SwetoDblp ontology of Computer Science publications

Boanerges Aleman-Meza [a,*], Farshad Hakimpour [a], I. Budak Arpinar [a], Amit P. Sheth [b]

[a] *LSDIS Lab, Computer Science Department, University of Georgia, Athens, GA, USA*
[b] *Kno.e.sis Center, College of Engineering and Computer Science, Wright State University, Dayton, OH, USA*

Abstract

SwetoDblp is a large populated ontology with a shallow schema yet a large number of real-world instance data. We describe how such ontology is built from an XML source and how it can be maintained. Instead of a one-to-one mapping from XML to RDF, the creation of the ontology emphasizes the addition of relationships and the value of URIs. SwetoDblp is publicly available online. We also summarize research efforts that have used or are using this freely available community resource.

Keywords: Ontology; Semantic analytics; RDF; XML; Ontology population

1. Introduction

Semantic technologies are gaining wider use in Web applications [16,8,10]. Development of Semantic Web applications typically involves processing of data represented using or supported by ontologies, which are a central part of Semantic Web technologies. Both shallow and deep ontologies are needed [13]. Shallow ontologies contain large amounts of data and the concepts and relations are unlikely to change. Deep ontologies contain smaller (or not any) amounts of data but the actual concepts and relations require extensive efforts on their building and maintenance. Whether represented in RDF or OWL, real-world and medium-to-large datasets are required by practical applications. In particular, ontologies with a substantial number of relationships among instance data provide the means for discovery and analytics. In this paper, we describe our approach on building a shallow ontology called SwetoDblp consisting of bibliography of Computer Science publications where the main data source is the DBLP bibliography data (http://dblp.uni-trier.de/).

SwetoDblp builds upon our previous experience on creating and using Semantic Web Technology Evaluation Ontology (SWETO) [1]. The creation of SwetoDblp is done through a SAX-parsing process that performs various domain-specific transformations on a large XML document (available at the DBLP website) to produce RDF. The schema-vocabulary part of the ontology is a subset of an ontology used by the back-end system of the LSDIS Lab's publications library. This schema adopts major concepts and relationships from FOAF and Dublin Core and extends them where needed. In addition, we used OWL vocabulary to indicate equivalence of classes and relations to six other vocabularies such as the AKTors publication ontology [14] (using `owl:equivalentClass` and `owl:equivalentProperty`).

SwetoDblp is publicly available for download together with the additional datasets that were used for its creation (http://lsdis.cs.uga.edu/projects/semdis/swetodblp/). The additional datasets facilitated the integration and addition of many relationships and entities in SwetoDblp. Thus, the resulting ontology is enriched by incorporating other data sources.

2. Ontology development

Our goal was to create a dataset in RDF from an XML document containing DBLP information about publications and their authors. This required mapping from syntax and structure of XML into abstract concepts and relationships in RDF. The hierarchical structure of XML documents implies relationships from parent to children elements. However, such relationships depend

* Corresponding author. Tel.: +1 706 542 4772; fax: +1 614 495 0023.
E-mail addresses: boanerg@cs.uga.edu (B. Aleman-Meza), farshad@uga.edu (F. Hakimpour), budak@cs.uga.edu (I. Budak Arpinar), amit.sheth@wright.edu (A.P. Sheth).

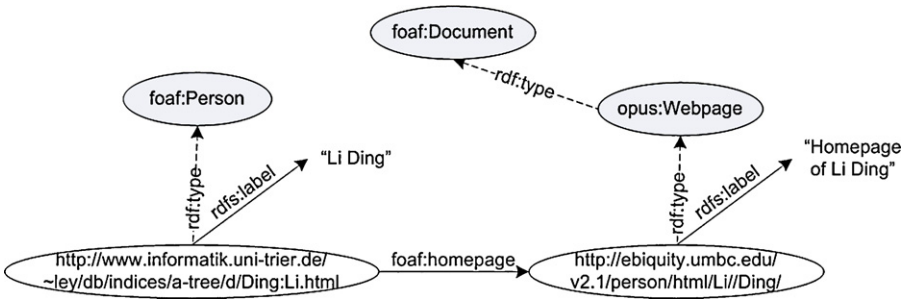

Fig. 1. Using existing vocabulary: example of foaf: homepage and foaf:Person.

upon human interpretation. Hence, one-to-one mappings from XML entities to RDF instances or literals would not be sufficient, as domain specific mappings might be needed. SwetoDblp goes beyond one-to-one mappings by taking care of special cases that help in producing an arguably more useful RDF dataset. We used the following guidelines for the creation of SwetoDblp.

- Creation of URIs that can be easily recognized and/or reused in other applications or datasets.
- Usage of existing vocabulary whenever is possible (such as FOAF and Dublin Core).
- Integration of relationships and entities from additional data sources.

These guidelines provide the general framework under which various domain specific mappings were implemented for the creation of SwetoDblp, as explained next.

2.1. Creation of URIs that have potential to be reused by other datasets

In the original XML document, the names of persons appear as plain literal values such as `<author> Li Ding </author>`. In SwetoDblp, each of these is represented as an RDF resource having its own URI. Our goal was to create URIs so that they can be reused by other datasets based on the assumption that the URI of choice will likely be the URL pointing to the author's DBLP entry on the Web. For example, the XML snippet above gets transformed into a resource with URI http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/d/Ding:Li.html (see Fig. 1). Similarly, the URI created for publications is the URL of the BibTeX entry at the DBLP site.

2.2. Usage of existing vocabulary

During the creation of SwetoDblp, we made an effort to reuse existing vocabulary whenever possible. For example, if the homepage of an author is available in the original XML document, then such relationship is kept in the resulting RDF by using `foaf:homepage`. In addition, the 'homepage' is represented as an RDF resource (with the URL as its URI); this domain-specific mapping automatically assigns a label to the homepage resource with the prefix "Homepage of". Fig. 1 illustrates an example of such mapping.

In very few cases, the data from DBLP indicates that a person can be referred to by more than one name. Examples include 'Tim Finin"and 'Timothy W. Finin". In SwetoDblp, such names are explicitly represented with an `owl:sameAs` relationship. This is the only relationship from the OWL vocabulary that is used in SwetoDblp instance data. We recognize that entity disambiguation techniques (or reference reconciliation) could further improve the quality of SwetoDblp yet this is out of the scope of this paper. Fig. 2 provides an example of relating two names of a person (hyphened last name) through an `owl:sameAs` relationship.

2.3. Inclusion of relationships and entities from other data sources

There are currently three other data sources used in the creation of SwetoDblp. The first is a Universities dataset that is used to determine and then explicitly add an affiliation relationship to a person either from the homepage of the person, or from 'note' elements appearing in the DBLP XML document such as `<note>The Open University, Milton Keynes, UK</note>`.

The universities dataset consists of two components. The first is a list of universities obtained from the Web source:

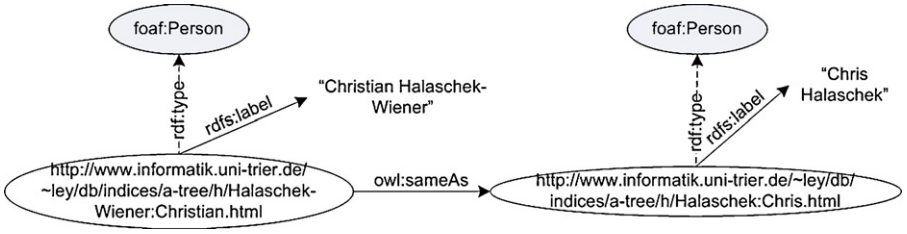

Fig. 2. Two author entities with a 'sameAs' relationship.

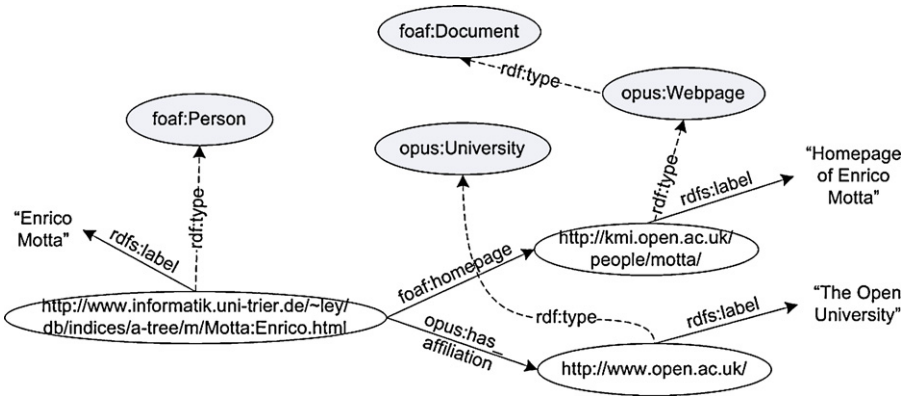

Fig. 3. Illustration of affiliation information extracted from a homepage and by using the Universities dataset.

http://www.google.com/intl/en/universities.html. The second component of the Universities dataset is a small, manually created list of universities containing synonyms and alternative spellings. It includes universities not listed in the Web source mentioned before. Fig. 3 illustrates an example of affiliation information added to a person by means of extracting it from his homepage and matching it against the Universities dataset. The usage of this dataset facilitated the integration of 4272 affiliation relationships.

The second dataset is about Publishers. This dataset is used to create a relationship from literal values such as `<publisher>McGraw-Hill</publisher>` to an RDF publisher entity with an URI that points to the actual website of the publishing company (e.g., www.pbg.mcgraw-hill.com/). This dataset was created manually with the most commonly appearing names of publishers in the original XML document from DBLP, but more publisher entities were added to cover all publishers that appear in DBLP data. Nevertheless, we could not locate the website of a small number of (arguably local or out of business) publishers. We assigned them an arbitrary URI using the http://example.org domain name as prefix.

The third dataset is of information about Series such as Lecture Notes in Computer Science and CEUR Workshops. This small dataset of 87 series entities was created manually to facilitate the creation of 'in series' relationships based on a lookup operation on literal values such as `<series>Dagstuhl Seminar Proceedings</series>`. A total of 5362 relationships were added from publication to series in SwetoDblp.

These datasets are all represented in RDF to allow for easy inclusion of synonyms. A lookup operation on the respective datasets is in most cases the key to establish relationships that enrich SwetoDblp. However, the relationships added depend upon the coverage of these datasets, which at the moment is quite good for the Series and Publisher datasets. The coverage of the Universities dataset mainly depends upon the Web source used to collect universities that consists of 781 Universities from all over the world. In fact, we have incrementally updated the Universities dataset with more entities and synonyms. Although this requires manual addition or update, the benefit is that such update is done only once regardless of how many times a given synonym appears during the conversion process.

2.4. Size and statistics of SwetoDblp

Number of available large size ontologies has recently been increasing. UniProt (www.pir.uniprot.org/) and Glyco/Propreo [12] are ontologies with well over one million entities. Other large ontologies such as TAP [4] and Lehigh Benchmark (swat.cse.lehigh.edu/projects/lubm/) have also proven useful for developments and evaluations in Semantic Web research. Lehigh Benchmark is a suitable dataset for performance evaluation but it is a synthetic dataset. Real world datasets are essential for evaluation in applications development where result of a query on real data is critical. SwetoDblp is a real world data set with a large number of entities. Table 1 provides a summary of size and statistics of the major types of entities and relationships in SwetoDblp (see also Figs. 4 and 5).

2.5. Maintenance

Our goal is to update SwetoDblp on a regular basis (e.g., monthly) as newer versions of the XML document become avail-

Table 1
Size and statistics of SwetoDblp in terms of number of entities and relationships

Class	Number of entities
foaf:Person	485,577
Articles in proceedings	482,641
Journal articles	296,212
Webpage of person	8,720
Book chapter	2,530
Book	1,214
Proceedings	7,935
Property	**Number of relationships**
Authored (or edited) publication	1,896,918
Contained in proceedings	362,907
Cites publication	112,373
dc:publisher	9,351
foaf:homepage	8,720
In series	5,381
owl:sameAs	1,647
Affiliation (foaf:workplaceHomepage)	4,554
Chapter of	1,248

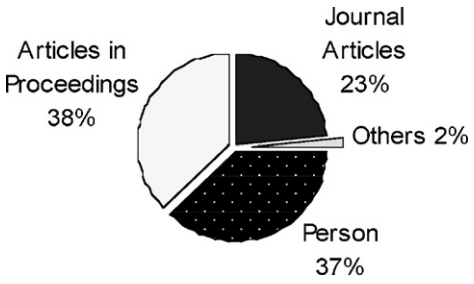

Fig. 4. Number of instances in the dataset.

able. We are also keeping 'frozen' versions of the dataset. The task of maintenance is straightforward. It involves processing the latest XML data from DBLP and updating, if necessary, the datasets of Publishers, Universities and Series. In fact, this allows for incremental maintenance of these datasets such as the additions of synonyms. This task took less than half an hour in the last update of the latest version of SwetoDblp (as of October 2006). SwetoDblp is available under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 license.

3. Related work

An earlier effort for representing DBLP data in RDF uses XSL transformations (http://sw.deri.org/~aharth/2004/07/dblp/). Such transformation requires a machine with several gigabytes of memory (as indicated by the authors). Our method to create SwetoDblp works well on a machine with one gigabyte of memory (it takes about 6 min to complete). Furthermore, their conversion performs only one of the domain specific mappings of Section 2 (i.e., homepage of person entities) and although it creates a URI for each person, we believe that a URI that points to the actual DBLP entry on the Web would be more useful.

Our previous work of SWETO ontology [1] contained only a subset of DBLP data and did not include any of the various improvements listed in Section 2. SWETO was created using data extraction tools of Semagix Freedom [15] whereas SwetoDblp is created without commercial tools.

D2RQ is a general toolkit that can be used for mapping relational data or XML to RDF and has demonstrated conversion of the DBLP data in XML to RDF [3]. However, such conversion does not perform any of the domain specific mappings mentioned in Section 2. XSLT processing can be used to convert XML to RDF (e.g., GRDDL [6,17]). However, most of the domain specific processing in our method would not be possible by using XSLT alone. For example, the creation of affiliation relationships requires external datasets of university names and their synonyms.

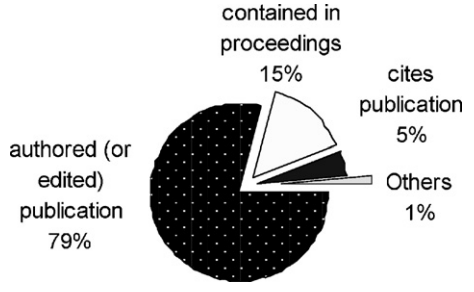

Fig. 5. Number of relations in the dataset.

Mappings from XML to ontologies can be automatically generated by relying on DTD and using synonyms to facilitate finding better matches from element names in the DTD to concepts in the target ontology [19]. However, human needs to select the mappings, and some of the domain-specific transformations in our approach would not be possible following that approach. Translation rules can also be generated from the DTD to generate RDF. A recent effort addresses semantic, structural and schematic translation rules [9]. However, such an approach does not consider that some literal values should become entities in RDF as in the cases of author or publisher names. Earlier work on creation of RDF from XML [7] seems to consider various aspects as we did. For example, a careful selection of the XML elements to be converted into RDF by keeping in mind the RDF Schema leads to useful RDF statements. Such work, however, did not put emphasis on the creation of URIs nor in the addition of entities and relationships by using external datasets.

Overall, existing methods would not be sufficient to create our target RDF output due to the fact that previous research has attempted to create generalized methods for converting XML to RDF. Our method contains various domain specific features that tie the conversion algorithm to the particular XML format used by DBLP yet we believe this produces a more useful dataset with relatively low human effort.

4. Applications

Based on our previous effort on the large ontology SWETO that is itself a precursor to SwetoDblp, we anticipate similar or wider usage for SwetoDblp. SWETO was used by researchers in our lab as well as researchers elsewhere (e.g., visualization of schema and instances [18]). So far, we are aware of the following projects/applications that make use of SwetoDblp.

- SwetoDblp is being used to test the iSPARQL query engine. This engine provides a myriad of similarity measures to search for similar objects in datasets. In the case of SwetoDblp, the interest is finding similar publications to a given publication request. SwetoDblp is also used for testing optimization techniques that have been implemented over the ARQ SPARQL engine (jena.sourceforge.net) where the goal is to improve query performance. More details on the iSPARQL project are available at www.ifi.unizh.ch/ddis/isparql.html.
- The AquaLog tool was built when there were not so many large ontologies available. The SwetoDblp and SWETO ontologies are being used to test the new generation tools beyond AquaLog that make use of large-scale ontology repositories (see PowerMap and PowerAqua at the AquaLog website, kmi.open.ac.uk/technologies/aqualog/).

- SwetoDblp has been used as the main dataset in our recent research on disambiguation of people names in text [5]. The domain of SwetoDblp is directly applicable for the scenario of disambiguating appearances of researchers names in posts of the DBWorld mailing list.
- SwetoDblp has been identified and is listed as a relevant dataset for the task of finding experts (see ExpertFinder initiative, www.rdfweb.org/topic/ExpertFinder).

In addition, research efforts that have made use of DBLP data (e.g., [2,11]) could benefit by using the semantic metadata in SwetoDblp. The scope and purpose of the ontology is not limited to analysis and querying of bibliographic data. We believe SwetoDblp can be a significant Semantic Web community resource for researchers in semantic search, digital libraries as well as semantic analytics and browsing.

Acknowledgment

This work is partially funded by NSF-ITRIDM Award#0325464 titled 'SemDIS: Discovering Complex Relationships in the Semantic Web.'

References

[1] B. Aleman-Meza, C. Halaschek, A. Sheth, I.B. Arpinar, G. Sannapareddy, SWETO: large-scale semantic web test-bed, in: Int'l Workshop on Ontology in Action, Banff, Canada, 2004, pp. 490–493.

[2] O. Alonso, S. Banerjee, M. Drake, GIO: a semantic web application using the information grid framework, in: 15th Int'l Conference on World Wide Web, Edinburgh, Scotland, May, 2006, pp. 857–858.

[3] C. Bizer, D2R MRP—a database to RDF mapping language, in: 12th International World Wide Web Conference (Poster), Budapest, Hungary, 2003.

[4] R.V. Guha, R. McCool, TAP: a semantic web test-bed, J. Web Seman. l (1) (2003) 81–87.

[5] J. Hassell, B. Aleman-Meza, I.B. Arpinar, Ontology-driven automatic entity disambiguation in unstructured text, in: Fifth International Semantic Web Conference, Athens, GA, USA, November 5–9, 2006.

[6] D. Hazaël-Massieux, Bridging XHTML, XML and RDF with GRDDL, XTech 2005: XML, the Web and Beyond, Amsterdam, Netherlands, 2005.

[7] M. Klein, Interpreting XML documents via an RDF schema ontology, in: 13th Int'l Workshop on Database and Expert Systems Applications, Aix-en-Provence, France, September 2–6, 2002, pp. 889–894.

[8] Y.L. Lee, Apps Make Semantic Web a Reality, SD Times, 2005.

[9] C. Li, T.W. Ling, From XML to semantic Web, in: 10th International Conference on Database Systems for Advanced Applications, Beijing, China, April 17–20, 2005, pp. 582–587.

[10] E. Miller, The semantic Web is here, in: Keynote at the Semantic Technology Conference 2005, San Francisco, California, USA, 2005.

[11] M.A. Nascimento, J. Sander, J. Pound, Analysis of SIGMOD's co-authorship graph, SIGMOD Rec. 32 (September (3)) (2003) 8–10.

[12] S.S. Sahoo, C. Thomas, A.P. Sheth, W.S. York, S. Tartir, Knowledge modeling and its application in life sciences: a tale of two ontologies, in: 15th International World Wide Web Conference, Edinburgh, Scotland, May 23–26, 2006.

[13] N.R. Shadbolt, T. Berners-Lee, W. Hall, The semantic Web revisited, IEEE Intell. Syst. 21 (May/June (3)) (2006) 96–101.

[14] N.R. Shadbolt, N. Gibbins, H. Glaser, S. Harris, m.c. Schraefel, Walking through CS AKTive Space: a demonstration of an integrated Semantic Web application, J. Web Seman. 1 (4) (2004) 415–419.

[15] A.P. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, Y. Warke, Managing semantic content for the Web, IEEE Internet Comput. (July/August) (2002) 80–87.

[16] A.P. Sheth, From semantic search & integration to analytics, in: Proceedings of in Semantic Interoperability and Integration, IBFI, Schloss Dagstuhl, Germany, September 19–24, 2004.

[17] C.M. Sperberg-McQueen, E. Miller, On Mapping from Colloquial XML to RDF using XSLT, Extreme Markup Languages 2004, Montreal, Quebec, Canada, 2004.

[18] K. Tu, M. Xiong, L. Zhang, H. Zhu, J. Zhang, Y. Yu, Towards imaging large-scale ontologies for quick understanding and analysis, in: Fourth International Semantic Web Conference, 2005, pp. 702–715.

[19] L. Xiao, L. Zhang, G. Huang, B. Shi, Automatic mapping from XML documents to ontologies, in: Fourth Int'l Conference on Computer and Information Technology, Wuhan, China, September 14–16, 2004.