

## HIVEing Across the U.S. DataNets

Jane Greenberg\*, Chelcie Rowell, Koushyar Rajavi  
Metadata Research Center  
School of Information and Library Science  
University of North Carolina at Chapel Hill  
{janeg@email.unc.edu; chelcie@live.unc.edu;  
koushyar@live.unc.edu}

Mike Conway, Howard Lander  
DataNet Federation  
Consortium/Data Bridge  
Renaissance Computing Institute  
{michael\_conway@unc.edu;  
howard@renci.org}

### 1. Advancing Data Management Infrastructures via Metadata Vocabularies

National, federally funded data programs, such as the NSF funded DataNets, seek a sustainable, national, and global infrastructure for data-driven research.<sup>1</sup> The emphasis is on digital data and cultivating new science and innovation.

Growing attention toward data management includes a range of conferences, symposia, and workshops helping to deter a silo effect. As significant as the current sharing of ideas and approaches may be, a sustainable and effective data-driven cyber-infrastructure requires more research targeting data sharing architectures and interoperable components.

Metadata vocabularies are crucial for interoperability both within and across *all* data management environments. Metadata vocabularies promote greater consistency across data grids, repositories, and hubs, and can contribute to an architecture supporting an unified set of services and interfaces. Specific to this goal, the **Helping Interdisciplinary Vocabulary Engineering (HIVE)**<sup>2</sup> approach supports the dynamic and interoperable application of metadata vocabularies—specifically controlled vocabularies. Controlled vocabularies in this realm represent topical, temporal, geospatial, and taxonomic concepts.

The paper introduces the HIVE approach and reports on HIVE R&D relating to DataOne and the DataNet Federation Consortium (DFC) DataNets. The chief goal for HIVE, as presented here, is to explore means for advancing interoperability between DataOne and the DFC. The work is sparking a bona fide, meaningful relationship between these two DataNets. Another significant outcome is the development of a framework for researching metadata (controlled) vocabulary challenges and broader interoperability questions for data management.

### 2. Metadata Vocabularies and Interdisciplinarity

Metadata vocabularies continue to proliferate in connection with the growing digital data deluge<sup>3</sup>. Metadata registration systems have been launched to promote vocabulary

---

<sup>1</sup> Sustainable Digital Data Preservation and Access Network Partners (DataNet) Program Summary, National Science Foundation. September 28, 2007: [http://www.nsf.gov/funding/pgm\\_summ.jsp?pims\\_id=503141](http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141).

<sup>2</sup> Helping Interdisciplinary Vocabulary Engineering (HIVE): <http://hive.nescent.org/>.

<sup>3</sup> Willis, C., Greenberg, J., and White, H. (2012). Analysis and Synthesis of Metadata Goals for Scientific Data. *Journal of the American Society for Information Science and Technology*, 63 (8): 1505–1520.

sharing.<sup>4</sup> Unfortunately, integrating multiple vocabulary systems within a data network has been impeded by **cost**, **interoperability**, and **use** constraints.<sup>5</sup>

- It is prohibitively expensive and simply not strategic to maintain a nationally or internationally endorsed metadata vocabulary at a project or DataNet level.
- Although metadata vocabularies are increasingly accessible via registries, or at least as digital content, seamless interoperability and easy use within a single framework is still quite limited.

The challenges noted here are magnified in the interdisciplinary world aiming for new science and innovation.

### 3. HIVE Approach

The HIVE project is addressing above outlined metadata vocabulary challenges via dynamic, automatic metadata generation, using multiple vocabularies and the Kea++/Maui machine learning algorithm.<sup>6</sup> HIVE was launched to demonstrate an approach using linked open data (LOD), specifically the linked open vocabularies (LOVs) encoded in the Simple Knowledge Organization System (SKOS) language, a World Wide Web Consortium (W3C) standard. LOV, combined automatic indexing, and machine learning based on expert cataloging provide a cost effective and efficient approach for multiple controlled vocabularies. A HIVE system (also referred to as HIVE) is particularly valuable for digital, interdisciplinary data collections where textual components can be leveraged to aid metadata description.

HIVE is supported by a suite of resources (open source code, instructions for installing HIVE, etc.) and a community communication platform (listserv, wiki, etc.) to support ongoing HIVE use and continued implementations. The HIVE infrastructure supports:

- Two machine learning algorithms, Keyphrase Extraction Algorithm (Kea++) and Maui, developed at the University of Waikato, New Zealand.<sup>7,8</sup>
- Downloading/term selection for multiple SKOSed concepts in SKOS-Resource Description Framework/eXtensible Markup Language (RDF/XML), SKOS-N Triples, Dublin Core/XML, Metadata Object Description Schema (MODS)/XML, and Machine Readable Cataloging (MARC/XML) (see Figure 1, screen shot from demonstration system).
- A service used by both information professionals and researchers, even via the demonstrative status of HIVE.

A Spanish language HIVE demonstration system collaboratively implemented by the Metadata Research Center (MRC) at the School of Information and Library Science, University of North

---

<sup>4</sup> Metadata registry examples: NCBO BioPortal <http://bioportal.bioontology.org/>; Open Metadata Registry (OMR): <http://metadataregistry.org>; Linked Open Vocabularies project (LOV): <http://labs.mondeca.com/dataset/lov>.

<sup>5</sup> Greenberg, J. et al. (2011). HIVE: Helping Interdisciplinary Vocabulary Engineering. *Bulletin of the American Society for Information Science and Technology*, 37 (4): [http://www.asis.org/Bulletin/Apr-11/AprMay11\\_Greenberg\\_etAl.html](http://www.asis.org/Bulletin/Apr-11/AprMay11_Greenberg_etAl.html).

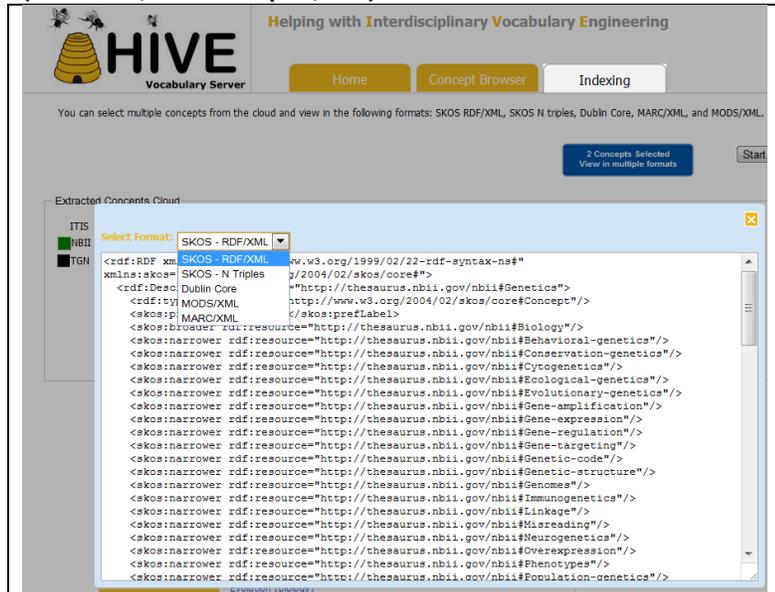
<sup>6</sup> Medelyan, O. and Whitten I.A. (2008). Domain independent automatic keyphrase indexing with small training sets. *Journal of the American Society for Information Science and Technology*, (59) 7: 1026-1040.

<sup>7</sup> Kea-algorithm: Automatic keyphrase algorithm: <http://code.google.com/p/kea-algorithm/>.

<sup>8</sup> Maui-indexer: <http://code.google.com/p/maui-indexer/>.

Carolina at Chapel Hill, and Tecnologías Aplicadas a la Información y la Documentación (TECNODOC), Library and Information Science Department, Universidad Carlos III de Madrid, Spain, with four vocabularies.

**Figure 1: HIVE demonstration system term download options (SKOS-RDF, SKOS N Triples, etc.)**



HIVE was launched by the Metadata Research Center, School of Information and Library Science, University of North Carolina at Chapel Hill (MRC/UNC-CH), in collaboration with the National Evolutionary Synthesis Center (NESCent). First phase nationally recognized vocabulary partners include the US Geological Survey, Library of Congress, the Getty Research Institute, and National Library of Spain. As the project continued, the National Library of Medicine, the UN Food and Agricultural Organization, the National Agricultural Library, and the partnership overseeing the Unified Astronomy Thesaurus formed a cadre of additional constituent vocabulary partners.

HIVE data-driven instances have been implemented for the 1.) Long Term Ecological Research Network(LTER), 2.) the USGS Thesaurus, and the 3.) the Dryad repository—via a curator prototype. HIVE instances, at various stages of use and experimentation, have also been implemented for the 1.) Library of Congress Web Archives, 2.) Yale University Library Cross Collection Discovery project, 3.) Columbia University Human Rights Web Archiving Project and Oral History Transcripts, 4.) Universidad Carlos III de Madrid (UC3M) demonstration system, and 5.) Institute of Legal Information Theory and Techniques, National Research Council, Italy.

#### **4. Advancing Interoperability and Interdisciplinarity Across DataNets**

HIVE is now being explored in more detail in the DataOne and the DFC DataNet communities. This work is motivated by the HIVE approach and provision of a framework supporting interoperability and interdisciplinarity across the DataNet environments. Effective interoperability among existing DataNets requires data gathering and examination on many levels. Two key efforts are taking place as part of this work: 1.) A HIVE instance is being

integrated with iRODS, the system underlying the DFC, 2.) A metadata vocabulary/infrastructure investigation is being conducted in both the DataOne and the DFC communities. This investigation will provide data useful for developing a sustainable, interoperable infrastructure, including how HIVE may better support metadata vocabulary needs.

#### 4.1 Metadata vocabularies: hosting and use

An early-stage prototype of HIVE integrated with iRODS is represented in Figure 2. iRODS provides free form attribute-value-unit metadata values (an AVU) that may be attached to data object (files) and collections (directories). The vision is that HIVE integration will accelerate the ability to utilize controlled vocabularies, as well as keyword extraction and search capabilities, within iRODS. The strategy is to serialize SKOS vocabulary terms in the iRODS AVU format as a canonical metadata store. These assigned terms may then be extracted by indexing processes to populate an external triple-store, providing SPARQL query access to iRODS collections. A HIVE-specific user scenario for iRODS is outlined in Table 1. Phase 2 will be to sweep the collections and populate a triple-store with the marked up collections and files. This will provide the necessary framework to support SPARQL queries to find iRODS collections. A current goal is to identify a unified URI for iRODS files and collections to link this data. Some of this work may be included in the DataBook application within VIVO.

Figure 2: HIVE in iRODS

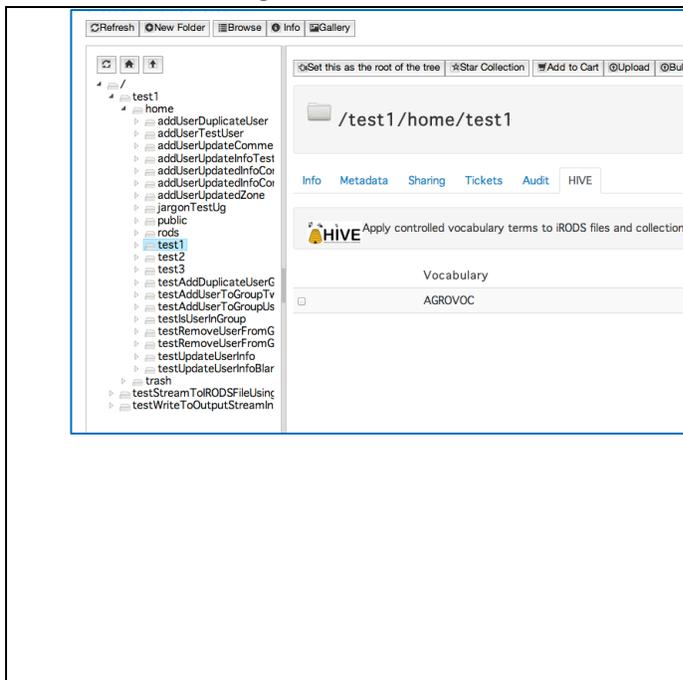


Table 1: HIVE/iRODS User Scenario

1. User goes to a collection.
2. User selects HIVE tab.
3. User selects a vocabulary.
  - a metadata value [iRODSUserTagging:HIVE:Vocabulary AVU](#) is added with the vocabulary name as the attribute.
4. User will navigate through tree, search, and browse by selecting narrower, broader, related as in the HIVE demo.
5. Upon arrival at a desired term, the user will select an 'apply button'.
6. The RDF triple associated with the vocabulary is stored as an AVU.
  - subject is implied, it is the target data object or collection.
  - predicate is stored in 'attribute' of AVU
  - object is the URI of the vocabulary term, stored in 'value' of AVU.
  - AVU unit is of type [iRODSUserTagging:HIVE:RDF](#).

## 5. Toward a framework for study controlled vocabulary use

A web survey is being distributed to DataONE and DFC stakeholders (data contributors, data curators, NSF DataNet Partner administrators, and repository infrastructure developers). The survey is gathering information on:

- Controlled vocabularies (a type of metadata vocabularies) are in use by different DataONE member nodes and DFC data grids.
- Purposes these controlled vocabularies serve (e.g. subject description of datasets or description of analytical processes or protocols that have been applied to certain datasets).
- Facilitators and inhibitors of controlled vocabulary use by data contributors.
- Facilitators and inhibitors of controlled vocabulary use by data curators, NSF DataNet Partner administrators, and/or repository infrastructure developers.

The survey work provides a framework studying controlled vocabulary use across all DataNets and other data environments.

The survey uses demographic information (such as discipline, DataNet affiliation, and DataNet role) to determine the question path. There are common questions that cut across all participant communities, such as knowledge of selected controlled vocabularies, although a large set of data contributor questions, differ from those presented to data curators, NSF DataNet Partner administrators, and repository infrastructure developers.

Given the very early stage of this work, it is premature to make any conclusions. It is very likely that more robust reporting will be supported by early to mid-March, 2013, specifically insight into the following questions about vocabularies used in these environments:

- Availability of a controlled vocabulary's values as Uniform Resource Identifiers (URIs)
- Vocabulary currency and update frequency.
- Data storage for controlled vocabularies (e.g. spreadsheet, relational database, thesaurus software, Web).
- Openness to term suggestions.
- Whether a controlled vocabulary is maintained in-house or out-of-house by the repository, and where term data is being stored.

## 6. Conclusions

The HIVE activities presented in this paper are being pursued via the DataONE and DFC DataNets. This work has implications for other DataNet communities. The documentation and sharing of HIVE implementations are useful for other initiatives that seek means for automatic indexing using multiple vocabularies, and for leveraging existing linked open vocabularies (LOV). Finally, and perhaps, most significantly, these collective HIVE efforts serve as a framework for helping us better understand metadata vocabulary challenges and broader interoperability questions that impact the management of scientific data.