

# Bibster - A Semantics-Based Bibliographic Peer-to-Peer System

Jeen Broekstra <sup>c</sup>      Marc Ehrig <sup>b</sup>      Peter Haase <sup>b</sup>  
Frank van Harmelen <sup>a</sup>      Maarten Menken <sup>a</sup>      Peter Mika <sup>a</sup>  
Bjorn Schnizler <sup>b</sup>      Ronny Siebes <sup>a</sup>

<sup>a</sup> Vrije Universiteit Amsterdam, The Netherlands

<sup>b</sup> Institute AIFB, University of Karlsruhe, Germany

<sup>c</sup> Aduna, Amersfoort, The Netherlands

The full version of this paper appeared in the proceedings of the Second Workshop on Semantics in peer-to-peer and Grid Computing (SemPGRID'04) held at the WWW'04 conference in New York [1]. In that paper, we describe the design and implementation of Bibster, a Peer-to-Peer system for exchanging bibliographic data among Computer Science researchers. Bibster is fully implemented on top of the JXTA platform<sup>1</sup>, and is now being rolled out for field testing. This submission is also a kind request for participating in the system test<sup>2</sup>.

Currently, many researchers in Computer Science keep lists of bibliographic metadata in BibTeX format, that they must 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 domain an interesting use case for a semantics-based peer-to-peer system:

- 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<sup>3</sup>.
- 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), much of this information is very volatile and does not allow user specific add-ons, like private comments.

Next, we describe how the use of ontologies are used in all the steps of Bibster: importing data, formulating queries, routing queries, and processing answers.

Firstly, the system enables users to import their own bibliographic metadata into a local repository. This bibliographic metadata is made available under two common ontologies: the first ontology (SWRC<sup>4</sup>) describes different generic aspects

---

<sup>1</sup><http://www.jxta.org>

<sup>2</sup><http://bibster.semanticweb.org>

<sup>3</sup><http://www.informatik.uni-trier.de/~ley/db/>

<sup>4</sup><http://ontobroker.semanticweb.org/ontos/swrc.html>

of bibliographic metadata, the second ontology (ACM Topic Hierarchy<sup>5</sup>) describes specific categories of literature for the Computer Science domain. Bibliographic entries that are made available to Bibster by a user are automatically classified under these two ontologies.

Secondly, users can send queries to other peers looking for bibliographic metadata. These queries are formulated in terms of the two ontologies and are translated into the RDF query language SeRQL to be answered by the different peers in the network.

Thirdly, these queries need to be routed across the peer-network, and again the ontologies play a crucial role. Queries are routed through the network depending on the expertise models of the peers. Such an expertise model describes which concepts from the ACM ontology a peer can answer queries on. A matching function determines how closely the semantic content of a query matches the expertise model of each peer. Routing is then done on the basis of this semantic ranking.

Finally, answers are returned for a query. Due to the distributed nature and potentially large size of the peer-to-peer network, this answer set might be very large, and contain many duplicate answers. Because of the semi-structured nature of bibliographic metadata, such duplicates are often not exactly identical copies. Again in this step, we exploit ontologies, this time to measure the semantic similarity between the different answers, and to remove apparent duplicates as identified by the similarity function.

In order to measure the effectiveness of our semantics-based approach, we are currently doing an extensive evaluation study, measuring both user-related aspects (such as user satisfaction with interface or performance), and system-related aspects (such as average number of hops for a query or number of duplicates detected).

The Bibster system is one of the first ontology-based peer-to-peer systems ready for fielded deployment, which uses ontologies in all its steps. Particularly interesting will be to see how its performance will compare to related systems such as P-Grid and Edutella.

## References

- [1] Jeen Broekstra, Marc Ehrig, Peter Haase, Frank van Harmelen, Maarten Menken, Peter Mika, Björn Schnizler, and Ronny Siebes. Bibster - a semantics-based bibliographic peer-to-peer system. In *Proceedings of SemPGRID '04, 2nd Workshop on Semantics in Peer-to-Peer and Grid Computing*, pages 3–22, New York, USA, May 2004.

---

<sup>5</sup><http://www.acm.org/class/1998/>