Semantic Citation

Ying Ding
Indiana University
1320 E 10th
Bloomington, IN
+1-812-855-5388
dingying@indiana.edu

Deepak Konidena
Indiana University
1320 E 10th
Bloomington, IN
+1-812-855-5388
bkoniden@umail.iu.edu

Yuyin Sun
Indiana University
1320 E 10th
Bloomington, IN
+1-812-855-5388
sunyuyin.david@gmail.com

Shanshan Chen
Indiana University
1320 E 10th
Bloomington, IN
+1-812-855-5388
chenshan@indiana.edu

ABSTRACT

Scholarly papers are the backbone of science and play an important role in the accumulation and dissemination of knowledge and innovation in the academy. Yet many research publications are currently just a "bag of strings" where valuable data and potentially knowledge are hidden. This demo provides referencing services to linking bibliographical papers and citations with existing Linked Open Data. It aims to convert current bibliographical data in various digital library databases into semantic bibliographical data to enable research profiling and intelligent knowledge discovery.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks

General Terms

Measurement, Design,

Keywords

semantic web, citation analysis, Linked Open Data (LOD)

1. INTRODUCTION

Scholarly papers are the backbone of science and play an important role in the accumulation and dissemination of knowledge and innovation in the academy. Scientific works are largely published in the form of journal articles, conference papers or books, but current publication protocols do not provide adequate support for linking or cross-referencing data and metadata. This phenomenon has significantly hindered data sharing and research networking. Semantic Web development provides enabling technologies for data integration and knowledge discovery. This demo, along with the coming project, will transform these bag-of-string research articles into semantic data to facilitate data cross-reference, data integration, data analysis and knowledge discovery.

Adding semantics to bibliographical data and citation data is an on-going research effort in the semantic web area. Some related works are:

- The Semantic Web for Research Communities Ontology (SWRC) models key entities in a typical research community, such as persons, organizations, publications and their relationships [1]. Used in numerous applications, its focus is to provide structured metadata for Web portals. The Publication concept subsumes all types of research publications designed in close correspondence with BibTEX, where modularization is realized by owl:imports statements. It is beneficial to have a modularized ontology design which significantly facilitates the reuse of ontologies and eases maintenance efforts;
- The Bibliographic Ontology (BIBO) provides main concepts and properties for describing citations and bibliographic references [2]. It models components of citing documents in RDF, in terms of pages, titles, abstracts, DOIs, editors, journals and so forth. BIBO also covers items outside of conventional scholarly publications including radio broadcasts and legal documents. It can be extended to include other vocabularies, such as the Dublin Core (DC) metadata and Friend Of a Friend (FOAF);
- The Citation Typing Ontology (CiTO) is used to describe the nature of citations in scientific articles [3]. It captures the intent of citations and permits authors to provide reasons for their citations, such as confirms, corrects, credits, critiques, disagreeWith, discusses, extends obtainsBackgroundForm, and so on. CiTO also provides a way to represent citation frequency for research papers, including the in-text citation frequency and the global citation frequency determined by third-party authorities such as Web of Science, Scopus and Google Scholar; and
- Other related schemas are BibTEX, a metadata format for modeling bibliography entries used within the LATEX document system; the Dublin Core (DC) metadata standard, an attribute value-based set for describing a wide range of resources; Friend of a Friend (FOAF), a way to create machine-readable Web documents for personal networks; Semantically Interlinked Online Communities (SIOC), a schema that describes discussion forums and online community sites; and Functional Requirements for Bibliographic Records (FRBR), a metadata covering various types of publishing items particularly for multimedia works [4].

This demo reports one of the on-going efforts of a project aiming to adding semantics to citation data to enable intelligent citation analysis. It focuses the cross-linking and cross-referencing of different bibliographic data, such as papers and citations, with the existing Linked Open Data to enable broad data integration.

2. Interlinking bibliographical data

Data has been downloaded from Web of Science, one of the major citation databases in the world, several services programmed in Perl to generate the cross-linking and references. For instance, for papers and citations published in the computer science area:

- Author names of the papers and citations were replaced by their corresponding URLs in DBLP, or Google Scholar
- Articles were referenced by their DOIs
- Journals and conference proceedings were referenced by their ISSNs
- CiTo is used to represent the citing relations between papers and citations

The whole infrastructure is illustrated in Figure 1 where data from Web Of Science (WOS), Scopus and other bibliographical data will be converted into RDF triples based on selected bibliographical ontologies. Then instances inside the publications will be hyperlinked, annotated and further linked to other existing semantic data, such as Linked Open Data sets. Afterwards, visualization and query can be performed to analyze these semantic citation data. This service infrastructure will contain the following major services: Author Reference Service (to add URL to authors), Journal Reference Service (to add DOIs or URL for journals), Article Reference Service (to add URL to articles), and Citation-paper Reference Service (to use RDF links to link citations and papers).

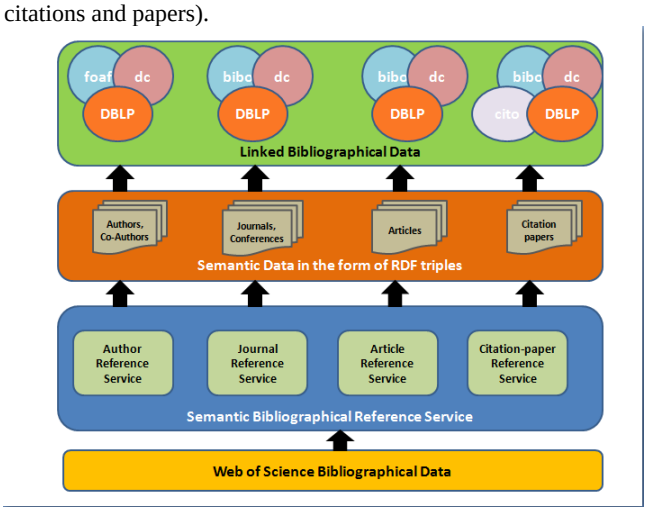

Figure 1. Semantic Bibliographical Reference Service Infrastructure

In this demo, we did the test on data from the Web of Science data were converted based on our reference services and stored in Jena RDF triples. This semantic repository will be used to provide information retrieval, inference, and statistics interfaces to enable intelligent semantic queries, reasoning, and semantic citation analysis. The example of part of RDF triples for one bibliographical paper is shown here:

<doi:10.1007/s00778-008-0125-y> a bibo:Article ;
dc:title "SW-Store: a vertically partitioned DBMS for Semantic Web data management"@en ;
dc:date "2009" ;
dc:subject "semantic web" ;
dc:subject "data management" ;
dc:subject "database" ;
dc:isPartOf <urn:issn:1066-8888> ;
bibo:volume "42" ;
bibo:issue "4" ;
bibo:pageStart "388" ;
bibo:pageEnd "391" ;
dc:creator <http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/a/Abadi:Daniel_J=.html> ;
bibo:authorList (<http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Marcus_0002:Adam.html>
<http://www.informatik.uni- trier.de/~ley/db/indices/a-tree/h/Hollenbach:Katherine_J=.html>) .

3. CONCLUSION

This paper describes a novel referencing infrastructure to generate semantic-powered cross-referencing and bring millions of bibliographical papers and their citations into semantic linked data. In the future, research profiling and intelligent bibliographical searching can be enabled: for example, queries like give me: ***the authors who use dataset X; the authors who develop algorithm Y; the authors who cite paper Z; the collaborators of author A; the highly cited domain experts in field A; the potential collaborators in my field; how many times dataset X has been cited/used; the names that author A acknowledged***, and so on can be processed. Furthermore, semantic mining on novelty detection, hot topic detection, dynamics of research, and topic clustering based on the large-scale Terabyte database can be conducted later on. In the future, with the continuous accumulation of data, we hope to cluster the research terms periodically and tracking timeline of topic, ranking authors together with their topics. We also need to test the scalability and efficiency of our approaches and link our data with other Linked Open Data sets.

4. REFERENCES

[1] Sure, Y., Bloehdorn, S., Hasse, P., Hartmann, J., & Oberle, D. (2005). The SWRC ontology: Semantic web for research communities. *Lecture Notes in Computer Science, 3808*, 218-231.

[2] Dabrowski, M., Synak, M., & Kruk, S. R. (2009). Bibliographic ontology. In S. R. Kruk & B. McDaniel (Eds.), *Semantic Digital Libraries* (pp. 103-122). New York: Springer.

[3] Shotton, D. (2008). CiTO, the Citation Typing Ontology, and its use for annotation of reference lists and visualization of citation networks. Retrieved July 12, 2009 from http://imageweb.zoo.ox.ac.uk/pub/2008/publications/Shotton_ISMB_BioOntology_CiTO_final_postprint.pdf

[4] Chaudhri, T. (2009). Assessing FRBR in Dublin Core application profiles. *Ariadne, 58*. Retrieved July 11, 2009 from http://www.ariadne.ac.uk/issue58/chaudhri