HIVE: Helping Interdisciplinary Vocabulary Engineering

by Jane Greenberg, Robert Losee, José Ramón Pérez Agüera, Ryan Scherle, Hollie White and Craig Willis

EDITOR'S SUMMARY

HIVE (Helping Interdisciplinary Vocabulary Engineering) is an effort to automatically generate metadata for content, drawing descriptor terms from multiple vocabularies encoded as Simple Knowledge Organization Systems (SKOS). The effort is a response to the challenges of interoperability, cost and usability of multiple terminology sets often needed to adequately describe digital resources. By offering access to more than one vocabulary with useful descriptors for a broad domain, HIVE enables aggregating the best terms to describe resources and automatically apply metadata. HIVE offers knowledge management value for multidisciplinary digital collections while demonstrating the expanded potential use of SKOS. The initiative is headed by the Metadata Research Center at the University of North Carolina's School of Information and Library Science working with several institutional partners. Conferences and workshops are scheduled to inform interested developers and users, who are invited to try out, contribute to and evaluate the system.

KEYWORDS

SKOS
knowledge organization systems
interoperability
metadata
automatic categorization
interdisciplinarity

HIVE (Helping Interdisciplinary Vocabulary Engineering) [1, 2] is both a model and a system that supports automatic metadata generation by drawing descriptors from multiple Simple Knowledge Organization System (SKOS)-encoded controlled vocabularies. This *Bulletin* article introduces the HIVE initiative. Project goals, plans and system design are presented. The conclusion highlights next steps.

Introduction: KOS Challenges

Knowledge organization systems (KOS) encode rich knowledge structures. As metadata systems, they enrich resource description and aid information retrieval [3]. Although these benefits are well known, a host of historical problems hinder their general use. Existing challenges are magnified when simultaneous use of multiple KOS is desired for representing interdisciplinary collections. Present challenges reveal *cost*, *interoperability* and *usability* constraints.

- **Cost:** KOS are expensive to create, maintain and use within most information infrastructures. KOS creation and maintenance requires domain and technical experts, in addition to financial and technical resources.
- **Interoperability:** KOS are frequently developed in silos. As a result, customized software, inconsistent semantics and different syntactic presentations of word forms limit interoperability – even in cases where NISO, ISO or other standards are followed.

Five of the authors are affiliated with the School of Information and Library Science, University of North Carolina at Chapel Hill. Jane Greenberg (janeg<at>email.unc.edu) is a professor in the school as well as director of the school's Metadata Research Center. Robert Losee (losee<at>ils.unc.edu) is a professor. José Ramón Pérez Agüera (jose.aguera<at>gmail.com) is a research affiliate; Hollie White (hcwhite1<at>email.unc.edu) is a doctoral fellow at the Metadata Research Center; and Craig Willis (craig.willis<at>unc.edu) is a research assistant there. Ryan Scherle is data repository architect, National Evolutionary Synthesis Center (NESCent). He can be reached by email at res20<at>duke.edu.

CONTENTS   < PREVIOUS PAGE   NEXT PAGE >   NEXT ARTICLE >

- **Usability:** Poor interface design limits public access to KOS, interfering with user benefits during search. Limited system functionalities impede the ways in which information professionals can work with and share KOS.

The Dryad Case

The problems outlined above reflect the challenges encountered when pursuing a controlled vocabulary system to support the Dryad repository [4]. Dryad is an international repository of data underlying peer-reviewed articles in the basic and applied biosciences. Dryad development is led by a partnership involving the National Evolutionary Synthesis Center (NESCent) and the Metadata Research Center at the School of Information and Library Science, University of North Carolina at Chapel Hill (UNC/CH). Additional partners include North Carolina State University, the University of New Mexico, Yale University, the British Library and Oxford University. Partner journals and societies are listed at http://datadryad.org/partners.

As Dryad planning began, it quickly became apparent that a single controlled vocabulary could not adequately represent the disciplines and varying conceptual needs comprising biological science research, such as research method, geographical location and taxon. An exploratory mapping experiment confirmed this point. A sample of 600 keywords, drawn from five Dryad partner journal articles, were categorized by facets (topical, geographical, methodological, temporal and taxonomical) and then mapped to terms recorded in 10 controlled vocabulary sources such as *Medical Subject Headings* (MeSH), the National Biological Information Infrastructure's *Biocomplexity Thesaurus* (NBII Thesaurus), *Library of Congress Subject Headings* (LCSH) and Getty's *Thesaurus of Geographic Names* – to name a few. Topical term exact matches (keyword to controlled vocabulary term) were found for 17% to 35% of the keywords when they were matched against the five major vocabularies, while less than 10% of the topical keywords mapped to two or more vocabularies. The results confirmed that each vocabulary had unique and valuable terms for representing Dryad content. The results further indicated that multiple controlled vocabularies combined might result in a single-collective KOS. Additional reporting on this study is found in [5].

A Need for HIVE

Conceptually, working with multiple KOS for representing interdisciplinary content makes sense, particularly if the vocabularies represent sub-components of a larger universe and do not overlap significantly. A major drawback is that, despite many being available online, there is no standardized way to access vocabulary information. This barrier forces KOS use to remain heavily manual. Many vocabularies, and other KOS for that matter, require the metadata creator to search, click, copy, insert and even re-type desired descriptors. The cost of using multiple controlled vocabularies in this current state is prohibitive. The cost increases with the number of KOS. In fact, most digital initiatives that consider employing a KOS use one, or at most two systems. This restriction seems quite limiting, if not primitive, particularly when the content and the KOS are digital and could be primed for machine processing.

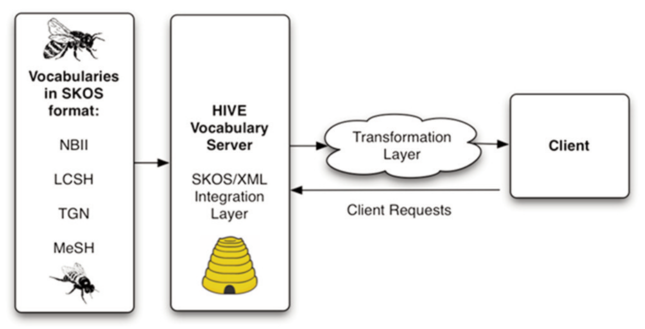
With the exception of the MAchine Research Cataloging (MARC) formats for Library of Congress authority files (names and subject) and the Library of Congress Classification (LCC) system, standards developed for KOS have been primarily intellectual, focusing on conceptual relationships rather than on the encoding necessary to support machine processing and interoperability. Vocabularies that are machine-readable are often idiosyncratic, requiring specialized procedures for working with each vocabulary. Even those that are accessible in XML (eXtended Mark-up Language) use different models/structures, which certainly impedes interoperability.

Fairly recent developments within the World Wide Web (WC3) consortium offer new solutions, such as the Web Ontology Language (OWL) and SKOS. SKOS is appealing because of its simplicity and its support of traditional KOS such as thesauri. The simplicity supports a universal application for KOS, making them easily machine manipulatable and interoperable during a metadata generation activity. The development of SKOS, along with the desire for efficient and affordable access to multiple KOS during metadata generation, supports the HIVE initiative.

The HIVE Initiative

HIVE [4] is led by the Metadata Research Center at the School of Information and Library Science, University of North Carolina at Chapel

FIGURE 1. The HIVE Model



Hill, and NESCent. It is funded by the Institute of Museum and Library Services. The HIVE initiative includes formal vocabulary partners and workshop hosts. Formal vocabulary partners include the Library of Congress, the United States Geological Survey and the Getty Research Institute. Workshop partners include Columbia University; George Washington University; Universidad Carlos III de Madrid, Spain; University of California, San Diego; University of North Texas; and the University of North Carolina at Chapel Hill.

HIVE is both a model and a system that includes a vocabulary server supporting automatic metadata generation by simultaneously drawing descriptors from multiple SKOS-encoded controlled vocabularies. The acronym imbues a thoughtful image – a bee (representing an algorithmic path) visits different KOS (vocabularies) for nectar and brings back selected terms to the HIVE (see Figure 1).

Term selection and assignment is expedited by dynamically selecting relevant concepts from multiple KOS during the metadata generation. The process has implications for a growing range of information systems, particularly institutional and other multi- and interdisciplinary digital repository collections. HIVE is being pursued in this larger context although Dryad is the focus of its attention.

HIVE Goals and Plans

The overriding goal of HIVE is to improve the state of access and use of structured KOS, specifically controlled vocabularies in the digital environment. The specific goals are to do the following:

- Provide an affordable approach for generating subject metadata using automatic metadata generation techniques and pulling concepts from multiple subject thesauri and vocabularies.
- Create an interoperable vocabulary server using the Simple Knowledge Organization Systems (SKOS).
- Create a usable and functional system that will aid resource catalogers and resource authors creating subject metadata.
- Develop and deliver a series of workshops on SKOS and the HIVE model.
- Evaluate the effectiveness and usability of HIVE for the Dryad repository and the larger library, museum and archival communities.

The HIVE project includes three components: building HIVE, sharing HIVE and evaluating HIVE.

- **Building HIVE** focuses on developing the vocabulary server. This component has been completed, although development is ongoing. The server provides efficient, affordable, interoperable and user-friendly access to multiple controlled vocabularies during metadata creation activities. The HIVE software is open source and is available via the GoogleCode repository at http://hive-mrc.googlecode.com/.

- **Sharing HIVE** focuses on educating library, museum and archival professionals about and empowering them through new enabling technologies that can assist them with developing and using controlled vocabularies. This part of the HIVE project is being pursued via a series of workshops. The positive response from the HIVE presentation at the Code4Lib 2010 conference has also made us aware of interest among system developers to learn more about the HIVE software; therefore, the initial workshop plans have been extended to include developer-oriented instruction. To date, the HIVE team has held two workshops, and others are forthcoming. For more information on HIVE workshops, please consult the HIVE wiki at https://www.nescent.org/sites/hive/Category:Workshops.

- **Evaluating HIVE** supports the evaluation of HIVE using automatic information retrieval approaches, through studying resource authors and scientists contributing to the Dryad repository and by surveying members of the LIS professional community about the usefulness of HIVE for their daily work.

CONTENTS
TOP OF ARTICLE
< PREVIOUS PAGE
NEXT PAGE >
NEXT ARTICLE >

HIVE Design

HIVE processes and procedures and SKOS server technology are open and accessible to any library or digital initiative that may benefit from access to multiple controlled vocabularies for indexing and accessing collection holdings. HIVE implements the technological infrastructure to store millions of concepts from different vocabularies and make them available on the web by a simple HTTP call. Vocabularies can be imported in HIVE using SKOS RDF/XML format. As a result of this work, sharing concepts and vocabularies on the web becomes easy and straightforward.

The HIVE software is divided into three different modules: HIVE Core, HIVE Web and HIVE REST.

- **HIVE Core** implements the main features of the system, like automatic metadata extraction and topic detection – using KEA (keyphrase extraction algorithm) (www.nzdl.org/Kea/). Concept retrieval is supported via Lucene, and RDF storage and management uses SESAME/Elmo.
- **HIVE Web** includes a web user interface based on the Google Web Toolkit, offering human users a comfortable interface to browse and search through vocabularies.
- **HIVE REST** provides a machine-oriented interface based on Web Services to ease integration with third party software.

Conclusion and Next Steps

This *Bulletin* article provides an overview of HIVE, noting rationales and outlining goals and plans, and offers an overview of system design. Next steps for HIVE include continuing evaluation efforts, adding new vocabularies to the HIVE system and pursuing partnerships to extend the use of HIVE in practitioner settings.

A current focus is exploring means for automatically synchronizing the HIVE vocabulary instances with the official instances maintained by our vocabulary partners. The goal is to support automatic periodic updates via the Uniform Resource Identifiers (URIs) for concepts. We are examining the Atom feed from id.loc.gov (http://id.loc.gov/authorities/feed/) as a solution for this need. The feed orders concepts by the atom:updated data element. The work with LCSH will inform the design for automatic updates with other vocabularies.

HIVE growth has also included supporting the expanding community of system developers. There has been a growing interest in the HIVE software, with prototype installations at organizations such as the Long Term Ecological Research network, the Library of Congress and the United States Geological Survey.

The HIVE work has been exciting and has required a team approach by people from a range of disciplines (information and library science, computer science and scientific domains). HIVE is real-world applications using Semantic Web technology and linking data to solve data sharing problems. Projects under its umbrella provide a framework for advancing scientific knowledge, not only within evolutionary biology through Dryad, but also about digital applications for sharing data.

Acknowledgements

HIVE is supported by IMLS grant, LG-07-08-0120-08 (Helping Interdisciplinary Vocabulary Engineering – HIVE). We would like to acknowledge Lina Huang, SILS graduate (2010), for her work on HIVE's design during her master's program. ■

Resources Mentioned in the Article

[1] *HIVE Project wiki:* https://www.nescent.org/sites/hive/Main_Page.

[2] *HIVE demonstration system:* http://hive.nescent.org:9090/home.html.

[3] Hodge, G. (April 2000). *Systems of knowledge organization for digital libraries: Beyond traditional authority files.* Retrieved March 18, 2011, from www.clir.org/pubs/reports/pub91/contents.html.

[4] *Dryad Repository:* http://datadryad.org.

[5] Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: An analysis of the Dryad Repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging & Classification Quarterly, 47*(3), 380-402. DOI: 10.1080/01639370902737547.

CONTENTS
TOP OF ARTICLE
< PREVIOUS PAGE
NEXT PAGE >
NEXT ARTICLE >