Publishing Bibliographic Data on the Semantic Web using BibBase

Reynold S. Xin[×], Oktie Hassanzadeh[+], Christian Fritz[§]
Shirin Sohrabi[+], Yang Yang[+], Minghua Zhao[+], Renée J. Miller[+]

[+]Department of Computer Science, University of Toronto
{oktie,sohrabi,c7yangya,mzhao,miller}@cs.toronto.edu
[×]Department of EECS, University of California, Berkeley
rxin@berkeley.edu
[§]Information Sciences Institute, University of Southern California
fritz@isi.edu

Abstract. We present BibBase, a system for publishing and managing bibliographic data available in BibTeX files on the Semantic Web. BibBase uses a powerful yet light-weight approach to transform BibTeX files into rich *Linked Data* as well as custom HTML and RSS code that can readily be integrated within a user's website. The data can instantly be queried online on the system's SPARQL endpoint. In this demo, we present a brief overview of the features of our system and outline a few challenges in the design and implementation of such a system.

Keywords: Bibliographic Data Management, Linked Data, Data Integration

1 Introduction

Management of bibliographic data has received significant attention in the research community. Many online systems have been designed specifically for this purpose, e.g., BibSonomy [9] and CiteSeer [10]. The work in the semantic web community in this area has also resulted in several tools (such as BiBTeX to RDF conversion tools [5]), ontologies (such as SWRC [7] and the Bibliographic Ontology [6]) and data sources (such as DBLP Berlin [8]). These systems, tools, and data sources are widely being used and have considerably simplified and enhanced many bibliographic data management tasks such as data curation, storage, retrieval, and sharing of bibliographic data.

Despite the success of the above-mentioned systems, very few individuals and research groups publish their bibliographic data on their websites in a structured format, particularly following the principles of Linked Data [1] which mandate the use of HTTP dereferenceable URIs and structured (RDF) data to convey the semantics of the data This is mainly due to the fact that existing systems either are not designed to be used within an external website, or they require expert users to set up complex software systems on machines that meet the requirements of this software. BibBase aims to fill this gap by providing several distinctive features that our demo will illustrate.

2 Light-weight Linked Data publication

BibBase makes it easy for scientists to maintain publication lists on their personal web site. Scientists simply maintain a BiBTeX file of their publications, and BibBase does the rest. When a user visits a publication page, BibBase dynamically generates an up-to-date HTML page from the BiBTeX file, as well as rich Linked Data with resolvable URIs that can be queried instantly on the system's SPARQL endpoint. We have chosen to use an augmented version of MIT's BiBTeX ontology definition to publish data in RDF[1].

Compared to existing Linked Data publication tools, this approach is notably easy-to-use and light-weight, and allows non-expert users to create a rich linked data source without any specific server requirements, the need to set up a new system, or define complex mapping rules. All they need to know is how to create and maintain a BiBTeX file and there are tools to help with that.

It is important to note that this ease of use does not sacrifice the quality of the published data. In fact, although the system is light-weight on the users' side, BibBase performs complex processing of the data in the back-end. When a new or updated BiBTeX file arrives, the system transforms the data into several structured formats using our ontology, assigns URIs to all the objects (authors, papers, venues, etc.), performs duplicate detection and semantic linkage, and maintains and publishes provenance information.

3 Duplicate Detection

BibBase needs to deal with several issues related to the heterogeneity of records in a single BiBTeX file, and across multiple BiBTeX files. BibBase uses existing duplicate detection techniques in addition to a novel way of managing duplicated data following the Linked Data principles.

Within a single BiBTeX file, the system uses a set of rules to identify duplicates and fix errors. For example, if a BiBTeX file has two occurrences of author names "J. B. Smith" and "John B. Smith", the system matches the two author names and creates only a single author object. In this example, the assumption is that the combination of the first letter of first name, middle name, and last name, "JBSmith", is a unique identifier for a person in a single file.

For identification of duplicates across multiple BiBTeX files, the assumptions made for local duplicate detection may not hold. Within different publication lists, "JBSmith" may (or may not) refer to the same author. BibBase deals with this type of uncertainty by having a *disambiguation* page on the HTML interface that informs the users looking for author name "J. B. Smith" (by looking up the URI `http://data.bibbase.org/author/j-b-smith`) of the existence of all the entities with the same identifier, and having `rdfs:seeAlso` properties that link to related author entities on the RDF interface.

[1] Notably, the MIT's BiBTeX ontology (`http://zeitkunst.org/bibtex/0.1/`) is extended to allow description of the order of authors, unlike some widely-used bibliographic ontologies. We also provide `owl:sameAs` and `umbel:isLike` links to the other existing bibliographic ontologies. The new ontology definition is available at `http://data.bibbase.org/ontology`.

Duplicate detection, also known as *entity resolution*, *record linkage*, or *reference reconciliation* is a well-studied problem and an active research area [3]. We use some of the existing techniques to define local and global duplicate detection rules, for example using fuzzy string similarity measures [2] or semantic knowledge for matching conference names and paper titles [4].

4 Discovering semantic links to external data sources

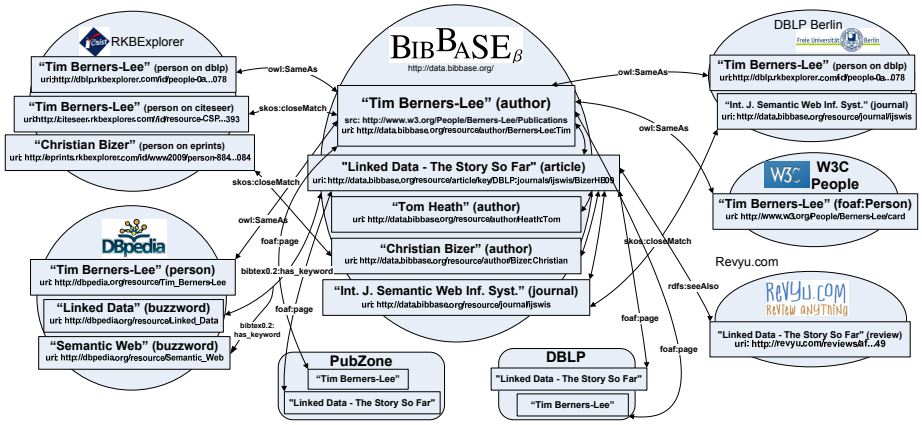

Fig. 1. Sample entities in BibBase interlinked with several related data sources.

In order to publish our data *in the Web*, not just *on the Web*, to avoid creation of an isolated data silo, we need to discover links from the entities in BibBase to entities from external data sources. Figure 1 shows a sample of entities in BibBase and several possible links to related Linked Data sources and web pages. In order to discover such links, similar to our duplicate detection approach, we can leverage online and offline solutions. The online approach mainly uses a dictionary of terms and strings that can be mapped to external data sets. A similar approach is used to match abbreviated venues, such as "ISWC" to "International Semantic Web Conference". The dictionaries (or ontology tables) are maintained inside BibBase, and derived from sources such as DBpedia, Wordnet, and DBLP. We also allow the users to extend the dictionaries by `@string` definitions in their BiBTeX files. Offline link discovery is performed using existing link discovery tools [4]. Our demo will allow users to interactively add BiBTeX entries, then view and query the semantically annotated entry and discovered links.

5 Additional Features

The success of BibBase as a Linked Data source depends on scientists using BibBase for their publications pages. To further entice scientists to do so, BibBase sports a number of additional features that make it an attractive proposition.

- Storage and publication of provenance information, i.e., metadata about the source of each entity and each link in the data.
- Dynamic grouping of entities based on attributes (e.g., by year or keyword).
- An RSS feed, allowing anyone to receive notifications whenever a specified scientist publishes a new paper.
- A DBLP fetch tool that allows scientists who do not yet have a BiBTeX file to obtain their DBLP publications to start using BibBase right away.
- Statistics regarding users, page views, and paper downloads.

We enable users to provide feedback on the quality of data and links. By providing feedback, users will not only improve the quality of the data published on their own websites, they will also help create a very high-quality data source in the long run that could become a benchmark for the notoriously hard task of evaluating duplicate detection and semantic link discovery systems.

6 Conclusion

In this demonstration, we will present BibBase, a system for light-weight publication of bibliographic data on personal or research group websites, and management of the data using existing semantic technologies as a result of the complex *triplification* performed inside the system. BibBase extends the Linked Data cloud with a data source that unlike existing bibliographic data sources, allows online manipulation of the data by non-expert users. We plan to continue to extend the features of BibBase. A list of currently implemented and upcoming experimental features is available at `http://wiki.bibbase.org`.

References

1. T. Berners-Lee. Linked Data - Design Issues. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006. [Online; accessed 14-June-2010].
2. A. Chandel, O. Hassanzadeh, N. Koudas, M. Sadoghi, and D. Srivastava. Benchmarking Declarative Approximate Selection Predicates. In *ACM SIGMOD Int'l Conf. on the Mgmt. of Data*, pages 353–364, 2007.
3. A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007.
4. O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang. A Framework for Semantic Link Discovery over Relational Data. In *Proc. of the Conf. on Information and Knowledge Management (CIKM)*, pages 1027–1036, 2009.
5. I. Herman. BibTeX in RDF. `http://ivan-herman.name/2007/01/13/bibtex-in-rdf/`, 2007. [Online; accessed 14-June-2010].
6. `http://bibliontology.com/`.
7. `http://ontoware.org/swrc/`.
8. `http://www4.wiwiss.fu-berlin.de/dblp/`.
9. `http://www.bibsonomy.org/`.
10. `http://citeseer.ist.psu.edu/`.