Semantically Linking and Browsing Provenance Logs for E-science

Jun Zhao, Carole Goble, Robert Stevens, and Sean Bechhofer

Department of Computer Science
University of Manchester
Oxford Road
Manchester
United Kingdom M13 9PL
{zhaoj,carole,robert.stevens,seanb}@cs.man.ac.uk

Abstract. e-Science experiments are those performed using computer-based resources such as database searches, simulations or other applications. Like their laboratory based counterparts, the data associated with an e-Science experiment are of reduced value if other scientists are not able to identify the origin, or provenance, of those data. Provenance is the term given to metadata about experiment processes, the derivation paths of data, and the sources and quality of experimental components, which includes the scientists themselves, related literature, etc. Consequently provenance metadata are valuable resources for e-Scientists to repeat experiments, track versions of data and experiment runs, verify experiment results, and as a source of experimental insight. One specific kind of *in silico* experiment is a workflow. In this paper we describe how we can assemble a Semantic Web of workflow provenance logs that allows a bioinformatician to browse and navigate between experimental components by generating hyperlinks based on semantic annotations associated with them. By associating well-formalized semantics with workflow logs we take a step towards integration of process provenance information and improved knowledge discovery.

1 Introduction

e-Science refers to science by scientists working collaboratively in large distributed project teams in order to solve scientific problems [1] that uses electronic resources (instruments, sensors, databases, computational methods, computers). e-Science *in silico* experiments complement traditional lab work through the use of computer-based information repositories and computational analysis to: test a hypothesis; derive a summary; search for patterns; or demonstrate a known fact [2]. Bioinformaticians, for example, perform analyses (*in silico* experiments) by submitting data (often biological sequences such as DNA or protein) to a succession of analysis tools and databases [3]. Currently when performing *in silico* experiments, bioinformaticians spend a great deal of time manually and repeatedly coordinating tools or applications to produce results.

M. Bouzeghoub et al. (Eds.): ICSNW 2004, LNCS 3226, pp. 158–176, 2004.

A new software infrastructure, to form a virtual organization for sets of heterogeneous distributed data collections, computing resources and people and machines for e-Science is needed. The Grid is such a proposed infrastructure [4]. myGrid [1] is a project, aiming to provide middleware infrastructure for the Grid to create and orchestrate *in silico* experiments for bioinformaticans working in functional genomics, and manage and exploit the results from these experiments. myGrid builds sets of middleware services to automate the experiment process as *workflows* [3].

Like their laboratory based counterparts, the data associated with an e-Science experiments are of reduced value if other scientists are not able to identify the origin, or provenance, of those data. Provenance is the term given to metadata about experiment processes, the derivation paths of data, and the sources and quality of experimental components, which includes the scientists themselves, related literature, etc. For an *in silico* experiment, knowledge to be shared amongst scientists includes:

- data results and the origin of the data results: Data are of reduced value without provenance information, which are valuable resources for results verification.
- a log of all the computational processes used: This recording is analogous to the conventional recording of materials and methods in a scientist's log-book at the laboratory bench. It promotes the sharing and re-use of experimental knowledge as well as good scientific practise.

Provenance also provides "recipes" for workflow designs by enabling tracing details of workflow executions and versions of data and services. When performing *in silico* experiments by hand, such provenance recordings are often ignored. These make experiment reproduction and experiment design difficult for novice bioinformaticians. Moreover, public bioinformatics databases are frequently updated with newly discovered data resources, which may lead to different results when repeating workflow runs. Provenance information is needed to explain and analyze the impact of changes.

1.1 A Scenario

In a typical *in silico* experiment, a biologist has been using a microarray study to investigate the differences in levels of gene expression between individuals with and without the disease. The up-regulated genes[2] are identified by keys to a proprietary database. The bioinformatician needs to map these proprietary identifiers (e.g. Affymetrix_probe_set_id) onto those used in other bioinformatics databases (e.g. GenBank_id). These database resources are then used to retrieve additional data recorded about the gene, and it products. The bioinformatician will take the DNA or protein sequence from these databases and submit them to other tools, collecting further information. Most of these analyses are performed

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

[2] Genes that have been switched on

using a separate service, and generates their own collections of results. In turn, many of these are fed into other services. In myGrid such an analysis is captured as a workflow and a collection of tools collect and co-ordinate the data generated by that workflow.

Such an experiment may be performed many times, using different datasets, in different projects or investigated with differing topics. At the end, a large repository of records about experiments will be collected. This information can be viewed from four different levels:

Organization level, which records the workflow user and creator, their organization, project, the hypothesis for this experiment/project, the experiment design, etc. For example, Dr Stevens works for the University of Manchester; he designed the workflow W21 whose purpose is to characterise an Affymetrix_probe_set_id with its predicted genomic function; this workflow is part of a greater experiment to identify the genes associated with Grave's Disease;

Process level, that collects how, when and where the workflow is run, what data are used and generated, which computational services are invoked, and the input and output data for service invocation. For example, BLAST version 3.1 run at the NCBI over SWISS-PROT version 41 on 04/05/2004 at 13:34 GMT was invoked with gene sequence identified by 1020_s_at, and successfully executed in 2.1 seconds;

Data level, most of which are inferred from the process level provenance, and describes derivation path of the data results from services. For example, a collection of pairwise sequence alignments were generated by BLAST, version 3.1, run at the NCBI over SWISS-PROT version 41, etc for a gene sequence 1020_s_at;

Knowledge level, mostly in the form of annotations, either in free text or in a structured/semi-structured form. For example, the data level view above, has the knowledge level view that the result set of *protein sequences* are *similar to* the input *gene sequence* 1020_s_at.

When a biologist reviews his or her work, these data can be used like a lab-book. Typical questions to such a resource would be:

- Which experiments used a particular workflow?
- Which experiments used gene *x*?
- Which experiments were performed upon Grave's disease and which upon Williams-Beuren syndrome?
- Which experiments investigated signal peptides in T-lymphocytes?

These questions are skewed towards descriptions of *what* the experiments are; *when*, *where* and *how* they are enacted; and *why* and for *whom* experiments were performed. The questions are about the analytical process. In a large collection of such data, simple keyword searches will not support such queries - we need knowledge that sequence 1020_s_at turns out to be *T-lymphocyte*, for example. Consequently, we need to include knowledge level provenance over these process provenance data in order to answer such questions.

An early version of the myGrid middleware generated process provenance logs as XML documents. Our ambition was to generate a web of these documents, together with related experimental components such as the home page of the scientist or the XML document describing the workflow script and so on, that could be navigated by the scientists. Biologists are familiar with the notion of *query by navigation*. Their standard way of interconnecting data sources is through web browsers and point and click navigation [5]. We emulate this by providing a browser-based navigation experience through a provenance document collection. We present the experimental holdings we wish to browse - logs, users' home pages, workflow scripts, data results, publications, notes - as a document set. We then dynamically generate a hypertext of these documents so that scientists navigate through a web of experiment data that connects enactment logs with an experiment's documents.

To do this we need a link generation mechanism. In particular we investigated the generation of links by using shared concepts associated with a document. By associating documents with domain concepts drawn from an ontology, logs can be dynamically linked together based on their shared semantic concepts. For example, even though two provenance logs have inputs with different values, they can be linked together based on the fact that both inputs are *Gene_ids*, the same semantic concept. This provides a higher level linking between these logs than just linking them based on the same data value recorded in the logs. In effect the ontology is acting as a link model for the provenance document collection.

As shown in Fig 1, the result is a dynamically generated Semantic Web of provenance, linking together a process provenance log with other logs, related workflow design templates, experiment notes, relevant literature, the scientists' home pages, as well as some organisational information about the overall experiment, through their shared semantics.

The rest of the paper is organised as follows. A brief description of the relevant myGrid components and the generation of process provenance, without semantic linking, is described in Sect 2. Section 3 describes how we build our semantically linked Web of experimental documents. Section 4 describes how such a Semantic Web is populated. Some related works are introduced in Sect 5. We close in Sect 6 with a discussion of our experiences and some pointers of our move to an Resource Description Framework (RDF) [6] based provenance model.

2 Provenance in myGrid

The myGrid middleware framework employs a service-based architecture, built on Web Services [7]. The platform is being used for biological investigations into gene expression and Single Nucleotide Polymorphism (SNP) analysis for Grave's disease [2] and Williams-Beuren Syndrome [3]. In addition to a workbench for creating and executing experiments, the primary services to support *in silico* experiments fall into four categories:

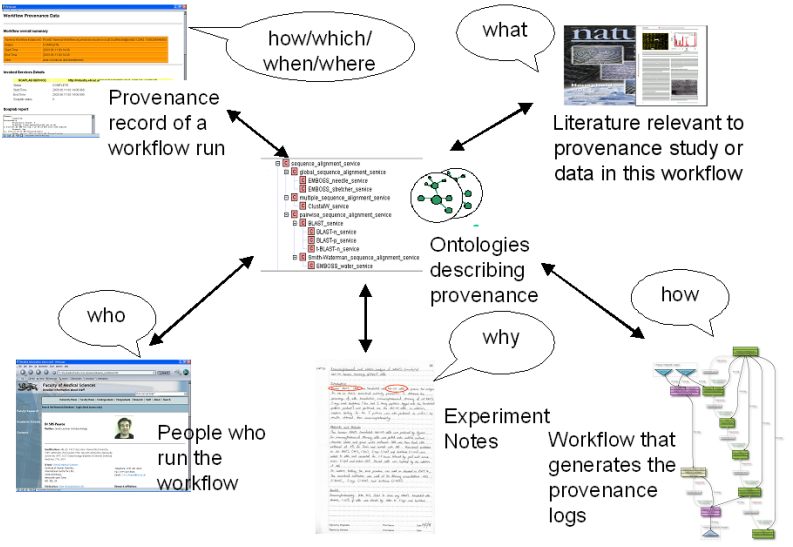

Fig. 1. A Web of provenance connected through the "Semantic Glue"

1. services that are the **tools** that will constitute the experiments, that is: external third party services such databases, computational analyses, simulations etc, presented as web services;
2. services for **forming and executing experiments**, primarily workflow management services for creating and enacting workflows, Taverna/Freefluo [3] which enact the XML-based workflow language XScufl (XML Simple Conceptual Unified Flow Language) [8], and myGrid Information Repositories (mIRs) for storage;
3. services for supporting the **e-Science scientific method**, specifically provenance management [9] and change notification [10];
4. **semantic services** for discovering services and workflows, and managing metadata, such as semantically annotated service registries and federated personalised views over those registries.

myGrid services can be described as semantically aware. Ontologies represented using the DAML+OIL ontology language [11] are used to describe the inputs, outputs and tasks of web services and workflows, as well as the semantic characterisation of an object in the mIR. An ontology captures a community's understanding or knowledge of a domain and provides a common, shared understanding of that domain. For the purposes of this paper an ontology is a hierarchical classification of domain specific concepts, a set of their properties, and their internal relationships. DAML+OIL is an ontology language developed

[3] Taverna and FreeFluo are both open source projects available from `http://taverna.sourceforge.net` and `http://freefluo.sourceforge.net`

for the Semantic Web. Its successor OWL (Web Ontology Language) [12] is now a W3C Recommendation. DAML+OIL is based on a Description Logic, whose reasoning abilities support automated classification of terms into hierarchies [13]. myGrid uses DAML+OIL ontologies to annotate web services stored in a registry to enable service and workflow discovery and composition. These DAML+OIL semantic descriptions build on the work of the DAML-S coalition and have been used to guide the construction of workflows by constraining the choice to those services, which have semantically compatible inputs and outputs. Similarly semantic description of workflows has been used to discover relevant workflows given an item of data selected from the mIR. An ontology service provides a single point of reference for these concepts and supports reasoning over concept expressions. The use of semantic web technology such as ontology services makes this an early example of a *Semantic Grid* [14].

2.1 Producing Process Provenance

Process level workflow provenance is generated as a process graph, with services as nodes and data as arcs. The workflow description is submitted to the Freefluo workflow enactment engine to "run" the experiment. Each workflow run consumes two XML documents:

1. An XScufl document gives the low-level workflow definition, describes the ordering of process invocations and data passes, and hence plays a similar role to the DAML-S process model [15].
2. A Web service information (ws-info) document contains ontological descriptions that are associated with the inputs, outputs and services in a workflow, similar to the DAML-S profile [16]. This document has two roles: the myGrid registry uses it to advertise and hence discover Web services based on their semantics; and the myGrid workbench environment uses it when finding suitable data inputs that semantically match those of a workflow.

As shown in Fig 2 (excluding the dashed box), before running a workflow, users interact with the myGrid workbench and choose workflows through their semantic descriptions stored in the ws-info files. When a workflow is executed, the workbench retrieves from the mIR the XScufl document , and the required resources, such as input/output data and parameters, based on the ws-info file. The workbench invokes the Freefluo enactment engine with the XScufl file and an XML document containing the input data. The process provenance logs are generated at the same time as data results, in the form of XML, by the workflow enactment engine. At the end, the process provenance and the data results are returned to the mIR through the workbench.

Figure 3 shows the content of an example provenance. This is interpreted from the original XML format provenance logs. This process provenance records: the start time, end time, the user of this workflow, the service invoked in this workflow (getSequence); the parameters of the service invocations; and the inputs and outputs and their metadata of the workflow.

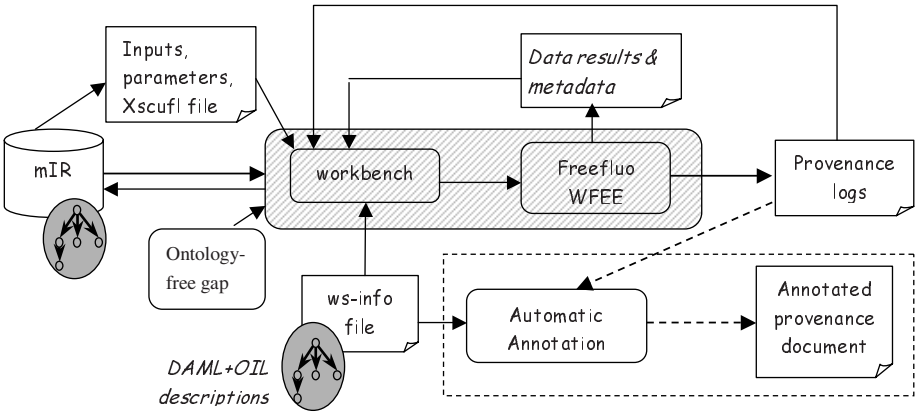

Fig. 2. The generation of provenance in ^{my}Grid.

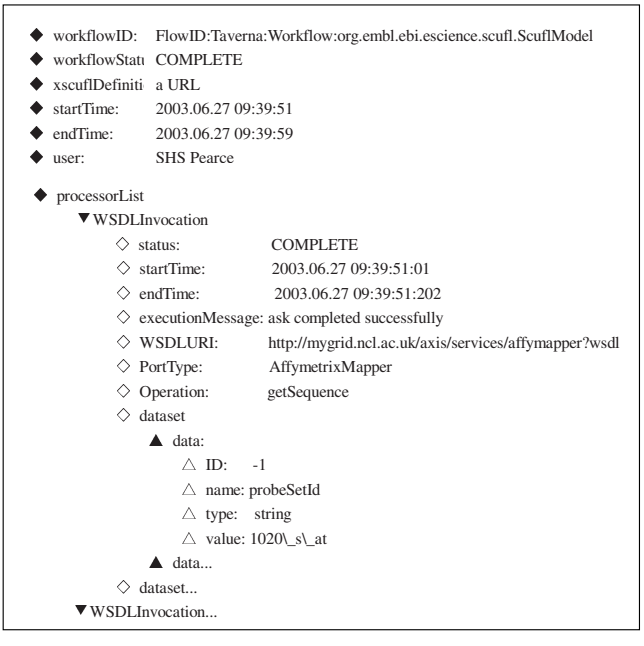

Fig. 3. An example provenance fragment.

An interface for navigating and querying these provenance logs is required by bioinformaticians. Similar experiments, using similar inputs, etc. need to be linked together semantically, at the knowledge level, in order to answer the kinds of tasks described in Sect 1. For example, what are the experiment runs using *BLAST* Web services with a *White_blood_cell* as input? Or how to find workflow runs whose outputs *contain* the *Single_nucleotide_polymorphism*? Also what is

the literature that users refer to when designing an experiment for invoking sequence alignment services; what are the notes about experiment conclusions, etc. Simple keyword searching, such as asking for a particular data value to find all similar experiments, will only find experiments with that data item and not those with data values that are instances of the same concept.

To enable such tasks, we integrated explicit domain semantics with our provenance logs and presented our Web of semantically linked provenance in a browser environment. For this we used the COHSE (Conceptual Open Hypermedia Services Environment) semantic hyperlink generation tool powered by the ontologies used by the myGrid registry for service and workflow discovery and type management.

3 A Semantic Web of Provenance

In this section we describe the COHSE semantic hyperlink tool, and show it in use. In Sect 4 we give details of the ontologies we used and how we annotated the logs with the ontologies.

Detailed descriptions of the COHSE system can be found in [17]. Briefly, the COHSE approach consists of a *COHSE agent* that augments documents with links based on the semantic content of those documents. Figure 4 shows the three technologies of COHSE. *An Ontology Service*, uses rich knowledge representation techniques and reasoning abilities to provide a machine processable semantics to the conceptual metadata associated with documents and between concepts. *An Annotation Service* plays two roles: (a) annotates documents or sections of documents with concepts from the loaded ontology and (b) maintains a link base that stores bi-directional mappings between concepts and target URIs (Unique Resource Identifiers). *A Link Service* fetches target URIs for the concepts associated with documents and displays the links, supported by the reasoning of the Ontology Service.

Semantic annotations in COHSE refer to the process of identifying and associating ontological concepts with documents. These semantic metadata are not embedded in the documents. It is a Link Base that maintains the mapping from target links to concepts by storing: the URL of the document, an XPointer expression indicating the fragment of the resource being annotated, a simple textual description of the annotations, the DAML+OIL concept associated with the annotation, and the ontology selected for this annotation. This guarantees that in the process of conceptual linking, the retrieval of target links is based on explicit ontology context.

The Link Service in COHSE has the basic task of generating and presenting links to Web pages on behalf of both authors and readers. When viewing a document COHSE, the Link Service obtains the concepts that will form the link source anchors either through the *lexicon*, which holds some language terms and their mappings to the ontological concepts loaded by the Ontology Service; or through *semantic annotations*. After identifying all the concepts in the page, the Link Service in COHSE contacts the Annotation Service to identify target

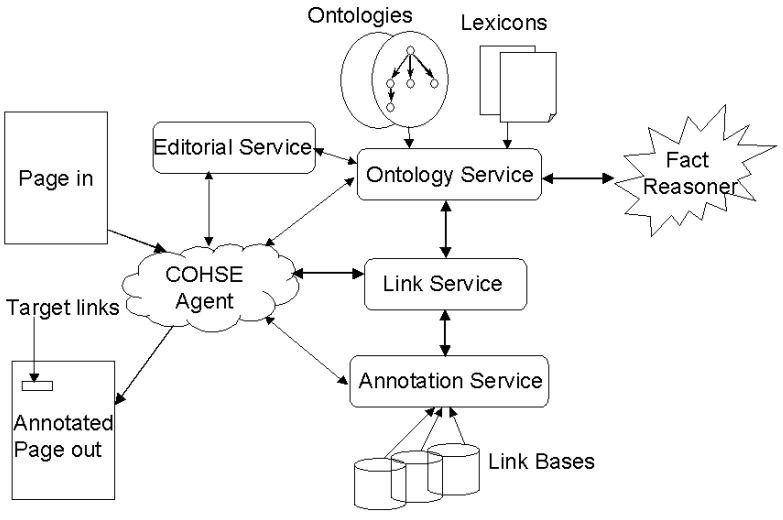

Fig. 4. The COHSE architecture.

URLs with corresponding concepts, as well as subsuming and subsumed concepts inferred by the Ontology Service.

The system employs either a specialist browser (based on Mozilla[4]) or a proxy through which all http requests are routed. In the Mozilla version, the DOM is manipulated to present the inferred links; in the proxy version the service creates a new document.

3.1 Generating Links

There are two types of entry point mechanisms to access the concepts that will form the link source anchors within a document:

1. Through the use of some language terms within the document, held in a lexicon, mapped to the concepts within an ontology provided by the Ontology Service. For example, the word *probeSetId* maps to the concept *Affymetrix_probe_set_id*. The relevant concepts in the ontology can then be used to determine appropriate targets for links out of the given document.
2. Through the use of explicitly asserted ontological annotations on regions of the document provided by an annotation service, e.g. a region is annotated directly with the concept *lymphocyte*. This approach relies on the ability to annotate resources with semantic metadata – where by semantic metadata we mean the explicit binding of concepts to resources rather than the use of terms and words as simple proxies for the concepts.

[4] `http://www.mozilla.org`

Key to the novelty of the COHSE approach is the provision of an *editorial component* within the agent. This component uses information within the ontology (such as hierarchical classification) in order to determine whether the links are suitable or to perhaps expand or cull the set of possible targets. Figure 4 shows a simplified view of the basic architecture of the system. The explicit annotations can then help guide the editorial component in its linking strategy. For example, if a passage in a web page has been annotated as being about a particular subject, say *Sequence_alignment*, the editorial component may know that there are certain terms that should be focused on within the context of that annotation (say the term *pair-wise*) – an example of an agent using semantic information to make decisions as to its behaviour.

Once an ontology is loaded, the COHSE agent contacts the Ontology Service to obtain a list of terms that are used in the lexicon to represent concepts in the ontology. One concept usually has more than one language terms corresponding to it. For example, the concept of *researcher* may have different expressions of researcher in different languages or different ways of capitalization. With the help of the lexicon we map all these into one concept *researcher*. When a new web page is viewed by COHSE, the agent grasps all the words in the page. For each word that corresponds to the term in the lexicon, the agent associates the concept with it. At the same time the agent contacts the Annotation Service to determine whether any region in the document has been annotated with any concept. Having identified all the concepts in the page, the agent contacts the Annotation Service to identify target links to other URLs with corresponding concepts as inferred by the Ontology Service using an inference reasoning service for the concepts provided by the FaCT reasoner [18]. Consequently, we are able to link web pages together based on the concepts associated with its documents. By changing the ontology we change the link anchors and link targets for the same web pages, getting different link *views* over those resources.

When the semantically annotated provenance logs are viewed through COHSE, the Link Service can identify bioinformatics concepts from the provenance log supported by the lexicon and query all the target URIs for each concept from the link base. In this way, a Web of provenance logs is built by semantically linked together. An example result is shown in Fig 5. The bioinformatics concept for the input data, 1020_s_at, — *Affymetrix_probe_set_id* — is attached to the provenance log. Since provenance logs are linked together based on the annotated semantic concepts, instead of the same data values in the logs, scientists are able to navigate from the current log to others which also use data of the type *Affymetrix_probe_set_ids* or its parent concept *Gene_id*, either as an input or an output. Using the ontology structure, links to subsumed and subsuming concepts can be fetched, supported by the reasoning ability of the Ontology Service of COHSE. Thus, users can browse the Web of provenance starting from provenance logs invoking *sequence alignment* services, to provenance logs invoking *BLAST* services or other *multiple sequence alignment services*, both of which are subsumed concepts of *sequence alignment*.

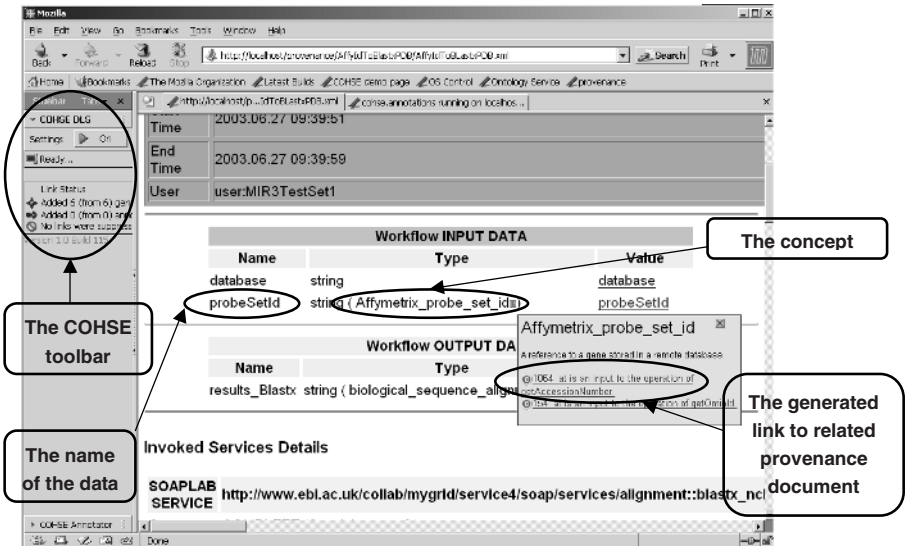

Fig. 5. Provenance discovery through semantic generation of links between provenance logs.

This Semantic Web of provenance is not limited to just provenance logs. COHSE can provide links from provenance logs to any document that can be displayed in a web browser: from a researcher's HTML home page, to a Word document or a PDF file. For example, from a workflow provenance log having invoked a *sequence alignment* service by a human geneticist, users can find links to the home page of this person describing his or her research activities in human genetics; and to a PDF file about *BLAST*, which introduces sequence alignment, etc. We imported the Gene Ontology into COHSE, for example, and were immediately able to present hyperlinks between documents based on shared GO terms.

By changing the annotation ontology, scientists can also discover computational resources beyond provenance logs. For example, when choosing viewing the Web of provenance logs with the *generic* ontology, a biologist can query the other logs that also performed upon the Grave's disease or upon the Williams-Beuren syndrome. If interested, he/she can also check out how other scientists performed the experiments which have a similar research hypothesis as his/hers. When the *biological* ontology is chosen, a higher level linking between these logs can be obtained. A biologists can query all the provenance logs which use a *Gene_id* as an input, instead of a bioinformatics term *Affymetrix_probe_set_id* (since *Affymetrix_probe_set_id* is a kind of *Gene_id*). The more general term maybe more comprehensible to a pure biologist, as opposed to a bioinformatician.

Thus this semantically linked Web of provenance logs can help answer the knowledge level questions raised in Sect 1.

4 Realising a Semantic Web of Provenance

Here we describe what and how ontologies are used to annotate the provenance logs and how the process provenance logs are semi-automatically populated with semantic concepts.

4.1 Domain Ontologies

We used three categories of ontologies, built with DAML+OIL:

- **Generic Ontology**, mainly formalizes concepts about organization, people, research topic and publishing, which supports general linking between provenance logs and documents, like experimental notes, research papers; and links between users who executed the workflows based on the relationships of their organizations or research topics. This small ontology shows a similar schema to the Dublin Core [19].
- **Bioinformatics Ontology**, conceptualizes a precise classification of bioinformatics data, web services and workflows. Services and data concepts are classified in multiple ways, for example, bioinformatics services can be classified by operation (alignment, pairwise, multiple), by the type of data source (protein, nucleotide, sequence) or by algorithm (SmithWaterman, BLAST [16].
- **Biological Ontology**, is based on the TAMBIS (Transparent Access to Bioinformatics Information Sources)[5] ontology. This ontology is used to annotate data with biological concepts, like *protein*, *acid* and *sequence*. Though large, well-established molecular ontologies as Gene Ontology [20] already exist, they mainly formalize specific data instances instead of more abstract types of data, like the biological entities processed by the bioinformatics Web services used in ^{my}Grid.

4.2 Acquiring the Semantic Annotations

As described in Sect 2, the process provenance logs are automatically recorded by the Freefluo workflow engine. The Freefluo workflow enactment engine is neutral about the services it enacts and only deals with WSDL (Web Service Definition Language) [21]. The ontological information in the ws-info files was not passed through the Freefluo engine to emerge in the logs or attached to the data, as shown in the Fig 2. So no domain semantics are associated with provenance entities even though the semantic information is available in ws-info files.

Here we extend the role of these ws-info files to provide domain semantics for provenance entities. This abandoned information is recovered by a post-processing annotation phase, as highlighted in the dashed box in Fig 2. From a close study of a ws-info file and its corresponding provenance log, shown in Fig 6, the semantic concept for the data 1020_s_at recorded in the provenance log can

[5] `http://img.cs.man.ac.uk/stevens/tambis-oil.html`

be retrieved from the ws-info file as *Affymetrix_probe_set_id*, based on the name of the data entity — *probeSetId*. While querying the corresponding ws-info files, the XML provenance logs can be semantically integrated with ontological concepts, defined mainly in the bioinformatics ontology. In this way, we have entries into the concepts with the instances of concepts in the provenance logs. Thus, we extend the process provenance log with a process of semantic annotation performed after the workflow run and before the provenance document is archived in the mIR, ready to be viewed by COHSE. In our prototype shown here in Fig 5 we have embedded the concepts within the document; other implementation store the concept externally using the COHSE Link Base. When the number of provenance logs increases, a Semantic Web of provenance is automatically and dynamically added to by this automatic annotation process.

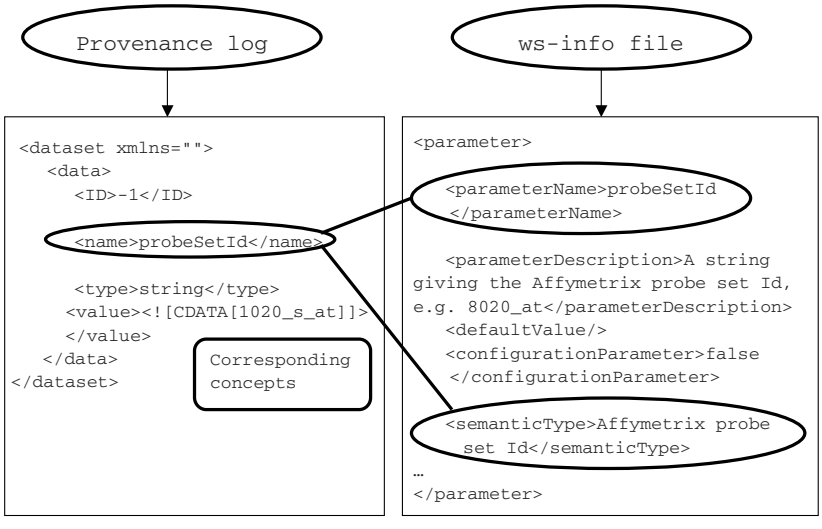

Fig. 6. The ontology concept in the corresponding ws-info of a provenance log. The annotation process transfers the concept from the ws-info file to the provenance log.

Despite the promise of this approach, there are problems. Semantically linking provenance logs based on a single *bioinformatics* ontology shows limitations for some of our users. Instead of viewing provenance logs from a bioinformatics point of view, they are more interested in more purely biological view of provenance, using semantics like *protein kinase*, *DNA binding*, *transcription*, or *carbohydrate metabolism*, alongside the bioinformatics or experimental methodological view of the data. This is because we can automatically annotate at the process and data level of provenance, but not the knowledge level.

We address this problem by semantically annotating provenance logs with *Biological* and *Generic* ontologies. This annotation is mainly a manual process. It depends on the authors of annotations to discover the *biological* relationship

between data, like a DNA sequence *encodes* a protein, or a gene *expresses* a protein, a semantic property from the Biological ontology. It also depends on authors' to identify semantic concepts from documents and associate discovered resources, like reference literatures, with provenance logs. This handcrafted process has high requirements for expert knowledge of authors and demands significant effort. Also how much we should trust the manual annotations contributed by distributed users is still an open research topic. Figure 7 shows the results of linking two pages with biological ontologies.

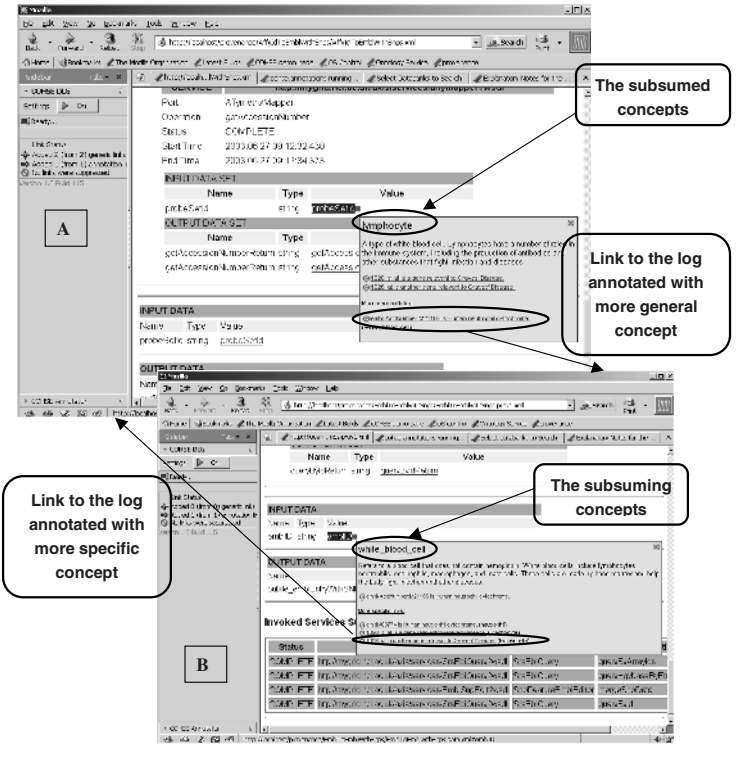

Fig. 7. Generated links based on conceptual annotations. Page A probeSetId is annotated with the concept *lymphocyte*, generating links to other pages annotated by lymphocyte and pages annotated with more general concept including page B annotated with *white_blood_cell*. Page B emblID annotated with the concept *white_blood_cell* generating links to other pages annotated by the same concept and pages annotated with more specialised concept including page A annotated with *lymphocyte*

.

5 Related Work

Several ontology-driven annotation tools are available, such as Magpie [22] and MnM [23]. Though Magpie is a very light annotation tool, the only annotation mechanism Magpie supports is through lexicon mapping. If we want to apply Magpie to annotating the provenance logs, first we have to populate the lexicon. But, it is difficult to build a lexicon of mapping each piece of computation data to its corresponding concept. This lexicon has to be updated each time a new piece of data needs to be annotated. MnM can support marking-up the document with semantic metadata, but its linking service can not provide target links to the subsuming/subsumed concept of the annotated concept. COHSE, on the other hand, supported by an ontology reasoner, is able to present target links to both subsuming and subsumed concepts of the annotated concept.

There are some other projects supporting knowledge discovery from provenance logs from different approaches. The Provenance Aware Service Oriented Architecture (PASOA) project aims to automatically record provenance logs and apply some reasoning algorithms over provenance logs by a provenance service client and server [24]. Their provenance logs mainly trace the process of service invocations generating a particular data in the Grid for a workflow enactment, which are similar to our process level provenance logs. They provide a tree-like provenance browsing environment, in a XML format, organized by the services and data in a workflow, without direct support for browsing linked provenance in a usable interface for e-Scientists. The focus of this project is to discover the quality and accuracy of data and services, and automatic re-execution of services in the context of e-Science based on provenance logs, instead of the semantic relationships among provenance logs, as in myGrid.

Chimera [6] records virtual data provenance to discover available methods and fulfill on-demand data generation in the Grid computing environment. Chimera comprises two main components: a virtual data catalog and a virtual data language interpreter. Virtual Data Catalog records transformation, derivation and data underpinned by a compact virtual data schema, which has a direct support for data linking. Though myGrid does not have direct data derivation graph in the first release, the current myGrid achieves similar results by building a RDF data provenance graph. Chimera Virtual Data Language supports data definition and query statements, written in XML. The current querying over provenance logs is mainly keyword based, rather than from semantic point of view. Chimera contains the mechanism to locate the "recipe" to produce a given logical file, in the form of an abstract program execution graph, to support job scheduling and planning in the Grid computing environment.

6 Discussion and Future Work

By annotating provenance logs with concepts drawn from the myGrid ontology, we built a dynamically generated hypertext of provenance documents, data,

[6] `http://www.griphyn.org/chimera/`.

services and workflows based on their associated concepts and reasoning over the ontology. We demonstrated a first implementation of semantically linking provenance logs and showed its utility in enhancing the acquisition of knowledge from data generated in myGrid. During this process, however, we came across two kinds of problems:

Knowledge provenance acquisition. The approach uses the bioinformatics ontology to annotated provenance logs with process provenance. In this way we hoped to simply reuse the bioinformatics ontology for a different task. However, it was not as reusable as first hoped in that it was good for discovering services but poor for linking logs. This ontology is too abstract to annotate data based on the *value* of the data. It only supports annotation based on the *type* of the data. Users can associate a bioinformatics concept — *Affymetrix_probe_set_id* — with the data 1020_s_at that is explicitly specified with this semantic type; but they can not say this data is a *lymphocyte*. This requires high level of text-mining technique to recognize the semantic meaning of this data or depends on time-consuming and error-prone manual annotations.

Difficulty of obtaining semantic annotations. Because our workflow enactment engine, Freefluo, is generic and neutral about the services that it enacts, and it does not preserve the ontological annotations associated with the services and their parameters. Thus we were throwing away valuable semantic annotations and were forced to recover them through post-analysis.

The work described here relates to myGrid version 1. In version 2 a Taverna Scufl workbench is extended to replace the previous workbench. It runs a workflow execution based on the XScufl script, supported by the underlying Freefluo enactment engine. Bioinformaticians can search for workflows by keywords as well as semantic concepts in Taverna, to obtain a XScufl script and a workflow graph as a recipe for experiment design. The semantic annotations for services and workflows are a) kept in the registry (for 3rd party semantics) and b) encapsulated in the metadata of services, identified by a Life Science Identifiers (LSIDs) [25]. LSID is a special Universal Resource Name (URN) proposed by the I3C consortium. The LSID resolution protocol promises persistent identification of objects. myGrid is extending the mIR with a LSID authority, to assign and resolve the LSID for each myGrid entity. The mIR exposes both a LSID data interface and a LSID metadata interface for data and metadata in the repository, which are moving to the RDF format. Thus given an LSID, both the data and its metadata can be retrieved from the mIR by resolving the LSID.

Taverna generates RDF provenance logs, which is used as a model to integrate more expressible ontological concepts with experiment resources. RDF is recommended by the W3C consortium as one of the core technologies for the Semantic Web. Each RDF triple consists of a subject, a predicate and an object, with the predicate linking the two others. A set of RDF triples forms a RDF graph model. RDF can identify resources by URI, of which URN is a type. Thus RDF can support building a graph of myGrid entities, identified by their LSIDs.

Currently the RDF provenance logs are being imported to Haystack[7], which provides a RDF file navigation and visualisation environment. In Haystack we achieve providing different views of linked provenance logs according to users' requirements. As a flexible data model, RDF can also be compatible with the syntax of XML and OWL, which promises the compatibility of our extension with existing myGrid provenance and the feasibility of the integration of ontological concepts with data.

Even if it becomes easier to identify semantic metadata for experiment resources, it is still not straightforward to annotate provenance logs with semantics. Logs are generated by the Freefluo workflow engine, which neither cares nor knows about semantics. One possible approach can be to keep the semantics with service descriptions and input parameters intact before and during the workflow execution in the workflow engine. At the end of the execution, semantics for outputs are queried from the service descriptions and attached to the result data. Thus in the end we can have semantics for services and data in the provenance logs. Another is to post-annotate the logs as what we did in the first release of myGrid, but the semantics (as part of metadata) are identified by resolving the LSIDs which are used to identify experiment resources.

The outcome of our experiments suggests that a dynamically generated Semantic Web of provenance records is a viable possibility. However, as with all the Semantic Web applications, the old story of where to get the annotations and what are the effective ontologies still stands, even for this closed problem in a closed domain.

Acknowledgments. The myGrid project, grant number GR/R67743, is funded under the UK e-Science programme by the EPSRC. The authors would like to acknowledge other members of the myGrid team for their contributions; and 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 implementation process.

References

1. Fox, G., Walker, D.: e-Science gap analysis. Technical report, Indiana University and Cardiff University, UK e-Science Center (2003)
2. Stevens, R., Glover, K., Greenhalgh, C., Jennings, C., Li, P., Radenkovic, M., Wipat, A.: Performing in silico experiments on the Grid: a users perspective. In Cox, S.J., ed.: UK e-Science All Hands Meeting 2003 Editors. (2003) 43–50
3. Stevens, R., Tipney, H., Wroe, C., Oinn, T., Senger, M., Lord, P., Goble, C., Brass, A., Tassabehji, M.: Exploring Williams-Beuren Syndrome using mygrid. In: Proceedings of 12th International Conference on Intelligent Systems in Molecular Biology. (2004)
4. Foster, I., Kesselman, C., eds.: Blueprint for a new computing infrastructure . Second edn. Morgan Kaufmann Publishers (2003)

[7] `http://haystack.lcs.mit.edu/`

5. Moreau, L., Miles, S., Goble, C., Greenwood, M., Dialani, V., Addis, M., Alpdemir, N., Cawley, R., De Roure, D., Ferris, J., Gaizauskas, R., Glover, K., Greenhalgh, C., Li, P., Liu, X., Lord, P., Luck, M., Marvin, D., Oinn, T., Paton, N., Pettifer, S., Radenkovic, M.V., Roberts, A., Robinson, A., Rodden, T., Senger, M., Sharman, N., Stevens, R., Warboys, B., Wipat, A., Wroe, C.: On the use of agents in a bioInformatics Grid. In: The Third IEEE/ACM CCGRID'2003 Workshop on Agent Based Cluster and Grid Computing, Tokyo, Japan (2003) 653–661
6. Klyne, G., Carroll, J.J.: Resource description framework (RDF): concepts and abstract syntax. W3C Proposed Recommendation. Available at `http://www.w3.org/TR/2003/PR-rdf-concepts-20031215/` (2003)
7. Lord, P., Wroe, C., Stevens, R., Goble, C., Miles, S., Moreau, L., Decker, K., Payne, T., Papay, J.: Semantic and personalised service discovery. In: Proceedings of Workshop on Knowledge Grid and Grid Intelligence (KGGI'03), in conjunction with 2003 IEEE/WIC International Conference on Web Intelligence/Intelligent Agent Technology. (2003) 100 – 107
8. Addis, M., Ferris, J., Greenwood, M., Li, P., Marvin, D., Oinn, T., Wipat, A.: Experiences with e-Science workflow specification and enactment in bioinformatics. In Cox, S.J., ed.: Proc UK e-Science All Hands Meeting 2003. (2003) 459–466
9. Greenwood, M., Goble, C., Stevens, R., Zhao, J., Addis, M., Marvin, D., Moreau, L., Oinn, T.: Provenance of e-science experiments - experience from bioinformatics. In Cox, S.J., ed.: UK e-Science All Hands Meeting 2003 Editors. (2003) 223–226
10. Krishna, A., Tan, V., Lawley, R., Miles, S., Moreau, L.: mygrid notification service. In Cox, S.J., ed.: UK e-Science All Hands Meeting 2003 Editors. (2003) 475–482
11. Horrocks, I.: DAML+OIL: a Description Logic for the Semantic Web. The IEEE Computer Society Technical Committee on Data Engineering **25** (2002) 4–9
12. Horrocks, I., Patel-Schneider, P.F.: A proposal for an owl rules language (2004)
13. Baader, F., Horrocks, I., Sattler, U.: Description Logics as ontology languages for the Semantic Web. In Hutter, D., Stephan, W., eds.: Festschrift in honor of Jörg Siekmann. Lecture Notes in Artificial Intelligence, Springer (2003) To appear.
14. Wroe, C., Goble, C., Greenwood, M., Lord, P., Miles, S., Papay, J., Payne, T., Moreau, L. In: Automating experiments using semantic data on a bioinformatics Grid. Volume 19. IEEE Intelligent Systems, Special Issue on E-Science (2004) 48–55
15. Ankolekar, A.: The DAML Services Coalition DAML-S: Web Service description for the Semantic Web. In: The First International Semantic Web Conference (ISWC), Sardinia (Italy) (2002)
16. Wroe, C., Stevens, R., Goble, C., Greenwood, M.: A suite of DAML+OIL ontologies to describe bioinformatics Web Services and data. International Journal of Cooperative Information Systems **12** (2003) 197–224
17. Bechhofer, S., Goble, C., Carr, L., Kampa, S., Hall, W., Roure, D.D.: COHSE: Conceptual Open Hypermedia Service. Volume 96. IOS Press, Frontiers in Artifical Intelligence and Applications (2003)
18. Horrocks, I.: The FaCT system. In de Swart, H., ed.: Automated Reasoning with Analytic Tableaux and Related Methods: International Conference Tableaux'98. Number 1397 in Lecture Notes in Artificial Intelligence, Springer-Verlag (1998) 307–312
19. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin Core metadata for resource discovery. In: The Internet Society. (1998)
20. Ashburner, M., et al: Gene Ontology: tool for the unification of biology. Nature Genetics **25** (2000) 25–29

21. Christensen, E., Curbera, F., Meredith, G., Weerawarana, S.: Web Services Description Language (WSDL) 1.1 (2001) W3C Note.
22. Dzbor, M., Domingue, J.B., Motta, E.: Magpie - towards a semantic web browser. In: The 2nd International. Semantic Web Conference, Florida US (2003) 255 – 265
23. Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., Ciravegna, F.: Mnm: ontology driven semi-automatic and automatic support for semantic markup. In Gomez-Perez, A., ed.: The 13th International Conference on Knowledge Engineering and Management (EKAW 2002). (2002)
24. Szomszor, M., Moreau, L.: Recording and reasoning over data provenance in Web and Grid Services. In: ODBASE. (2003)
25. Clark, T., Martin, S., Liefeld, T.: Globally distributed object identification for biological knowledgebases. Briefings in Bioinformatics **5** (2004) 59–70