Workflow Management Systems in supporto alla ricerca biomedica:

verso l'automazione dell'accesso ai dati e falla loro analisi in rete

Paolo Romano

S.C. Bioinformatica e Proteomica Strutturale

Istituto Nazionale per la Ricerca sul Cancro

(paolo.romano@istge.it)

Sommario

- Integrazione dei dati in biologia
- Strumenti e metodologia per l'automazione dei processi d'analisi
- Conclusioni

L’informazione biologica

La ricerca biomedica produce un’enorme quantità di dati

- La genomica e la proteomica, contribuiscono alla realizzazione di banche dati di rilevanti dimensioni
- Altri settori contribuiranno in futuro con quantità di dati ancora superiori:
  - analisi di variabilità (mutazioni, polimorfismi)
  - percorsi biochimici / metabolici
  - strumenti HT (microarray)
  - ….

Dimensione informazioni biologia

- 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 (http://www3.ebi.ac.uk/Services/DBStats/)
- ArrayExpress (21/12/2005):
  - Esperimenti 1,187 (circa 800 Gb)
  - Aumento +100% da Ottobre 2004 a Ottobre 2005 (http://www.ebi.ac.uk/arrayexpress/Help/stats/index.html)
- Nucleic Acids Research Supplement (2006)
  - 858 database di biologia molecolare (http://nar.oxfordjournals.org/cgi/content/full/33/suppl_1/D5/DC1)
- SRS public sites (2006)
  - 1,300 librerie (http://downloads.lionbio.co.uk/publisrs.html)

Banche dati in biologia

- **Eterogenee**:
  - Poche banche dati sono gestite in modo omogeneo
  - Banche dati secondarie di ottima qualità (annotazione estesa, controllo accurato)
  - Molte banche dati specializzate:
    - gene/genoma,
    - organismo,
    - patologia,
  - Molte banche dati sviluppate da singoli ricercatori o piccoli gruppi di ricerca

- **Distribuite**:
  - Strutture dati, DBMS e metodi di distribuzione sono differenti
  - Informazioni e significati (semantica) sono differenti

Obiettivi dell'integrazione

L'integrazione dei dati è necessaria per:

- Ottenere una visione complessiva e più precisa delle informazioni disponibili
- Eseguire in maniera integrata query e/o analisi dati che coinvolgono più database e software
- Eseguire con efficienza analisi che coinvolgono grosse quantità di dati
- Realizzare un effettivo data mining

Specificità dell'integrazione

In ambito biologico:

- Una pre-analisi delle informazioni è impossibile: dati e conoscenze cambiano frequentemente e rapidamente
- La complessità delle informazioni non permette di creare modelli validi in diversi ambiti e 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

- Strumenti tradizionali (data warehouse, software di integrazione SRS) pongono problemi: dimensione, aggiornamento, struttura variabile
- L'integrazione deve essere sviluppata con sistemi flessibili, adattabili ed espandibili

Verso l'automazione dei processi

Una possibile metodologia:

- 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 e Web Services in bioinformatica

Linguaggi XML

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

Web Services

- EMBOSS, XEMBL, Interpro (EBI)
- eUtils (NCBI)
- caBIO (NCICB)
- KEGG API
- GeneCruiser, Biosphere (microarray)
- SIMAP (proteine)
- Cataloghi CABRI (risorse biologiche)
- Mutazioni TP53

Workflow

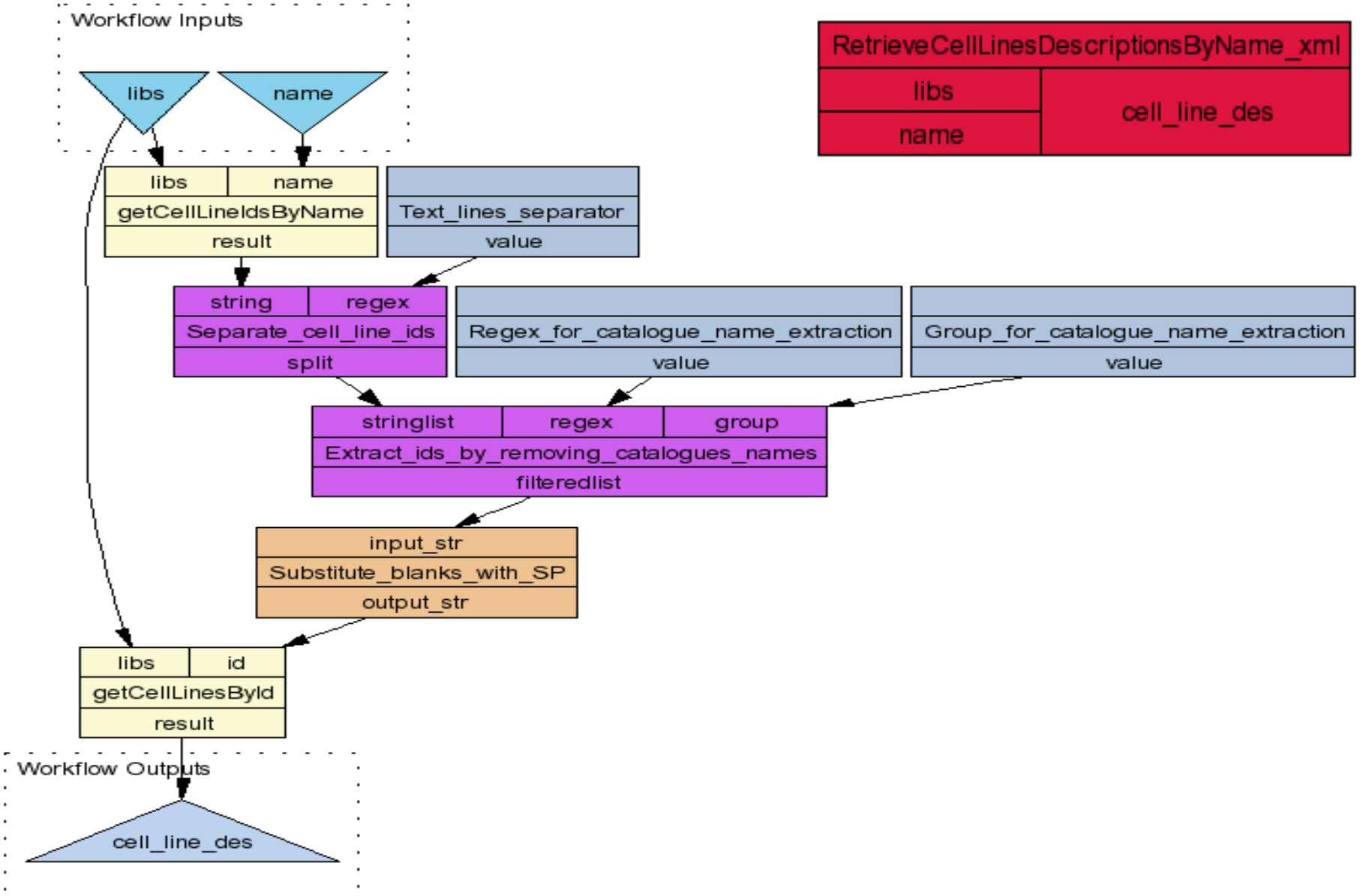
"**A computerized facilitation or automation of a business process, in whole or part**". (Workflow Management Coalition)

Obiettivo:

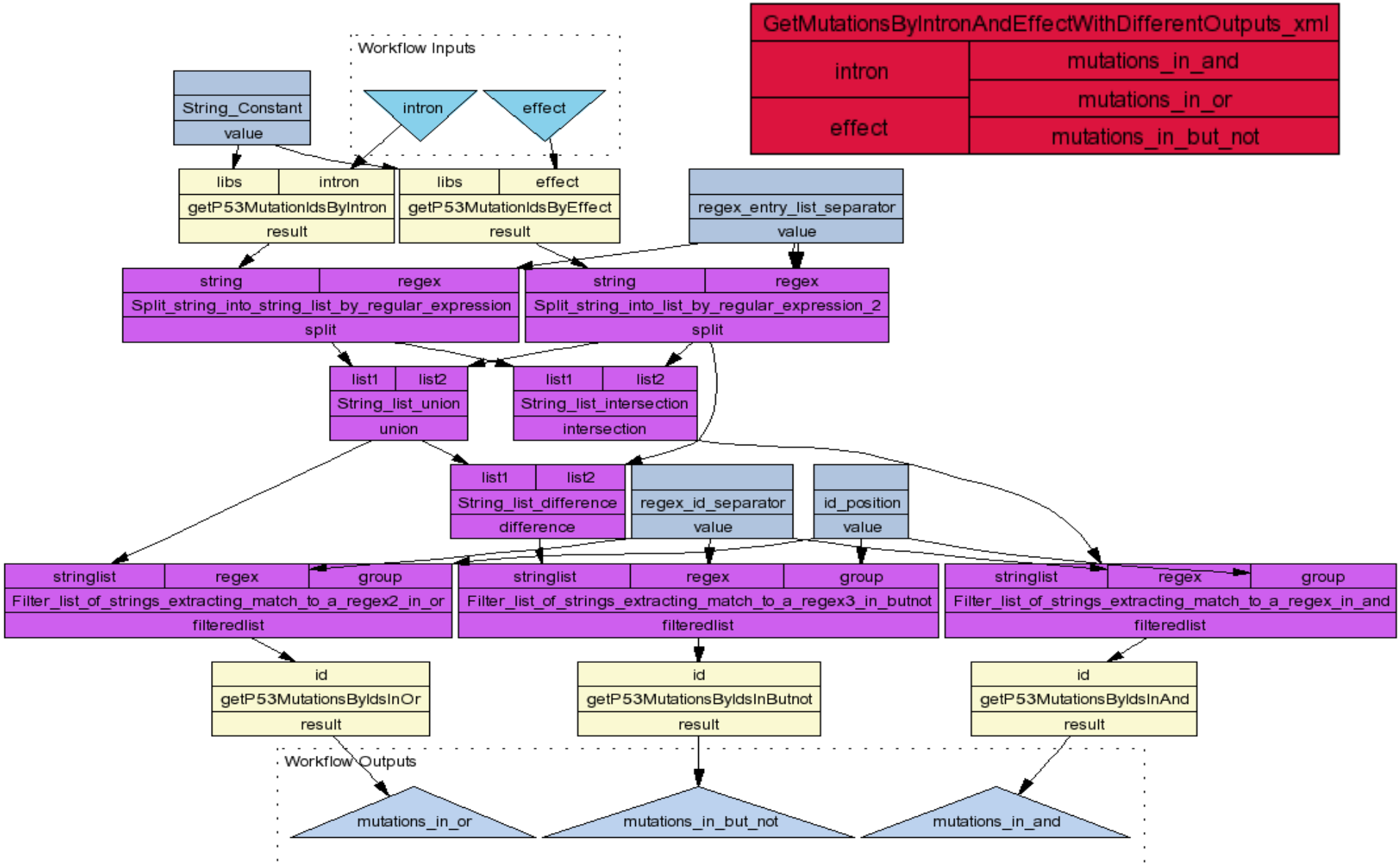
- implementazione di processi di analisi dei dati in ambienti standardizzati

Vantaggi principali:

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



Workflow per database CABRI



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.

Presentano diverse tipologie e utilizzano diversi standard.

biowep: obiettivi

Workflow Enacment Portal for Bioinformatics

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: caratteristiche

Il sistema:

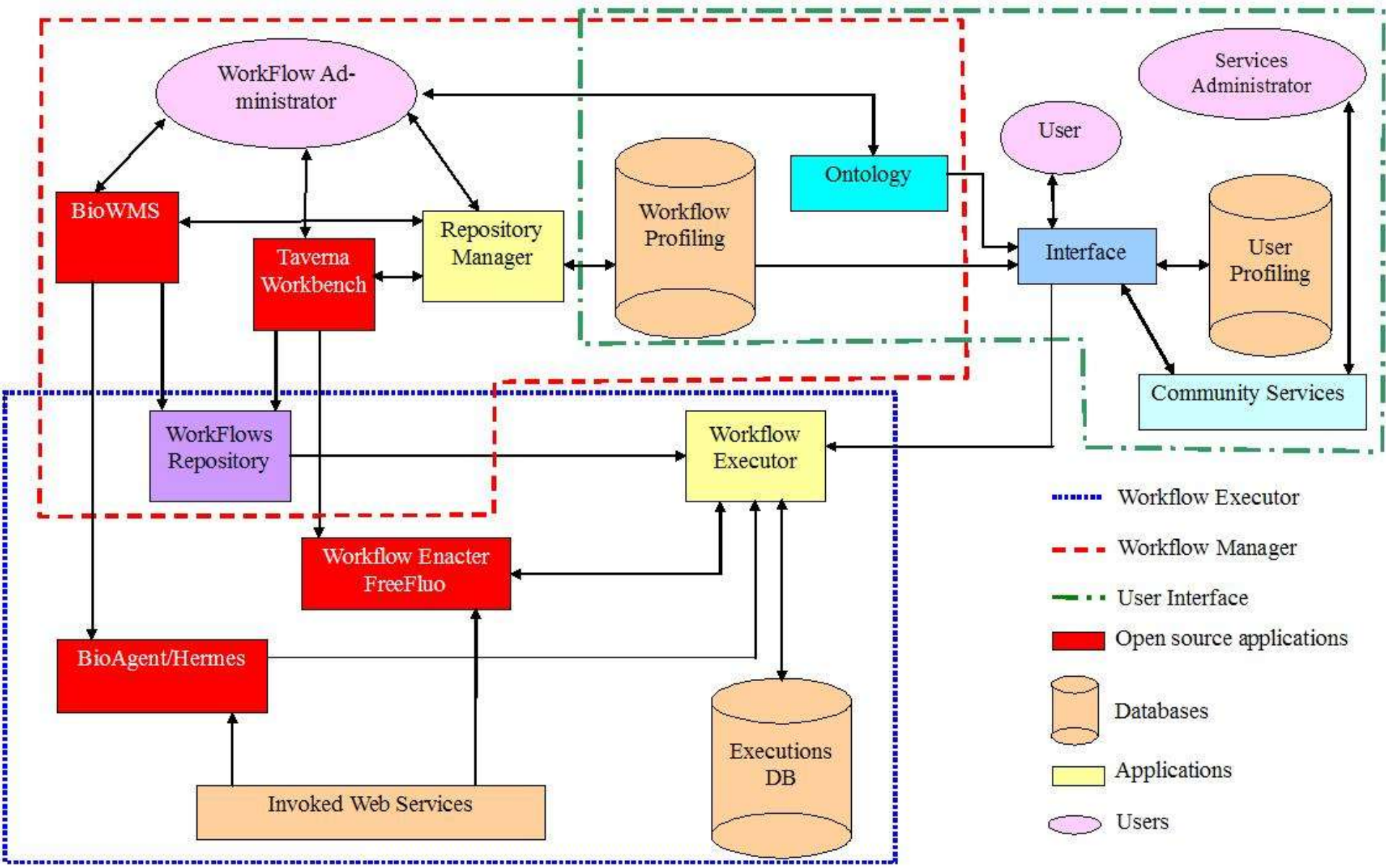
- è scritto in java + javascript
- è parzialmente basato su software open source (Taverna WB, FreeFluo, Tomcat e mySQL)
- accede a dati e analisi disponibili tramite Web Services
- può svolgere anche elaborazione locali
- archivia i workflow in formato Scufl o XPDL
- è distribuito free (licenza LGPL)

Prototipo disponibile on-line:

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

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

biowep: architettura



biowep: workflow e utenza

I workflow sono:

- creati dall'amministratore o inviati da utenti con Taverna o BioWMS
- archiviati in formato Scufl o XPDL
- aggiornati (workflow vs versione)
- annotati sulla base di un'ontologia dei task in bioinformatica (dominio, task, dati I/O)

Gli utenti:

- accedono in remoto al sistema
- sono registrati (accesso controllato)
- lavorano in un proprio ambiente (sessioni, risultati, wf)

Conclusioni

- Metodologia, tecnologia e standard per l'automazione delle procedure di ricerca, integrazione e analisi dei dati biomedici
- Opportunità per:
  - Sviluppo di strumenti software (WMSs, portali)
  - Sviluppo di workflow d'analisi
  - Attività di supporto/consulenza

Collaborazioni

IST, Genova
Paolo Romano,
Domenico Marra,
Chiara Rasi,
Ulrich Pfeffer,
Valentina Mirisola,
M. Assunta Manniello,
Gilberto Fronza

DIST, Università di Genova
Ivan Porro,
Silvia Scaglione

DISCo, Università di Milano Bicocca
Guglielmo Bertolini,
Flavio De Paoli,
Giancarlo Mauri

ITB, CNR, Milano
Luciano Milanesi

DMI, Università di Camerino (MC)
Emanuela Merelli,
Ezio Bartocci

ISMAC, CNR, Genova
Patrizio Arrigo

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