Available online at www.sciencedirect.com

SCIENCE DIRECT®

Web Semantics: Science, Services and Agents
on the World Wide Web 2 (2004) 99–103

JOURNAL OF Web Semantics

www.elsevier.com/locate/websem

Bibster—a semantics-based bibliographic Peer-to-Peer system

Peter Haase[a,*], Björn Schnizler[a], Jeen Broekstra[b], Marc Ehrig[a], Frank van Harmelen[b], Maarten Menken[b], Peter Mika[b], Michal Plechawski[c], Pawel Pyszlak[c], Ronny Siebes[b], Steffen Staab[a], Christoph Tempich[a]

[a] *Institute AIFB, University of Karlsruhe, D-76128 Karlsruhe, Germany*
[b] *Vrije Universiteit Amsterdam, The Netherlands*
[c] *Empolis, Warsaw, Poland*

Received 7 July 2004; received in revised form 7 September 2004; accepted 29 September 2004

Abstract

This paper describes Bibster, a Peer-to-Peer system for exchanging bibliographic metadata among researchers. We show how Bibster exploits ontologies in data-representation, query formulation, query routing, and query result presentation. The Bibster system is freely available and is used by researchers across multiple organizations.

Keywords: Peer-to-peer; Semantic web ontologies; Knowledge management

1. Introduction

In this paper, we describe the Bibster system[1], an application of the use of semantics in Peer-to-Peer systems. Bibster is aimed at researchers that share bibliographic metadata. Currently, many researchers in computer science keep lists of bibliographic metadata in BibTeX format that they must laboriously maintain manually, for which they do not have an easy overview, and that has greatly varying quality. At the same time, many researchers are willing to share these resources, provided they do not have to invest work in doing so. The following characteristics make this scenario an interesting use case for a semantics-based Peer-to-Peer system. First, a centralized solution does not exist and cannot exist, because of the multitude of informal workshops that researchers refer to, but that do not show up in centralized resources, such as DBLP[2]. Sec-

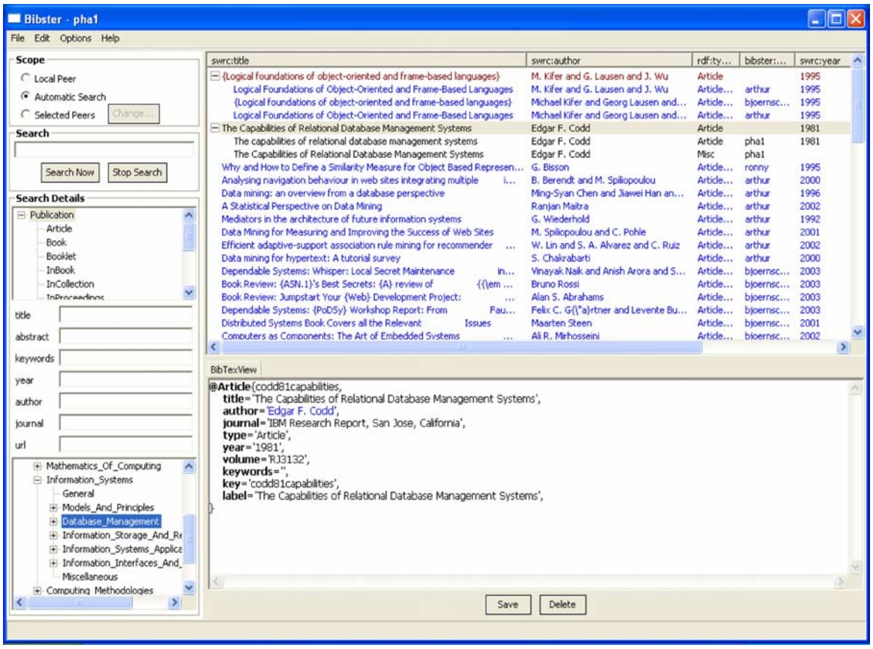
[*] Corresponding author. Tel.: +49 721 6083705; fax: +49 721 693717.

E-mail addresses: haase@aifb.uni-karlsruhe.de (P. Haase), schnizler@iw.uka.de (B. Schnizler), jbroeks@cs.vu.nl (J. Broekstra), ehrig@aifb.uni-karlsruhe.de (M. Ehrig), frankh@cs.vu.nl (F. van Harmelen), mrmenken@cs.vu.nl (M. Menken), pmika@cs.vu.nl (P. Mika), mpl@empolis.pl (M. Plechawski), pap@empolis.pl (P. Pyszlak), ronny@cs.vu.nl (R. Siebes), staab@aifb.uni-karlsruhe.de (S. Staab), tempich@aifb.uni-karlsruhe.de (C. Tempich).

[1] Bibster is freely available for download under http://bibster.semanticweb.org/.

[2] http://www.informatik.uni-trier.de/~ley/db/.

1570-8268/$ – see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.websem.2004.09.006



swrc:title	swrc:author	rdf:ty...	bibster:...	swrc:year
⊟ {Logical foundations of object-oriented and frame-based languages}	M. Kifer and G. Lausen and J. Wu	Article		1995
Logical Foundations of Object-Oriented and Frame-Based Languages	M. Kifer and G. Lausen and J. Wu	Article...	arthur	1995
{Logical foundations of object-oriented and frame-based languages}	Michael Kifer and Georg Lausen and...	Article...	bjoernsc...	1995
Logical Foundations of Object-Oriented and Frame-Based Languages	Michael Kifer and Georg Lausen and...	Article...	arthur	1995
⊟ The Capabilities of Relational Database Management Systems	Edgar F. Codd	Article		1981
The capabilities of relational database management systems	Edgar F. Codd	Article	pha1	1981
The Capabilities of Relational Database Management Systems	Edgar F. Codd	Misc	pha1	
Why and How to Define a Similarity Measure for Object Based Represen...	G. Bisson	Article...	ronny	1995
Analysing navigation behaviour in web sites integrating multiple i...	B. Berendt and M. Spiliopoulou	Article...	arthur	2000
Data mining: an overview from a database perspective	Ming-Syan Chen and Jiawei Han an...	Article...	arthur	1996
A Statistical Perspective on Data Mining	Ranjan Maitra	Article...	arthur	2002
Mediators in the architecture of future information systems	G. Wiederhold	Article...	arthur	1992
Data Mining for Measuring and Improving the Success of Web Sites	M. Spiliopoulou and C. Pohle	Article...	arthur	2001
Efficient adaptive-support association rule mining for recommender ...	W. Lin and S. A. Alvarez and C. Ruiz	Article...	arthur	2002
Data mining for hypertext: A tutorial survey	S. Chakrabarti	Article...	arthur	2000
Dependable Systems: Whisper: Local Secret Maintenance in...	Vinayak Naik and Anish Arora and S...	Article...	bjoernsc...	2003
Book Review: {ASN.1}'s Best Secrets: {A} review of {{\em ...	Bruno Rossi	Article...	bjoernsc...	2003
Book Review: Jumpstart Your {Web} Development Project: ...	Alan S. Abrahams	Article...	bjoernsc...	2003
Dependable Systems: {PoDSy} Workshop Report: From Fau...	Felix C. G{\"a}rtner and Levente Bu...	Article...	bjoernsc...	2003
Distributed Systems Book Covers all the Relevant Issues	Maarten Steen	Article...	bjoernsc...	2001
Computers as Components: The Art of Embedded Systems ...	Ali R. Mirhosseini	Article...	bjoernsc...	2002

```
@Article{codd81capabilities,
  title='The Capabilities of Relational Database Management Systems',
  author='Edgar F. Codd',
  journal='IBM Research Report, San Jose, California',
  type='Article',
  year='1981',
  volume='RJ3132',
  keywords='',
  key='codd81capabilities',
  label='The Capabilities of Relational Database Management Systems',
}
```

Fig. 1. Screenshot of the Bibster system.

ond, the use of Semantic Web technology is crucial in this setting. Although a small common-core ontology of bibliographic information exists (title, author/editor, etc.), but much of this information is very volatile and users define arbitrary add-ons, like including URLs.

Ontologies are crucial throughout the usage of Bibster, viz., for importing data, formulating and routing queries, and processing answers. Fig. 1 shows how these steps are realized in the user interface of Bibster. The *Scope* widget allows for defining the targeted peers, the *Search* and *Search Details* widgets allow for keyword and semantic search; *Results Table* and *BibTeXView* widgets allow for browsing and re-using query results. In the following, we will describe the use of semantic methods in each of these steps. A detailed presentation of the architecture and methods of the Bibster system can be found in [1,2].

2. Bibster architecture and modules

The Bibster system has been implemented as an instance of the SWAP System architecture as introduced in [1]. Fig. 2 shows a high-level design of the architecture of a single node in the Peer-to-Peer system. We will now briefly present the individual components as instantiated for the Bibster system.

The *Communication Adapter* is responsible for the network communication between peers. It serves as a transport layer for other parts of the system, for sending and forwarding queries. It hides and encapsulates all low-level communication details from the rest of the system. In the specific implementation of the Bibster system we use JXTA as the communication platform.

The *Knowledge Sources* in the Bibster system are sources of bibliographic metadata, such as BibTeX files stored locally in the file system of the user. The *Knowledge Source Integrator* is responsible for the extraction and integration of internal and external knowledge sources into the Local Node Repository. The *Local Node Repository* provides the following functionality: (1) It serves as storage for and provides views on the available knowledge, (2) it supports query formulation and processing, (3) it serves as a registry of peer descriptions as the basis for peer selection.

The task of the *Informer* is to proactively advertise the available knowledge of a peer in the Peer-to-Peer network and to discover peers with knowledge that may

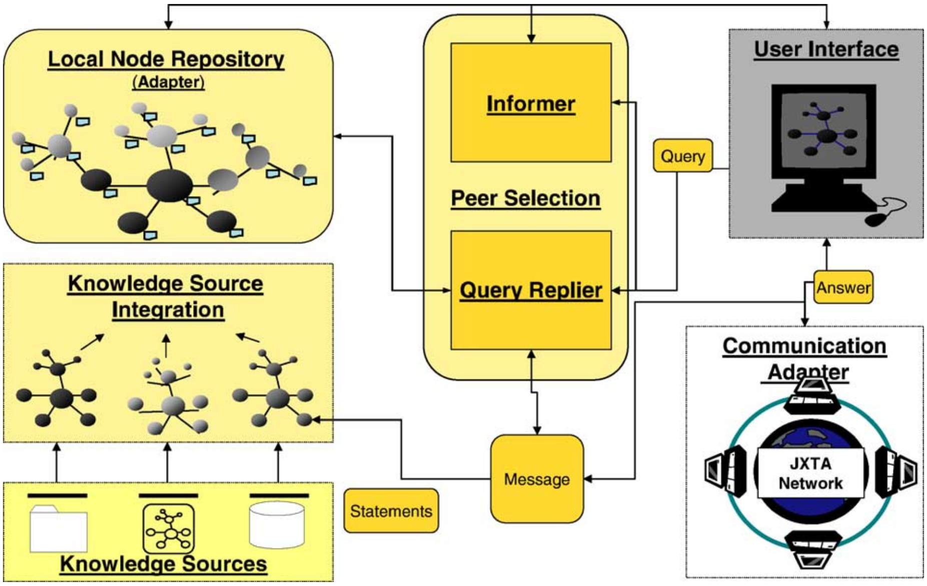


Fig. 2. SWAP system architecture.

be relevant for answering the user's queries. This is realized by sending advertisements about the expertise of a peer. In the Bibster system, these expertise descriptions contain a set of topics that the peer is an expert in. The *Query Replier* is the coordinating component controling the process of distributing queries. It receives queries from the user interface or from other peers. It initiates the processing of the queries locally and may decide to forward them to other peers, according to the peer selection model. The *User Interface* allows the user to import, create and edit and export bibliographic metadata as well as to easily formulate queries, as shown in Fig. 1.

3. Semantic methods in Bibster

3.1. Import and semantic representation of bibliographic metadata

Many researchers have accumulated extensive collections of BibTeX files for their bibliographic references. The Bibster system allows users to import their own bibliographic metadata into a local RDF repository. Bibliographic entries made available to Bibster by a user are automatically aligned to two common ontologies: The first ontology—the Semantic Web Research Community Ontology (SWRC[3])—describes different generic aspects of bibliographic metadata, such as publications, persons, organizations, etc. and the relations between them. The second ontology—the ACM Topic Hierarchy[4]—describes specific categories of literature for the Computer Science domain. The bibliographic entries are classified automatically during import, but can be reclassified manually in the user interface of Bibster via drag and drop.

3.2. Semantic querying using SeRQL

Queries are formulated in terms of the two ontologies: queries may concern fields like author, publication type, etc. (using terms from the SWRC ontology) or queries may concern specific Computer Science terms (using the ACM Topic Hierarchy). We use SeRQL[5] to query the RDF repository.

When querying the bibliographic metadata, we make use of the following characteristics of SeRQL: (1) It is a functional and compositional language, meaning

[3] http://www.semanticweb.org/ontologies/swrc-onto-2001-12-11.daml.

[4] http://www.acm.org/class/1998/.

[5] Sesame RDF Query Language, http://www.openrdf.org/doc/SeRQLmanual.html.

that each query returns an RDF graph, which may be shipped between peers, integrated into the local repository, or queried again. (2) It allows to formulate path expressions for navigating the RDF graph, i.e. the combination of SWRC and ACM topic hierarchy. (3) It is aware of the schema, allowing us to query against the concept hierarchy of the used ontologies. (4) It is able to deal with *optional* values. This is important as BibTeX entries may be incomplete, e.g. a publisher field may be given or not.

3.3. Expertise-based peer selection

In the Bibster system, the user can specify the scope of a query: He can either query the local knowledge, direct the query to a selected set of peers, or can query the entire peer network. For the latter option, the scalability of the Peer-to-Peer network is essentially determined by the way how the queries are propagated in the network. Peer-to-Peer networks that broadcast all queries to all peers do not scale—intelligent query routing and network topologies are required to be able to route queries to a relevant subset of peers that are able to answer the queries. In the Bibster system, we apply the model of expertise-based peer selection as proposed in [3]. Based on this model, peers advertise semantic descriptions of their expertise specified in terms of the ACM topic hierarchy. The knowledge about the expertise of other peers forms a semantic topology, in which peers with a similar expertise are clustered. To determine an appropriate set of peers to forward a query to, a matching function determines how closely the semantic content of a query that references an ACM topic matches the expertise of a peer.

3.4. Semantic duplicate detection

Due to the distributed nature and potentially large size of the Peer-to-Peer network, the returned result set for a query might be large and contain duplicate answers. Furthermore, because of the heterogeneous and possibly even contradicting representation, such duplicates are often not exactly identical copies. Here, ontologies help to measure the semantic similarity between the different answers and to remove apparent duplicates as identified by the similarity function. Bibster uses specific similarity functions that operate on various properties of the publications and on different levels of similarity. For example, to determine the similarity of two publications we calculate the *syntactic similarity* of the titles using the Levenshtein similarity measure. At the *graph level*, we determine how resources are interlinked, e.g. to compare publications based on their co-authorship. On the *ontology level*, we apply a hierarchical similarity function to determine the taxonomic similarity of the classifications of publications according to the ACM topic hierarchy. Finally, we apply background knowledge about the bibliographic domain that considers, for example, that publication types are often provided as Misc, if their type is unknown.

From the variety of individual similarity functions, an overall value is obtained with an aggregated similarity function by means of a weighted average. As duplicates we consider those pairs of resources whose similarity is larger than a defined threshold. Instead of presenting all individual resources of the query result, the duplicates are then visualized as one merged resource that represents the combined knowledge about the publication.

4. Evaluation and conclusion

The Bibster system is currently being evaluated by means of a public field experiment. The user actions and system events are continuously logged and analyzed to evaluate the user behavior and system performance. We have analyzed the results for a period of 1 month (June 2004) and we have obtained the following interesting results: 53 peers used the Bibster system and shared more than 33,000 bibliographic entries. Eight peers shared more than 1000 items each. The users performed a total of 700 queries. The SWRC ontology was used for about an half of all queries. Most searches concerned queries for authors (144). In 101 queries, the users asked for topics of the ACM topic hierarchy. With respect to query routing, with the expertise-based peer selection we were able to reduce the number of messages by about 50%, while retaining the same recall of documents compared with a naive broadcasting approach.

With our case study, we have shown that for Bibster and similar applications the usage of Semantic Web technologies and ontologies provide an *added value*—in fact, it is almost a strict requirement given its semi-structured, volatile data structures. Semantic structures serve important user concerns like high quality duplicate detection or comprehensive searching capabilities.

Acknowledgments

Research reported in this paper has been partially financed by the EU in the IST Project SWAP (IST-2001-34103).

References

[1] M. Ehrig, P. Haase, F. van Harmelen, R. Siebes, S. Staab, H. Stuckenschmidt, R. Studer, C. Tempich, The swap data and metadata model for semantics-based peer-to-peer systems, in: M. Schillo, M. Klusch, J.P. Miiller, H. Tianfield (Eds.), Proceedings of MATES-2003. First German Conference on Multiagent Technologies, Volume 2831 of LNAI, Erfurt, Germany, Springer, 2003, pp. 144–155.

[2] P. Haase, J. Broekstra, M. Ehrig, M. Menken, P. Mika, M. Plechawski, P. Pyszlak, B. Schnizler, R. Siebes, S. Staab, C. Tempich, Bibster—a semantics-based bib liographic peer-to-peer system, in: Proceedings of the Third International Semantic Web Conference, Hiroshima, Japan, 2004.

[3] P. Haase, R. Siebes, F. van Harmelen, Peer selection in peer-to-peer networks with semantic topologies., in: International Conference on Semantics of a Networked World: Semantics for Grid Databases, Paris, 2004.