Undefined 0 (0) 1
IOS Press

Publishing Bibliographic Data on the Semantic Web using BibBase

Editor(s): Carsten Keßler, University of Münster, Germany; Mathieu d'Aquin, The Open University, UK; Stefan Dietze, L3S Research Center, Germany
Solicited review(s): Kai Eckert, University of Mannheim, Germany; Antoine Isaac, Vrije Universiteit Amsterdam, The Netherlands; Jan Brase, German National Library of Science and Technology, Germany

Reynold S. Xin [a], Oktie Hassanzadeh [b,*], Christian Fritz [c], Shirin Sohrabi [b] and Renée J. Miller [b]
[a] *Department of EECS, University of California, Berkeley, California, USA*
E-mail: rxin@cs.berkeley.edu
[b] *Department of Computer Science, University of Toronto, 10 King's College Rd., Toronto, Ontario, M5S 3G4, Canada*
E-mail: {oktie,shirin,miller}@cs.toronto.edu
[c] *Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, California, USA*
E-mail: cfritz@parc.com

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 *Linked Data* as well as custom HTML code and RSS feed 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 paper, we present an overview of several features of our system. We outline several challenges involved in on-the-fly transformation of highly heterogeneous BiBTeX files into high-quality Linked Data, and present our solution to these challenges.

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 [16], CiteSeer [17], CiteULike [18], EPrints [20], Mendeley [22], PubZone [21], refbase [27] and RefWorks [23]. The work in the Semantic Web community in this area has also resulted in several systems (such as VIVO [10]), tools (such as BiBTeX to RDF conversion tools [13]), ontologies (such as the BIBO [26] and SWRC [25]) and data sources (such as DBLP Berlin [19] and RKBExplorer [24]). 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

*Corresponding author. E-mail: oktie@cs.toronto.edu.

require expert users to set up complex software systems on machines that meet the requirements of this software. BibBase [15] aims to fill this gap by providing several distinctive features as described in the following sections.

2. Light-Weight Linked Data Publication

BibBase makes it easy for scientists to maintain publication lists on their personal website. 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 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 particular, we use LinQuer [7] to specify linkage requirements in LinQL query language, and translate the specification into standard SQL queries that can run over our relational backend.

In addition to the definition of rules and online duplicate detection (i.e., detection of duplicates while the input BibTeX file is being processed), we use some graph-based duplicate detection techniques [11] (also known as *collective entity resolution* [2]) to identify duplicates in an offline manner (i.e., after the input file is submitted and processed). Such techniques take advantage of structure of the data and the currently identified duplicates to iteratively enhance the quality of the identified duplicates. For example, two authors with similar names that have many common co-authors (or co-authors that are identified as duplicates in previous phases) are more likely to be duplicates. Similarly, two papers with similar titles that have the same set of authors (or authors that are identified as duplicates) are more likely to refer to the same publication. In order to avoid loss of user data as a result of imperfect data cleaning, the results of this process will

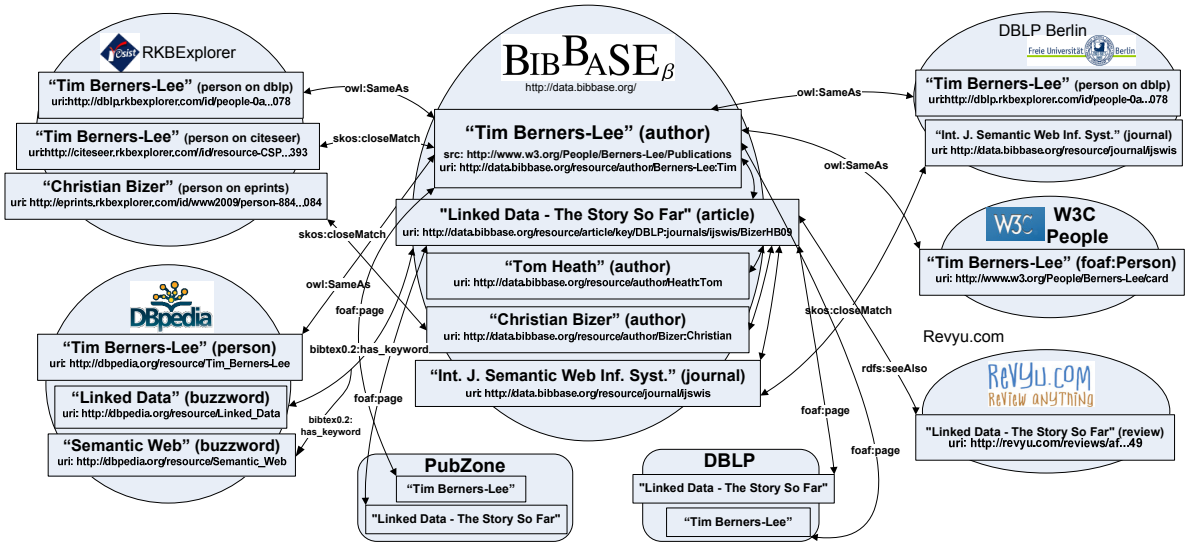

Fig. 1. Sample entities in BibBase interlinked with several related data sources.

be published as additional data on our system that result in disambiguation pages or `skos:closeMatch` predicates.

4. Discovering 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 [7,12], and more complex linkage requirements that require a longer processing time, and therefore cannot be performed online when the BibTeX files are submitted. Offline link discovery tasks are scheduled once a BibTeX file is submitted (or updated) or a change in the target data sources are detected. We use the Celery distributed task queue [14] for task management of offline duplicate detection and link discovery processes.

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 are able to see the source of entities and the facts about the entities. As a result, 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.

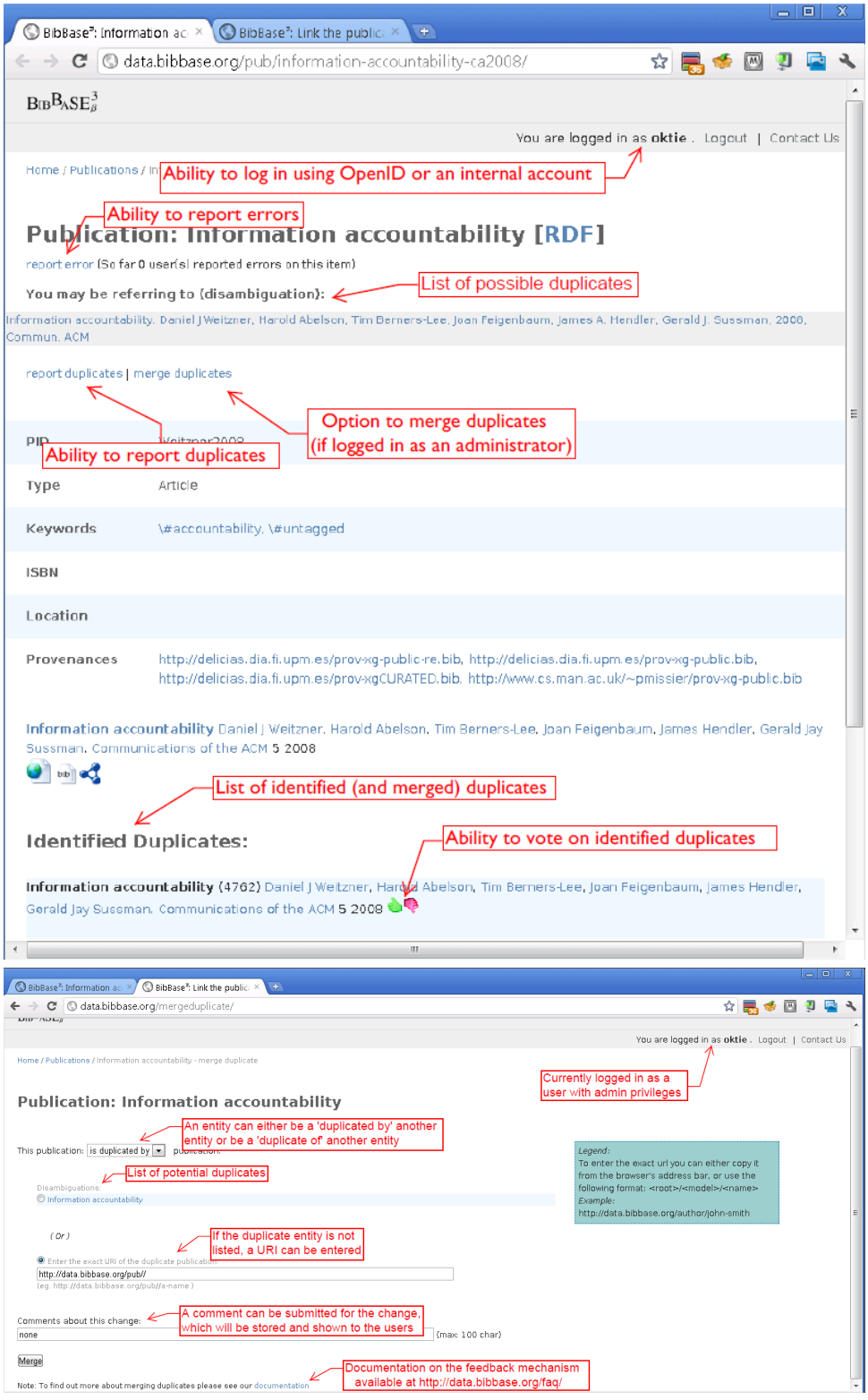

Fig. 2. Data browse interface (top) and admin interface for merge (bottom).

For example, as stated in Section 3, users can distinguish different authors with the same name by adding a number to the end of their names. Similarly, as stated in Section 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, where the users can report errors such as broken links, typos, or wrong duplicate detection outside the scope of their own BiBTeX entries. Furthermore, the users are able to vote for or against existing identified duplicates or discovered links. Figure 2 shows a snapshot of our Web interface for a publication that has a potential duplicate, and the admin interface for merging the duplicates. Currently, there are three class of users in our system:

- Anonymous users: all users can report errors or vote. The IP address of the users are recorded, and each IP address will not be able to vote more than once for each link, duplicate, or entity.
- Authenticated Users: users can also log in using their Open ID (an open standard that allows users to be authenticated in a decentralized manner, using their existing Google accounts for example). Feedback from the authenticated users are marked accordingly and verified with higher priority.
- Administrators: trusted users are assigned as administrators who can log in to verify the provided feedback, and perform specific actions based on the provided feedback. Our admin interface ranks the feedbacks on duplicates and external links based on type of the users (whether they are authenticated), and other criteria, and allows efficient verification of the users' feedback.

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 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 and build on one that is specifically designed for BiBTeX data (as opposed to the more general vocabularies), namely MIT's BiBTeX ontology [9]. We have extended 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 [8]). Some of these changes include:

- Addition of several new classes such as `Keyword`, `Organization`, `Language`, `Journal` and `Author` that were defined as properties with string value ranges in the previous version. This facilitates grouping and querying of these attributes of publications.
- Addition of a new class `Authorship` with properties `hasPosition` and `hasAuthor` to record the order of authors for a publication.
- Addition of new properties `hasAuthorship` and `firstAuthor` for publication entries in order to facilitate querying based on order of authors or the first author only.

The new namespace (BiBTeX 2.0) is available at `http://data.bibbase.org/ontology/`. The OWL-Lite version of the ontology is available at `http://data.bibbase.org/ontology/owl/`. We also publish the relationship between the classes and properties of our ontology with the two other widely-used bibliographic ontologies, namely BIBO [26] and SWRC [25]. Figure 3 shows a subset of the relationship between these ontologies. Note that several classes in our extended ontology are also present in these ontologies, whereas some do not exist in these or other ontologies.

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.

BibTeX 2.0	BIBO	SWRC
http://data.bibbase.org/ontology/#Publication	http://purl.org/ontology/bibo/Document	http://swrc.ontoware.org/ontology#Publication
http://data.bibbase.org/ontology/#Article	http://purl.org/ontology/bibo/Article	http://swrc.ontoware.org/ontology#Article
http://data.bibbase.org/ontology/#Book	http://purl.org/ontology/bibo/Book	http://swrc.ontoware.org/ontology#Book
http://data.bibbase.org/ontology/#hasTitle	http://purl.org/dc/terms/title	http://swrc.ontoware.org/ontology#title
http://data.bibbase.org/ontology/#hasPages	http://purl.org/ontology/bibo/pages	http://swrc.ontoware.org/ontology#pages
http://data.bibbase.org/ontology/#hasVolume	http://purl.org/ontology/bibo/volume	http://swrc.ontoware.org/ontology#volume
http://data.bibbase.org/ontology/#hasAddress	Does not exist	http://swrc.ontoware.org/ontology#address"
http://data.bibbase.org/ontology/#hasEdition	http://purl.org/ontology/bibo/edition	http://swrc.ontoware.org/ontology#edition
http://data.bibbase.org/ontology/#hasLocation	Does not exist	http://swrc.ontoware.org/ontology#location
http://data.bibbase.org/ontology/#hasPublisher	http://purl.org/dc/terms/publisher	http://purl.org/dc/elements/1.1/publisher
http://data.bibbase.org/ontology/#hasJournal	Does not exist	http://swrc.ontoware.org/ontology#journal
http://data.bibbase.org/ontology/#hasSchool	Does not exist	http://swrc.ontoware.org/ontology#school
http://data.bibbase.org/ontology/#hasLanguage	http://purl.org/dc/terms/language	http://purl.org/dc/elements/1.1/language
http://data.bibbase.org/ontology/#hasFirstAuthor	Does not exist	Does not exist
http://data.bibbase.org/ontology/#hasAuthorship	Does not exist	Does not exist
http://data.bibbase.org/ontology/#Authorship	Does not exist	Does not exist

Fig. 3. A subset of the comparison between the BibTeX 2.0 ontology and two other bibliographic ontologies BIBO [26] and SWRC [25].

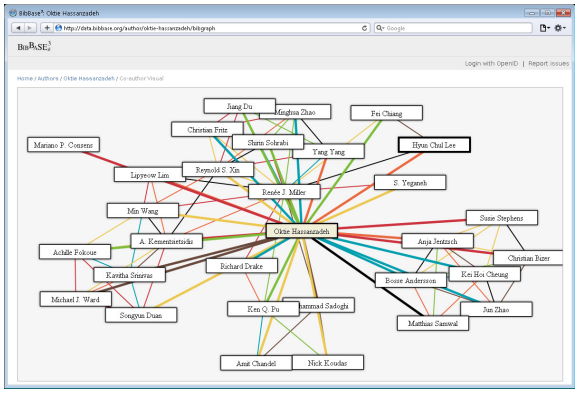

Fig. 4. Co-author graph on BibBase data browse interface

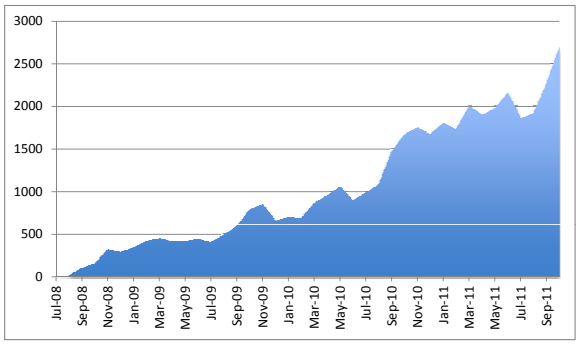

Fig. 5. BibBase.org usage statistics since July 2008

- A feature that allows fetching publications directly from Mendeley [22] users or groups.
- Statistics regarding users, page views, and paper downloads, including a list of most popular papers.
- A co-author visualization interface that shows a graph for each entity of type `author` on the data browse interface. Figure 4 shows an example of this co-author graph.

We are currently working on a number of new features as described in Section 9.

8. Online Demo

The HTML and RSS interface of BibBase, available at `http://bibbase.org`, have been running since 2008, and are in active use by several groups and individuals. A list of current users of BibBase along with the URLs of their pages and visitor statistics can be found at `http://purl.org/bibbase/stats`. Figure 5 shows the number of unique visitors of BibBase-powered pages since July 2008, which shows a fairly robust and continuous growth in terms of traffic and users. In addition, the latest statistics on the number of entities in our data server can be found on: `http://data.bibbase.org/stats/`. As of November 2011, the database contains 7,538 Publications, 8,937 Authors, and 261 'Provenance' entities (meaning 261 uploaded BibTeX files). Documentation and current status of the experimental features of the data interface are available at `http://wiki.bibbase.org`.

9. Future Work

We are actively working on additional features ranging from minor improvements to new major applications of state-of-the-art data and knowledge management techniques over the data. We are working on extensions to the data browse interface, 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.

Some of the more advanced features that we intend to add to BibBase in the future include:

- The ability for visitors of a BibBase-rendered page to "recommend" papers. This again is done in a declarative fashion, such that the fact of the recommendation can serve various purposes in the future. For instance, a student might be interested in all the papers her advisor recommends that have a certain keyword, or, to get a first overview of a new topic, someone may want to see the most recommended papers in that field.

  Recommendations could be extended with in-depth reviews of the paper, providing an extension of the usually rather opaque peer-reviewing system into the Semantic Web. For this to be efficient, it is important to uniquely identify publications, rather than establishing various different sites on the Web where the same paper may be discussed. It also allows for an older publication to be reviewed over and over again in the light of new results or new work that builds on top of this.
- Another possible extension is to perform disambiguation of references in a paper and assert new links from them. This way it would be possible to explore the trace of references which in turn can be used in various ways, including, e.g., for analysing the impact past work has had on a field, or how cross-fertilization happens between areas.

Linked publication data opens up a lot of interesting opportunities to build exciting applications. Such applications can only work well if it is easy for authors to provide that data, which is the main goals of BibBase.

10. 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 of data sets with a data source that unlike existing bibliographic 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`.

11. Acknowledgements

This work is partially supported by NSERC BIN. We thank Yang (Jack) Yang for his work on the initial implementation of the data interface, Nataliya Prokoshyna for implementing the user feedback interface, and Zheng Xiong for his work on the co-author visualization interface. We also thank Jan Brase, Kai Eckert, and Antoine Isaac for their thoughtful reviews.

References

[1] T. Berners-Lee. Linked Data - Design Issues. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006. [Online; accessed 8-6-2011].

[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 8-6-2011].

[4] A. Chandel, O. Hassanzadeh, N. Koudas, M. Sadoghi, and D. Srivastava. Benchmarking Declarative Approximate Selection Predicates. In C. Y. Chan, B. C. Ooi, and A. Zhou, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 353–364. ACM, 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 D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin, editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1027–1036. ACM, 2009.

[7] O. Hassanzadeh, R. Xin, R. J. Miller, A. Kementsietsidis, L. Lim, and M. Wang. Linkage Query Writer. *Proceedings of the VLDB Endowment (PVLDB)*, 2(2):1590–1593, 2009.

[8] I. Herman. BibTeX in RDF. `http://ivan-herman.name/2007/01/13/bibtex-in-rdf/`, 2007. [Online; accessed 8-6-2011].

[9] N. Knouf. BibTeX Ontology. `http://purl.org/net/nknouf/ns/bibtex`. [Online; accessed 8-6-2011].

[10] D. B. Krafft, N. A. Cappadona, B. Caruso, J. Corson-Rikert, M. Devare, B. J. Lowe, and VIVO Collaboration. VIVO: Enabling National Networking of Scientists. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, 2010.

[11] F. Naumann and M. Herschel. *An Introduction to Duplicate Detection*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2010.

[12] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and Maintaining Links on the Web of Data. In A. Bernstein, D. R. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, and K. Thirunarayan, editors, *Proceedings of the 8th International Semantic Web Conference (ISWC)*, volume 5823 of *Lecture Notes in Computer Science*, pages 650–665. Springer, 2009.

[13] List of BibTeX to RDF Conversion Tools. http://purl.org/bibbase/other-bibtex2rdf-tools. [Online; accessed 10-11-2011].

[14] Celery: Distributed Task Queue. http://www.celeryproject.org/. [Online; accessed 31-10-2011].

[15] BibBase. http://www.bibbase.org/. [Online; accessed 10-11-2011].

[16] BibSonomy. http://www.bibsonomy.org/. [Online; accessed 31-10-2011].

[17] CiteSeerX. http://citeseer.ist.psu.edu/. [Online; accessed 31-10-2011].

[18] citeulike. http://www.citeulike.org/. [Online; accessed 31-10-2011].

[19] D2R Server publishing the DBLP Bibliography Database. http://www4.wiwiss.fu-berlin.de/dblp/. [Online; accessed 31-10-2011].

[20] EPrints. http://www.eprints.org/. [Online; accessed 31-10-2011].

[21] PubZone. http://www.pubzone.com/. [Online; accessed 31-10-2011].

[22] Mendeley. http://www.mendeley.com/. [Online; accessed 31-10-2011].

[23] RefWorks. http://www.refworks.com/. [Online; accessed 31-10-2011].

[24] RKB Explorer. http://www.rkbexplorer.com/. [Online; accessed 31-10-2011].

[25] SWRC Ontology. http://ontoware.org/swrc/. [Online; accessed 31-10-2011].

[26] The Bibliographic Ontology. http://bibliontology.com/. [Online; accessed 14-10-2011].

[27] Web Reference Database (refbase). http://www.refbase.net/. [Online; accessed 31-10-2011].