Annotating, linking and browsing provenance logs for e-Science

Jun Zhao, Carole Goble, Mark Greenwood, Chris Wroe, Robert Stevens
Department of Computer Science, University of Manchester, Oxford Road,
Manchester, M13 9PL <zhao/carole/markg/cwroe/stevensr>@cs.man.ac.uk

Abstract

Like experiments performed at a laboratory bench, the results of an e-science *in silico* experiment are of limited value if other scientists are not able to identify the origin, or ***provenance***, of those results. For e-Science, we need more systematic provenance logs across a range of e-Science activities and disciplines as well as a more informed understanding of the information in these provenance data. Semantic Web technology, which enables data to be linked and defined in a way for more effective discovery, integration and cooperation across computers and people, provides an appropriate solution for our current requirement. In this paper we show how we used the COHSE conceptual open hypermedia system to build a dynamically generated hypertext of web of provenance documents arising from the myGrid project based on associated concepts and reasoning over the ontology.

1. Introduction

e-Science is the use of electronic resources -- instruments, sensors, databases, computational methods, computers – by scientists working collaboratively in large distributed project teams in order to solve scientific problems. An *in silico* experiment is a procedure that uses computer-based information repositories and computational analysis to test a hypothesis, derive a summary, search for patterns, or demonstrate a known fact. The myGrid project[1] [Goble03] is developing high-level service-based middleware to support the construction, management and sharing of data-intensive *in silico* experiments in biology.

Like experiments performed at a laboratory bench, the results of an e-science *in silico* experiment are of limited value if other scientists are not able to identify the origin, or ***provenance***, of those results. Thus, provenance is a kind of metadata, recording the process of biological experiments for e-Science, the purpose and results of experiments as well as annotations and notes about experiments by scientists [Buneman02].

Provenance logs in the biological community are expected to record the process of biological experiments, which cover the person who involves in the experiment (who); the materials and methods used in the experiment (what and how); the purpose of running this experiment (why); as well as the results and conclusions of the experiment (what). For biologists, the focus is not just the intention and results of experiments but also the understanding of the "*how to*" of experiments. Thus there are two major forms:

- an *annotation* attached to an object or collection of objects, such as a database entry, in a structured, semi-structured or free text form;
- a *derivation path* such as a workflow, a database query, or a program & its parameters. The derivation path could be the logging of services invoked by a workflow, the logging of an evolution of a workflow (e.g. substituting a BLAST algorithm for a PSI-BLAST algorithm) or recording alterations of the parameters of an activity while a workflow is being enacted.

Currently only sporadic and incomplete provenance is available for *in silico* experiments. For e-Science, we need more systematic provenance logs across a range of e-Science activities and disciplines as well as a more informed understanding of the information in these provenance data. Semantic Web technology, which enables data to be linked and defined in a way for more effective discovery, integration and cooperation across computers and people, provides an appropriate solution for our current requirement.

Provenance generation, management and discovery is a core component of myGrid . Not only do we generate provenance logs that can fulfill many of the functions of a conventional lab book but in a machine accessible way, but also to annotate and link this provenance data to form a web of provenance experimental holdings. myGrid makes liberal use of ontologies to annotate, discover and manage its various components. By annotating provenance logs with concepts drawn from an ontology, we propose that we can link the records to each other and or other experimental holdings in myGrid by means of inference over these associated concepts. The COHSE (Conceptual Open Hypermedia Services Environment) system provides an open hypermedia environment, enabling documents to be annotated and linked based on the concepts associated with document contents. Using COHSE we aim to build a dynamically generated hypertext of web of provenance documents, data, services and workflows based on associated concepts and reasoning over the ontology.

The rest of the paper is organised as follows: in section 2 we discuss provenance records in myGrid, its generation, its management, and its information model. In section 3 we introduce the COHSE system. In section 4 we demonstrate our on-going experiments with browsing,

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

linking and annotating provenance logs. Section 5 gives related work and section 6 concludes with a discussion and future plans.

2. Provenance and myGrid

In silico experiments in myGrid are primarily made up of four components:

- *Services* such as databases, applications, text extraction and so on, that a scientist uses as his basic tools. Services might be personal to the scientist, local to the enterprise or global serving the community. In myGrid these are presented as Web (eventually Grid service).
- *Workflows* are our primary experimental mechanism for integrating services, representing a protocol that orchestrates and enacts a range of remote and local disparate services. Workflows are used to derive outputs from inputs. The outputs of one workflow or service may form the inputs to another so that a complete in silico experiment includes a network of related workflow invocations. Any output can be associated with its corresponding workflow invocation record and the associated provenance data. In this way the detail of how an output is related to its inputs is permanently recorded.
- *Provenance* records metadata about experiments including: records of enacted workflows: data results, a history of services invoked by the workflow engine, instances of services invoked, parameters set for an application, notes commenting on the results, and so on.
- *"Glue" metadata* associates experimental components together to form an investigation web, including: notes describing objectives, applications, databases, and relevant papers, the web pages of important workers.

Services and workflows are published and discovered by a personalised federated semantic registry. Services and workflows can be advertised and found by semantic descriptions covering their inputs, outputs, function and so forth, drawn from an ontology [Wroe03]. A myGrid Information Repository (mIR) is a specialist data service that serves as a store for experimental components (data, workflow specification documents) and metadata about the experiments (provenance records and associations between experimental components). Each and every mIR entry can also be associated with (multiple) terms from an ontology. The myGrid ontology is a suite of ontologies managed by an ontology service and represented in DAML+OIL [Horrocks02]. It draws on community efforts such as DAML-S [DAML-S02], covering web services, bioinformatics, molecular biology, publications, organisations and research methods; see [Wroe03] for details. A workbench application, developed using NetBeans, has been prototyped as a platform for technical experimentation using biological investigations in gene expression and SNP analysis for Grave disease.

In this paper we concentrate on the provenance that arises from running workflows.

2.1 Generating workflow provenance

The myGrid environment uses the Freefluo workflow enactment engine [Freefluo], which can handle WSDL based web service invocation. It supports two XML workflow languages, one based on IBM's Web Service Flow Language and our own, XScufl, developed as part of a Taverna, a collaboration with the Human Genome Mapping Project. The XScufl workflow script can represent a workflow specification or *template* (with unbound service instances); a workflow instance (where the services are bound to concrete implementations); or be partially instantiated.

Each workflow run is based on two XML documents, both of which are stored in the mIR:

- A XScufl document giving the workflow definition, describing the process of services composition, and hence plays a similar role to the DAML-S process model; and
- A ws-info document which contains ontological descriptions associated with the inputs, outputs and services in a workflow, similar to the DAML-S profile. This document has two roles: the myGrid registry uses it to advertise and hence and discover workflows based on their semantics; the myGrid workbench environment uses it when interrogating the mIR for data inputs that semantically match a workflow, for providing configuration and defaults information for service parameters, and for identifying the data type of the workflow data results. Figure 5 gives a screenshot of a ws-info document.

When a workflow is executed, the workflow enactment engine extracts needed resources, such as input/output data and parameters, from the mIR based on the XScufl documents. The provenance logs are generated at the same time in the form of XML files by the workflow enactment engine, recording the start time, end time and service instances operated in this workflow. At the end of the workflow execution, the result data, metadata about the workflow and the provenance logs are put back into the mIR via the workbench in the appropriate semantic and MIME type provided in the ws-info files. All mIR objects carry provenance attributes: hence the provenance log has who created it, when, in what context, and so on. In addition, a set of metadata is associated with this workflow invocation instance: the input and output relationships between the workflow instance and data items, the 'is defined by' relationship between the workflow instance, the ws-info document and the XScufl script. See Figure 1.

Other annotations regarding the hypothesis of the experiment, thoughts and opinions by the scientist and quality of results are also stored as XML in the mIR or as regular web documents.

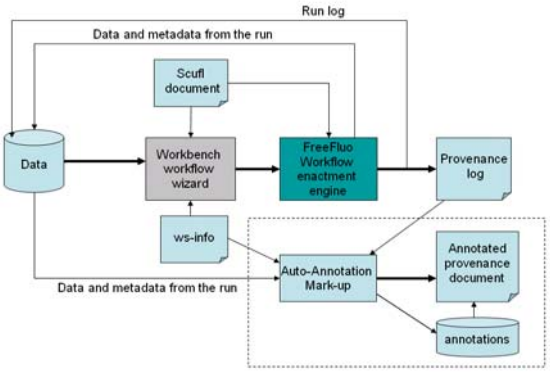

Figure 1: Generation of provenance in [my]Grid

From the view of FreeFluo each workflow execution includes more than one operation. In each service operation there are inputs, outputs and intermediate data. The information model of the workflow provenance logs currently includes metadata about inputs, outputs, parameter data as well as services operated in each workflow execution, see Figure 2.

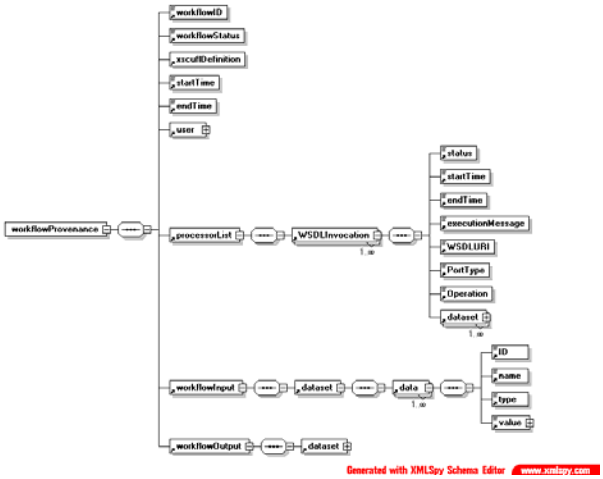

Figure 2: Information Model of Provenance

2.1 Exploiting workflow provenance

In the biological community the same experiment operated at different times may result in different results. Provenance documents provide record of the past experiments for biologists to understand how this piece of scientific data is generated by their colleagues. Sometimes biologists also prefer to get some information about the data, where it is generated by other scientists, the location of the provenance documents, whether there are any papers or web sites about the data.

How we can link provenance documents together and provide a platform for e-Scientists not only to browse them but also to link and annotate them?

We would like to build a web of related pages relevant to an experimental investigation, marked up with, and linked together using annotations drawn from shared ontologies (see Figure 3). This web includes not only the provenance record of a workflow run but also links to other provenance records of other related or unrelated workflow runs, diagrams of the workflow specifications, web pages about people who ran the workflow or have related study in provenance, literatures relevant to provenance study, notes of the experiment and so on. This is the idea behind a "web of science" as proposed by Hendler in [Hendler02].

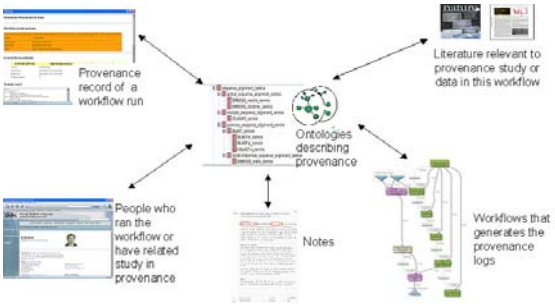

Figure 3: a Web of Experimental Holdings connected through shared concepts forming "semantic glue"

So we should like to link, browse and annotate provenance documents and external documents based on concepts.

3. Conceptual Open Hypermedia

The COHSE (Conceptual Open Hypermedia Services Environment) integrates three technologies to form a conceptual open hypermedia system for the web. An *Ontology Service*, using rich knowledge representation techniques and reasoning, gives machine processable semantics to the conceptual metadata associated with documents and between concepts. An *Annotation Service* annotates documents or sections of documents with a concept and maintains the mappings from concepts to annotated documents. A *Linking Service* generates target links for the concepts associated with web documents. The union of the three is a web-based authoring and browsing environment that we proposed would improve the quality, consistency and breadth of linking of web documents [Carr01].

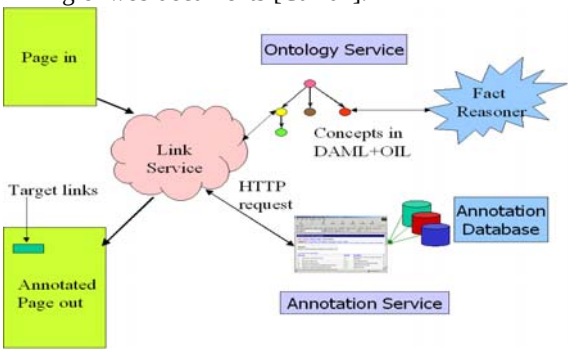

Figure 4. The COHSE architecture.

COHSE provides conceptual linking for documents based on the concepts associated. These intensional (rather than extensional) links need a sophisticated content understanding mechanism not only to recognize

the things in the documents but also to understand the relationships between them. The COHSE Agent has the basic task of generating and presenting links to web pages on behalf of both authors and readers. There are two types of mechanisms to get to the concepts that will form the link source anchors within a page:

i. Through the use of some language terms within the document and their mappings, held in a lexicon, to the concepts with an ontology provided by the Ontology Service;
ii. Through the use of ontological annotations on regions of the document provided by an annotation service. Link targets are found via the Annotation Service, which maintains mappings between resources and the concepts in the ontology. The ontologies are represented in DAML+OIL, built with ontology editor OilEd [2]. The concept to resource mapping provides candidate targets for links (as in many other resource discovery systems. The resource to concept mapping provides candidate source anchors for links. These mappings are between XPointer and a DAML+OIL expression, stored in a link-base held as RDF and implemented by a mySQL database. Figure 4 gives the COHSE architecture.

The COHSE agent can be embedded within a specialized web browser such as Mozilla or a proxy that supports any web browser; both implementations exist. The ontologies used are logic models for concepts and their relationships using the controlled, common sharing language DAML+OIL, which enables a logic reasoning service on the top of concepts as well as a shared communication and understanding across people and computers.

4. COHSEing provenance documents

In order to realize the annotation and linking of provenance documents, we needed to do three things:

1. Export the provenance documents into the COHSE environment; that is extract the XML documents from the mIR;
2. Prepare ontologies for annotating the documents and related web pages;
3. Annotate the documents with concepts drawn from the [my]Grid ontology.

4.1 Acquiring an Ontology

We used the [my]Grid generic and domain specific ontologies annotating the provenance documents.

The ***Generic Ontology*** mainly includes concepts about organization, users and publishing, providing general linking between provenance documents, and links between the related users who executed the workflows based on the relationships of their organizations or research topics.

[2] http://oiled.man.ac.uk

The ***Domain Ontologies*** include concepts about bioinformatics and services in [my]Grid workflows as well as biological concepts, mostly coming from the [my]Grid ontologies.

4.1 Acquiring Annotations

The provenance logs produced by FreeFluo have no associated explicit concepts. This is because the enactment engine is intended as a general purpose engine. We were required to post annotate the documents with the corresponding concepts either through manual annotations or a deep annotation process.

Manual annotation: We apply handcrafted annotations to the provenance documents. The scientist recognizes the semantic concepts for data and services in the logs and manually annotates these data with the corresponding concepts. However, the authoring process is quite time consuming.

Automatic annotation: We need to automatically automate the provenance documents with concepts. Semantic concepts about services and data are provided in the ws-info documents (figure 5) which, as we described in section 3.1, provide appropriate semantic and mime type information for workflow and service inputs and outputs during workflow execution. These ws-info files are linked to the provenance logs in the [my]Grid Information Repository. Hence we can recognize the concepts about the data and services from the ws-info files and associate the concepts with these data (Figure 1, dashed-boxed area). In this way we have entries into the concepts not only with the language terms but also with the instances of concepts. This annotation is now done post the workflow run and pre the provenance document's archiving in the mIR as part of the [my]Grid workflow pipeline. Consequently, we can export the provenance log as an already annotated document into COHSE.

Figure 5: Corresponding ws-info file for provenance log document.

4.2 Using COHSE

We now show COHSE in action over our provenance documents. The concepts `lymphocyte` and `neutrophil` are both subsumed by the concept `white blood cell` in the Domain Ontology. Figure 6 shows a provenance document (A) that includes an input to the service *AffyMetrixMapper* that is a *ProbeSetId* that has been annotated by the concept `lymphocyte`. When the scientist clicks on the annotation icon (a "C" icon) next to the link anchor, the links that are generated are to other documents that also annotated as `lymphocyte`. The "More General Links" refer to other documents labelled with subsuming concepts, here `white blood cell`. On (B), a link anchor is generated for the subsuming concept (`white blood cell`). Links to documents annotated with more specific concepts (`lymphocyte` and `neutrophil`) are displayed as "More specific Links" in the popup window.

Figure 7 shows that other kinds of documents can be annotated and linked into the provenance logs with the help of the Generic Ontology. The web page of the *Institute of Human Genetic in the University of Newcastle* is linked to the provenance logs based on the common annotated concept `Human Genetics.` Also links to some other human genetics related literatures are provided for the `Human Genetics` link anchor.

These two figures also demonstrate different views of linking between documents due to different ontologies applied for conceptual linking. As introduced above, we used two ontologies in this project. By choosing one ontology for conceptual linking each time, different link anchors are recognized by the Linking Service in COHSE and different target links are provided for different concepts.

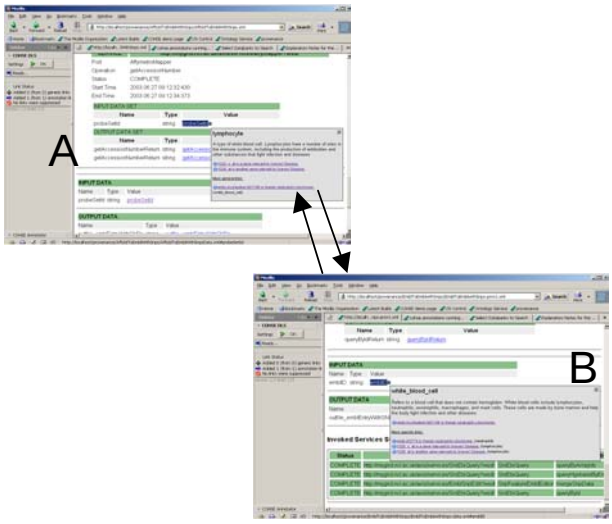

Figure 6: Generated links between provenance documents

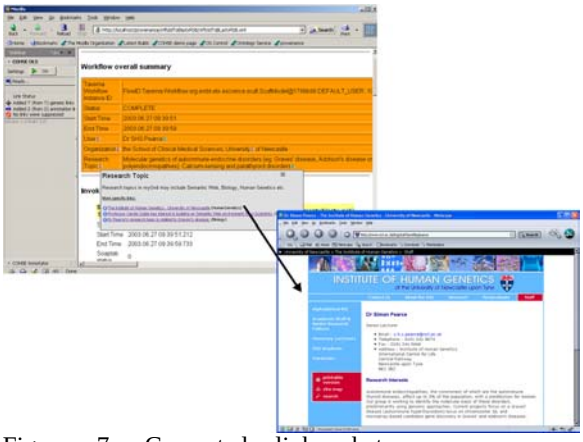

Figure 7: Generated links between provenance documents and other kinds of documents

5. Related Work

Geodise is a project concerned enhancing the engineer design process and optimization through knowledge management and ontology techniques, such as ontology construction and semantic annotation. It links and annotates engineer design logs with domain and task ontology, which is similar to what we are doing with the provenance documents in myGrid project. Geodise realizes a kind of automatic annotation through the support from its ontology editor OntoMat [Handschuh02], which enables the reasoning service based on instances of concepts, and the ontological-driven data management [Chen03]. Geodise does not intend to provide an open hypermedia environment to end users, that is, it does not provide target links in the annotated pages as what COHSE can do. Instead it stores the semantically enriched logs to the knowledge repository and emphasizes more on knowledge query and reuse. OntoMat does not support annotating a region of the resource pages, instead it can only realize annotating a whole log page with a concept or an instance of concept.

Our work is aimed at providing an open hypermedia environment for e-Scientists by providing knowledge discovery ability through the reasoning service of the ontologies. Thanks to the workbench in myGrid, we can realize the automatic annotation process by extracting the semantics simultaneously with the generation of provenance documents.

There is research work involving in the provenance archiving and retrieval. The PENN Database Research Group applies an approach for data archiving based on the key constraints. This approach uses a time-stamp for persistent records that appear in different versions of documents [Buneman02]. It is based on the assumption that most databases for scientific data employ a well-organized schema, which has key constraints for their data. The application of key constraints for provenance archiving can be especially applicable for well-structured scientific data and save significant space by converging persistent data using time-stamp. You can trace the

provenance of data by checking the time-stamping value of the record.

ESSW (Earth System Science Workbench) [Frew01] provides an approach similar to the FreeFluo workflow enactment engine to log experiments and their relationships. It generates an experiment based on a metadata template and adds each new experiment to an "Experiment" table along with the metadata about their inputs and outputs to the "ScienceObject" table in their relational database. However up until now no EESW tools exploited their metadata base to support the reuse and justify of their experiments.

6. Discussion

Early results are promising. The COHSE linking service offers a rich browsing environment for complex scientific data. By clinking on an annotation we are able to navigate to other workflow records semantically associated with that concept, or any other document such as personal notes, papers or the home page of the author of the experiment. Thus we hope to automate and simplify the process for biologists to find the related provenance documents for certain scientific data generated from previous experiment process or by other colleague and to be informed with changes and improvements in related research areas from the annotations provided by colleagues.

Work is underway to more extensively populate the COHSE document and link bases and undertake some formative evaluations with our scientific collaborators.

The automated “deep” annotation of the logs has been effective in associating concepts with their instances in the logs. We realised that much of the information was available to undertake this but was effectively “thrown away” once the workflows were executed. Whilst attempting to automate the annotations we came to realise that the bioinformatics ontology, describing the inputs and outputs of the services and workflows, only provided very simple link opportunities for the logs. On reflection this is obvious – the information relates to bioinformatics *types* such as EMBL_record or ProbeId, independent of the data inputs and outputs themselves. It is the data that offers riches in description. Bioinformatics concepts are crucial for typing the inputs and outputs of workflows, which is what they are used for. The biological ontology provides a more deep annotation to data using a more comprehensive and complex domain knowledge than bioinformatics, which enables rich reasoning opportunities between the data annotated. However, that a ProbeID is actually a lymphocyte requires either hand annotation or sophisticated and specific automatic annotation at some point in the provenance document annotation pipeline. The investigation of “painless” or somehow incidental manual annotation is high on our agenda and a real challenge.

COHSE is a standalone document browsing environment. Provenance records are immutable, so there is no problem with inconsistencies with the mIR. However, we could integrate the browsing facilities into the myGrid workbench.

Life Science Identifers (LSID) [LSID] are persistent, location-independent, resource identifiers for uniquely naming biologically significant resources including but not limited to individual genes or proteins, or data objects that encode information about them. myGrid has adopted LSIDs as a unique naming scheme for data objects in external databases as well as objects in the mIR. The addition of an LSID resolver in COHSE would make it a portal interconnecting the original databases and the mIR entries via their provenance documents.

Acknowledgements

The 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; and Yeliz Yeslida for her help in getting COHSE up and running.

References

[Buneman02] Buneman P, Khanna S and Tajima K and Tan W-C Archiving Scientific Data Proceedings of ACM SIGMOD International Conference on Management of Data 2002

[Carr01] Carr L., Bechhofer S., Goble C.A., Hall W. *Conceptual Linking: Ontology-based Open Hypermedia*. WWW10, Tenth World Wide Web Conference, Hong Kong, May 2001

[Chen03] Chen L, Shadbolt N.R., Goble C, Tao F, Cox S.J, Puleston C, Smart PR Towards a Knowledge-based Approach to Semantic Service Composition in proceedings 2nd International Semantic Web Conference, Florida, USA, October 2003

[DAML-S02] The DAML Services Coalition DAML-S: Web Service Description for the Semantic Web, *The First International Semantic Web Conference (ISWC)*, Sardinia (Italy), June, 2002.

[FreeFluo] http://freefluo.souceforge.org

[Frew01] Frew, J. and Bose, R Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products the 13th International Conference on Scientific and Statistical Database Management, Fairfax, VA, 2001

[Goble03] Stevens R, Robinson A, and Goble C myGrid*: Personalised Bioinformatics on the Information Grid* in proceedings of 11th International Conference on Intelligent Systems in Molecular Biology, 29th June–3rd July 2003.

[Handschuh02] Handschuh S, Staab S Authoring and Annotation of Web Pages in CREAM in Proc 11th Intl Conf WWW 2002, Hawaii

[Hendler02] James Hendler, Science and The Semantic Web, Science, Jan 24, 2003

[Horrocks02] Horrocks I. DAML+OIL: a reason-able web ontology language. In *Proc. of EDBT 2002*, March 2002.

[LSID] http://www.i3c.org/wgr/ta/resources/lsid/docs/index.htm