Semantic Web Applications to E-Science *in silico* Experiments

Jun Zhao
University of Manchester
Kilburn Building, Oxford Road
Manchester, United Kingdom
zhaoj@cs.man.ac.uk

Carole Goble
University of Manchester
Kilburn Building, Oxford Road
Manchester, United Kingdom
carole@cs.man.ac.uk

Robert Stevens
University of Manchester
Kilburn Building, Oxford Road
Manchester, United Kingdom
robert.stevens@cs.man.ac.uk

ABSTRACT

This paper explains our research and implementations of manual, automatic and deep annotations of provenance logs for e-Science *in silico* experiments. Compared to annotating general Web documents, annotations for scientific data require more sophisticated professional knowledge to recognize concepts from documents, and more complex text extraction and mapping mechanisms. A simple automatic annotation approach based on "lexicons" and a deep annotation implemented by semantically populating, translating and annotating provenance logs are introduced in this paper. We used COHSE (Conceptual Open Hypermedia Services Environment) to annotate and browse provenance logs from myGrid [1] project, which are conceptually linked together as a hypertext Web of provenance logs and experiment resources, based on the associated conceptual metadata and reasoning over these metadata.

Categories and Subject Descriptors

E.5 [**Data**]: Files; H.2.8 [**Information Systems**]: Database Management—*Database Applications*; H.5.4 [**Information Systems**]: Hypertext/Hypermedia

General Terms

Experimentation, Human Factors

Keywords

provenance, integration, annotation, ontology, Semantic Web, e-Science

1. INTRODUCTION

e-Science *in silico* experiments complement traditional lab work by using computer-based information repositories and computational analysis to test a hypothesis, search for patterns, or demonstrate a known fact. myGrid, as a pilot e-Science research project in the U.K., intends to build a personalized workbench for biologists and bioinformaticians to enact *in silico* experiments, and to organize variety of experiment resources (like data, services, experiment designs, conclusions, etc). Data results from *in silico* experiments are of reduced value without provenance, which records the lineage of data, experiments, and the context information of experiments. Provenance is helpful to evaluate a scientific hypothesis, to share results and avoid duplication of work, to verify the credit and accountability of scientific discoveries and to trace the ownership and version of data.

Complex interlinking information reside in collections of provenance logs. In practice, users expect higher level linkings between provenance logs to support drawing experiment conclusions. The Semantic Web, as an extension to the current Web, greatly succeeds in defining and linking documents or resources based on the associated conceptual metadata for more effective knowledge discovery, integration and cooperation across computers and people. By annotating provenance logs with a suite of commonly shared bioinformatics and other generic ontologies [2], we aim to *conceptually* link provenance logs and other experiment resources together, to give a higher level of view over these documents; and provide a friendly interface for querying by navigation through these documents, supported by COHSE [1].

2. COHSE

COHSE integrates three technologies: **an Ontology Service**, which uses rich knowledge representation techniques and reasoning abilities to produce a machine processable semantics to the conceptual metadata associated with documents, and the relationships between these concepts; **an Annotation Service**, which annotates documents or sections of documents with concepts from the ontology and maintaining the mappings from concepts to documents; and **a Link Service**, which generates target links for the concepts associated with Web documents and displays links in a hierarchy supported by the reasoning of the Ontology Service. COHSE supports both **proxy-based** and **browser-based** annotation approaches. In this project, we adopted the browser-based approach for annotation and conceptual linking.

Annotations are stored and managed as independent objects in COHSE. A central link-base, implemented in a mySQL database, holds all the annotations and provides an HTTP/CGI interface for inserting and querying the annotations. When COHSE loads a new Web page, the Link Service retrieves all the words that appear in the page. By identifying all the concepts in the page through the lexicon and the Annotation Service, the Link Service in COHSE contacts the Annotation Service to identify target links to other URLs with corresponding concepts as inferred by the Ontology Service. Consequently, we are able to link Web pages together based on the concepts associated with their contents. By changing the ontology, we can get different link "views" over those resources.

[1] http://www.mygrid.org.uk

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

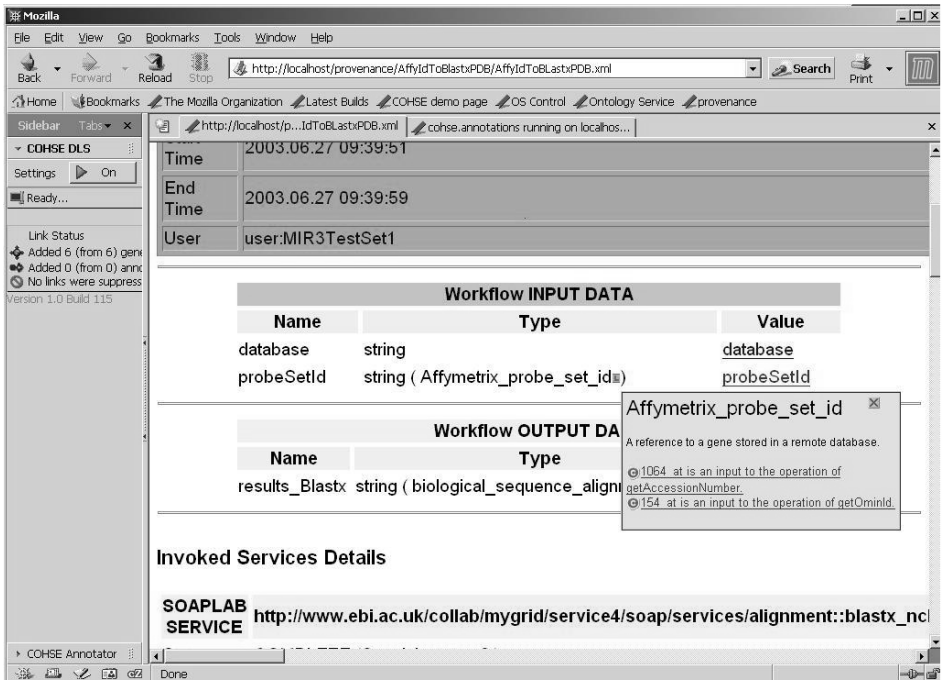

Figure 1: Linked Logs through Semantic Annotations.

3. ANNOTATING PROVENANCE LOGS

When a biologist, whose research interest is human genetics, plans to run experiments in the myGrid workbench, in order to find new genes in the critical region that leads to a DiseaseX, he wants to navigate linked provenance logs from repeated experiment runs in order to compare and verify the results, and test his hypothesis. Also he hopes to discover supporting materials, like research literature, experience materials from his colleagues, etc.

By a manual process annotation authors can recognize concepts for scientific data and services from documents and manually annotate the corresponding concepts with these objects. Thus provenance logs are conceptually linked together based on the associated semantic metadata. But the process of manual annotation is time-consuming and error-prone, which motives our research in automatic and deep annotation.

In the process of automatic annotations, language terms are recognized by the control of DOM (Document Object Model) objects in documents and unsupervisedly mapped to lexicons in an ontology. So the home page of the biologist's colleague (a human geneticist) can be linked to the provenance log run by him, based on the underlying semantic concepts in respective document and their relationships that "Human Geneticist" is somebody who *studies* "Human Genetics". But this automatic annotation currently is not intelligent enough for complex term mappings and can easily result in redundant linkings between documents.

In myGrid project, for each experiment run, there is a corresponding file which provides semantic information about the data and services invoked in the experiment. "Deep annotation" is a process exploiting the semantic information from these files, for each data and service object in the experiments, based on the unique identifiers of these objects which are applied in both information resources. Thus deep annotation saves users from recognizing and annotating documents by hand, and applies more professional concepts, like bioinformatics terms, to complex scientific data. As shown in Figure 1, provenance log of an experiment run by the biologist which uses a *Affymetrix_probe_set_id* (a semantic concept) as input is semantically linked with other provenance logs which use *Affymetrix_probe_set_id* as either input, output or parameter for services invoked in an experiment. Links to these logs are provided to the biologist in the pop-up window (Figure 1) nearby the annotated concept in the log. So the user can browse these conceptually linked logs by just pointing and clicking.

Our semantically linked provenance logs successfully provide a higher level view between these logs and experiment resources. Currently, this deep annotation design is based on a single bioinformatics ontology. myGrid is working on building an annotation template which supports annotating with multiple ontologies and expects to improve the provenance model with richer semantic ability.

4. ACKNOWLEDGMENTS

myGrid project, grant number GR/R67743, is funded under the UK e-Science programme by the EPSRC. The authors would like to acknowledge the other members of the myGrid team for their contributions; Sean Bechhofer who developed the COHSE system; Yeliz Yeslida for her help in getting COHSE up and running. We also thank Mark Greenwood and Chris Wroe for their help during our annotation implemetation process.

5. REFERENCES

[1] S. Bechhofer, L. Carr, C. Goble, and W. Hall. Conceptual Open Hypermedia = The Semantic Web? In *SemWeb2001, The Second International Workshop on the Semantic Web*, Hong Kong, May 2001.

[2] C. Wroe, R. Stevens, C. Goble, and M. Greenwood. A suite of DAML+OIL Ontologies to Describe Bioinformatics Web Services and Data. *International Journal of Cooperative Information Systems*, 12(2):197–224, 2003.