DCPAPERS

Architecting a Cross-Disciplinary Thesaurus for the Semantic Web

W. Davenport Robertson
National Institute of Environmental Health Sciences Library, Research Triangle Park, NC, USA
Mail:robert11@niehs.nih.gov

Jane Greenberg
School of Information and Library Science,
University of North Carolina at Chapel Hill, NC, USA
Mail:janeg@ils.unc.edu

Abstract: An environmental health science thesaurus is needed to facilitate Semantic Web operations and aid with problem solving in this important cross-domain area. This paper demonstrates the need for an environmental health science thesaurus and reviews metadata generation research that highlights the importance of subject metadata. A conceptual design for building a cross-disciplinary metathesaurus using shared ontologies and other Semantic Web technologies is presented. The design emphasizes a dynamic, distributed approach to thesaurus construction, and builds on Semantic Web developments, especially the integration of multiple ontologies. The paper concludes by identifying candidate terminological sources and identifying a major challenge.

Keywords: Semantic Web; thesaurus; metathesaurus; metadata generation; multi-disciplinary; decentralized semantics; ontology; environmental health.

1 Introduction

The field of environmental health is multi-disciplinary. It is the study of human health and disease as a reflection of three attributes: environmental factors, individual susceptibility (genetics), and age. Despite the overarching importance of this field to all people,, there is no single, comprehensive thesaurus for the study of environmental health. There are thesauri that provide vocabulary terms describing parts of the field (e.g. MeSH, Medical Subject Headings, http://www.nlm.nih.gov/mesh/meshhome.html), but in order to ensure that the general public can find accurate environmental health information on the Web site of the National Institute of Environmental Health Sciences (NIEHS), a comprehensive cross-disciplinary thesaurus is needed. Building upon previous work on metadata generation, we propose an architecture for a dynamically generated metathesaurus that would take advantage of other thesauri using a decentralized semantics approach.

2 Semantic Web Scenario

The popular and often repeated Semantic Web scenario described by Berners-Lee *et al* (**1**) focuses on scheduling a medical appointment. Semantic Web scenarios can include problem-solving in environmental health sciences and other fields that are multi-disciplinary in nature. To illustrate this point, let's say you have just found an old box of mothballs in your grandmother's closet. Should you leave them, throw them away, or what? You go online to your Semantic Web agent and ask, "What should I do with old mothballs?" It promptly retrieves information found at the NIEHS Web site saying that one of the active ingredients in mothballs has been shown to cause cancer, and the agent adds that there might be other serious effects of long-term exposure. The agent, knowing your Zip code, looks up recycling centers for hazardous waste and gives you the location and hours of operation for the one nearest to where you live. It asks if anyone has eaten any of the mothballs, and, if you answer Yes, the agent immediately connects you with a person at your nearest Poison Control Center for a live chat. If you answer No, the agent asks if anyone has had long-term exposure to the mothballs, and--depending on your answer--it supplies you with the name and contact information for a nearby environmental health clinician. Then the agent looks in your calendar and the clinician's calendar to set up an appointment.

https://doi.org/10.23106/dcmi.952107948

This hypothetical scenario would be accomplished through the use of standard enabling technologies such as RDF and OWL that facilitate the use of metadata. Implementation of an environmental health thesaurus is fundamental for providing the high quality subject metadata upon which this scenario is based.

3 Previous Work on Metadata

As reported in previous Dublin Core conferences, NIEHS and the University of North Carolina at Chapel Hill have been collaborating through the Metadata Generation Research (MGR) Project (http://www.ils.unc.edu/mrc/mgr_index.htm). The MGR project aims to improve metadata generation and enhance the discovery of NIEHS Web resources. A number of important research activities and findings of this work specifically focus on subject metadata. In fact, the need for *subject* metadata to facilitate resource discovery of NIEHS Web resources was a key reason for the development of an environmental health Dublin Core application profile (**2**). The motivation stems from examination of NIEHS Web logs, which showed the popularity of subject searches on the NIEHS Web site, but also showed user frustration from not typing in the appropriate terms or from misspelling. A well documented example is the misspellings for the word "tattoo" (e.g., tatoo, tatoe, etc.). An examination of author generated metadata found that authors would like more guidance when creating subject metadata for their resources (**3**). Research on collaborative metadata generation found that authors would also like cataloger assistance when creating *subject* metadata more than any other Dublin Core metadata element (**4**).

Recent resource discovery research also shows the importance of subject metadata (**5).** Preliminary findings from a recent user study showed that users were rarely satisfied when examining surrogates returned by Google, after searching the NIEHS Web site. This work includes an examination of user-selected document features, which aid in relevance decision making, and clearly indicates that subject metadata is key to document evaluation.

The importance of high quality subject metadata for accurate resource discovery cannot be exaggerated. To enhance current resource discovery activities and facilitate Semantic Web operations, the NIEHS Web site needs a high quality subject authority list--an Environmental Health (EH) thesaurus. This tool could be used by the content creators for selecting the terms to input for the subject metadata, and it could be used by Semantic Web agents to perform tasks, or by the public when searching the NIEHS Web site.

4 Needs Assessment for an Environmental Health Thesaurus

The NIEHS conducted a needs assessment and feasibility study for the creation of an Environmental Health (EH) thesaurus in the fall of 2003. The conclusion was that NIEHS has a strong need for a thesaurus and that no existing thesaurus covers the field of environmental health adequately for NIEHS. A key reason is the cross-disciplinary nature of the field. NIEHS is one of 27 national institutes and centers of the National Institutes of Health (NIH), but it is one of the few that crosses disciplinary lines. Most of the NIH Institutes are devoted to specific diseases or organ systems. Environmental health encompasses many diseases and any organ system, as well as genetics and aging. It also includes chemicals, occupations, lifestyles, and many other facets. Thus there is a major domain challenge when attempting to create any kind of ontology for environmental health. That notwithstanding, the field of environmental health is significant to our daily existence and the development of a standard vocabulary is crucial for ensuring our well-being.

To reach that goal, one recommendation of the needs assessment was to license selected thesauri, in whole or in part, and merge them into a centralized EH thesaurus on a single server. The result would be somewhat similar to the National Library of Medicine's Unified Medical Language System (UMLS) in structure (being made up of multiple individual thesauri, vocabularies, etc.), but not in subject coverage.

However, as explained below, another approach is to take advantage of Semantic Web technologies to create a dynamic and decentralized thesaurus that would assist both metadata generation and searching.

https://doi.org/10.23106/dcmi.952107948

5 Architecture of a Decentralized Thesaurus

5.1 Decentralized Semantics Approach

We propose an EH metathesaurus that takes advantage of Semantic Web technologies to encompass a variety of related subject ontologies, thesauri, and vocabularies created for other organizations and located on remote servers (Figure 1). This decentralized approach would depend upon sharing resources "owned" by heterogeneous communities. In this sense, the design depends upon the critical attribute of interoperability. The one centralized aspect of the metathesaurus is that it would contain a link repository for all of the component ontologies. In Figure 1, the Metathesaurus Ontology Server is queried for a term, and it sends out the query to the linked thesauri (A-F) resident on remote servers (1-6). They, in turn, reply with the appropriate terms, whether equivalent, hierarchical or associative.

The implementation of this design would have major implications for metadata generation. For example, at NIEHS a Web page author would begin generating metadata by using our input form. When the author gets to the DC Subject Element, he/she inputs a keyword and issues the command to look it up in the metathesaurus. The software agent then searches the component ontologies for that term and returns its findings for the author to review. The author selects the preferred term as appropriate and populates the subject field in the form with it.

Another area where this design would have a significant impact is in the information searching process. In this example, the author would have supplied in the metadata subject element field whatever keyword he/she thought of, without reference to the metathesaurus. But when someone from the public comes to the NIEHS Web site and inputs a natural language term, the metathesaurus search agent searches the component ontologies and retrieves all the equivalent terms from the thesauri. The agent executes the search form with these terms and—if one of them is the one the author used for indexing—the appropriate Web page is found.

Most importantly, this type of metathesaurus is fundamental to the Semantic Web in order to accomplish discovery operations like the mothball scenario presented at the beginning of this paper and others that cut across many disciplines. This design incorporates the interoperability principles for shared ontologies expressed in the W3C's *OWL Web Ontology Language Use Cases and Requirements* (http://www.w3.org/TR/webont-req/) and additional guides. It is complementary to design concepts espoused by Hendler (**6**) and others (**7**) who advocate the importance of multiple ontology integration for the success of the Semantic Web.

A significant advantage of a decentralized metathesaurus is that no member organization has to spend resources to update the overall contents (the terms and the syndetic structure). Each participating organization would continue updating its own component just as it normally would. Another advantage is that other organizations can develop additional metathesauri by selecting EH component thesauri and incorporating additional thesauri and ontologies. An example of another important cross-discipline where this would be helpful is the area of bioterrorism research.

5.2 Decentralized Semantics Requirements

Standards for implementing a thesaurus on the Semantic Web are under development. The Semantic Web Best Practices and Deployment (SWBPD) Working Group includes the publishing of ontologies/vocabularies as a focus area in its charter (http://www.w3.org/2003/12/swa/swbpd-charter). Its Thesaurus Task Force has outlined objectives (http://www.w3.org/2004/03/thes-tf/mission). These resources have potential for setting the relevant standards in the future, but they have barely started.

An example of a consolidated registry of disparate but related ontologies is The Open Biological Ontologies (OBO) site (http://obo.sourceforge.net/) which describes itself as "an umbrella web address for well-structured controlled vocabularies for shared use across different biological domains." It has links to many ontologies that conform to the specified standards and requirements.

https://doi.org/10.23106/dcmi.952107948

The criteria that are fundamental to our design for a decentralized metathesaurus on the Semantic Web are as follows:

- The components must be open, preferably open source.
- They must be listed in a standard registry.
- They will use shared enabling technologies and standards (**8**)
    - Uniform Resource Identifiers (URIs)
    - Resource Descriptor Format (RDF), RDFS, and XML
    - OWL Web Ontology Language

By adhering to these standards, information can be retrieved from the component ontologies in a consistent manner. There are mechanisms within OWL (equivalentClass, sameAs, differentFrom) that can be used for ontology mapping. Of course, software has to be written to perform the retrieval. The MindSwap Group at the University of Maryland has developed a software package called SWOOP (**9**) that appears to integrate ontologies and make them searchable as one. A useful example of deriving a domain ontology from a thesaurus using RDFS is reported by Lauser and others (**10**) in their description of the FAO's project using the AGROVOC multilingual agricultural thesaurus. A helpful resource for understanding ontologies and interoperability structures is Jacob's article (**11**).

6 Conclusion

From the foregoing examples, it is clear that shared ontologies can be used to create a new, comprehensive ontology. There is still much work to be done, not only in further designing this architecture for a metathesaurus but also in developing the Semantic Web standards and the necessary tools. Research into the feasibility of using a tool such as SWOOP for this purpose needs to be carried out. As we apply the decentralized design to the case of environmental health, we must choose the component domain thesauri that will comprise the EH metathesaurus. Some of the candidates and their primary areas of subject coverage are as follows:

- U. S. Environmental Protection Agency Terminology Reference System (and its incorporation of the General Multilingual Environmental Thesaurus, GEMETS), for environmental pollution and cleanup terminology. (http://www.epa.gov/trs/index.htm)
- Open Biological Ontologies, for genomic and biological taxonomic areas. (http://obo.sourceforge.net/)
- National Cancer Institute Metathesaurus (and its incorporation of MESH and the Unified Medical Language System, UMLS), for cancer and other disease terminology. (http://ncimeta.nci.nih.gov/indexMetaphrase.html)
- National Agricultural Library Agricultural Thesaurus, for additional biological terminology and for terminology involving farm workers. (http://agclass.nal.usda.gov/agt/agt.htm)
- National Biological Information Infrastructure Biocomplexity Thesaurus, for biological taxonomic terms.
- Getty Thesaurus of Geographic Names ® for geographic names. A proprietary thesaurus like this one would present certain obstacles in an open source environment. (http://www.getty.edu/research/conducting_research/vocabularies/tgn/)
- National Library of Medicine Haz-Map, for hazardous materials and for occupations. (http://hazmap.nlm.nih.gov/)
- National Library of Medicine Toxicology Glossary, for toxic effects terminology (http://sis.nlm.nih.gov/Glossary/main.html)

The major challenge in implementing a decentralized metathesaurus of this type is convincing the participating organizations to convert existing thesauri to ontologies that adhere to the prescribed standards for the Semantic Web. It will require significant effort to illustrate the benefits to them and eventually arrive at a consensus that this model will benefit them and serve their missions. If we are successful in applying this approach to the field of environmental health, other cross-disciplinary fields may be able to adapt it to serve their needs and contribute to the development of the Semantic Web.

https://doi.org/10.23106/dcmi.952107948

References

1. Berners-Lee T., Hendler J., and Lassila O. The Semantic Web. Scientific American, 284 (5):35-43, 2001.

2. Robertson, W. D., Leadem, E. M., Dube, J. and Greenberg, J. Design and Implementation of the National Institute of Environmental Health Sciences Dublin Core Metadata Schema. In DC-2001 Proceedings of the International Conference on Dublin Core and Metadata Applications 2001, Tokyo, Japan. Tokyo: National Institute of Informatics, pp. 193-199, 2001. http://www.nii.ac.jp/dc2001/proceedings/product/paper-29.pdf

3. Greenberg, J., Pattuelli, M. C., Parsia, B., and Robertson, W. D. Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. Journal of Digital Information, 2(2), Article No. 78, 2001-11-06. http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg/

4. Greenberg, J. and Robertson, W. D. Semantic Web Construction: An Inquiry of Authors' Views on Collaborative Metadata Generation. DC-2002: Metadata for e-Communities: Supporting Diversity and Convergence. In Proceedings of the International Conference on Dublin Core and Metadata for e-Communities, 2002, Florence, Italy. October 13-17. Firenze: Firenze University Press, pp. 45-52, 2002. http://www.bncf.net/dc2002/program/ft/paper5.pdf

5. Crystal, A. and Greenberg, J. (Draft) User-Centered Metadata: an Exploratory Study of Relevance Judgments and Metadata Usage on the Web. 2004.

6. Hendler, J. Agents and the Semantic Web. Preprint of paper published in IEEE Intelligent Systems Journal, 16 (2): 30-37, 2001. http://www.cs.umd.edu/users/hendler/AgentWeb.html

7. Noy, N. F. and McGuinness, D. L. Ontology Development 101: a Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001. http://www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html

8. Miller, E. and Swick, R. An Overview of W3C Semantic Web Activity. Bulletin of the American Society for Information Science and Technology, 29 (3):8-10, 2003. http://www.asis.org/Bulletin/Apr-03/MillerSwick.pdf

9. Kalyanpur, A., Hashmi, N., Golbeck, N., and Parsia, B. Lifecycle of a Casual Web Ontology Development Process. In Proceedings of the WWW2004 Workshop on Application Design, Development and Implementation Issues in the Semantic Web, May 18, 2004, 2004. http://www.mindswap.org/~aditkal/WWW04_COD.pdf

10. Lauser, B., Wildemann, T., Poulos, A., Fisseha, F., Keizer, J., and Katz, S. A Comprehensive Framework for Building Multilingual Domain Ontologies: Creating a Prototype Biosecurity Ontology. In Proceedings of the International Conference on Dublin Core and Metadata for e-Communities, 2002, Florence, Italy. October 13-17. Firenze: Firenze University Press, pp 113-123, 2002. http://www.bncf.net/dc2002/program/ft/paper13.pdf

11. Jacob, E. Ontologies and the Semantic Web. Bulletin of the American Society for Information Science and Technology, 29 (3):19-22, 2003. http://www.asis.org/Bulletin/Apr-03/Jacob.pdf

https://doi.org/10.23106/dcmi.952107948