Semi-automatic and efficient annotation of bioinformatics processors in the biowep workflow enactment portal

ID - 181

Romano Paolo[1], Marra Domenico[1], Rasi Chiara[1]

[1]Bioinformatics and Structural Proteomics, National Cancer Research Institute. Genova

Motivation

Workflow systems are a valid option to coordinate access to and retrieve data from standardized Web Services. Various workflow management systems (WMS) for bioinformatics have been developed. Nevertheless, creation of workflows can be difficult, since it implies knowledge of available Web Services and data formats, not mentioning programming skills. Workflows enactment portals are therefore being developed.

Biowep, a workflow enactment portal for bioinformatics, has been made available on-line to all researchers [1,2]. It allows the carrying out of predefined workflows and the storing and retrieval of workflows executions and related results. It supports the annotation of workflows components through an ontology for bioinformatics tasks. Search and selection of workflows can be done on the basis of their annotation. Biowep makes use of open source: the WMS Taverna [3] and mySQL.

Here, we present bioweps Workflow Repository Manager (WfRM), a web application for the management of workflows in the workflow repository. WfRM supports the semi-automatic, efficient insertion, update and annotation of workflows described with XScufl, a workflows language developed within the myGrid initiative [4,5].

Methods

WfRM has been implemented as a front-end for the administration of biowep. It has been written by using JavaServer Pages (JSP) technology, that provides a fast, simplified, both server- and platform-independent way to create dynamic web content.

WfRM provides a user-centred interface for the uploading of workflows written in the XScufl language. It includes a Java class backend component connecting the interface with the workflow repository, based on a MySQL database.

Uploaded workflows are first stored in the working directory, then syntactically validated and finally parsed by using a set of SAX-based classes. These return workflows values to the client application, therefore promoting an application-driven insertion of basic data, such as workflows name, description and author, in the database. Other information, such as the workflows application domains, must be added by the user.

In our db schema, a distinction is made between a workflow and its implementations, that we call versions. A workflow is only conceptually described, on the basis of its goals, and it does not refer to any actual file. Instead, each version is strictly linked to one file, that can be enacted and give results. Versions can differ among them, e.g., for accessed Web Services, offering alternative, but equivalent, services, and local elaboration processors, that can be modified by keeping the same function. So, WfRM makes a distinction between uploading a new workflow, in which case the associated file is assigned to a first version of the workflow, or a new version of an existing workflow.

Submitted files include a description of processors, their links, and the overall inputs and outputs of the workflow. This information is semi-automatically and efficiently annotated by WfRM through a classification of bioinformatics data and tasks. We choose to annotate the overall workflow and the most significant processors (the choice of which is left to the user). Annotation is then inserted into the database, while the workflow itself is not changed. A Java applet provides the researcher an exploratory tool for identifying and selecting the best definitions for annotating application domain, elaboration task and input and output data types.

Annotations may be updated (inserted, modified or deleted) at any time.

Our classification of bioinformatics tasks and data was derived from the original myGrid ontology [6], that has been reorganized, by using a different hierarchy, and expanded, by adding biological resources and images data types. This annotation is also used when searching for workflows in the repository.

Results

We presented WfRM, a user-friendly interface implemented as a tool for the efficient and semi-automatic management of information within biowep workflow repository.

Before, insertion of workflows in the repository was a complex and time-consuming procedure, requiring manual updating of database contents. Now, system maintenance is easy and intuitive. Workflows basic data is collected, processors are annotated by a proper ontology, and the database is updated in a coherent and effective way.

Planned development of WfRM are aimed at overcoming a limitation of the XScufl language, which does not allow to associate any data type to processors I/O.

Associating semantic information to processors I/O would be useful for selecting workflows, but also to support the creation of workflows having well interconnected processors and their validation at enactment time.

Improvement of the bioinformatics data and tasks ontology is also planned, with the aim of broadening its application domain and comparing with similar ontologies and classifications.

Availability: http://bioinformatics.istge.it/biowep/

Email: paolo.romano@istge.it