Greenberg, J. (2014). Metadata Capital via a linked data HIVE. Advances In Classification Research Online, 24(1). doi:10.7152/acro.v24i1.14680

Metadata Capital via a linked data HIVE

Jane Greenberg
Metadata Research Center,
School of Information and Library Science,
University of North Carolina at Chapel Hill
janeg@email.unc.edu;

ABSTRACT
This paper explores metadata capital via linked open metadata vocabularies, specifically via the HIVE (Helping Interdisciplinary Vocabulary Engineering) initiative in the U.S. DataNet Federation Consortium (DFC). Formulas representing 'Capital-sigma notation' and 'Successive growth rates' are introduced as potential means for quantifying metadata capital. A conclusion summarizes this paper and identifies next steps.

Keywords
Metadata capital; Metadata vocabularies; Linked data, Linked open vocabularies (LOV)

INTRODUCTION

Big data and linked data fervor have ignited a sincere interest in metadata in nearly every domain and societal sector. Perhaps most telling of metadata's reach is the Snowden case revealing the U.S. National Security Administration's (NSA) access to phone communication metadata (PRISM Whistle Blower, 2013). A more hospitable environ includes national and global open data initiatives (e.g., U.S. DataNets and the Research Data Alliance) that are advancing approaches to sharing metadata vocabularies. These initiatives recognize that metadata vocabularies are necessary for building a sustainable data-driven cyberinfrastructure and can aid in cultivating new science. In this context, linked open vocabularies, also known as LOV, manifest as metadata capital (Greenberg, et al, 2013b) and have implications for big data analytics.

This paper explores metadata capital via linked open metadata vocabularies and specifically the HIVE (Helping Interdisciplinary Vocabulary Engineering) initiative in the U.S. DataNet Federation Consortium (DFC). The paper is presented to enable further dialog and consider approaches for quantifying metadata value and building capital via reuse. Long term goals include contributing to a sustainable cyberinfrastructure where metadata vocabularies enable robust approaches for big data analytics.

METADATA VOCABULARIES, COST, AND CAPITAL

It's well known that 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" (Greenberg, et al, 2013a). Benefits aside, metadata generation requires resources. Regardless of how metadata is produced (e.g., human, machine, or mixed means), there is a cost; and this cost is greater when the agency engaged in metadata generation is also responsible for developing and maintaining the metadata standard or vocabulary. Resource investment in metadata can yield positive returns when metadata quality is good and the metadata is reused. In other words, in addition to simply improving systems functionalities, metadata value can increase via reuse.

Metadata capital reflects this idea, although exploration of this topic is limited. Quantifying metadata reuse can allow for more rigorous analysis and evaluation of metadata capital, and enable a deeper understanding of this concept. The next section introduces the U.S. DataNet Federation Consortium, and explores the early implementation of HIVE as a mean for further study of metadata capital.

THE DATANET FEDERATION CONSORTIUM

The DataNet Federation Consortium (DFC) is one of five DataNet projects under the National Science Foundation DataNet initiative. DFC seeks to "..assemble national data infrastructure that enables collaborative research, through federation of existing data management infrastructure..." (Moore, et al, 2012). The iRODS data grid underlies the DFC foundation. iRODS provides interoperability mechanisms necessary for federating existing data management systems and global collaboration. The DFC requires domain knowledge to: 1.) access community resources and discover and retrieve relevant input data sets; 2.) execute a data-driven research analysis; and 3.) manage research results in compliance with NSF data management plans (Conway, et al, 2013). Metadata vocabularies are crucial in these knowledge domain contexts, and HIVE functionalities are being implemented in iRODS to support the DFC knowledge domains.

HIVE and the DFC

HIVE demonstrates an approach using 1.) linked open vocabularies (LOVs) encoded in the Simple Knowledge Organization System (SKOS) language, a World Wide Web Consortium (W3C) standard, and 2.) machine learning (Greenberg, et al, 2011). LOV combines automatic indexing based on machine learning that draws from expert catalogers. The approach provides cost effective means for using multiple controlled vocabularies in an interdisciplinary setting. HIVE is supported by a suite of resources (open

ISSN: 2324-9773

source code, installation instructions, etc.) and a community communication platform (listserv, wiki, etc.) to support ongoing HIVE use and continued implementations (https://code.google.com/p/hive-mrc/).

A HIVE beta system has been integrated into iRODS. The system uses attribute-value-unit metadata values (AVUs)—essentially triples, data object (files), and collections. Currently, two SKOS vocabularies have been implemented: the Food and Agriculture Thesaurus (FOA), *AGROVOC*, (English language component), maintained by the FAO; and the *Unified Astronomy Thesaurus* (UAT), maintained by a Smithsonian/NASA partnership.

It is neither feasible nor practical for the DFC to maintain either of these vocabularies, especially when the maintenance agencies have solid procedures and domain experts to oversee vocabulary development. Rather, a DFC priority is to implement processes to feed automatic updates into the DFC-HIVE. The general cost of vocabulary development and maintenance is one aspect of the equation. DFC metadata costs begin via the initial implementation of a vocabulary into the DFC HIVE. Metadata capital can build via the assignment of vocabulary terms to DFC datasets, and the reuse of metadata over-and-over again, and data sharing amongst scientist.

Quantifying Capital

Measuring metadata vocabulary reuse can lead to a better understanding of metadata value and demonstrate a form of capital. The idea relates, on some level, to when one puts a valuable object (or money) in a bank, and the value of the object increases overtime. In the case of metadata capital, the value increases with reuse, although time is a similar factor.

The early stage of this work considers two formulas. Example 1 presents a simple **Capital-sigma notation**. Here, the value increases by one for each successive us of the metadata vocabulary term. Capital-sigma notation may have an application for metadata capital, although the successive value likely needs to be some higher quantification, or potentially an exponential number.

Example 1:
Capital-sigma notation for Metadata Capital

$$\sum_{i=m}^{n} a_i = a_m + a_{m+1} + a_{m+2} + \ldots$$

Another candidate formula targets **successive growth rates** using approximations and the theta notation. This approach account for an incremental increase over time.

Example 2:
Successive growth rates for Metadata Capital

$$\sum_{i=1}^{n} i^c = \Theta\left(n^{c+1}\right)$$

The formulas presented in this section (Examples 1 and 2) are early considerations. Work is underway at the SILS Metadata Research Center to further determine the initial cost of a DFC vocabulary integration within HIVE, and to map out several reuse iterations. A SIG/CR presentation will provide an opportunity to share this progress and discuss implications of this work.

CONCLUSION AND NEXT STEPS

This paper introduced the notion of metadata capital in the context of linked open metadata vocabularies, specifically the HIVE (Helping Interdisciplinary Vocabulary Engineering) initiative in the U.S. DataNet Federation Consortium (DFC). Two formulas for potentially quantifying metadata capital are introduced. An expanded version of this work will delve into linked data, and address the metadata vocabulary life-cycle, in order to target where 'reuse' is a manifest of metadata capital. Additionally, bona fide examples articulating the formula variables will help demonstrate how we may begin to quantify metadata capital. Finally, attention to role and potential of metadata vocabularies for bid data analytics will be discussed.

ACKNOWLEDGMENTS

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

REFERENCES

Conway, M., Greenberg, J., Moore, R., Whitton, M., and Zhang, L. (2013). Advancing the DFC Semantic Technology Platform via HIVE Innovation. *MTSR 2013 Proceedings.* MTSR 2013: 7th Metadata and Semantics Research Conference November 19-22, 2013 Alexander

Greenberg, J., Losee, R., Pérez Agüera, J.R., Scherle, R., White, H., and Willis, C. (2011). HIVE: Helping Interdisciplinary Vocabulary Engineering. *Bulletin of the American Society for Information Science and Technology, 37* (4): 23-26.

Greenberg et al, (2013a). Greenberg, J., Rowell, C., Rajavi, K., Conway, M., and Lander, H. (2013). HIVEing Across U.S. DataNets. Research Data Management Implementations Workshop, NSF/Coalition for Academic Scientific Computation (CASC), Arlington, VA, March 13-15, 2013: http://tinyurl.com/d85kywg.

Greenberg et al, (2013b). Greenberg, J., Swauger, S., and Feinstein, E. (2013). Metadata Capital in a Data Repository. *DC 2013: Proceedings of the International Conference on Dublin Core and Metadata Applications.* Lisbon, Portugal, September 2-6, 2013.

Moore, R.W., et al, (2012). DataNet Federation Consortium Vision and Rationale [Project Proposal]:

ISSN: 2324-9773

http://datafed.org/dev/wp-content/uploads/2012/04/DFCproposal.pdf.

PRISM Whistle Blower. Interview with Edward Snowden, June 6, 2013, Hong Kong. (Interviewer: Glenn Greenwald, Filmmaker: Laura Poitra): Available at: http://www.theguardian.com/world/video/2013/jun/09/nsa-whistleblower-edward-snowden-interview-video.

ISSN: 2324-9773