È anche possibile scrivere propri script con beanshell

Taverna: GUI

La Graphical User Interface (GUI) comprende:

- Advanced Model Explorer (AME)
- Workflow diagram
- Available services
- Run workflow
- Enactor invocation

Opzioni: default services, workflow editor, debug
Feature extra: FETA search engine

Taverna Scufl Workbench v1.3

myGrid

<http://taverna.sf.net>

djm,jf,mg,pl,ktg,mp,ms,tmo,mf,ek,pa, jb et al.

actme

Advanced model explorer

Workflow Object properties

Load Load from web Save New subworkflow Offline Reset

Workflow object Retries Delay Back... Thre... Critical

- Workflow model
 - Workflow inputs
 - Workflow outputs
 - Processors
 - Data links
 - Control links

Workflow diagram

Save as Configure diagram

Rendering done.

Available services

Search list Watch loads

Available Processors

Local Services

Notification Processor

Local Java widgets

list

io

metadata

xml

ui

ncbi

conditional

net

text

Byte[] to String

String list union

Concatenate two strings

String list difference

Filter list of strings by regex

Split string into string list by regular expression

Pad numeral with leading 0s

Filter list of strings extracting match to a regex

String list intersection

biojava

jdbc

base64

moby

test

Beanshell scripting host

String Constant

Biomart registry @ <http://www.ebi.ac.uk/~tmo/defaultMartRegist>WSDL @ <http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoVi>

porttype: GoVi [RPC]

createSession

destroySession

getChildren

addTerm

- Workflow Editor (BETA)
- Available services (no defaults)
- Advanced model explorer
- <> DEBUG - Workflow XML preview
- DEBUG - Enactor event monitor
- Dot text
- Available services
- Workflow diagram
- Run workflow

orkbench v1.3

<http://taverna.sf.net>

djm,jf,mg,pl,ktg,mp,ms,tmo,mf,ek,pa, jb et al.

lorer

cles

Save New subworkflow Offline Reset

Workflow object

- Workflow model
 - Workflow inputs
 - Workflow outputs
 - Processors
 - Data links
 - Control links

Retries

Delay

Back...

Thre...

Critical

Workflow diagram

Save as Configure diagram

Rendering done.

Available services

Search list

☒ Watch loads

Available Processors

Local Services

Notification Processor

Local Java widgets

list

io

metadata

xml

ui

ncbi

conditional

net

text

Byte[] to String

String list union

Concatenate two strings

String list difference

Filter list of strings by regex

Split string into string list by regular expression

Pad numeral with leading 0s

Filter list of strings extracting match to a regex

String list intersection

biojava

jdbc

base64

moby

test

Beanshell scripting host

String Constant

Biomart registry @ <http://www.ebi.ac.uk/~tmo/defaultMartRegist>WSDL @ <http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoVi>

porttype: GoViz [RPC]

createSession

destroySession

getChildren

addTerm

Advanced model explorer

Workflow | Object properties

Load | Load from web | Save | New subworkflow | Offline | Reset

Workflow object: Retries Delay Back... Thre... Critical

- Workflow model
 - Workflow inputs
 - libs
 - name
 - Workflow outputs
 - cell_line_des
 - Processors
 - Text_lines_separator : \n
 - Group_for_catalogue_name_extraction : 2
 - Regex_for_catalogue_name_extraction : ([
 - Substitute_blanks_with_SP
 - Extract_ids_by_removing_catalogues_name
 - Separate_cell_line_ids

Workflow diagram

Save as | Configure diagram

Text_lines_separator → Separate_cell_line_ids → Group_for_catalogue_name_extraction → Extract_ids_by_removing_catalogues_name → Separate_cell_line_ids

Workflow outputs: cell_line_des

Rendering done.

Available services

Search list | Watch loads

Available Processors

- Local Services
 - Biomart registry @ <http://www.ebi.ac.uk/~tmo/defaultMartRegistry...>
 - WSDL @ <http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoViz.jw...>
 - WSDL @ <http://www.ebi.ac.uk/xembl/XEMBL.wsdl>
 - WSDL @ <http://soap.genome.jp/KEGG.wsdl>
 - WSDL @ <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.wsdl>
 - WSDL @ <http://soap.bind.ca/wsdl/bind.wsdl>

Open Workflow

Cerca in: workflows

- Documenti recenti
- Desktop
- Documenti
- Risorse del computer
- Risorse di rete

- GetTp53MutationsByExonAndEffect2.xml
- GetTP53MutationsByExonAndEffect2a.xml
- GetTp53MutationsByIntron.xml
- GetTp53MutationsByIntronAndEffect.xml
- GetTp53MutationsByIntronAndEffect2.xml
- IterationStrategyExample.xml
- origwrf.xml
- RetrieveBacteriaDescriptionsByName.xml
- RetrieveCellLineAndRelatedTp53MutationWithRelatedPubmedInXmlFormat.xml
- RetrieveCellLinesDescriptionsByName.xml**
- RetrievePMIDsFromCellLinesDescriptions.xml
- RetrieveTP53MutationAndRelatedCellLine.xml
- ShowGeneOntologyContext.xml
- test.xml

Nome file: RetrieveCellLinesDescriptionsByName.xml

Tipo file: Filter for extensions : xml

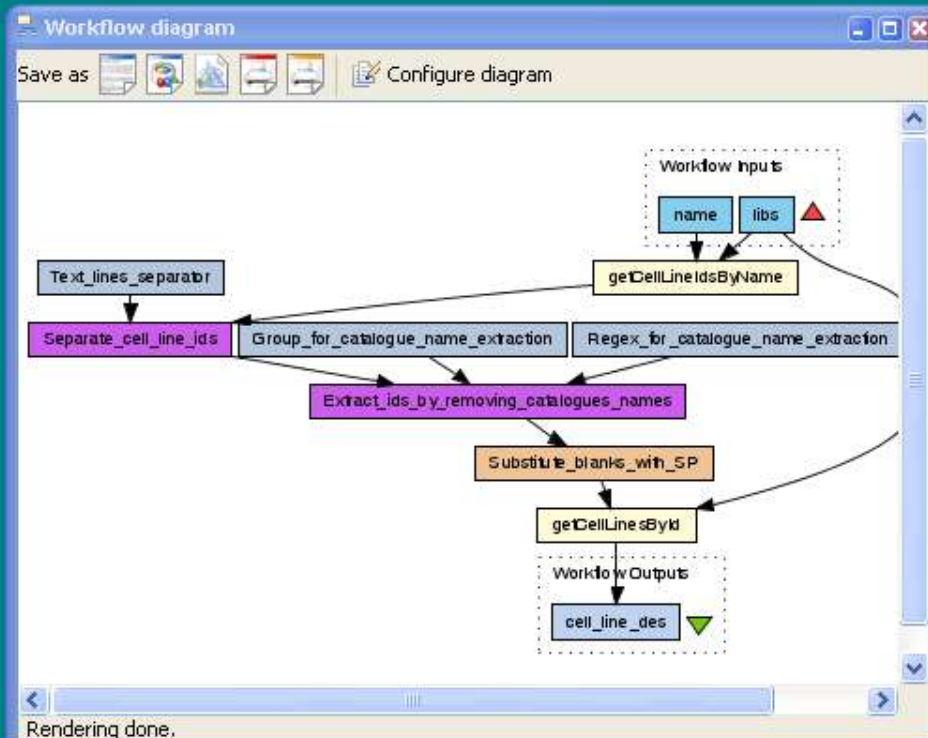
Apri | Annulla

Advanced model explorer

Workflow | Object properties

Load | Load from web | Save | New subworkflow | Offline | Reset

Workflow object	Retries	Delay	Back...	Thre...	Critical
Workflow model					
Workflow inputs					
libs					
name					
Workflow outputs					
cell_line_des					
Processors					
Text_lines_separator : \n	0	0	1	1	☐
Group_for_catalogue_name_extraction : 2	0	0	1	1	☐
Regex_for_catalogue_name_extraction : ([A	0	0	1	1	☐
Substitute_blanks_with_SP	0	0	1	1	☐
Extract_ids_by_removing_catalogues_names	0	0	1	1	☐
Separate_cell_line_ids	0	0	1	1	☐



Available services

Search list | Watch loads

Available Processors

- Local Services
- Biomart registry @ <http://www.ebi.ac.uk/~tmo/defaultMartRegistry>
- WSDL @ <http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoViz.jw>
- WSDL @ <http://www.ebi.ac.uk/xembl/XEMBL.wsdl>
- WSDL @ <http://soap.genome.jp/KEGG.wsdl>
- WSDL @ <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.wsdl>
- WSDL @ <http://soap.bind.ca/wsdl/bind.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSFasta.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSWUBlast.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSInterProScan.wsdl>
- Soaplab @ <http://www.ebi.ac.uk/soaplab/services/>
- Biomoby @ <http://mobycentral.icapture.ubc.ca/cgi-bin/MOBY05/moby>
- SeqHound @ seqhound.blueprint.org
- Soaplab @ <http://www.o2i.it:8080/axis/services/>

The screenshot displays the Scufi Workflowbench v1.3.1 interface. The main window is titled 'Advanced model explorer' and shows a tree view of the workflow object. The 'Workflow inputs' section is expanded, showing the 'libs' input. The 'Run Workflow' dialog box is open, showing the 'Workflow Input : libs' configuration. The dialog includes a table for 'Input Metadata' and a 'Description' section. The 'Workflow diagram' panel at the bottom shows the workflow flow, including processors like 'Text_lines_separator', 'Separate_cell_line_ids', 'Group_for_catalogue_name_extraction', 'Regex_for_catalogue_name_extraction', 'Substitute_blanks_with_SP', 'Extract_ids_by_removing_catalogues_name', and 'Separate_cell_line_ids'.

Workflow Input : libs

Input Metadata	
Semantic type	http://www.mygrid.org.uk/ontology#CABRI_cell_lines_catalogue
Syntactic type	text/plain/text/plain

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 Feb 15, 2005, possible values are: - 'iclc' (i.e., the Interlab Cell Line Collection, www.iclc.it) - 'ecacc_cell' (i.e., the European Collection of Cell Cultures, www.ecacc.org.uk) - 'dsmz_mutz' (i.e., the collection of human and animal cell cultures of the DSMZ, www.dsmz.de) Catalogues can be added (or, rarely, removed) without notice. See www.cabri.org for further information.

Instructions

To input data into this workflow you must first create either a single item or a list. Having done this you can select the item from the tree to the left of this panel and either enter the data manually, upload from a file on your local machine or load from a location on the internet. When all workflow inputs have been populated as required you can click the 'run workflow' button to

Workflow diagram

```

graph TD
    Text_lines_separator[Text_lines_separator] --> Separate_cell_line_ids[Separate_cell_line_ids]
    Group_for_catalogue_name_extraction[Group_for_catalogue_name_extraction] --> Separate_cell_line_ids
    Separate_cell_line_ids --> Extract_ids_by_removing_catalogues_name[Extract_ids_by_removing_catalogues_name]
    Extract_ids_by_removing_catalogues_name --> Separate_cell_line_ids
    Separate_cell_line_ids --> cell_line_des[cell_line_des]
  
```

Rendering done.

The screenshot displays the Scufl Workflow v1.3.1 application interface. The main window is titled 'Advanced model explorer' and contains a 'Workflow' tab with 'Object properties' and 'Workflow diagram' views. The 'Workflow diagram' view shows a flowchart with processors like 'Text_lines_separator', 'Separate_cell_line_ids', 'Group_for_catalogue_name_extraction', 'Regex_for_catalogue_name_extraction', 'Substitute_blanks_with_SP', 'Extract_ids_by_removing_catalogues_name', and 'Separate_cell_line_ids'. The 'Run Workflow' dialog box is open, showing the 'Input Document' section with a tree view of 'libs'. A context menu is visible over the 'libs' input, with options 'New Input Value' and 'New List'. The 'New Input Value' option is selected, and the 'Input metadata' section is displayed, showing the 'Semantic type' as 'http://www.mygrid.org.uk/ontology#CABRI_cell_lines_catalogue' and the 'Syntactic type' as 'text/plain/text/plain'. The 'Description' section provides details about the input, including the name(s) of the CABRI human and animal cell lines catalogues involved in the query. The 'Instructions' section explains how to input data into this workflow.

Run Workflow

Load Inputs New Input New List Remove

Input Document

libs

New Input Value

New List

Input metadata

Semantic type	http://www.mygrid.org.uk/ontology#CABRI_cell_lines_catalogue
Syntactic type	text/plain/text/plain

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 Feb 15, 2005, possible values are: - 'iclc' (i.e., the Interlab Cell Line Collection, www.iclc.it) - 'ecacc_cell' (i.e., the European Collection of Cell Cultures, www.ecacc.org.uk) - 'dsmz_mutz' (i.e., the collection of human and animal cell cultures of the DSMZ, www.dsmz.de) Catalogues can be added (or, rarely, removed) without notice. See www.cabri.org for further information.

Instructions

To input data into this workflow you must first create either a single item or a list. Having done this you can select the item from the tree to the left of this panel and either enter the data manually, upload from a file on your local machine or load from a location on the internet. When all workflow inputs have been populated as required you can click the 'run workflow' button to

Run Workflow

Workflow diagram

Save as

Configure

Text_lines_separator

Separate_cell_line_ids

Group_for_catalogue_name_extraction

Extract_ids_by_removing_catalogues_name

cell_line_des

Rendering done.

The screenshot displays the Scufl Workflow v1.3.1 interface. The main window is titled "Advanced model explorer" and contains a "Workflow" tab and an "Object properties" tab. The "Workflow" tab shows a tree view of the workflow object, including "Workflow model", "Workflow inputs", "Workflow outputs", and "Processors". The "Workflow inputs" section is expanded, showing a list of inputs: "libs", "name", and "vero". The "Workflow outputs" section is also expanded, showing a list of outputs: "cell_line_des". The "Processors" section is expanded, showing a list of processors: "Text_lines_separator", "Group_for_catalogue_name_extraction", "Regex_for_catalogue_name_extraction", "Substitute_blanks_with_SP", "Extract_ids_by_removing_catalogues_name", and "Separate_cell_line_ids".

The "Run Workflow" dialog box is open, showing a "Load Inputs" section with a "New Input" button and a "New List" button. The "Input Document" section shows a list of inputs: "libs", "name", and "vero". The "Load" section shows a list of outputs: "cell_line_des". The "Run Workflow" button is visible at the bottom right of the dialog box.

The "Workflow diagram" section shows a flowchart of the workflow. The flow starts with "Text_lines_separator", which connects to "Separate_cell_line_ids". "Separate_cell_line_ids" then connects to "Group_for_catalogue_name_extraction". "Group_for_catalogue_name_extraction" connects to "Extract_ids_by_removing_catalogues_name". The final output is "cell_line_des".

Available services:

- o/defaultMartRegistry...
- service1/goviz/GoViz.jw...
- wsdl
- teutils/soap/eutils.wsdl
- n:Dbfetch?wsdl
- wsdl
- t.wsdl
- oScan.wsdl
- vices/
- a/cgi-bin/MOBY05/moby...
- es/

Rendering done.

Rendering done.

Advanced model explorer

Workflow Object properties

Load Load from web

Workflow object

- Workflow model
 - Workflow inputs
 - libs
 - name
 - Workflow outputs
 - cell_line_des
 - Processors
 - Text_lines_separat
 - Group_for_catalo
 - Regex_for_catalo
 - Substitute_blank
 - Extract_ids_by_r
 - Separate_cell_lin

Enactor invocation

< > Save as XML Save to disk Save to disk as website Excel

Status Results Process report

Processor stati

...	Name	Last event	Event timestamp	Event detail	Breakpoint
	Text_lines_separator	ProcessComplete	8-mar-2006 13.33.31		
	Group_for_catalogue_nam...	ProcessComplete	8-mar-2006 13.33.31		
	Regex_for_catalogue_nam...	ProcessComplete	8-mar-2006 13.33.31		
	Substitute_blanks_with_SP	ProcessComplete	8-mar-2006 13.33.31		
	Extract_ids_by_removing_...	ProcessComplete	8-mar-2006 13.33.31		
	Separate cell line ids	ProcessComplete	8-mar-2006 13.33.31		

Graph Intermediate inputs Intermediate outputs

```
graph TD; Inputs["Inputs<br/>name"] --> getCellLineIdsByName["getCellLineIdsByName"]; getCellLineIdsByName --> Separate_cell_line_ids["Separate_cell_line_ids"]; Separate_cell_line_ids --> Extract_ids_by_removing_catalogues_names["Extract_ids_by_removing_catalogues_names"]; Extract_ids_by_removing_catalogues_names --> Substitute_blanks_with_SP["Substitute_blanks_with_SP"]; Substitute_blanks_with_SP --> getCellLinesById["getCellLinesById"]; getCellLinesById --> Outputs["Outputs<br/>cell_line_des"]; Inputs --> getCellLinesById;
```

Rendering done.

The screenshot displays the Scufl Workflow v1.3.1 interface. The main window is titled 'Advanced model explorer' and shows a workflow object tree on the left. The 'Run Workflow' window is open, displaying the 'Enactor invocation' results for the workflow object 'cell_line_des'.

Workflow object: cell_line_des

Enactor invocation:

- Save as XML
- Save to disk
- Save to disk as website
- Excel

Status: Results

cell_line_des

List

- urn:lsid:www.mygrid.org.uk:document
- Click to view... <urn:lsid:www.mygrid.org.uk:lsid:www.mygrid.org.uk:document>

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; split co

Viruses Search for viruses was not performed

Properties indicator line for mycoplasma testing; virology; virus titration; viru

Release_conditions Cell line available for distribution. For non-commercial inves

Hazard -

Species_validation Validated by isoenzymes: confirmed as monkey with AST, MD, NP,

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]
 Natu

Run Workflow

Workflow outputs

- cell_line_des

Rendering done.

Advanced model explorer

Workflow Object properties

Load Load from web

Workflow object

- Workflow model
 - Workflow inputs
 - libs
 - name
 - Workflow outputs
 - cell_line_des
 - Processors
 - Text_lines_separ
 - Group_for_catalo
 - Regex_for_catalo
 - Substitute_blank
 - Extract_ids_by_r
 - Separate_cell_lin

Enactor invocation

< > Save as XML Save to disk Save to disk as website Excel

Status Results Process report

Processor stati

...	Name	Last event	Event timestamp	Event detail	Breakpoint
	Regex_for_catalogue_nam...	ProcessComplete	8-mar-2006 13.33.31		
	Substitute_blanks_with_SP	ProcessComplete	8-mar-2006 13.33.31		
	Extract_ids_by_removing_...	ProcessComplete	8-mar-2006 13.33.31		
	Separate_cell_line_ids	ProcessComplete	8-mar-2006 13.33.31		
	getCellLineIdsByName	ProcessComplete	8-mar-2006 13.33.31		
	getCellLinesById	ProcessComplete	8-mar-2006 13.33.32		

Graph Intermediate inputs Intermediate outputs

result

ICLC:ICLC ATL95005
<urn:lsid:www.mygrid.org.uk:lsdocument:7Y4DDO713113>

ICLC:ICLC ATL95005

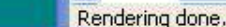
Workflow diagram

Save as

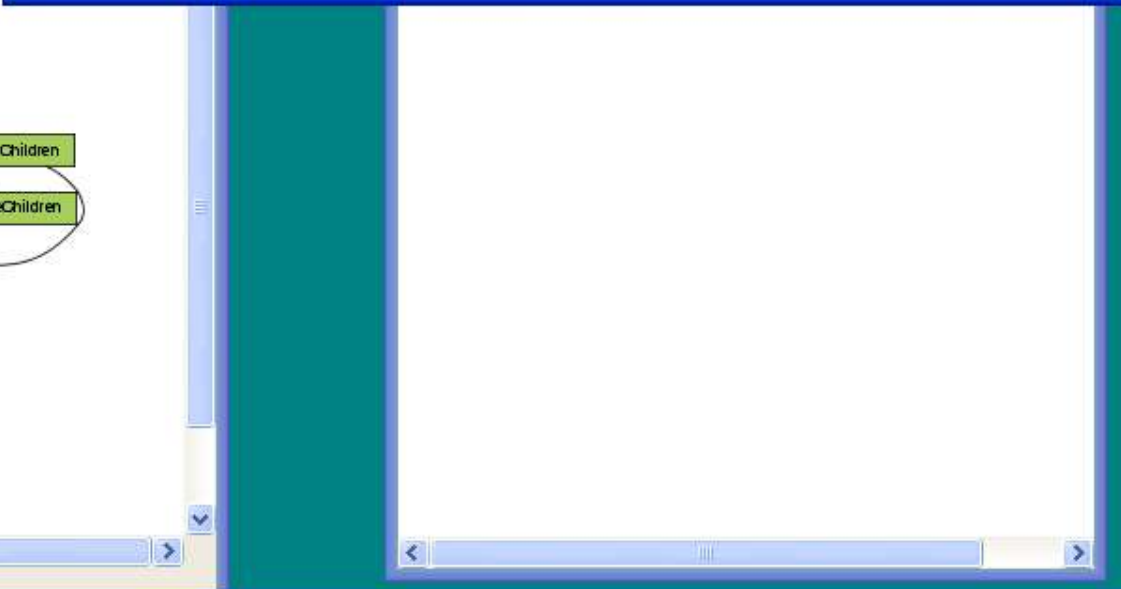
Text_lines_separatr

Separate_cell_line_ids

Rendering done.



The screenshot shows the 'Run Workflow' window. The toolbar includes 'Load Inputs', 'New Input', 'New List', and 'Remove'. The 'Input Document' tab is selected, displaying a tree structure with 'termID' and 'GO:0007601'. The 'Load' tab is also visible, showing 'Load from URL'. A 'Run Workflow' button is located at the bottom right.



Enactor invocation

Workflow Status : Running

Pause

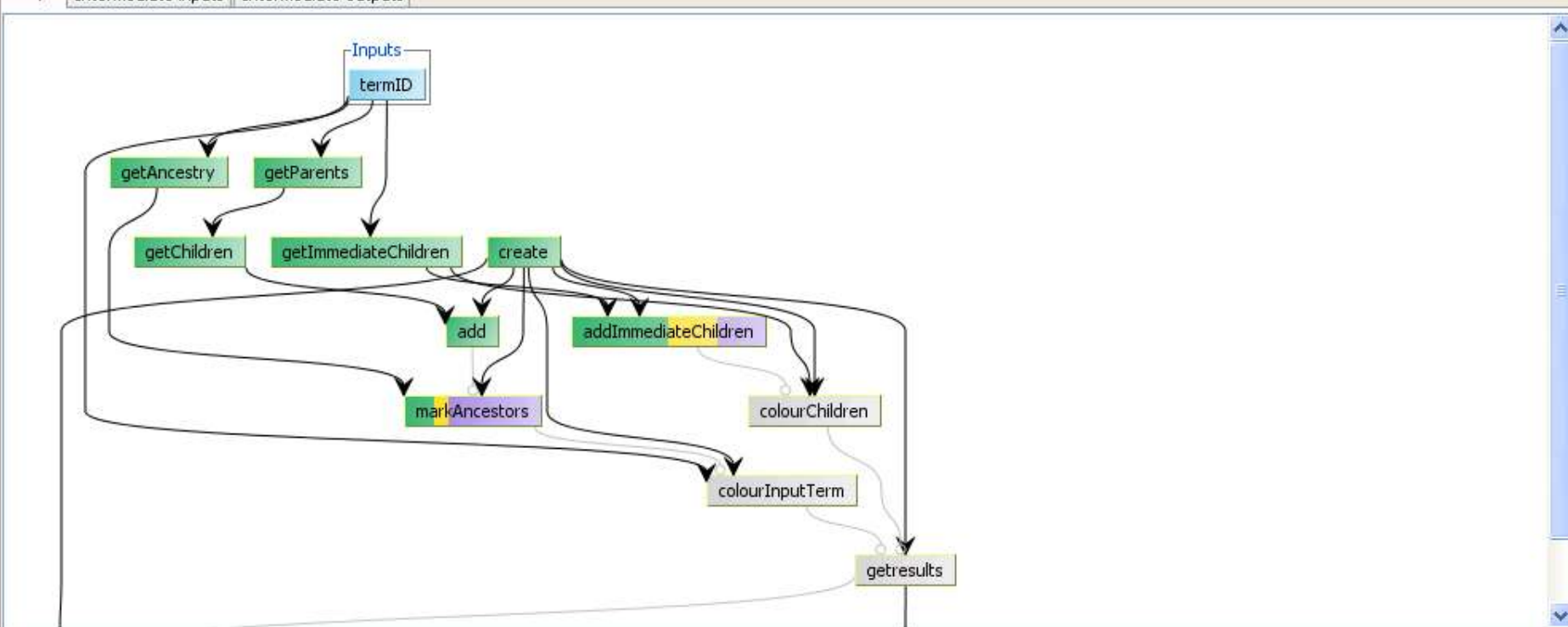
Stop

Status

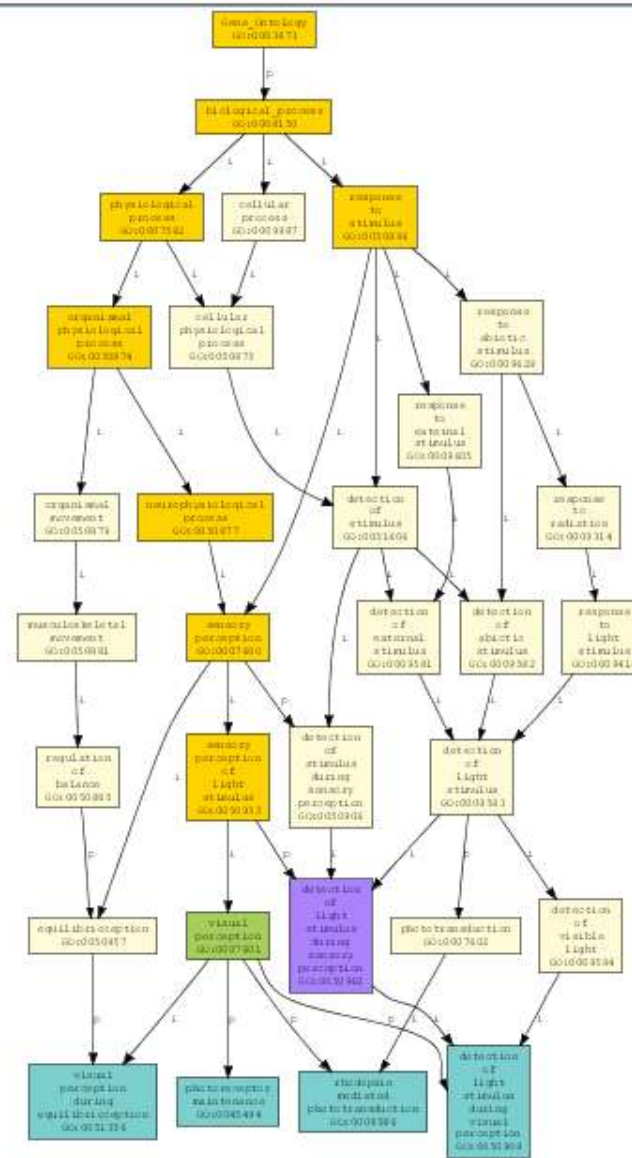
Processor stati

...	Name	Last event	Event timestamp	Event detail	Breakpoint
	ancestorColour	ProcessComplete	8-mar-2006 14.20.21		
	childColour	ProcessComplete	8-mar-2006 14.20.21		
	inputTermColour	ProcessComplete	8-mar-2006 14.20.21		
	getresults	ProcessScheduled	8-mar-2006 14.20.21		
	getImmediateChildren	ProcessComplete	8-mar-2006 14.20.24		
	getChildren	ProcessComplete	8-mar-2006 14.20.37		
	create	ProcessComplete	8-mar-2006 14.20.24		
	getAncestry	ProcessComplete	8-mar-2006 14.20.24		
	finish	ProcessScheduled	8-mar-2006 14.20.21		
	markAncestors	InvokingWithIteration	8-mar-2006 14.20.38	IterationNumber='3' IterationTotal='9' ActiveWorkers...	
	add	ProcessComplete	8-mar-2006 14.20.38		

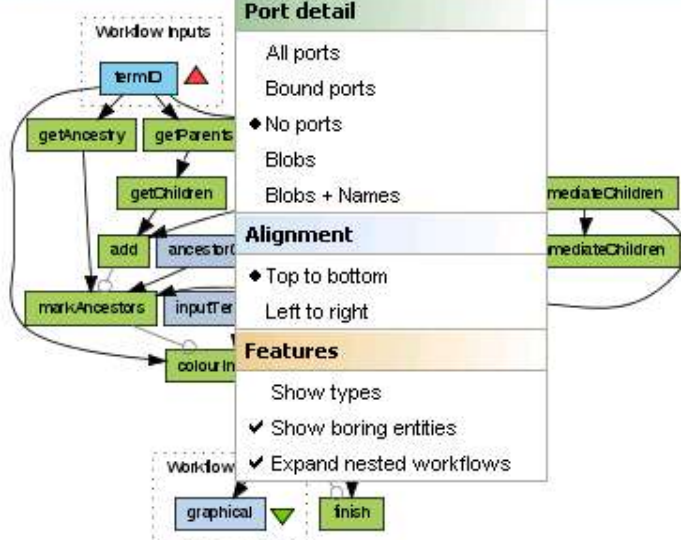
Graph Intermediate inputs Intermediate outputs



Click to view...
<http://www.mygr...>
 Save to file
 Viewers
 X-Graphviz
 Text
 Table



Workflow diagram

Save as       Configure diagram

Workflow diagram

Save as



Configure diagram

Port detail

All ports

◆ Bound ports

No ports

Blobs

Blobs + Names

Alignment

◆ Top to bottom

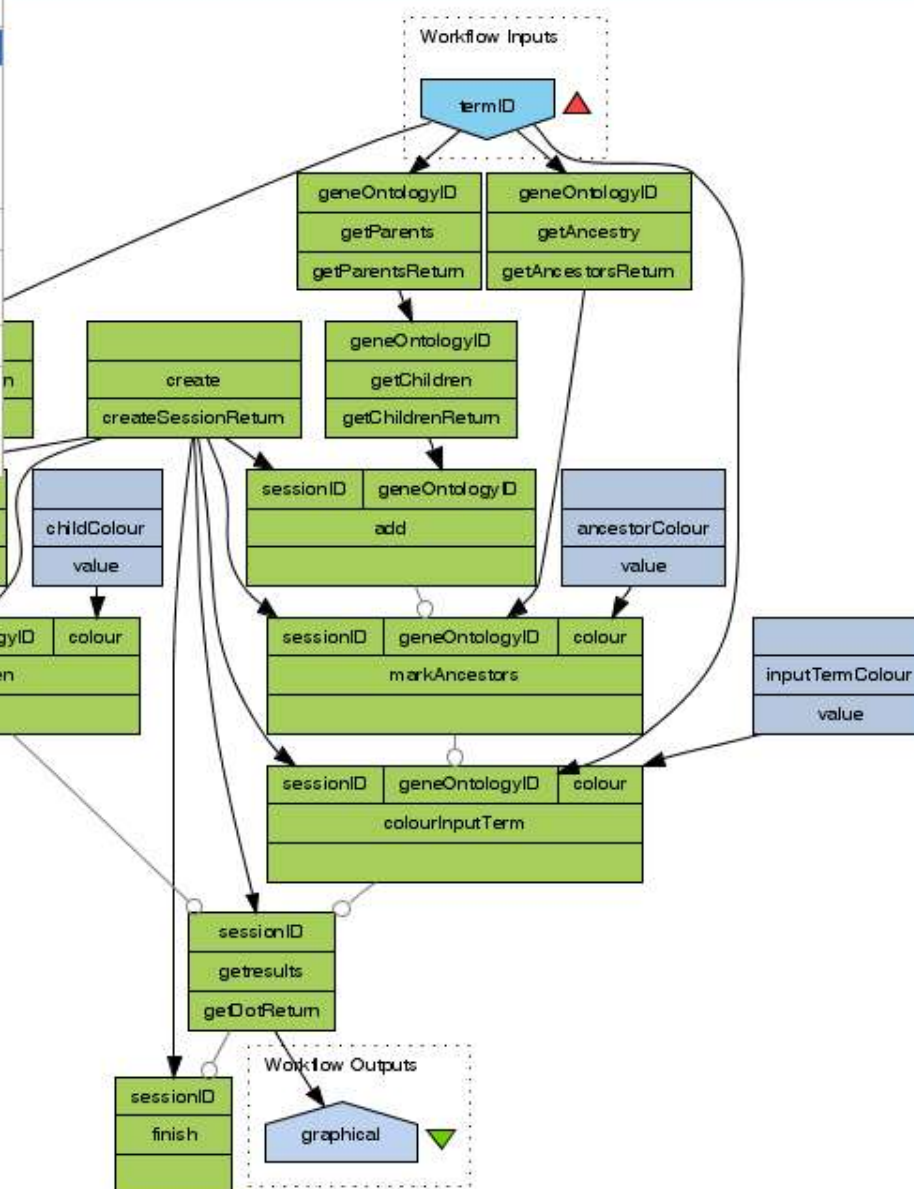
Left to right

Features

Show types

✓ Show boring entities

✓ Expand nested workflows



Workflow diagram

Save as      Configure diagram

Port detail

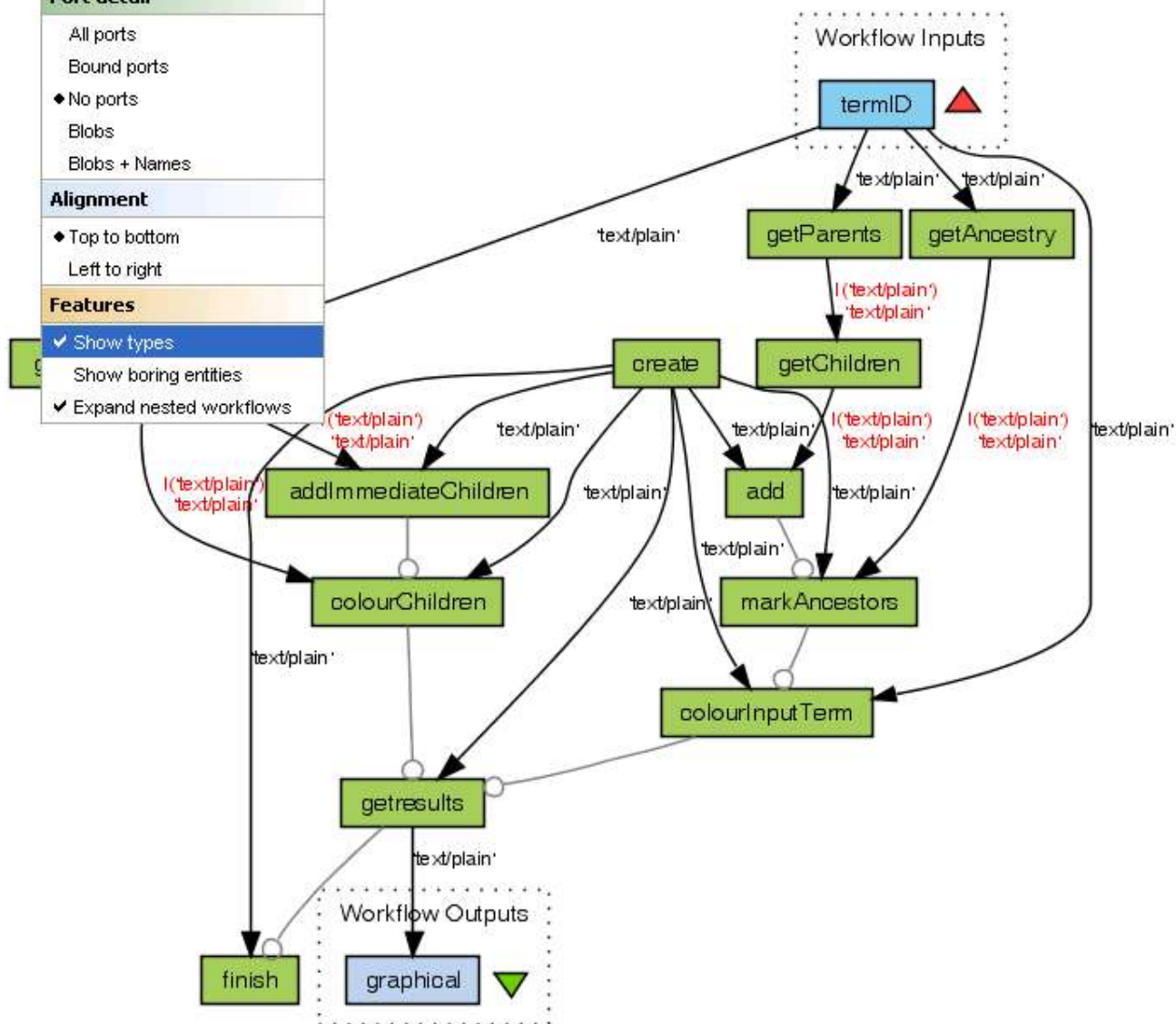
- All ports
- Bound ports
- ◆ No ports
- Blobs
- Blobs + Names

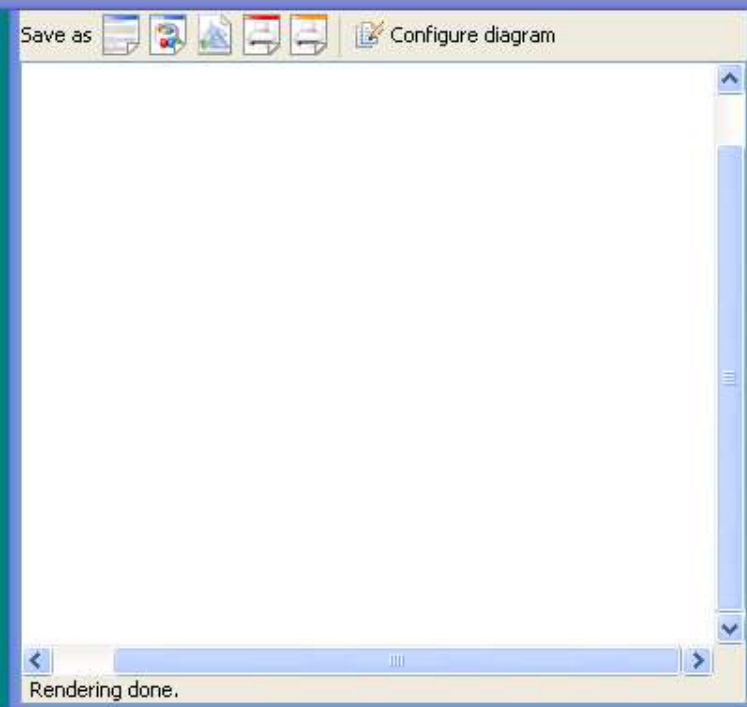
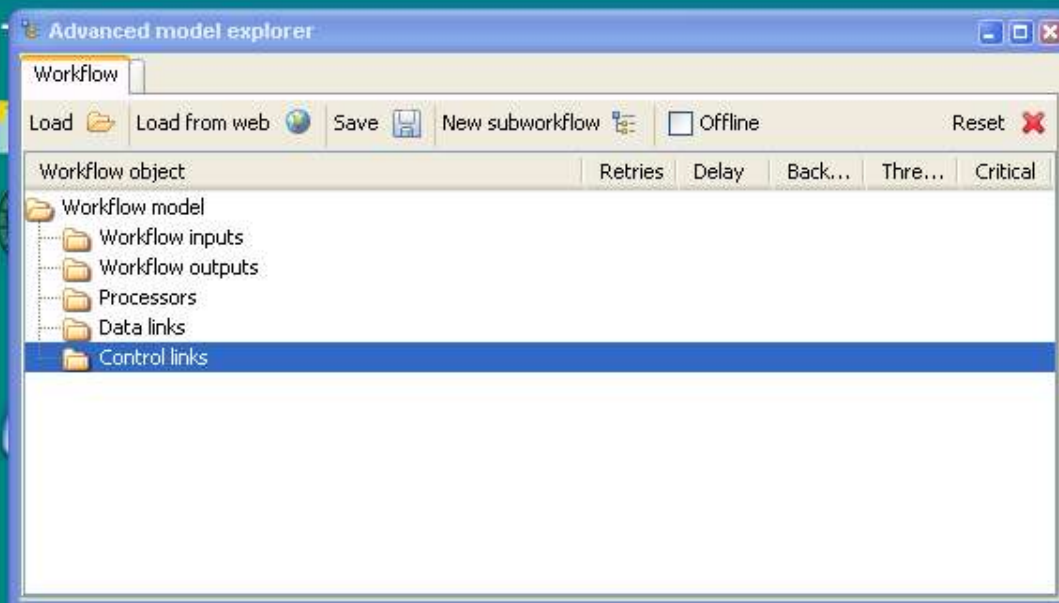
Alignment

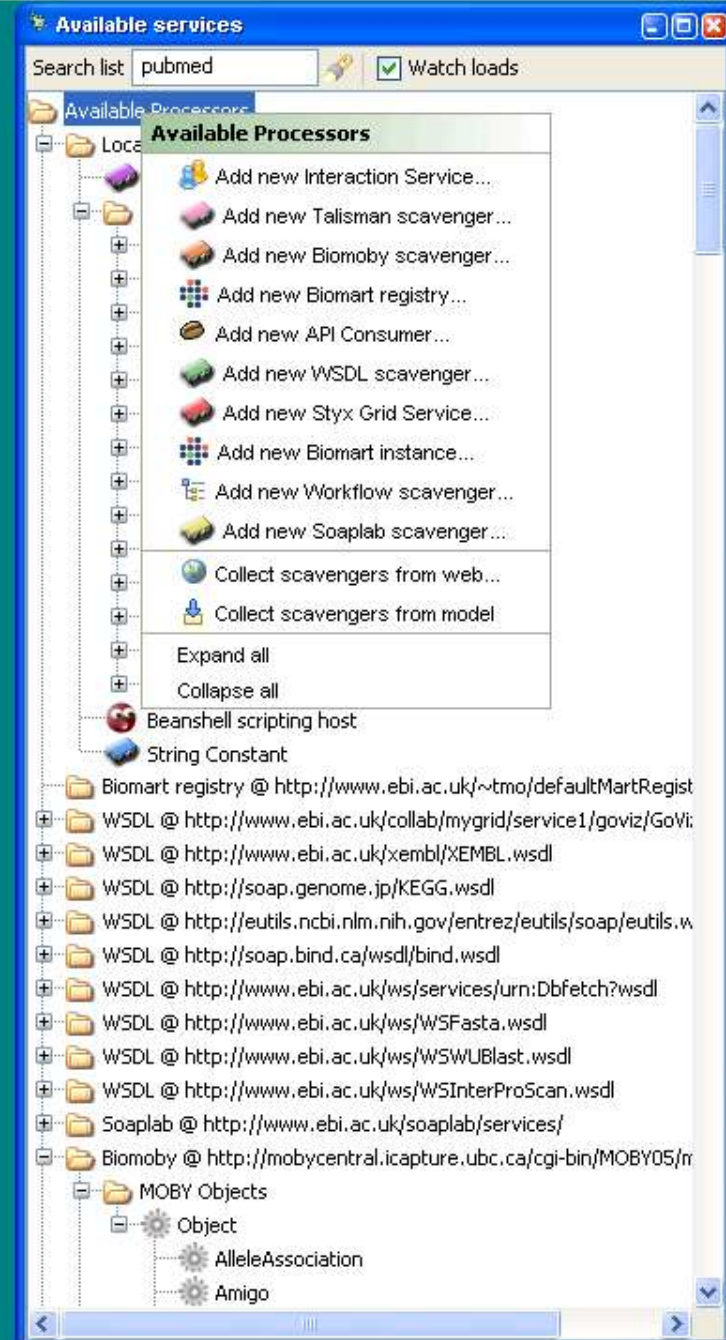
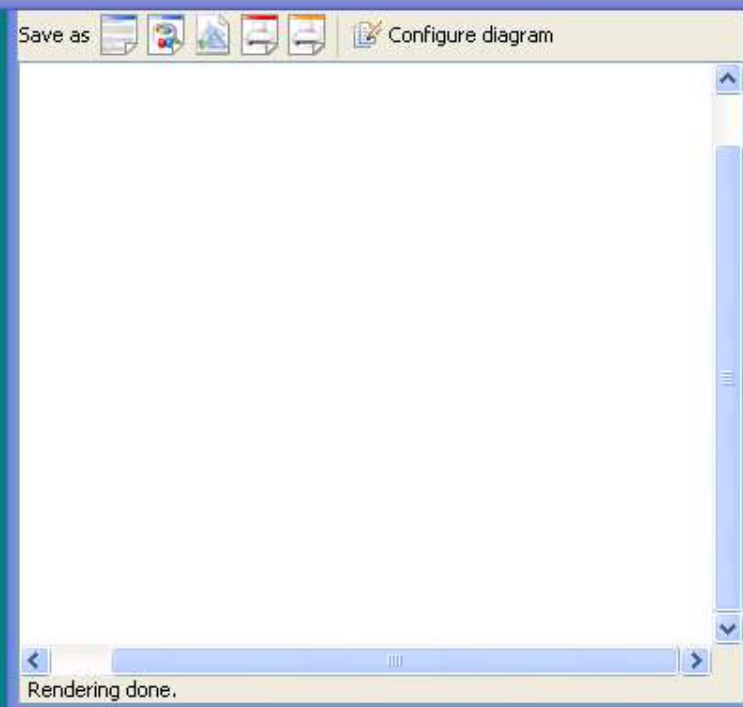
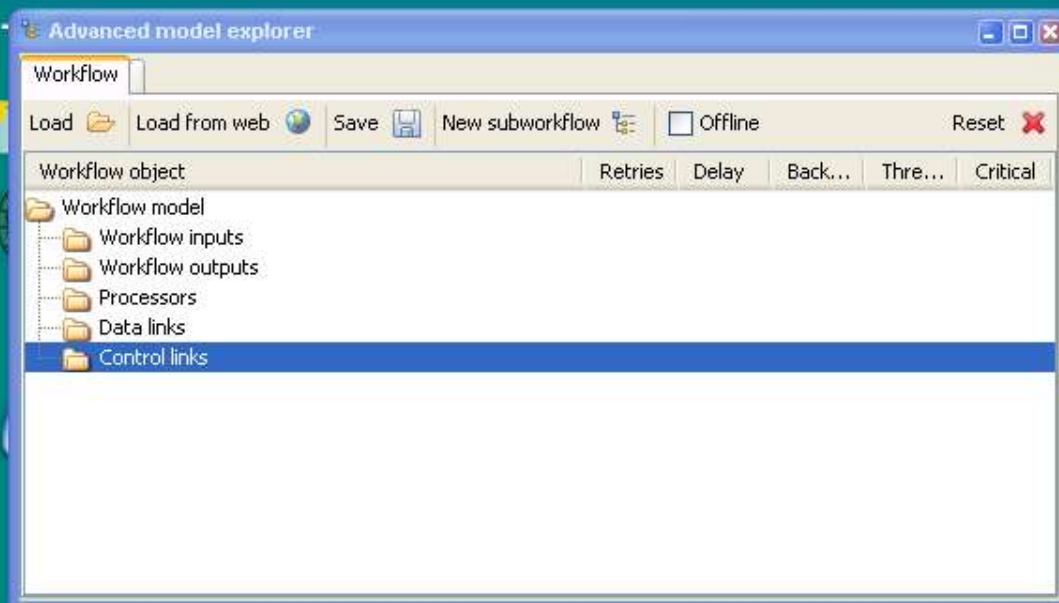
- ◆ Top to bottom
- Left to right

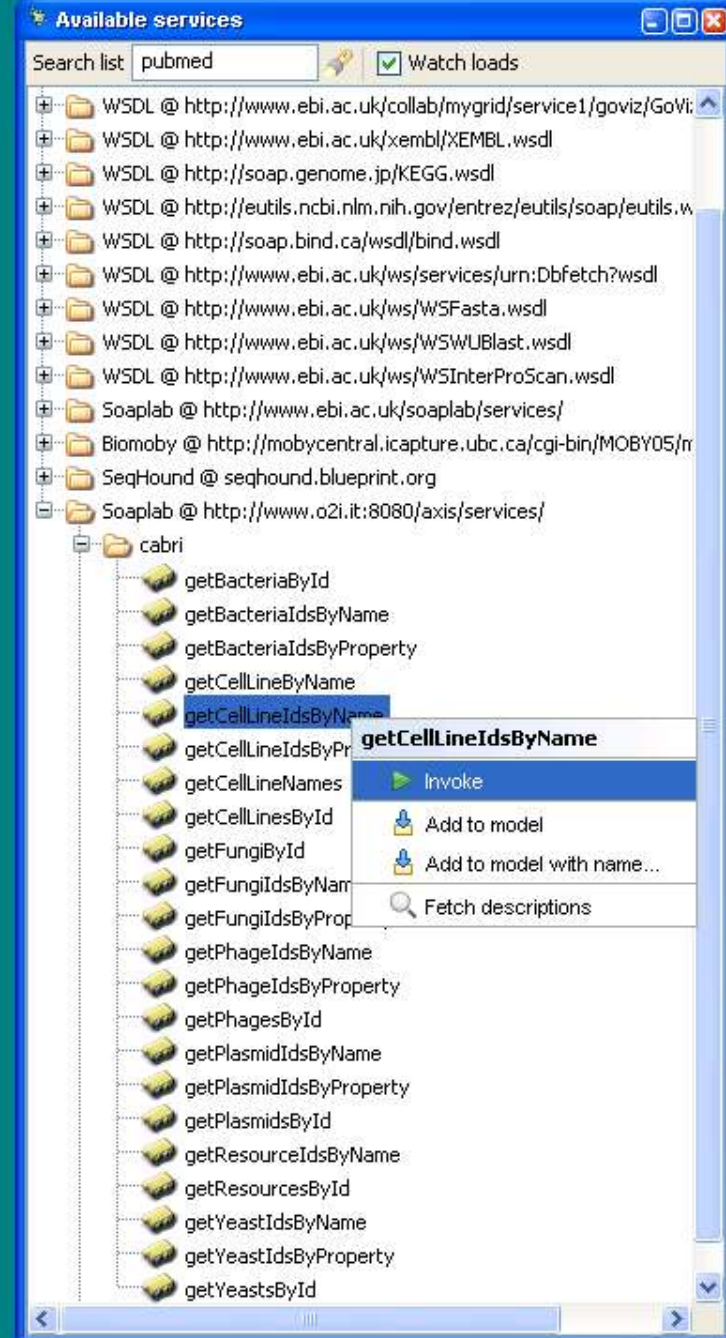
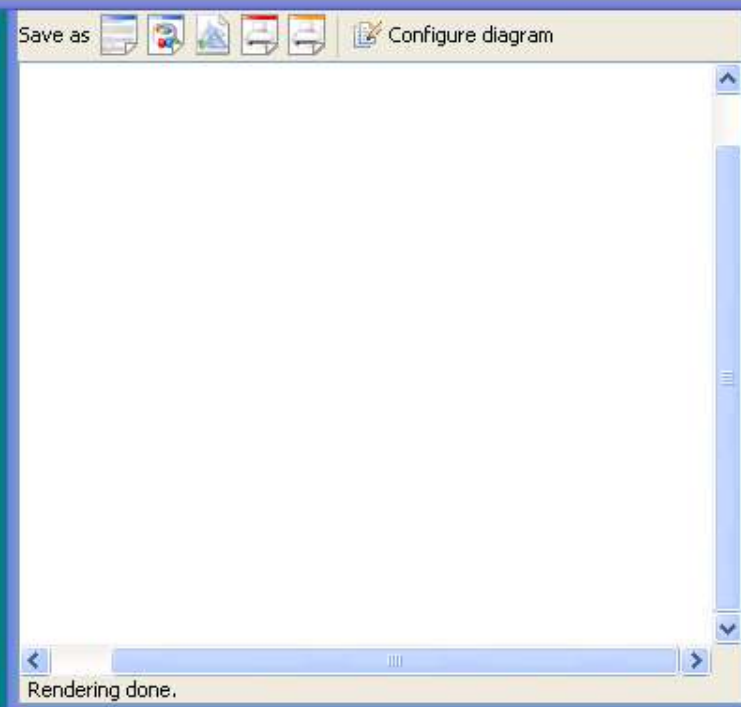
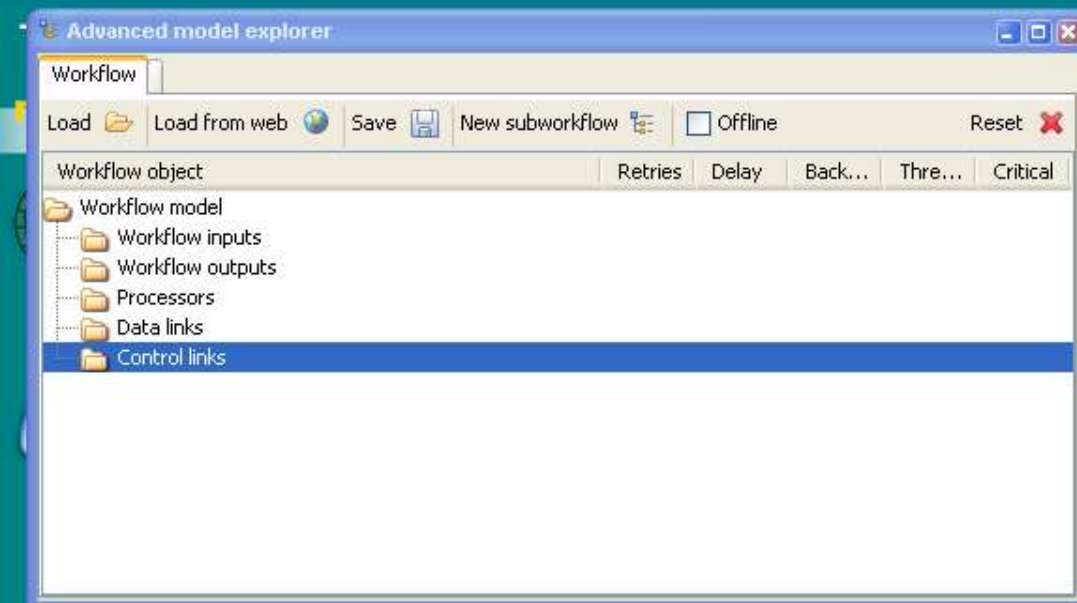
Features

- ✓ Show types
- Show boring entities
- ✓ Expand nested workflows







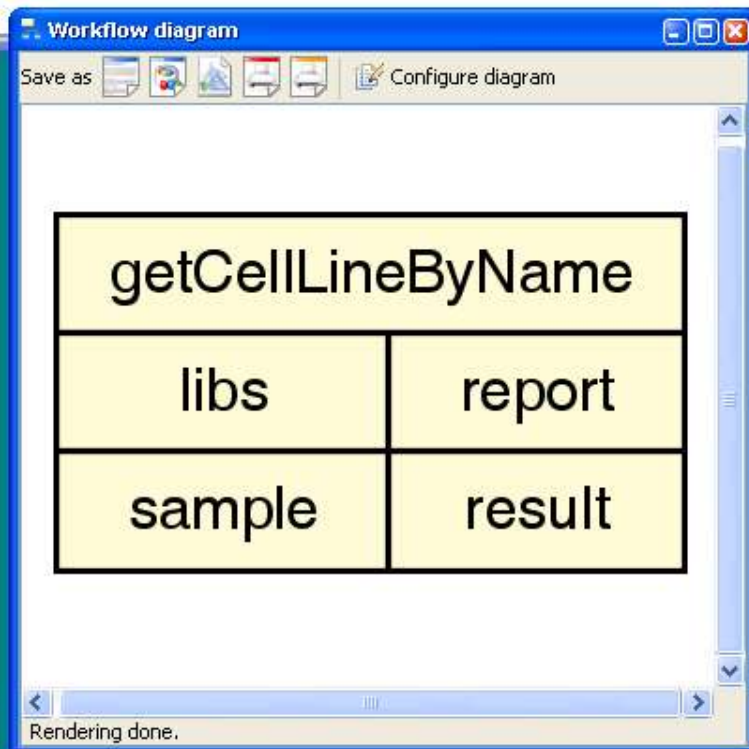


Advanced model explorer

Workflow

Load Load from web Save New subworkflow Offline Reset

Workflow object	Retries	Delay	Back...	Thre...	Critical
Workflow model					
Workflow inputs					
Workflow outputs					
Processors					
getCellLineByName	0	0	1	1	☐
libs 'text/plain'					
sample 'text/plain'					
report 'text/plain'					
result 'text/plain'					
Data links					
Control links					



Available services

Search list pubmed Watch loads

- WSDL @ http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoVi
- WSDL @ http://www.ebi.ac.uk/xembl/XEMBL.wsdl
- WSDL @ http://soap.genome.jp/KEGG.wsdl
- WSDL @ http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.w
- WSDL @ http://soap.bind.ca/wsdl/bind.wsdl
- WSDL @ http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl
- WSDL @ http://www.ebi.ac.uk/ws/WSFasta.wsdl
- WSDL @ http://www.ebi.ac.uk/ws/WSWUBlast.wsdl
- WSDL @ http://www.ebi.ac.uk/ws/WSInterProScan.wsdl
- Soaplab @ http://www.ebi.ac.uk/soaplab/services/
- Biomoby @ http://mobycentral.icapture.ubc.ca/cgi-bin/MOBY05/n
- SeqHound @ seqhound.blueprint.org
- Soaplab @ http://www.o2i.it:8080/axis/services/
- cabri
 - getBacteriaById
 - getBacteriaIdsByName
 - getBacteriaIdsByProperty
 - getCellLineByName
 - getCellLineIdsByName
 - getCellLineIdsByProperty
 - getCellLineNames
 - getCellLinesById
 - getFungiById
 - getFungiIdsByName
 - getFungiIdsByProperty
 - getPhageIdsByName
 - getPhageIdsByProperty
 - getPhagesById
 - getPlasmidIdsByName
 - getPlasmidIdsByProperty
 - getPlasmidsById
 - getResourceIdsByName
 - getResourcesById
 - getYeastIdsByName
 - getYeastIdsByProperty
 - getYeastsById

Advanced model explorer

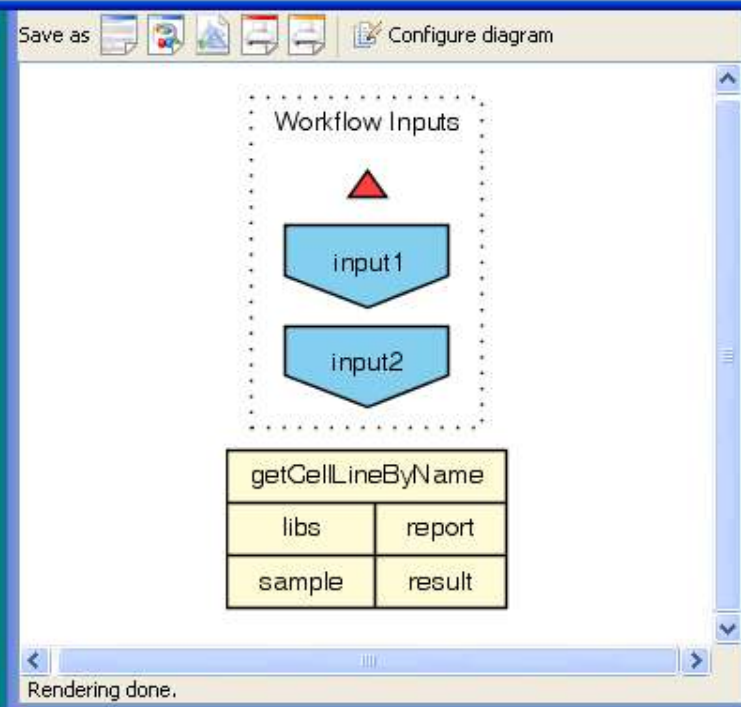
Workflow

Load Load from web Save New subworkflow Offline Reset

Workflow object	Retries	Delay	Back...	Thre...	Critical
Workflow model					
Workflow inputs					
input1					
input2					
Workflow outputs					
Process					
g	0	0	1	1	☐
libs 'text/plain'					
sample 'text/plain'					
report 'text/plain'					
result 'text/plain'					
Data links					
Control links					

Workflow outputs

Create New Output...



Available services

Search list pubmed Watch loads

- WSDL @ <http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoVi>
- WSDL @ <http://www.ebi.ac.uk/xembl/XEMBL.wsdl>
- WSDL @ <http://soap.genome.jp/KEGG.wsdl>
- WSDL @ <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.w>
- WSDL @ <http://soap.bind.ca/wsdl/bind.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSFasta.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSWUBlast.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSInterProScan.wsdl>
- Soaplab @ <http://www.ebi.ac.uk/soaplab/services/>
- Biomoby @ <http://mobycentral.icapture.ubc.ca/cgi-bin/MOBY05/n>
- SeqHound @ seqhound.blueprint.org
- Soaplab @ <http://www.o2i.it:8080/axis/services/>
- cabri
 - getBacteriaById
 - getBacteriaIdsByName
 - getBacteriaIdsByProperty
 - getCellLineByName
 - getCellLineIdsByName
 - getCellLineIdsByProperty
 - getCellLineNames
 - getCellLinesById
 - getFungiById
 - getFungiIdsByName
 - getFungiIdsByProperty
 - getPhageIdsByName
 - getPhageIdsByProperty
 - getPhagesById
 - getPlasmidIdsByName
 - getPlasmidIdsByProperty
 - getPlasmidsById
 - getResourceIdsByName
 - getResourcesById
 - getYeastIdsByName
 - getYeastIdsByProperty
 - getYeastsById

Advanced model explorer

Workflow ▼ Metadata for 'input1'

Load Load from web Save New subworkflow ☐ Offline Reset

Workflow object Retries Delay Back... Thre... Critical

Workflow model

Workflow inputs

input1

Workflow Input : input1

Remove from model

Edit metadata...

Connect to... 0 0 1 1 ☐

Workflow outputs ▶

Processors

getCellLineByName ▶ **Choose an Input**

libs

sample

Data links

Control links

Workflow Outputs

output

getCellLineByName

libs	report
sample	result

Rendering done.

Available services

Search list ☒ Watch loads

- WSDL @ <http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoVi>
- WSDL @ <http://www.ebi.ac.uk/xembl/XEMBL.wsdl>
- WSDL @ <http://soap.genome.jp/KEGG.wsdl>
- WSDL @ <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.w>
- WSDL @ <http://soap.bind.ca/wsdl/bind.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSFasta.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSWUBlast.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSInterProScan.wsdl>
- Soaplab @ <http://www.ebi.ac.uk/soaplab/services/>
- Biomoby @ <http://mobycentral.icapture.ubc.ca/cgi-bin/MOBY05/n>
- SeqHound @ seqhound.blueprint.org
- Soaplab @ <http://www.o2i.it:8080/axis/services/>

cabri

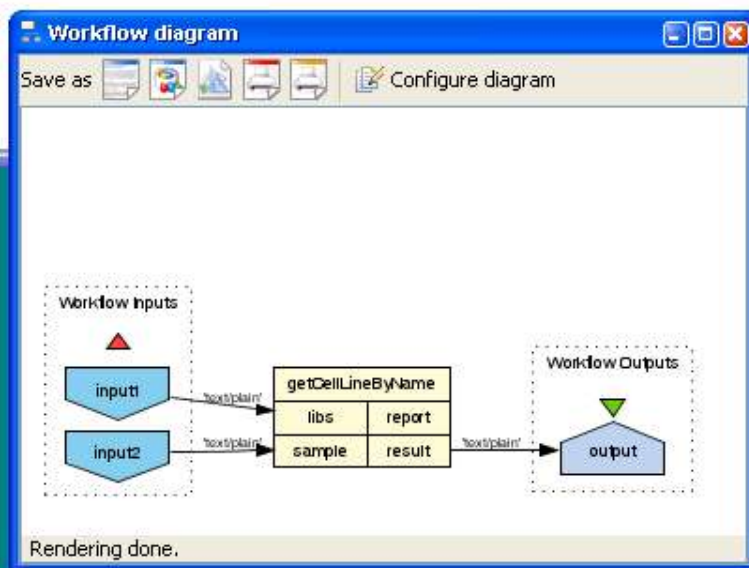
- getBacteriaById
- getBacteriaIdsByName
- getBacteriaIdsByProperty
- getCellLineByName**
- getCellLineIdsByName
- getCellLineIdsByProperty
- getCellLineNames
- getCellLinesById
- getFungiById
- getFungiIdsByName
- getFungiIdsByProperty
- getPhageIdsByName
- getPhageIdsByProperty
- getPhagesById
- getPlasmidIdsByName
- getPlasmidIdsByProperty
- getPlasmidsById
- getResourceIdsByName
- getResourcesById
- getYeastIdsByName
- getYeastIdsByProperty
- getYeastsById

Advanced model explorer

Workflow

Load Load from web Save New subworkflow Offline Reset

Workflow object	Retries	Delay	Back...	Thre...	Critical
Workflow model					
Workflow inputs					
input1					
input2					
Workflow outputs					
output					
Processors					
getCellLineByName	0	0	1	1	☐
libs 'text/plain'					
sample 'text/plain'					
report 'text/plain'					
result 'text/plain'					
Data links					
input1-getCellLineByName:libs					
input2-getCellLineByName:sample					
getCellLineByName:result-output					
Control links					



Available services

Search list pubmed Watch loads

- WSDL @ <http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoVi>
- WSDL @ <http://www.ebi.ac.uk/xembl/XEMBL.wsdl>
- WSDL @ <http://soap.genome.jp/KEGG.wsdl>
- WSDL @ <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.w>
- WSDL @ <http://soap.bind.ca/wsdl/bind.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSFasta.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSWUBlast.wsdl>
- WSDL @ <http://www.ebi.ac.uk/ws/WSInterProScan.wsdl>
- Soaplab @ <http://www.ebi.ac.uk/soaplab/services/>
- Biomoby @ <http://mobycentral.icapture.ubc.ca/cgi-bin/MOBY05/n>
- SeqHound @ seqhound.blueprint.org
- Soaplab @ <http://www.o2i.it:8080/axis/services/>
- cabri
 - getBacteriaById
 - getBacteriaIdsByName
 - getBacteriaIdsByProperty
 - getCellLineByName**
 - getCellLineIdsByName
 - getCellLineIdsByProperty
 - getCellLineNames
 - getCellLinesById
 - getFungiById
 - getFungiIdsByName
 - getFungiIdsByProperty
 - getPhageIdsByName
 - getPhageIdsByProperty
 - getPhagesById
 - getPlasmidIdsByName
 - getPlasmidIdsByProperty
 - getPlasmidsById
 - getResourceIdsByName
 - getResourcesById
 - getYeastIdsByName
 - getYeastIdsByProperty
 - getYeastsById

Taverna Scufl Workbench v1.3

myGrid

<http://taverna.sf.net>

djm,jf,mg,pl,ktg,mp,ms,tmo,mf,ek,pa, jb et al.

Enactment

Advanced model explorer

Workflow: Metadata for 'RetrieveCellLinesDescriptionsByName.xml'

Load Load from web Save New subworkflow Offline Reset

Workflow object	Retries	Delay	Back...	Thre...	Critical
Workflow model					
Workflow inputs					
Workflow outputs					
Processors					
Data links					
Control links					

Workflow diagram

Save as       Configure diagram

Available services

Search list ☒ Watch loads

Available Processors

- Add new Interaction Service... /defaultMartRegistry.xml
- Add new Talisman scavenger... service1/goviz/GoViz.jws?wsdl
- Add new Biomoby scavenger... /wsdl
- Add new Biomart registry... eutils/soap/eutils.wsdl
- Add new API Consumer... n:Dbfetch?wsdl
- Add new WSDL scavenger... sdl
- Add new Styx Grid Service... .wsdl
- Add new Biomart instance... .wsdl
- Add new Workflow scavenger... Scan.wsdl
- Add new Soaplab scavenger... ices/
- Collect scavengers from web... /cgi-bin/MOBY05/mobycentral.pl
- Collect scavengers from model

Expand all
Collapse all

Taverna Scufl Workbench v1.3

myGrid

<http://taverna.sf.net>

djm,jf,mg,pl,ktg,mp,ms,tmo,mf,ek,pa, jb et al.

Enactment

Advanced model explorer

Workflow: Metadata for 'RetrieveCellLinesDescriptionsByName.xml'

Load Load from web Save New subworkflow Offline Reset

Workflow object	Retries	Delay	Back...	Thre...	Critical
Workflow model					
Workflow inputs					
Workflow outputs					
Processors					
Data links					
Control links					

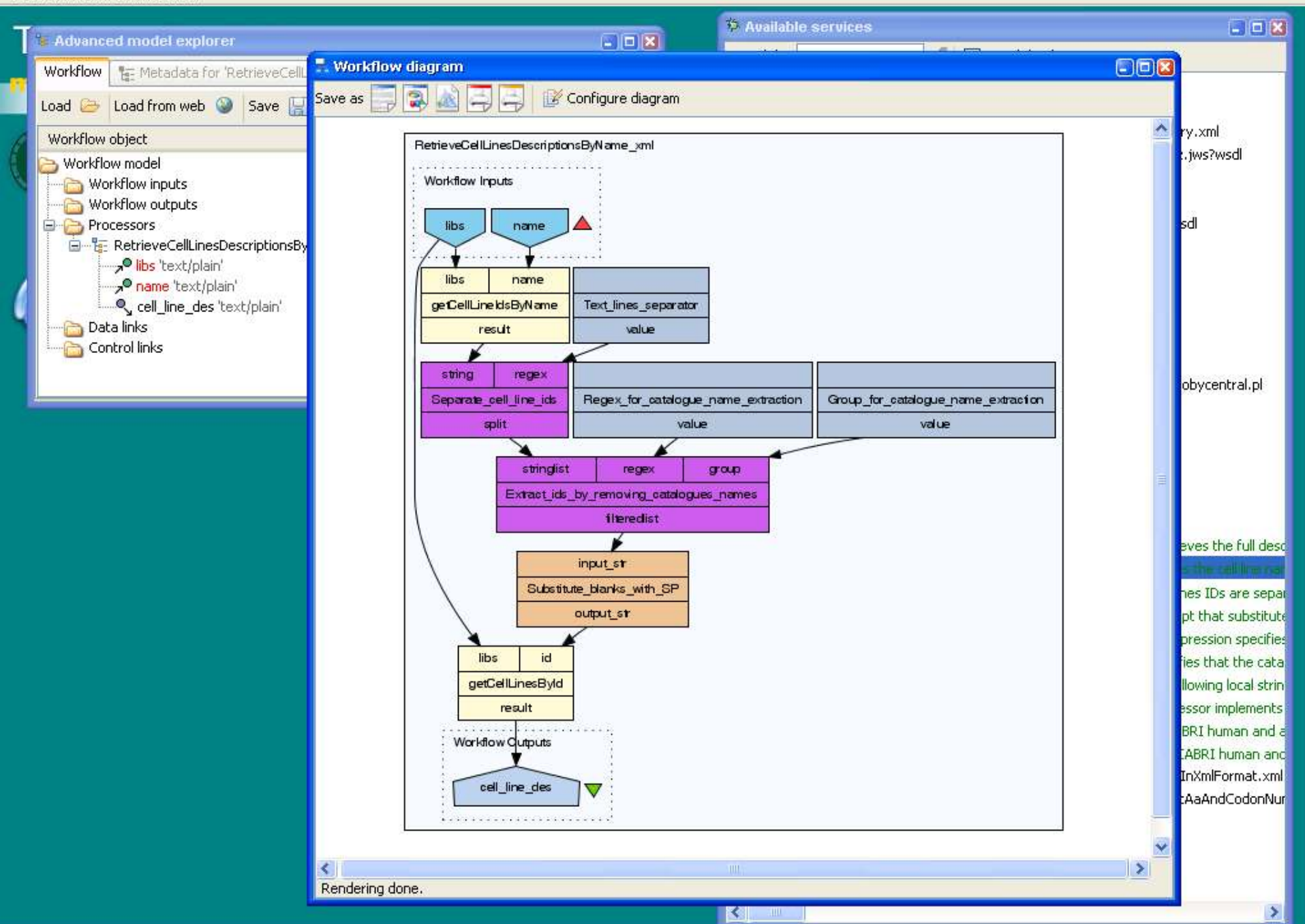
Workflow diagram

Save as Configure diagram

Available services

Search list Watch loads

- Available Processors
 - Local Services
 - Biomart registry @ <http://www.ebi.ac.uk/~tmo/defaultMartRegistry.xml>
 - WSDL @ <http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoViz.jws?wsdl>
 - WSDL @ <http://www.ebi.ac.uk/xembl/XEMBL.wsdl>
 - WSDL @ <http://soap.genome.jp/KEGG.wsdl>
 - WSDL @ <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.wsdl>
 - WSDL @ <http://soap.bind.ca/wsdl/bind.wsdl>
 - WSDL @ <http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl>
 - WSDL @ <http://www.ebi.ac.uk/ws/WSFasta.wsdl>
 - WSDL @ <http://www.ebi.ac.uk/ws/WSWUBlast.wsdl>
 - WSDL @ <http://www.ebi.ac.uk/ws/WSInterProScan.wsdl>
 - Soaplab @ <http://www.ebi.ac.uk/soaplab/services/>
 - Biomoby @ <http://mobycentral.icapture.ubc.ca/cgi-bin/MOBY05/mobycentral.pl>
 - SeqHound @ seqhound.blueprint.org
 - Web crawl @ <http://www.o2i.it/workflows/index2.html>
 - GetMutationsByExon.xml
 - Invoke
 - Add to model
 - Add to model with name...
 - Import workflow...
 - Fetch descriptions
 - GetMutationsById
 - GetMutationsByInt
 - GetTP53Mutation
 - RetrieveBacterial
 - RetrieveCellLines
 - Text_lines
 - Substitute
 - Regex_for_catalogue_name_extraction - This regular expression specifies
 - Group_for_catalogue_name_extraction - This data specifies that the cata
 - Separate_cell_line_ids - This processor implements the following local strin
 - Extract_ids_by_removing_catalogues_names - This processor implements
 - getCellLinesById - Get cell line descriptions by Id from CABRI human and a
 - getCellLineIdsByName - Get cell lines' IDs by name from CABRI human and
 - RetrieveCellLineAndRelatedTp53MutationWithRelatedPubmedInXmlFormat.xml
 - GetTP53MutationFunctionEntriesAndTP53CellLineEntriesByMutAaAndCodonNur
 - GetTP53MutationFunctionEntriesByExon.xml -



Advanced model explorer

Workflow: Metadata for 'RetrieveCellLinesDescriptionsByName.xml'

Load Load from web Save New subworkflow Offline Reset

Workflow object	Retries	Delay	Back...	Thre...	Critical
Workflow model					
Workflow inputs					
Workflow outputs					
Processors					
RetrieveCellLinesDescriptionsByNamr	0	0	1	1	☐
libs 'text/plain'					
name 'text/plain'					
cell_line_des 'text/plain'					
Data links					
Control links					

Workflow diagram

Save as Configure diagram

Rendering done.

Available services

Search list Watch loads

Available Processors

- Local Services
 - Biomart registry @ <http://www.ebi.ac.uk/~tmo/defaultMartRegistry.xml>
 - WSDL @ <http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoViz.jws?wsdl>
 - WSDL @ <http://www.ebi.ac.uk/xembl/XEMBL.wsdl>
 - WSDL @ <http://soap.genome.jp/KEGG.wsdl>
 - WSDL @ <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.wsdl>
 - WSDL @ <http://soap.bind.ca/wsdl/bind.wsdl>
 - WSDL @ <http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl>
 - WSDL @ <http://www.ebi.ac.uk/ws/WSFasta.wsdl>
 - WSDL @ <http://www.ebi.ac.uk/ws/WSWUBlast.wsdl>
 - WSDL @ <http://www.ebi.ac.uk/ws/WSInterProScan.wsdl>
 - Soaplab @ <http://www.ebi.ac.uk/soaplab/services/>
 - Biomoby @ <http://mobycentral.icapture.ubc.ca/cgi-bin/MOBY05/mobycentral.pl>
 - SeqHound @ seqhound.blueprint.org
 - Web crawl @ <http://www.o2i.it/workflows/index2.html>
- GetMutationsByExon.xml -
- GetMutationsByIntronAndEffect.xml -
- GetMutationsByIntron.xml -
- GetTP53MutationsByIntronAndEffect2.xml -
- RetrieveBacteriaDescriptionsByName.xml - This workflow retrieves the full description of a bacteria from the NCBI database.
- RetrieveCellLinesDescriptionsByName.xml - This workflow takes the cell line name as input and returns the cell line description from the CABRI database.
- Text_lines_separator - This string ("\n") specifies that cell lines IDs are separated by a newline character.
- Substitute_blanks_with_SP - This is a trivial beanshell script that substitutes spaces with underscores.
- Regex_for_catalogue_name_extraction - This regular expression specifies the pattern for extracting the catalogue name from the cell line ID.
- Group_for_catalogue_name_extraction - This data specifies that the catalogue name is a group of characters.
- Separate_cell_line_ids - This processor implements the following local string: `String[] ids = id.split("_");`
- Extract_ids_by_removing_catalogues_names - This processor implements the following local string: `String[] ids = id.split("_");`
- getCellLinesById - Get cell line descriptions by Id from CABRI human and mouse databases.
- getCellLineIdsByName - Get cell lines' IDs by name from CABRI human and mouse databases.
- RetrieveCellLineAndRelatedTp53MutationWithRelatedPubmedInXmlFormat.xml -
- GetTP53MutationFunctionEntriesAndTP53CellLineEntriesByMutAaAndCodonNumber.xml -
- GetTP53MutationFunctionEntriesByExon.xml -

Obiettivi del sistema:

- Mettere a disposizione di ricercatori non esperti un insieme predefinito di workflow (testati, validati, annotati, mantenuti)
- Consentire la ricerca e la selezione dei workflow sulla base di una loro annotazione basata su una semplice ontologia dei processori bioinformatici (dominio, task, i/o) e della tipologia dell'utente (interessi e ruolo)
- Consentire l'esecuzione interattiva dei workflow selezionati
- Consentire la memorizzazione e il recupero dei risultati dei workflow eseguiti

biowep

Il sistema:

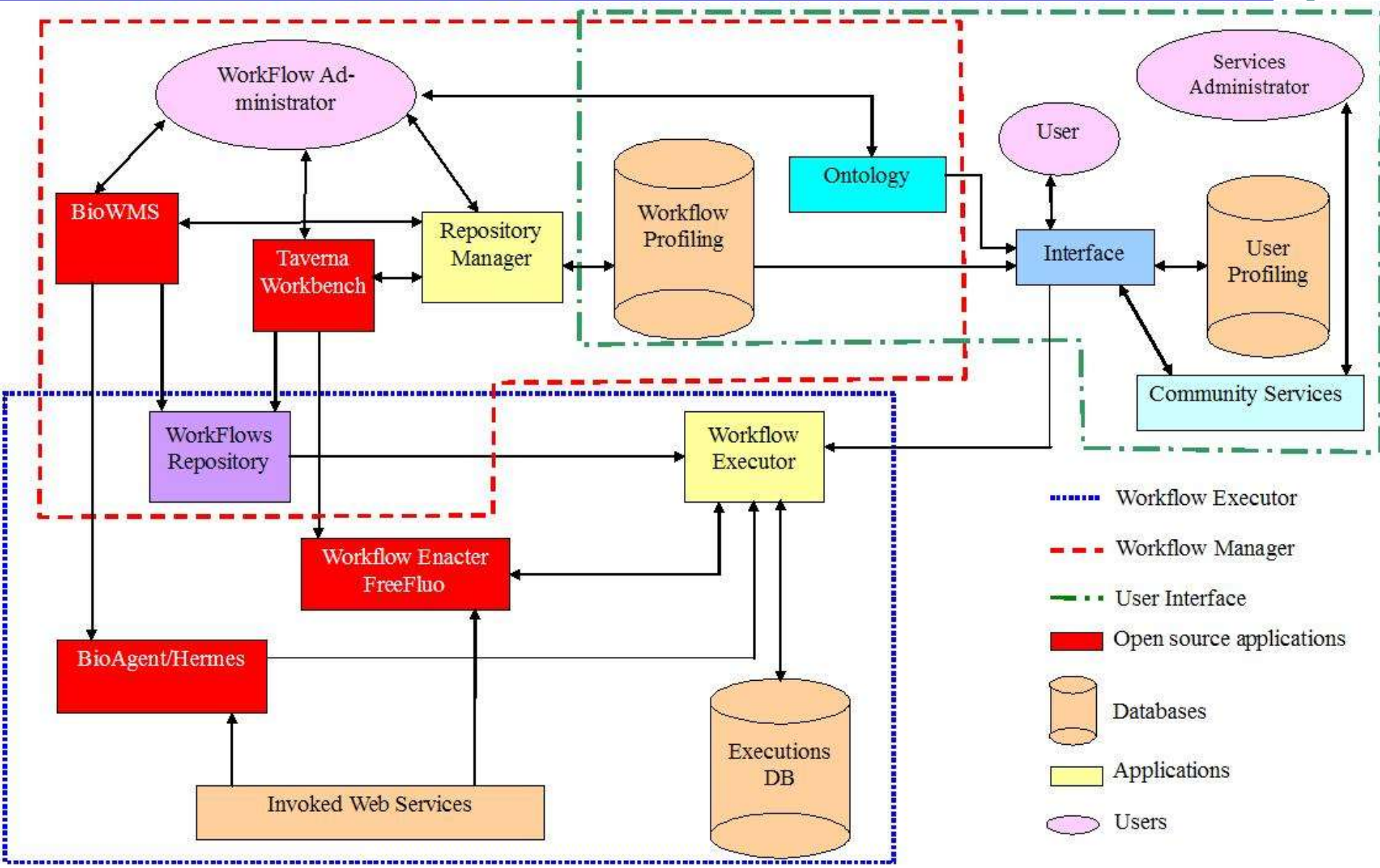
- è parzialmente basato su software open source (Taverna WB, FreeFluo e MySQL)
- accede a dati e analisi disponibili tramite Web Services
- archivia i workflow in formato Scufi o XPD
- è disponibile (licenza LGPL)

Prototipo disponibile on-line:

<http://bioinformatics.istge.it/biowep/> (sito supporto)

<http://bioinformatics.istge.it:8080/biowep/> (portale)

Architettura di biowep



Workflow predefiniti

I workflow sono:

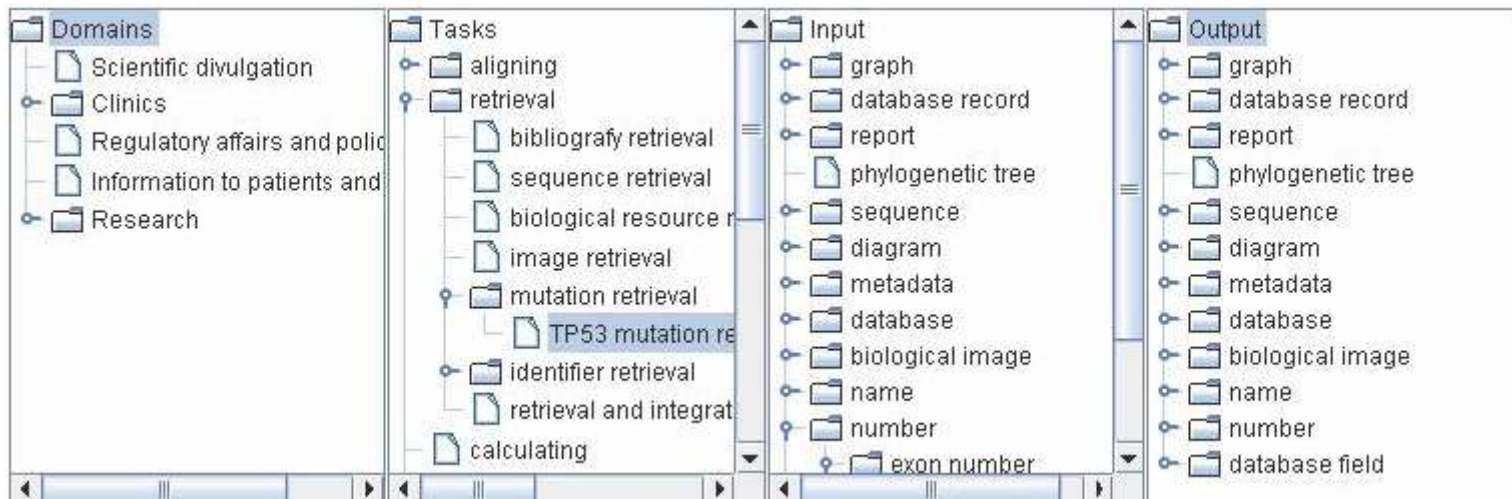
- creati dall'amministratore con Taverna o BioWMS
- archiviati in formato Scufi o XPD
- aggiornati (workflow vs versione)
- possono essere sottomessi da utenti o service providers tramite il sito di supporto

Annotazione dei workflow

I workflow sono annotati sulla base di:

- una semplice ontologia per processori bioinformatici:
 - dominio applicativo
 - task
 - input e output
- l'ontologia deriva da quella di Taverna:
 - nuova struttura
 - alcune aggiunte (biological resources, images, ...)
 - in via di ulteriore sviluppo e adattamento

Annotazione dei workflow



Dominio applicativo	Tipo di elaborazione	Dato in input	Dato in output
---------------------	----------------------	---------------	----------------

Registrazione e profilo utenti

Gli utenti sono classificati in base a:

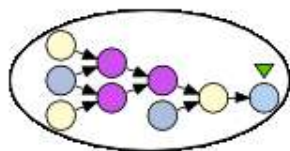
- ruolo ricoperto
 - informatico / medico / ricercatore / paziente / giornalista / ...
- ambito di ricerca di interesse

I workflow sono proposti sulla base di:

- ruolo / ambito di interesse
- precedenti esecuzioni



Indirizzo http://bioinformatics.istge.it:8080/biowep/ Vai



biowep: a Workflow Enactment Portal for Bioinformatics

Please login

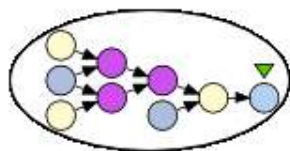
Username:

Password:

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:

Choosing your Username and Password

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

Capitalization matters for your 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

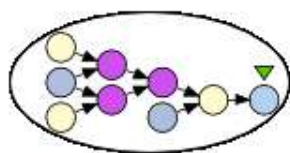
Describing yourself

Please give us some information about you.

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

Do you want to receive news through your e-mail address?

☒ Yes ☐ No



biowep: a Workflow Enactment Portal for Bioinformatics

USER: PaoloR

Clone Window

logout

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

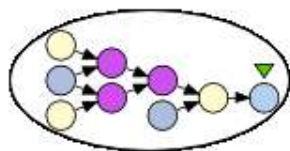
Temporary saved results

Persistently saved results

Edit your profile

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

Clone Window

logout

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

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 substituted by a '_SP_' characters string before submitting 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

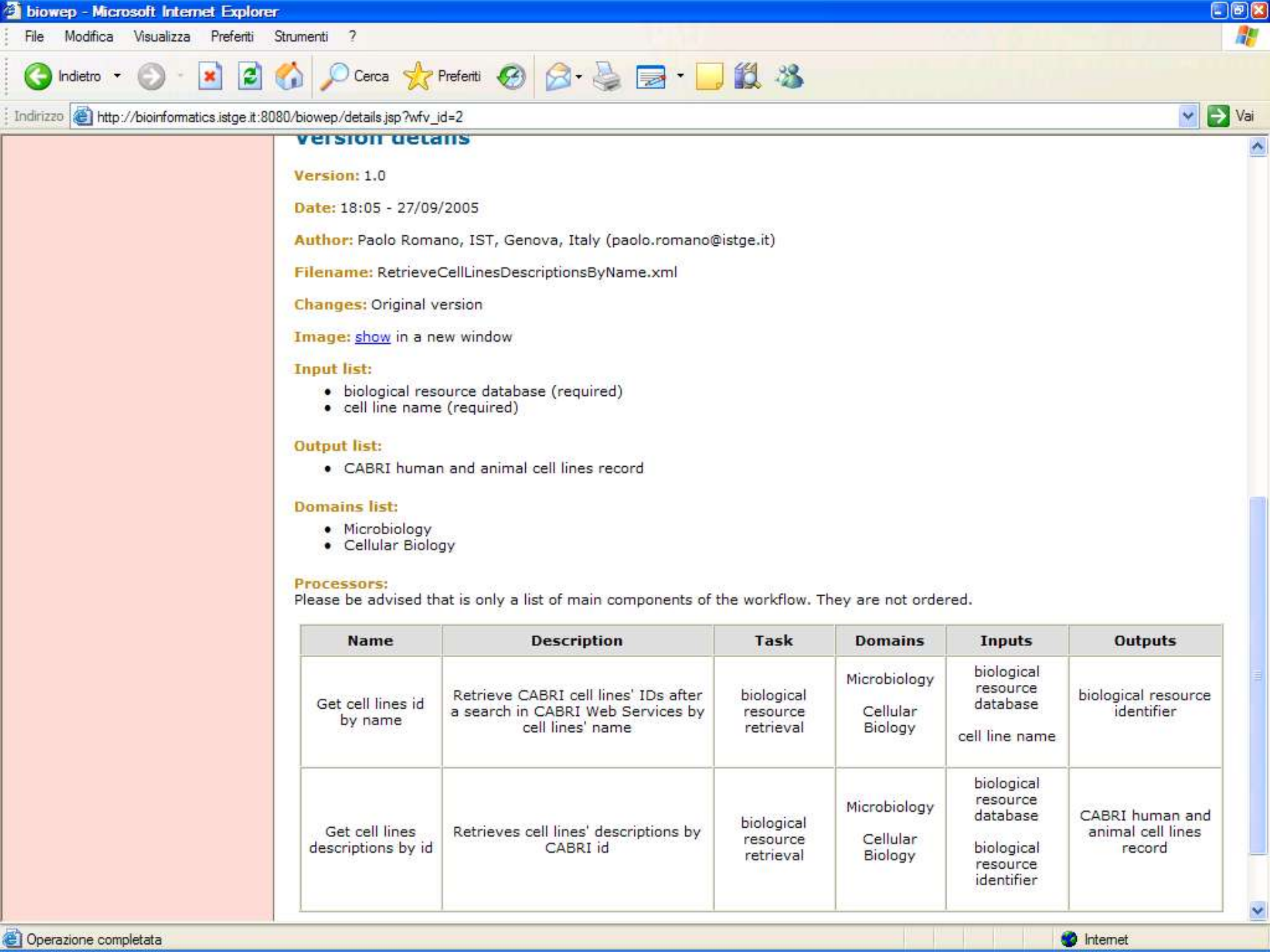
Author: Paolo Romano, IST, Genova, Italy (paolo.romano@istge.it)

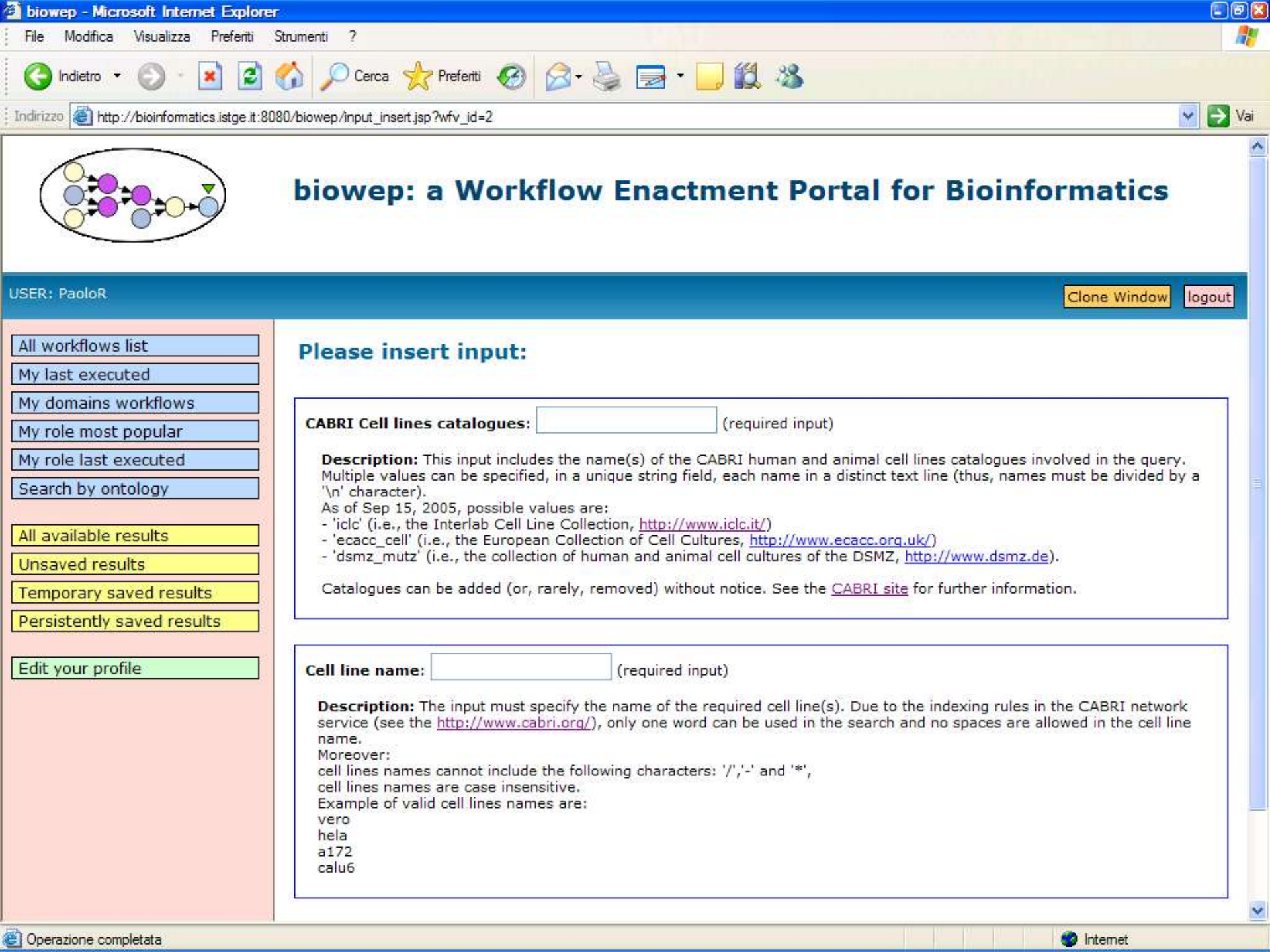
Roles list:

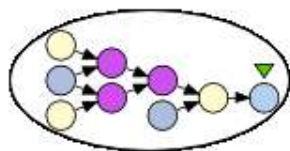
- researcher - molecular biologist
- researcher - cellular biologist
- researcher - structural biologist
- researcher - microbiologist
- researcher - immunologist
- researcher - virologist

Version details

Version: 1.0







biowep: a Workflow Enactment Portal for Bioinformatics

Username: PaoloR

logout

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

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](#)

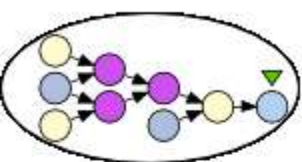
☐ [getCellLinesById](#)

Select All

Save selected (permanently)

Save selected (temporarily)

Delete selected



biowep: a Workflow Enactment Portal for Bioinformatics

name: PaoloR

Clone Window

logout

- workflows list
- last executed
- domains workflows
- role most popular
- role last executed
- arch by ontology
- available results
- saved results
- temporary saved results
- persistently saved results
- edit your profile

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

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

[Select All](#)[Save selected \(permanently\)](#)[Save selected \(temporarily\)](#)[Delete selected](#)

Legenda:

- not saved
- temporarily saved
- persistently saved

biowep: a Workflow Enactment Portal for Bioinformatics

Jusername: domenico

Clone Window

logout

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
 - bibliografy retrieval
 - sequence retrieval
 - biological resource r
 - image retrieval
 - mutation retrieval
 - TP53 mutation re
 - identifier retrieval
 - retrieval and integrat
- calculating

Input

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

Output

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

search

"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

Prossimi sviluppi

- Supporto ai ricercatori (manuali, FAQ, mailing list, ...)
- Inserimento di workflow sottomessi da ricercatori esterni
- Creazione di nuovi workflow
- Estensione di biowep con nuovi WMS / enacters

Collaborazioni

IST, Genova

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

ISMAL, CNR, Genova

Patrizio Arrigo,
Matteo Fattore

ITB, CNR, Milano

Luciano Milanese

DISCo, Università di Milano Bicocca

Guglielmo Bertolini,
Flavio De Paoli,
Giancarlo Mauri

DIST, Università di Genova

Ivan Porro,
Silvia Scaglione

DMI, Università di Camerino (MC)

Emanuela Merelli,
Ezio Bartocci

Lavoro svolto nell'ambito dei progetti "Oncology over Internet" (MIUR, 2002 – 2005) e Laboratorio Interdisciplinare di Tecnologie Bioinformatiche (MIUR, 2005-2008).