Describing Web Services for user-oriented retrieval

Duncan Hull, Robert Stevens, and Phillip Lord

School of Computer Science, University of Manchester, Oxford Road, Manchester,
UK. M13 9PL

Abstract. As of 2005, there are over 1 000 publicly available Web Services in the Life Sciences community and this number continues to grow. Describing them so that users can retrieve the service required to perform a given task is now a key problem in managing these services. This paper outlines some of the requirements and challenges for semantically describing Web Services for user-oriented retrieval in the Life Sciences domain. We briefly discuss the OWL-S, WSMO and ^{my}Grid ontologies and their suitability to tackling this problem.

1 Introduction

In *e-science*, scientists use computational techniques to perform experiments, by generating and testing hypotheses using software tools, databases and the Web. For example, life scientists working in biology and medicine can investigate properties of genes, sequences of DNA, by querying over 700 [1] separate and publicly available databases. One of these databases, GenBank, has experienced rapid and sustained growth and this pattern is typical of life science data.

Consequently, informatics or information science has become an integral part of experimental biology. The combination of informatics and biology, known as bioinformatics [7], relies on techniques for describing and performing experiments that retrieve data from distributed resources. Such "dry" laboratory techniques, or *in silico* experiments, are an important tool for biologists because they can validate and motivate experiments carried out in a conventional "wet" laboratory, *in vitro* or *in vivo*.

Currently, bioinformaticians often use *ad hoc* scripts and programs to perform these experiments because the service providers expose many different programmatic interfaces to their resources. Where programmatic interfaces are not provided, scientists are forced to "screen-scrape" or "cut-and-paste" data between web-based forms. Screen scraping is notoriously fragile [10] and cut-and-paste experiments can be laborious, or impossible, to repeat. This unfortunate situation could change as the Web Services protocol stack becomes more widely understood and adopted, and an increasing number of database and tool providers in biology recognise the benefits of providing Web Service interfaces to their resources. As of 2005, the ^{my}Grid project has a registry of over 1 000 of these publicly accessible Web Services [8], provided by a wide range of third parties.

The number of available services will continue growing as long as the impetus behind the technology grows [3].

In the life sciences, we now have, or are quickly becoming, a "bioinformatics nation" [10]. In this nation, previously fragmented and rival city-states provide standard service interfaces to their resources which unites them. Performing *in silico* experiments is now potentially quicker and easier than the myriad of techniques used to integrate these resources in the past. One way of formalising and executing these *in silico* experiments is to pipe together inputs and outputs of consecutive Web Services in a workflow environment and this is the approach taken by several projects in the life sciences [11, 2]. These efforts are a progression towards more fully exploiting the wealth of biological data on the web.

Despite having over 1 000 web services, there is currently no globally agreed model for describing the structure, type and semantics of data passed between services. For example when SOAP is used in the messaging layer, the body of the message typically contains weakly or implicitly typed data, with many complex legacy and flat-file formats described simply and naively as `xsd:string`. Forcing every Web Service to pass messages validated against a global type system, such as XML schema or RDF schema, is desirable because the messages are simpler and safer to process by the services that consume them. This approach is taken by the MOBY-S [13] and S-MOBY projects [9]. However, such an approach provides a barrier to service registration which some service providers may not cross due to their reluctance to spend time and resources describing inputs and outputs conforming to some pre-agreed global data model.

The lack of a globally accepted data model has two important consequences. Firstly, to help scientists find services they need, semantic annotations, over and above the WSDL descriptions, are required. Secondly, users often need to join two services together that have *closely related* inputs and outputs but require some kind of alignment to mediate between them. Because these relationships are not explicitly stated, the scientist constructing the workflow has to use their knowledge to align services in the way they require. Currently these alignments are performed with "shim" services [5] which are a subclass of mediators. Like mediators, shims create information for a higher level of application [12], in bioinformatics they are an unavoidable feature of the experimental process. The problem of shim services is not unique to bioinformatics, scientists using workflows of Web Services to predict earthquakes have experienced similar problems [6], so it seems likely that these problems are general rather than specific to life sciences.

This paper will discuss some of the challenges to developing ontologies of Web Services generally. Section 2 describes scenario which illustrates some requirements for semantically describing Web Services. Section 3 looks at ways of querying descriptions to allow users to discover services. Finally we draw some conclusions in section 4.

2 Describing Services: A motivating scenario

A scenario which illustrates some of the requirements of describing services is shown in Figure 1. In this scenario, a user wishes to join a service which searches a database of DNA sequences, such as GenBank into a service that consumes protein sequences, such as BLASTp. This scenario captures some, but not all of the requirements of semantically describing Web Services, more scenarios are outlined in the Knowledge Web deliverables [4].

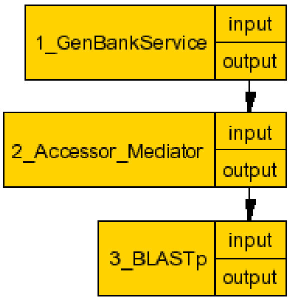

Fig. 1. A mediation scenario in bioinformatics: A `GenBankService` (1) producing a GenBank record needs to be plugged into a `BLASTp` service (3), which accepts a protein sequence. Getting services (1) and (3) to interoperate requires an Accessor_mediator service (2), to extract the protein sequence from the GenBank record.

The inputs and outputs of service 1 and 3 are closely related but can not be joined together directly without alignment and mediation. The GenBankService produces an output of semantic type `GenBank record` and the BLASTp service accepts input of `protein sequence`. Using semantic descriptions of the services, it is possible to reason that a `GenBank record` *hasPart* `protein sequence` and then, using these ontology terms, retrieve a service from that extracts the protein sequence from the GenBank record. These terms need to allow a user, or possibly an agent, to recognise the following:

- Services 1 and 3 are *closely related*
- The "gap" between 1 and 3, can be aligned with service 2
- Service 2 is "experimentally neutral" to insert, and will not adversely affect the outcome of the workflow

Additionally, the ontology used to describe these services needs to be simple enough for domain experts to use. It needs to scale to annotate hundreds of services in the registry, and the subset of these services that align and mediate between them, in the worse case the number of shims is quadratic. In practice, the shims or mediators will be far fewer than as we only need to describe the services that bridge close relationships and common paths.

There are several ontologies for describing Web Services. The OWL-S and WSMO ontologies are both complex and highly expressive. However, in our experience, they are not yet practical for manually describing large numbers of services and contain many features that are not immediately required by the life sciences community. For example, many bioinformatics services are *stateless* - their primary purpose is to retrieve information, not change its state. Consequently, these services do not have pre and post conditions. The approach taken to describing services in ^{my}Grid is much more lightweight; Web Services are annotated with terms from an OWL ontology [14], which is then serialised as RDF. Although these descriptions have limited expressivity, they are sufficient for describing the services outlined in Figure 1 and their simplicity makes them more practical for describing large numbers of services.

3 Discovering Web Services

Once services have been described in RDF and stored in a registry, it is possible to retrieve services using RDQL queries, which is the approach taken in the Feta [8] component of ^{my}Grid. In the case of our scenario outlined in Figure 1, the service required (service 2) could be retrieved by the following query:

$$acceptsInput < GenBankRecord > producesOutput < Protein\_sequence >$$

This query can be answered using RDFS entailment over the named ontology classes, and this is sufficient for the immediate needs of service discovery in bioinformatics. It remains to be seen wether this approach is:

1. Scalable as the number of services increases
2. Usable by annotators describing services
3. Expressive enough to meet the needs of end-users performing *in silico* experiments

However, the design of the ^{my}Grid Feta architecture allows the use of OWL-DL in place of RDQL, if this is required in the future.

4 Conclusions

Annotating services with terms from an ontology can make up for the lack of a globally accepted model or type system for life science data. In our experience, the OWL-S and WSMO ontologies are too heavyweight to be practically applicable to the needs of bioinformatics which requires, simple ontologies to describe large number of stateless services and the mediators or shims that join them.

4.1 Acknowledgements

The authors would like to acknowledge Jeff Pan for his comments on drafts of this paper. This work is supported by the UK e-Science programme, EPSRC grant reference GR/R67743/01.

References

1. M. Y. Galperin. The molecular biology database collection: 2005 update. *Nucleic Acids Research*, 33:D5–D24, 2005. Database issue.
2. H. T. Gao, J. Hayes, and H. Cai;. Integrating biological research through web services. *IEEE Computer*, 38(3):26–31, March 2005.
3. D. Greenbaum, A. Smith, and M. Gerstein. Editorial: Impediments to database interoperation: legal issues and security concerns. *Nucleic Acids Research*, 33:D3–D4, 2005. doi:10.1093/nar/gki134.
4. D. Hull, M. Karemba, and J. Pan. Use Cases for Semantic Web Services from Bioinformatics, 2005. Knowledge Web deliverable: Workpackage 2.4 Semantic Web Services: In press.
5. D. Hull, R. Stevens, P. Lord, C. Wroe, and C. Goble. Treating shimantic web syndrome with ontologies. In *First AKT workshop on Semantic Web Services (AKT-SWS04) KMi, The Open University, Milton Keynes, UK. December 8, 2004*, 2004. Workshop proceedings CEUR-WS.org ISSN:1613-0073.
6. J. Kim, Y. Gil, and M. Spraragen. A knowledge-based approach to interactive workflow composition. In *In Workshop: Planning and Scheduling for Web and Grid Services, at the 14th International Conference on Automatic Planning and Scheduling (ICAPS 04); Whistler, Candada*, 2004.
7. A. M. Lesk. *Introduction to Bioinformatics.* Oxford University Press, 2002. ISBN 0199251967.
8. P. Lord, P. Alper, C. Wroe, and C. Goble. Feta: A light-wieght architecture for user oriented semantic service discovery. In *European Semantic Web Conference*, 2005. ESWC 2005 :: 2nd European Semantic Web Conference 2005, Heraklion, Greece 29th May to 1st June 2005.
9. P. Lord, S. Bechhofer, M. D. Wilkinson, G. Schiltz, D. Gessler, D. Hull, C. Goble, and L. Stein. Applying semantic web services to bioinformatics: Experiences gained, lessons learnt. 2004. Proceedings of the 3rd International Semantic Web Conference, Hiroshima, Japan.
10. L. Stein. Creating a bioinformatics nation. *Nature*, (417):119–120, May 2002.
11. R. D. Stevens, H. J. Tipney, C. Wroe, T. Oinn, M. Senger, P. W. Lord, C. A. Goble, A. Brass, and M. Tassabehji. Exploring Williams-Beuren Syndrome Using myGrid. In *Intelligent Systems for Molecular Biology, Glasgow, UK.*, volume 20, 2004. ISSN 1367-4803.
12. G. Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, 1992.
13. M. Wilkinson and M. Links. BioMOBY: An Open Source Biological Web Services proposal. *Briefings in Bioinformatics*, 3(4):331–341, 2002.
14. C. Wroe, R. Stevens, C. Goble, A. Roberts, and M. Greenwood. A Suite of DAML+OIL Ontologies to Describe Bioinformatics Web Services and Data. *International Journal of Cooperative Information Systems*, 12(4):197–224, June 2003.