Project no. 34721

TAGora

Semiotic Dynamics in Online Social Communities

http://www.tagora-project.eu

Sixth Framework Programme (FP6)

Future and Emerging Technologies of the Information Society Technologies (IST-FET Priority)

D2.1
First version of social tagging system for bibliographic data
and
First version of folksonomy peer-to-peer system for multimedia data

---

Period covered: from 01/06/2006 to 31/05/2007
Start date of project: June 1st, 2006
Due date of deliverable: May 31st, 2007
Distribution: Public

Date of preparation: 31/05/2007
Duration: 36 months
Actual submission date: May 31st, 2007
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: “Sapienza” Università di Roma
Lead contractor for this deliverable: University of Koblenz-Landau
and University of Kassel

Executive Summary

This deliverable summarizes the development of two systems:

BibSonomy – a social resource sharing system for bookmarks and publications.

SEA – a folksonomy-based peer-to-peer system for sharing of multimedia data.

The aim of these systems in Tagora is twofold: they shall support the collection of real-world user data for experiments, and provide a platform for experimenting how different kinds of user interaction influence the evolution of the resulting data over time.

The first system, BibSonomy, allows users collaborative organizing and sharing of bookmark collections and publication lists. A basic version of BibSonomy has been online before the start of the project. Within Tagora, we have extended its functionality to attract a significant number of users, and have provided means for a systematic generation of data for experiments. These datasets are updated on a half yearly basis and are publicly available for research purposes.

The second system, SEA, is a distributed collaborative tagging application focusing on multimedia data. In contrast to common online tagging applications it allows to tag any file on a local computer and exchange the data in a peer-to-peer network. The first prototype provides general tagging and metadata exchange functionalities. A file-browser-like interface is used to choose files and apply free text tags to them. Whole directories can be selected to tag all the contained files at once. The stored tagging data is easily browsable via the assigned tags and adding/removing tags is directly done by changing the tag list of one resource. General statistics like global tag assignments are additionally stored in distributed indices in the peer-to-peer network. The tagging data of the final system can then be used for further folksonomy analysis.

Contents

Chapter 1

BibSonomy - Social tagging for bibliographic data

The central component of this part of the deliverable is the social bookmarking system BibSonomy. Its installation is publicly accessible at `http://www.bibsonomy.org` *. This chapter complements the deliverable by briefly describing the main steps that we followed towards its implementation within the Tagora project.*

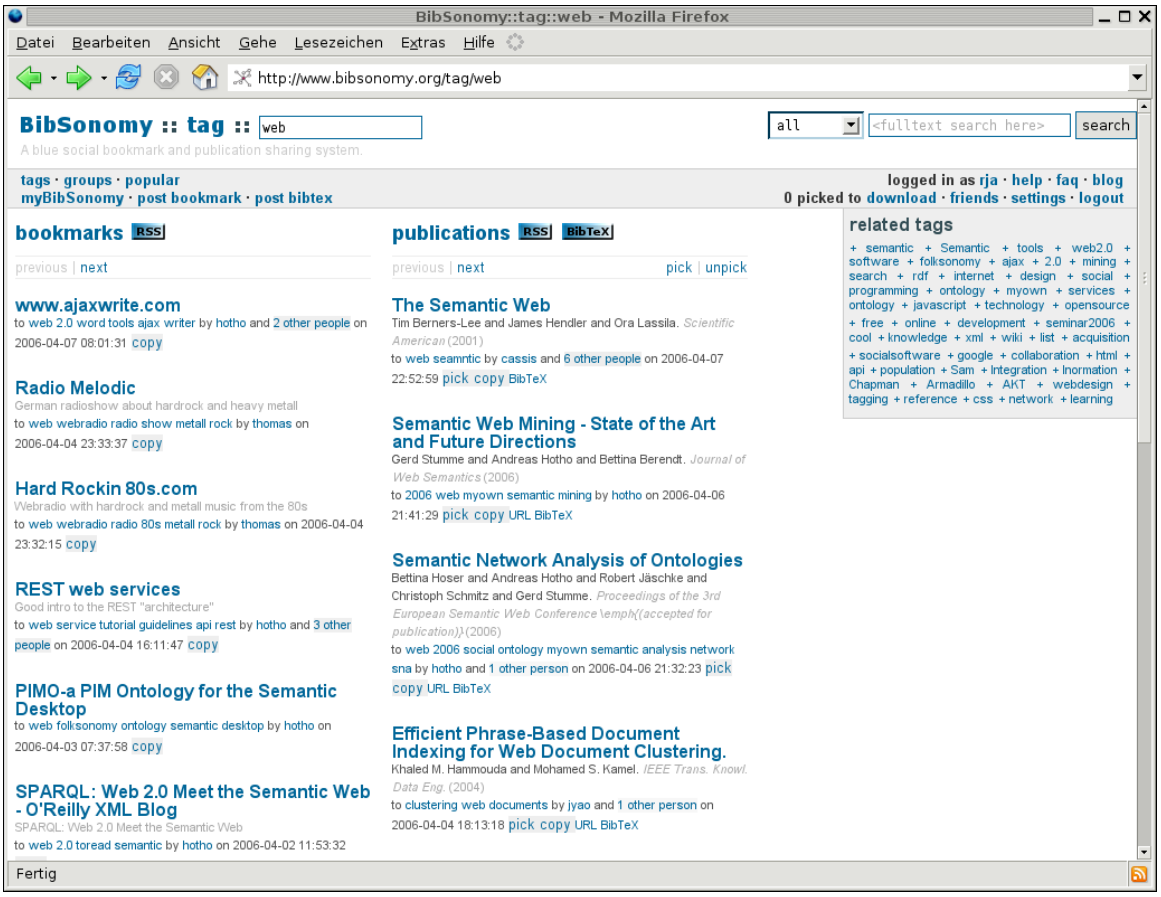

Figure 1.1: BibSonomy displays bookmarks and BibTEX based bibliographic references simultaneously.

1.1 Motivation

Social bookmark tools are rapidly emerging on the Web. Web services such as Connotea or CiteULike for tagging bibliographic references offer powerful functionality regarding knowledge sharing and managing. The motivation for implementing our own tagging system is threefold. Although crawling activities offer the possibility to get access to data of existing web services as mentioned above, an unfettered access to the complete data is not guaranteed. The implementation of an own bibliographic sharing system offers us the full access to data and thus the possibility to observe developments of social structures from the beginning. In addition to this, our system can be used as experimental environment to manipulate variables like tag suggestions, e.g. to prove the direct impact of evolutional changes in applications for build up cause and effect relations.

Our bibliographic sharing system BibSonomy enables the user to manage and collect publications in form of BibTEX entries and bookmarks and is publicly available on the web on `http://www.bibsonomy.org`. We recommend that you have a look at the online system before continuing to read.

1.2 System overview

The prototype that existed at the beginning of the project (Hotho et al., 2006) offered basic functionality for social tagging, but did not provide the power of systems like CiteULike or Connotea. For example the initial version of BibSonomy did not offer a classified view of the content like CiteULike it does. A further option, which is offered by Connotea (but not by the initial version of BibSonomy) regards the suggestions of related tags and users concerning a chosen publication entry.

The data model of the publication part is based on BibTEX, a popular literature management system for LATEX.

Figure 1.1 shows a typical list of bookmark and publication posts containing the tag *web*. The page is divided into four parts: the header (showing navigation links and search boxes), two lists of posts – one for bookmarks and one for publications – sorted by date in descending order, and a list of tags related to the posts. This scheme holds for all pages showing posts and allows for navigation in all dimensions of the folksonomy.

REST web services
Good intro to the REST "architecture"
to web service tutorial guidelines api rest by hotho and 3 other people on 2006-04-04 16:11:47 copy

Figure 1.2: A single bookmark post.

Semantic Network Analysis of Ontologies
Bettina Hoser and Andreas Hotho and Robert Jäschke and Christoph Schmitz and Gerd Stumme. *Proceedings of the 3rd European Semantic Web Conference \emph{(accepted for publication)}* (2006)
to web 2006 social ontology myown semantic analysis network sna by hotho and 1 other person on 2006-04-06 21:32:23 pick copy URL BibTeX

Figure 1.3: A single publication post.

Figures 1.2 and 1.3 present a detailed view of bookmark and publication posts from Figure 1.1. The first line of the bookmark post shows in bold the title of the bookmark which has hyperlinks to its URL. The second line is an optional description assigned by the user. The last two lines first show the tags the user has assigned to this post, the user name, and how many users tagged that resource. These parts hyperlink to the corresponding tag pages of the user, the user's overview page, and a page showing all four posts related to this resource. The last part shows the posting date and time followed by actions the user can perform on this post. The structure of a publication post displayed in BibSonomy is very similar, showing the bibliographic details instead of the description. The title links to a page providing detailed information on that post. The actions for BibTEX items include picking the entry for later download, copying it, accessing the URL of the

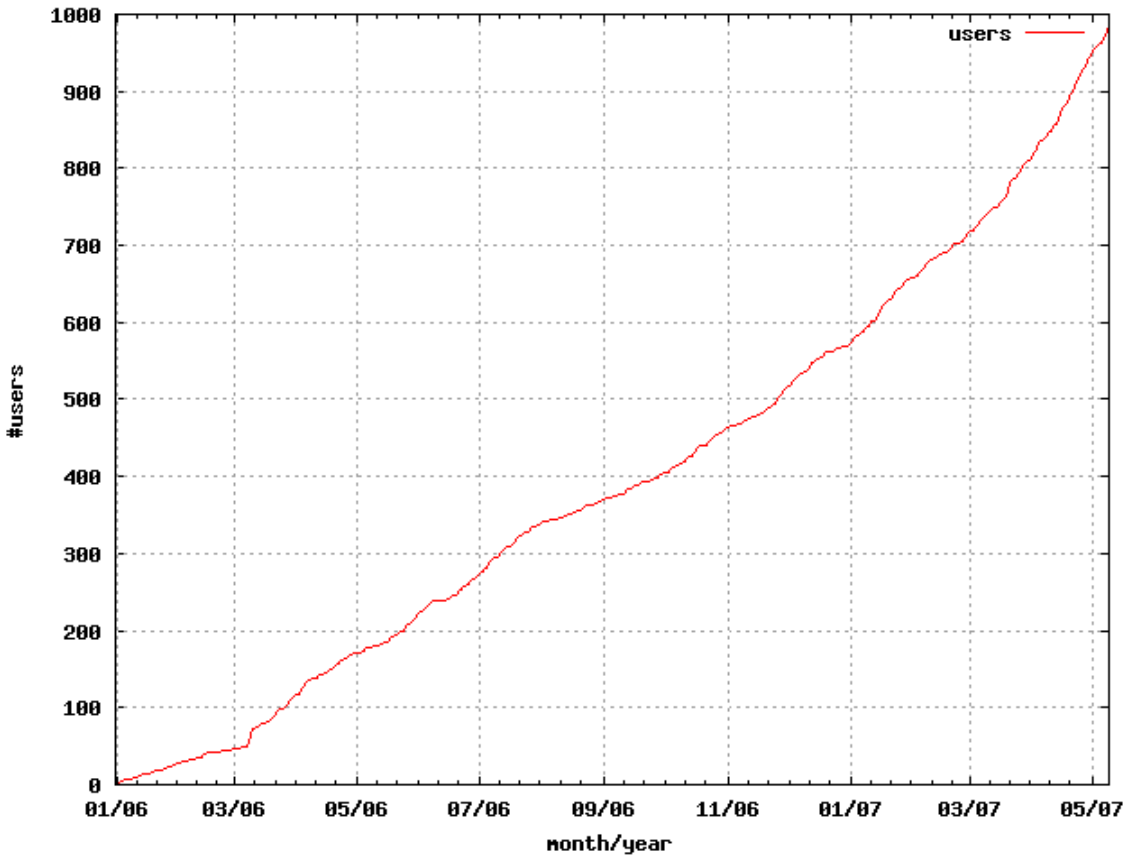

Figure 1.4: Growth of BibSonomy over the last 18 months.

entry or viewing the BibTEX source code. To enhance intuitive navigation, the user is supported by a tag cloud.

1.2.1 Extended functionality

BibSonomy offers a platform to realize sharing interests of a broad mass of users and guarantees scalability. A publication or a set of publications can be posted in different forms to our system. The most comfortable option is posting a BibTEX entry via a scraper (eg. for the ACM Digital Library and CiteSeer). This option offers a simply utilized feature, because the user only needs to mark a BibTEX entry on a homepage and press the post button. In addition, the user can post BibTEX snippets via copy & paste. All tag assignments to a resource can be modified later by the user. Further options offer renaming of tags.

A BibTEX entry can be marked with three different types of privacy. The first option contains marking a BibTEX entry with the label private and let a BibTEX entry viewable for its own. The second possibility contains the feature of sharing publications with friends. The – default – option of public posting offers the possibility to share content with all users of the BibSonomy community.

BibSonomy offers different export formats for data and metadata: Endnote, BibTEX, HTML, RTF, RDF, XML, RSS, and many others, see `http://www.bibsonomy.org/export/` for the full list. BibSonomy also allows to import data from del.icio.us.

Updates, questions and suggestions from users can be posted to `http://bibsonomy.blogspot.com/` for discussion. To support the user in frequently asked questions, an interactive tutorial is available on `http://www.bibsonomy.org/help` and provides the user in his first steps in social software. The tutorial serves as introduction to our system and gives an overview of the power

of BibSonomy's functionality. Beside technical support, the user gets an insight on BibSonomy add-ons.

Furthermore, the following extended functionalities have been added to BibSonomy: keyboard shortcut for BibSonomy posting in Firefox, OpenURL support, tag editor, tag hierarchy ("relations"), gnome desktop integration, spam filter, improvement of basket and group functionalities, tutorials, faq, extended help pages, migration to a new server to increase hardware redundancy, password forgotten functionality, improved relation management, information extraction for publications in unstructured text, full-text search. For more details see `http://bibsonomy.blogspot.com/`. More information about the functionality of BibSonomy in general can be found at `http://www.bibsonomy.org/help`.

1.2.2 Data Collection

To call attention to BibSonomy, we initially advertised it via mailing lists and presented it to the scientific community on conferences and workshops. The growth of BibSonomy in terms of users that posted at least once (excluding spammers) over the last 18 months is shown in Figure 1.4. A snapshot of the folksonomy is updated every half year to guarantee that project members work on a unique dataset. The datasets are online available and accessible as mentioned in D1.1.

To allow for tracking the system evolution over time at a fine-grained level, each BibTEX entry is labeled by a time stamp. The popularity of a BibTEX entry or a tag can be measured over time and allows us to track the evolution of the folksonomy. Additional information about the individual user behavior is logged in an Apache access log, including calls of the copying function. I. e., it is logged which user copies a BibTEX entry from another user.

A further user activity is recorded by logging the website from which a user post a BibTEX entry. The count of users, who also share the same publication, is appended and viewable on the system homepage. Attributes like author, pages, volume and time stamp of a BibTEX entries are logged in a database. This also includes detailed data of a user. E.g. the name, date of first login and email-address. We also log group memberships and the evolution of the friendship relation over time.

Chapter 2

Folksonomy peer-to-peer system for multimedia data

The central component of this part of the deliverable is a folksonomy peer-to-peer system for multimedia data. This chapter complements the deliverable by briefly describing the main steps that we followed towards its implementation within the Tagora project. A beta version of the software can be downloaded at `http://isweb.uni-koblenz.de/Projects/TAGORA`.

2.1 Motivation

The development of the peer-to-peer tagging system has different reasons. First, we want to promote tagging as an easy way for organizing and sharing personal data. In contrast to common centralized tagging systems a client application allows for better tagging flexibility because a user can tag virtually any local information object and does not need to sign up for multiple online accounts to share, for example, photos or bookmarks. Furthermore, only the tagging metadata needs to be shared while the actual information objects generally remain on the local machine (or can be downloaded on demand).

Another aspect is the gathering of additional tagging data for the analysis in workpackage 3. We are interested to see whether there are any significant differences between the tagging behavior in a centralized and in a decentralized system.

2.2 System overview

The prototype of the distributed tagging system basically allows to tag any local information object. As described in the technical annex the initial idea was to such a system on top of Bibster, a peer-to-peer sharing system for bibliographic data. Due to the limitations of Bibster which does not allow arbitrary tags and does not support statistics we developed a completely new and much more flexible architecture which allows the tagging and sharing of virtually any information object. Without that limitation to bibliographic data the focus was extended to multimedia data. This avoids on the one hand a direct competition with BibSonomy and provides on the other hand a much greater data basis for collecting tagging data.

At the moment all tagging data and tagged resources are publicly visible. This means that all users in the network can see all data of all other uses. This may be refined in the future to allow private data as well. But for the time being the main benefit of this system is that all data is publicly shared.

The prototype distributed tagging system has been developed in Java to allow its use on virtually any platform. It is based on the Semantic Exchange Architecture (SEA(Franz et al., 2006)) - initial work that defines the core components for a scalable peer-to-peer architecture for metadata ex-

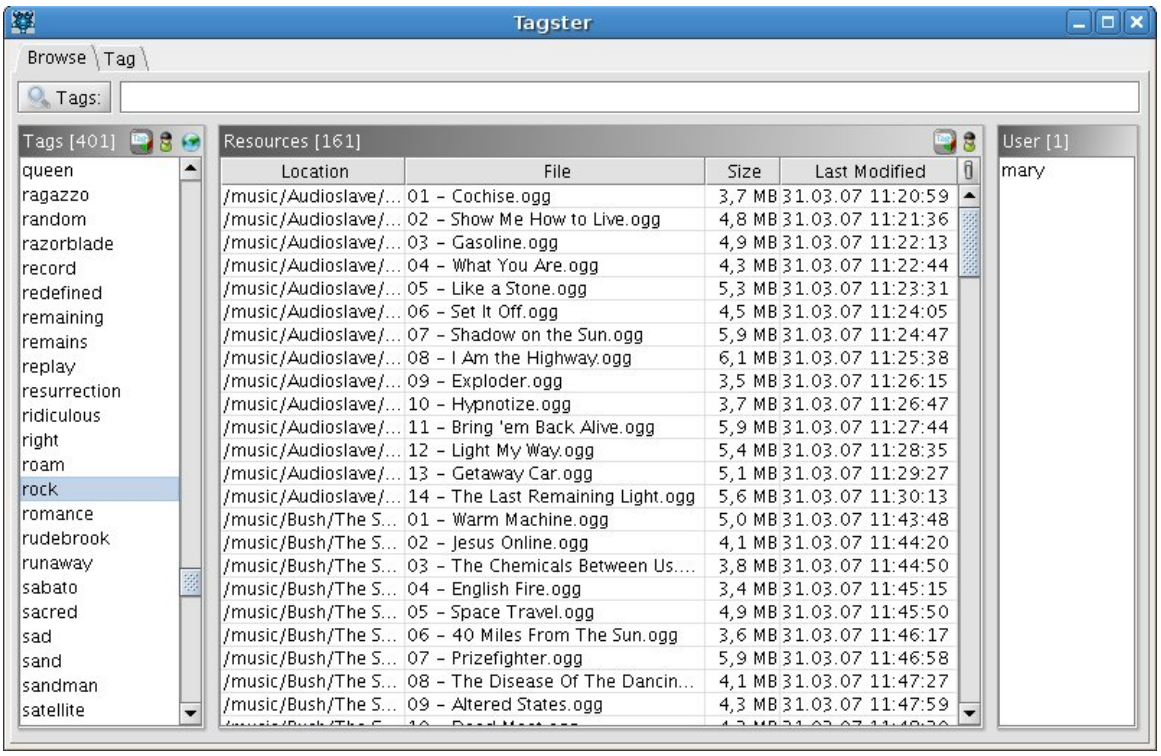

Figure 2.1: Peer-to-peer application - showing all tags and all resources tagged with “rock”.

change. The main components of this architecture are the data repository, the networking modules and the kernel which manages all module interaction.

All tagging data is represented as RDF[1] triples in respect to a tagging ontology. Sesame[2] is used as a persistent RDF data store. Queries against the repository are done with the SPARQL[3] query language which allows for flexible and complex queries on the tagging data.

The kernel, as the core module, has direct access to the repository and answers all requests coming from the client application or other peers. The network communication with other peer is done via two different networking layers. The first network layer is just responsible for direct interaction with other peers, e.g. downloading files. The second network layer is used to make certain tagging data statistics available in the peer-to-peer network. Such statistics can include the frequency of tag use, similar tags etc. The current prototype implements a distributed index structure that holds user-tag, user-resource, and resource-tag relations. For example, accessing a resource ID in the index will return all tags assigned to that resource and all users who tagged it.

The index has been realized with Bamboo[4], a distributed hashtable (DHT) implementation which allows for efficient access to the stored data. DHTs generally ensure that within a given id space, evenly distributed among all peers, all key/value pairs get assigned to exactly one responsible peer and can be stored and retrieved with at most $log(n)$ routing hops.

Another feature of the system is the detection of firewalls because users behind firewalls may not be able to communicate with other peers. So far, no special mechanism is implemented to circumvent firewalls directly but the user is informed about the connection problems and advised to open the necessary ports.

[1] http://www.w3.org/RDF/
[2] http://openrdf.org/
[3] http://www.w3.org/TR/rdf-sparql-query/
[4] http://bamboo-dht.org

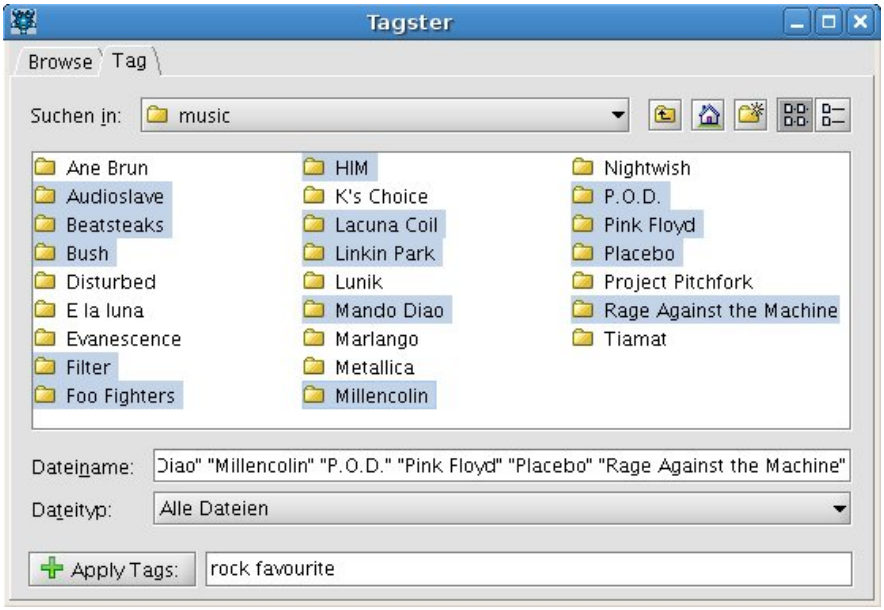

Figure 2.2: Peer-to-peer application - file browser like view for tagging data.

2.3 User Interaction

The setup of the system is intended to be as simple as possible. The user just downloads the java archive which can basically run on any platform. When the application is started for the first time, a user has to type in her name and specify where the local database should be stored. Thereafter, she can start to tag local files through a special file browser view. In addition to the usual directory and file view a text field allows to type in tags that should be assigned to the selected files (c.f. fig. 2.2). The application also supports the automatic extraction of tags from the file and path name and certain multimedia metadata like ID3 tags of audio files. For batch tagging of files the user just needs to select a whole folder and all contained files will be added with the specified tags.

The main application view displays all existing tags and the respective resources (c.f. fig. 2.1). If one tag is selected only those resources are shown that have the selected tag assigned. For each resource all other assigned tags are shown as well and can be edited directly by adding and removing tags.

Additionally, some multimedia data inspection features have been implemented to automatically extract specific multimedia metadata which can be directly added as tags. This helps to simplify the tagging process - especially for larger data sets like multiple music files extracted from a CD.

2.3.1 Data collection

Data from the peer-to-peer tagging system is also subject for further analysis. For each tag assignments the combination of user, tag and resource plus a time stamp is stored in each peers database. Additionally, information about the connected buddies is available. Currently, every peer's data has to be aggregated manually but future versions should include some more sophisticated mechanism to gather all of the distributed data.

Bibliography

Thomas Franz, Carsten Saathoff, Olaf Goerlitz, Christoph Ringelstein, and Steffen Staab. Sea: A lightweight and extensible semantic exchange architecture. In *Proceedings of the 2nd Workshop on Innovations in Web Infrastructure. 15th International World Wide Web Conference (Edinburgh, Scotland)*, 2006. URL `http://www.uni-koblenz.de/~saathoff/publications/sea.pdf`.

Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. BibSonomy: A social bookmark and publication sharing system. In Aldo de Moor, Simon Polovina, and Harry Delugach, editors, *Proceedings of the First Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, pages 87–102, Aalborg, 2006. Aalborg Universitetsforlag. ISBN 87-7307-769-0. URL `http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hotho2006bibsonomy.p%df`.