The Largest Scholarly Semantic Network...Ever.

Johan Bollen
Digital Library Research & Prototyping Team
Los Alamos National Laboratory
Los Alamos, NM 87545
jbollen@lanl.gov

Marko A. Rodriguez
Digital Library Research & Prototyping Team
Los Alamos National Laboratory
Los Alamos, NM 87545
marko@lanl.gov

Herbert Van de Sompel
Digital Library Research & Prototyping Team
Los Alamos National Laboratory
Los Alamos, NM 87545
herbertv@lanl.gov

ABSTRACT

Scholarly entities, such as articles, journals, authors and institutions, are now mostly ranked according to expert opinion and citation data. The Andrew W. Mellon Foundation funded MESUR project at the Los Alamos National Laboratory is developing metrics of scholarly impact that can rank a wide range of scholarly entities on the basis of their usage. The MESUR project starts with the creation of a semantic network model of the scholarly community that integrates bibliographic, citation, and usage data collected from publishers and repositories world-wide. It is estimated that this scholarly semantic network will include approximately 50 million articles, 1 million authors, 10,000 journals and conference proceedings, 500 million citations, and 1 billion usage-related events; the largest scholarly semantic network ever created. The developed scholarly semantic network will then serve as a standardized platform for the definition and validation of new metrics of scholarly impact. This poster describes the MESUR project's data aggregation and processing techniques including the OWL scholarly ontology that was developed to model the scholarly communication process.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks; H.3.7 [**Digital Libraries**]: Standards—*ontologies*

General Terms

Measurement, Standardization

Keywords

Resource Description Framework and Schema, Web Ontology Language, Semantic Networks

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

1. INTRODUCTION

The most commonly applied metric for determining the value of a journal and thus its authors and their institutions is the ISI[1] Impact Factor (IF) [6]. It is increasingly being used in promotion and funding decisions. It has also had a significant impact on the publication habits of researchers worldwide. There are however a number of well-documented limitations to the ISI IF; citation data lags scholarly trends due to publication delays, the ISI IF is calculated for only about 9,000 journals, journal level statistics do not accurately represent the value of a particular article, and the semantics of citation (e.g. disagreement vs. endorsement) is not always clear.

Whereas millions of articles are stored in repositories worldwide, an even large number of scholarly usage events occur on a daily basis, e.g. downloads and abstract views. This usage may provide a more accurate and refined insight into scholarly impact [4, 3] and at a shorter time-scale than citation data can provide [7, 5, 2]. However, in spite of numerous scientific explorations demonstrating the value of usage data, usage-based metrics of scholarly impact have not achieved any degree of community-acceptance. This can be attributed to the lack of standards to record, represent and aggregate usage data and the absence of a systematic investigation of the properties of various potential usage-based impact metrics.

The MESUR[2] project at the Digital Library Research and Prototyping team of the Los Alamos National Laboratory is in the process of constructing the largest scholarly semantic network ever created which integrates bibliographic and citation data with usage data obtained from various worldwide service providers, e.g. publishers, institutions, library consortia, etc. This scholarly semantic network provides a standardized framework to perform a 2-year systematic study of usage-based metrics which will result in a set of guidelines and specifications with regards to their properties and appropriate applications. The MESUR project will develop metrics using various algorithms drawn from graph theory, semantic network theory, and statistics, along with theoretical techniques developed internal to the project and cross-validated with existing metrics such as the ISI IF, the Usage Impact Factor [3], and the Y-Factor [1]. Figure 1 provides a general overview of the the various stages of the MESUR project.

The MESUR project seeks to aggregate bibliographic, ci-

[1]Now Thomson Scientific
[2]MEtrics from Scholarly Usage of Resources available at: http://www.mesur.org

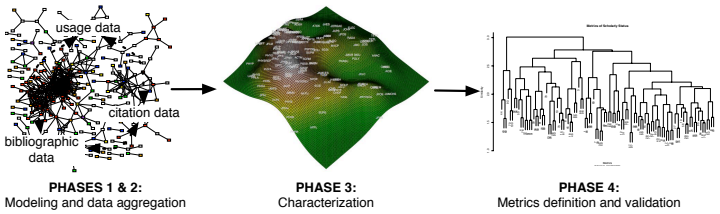

Figure 1: Overview of the MESUR project phases

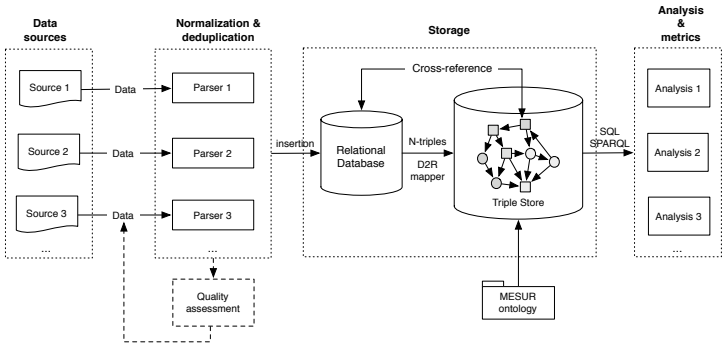

Figure 2: Data flow in the MESUR project

tation and usage data at a very large-scale and hence faces significant data management issues. Therefore, a primary component of this project is focused on data acquisition, de-duplication techniques, data quality measures, and mitigating the time and space limitations of modern triple store platforms. The project includes efforts to define provenance XML schemas, algorithms for uncertainty quantification, and a novel semantic query model that leverages both relational and triple store databases. Another significant component of the MESUR project is the development of a scholarly ontology that represents bibliographic, citation, usage concepts, along with concepts for expressing different artifact metrics.

The proposed poster is divided into two primary components. The first component will focus specifically on the MESUR project's data aggregation and processing methodology. This data flow model is diagrammed in Figure 2. The second component of the poster will present MESUR's scholarly OWL ontology [8]. The presentation of the ontology will demonstrate the novel query model developed by the MESUR project to handle the constraints of modern triple store platforms.

2. CONCLUSION

The MESUR project aims to produce a variety of community-accepted metrics of scholarly impact that each highlight different aspects of value in the scholarly community. This model can be juxtaposed to the citation-driven monoculture that presently prevails in the assessment of scholarly status. Furthermore, the MESUR project aims to contribute to the study of large-scale semantic networks. Along with novel models of scholarly evaluation, advances in semantic network analysis algorithms and large-scale data management techniques have and will continue to be produced.

3. ADDITIONAL AUTHORS

Additional authors: Lyudmila L. Balakireva (Digital Library Research & Prototyping Team, Los Alamos National Laboratory, email: `ludab@lanl.gov`) and Aric Hagberg (Mathematical Modeling and Analysis, Los Alamos National Laboratory, email: `hagberg@lanl.gov`).

4. REFERENCES

[1] J. Bollen, M. A. Rodriguez, and H. Van de Sompel. Journal status. *Scientometrics*, 69(3), December 2006.

[2] J. Bollen and H. Van de Sompel. Mapping the structure of science through usage. *Scientometrics*, 69(2), 2006.

[3] J. Bollen and H. Van de Sompel. Usage impact factor: the effects of sample characteristics on usage-based impact metrics. *cs.DL/0610154*, 2006.

[4] J. Bollen, H. Van de Sompel, J. Smith, and R. Luce. Toward alternative metrics of journal impact: a comparison of download and citation data. *Information Processing and Management*, 41(6):1419–1440, 2005.

[5] T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060 – 1072, 2006.

[6] E. Garfield. Journal impact factor: a brief review. *Canadian Medical Association Journal*, 161:979–980, 1999.

[7] M. J. Kurtz, G. Eichhorn, A. Accomazzi, C. S. Grant, M. Demleitner, and S. S. Murray. The bibliometric properties of article readership information. *Journal of the American Society for Information Science and Technology*, 56(2):111–128, 2005.

[8] M. A. Rodriguez, J. Bollen, and H. Van de Sompel. A practical ontology for the large-scale modeling of scholarly artifacts and their usage. In *Joint Conference on Digital Libraries (JCDL07)*, Vancouver, Canada, June 2007. IEEE/ACM.