Guided Composition of Tasks With Logical Information Systems - Application to Data Analysis Workflows in Bioinformatics*

Mouhamadou Ba

IRISA/INSA Rennes, 35043 Rennes, France
mouhamadou.ba@irisa.fr

Abstract. In a number of domains, particularly in bioinformatics, there is a need for complex data analysis. For that issue, elementary data analysis operations called tasks are composed as workflows. The composition of tasks is however difficult due to the distributed and heterogeneous resources of bioinformatics. This doctorial work will address the composition of tasks using Logical Information Systems (LIS). LIS let users build complex queries and updates over semantic web data through guided navigation, suggesting relevant pieces and updates at each step. The objective is to use semantics to describe bioinformatic tasks and to adapt the guided approach of Sewelis, a LIS semantic web tool, to the composition of tasks. We aim at providing a tool that supports guided composition of semantic web services in bioinformatics, and that will support biologists in designing workflows for complex data analysis.

1 Motivation and Research Questions

The Workflow Management Coalition (WfMC)[1] defines a workflow as "the automation of business processes, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules." Originally used by industry for business processes, workflows have been increasingly used to lead *in silico* experiments in scientific areas. Web services [1] are being used as components of workflows, they provide access to data sources and to tools to analyse data. The bioinformatic domain is much involved in the use of workflows of web services, for example complex data analysis is performed by composing various elementary data analysis operations (e.g. search for homologous sequences, transcription). In bioinformatics however, the resources are complex and heterogeneous. They are produced and maintained by groups localized around the world. The nature of the bioinformatic domain raises distribution and heterogeneity problems, making it difficult to compose tasks.

* This work started in October 2012 under an ARED funding from Region Bretagne and is supervised by Mireille Ducassé (IRISA/INSA Rennes) and Sébastien Ferré (IRISA/University Rennes 1)

[1] http://www.wfmc.org/

The semantic web [2] provides technologies that facilitate the composition of web services as workflows. Ontologies for example, are used to describe bioinformatic resources, including meta data, data types and tasks, which eases resource integration. Discovery helps to access services that will compose a workflow. The description of characteristics of services through technologies like RDF facilitates their discovery. Languages like OWL can be used to allow to constraint and reason on workflow management systems. Those technologies can also help manage tasks, results and data provenance. Semantic web technologies can support automation of some manual tasks (e.g. service selection) during workflow definition.

There are different approaches for workflow definition: the manual approach and the automatic approach. The manual approach requires users to entirely define the workflow. With that approach, the definition and update of workflows require too much training for end users. The automatic approach selects components and defines workflows in an automatic manner. That requires strong and complete specifications, which are themselves difficult to express.

Our goal is to provide an environment for the design of workflows in a semi-automatic approach that combines the advantages of manual and automatic approaches. We will use semantic web technologies to support guided composition of services. The work will be applied to the bioinformatic domain. We nevertheless aim to produce methods and tools that are generic and relevant to other fields.

2 State of the Art

A web service corresponds to a set of operations whose characteristics are generally described through an XML-based standard language. It is accessible through standardized web protocols such as SOAP (Simple Object Access Protocol). Web service technologies can serve as infrastructure for workflow development.

Many tools exist to define web services. Some of them operate at a syntactic level, others up to the semantic level. The XML standards WSDL (Web Service Definition Language) and UDDI (Universal Description Discovery and Integration) are defined for, respectively, the description of services and the publication and access to services. They operate at the syntactic level. New languages are proposed to add semantics to the definition of services, for example OWL-S, WSMO and SAWSDL. They lead to semantic web services. Some annotation models based on SAWSDL, OWL-S and WSMO for service annotation force users to think in terms of service interfaces, rather than of high-level functionality. Missier et al. [3], to increases the effectiveness of annotation models, define *Functional Units* (FU) as the elementary units of information used to describe a service. Ontologies like myGrid ontology [4] are also proposed to support the description of web services and data.

In bioinformatics, implementations of web services (e.g. BioMoby [5]) are proposed by institutes and web service providers. Those services are used by many systems for different needs, such as the definition of workflows [6].

Many languages are also proposed to define workflows. Wang et al. [7] present a survey of such languages. For example, the Scufl language (Semantic Conceptual Unified Flow Language) is provided to define scientific workflows in the context of the myGrid project[2]. In face of the great number of workflow languages, criteria are important to choose a language, for example complexity, semantic, license, stability, executability, generalizability, shareability. The quality and degree of automation in the workflow design process depend on the chosen language.

There are many approaches for semi-automatic composition of web services. Some of them use Semantic Web technologies and Artificial Intelligence techniques to assist users in web service selection and composition. Wang et al. [7] assess some of them using the following criteria: use of ontologies [8], filtering of inappropriate services, suggestion of partial plans, checking of the composition validity, use of a planning strategy, use of a modeling environment [8], control constructs and executable results.

Taverna is a component of the myGrid project. MyGrid aims at developing a middleware to support data intensive *in silico* experiments in biology. Taverna [8] is a tool to compose and enact bioinformatic workflows. Its GUI allows biologists to create, execute and share workflows. However, while being much simpler than raw programming, Taverna and similar systems are still difficult to use for average biologists. In Taverna the creation of workflows is neither interactive nor guided enough, there are no automatic data mediation, and no suggestions are made during the workflow design process [7].

3 Approach and Research Methodology

The LIS team[3] has an expertise in guided approaches for data exploration and authoring. Logical Information Systems (LIS) let users build complex queries and updates over semantic web data through guided navigation, suggesting relevant pieces and updates at each step. That approach combines query search and faceted search and is implemented in Sewelis [9]. For example, Sewelis has been applied to the exploration of films and related people, and to the semantic annotation of comics panels.

Our work will address the design of workflows using the LIS approach. The design of workflows requires the location of the relevant tasks, the LIS approach will facilitate integration and management of tasks as well as the selection of tasks that can be matched together to form a workflow. We aim at extending the guided approach of Sewelis to the composition of tasks in order to make it easier for biologists. A visual environment will help users to design workflows. That environment will integrate Sewelis for the retrieval of tasks. We will take advantage of related work and tasks will be wrapped as semantic web services. The suggestion mechanisms and reasoning engine of Sewelis will be adapted

[2] http://www.mygrid.org.uk/
[3] http://www.irisa.fr/LIS/

to enable automatic parameter matching [10], selection of services and guided edition of workflows. We will address the following tasks:

Resource description: This task will be the semantic basis of our work. We envision the reuse of [3, 4] and we will adapt it for the Logical Information System we use. This part requires an in-depth study of [3, 4] and related work.

Resource editing and discovery: The objective of this task is to propose methods for guided search and editing on web services and data. Sewelis is a tool that supports easy and intuitive search on semantic web data. Semantic web services are semantic web data, thus Sewelis approach is applicable to discover them. However, web services are a particular kind of data, and they are diverse and heterogeneous. Their discovery depends not only on the representation of their characteristics and functionalities at the registration and update phases, but also on the techniques and algorithms used to match them at the retrieval phase. We will ensure that selected services for composition offer the required features.

Workflow language: The orchestration of web services involves a workflow definition language. Many languages are proposed, we want to choose a language that helps to be domain independent. The language should also allow the workflows to be enacted and shared.

Guided composition: We think that, at the architectural level, it is important to separate discovery and composition of web services. For users however the tasks supported by those components must be associated to improve interaction. The insertion of a new service in a workflow must depend on all services already in the workflow. The choice of a service should allow data links and suggestions of composition plans. Contextual information must allow search results and suggestions to be precise. The guided composition will be based on the guided discovery of Sewelis. The automation of tedious tasks and interaction during the process of defining workflow will be supported by resource description. Resource description will be tailored to support reasoning at a reasonable cost. We will adapt the user interface of Sewelis to the workflow edition maintaining its expressiveness and ease-of-use. The workflow will be expressed in the workflow definition language chosen in the previous task. The workflow edition is more difficult when the language is complex. A suitable level of abstraction for the users and simplifications on the patterns of the language will facilitate composition.

Workflows as services: We will adopt a recursive view on services. We will consider primitive services and complex services. A primitive service will be a task component of a workflow and a complex service will be defined as a workflow that can be used as a task component of another workflow. That view will facilitate reuse and composition.

4 Evaluation methodology

GenOuest[4] is a bioinformatic platform that provides a large collection of tools and services for data analysis. The platform offers a suitable environment to evaluate and validate our approach. We will use datasets and real cases of the GenOuest platform for evaluation. We will test our approach with biologist users of GenOuest and make a comparison with approaches of existing systems such as Taverna.

References

1. Alonso, G., Casati, F., Kuno, H., Machiraju, V.: Web Services: Concepts, Architectures and Applications. Springer (2003)
2. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web Technologies. CRC, Boca Raton, FL (2009)
3. Missier, P., Wolstencroft, K., Tanoh, F., Li, P., Bechhofer, S., Belhajjame, K., Pettifer, S., Goble, C.A.: Functional units: Abstractions for web service annotations. In: SERVICES, IEEE Computer Society (2010) 306–313
4. Wolstencroft, K., Alper, P., Hull, D., Wroe, C., Lord, P.W., Stevens, R.D., Goble, C.A.: The myGrid ontology: bioinformatics service discovery. Int. Journal of Bioinformatics Research and Applications **3**(3) (2007) 303–325
5. Wilkinson, M.D., Links, M.: Biomoby: An open source biological web services proposal. Briefings in Bioinformatics **3**(4) (2002) 331–341
6. Romano, P.: Automation of in-silico data analysis processes through workflow management systems. Brief Bioinform **9**(1) (2008) 57–68
7. Wang, Z., Miller, J.A., Kissinger, J.C., Wang, R., Brewer, D., Aurrecoechea, C.: Ws-biozard: A wizard for composing bioinformatics web services. In: SERVICES I, IEEE Computer Society (2008) 437–444
8. Oinn, T., Greenwood, M., Addis, M., Ferris, J., Glover, K., Goble, C., Hull, D., Marvin, D., Li, P., Lord, P.: Taverna: Lessons in creating a workflow environment for the life sciences. Concurrency and Computation: Practice and Experience **18**(10) (2006) 1067–1100
9. Ferré, S., Hermann, A.: Semantic search: Reconciling expressive querying and exploratory search. In Aroyo, L., Welty, C., eds.: Int. Semantic Web Conf. LNCS 7031, Springer (2011) 177–192
10. Lebreton, N., Blanchet, C., Claro, D.B., Chabalier, J., Burgun, A., Dameron, O.: Verification of parameters semantic compatibility for semi-automatic web service composition: a generic case study. In Taniar, D., Pardede, E., Nguyen, H.Q., Rahayu, J.W., Khalil, I., eds.: Int. Conf. on Information Integration and Web Based Applications and Services, ACM (2010) 845–848

[4] http://www.genouest.org/