BibBase Triplified*

http://data.bibbase.org

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

[+]Department of Computer Science
University of Toronto

[×]EECS Department
University of California, Berkeley

[§]Information Sciences Institute
University of Southern California

ABSTRACT

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

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services; H.2.4 [**Database Management**]: Systems

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, including but certainly not limited to, BibSonomy [13], CiteSeer [14], CiteULike [15], EPrints [16], Mendeley [17], PubZone [18], refbase [19] and RefWorks [20]. The work in the semantic web community in this area has also resulted in several tools (such as BiBTeX to RDF conversion tools [7]), ontologies (such as SWRC ontology [10]) and data sources (such as DBLP Berlin [11] and RKBExplorer [21]). 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], to provide users with HTTP dereferenceable URIs that provide structured (RDF) data as well as nicely formatted HTML pages. 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 [12] aims to fill this gap by providing several distinctive features as described in the following sections.

*Supported in part by NSERC.

I-SEMANTICS 2010, September 1-3, 2010 Graz, Austria

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. 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 a rich ontology, assigns URIs to all the objects (authors, papers, venues, etc.), performs duplicate detection and semantic linkage, and maintains and publishes provenance information as described below.

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. We refer to this phase as *local* duplicate detection. 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

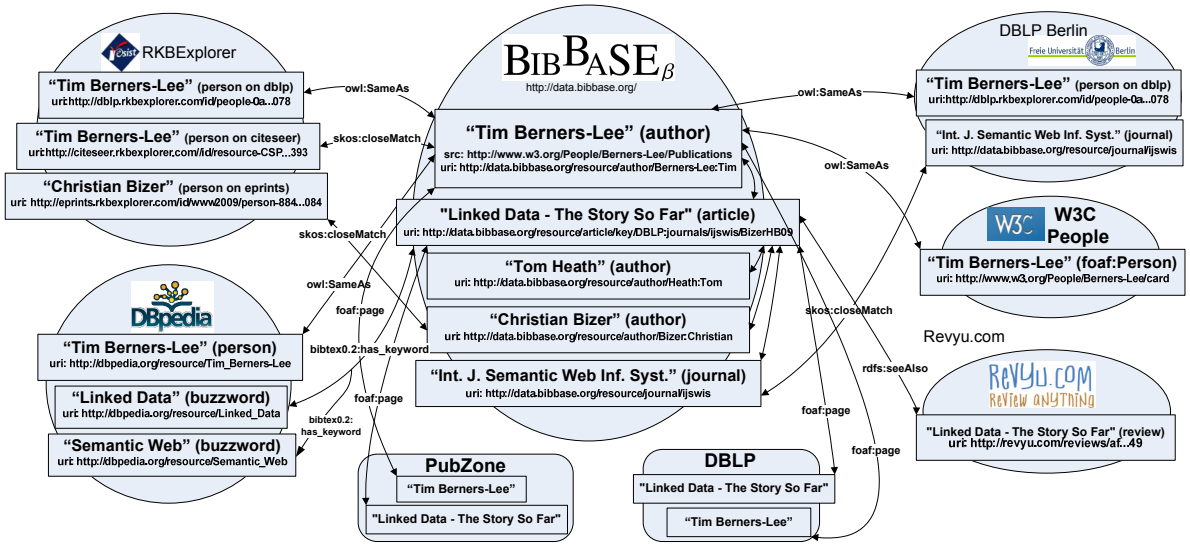

Figure 1: Sample entities in BibBase interlinked with several related data sources.

name, middle name, and last name, "JBSmith", is a unique identifier for a person in a single file. If this assumption does not hold for a specific user (which is unlikely) BibBase allows the user to distinguish two people with the same identifier by adding a number at the end of one of the author names.

For identification of duplicates across multiple BiBTeX files, which we call *global* duplicate detection, 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 `skos:closeMatch` and `rdfs:seeAlso` properties that link to related author entities on the RDF interface.

Duplicate detection, also known as *entity resolution*, *record linkage*, or *reference reconciliation* is a well-studied problem and an attractive research area [5]. We use some of the existing techniques to define local and global duplicate detection rules, for example using fuzzy string similarity measures [4] or semantic knowledge for matching conference names and paper titles [6]. In addition to the definition of rules and online duplicate detection, we also uses some graph-based duplicate detection techniques [8] (also known as *collective entity resolution* [2]) to identify duplicates in an offline manner. However, in order to avoid loss of user data as a result of imperfect data cleaning, the results of this process will be published as additional data on our system that result in disambiguation pages or `skos:closeMatch` predicates.

4. DISCOVERING SEMANTIC LINKS TO EXTERNAL DATA SOURCES

In order to publish our data *in the Web*, not just *on the Web*, to avoid creation of an isolated data silo, and to add BibBase data to the growing Linking Open Data cloud of data sources, 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. An important type of links comes from `keywords` in BiBTeX entries that can be used to relate publications to entries on DBpedia (and pages on Wikipedia), such as DBpedia entities of type `buzzword` shown in the example figure. 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, Freebase, Wordnet, and DBLP. We also allow the users to extend the dictionaries by `@string` definitions in their BiBTeX files, e.g.,

```
@string{ISWC={Proc. of the Int'l Semantic Web Conference(ISWC)}}
```

An offline link discovery can be performed using existing link discovery tools [6, 9].

5. PROVENANCE AND USER FEEDBACK

Another highlight of the features implemented in BibBase is storage and publication of provenance information, i.e., the source of each entity and each link in the data. This is of utmost importance in a system like BibBase where the data comes from several different users and BiBTeX files, and where (imperfect) automatic duplicate detection and linkage is performed over the data. Users will be able to see the source of entities and the facts about the entities. In addition, they will be able to fix their own BiBTeX files or provide feedback to the system and to other users who need to fix their files or provide additional information.

User feedback is another important aspect of BibBase Feedback is received in two forms. The first and major part of feedback comes from the BiBTeX files. For example, as stated in Sec. 3, users can distinguish different authors with the same name by adding a number to the end of their names. Similarly, as stated in Sec. 4, users can provide string equivalences to enhance our internal ontology tables for semantic linkage as well as duplicate detection. Another form of feedback comes from the web interface (currently under development), where the users can report broken links,

typos, or wrong duplicate detection outside the scope of their own BiBTeX entries.

By providing feedback, users will not only increase the quality of the data published on their own websites, they will also 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. BIBTEX ONTOLOGY DEFINITION

Using terms from existing vocabularies to publish data in RDF, is recognized as a "best practice" by the linked data community [3]. Several different vocabularies exist for bibliographic data. We have chosen to use one that is specifically designed for BiBTeX data (as opposed to the more general vocabularies), namely MIT's BiBTeX ontology definition available at `http://zeitkunst.org/bibtex/0.1/`. However, we also need to extend the vocabulary in several aspects to meet the requirements of our system, and address some shortcomings of existing ontologies (e.g., those noted by Herman [7]). Some of these changes include addition of a new class `keyword` and addition of a new property `firstAuthor` for publication entries which facilitate grouping and querying publications based on keywords and first authors. The new namespace will be available at `http://zeitkunst.org/bibtex/0.2/`.

7. 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 solution for this purpose.

- Dynamic, multi-level grouping of publications based on different attributes (e.g., by year or keyword).
- Customizable appearance via CSS style sheets.
- 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, including a list of most popular papers.

We are currently working on a few additional features such as a generic keyword search interface that accounts for spelling errors, abbreviations, and semantic mismatches, and a visual navigator for the RDF data specifically designed to find correlations between the authors, papers, and keywords. We will also create an RSS feed for every keyword being used, so that anyone can be notified when new papers in that area are published.

8. ONLINE DEMO

The HTML and RSS interface of BibBase, available at `http://bibbase.org`, is fully functional and in active use by several groups and individuals. As of June 2010, the data interface, `http://data.bibbase.org`, is still in the experimental phase and the available online demo only showcases a subset of the features explained in this report. Documentation and current status of the experimental features of the data interface are available at `http://wiki.bibbase.org`.

9. CONCLUSION

We presented 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 linking open 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`.

10. REFERENCES

[1] T. Berners-Lee. Linked Data - Design Issues. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006. [Online; accessed 14-June-2010].

[2] I. Bhattacharya and L. Getoor. Collective Entity Resolution in Relational Data. *IEEE Data Engineering Bulletin*, 29(2):4–12, 2006.

[3] C. Bizer, R. Cyganiak, and T. Heath. How to Publish Linked Data on the Web. `http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/`, 2007. [Online; accessed 14-June-2010].

[4] 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.

[5] 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.

[6] 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.

[7] I. Herman. BibTeX in RDF. `http://ivan-herman.name/2007/01/13/bibtex-in-rdf/`, 2007. [Online; accessed 14-June-2010].

[8] M. Herschel and F. Naumann. Scaling Up Duplicate Detection in Graph Data. In *Proc. of the Conf. on Information and Knowledge Management (CIKM)*, pages 1325–1326, 2008.

[9] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and Maintaining Links on the Web of Data. In *Proc. of the Int'l Semantic Web Conference (ISWC)*, pages 650–665, 2009.

[10] `http://ontoware.org/swrc/`.

[11] `http://www4.wiwiss.fu-berlin.de/dblp/`.

[12] `http://www.bibbase.org/`.

[13] `http://www.bibsonomy.org/`.

[14] `http://citeseer.ist.psu.edu/`.

[15] `http://www.citeulike.org/`.

[16] `http://www.eprints.org/`.

[17] `http://www.mendeley.com/`.

[18] `http://www.pubzone.com/`.

[19] `http://www.refbase.net/`.

[20] `http://www.refworks.com/`.

[21] `http://www.rkbexplorer.com/`.