Ontology Driven Dynamic Linking of Biology Resources

S.K. Bechhofer, R.D. Stevens, and P.W. Lord

Pacific Symposium on Biocomputing 10:79-90(2005)

ONTOLOGY DRIVEN DYNAMIC LINKING OF BIOLOGY RESOURCES

S. K. BECHHOFER, R. D. STEVENS AND P. W. LORD

Department of Computer Science
University of Manchester
Oxford Road
Manchester
UK, M13 9PL
E-mail: seanb@cs.man.ac.uk

Biologists were early adopters of the Web and continue to use it as the primary means of delivering data, tools and knowledge to their community. The Web is made by the links between pages, yet these links have many limitations: they are static and maintained by hand; they can only link one lexical item to another single resource; ownership is necessary for the placement of link anchors and the link mechanism is essentially inflexible. Dynamic linking services, supported by ontologies, offer a mechanism to overcome such restrictions. The **C**onceptual **O**pen **H**ypermedia **Se**rvice (COHSE) system enhances web resources through the dynamic addition of hypertext links. These links are derived through the use of an ontology and associated lexicon along with a mapping from concepts to possible link targets. We describe an application of COHSE to Bioinformatics, using the **G**ene **O**ntology (GO) as an ontology and associated keyword mappings and GO associations as link targets. The resulting demonstrator (referred to here as GOHSE) provides both glossary functionality and the possibility of building knowledge based hypertext structures linking bioinformatics resources.

1. Introduction

This paper investigates the use of ontology driven open hypermedia within bioinformatics. Using this technology it is possible to separate links from document resources and consequently provide a rich, dynamically linked collection of biology oriented documents. By driving the dynamic formation of links through an ontology we are taking advantage of the common understanding provided by an ontology. The relationships between the concepts in the ontology add a further dimension to the flexibility of linking, which can help to enhance the Web, the primary mechanism used within biology for delivery of data, tools and knowledge. As a discipline, bioinformatics relies on the knowledge held within its documents (Web pages,

database entries, books or articles). Query by navigation, via links between these documents and others is still fundamental to practical bioinformatics. It is the links between biology documents that provide the utility to both humans and machines. Common usage of the Web involves embedding links within documents. There are, however, a number of limitations to this approach, that can be extended to other representations of linked documents, such as PDF.

Hard Coding: Links are hand-crafted and hard coded in the HTML encoding of a page. An anchor is placed around the source object in the originating page and the location of the end-point is included in the link. This end-point, or target, of a link can be a page, an anchor placed within a page, or perhaps some dynamically evoked service such as a query. The link is a static, inflexible entity that is intimately bound with the source node.

Format Restrictions: Documents need to be written in a particular format (e.g. HTML or PDF) in order to support the addition of links.

Ownership: Ownership of the page is required in order to place an anchor in a page. It is, of course, possible to point to targets on other pages without ownership, but in order to insert a link source anchor, ownership is required.

Legacy resources: It can be difficult to deal with legacy material – when the view of a world changes, old pages might need to be updated with new links.

Maintenance: There is a weight of maintenance in creating and updating links in pages. This is due in part to the hard coding and ownership issues described above.

Link targets: Current Web links are restricted to point to point linking; there is only one target. Web links are essentially unary with no explicit inverse link (although browsers offer a "back" button that will take the user back to the originating point). Binary or n-ary links would allow greater flexibility in linking by offering more choice of targets for each link.

Dynamic linking services, supported by ontologies, offer a mechanism to overcome such restrictions. The Conceptual Open Hypermedia Service (COHSE)[3] system enhances document resources through the dynamic ad-

dition of hypertext links. These links are derived through the use of an ontology and associated lexicon along with a mapping from concepts to possible link targets. We describe an application of the COHSE architecture to Bioinformatics, using the Gene Ontology (GO)[12] as an ontology and GO associations as link targets. The resulting demonstrator (referred to here as GOHSE) provides both glossary functionality and the possibility of building dynamic hypertext structures linking bioinformatics documents. Ontology driven dynamic linking offers a vision of biology documents dynamically linked to multiple resources based on a common understanding of a domain based upon ontologies.

The following scenario describes the added value that the COHSE system can provide. A biologist is reading a Web page about cellular structure, e.g. `http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/C/CellularRespiration.html`. When viewed using a traditional browser, she will see static links (as inserted by the author) contained within the page. She then employs the COHSE agent (making use of the cellular component of GO) to assist browsing. Now, as well as the static links contained within the page, the lexical items within the page corresponding to GO cellular component terms or their synonyms are also highlighted as link sources – for example, the term "cytochrome c oxidase", which is a synonym of the GO term respiratory chain complex IV (sensu Eukarya) [GO:0005751]). Next to this highlighted term is an icon indicating that a definition is available. She clicks on this icon and sees a pop up definition of the term, taken from the GO. Also shown are a number of further link targets, taken from a link base, offering her a range of resources related to the term, such as `COX3_YEAST` in the SwissProt[11] database. In addition, if the number of resources found for the term is below a threshold, the agent will use the taxonomic structure of the GO cellular component ontology to find more general or more specific terms that may provide appropriate resources. Clicking upon `COX3_YEAST`, she is taken to the appropriate UniProt/SWISS-PROT entry. This resource can itself then be dynamically linked to terms etc. within the GO cellular component ontology.

In the following sections, we discuss how this scenario is realised. In Section 2 we describe the Gene Ontology and how it makes a suitable resource for the open, dynamic linking of biology document resources. We then introduce the COHSE system in Section 3 and describe the combination with GO in Section 4. We conclude with discussion and pointers to future work.

2. GO

The Gene Ontology (GO)[a] is a collaborative effort to address the need for consistent descriptions of the major attributes of gene products in different databases[12]. Figure 1 shows a portion of the cellular component ontology from GO. Each term has an associated textual definition describing the term along with subsumption and partitive relationships with other terms in the ontology. Each term within GO also has associated synonyms that represent alternative, equally valid terms for the concept within the ontology. In addition, a number of mappings between other vocabularies or classification systems and GO are available[b].

```
respiratory chain complex IV (sensu Eukarya)
 Accession: GO:0005751
 Aspect: cellular_component
 Synonyms:
          o cytochrome c oxidase
          o GO:0005752
 Definition:
          o A part of the respiratory chain, containing the 13
          polypeptide subunits of cytochrome c oxidase, including
          cytochrome a and cytochrome a3. Catalyzes the oxidation of
          reduced cytochrome c by dioxygen (O2). Found in eukaryotes.
 hierarchy
 * GO:0003673 : Gene_Ontology ( 146200 )
   o GO:0005575 : cellular_component ( 79199 )
     + GO:0005623 : cell ( 56534 )
       # GO:0005622 : intracellular ( 46101 )
       * GO:0005737 : cytoplasm ( 35977 )
         o GO:0005739 : mitochondrion ( 12311 )
          + GO:0005740 : mitochondrial membrane ( 979 )
            # GO:0005743 : mitochondrial inner membrane ( 775 )
             * GO:0005746 : mitochondrial electron transport chain ( 211 )
               o GO:0005751 : respiratory chain complex IV (sensu Eukarya) ( 54 )
       * GO:0045277 : respiratory chain complex IV ( 55 )
         o GO:0005751 : respiratory chain complex IV (sensu Eukarya) ( 54 )
      # GO:0016020 : membrane ( 13431 )
       * GO:0019866 : inner membrane ( 803 )
         o GO:0005743 : mitochondrial inner membrane ( 775 )
          + GO:0005746 : mitochondrial electron transport chain ( 211 )
            # GO:0005751 : respiratory chain complex IV (sensu Eukarya) ( 54 )
       * GO:0005740 : mitochondrial membrane ( 979 )
         o GO:0005743 : mitochondrial inner membrane ( 775 )
          + GO:0005746 : mitochondrial electron transport chain ( 211 )
            # GO:0005751 : respiratory chain complex IV (sensu Eukarya) ( 54 )
```

Figure 1. GO Ontology fragment

Collaborating databases annotate their gene products with appropriate GO terms, providing the consistency of annotation needed for reliable querying of databases. All the entries from the 16 collaborating databases either contain GO identifiers (that map to GO terms) or their equivalent mappings to internal database keyword lists or GO synonyms. In addition, other Web pages about biology, articles in on-line databases such as

[a] `http://www.geneontology.org`
[b] See `http://www.geneontology.org/GO.indices.html`

Pubmed, etc. also use these GO terms, synonyms and keyword mappings. Finally, there are tools, such as the GO Amigo browser, that are also oriented towards the Gene Ontology and offer further mechanism for linking via the GO terminologies.

GO offers a huge resource of community knowledge, but no one organisation owns all the "documents" that use the GO vocabularies and definitions. The use of COHSE together with GO offers a mechanism by which any Web page or other document containing lexical items matching a GO term or any of its equivalents can be automatically linked to not only a definition of the term from GO, but also a legion of other resources, based upon GO. As a consequence, we can dynamically generate a rich, flexible web of biology resources.

3. COHSE

Detailed descriptions of COHSE[c] can be found elsewhere[3], but we give here a brief overview of the basic approach and architecture.

Open Hypermedia Systems[6,9] seek to solve some of the problems outlined in Section 1. Rather than embedding links in the documents, which is inflexible, links are considered *first class citizens*. They are stored and managed separately from the documents and can thus be stored, transported, shared and searched separately from the document itself. The Distributed Link Service (DLS)[4], developed by the University of Southampton is a service that adopts this approach, and provides dynamic linking of documents. Links are taken from a link base, and can be either *specific*, where the source of the link is given by addressing a particular fragment of a resource, or *generic*, where the source is given by some selection, e.g. a word or phrase. Documents and linkbases are dynamically brought together by the DLS, which then adds appropriate links to documents.

COHSE extends the DLS with *ontological services*, providing information relating to an ontology. These services include mappings between concepts and lexical labels (synonyms). For example, GO tells us that cytochrome c oxidase is a synonym of respiratory chain complex IV (sensu Eukarya (See Figure 1). In this way, the terms and their synonyms in the ontology form the means by which the DLS *generically* finds lexical items within a document from which links can be made. The services also provide information about relationships, such as sub- and super-classes – here respiratory chain complex IV (sensu Eukarya) is a sub-class of respiratory chain

[c] http://cohse.semanticweb.org

complex IV. The use of an ontology helps to bridge gaps[1] between the terms used in example web pages (e.g. in this case cytochrome c oxidase), and those used to index other bioinformatics resources, such as the more specialised or generalised GO terms. We can loosen the restriction of linking only to Eukaryote complexes, and consider also linking to all complexes.

COHSE thus extends the notion of generic linking – the key point being that the ontology provides the link service with more opportunities for identifying link sources. As the ontology contains the terms that inform the DLS about the lexical items that may become links, there is no longer a need to own the page in order to make the link from the source to the target – this is taken care of by the DLS. Furthermore, the effort in providing the source of links moves from the document author to the creator(s) of the ontologies that are used by COHSE. In this case, these are the creators of GO – an ontology supported by a wide community of biologists and database curators that form a consensus in the ontological representation of domain understanding.

The system is implemented as a *COHSE agent*, along with two supporting services: the *Ontology Service* (OS) and *Annotation Service* (AS). The agent augments documents with links based on the semantic content of those documents. The Ontology Service delivers ontological information (as introduced above) in a dynamic fashion[2] to the DLS. The Annotation Service associates concepts with resources and provides mechanisms for querying those associations.

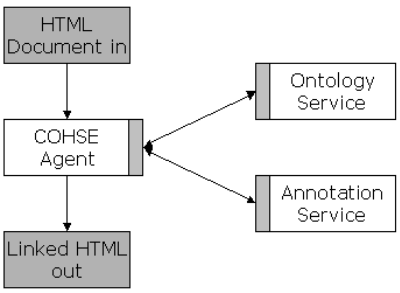

Figure 2. COHSE Architecture

In the implementation, the agent is attached to a proxy through which all HTTP requests are routed. The agent first contacts the OS to obtain a collection of relevant lexical items. As documents come through the proxy,

the agent then looks for these items. Any that are found in the documents provide potential link sources. For each source, a link is then added that includes:

(1) The concept to which the lexical item resolves (in the case of GOHSE this will be the GO term).
(2) A description of the term.
(3) A collection of link targets associated with that term.

Items 1 and 2 are supplied by the OS. The targets in 3 are supplied via calls to the AS. The concepts in the ontology are used to determine appropriate targets for links out of the given document. Within COHSE, the AS plays two roles. It maps resources or documents to concepts (via explicit annotations that have been made), and maps concepts to documents. It is this latter functionality that allows us to provide potential link targets once a link source has been found. For the concept to document mapping, we can either rely on a reversal of the explicit document to concept mapping, or provide targets through the use of external resources. For example, the AS provides potential link targets through queries to the GO database. Once a link source and associated GO term has been identified, we can query the GO database for proteins that have been annotated with that term, in the UniProt/SWISS-PROT database. Given the identifiers for those proteins, we can then produce URLs allowing browsing of those proteins via UniProt's web front end. In addition, we can provide links to the AmiGO or MGI GO browsers. Alternative mechanisms that we have explored for target retrieval include using external search engines (such as Google or the Amazon catalogue) with the query being based on keywords associated with a concept.

Central to the COHSE agent is the provision of an *editorial component* within the agent. This component uses information within the ontology (such as hierarchical classification) in order to either determine whether the generated links are suitable or to expand or cull the set of possible targets. Figure 2 shows a simplified view of the basic architecture of the system.

Both OS and AS are presented to the COHSE agent using simple CGI interfaces. This allows the system to make use of existing protocols and results in a relatively lightweight, loosely-coupled, and open architecture. In this demonstrator, we use a proxy to give access to COHSE. There are potential disadvantages in terms of scalability, but the use of a proxy allows users to access the demonstration without requiring local installation

of software or browser plugins and avoids problems of developing bespoke implementations for different browsing platforms.[d].

4. COHSE+GO = GOHSE

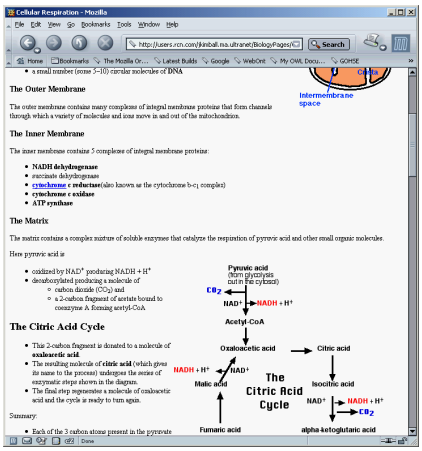

Figure 3. Before Proxy Linking

For the purposes of the GOHSE demonstration, the cellular component hierarchy of GO provides the ontology, while link targets are derived from the GO annotations in a variety of resources.

The concept taxonomy, along with term synonyms, is loaded into the COHSE OS. Annotation retrieval in the AS is implemented by returning UniProt/SWISS-PROT GO associations. For any given GO term, the AS returns URIs providing access to a number of potential targets:

- The AmiGO browser focused on the term.
- The Gene Ontology Browser focused on the term
- Any UniProt/SWISS-PROT entries known to be annotated with the term.

Both the ontology and the annotations are obtained dynamically from a (local) copy of the GO database. The ontology is produced via an on-the-fly translation to the recently developed W3C Web Ontology Language

[d]A Mozilla plug-in providing the COHSE agent functionality is, also available. See `http://cohse.man.ac.uk/` for details of available software.

OWL[8] (the format expected by the Ontology Service) and annotations are obtained via appropriate queries.

To use the system, the user first configures their web browser to use the GOHSE proxy. The user can then set up appropriate options concerning the behaviour of the proxy, including the number of times an individual link source should be identified. When the user requests a particular Web page, the proxy will attempt to find words or terms in the page that correspond to GO terms (via the lexical information held in the OS). For any matches to terms or term variants found, a link is added to the page providing access to the term and targets as described in Section 3. This presentation of the information is as a "linkbox" which pops up when the link is selected.

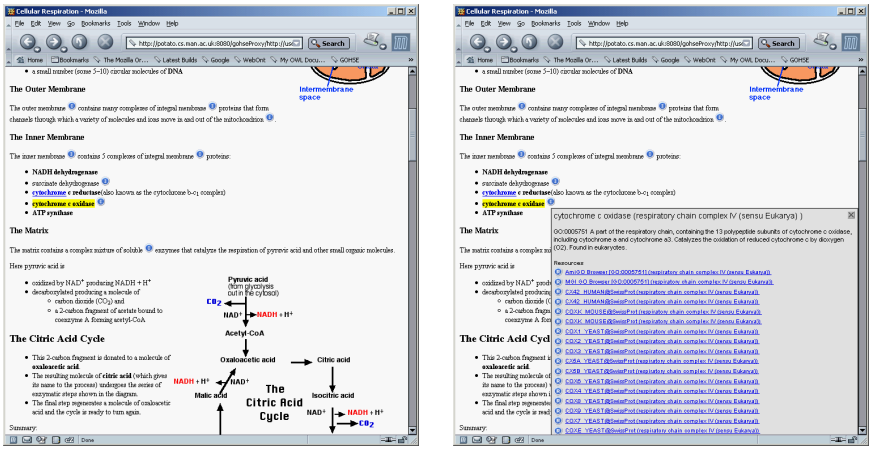

Figure 4. After Proxy Linking

The resulting additional links can help in providing both explanations of relevant terms and links to relevant materials. A key aspect of the system is its open-ness – we do **not** need to have resources under our control in order to add links to them. Thus third-party resources can be enhanced with additional links, allowing us to construct hypertexts using existing web resources. Figures 3 and 4 show the effects of the proxy – the scenario here is exactly as described in Section 1. In Figure 3 we see the original web page. In Figure 4 (left hand image), terms have been identified and marked (with small icons). On clicking on an icon (the right hand image) we see that the term maps to the GO term respiratory chain complex IV (sensu Eukarya) [GO:0005751]), and that a number of resources relating to

this are available.

5. Discussion

We have described an application of the COHSE infrastructure to support ontology driven browsing of biology document resources on the Web. In particular, the *dynamic* nature of the linking process helps to alleviate some of the problems with traditional Web linking, which can be static, restricted and inflexible.

Linking is based upon a conceptual model provided by an ontology, where the definitions and structure of the ontology, together with the lexical labels drive the consistency of link provision and dynamic aspects of the linking. The Gene Ontology has already been developed and encapsulates a great deal of shared knowledge about important concepts in the domain. Although this was not necessarily the purpose for which GO was designed, by using GO within this application, we are able to access a wealth of domain knowledge and gain added value "for free". The ontological resource that drives the linking of documents has already been created and its existence independent of the documents it links means that linking is consistent between documents (for a given version of the ontology).

The use of GO in this context illustrates one of the benefits of the Semantic Web approach: a computationally amenable representation of the content and facilities of documents and services[7]. GO provides an encoding of some domain knowledge (concept synonyms and taxonomy) in a *machine processable* fashion. By making this information available to applications, we are able to use this to support the presentation and browsing of resources. Note also that the approach used in COHSE is generic – we are not bound to the use of the Gene Ontology, but could use any other ontology appropriate for the domain, for example MGED[10]. Indeed, given an ontology relating to *any* domain, we can use COHSE to dynamically link suitable resources.

We note that the separation of the maintenance of link data from the underlying content enables us to manage the task of updating databases (or at least their web representations) as new knowledge becomes available. For example, while UniProt/SWISS-PROT links directly to the Gene Ontology, its predecessor, SWISS-PROT, did not. Using GOHSE, we can synthesize these links before the underlying data source provides them. Similarly, by extending GOHSE to recognise UniProt/SWISS-PROT identifiers, we can link between free text resources, such as PubMed, GO concepts, and the

underlying protein data sources. This feature of open hypermedia systems in general, and GOHSE in particular, is of particular relevance to biological data where cross-linking is known to be fragile[13]. As a discipline, bioinformatics relies on access to knowledge held in its databases. A system such as COHSE, augmented by ontologies such as GO, provide a knowledge model to drive the linking of diverse, distributed resources according to that knowledge.

It is clear that one of the reasons that this approach works here is because we have a well-defined domain of interest, a community and (to a certain extent) agreement on the important terms and concepts within that community. This is where we believe Semantic Web technology will have its initial successes – within well-defined communities.

In the current implementation, the identification of potential link sources is done in a rather naive fashion – effectively through a straight lexical match. In the GOHSE setting, this produces reasonable results (from the technical point of view), largely because (as discussed above), GO tries to use terms commonly used in the domain and these provide a clear set of lexical items to act as potential link sources. We can, however, encounter problems when, for example, formatting information is included in the source, making the identification of lexical items harder. We are investigating the use of the GATE[5] framework in order to gain access to effective text processing components. These will provide the DLS greater flexibility in its use of the terms provided by the OS.

Once concepts have been identified within the page, navigation of the ontology is driven by the COHSE Agent rather than the user. The agent decides if sufficient link targets are available, and whether or not to traverse the hierarchy to obtain more candidates. It may be more profitable to allow the user to explicitly navigate or explore the ontology at this point. However, there are then questions as to how one exposes the ontological structure to the user. In a similar vein, COHSE is clearly a system in which *personalization* can play a part – different users will want to use different ontologies or annotation collections. The current architecture is rather inflexible in this respect, and we are investigating support for more effective personalization.

GOHSE does not provide us with any new knowledge – it simply allows us to organise and present what is already there. Nor is it, at present, a particularly sophisticated implementation and improvements can certainly be made. It does, however, allow us to link together diverse biology resources, including those not in our control, in a consistent fashion based

upon a community understanding of the domain.

Acknowledgments

Phil Lord is supported by the myGrid EPSRC E-science pilot (EPSRC GR/R67743). The original COHSE system was developed in collaboration with the University of Southampton. The authors would like to thank John Kimball for permission to use his pages in our examples.

References

1. Marcia J. Bates. Indexing and Access for Digital Libraries and the Internet: Human, Database and Domain Factors. *JASIS*, 49(13):1185–1205, 1998.
2. Sean Bechhofer and Carole Goble. Delivering Terminological Services. *AI*IA Notizie, Periodico dell'Associazione Italiana per l'intelligenza Artificiale.*, 12(1), March 1999.
3. L. Carr, S. Bechhofer, C. A. Goble, and W. Hall. Conceptual Linking: Ontology-based Open Hypermedia. In *Proceedings of WWW10, Tenth World Wide Web Conference*, Hong Kong, May 2001.
4. L. Carr, D. De Roure, W. Hall, , and G. Hill. The Distributed Link Service: A Tool for Publishers, Authors and Readers. *World Wide Web Journal*, 1(1):647–656, 1995.
5. H. Cunningham, D. Maynard, K. Bontchev, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, July 2002.
6. K. Grønbæk, L. Sloth, and Orbæk P. Webvise: Browser and Proxy Support for Open Hypermedia Structuring Mechanisms on the WWW. In *Proceedings of the Eighth International World Wide Web Conference*, pages 253–268, 1999.
7. J. Hendler. Science and The Semantic Web. *Science*, page 24, Jan 2003.
8. D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview. W3C Recommendation, World Wide Web Consortium, 2004. `http://www.w3.org/TR/owl-features/`.
9. K. Osterbye and U Wiil. The Flag Taxonomy of Open Hypermedia Systems. In *Proceedings of the 1996 ACM Hypertext Conference*, pages 129–139, 1996.
10. Chris Stoeckert and Helen Parkinson. The MGED Ontology: A framework for describing functional genomics experiments. *Comparative and Functional Genomics*, 4(1):127–132, 2002.
11. SWISS-PROT Annotated Protein Sequence Database. `http://www.expasy.org`.
12. The Gene Ontology Consortium. Gene Ontology: a tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
13. Jonathan D. Wren. 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics*, 20(5):668–672, 2004.