

Document Navigation: Ontologies or Knowledge Organisation Systems?

Simon Jupp^{*1}, Sean Bechhofer¹, Patty Kostkova², Robert Stevens¹, Yeliz Yesilada¹

¹ *University of Manchester, Oxford Road, UK*

² *City eHealth Research Centre, City University, London, UK*

Email addresses:

Manchester: first.last [at] manchester.ac.uk

London: patty [at] soi.city.ac.uk

Abstract

Bioinformatics relies heavily on web resources for information gathering. Ontologies are being developed to fill the background knowledge needed to drive Semantic Web applications. This paper discusses how formal ontologies are not always suited for document navigation on the web. Converting ontologies into a model with looser semantics, allows cheap and rapid generation of useful knowledge systems. The message is that ontologies are not the only knowledge artefact needed; vocabularies and other classification schemes with weaker semantics have their role and are the best solution in certain circumstances.

Introduction

Navigation via hypertext is a mainstay of the World Wide Web (WWW). The author owned and unary links of standard HTML often neither offer the link sources nor targets needed by a particular group. Conceptual hypermedia provides navigation between web resources, supported by a conceptual model. The content of the model is used to dynamically identify link sources in web documents, and also supply the link targets to relevant web-services. The field of bioinformatics relies heavily on web resources and the community is now rich in bio-medical ontologies that can be used to populate this conceptual model.

The ability to browse documents on the web via hyperlinks embedded in text is still a fundamental part of the information gathering process used by bioinformaticians. As successful as hypertext is, it is not without its limitations;

- **Hard Coding:** Links are hard coded into the HTML source of a document.
- **Ownership:** Ownership of the page is required to place links in pages.
- **Legacy:** Link target can be deprecated leaving invalid links on pages.
- **Unary targets:** The current web links are restricted to point-to-point linking; there is only one target.

Conceptual Open Hypermedia supports the construction of hypertext link structures built using information encoded in ontologies. Dynamic linking, supported by ontologies, offer a mechanism to help overcome some of these restrictions. The

Conceptual Open Hypermedia Service (COHSE)¹ (Carr 2001) system enhances document resources through the addition of hypertext links (see Figure 1). These links are generated based on a mapping between concepts found in the document and lexicons available from the ontology. Links can have multiple targets based on the type of concept identified and in addition the structure of the ontology facilitates navigation to further targets based on sub/super concepts asserted in the ontology.

The COHSE architecture has been demonstrated in several fields, the GOHSE (Bechhofer 2005) system was applied to bioinformatics using the Gene Ontology (GO) (Ashburner, Ball et al. 2000) as an ontology and GO associations² as link targets. The Sealife project³ is now looking to extend this work and provide an ontology that integrates many of the ontologies being developed in biomedicine, to aid query by navigation to both scientists and health care professionals in the study of infectious diseases.

One of the major obstacles at this stage is how to integrate all the necessary ontologies into a single model with appropriate semantics that suit navigation. We argue that the strict relationships held between concepts in ontologies are not well suited for navigational purposes. A thesaurus like artefact is better suited for this task, it allows us to capture relationships that are not formal or universal nor part of the integral definition of the term. Our goal is to benefit from the work being done in the bio-ontology community i.e. capturing specific domain knowledge, and bring this knowledge into a model that suits the application's needs.

The proposed solution is to convert relevant bio-ontologies, medical vocabularies, thesauri, taxonomies and other concept schemes into one large Knowledge Organisation System (KOS). The Simple Knowledge Organisation System (SKOS)⁴ is chosen as a model to hold this information. A use case from the Sealife project is used to demonstrate the application in the study of infectious disease.

Sealife Use Case

The Sealife project seeks to develop a series of browsers in the context of the Semantic Web and Semantic Grid. The grid offers an infrastructure for large scale *in silico* science via a large number of computational services. The Grid setting needs to be combined with the continuing presence and use of numbers of Web documents describing knowledge about biology. Ontologies and controlled vocabularies provide great benefits for describing and using their data. The Sealife browser aims to use these vocabularies and ontologies as description of knowledge in the life sciences to flexibly manage the inter-linking of these documents and services.

One example application is to provide dynamic hyper-linking of resources from the National electronic Library of Infection (NeLI)⁵ (Kostkova 2003) portal to other related resources on the web. NeLI is a digital library bringing together the best available on-line evidence-based, quality tagged resources on the investigation, treatment, prevention and control of infectious disease. Many documents on the NeLI site contain few, if any, hyperlinks to other resources on the web. It would take a large

¹ <http://cohse.cs.manchester.ac.uk/>

See <http://www.geneontology.org>²

³ <http://www.biotec.tu-dresden.de/sealife/>

⁴ <http://www.w3.org/2004/02/skos/>

⁵ <http://www.neli.org.uk/>

curational effort and cost to manually mark up these pages with links to other web resources. In addition to this problem, NeLI has a range of users; we want different link targets based on the kind of user browsing the NeLI site. COHSE can help to solve some of these problems. Terms from the ontologies can be used to identify concepts in web pages and create the link sources. Each link source can have multiple targets; the targets selected are tailored to suit the needs of each user group.

The ability to identify user groups is important. Users can range from members of the public, molecular biologists to clinicians and GPs. Each group has a different view of the bio-medical domain, and is therefore interested in different kinds of information. By providing alternative vocabularies for different users, the system can identify link sources relevant to that user and also provide multiple targets to relevant web resources. Table 1 shows four different user groups, some questions they might want answering and the different kinds of target sites a Sealife browser would offer them based on the type of user (Madle 2006).

The system is demonstrated with a simple use case involving a news site linking to NeLI. News sites are often the first to report on disease outbreaks via news feeds. Consider the scenario where a traveller is planning a trip to Namibia, only to find an article on the BBC website about a recent outbreak of Polio. COHSE can provide links to relevant resources that had not been included by the original author. Such resources could include information about the polio virus, its effect on humans, vaccination information and also geographical information about the local area. A family doctor, in contrast, might use a vocabulary skewed to their interests to link through to sites on drugs, details of symptoms and clinical presentations, treatment and local hospital facilities etc.

| User Group | Question | Targets |
|----------------------|--------------------------------------|--|
| Family Doctor (GP) | Tuberculosis drugs and side effects? | British National Formulary (BNF) |
| Clinicians | Tuberculosis treatments guidelines? | Public Health Observatories (PHO) |
| Molecular Biologists | Drug resistant tuberculosis species? | PubMed |
| General Public | What is tuberculosis? | Health Protection Agency (HPA) or the NHS direct online website. |

Table 1. NeLI users and example targets from the UK.

Figure 1 shows the system in action. The first image shows the original BBC article, the second shows dynamic links that have been added based on concepts held in the vocabulary. It also shows a link box that is dynamically generated when a link is clicked. The link box contains a textual description of the term and targets to multiple web resources. In addition to targets for the selected link the system can provide targets for broader, narrower and related resources. For example, NeLI has a web-service which takes terms from a NeLI vocabulary as inputs, this service is invoked when the “polio” link is selected and targets are returned which link to relevant documents from the NeLI portal. This simple demonstration shows how the addition of a navigational layer based on the semantic content of documents can be added to the existing web.

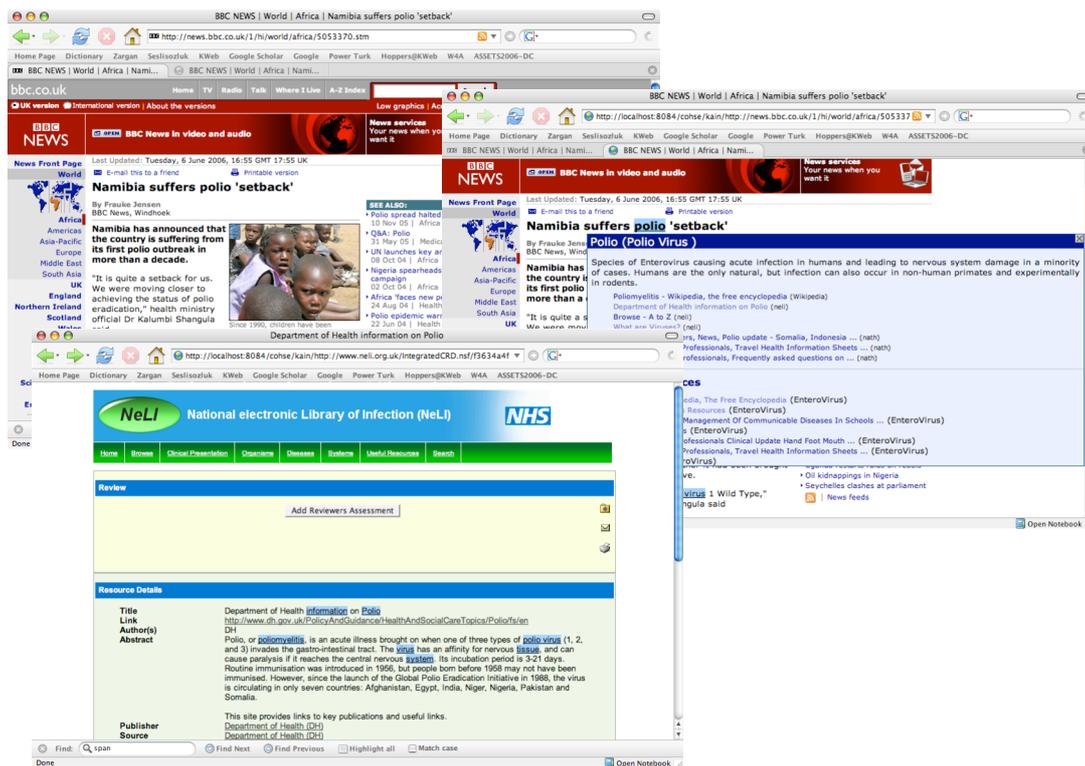


Fig. 1. Dynamic linking in action.

Gathering the background knowledge

For the Sealife browser to be useful across such a diverse subject as biology we, need a system to rapidly collect the current available resources together, and place them into a single representation that will facilitate navigation. The biomedical domain already has a rich collection of vocabularies and ontologies such as MesH⁶, UMLS⁷, GALEN⁸ and the OBO⁹ ontologies. There are also classification systems relating to genes, protein, drug and other terminological resources that would be useful to Sealife.

The languages used to represent ontologies vary considerably, and can range from simple taxonomy languages through to rich, formal logic based languages such as OWL. Increasingly strict semantics can remove ambiguity in the representation and facilitate the use of machine processing. Similarly, these languages can be used with varying degrees of ontological formality, not all OWL ontologies, for example, make rigorous ontological distinctions. Experience with COHSE has suggested that formal ontological distinctions and strict semantics are not *always* best suited to the task of navigating a collection of resources. Strict sub/super class relationships are not necessarily appropriate for navigation – rather, the looser notions of broader/narrower as found in vocabularies or thesauri provide the user with more appropriate linking.

SKOS is a model for representing classification systems, thesauri, taxonomies and other concept schemes. SKOS is currently undergoing standardisation by the W3C¹⁰

⁶ <http://www.nlm.nih.gov/mesh/>

⁷ <http://www.nlm.nih.gov/research/umls/>

⁸ <http://www.opengalen.org/>

⁹ <http://obofoundry.org/>

¹⁰ <http://www.w3.org/>

and has a RDF/XML representation that makes it well suited for semantic web applications. By representing the biological knowledge in SKOS we have a simple model that provides a lexical resource for identifying concepts in our documents, as well as a framework for asserting semantic relationships between concepts. SKOS has a set of properties that are well suited for supporting navigation. These include preferred labels, alternate labels (synonyms) and textual definitions for describing concepts as well as *'broader'*, *'narrower'* and *'related'*, for representing the relationships between concepts.

Converting ontologies to SKOS

The semantics of some biomedical terminologies are already relatively weak. A good example for such a terminology that is commonly used in medicine is the Medical Subject Headings (MeSH). The semantics of **A narrower B** simply means that users interested in **B** might also be interested in **A**. The MeSH terms found under accident include kinds of accidents – as expected (e.g. Traffic accidents), but also Accident prevention. This is not a good ontological distinction, but a valid one in the context of navigation and retrieval. In contrast in the Open Biomedical Ontologies (OBO), **A is-a B**, a common type of relationship in OBO ontologies implies that all **A**'s are also instances of **B**. This contrast in semantics means that conversions from MeSH into OBO are not possible without misinterpreting the intended semantics. Despite this, we see that many of the OBO ontologies share concepts with MeSH, especially the Disease Ontology¹¹. From a navigation point of view we would like to combine these resources to gain maximum benefit from efforts in MeSH and OBO development. By converting them both into SKOS we can use a single representation and use the lightweight semantics to build a larger and richer vocabulary.

The release of the 10 OBO relations (Smith 2005) gives OBO developers another level of expressivity in their ontologies. These relations have strong definitions with precise semantics; which are used to define relationships between terms in OBO ontologies. When converting OBO ontologies into SKOS we can use these relationships to assert *broader*, *narrower* and *related* relationships between SKOS concepts. Here is an example of the conversion one might make when mapping ontological properties to SKOS properties.

- rel:part_of -> skos:broader (e.g. finger part_of hand)
- rel:contains -> skos:narrower (e.g. skull contains brain)
- rel:has_name -> skos:related (e.g. Person has_name PersonName)

Another advantage when converting properties from ontologies to SKOS is the ability to assert the inverse. Consider an ontology where **Nucleus partOf cell**, from an ontological point of view this implies that every **Nucleus** is *partOf* some **Cell**. However, the inverse is not true, every **Cell** does not *havePart* **Nucleus**. When converting to a SKOS model we can assert the inverse using the *broader* property to say that **Nucleus** has a *broader* term called **Cell**, which is quite reasonable. When navigating around documents about cells, the system could then also provide links to documents about nuclei – users interested in cells are often also interested in nuclei.

If we use the polio use case example we can show that a great deal of information can be acquired about polio from the various vocabularies available. When the

¹¹ <http://diseaseontology.sourceforge.net>

semantics are strict we have to be very careful how we bring all this related information together. With all this information in SKOS, Sealife can benefit from many different knowledge resources. Table 2 outlines the results from searching polio against a varying set of ontologies and vocabularies alongside the SKOS property used to relate them.

| Source | Terms found | SKOS relation to "Poliovirus" |
|------------------|----------------------------|-------------------------------|
| MeSH | Brunhilde Virus | skos:altTerm |
| Disease Ontology | Spinal cord disease | skos:broaderThan |
| | Postpoliomyelitis Syndrome | skos:narrowerThan |
| SNOMED | Microorganism | skos:broaderThan |
| | Enterovirus | skos:broaderThan |

Table 2. Searching polio virus against different resources and converting intended semantics into SKOS semantics for navigation. (NLM 1960, Cote et al 1993)

There is, however, likely to be some trade off associated when bringing multiple resources together, it is possible that a lot of unwanted terms are returned, especially when using formal ontologies. Ontologies can benefit from an upper-ontology (Rector 2003) that contains abstract categories; these can be used to build formal definitions for classes. An ontological definition which states whether the concept is a physical or non-physical entity may be crucial to the design of a robust ontology, but is largely irrelevant from a navigational point of view. To overcome this we must remove these properties at the stage of conversion into SKOS, how we do the conversion from ontologies to SKOS is something for future work.

CONCLUSION

A large community of ontology developers and knowledge engineers is forming in the life sciences. It is hoped that they will deliver the infrastructure needed to realise a real semantic-web, where computers can begin to interpret and interoperate biological data automatically. If applications like Sealife are to demonstrate the early potential of Semantic Web technologies, then the trade off associated with relaxing the semantics in the background knowledge has to be acceptable.

The nature of formal ontologies can sometimes make it difficult to express relationships between concepts that experts from the domain would expect to find under some circumstances. Thesauri are much more suited to represent the way *words* and *language* are used in the field. The Sealife project will demonstrate how the effort and cost associated with building rich formal ontologies can also be used to feed into other knowledge artefacts, like thesaurae, vocabularies, classification schemes etc. in SKOS, which can then be used in different application scenarios.

ACKNOWLEDGEMENTS

Funding by the Sealife project (IST-2006-027269) for Simon Jupp is kindly acknowledged.

REFERENCES

- Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nat Genet* **25**(1): 25-9.
- Bechhofer, S. Stevens, R. Lord, P. (2005) *Ontology Driven Dynamix Linking of Biology Resources*. Pacific Symposium on Biocomputing, Hawaii.
- Carr, L. Bechhofer, S. Goble, C. Hall, W. (2001) *Conceptual Linking; Ontology-based Open Hypermedia*. WWW10, Tenth World Wide Web Conference, Hong Kong.
- Cote RA et al., eds. SNOME International: the systematized nomenclature of human and veterinary medicine. Vols I-IV. Northfield, IL, College of American Pathologists. 1993.
- G. Madle, P. Kostkova, J. Mani-Saada, A. Roy. (2006) *Lessons learned from Evaluation of the Use of the National electronic Library of Infection*, Health Informatics Journal, Special Issue, Healthcare Digital Libraries“, 12: 137-15
- National Library of Medicine. Medical subject headings: main heading, subheadings, and cross references used in the Index Medicus and the Natinal Library of Medicine Catalog. 1st ed. Washinton, DC:U.S. Department of Health, Education, and Welfare.1960
- P. Kostkova et al. (2003) *Agent-Based Up-to-date Data Management in National electronic Library for Communicable Disease*. In SI: "Applications of intelligent agents in health care", J. Nealon, T. Moreno Ed, in Whitestein Series in Software Agent Technologies, pages 103-122.
- Rector, A. (2003) *Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL*. Knowledge Capture, ACM, 121-128.
- Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C *Relations in Biomedical Ontologies*. *Genome Biology*, 2005, 6:R46