Semantically Linking Web Pages to Web Services in Bioinformatics

Karen Sutherland, Kenneth M^{c}Leod and Albert Burger

Contact: kjs1@macs.hw.ac.uk

Abstract:
A key application area of semantic technologies is the fast-developing field of bioinformatics. Sealife is a project within this field with the aim of creating semantics-based web browsing capabilities for the life sciences. This includes meaningfully linking significant terms from the text of a web page to executable web services. This requires the semantic mark-up of biological terms, linking them to biomedical ontologies, then discovering and executing services based on terms that interest the user. This paper deals with the process of discovery and execution of web services, providing an overview of our experience in using existing Semantic Web technology to achieve better integration of the current web with the rapidly growing e-Science web service infrastructure.

1 Introduction

One of the primary motivations for the Semantic Web is the promise of automating tasks that previously required human user intervention. Within the bioinformatics community, biologists have access to a wide range of web-based tools, but tend to use only those with which they are most familiar [7]. As no one tool can provide all the data needed, biologists must use several. Often they manually copy the output from one tool to be input for the second. This is repetitive, time-consuming and error-prone. It therefore makes sense to use Semantic Web technology to help biologists to discover and execute services automatically.

Much research focuses on the automation of workflow execution in bioinformatics, but does not link this to information that is browsed in web pages. For example, myExperiment[1] is essentially a repository of workflows that have been created within the Taverna Workbench[2]. Taverna itself allows workflows to be both created and executed, but is not inherently linked to pages a user may browse. Other projects, e.g. COHSE[3], focus on the semantic links between web pages, but do not take this further to link them to web services.

This paper considers the potential benefits of applying Semantic Web technologies to the life sciences domain, with the aim of linking web pages to the corresponding web service

[1] http://www.myexperiment.org/
[2] http://taverna.sourceforge.net/
[3] http://cohse.cs.manchester.ac.uk/

infrastructure. The wider context of the EU Sealife [6] project is introduced, however the paper focuses on Sealife's underlying functional model, the *Task Composition Manager* (TCM).

The paper is structured as follows: Section 2 describes the aims of the Sealife project and provides a use case; Section 3 discusses the implementation of workflow discovery and workflow enactment; future work is outlined in Section 4; and the paper is concluded in Section 5.

2 Sealife Overview and Use Case

As mentioned, Sealife is a project researching the possibility of linking web pages to web services through Semantic Web technologies. Figure 1 shows the main components of this interaction.

In Figure 1 a user browses web pages in which biological terms have been identified and annotated with semantic hyperlinks. This is achieved through the exploitation of text-mining techniques and a series of pre-existing ontologies, e.g. the Gene Ontology[4]. When the user selects a link on the page, the Sealife server adds the associated term (with corresponding ontology ID) to the user's personal 'shopping cart' (CART). The TCM then dynamically discovers services that require inputs of the same semantic type as those in the CART. For example, a DNA sequence comparison web service will be found if a DNA sequence is placed into the CART. The user can choose to execute any of these services through the TCM, and the result is presented through the browser. An example of this process is given in the following use case.

The use case is based upon the scenario of a researcher using the mouse embryo as a model for early human development. The researcher browses some literature that has been semantically marked up by Sealife, and clicks on some of the terms related to *cardiac muscle development*. This action adds these terms to the CART. When requested, the TCM suggests some services the researcher may wish to execute with these terms. One option is a workflow that offers to discover which human genes could be involved in the early development of cardiac muscle tissue.

In order to answer this query the TCM initially maps the human tissue onto the equivalent mouse tissue using a cross-species anatomy database, XSPAN[5]. A gene expression database, GXD[6], is then used to retrieve the mouse genes which are expressed in this structure. Once the mouse genes have been found, BLAST[7] can be used to find homologous human genes.

Manual execution of this sequence of events is time-consuming and repetitive. In contrast the automated process can produce results both quickly and efficiently.

[4] http://www.geneontology.org
[5] http://www.xspan.org
[6] http://www.informatics.jax.org
[7] http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

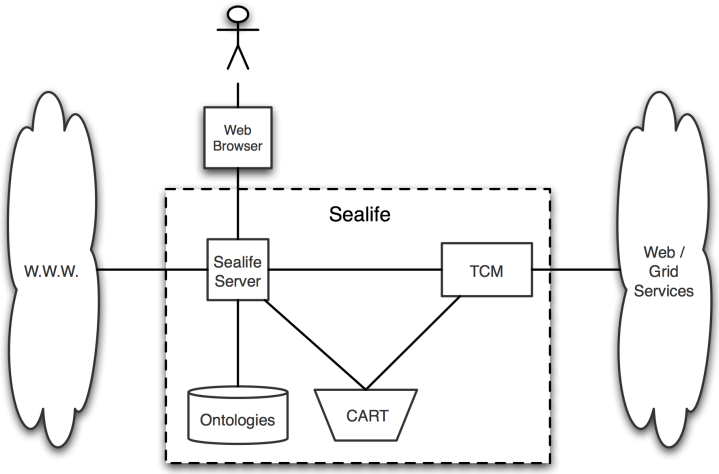

Abbildung 1: Sealife Architecture.

3 Implementation of the TCM

This section describes how the TCM works, including an outline of the work that needed to be carried out before implementation of the use case. It briefly details the technicalities of service/workflow discovery and implementation. It should be noted that this work has been implemented as a functional prototype.

3.1 Semantic Matching Prerequisites

In order to enable semantic linking between web pages and web services, some groundwork needs to be done. Firstly, the HTML of the web page is parsed on-demand and a text-miner identifies key biological terms using a range of ontologies. These terms then appear as clickable links, allowing the user to add items to the CART. This work is undertaken by a separate partner in the Sealife project.

In order to find services that semantically match the terms in the cart, it is necessary to assign a semantic 'type' to each term, e.g. 'cardiac muscle' can be classified as a 'tissue'. This type can then be used by the Service Discovery component of the TCM (which uses myGrid's Feta [5]) to locate services with an input that matches semantically. Feta only recognises semantic types that have been declared in the myGrid Domain Ontology, therefore this ontology has been extended to include all significant terms in our use cases.

Furthermore, it is necessary to semantically describe available web services in order to allow Feta to search through them. These *service descriptions* can also describe pre-canned workflows, which contain a series of services linked together to create a virtual experiment.

The service descriptions are written in XML and contain natural language descriptions of the service (or workflow) operation(s), and input/output parameters.

With the above requirements in place, it is possible to semantically search for services.

3.2 Service/Workflow Discovery

The task of service discovery is primarily undertaken by Feta. Feta searches the service descriptions looking for those that satisfy a given query. Queries are written in SeRQL[8].The TCM uses one pre-canned parameterised query to discover services that semantically match one or more of the CART items (i.e. the service has an input parameter with a semantic type identical to that of a CART item). Feta executes an instance of this query, building an RDF graph containing the answer(s).

The TCM parses the graph, obtaining a list of appropriate services. At this stage, the TCM uses other queries to gather further knowledge of the service(s), e.g. details of their input and output parameters. All of the information obtained from Feta is returned to the user.

3.3 Service/Workflow Enactment

The user is presented with a range of services/workflows that can be executed based on some of the items in their CART. If they choose to run a service/workflow, they must also select values for the input parameter(s). All of this information is then given to the TCM.

The TCM manages the process of service enactment using myGrid's workflow enactor. This enactor uses a workflow enactment language called *Scufl*[9]. A Scufl file contains information describing the location of a particular web service and, if necessary, how it should be linked to others. Frequently the output of one service does not match the input for the subsequent service and in this case a *shim* [3] is used to convert the output.

Once the workflow has been executed, the TCM parses the output and returns it to the browser.

4 Future Work

So far the basic features of the TCM have been described, in which the TCM uses items added to a user's shopping cart to both discover and execute services. In addition to this functionality, work is ongoing into enhancing the underlying sequence of events to give the user more help in finding services/workflows, and in analysing their output. This work utilises automated planning and argumentation.

[8] http://www.openrdf.org/doc/sesame/users/ch06.html
[9] http://www.gridworkflow.org/snips/gridworkflow/space/XScufl

The use of automated planning, in a modified form, can break down large goals into smaller steps executable by a workflow enactment engine. Hierarchical Task Network (HTN) planning [2] is particularly suitable for this task. The basic components of workflows (operations) are kept within a domain model. The intention is to query this domain model with the planner and translate the output into a Scufl workflow for execution.

Argumentation [1] is the use of computers in the process of arguing, either for helping humans to argue or by using computers to conduct the argument. The argumentation engine ASPIC[10] is currently being used to generate arguments on the validity of information provided by GXD (the gene expression database appearing in the use case) [4].

5 Discussion/Conclusion

The aim of this work was to provide a module within Sealife that would enable the semantic linking of web pages to web services. In order to achieve this goal, tools originating from the myGrid project were used, connected together and added to novel ideas and code. The outcome was a software module (the TCM) that semantically discovers web service operations based on items the user places in a shopping cart. Currently, we have a full implementation of this module. The next steps are to integrate this with the browser and text mining module of Sealife, at which point we will be able to carry out a full end user evaluation.

However, what has already become clear at this stage is that by deploying existing semantic web technologies, as well as AI planning and argumentation, it is possible to create novel functionality for the end user. More specifically, our prototype shows that it is possible to integrate web pages with web services in a dynamic approach that exploits the semantics of the domain, and which requires little technical understanding of the computational elements by the end user.

It has become obvious that significant efforts are necessary to capture all the relevant semantics of the domain. This not only includes the domain ontologies, such as mouse and human anatomy for our gene expression use case, but also ontologies to semantically describe services, capture the domain model used by the AI planning module, and represent the rules that govern the argumentation mechanism across resources. Not only are these required in isolation, they must also be consistent between each other. For example, the concept of an anatomical structure in the mouse anatomy ontology needs to be consistent with the same concept in the service descriptions, the planning domain model and the argumentation representations.

In this work terms were marked up according to the myGrid domain ontology. However, this ontology is too small to fully represent all aspects of the life sciences, but significant extension could make it overly complex and cumbersome. The limitations of the ontology also mean that there are gaps in the semantic mark-up of available services as all input and output parameters cannot be associated with a semantic type. This in turn means Feta can

[10] http://www.argumentation.org

only discover services with certain parameters that are marked up with this ontology.

Overall the process of creating service descriptions, identifying suitable semantic types for the parameters, and extending the ontology requires much effort and is prone to human error. Developing this framework on a larger scale is extremely demanding as consistent semantic mark-up of web services and web pages is almost non-existent, even just within the bioinformatics domain.

The idea of using semantic technology to automatically link the current web to the increasing range of web services is appealing. However it is unfortunately also premature. Much work needs to be performed to create a series of sophisticated, cross-linked ontologies so that terms in different resources can be mapped to each other. Within the domain of mammalian anatomy a reference ontology has been proposed which potentially can fulfil this function (i.e. linking more specific ontologies). This is known as CARO (Common Anatomy Reference Ontology)[11].

Until such ontologies become mature, and there is widespread co-operation to enable the dreams of the semantic web, works such as this one continue to present the expanding range of possibilities that could one day become reality.

Acknowledgements Funding for this work has been provided by the EU project Sealife (FP6-2006-IST-027269).

Literatur

[1] Carbogim, D.V., Robertson, D., Lee, J.: Argument-based applications to knowledge engineering. Knowledge Engineering Review 15(2), 119–149 (2000)

[2] Erol, K., Hendler, J., Nau, D.S.: UMCP: A Sound and Complete Procedure for Hierarchical Task Network Planning. In: 2nd International Conference on Artificial Intelligence Planning Systems, pp. 249-254 (1994)

[3] Hull, D., Stevens, R., Lord, P., Wroe, C. Goble, C.: Treating “Shimantic Web” Syndrome with Ontologies. In: 1st AKT workshop on Semantic Web Services (2004)

[4] McLeod, K., Burger, A.: Using argumentation to tackle inconsistency and incompleteness in online distributed life science resources. In: IADIS, International Conference Applied Computing, IADIS Press, 489–492 (2007)

[5] Wolstencroft, K., Alper, P., Hull, D., Wroe, C., Lord, P., Stevens, R., Goble,C.: The mygrid ontology: Bioinformatics service discovery. International Journal of Bioinformatics Research and Applications, 3(3), 303–325 (2007)

[6] Schroeder, M., Burger, A., Kostkova, P., Stevens, R., Habermann, B., Dieng-Kuntz, R.: Sealife: A Semantic Grid Browser for the Life Sciences. In: HealthGrid. IOS Press (2006)

[7] Stevens, R., Goble, C., Baker, P., Brass, A.: A classification of tasks in bioinformatics. Bioinformatics, 17(2), 180-188 (2001).

[11] http://www.bioontology.org/wiki/index.php/CARO:Main_Page