Advancing the DFC Semantic Technology Platform via HIVE Innovation

Mike C. Conway[1], Jane Greenberg[2], Reagan Moore[1],
Mary Whitton[1], and Le Zhang[2]

[1] The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
michael_conway@unc.edu, {rwmoore,whitton}@renci.org
[2] Metadata Research Center, School of Library and Information Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
janeg@email.unc.edu, lezha@live.unc.edu

Abstract. The DataNet Federation Consortium (DFC) is developing data grids for multidisciplinary research. As the DFC grid grows in size and number of disciplines, it becomes critical to address metadata management and findability challenges. The HIVE project is being integrated into the iRODS in the DFC architecture to provide a scaleable linked open data approach to scientific data sharing.

Keywords: HIVE, iRODS, semantic web, linked open data, SKOS.

1 Introduction

National and global cyberinfrastructure initiatives must manage large data collections across their entire life-cycle, enabling research and scientific discovery [1]. Data findability and access are crucial to these goals. As scientific research becomes multidisciplinary in nature, and data increase in volume and diversity, the data management challenges involved also become metadata management challenges [2]. Additionally, as infrastructure evolves through multiple efforts, such as the U.S. DataNet implementations and European Unions's INSPIRE initiative, data management challenges further reveal metadata interoperability challenges [3,4].

Semantic systems, specifically linked open data (LOD) vocabularies, can address these challenges and advance cyberinfrastructure development [5]. Open, shared semantics introduce new capabilities that ought to be explored with scientific data. Semantic systems enhance findability by defining standard vocabularies used within a community. Researchers within the community can then reference data sets using a consensus naming convention. Semantic systems enhance interoperability because they provide a way to translate between the vocabularies used by different communities. The DataNet Federation Consortium (DFC) recognizes LOD capabilities and team members are exploring open semantics via HIVE, a technology that supports the dynamic integration of linked data vocabularies encoded in the Simple Knowledge Organization System (SKOS) [6].

E. Garoufallou and J. Greenberg (Eds.): MTSR 2013, CCIS 390, pp. 14–21, 2013.

The DFC is extensively interdisciplinary, hosting data documenting ocean observations, hydrologic science, engineering education and archives, plant genome sequences, and social science research. There is no single semantic system that covers the wide variety of disciplines being integrated into the DFC, particularly at the granular level that data requires [7]. Creating a single vocabulary to enhance interoperability among the expanse of DFC datasets is prohibitively expensive, and neither practical nor feasible. DFC's approach is to instead leverage the universe of existing semantic systems that are already developed and maintained by other agencies. HIVE technology supports this goal by providing a simple means for integrating multiple disciplinary-specific vocabularies, on a basic level [8], and is being pursued as part of the DFC R&D efforts within the iRODS data grid.

This paper reports on this DFC/HIVE initiative. The paper is organized as follows: Section 2 reviews the relationship between semantics and data; section 3 provides an overview of HIVE; section 4 introduces the DFC and the iRODS platform (software technology) on which this system relies; section 5 reveals the DFC's semantic needs; section 6 reports project implementation progress; and section 7 presents a conclusion and identifies next steps.

2 Semantics and Data

Semantics systems used in the information/database community vary in scope, structure, domain and other aspects. The diversity is reflected in varied naming conventions referenced as ontologies, taxonomies, authority control lists, vocabularies, lexicons and the like [9]. Notwithstanding differences, these semantic systems all support similar functions, chiefly they aim to facilitate discovery, link related resources, and add context to a collection. Semantic systems also support interoperability, enabling the sharing, cross- searching and exchange of information that is being represented.

Semantic systems are crucial for data management. Historical examples commonly link back to Linnaean taxonomy, or delve as far back as Aristotle's naming of specimens in his *Historia Animalium* (History of Animals) [10], and his contribution to binomial means "two names" convention for naming specimens, a practice replicated in many scientific domains. The development and sharing of semantic systems in science thrives today, due to digital innovation and networked technologies. Examples include the National Center for Biological Ontologies (NCBO) [11], for registering scientific ontologies the collective effort to develop the Gene Ontology–GO [12], and the Marine Metadata Initiative, which creates, published, and makes accessible semantic systems, in an open format [13]. These developments provide infrastructure that can be leveraged to improve interoperability and the sharing and exchange of data. The DFC is advancing its semantic technology platform via integration of the HIVE system with the iRODS data management system.

3 An Overview of HIVE

HIVE is an acronym for Helping Interdisciplinary Vocabulary Engineering. As part of the HIVE project, a framework was developed to allow curators to manage multiple controlled vocabularies defined in SKOS [14]. The HIVE system includes support for importing SKOS vocabularies through an administrative toolkit. This import takes a SKOS vocabulary, and populates a `Sesame/Elmo` triple store. HIVE provides an easy API to interactively query and navigate across and within loaded vocabularies from a user interface [15].

As vocabularies are imported into HIVE, they are indexed to support concept retrieval, using the Lucene search engine. This index allows a curator to find appropriate terms across selected vocabularies using concepts specified as a free text search query, and then to select and navigate based on matching concepts.

HIVE also supports automatic term suggestion for documents using the `KEA++` and `MAUI` algorithms [16,17]. These machine learning algorithms are trained with a sample set of documents indexed by a human working with a designated vocabulary. The training enables the use of these learned patterns during the dynamic indexing activities, although candidate terms are drawn from the SKOS vocabulary, or multiple vocabularies, based on novel document content. This term suggestion process is applied at the time that a document is uploaded into a repository, allowing users to view suggested terms, and to select the terms that best apply. The DFC project is exploring automatic term extraction as a policy applied within the iRODS grid, and this is discussed in the next steps section.

HIVE provides a framework to manage open vocabularies, and includes the functionality to integrate multiple vocabularies into the metadata workflow aiding researchers and curators, providing consistent and enhancing findability in multidisciplinary research environment.

4 DFC and iRODS Technology

The DataNet Federation Consortium (DFC) is one of five DataNet projects under the National Science Foundation DataNet initiative. The stated goal of the DFC is to "..assemble national data infrastructure that enables collaborative research, through federation of existing data management infrastructure..." [18] The DFC is based on the iRODS data grid, which provides the interoperability mechanisms needed to federate existing data management systems with a national collaboration environment. Federation requires the encapsulation of three types of domain knowledge:

- knowledge required to access community resources and discover and retrieve relevant input data sets;
- knowledge needed to execute a data-driven research analysis;
- knowledge needed to manage research results in compliance with NSF data management plans.

Given these three types of knowledge, it becomes possible to support reproducible data-driven research. All of these types of knowledge encapsulation require the ability to manage domain vocabularies. Each domain uses different terms to describe the contents of data repositories, different terms to describe the operations performed in an analysis, and different terms for describing research results. The DFC offers an opportunity to investigate and prototype findability and interoperability solutions across deployed research infrastructure.

iRODS is an open-source, policy managed data grid, developed by the Data Intensive Cyber Environments group (DICE). iRODS is an acronym for the Integrated Rule-Oriented Data System [19]. The iRODS software manages a central metadata catalog of distributed data collections. iRODS virtualizes collections of data, presenting a uniform, abstract view across multiple physical storage architectures. An iRODS grid can exist on distributed storage nodes, and grids can be federated between organizations [20].

Central to iRODS is the concept of policy managed data. iRODS has a concept of microservices, which are defined as "... small, well-defined procedures/functions that perform a certain task... " [21]. These microservices can be chained together by rules to create complex actions. iRODS provides a rule engine that runs at each remote storage location. Policy Enforcement Points (PEPs) are defined within the software middleware that can trigger rule execution based on events within the grid. Administrators can thus define policies to control data management, which are enforced by the grid, and which can be verified by assessing the state of the catalog [22].

Through the iRODS catalog (ICAT), it is possible to manage and query for user defined metadata. The user defined metadata is stored as Attribute-Value-Unit (AVU) triples, and these arbitrary triples can be associated with collections and data objects contained by the grid. These AVUs are rudimentary and free-form, with no ability to express rich or linked relationships. On the other hand, these AVUs benefit from the policy management and preservation capabilities of iRODS. Layering the capabilities of the HIVE system above the facilities of iRODS adds metadata richness and expressivity to the native iRODS metadata approach.

5 DFC Semantic Needs

As iRODS and the DFC represent a policy managed, open architecture for scientific data management, any approaches to findability must follow a similar open approach. As new policies may be required by the inclusion of new scientific domains, new metadata vocabularies may also be required. HIVE is particularly attractive in that it operates as a 'container', neutral to the vocabularies it contains, able to be augmented with new vocabularies, and able to work across multiple vocabularies.

SKOS, and more broadly, RDF, with a resource centric approach, map well to the DFC architecture. iRODS maintains a global logical namespace over DFC collections. In effect, these collections and files are dereferenceable resources. Attaching SKOS vocabulary terms to these target collections and files can be accomplished with no alteration to the DFC architecture, since they are easily stored as Attribute-Value-Unit (AVUs) attached to DFC files and collections. This delineation of responsibilities is key, as iRODS archives have historically grown up to hundreds of millions of data objects [23]. Maintaining metadata and processing rich queries on such large repositories is a daunting challenge. Using an architecture where iRODS serves as a canonical metadata store, and indexes are generated and replicated as ephemeral entities is key to the DFC approach. It is important to maintain a clear separation of concerns between all of the elements of this solution.

Maintenance and storage of metadata in iRODS is simple and well understood. SKOS defines a fairly rich metadata schema in a format that is approachable and realistic. HIVE maintains dynamic access to libraries of controlled vocabularies, providing researchers and curators an user- friendly application to annotate and classify DFC collections. Triple store implementations via Jena, enables the storage and manipulation RDF data, and inferencing and querying. All of these separate, well-defined operations and the marriage of policy-managed preservation with linked open data approaches are showing promise in the first iteration of this integration.

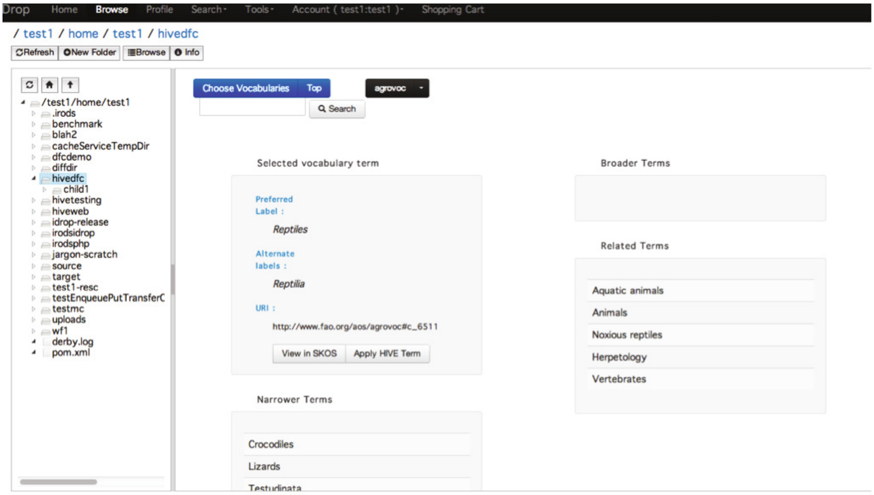
Improving findability through semantic metadata presents a classic information organization versus information retrieval dilemma. HIVE is attractive for DFC because it is oriented towards easy annotation by researchers and curators. HIVE translates easily to the DFC web interfaces, allowing search and exploration of appropriate terms across multiple configured vocabularies, without requiring any knowledge of, or exposure of the semantic nature of the underlying data. In addition, HIVE supports automatic term suggestion using the `KEA++` and `MAUI` frameworks. This term suggestion approach has been shown to provide consistent and accurate and useful annotations with reduced effort on researchers and curators [24]. Algorithmic term extraction approaches are especially appealing in the policy-managed DFC environment, and such metadata extraction is easily supported by the iRODS software as part of the data curation lifecycle.

6 Implementation Progress

The DFC has demonstrated iRODS/HIVE integration using a working system that implements three primary system components. These are:

- Interface integration to allow navigating and searching vocabularies stored in HIVE, and applying selected terms to the DFC grid;
- Indexing services to extract vocabulary metadata that was stored in the DFC grid, in order to populate a triple store with vocabulary and DFC metadata;
- Creation of a search service, and integration of search of the indexed semantic metadata, via `SPARQL`, into the user interface, allowing search and linking to the underlying DFC data associated with the `SPARQL` query results.

The HIVE service is integrated into the iRODS iDrop web interface, providing a metadata tab in the user interface when viewing grid data. This 'concept browser' view provides an intuitive way to navigate around the SKOS vocabulary, by moving to narrower, broader, or related terms in the selected set of



Preferred Label : Reptiles

Alternate labels : Reptilia

URI : http://www.fao.org/aos/agrovoc#c_6511

Related Terms: Aquatic animals, Animals, Noxious reptiles, Herpetology, Vertebrates

Narrower Terms: Crocodiles, Lizards, Testudinata

Fig. 1. HIVE Concept Browser in iDrop

vocabularies. When desired terms are located, they may be applied, resulting in the recording of AVU metadata within the DFC grid, essentially as a serialized RDF triple. The concept browser, through the HIVE API, masks the semantic nature of this data, and instead represents links between concepts in a pivot table arrangement, rather than a hierarchical tree. Comparisons between tree representations of controlled vocabularies and this concept browser arrangement are an interesting area for future study.

Once the metadata is serialized into the DFC grid, an indexing process extracts relevant metadata to populate an external triple store. Currently, the first phase uses a batch indexing mode, where the DFC master catalog is queried to locate such RDF triples. During this sweep operation, a `Jena` triple store is populated with the SKOS vocabularies, and the application of these vocabularies to iRODS logical resources (files and collections), along with other catalog metadata, expressed in an iRODS OWL schema. This iRODS system metadata can include characteristics of files and collections, such as ownership, access control relationships, provenance, and collection membership. Note here that sharing relationships between researchers can be extracted to FOAF, and developing this social dimension is an interesting future research topic.

To complete the first phase, the `SPARQL` query language, as implemented by `Jena`, was used to demonstrate metadata search based on SKOS vocabulary terms, including the ability to find collections annotated with a given term, as well as the ability to find collections related to a given term. This demonstrates the enhanced findability that results from semantic metadata, and highlights the potential utility of the HIVE + iRODS approach for richer queries. The search facility was implemented as a RESTful web service, and integrated into the search toolbar of the iDrop web interface. The search results included dereferenceable links that corresponded to the iRODS URI for the files or collections, and showed the ability to resolve and return the annotated data.

7 Conclusion

This paper provided an overview of DFC and the iRODS platform (software technology) and policy management mechanisms, discussed DFC's semantic needs, explained HIVE and the machine learning processes supported by `Kea++/MAUI`, and reported on DFC/iRODS HIVE implementation. The first phase of HIVE integration has successfully demonstrated an operating service for annotation of DFC data using controlled vocabularies defined with SKOS in a manner accessible to researchers and curators. This phase demonstrated a batch indexing mode, populating a triple store, as well as integrated `SPARQL` based query, and resolution of query results back to DFC content. The next phase of this work will add additional facilities for near-real-time indexing using asynchronous messaging. This will use the policy-managed aspects of iRODS to detect metadata activities, and publish changes to a topic queue. Pluggable indexers can run on an external message bus, allowing other metadata tools to be easily integrated.

As phase two of this work begins, we will add automatic extraction of terms to documents deposited into DFC working with `Kea++` and `MAUI` and generating AVU metadata that serialize the detected SKOS terms. These terms will be indexed, and used for searching. It is hypothesized that users of DFC search can then rate returned documents based on appropriateness, allowing the gradual improvement of metadata quality through actual use. The automatic term extraction approach lowers the cost of metadata annotation, enables greater interoperability across the full DFC and with other datagrids, and can improve findability of data. This marrying of automatic term extraction and interactive rating of metadata quality will be added to the iDrop web interface in the next phase. Planned assessments will allow us to study and improve this work, and contribute new functionalities to operating research grid.

HIVE code is open source. The GForge repository at the Renaissance Computing Institute (RENCI) holds the HIVE and HIVE-iRODS integration libraries at `https://code.renci.org/gf/project/irodshive/`. The iDrop browser implementation is available at `https://code.renci.org/gf/project/irodsidrop/`, with the HIVE integration on the git branch `1131-hive-integration`.

Acknowledgments. The work in this paper is supported, in part, by the U.S. National Science Foundation (OCI 0940841).

References

1. Moore, R.W., Whitton, M.: Data Intensive Cyber Environments Center: Six Month Report. University of North Carolina at Chapel Hill: dfc quarterly report (2012), `http://datafed.org/dev/wp-content/uploads/2012/04/DFC-quarter-2_7kag1.pdf`
2. Willis, C., Greenberg, J., White, H.: Analysis and Synthesis of Metadata Goals for Scientific Data. Journal of the American Society for Information Science and Technology 63(8), 1505–1520 (2012)

3. Lee, J.W., Zhang, J.T., Zimmerman, A.S., Lucia, A.: Datanet: an emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning. Aiche Journal 55, 2757–2764 (2009)
4. Portele, E. (ed.): GIGAS Technical Note - Data Harmonisation and Semantic Interoperability. Project co-funded by the European Commission within the Seventh Framework Programme, `http://inspire-forum.jrc.ec.europa.eu/pg/pages/view/9782/`
5. Janowicz, K., Schade, S., Bröring, A., Keßler, C., Maué, P., Stasch, C.: Semantic enablement for spatial data infrastructures. Transactions in GIS 14(2), 111–129 (2010)
6. Greenberg, J., Rowell, C., Rajavi, K., Conway, M., Lander, H.: HIVEing Across U.S. DataNets. In: Research Data Management Implementations Workshop, March 13-15. NSF/Coalition for Academic Scientific Computation (CASC), Arlington, VA (2013), `http://tinyurl.com/d85kywg`
7. Greenberg, J., Moore, R.W., Whitton, M.: Advancing Interoperability and Interdisciplinarity Across Datanets. NSF Supplement Proposal (2012)
8. Greenberg, J., Losee, R., Pérez Agüera, J.R., Scherle, R., White, H., Willis, C.: HIVE: Helping Interdisciplinary Vocabulary Engineering. Bulletin of the American Society for Information Science and Technology 37(4), 23–26 (2011)
9. Rowell, C., Greenberg, J.: Advancing Interoperability of NSF DataNet Partners Through Controlled Vocabularie, July 7-8, 2013. DataOne Users Group meeting. Chapel Hill, North Carolina (2012)
10. Historia animalium, vol. 2. Loeb Classical Library (1993)
11. National Center for Biological Ontologies: Bioportal, `http://bioportal.bioontology.org/`
12. Gene ontology, `http://www.geneontology.org/`
13. Marine Metadata Initiative: Vocabularies: Dictionaries, Ontologies, and More, `https://marinemetadata.org/guides/vocabs`
14. Greenberg, et al.: Ibid (2011)
15. Greenberg: (2009) presentation at, `http://www.cendi.gov/presentations/11-17-09_cendi_nfais_Greenberg_UNC.pdf`
16. Kea, `http://www.nzdl.org/Kea/download.html`
17. MAUI, `https://code.google.com/p/maui-indexer/`
18. Moore, R.W., et al.: DataNet Federation Consortium Vision and Rationale [Project Proposal] (2012), `http://datafed.org/dev/wp-content/uploads/2012/04/DFCproposal.pdf`
19. Rajasekar, A., Moore, R.W.: iRODS Primer: Integrated Rule-Oriented Data System Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan and Claypool Publishers (2010), `http://www.morganclaypool.com/doi/abs/10.2200/S00233ED1V01Y200912ICR012?journalCode=icr`
20. iRODS White Paper, `https://www.irods.org/pubs/DICE_RODs-paper.pdf`
21. iRODS Microservices Overview, `https://www.irods.org/index.php/Micro-Services`
22. Moore, Rajasekar, Marciano (2007), `https://www.irods.org/pubs/DICE_DigcCur-Trusted-Rep-07.pdf`
23. iRODS Fact Sheet, `https://www.irods.org/pubs/iRODS_Fact_Sheet-0907c.pdf`
24. White, H., Willis, C., Greenberg, J.: The HIVE impact: Contributing to Consistency via Automatic Indexing. In: iConference 2012, Toronto, ON, Canada, February 7-10 (2012), doi:10.1145/2132176.2132297