Collaboration Support for Bibliographic Data

Erik Wilde, Sai Anand, Thierry Bücheler, Nick Nabholz, Petra Zimmermann
Computer Engineering and Networks Laboratory (TIK)
Swiss Federal Institute of Technology, Zürich (ETH Zürich)

Abstract

In many research settings, bibliographies are a central resource for collecting information about related work, keeping track of the own research record, and annotating this information with remarks. By its very nature, this information should be shared between researchers within a research group and maybe in larger organizational units (for example research institutes) as well. However, most tools used for managing bibliographic data do not support collaboration. Using ShaRef, users can share bibliographic information, collaborate, and publish and export data using a variety of output channels. ShaRef's goal is to make sharing of and collaboration with bibliographic information easier than it is today.

1 Introduction

Research is based not only on innovation, but also depends heavily on knowing about other innovations and understanding the implications of these for the own work. Thus, research is a task which depends on knowing about other research, and in most cases today, this is accomplished by using bibliographies, which are often annotated with additional remarks. From the knowledge management point of view, bibliography management is the closest activity comparable to formal knowledge management that most researchers will ever practice [10].

Surprisingly, only few tools support collaboration and sharing with bibliographic data, most users today use stand-alone software tools, the most popular being BibTeX and EndNote. The *Shared References (ShaRef)* project's goal is to build a tool which is open, supports document preparation with both LaTeX and Office/OpenOffice, and supports collaboration and sharing with bibliographic data. Additionally, ShaRef is targeted to be a building block within a general information management infrastructure, and thus is designed to cooperate with other applications, for example content management systems for integrating publication data, or publication databases for purposes such as collecting publication data for compiling publication lists for research groups, institutes, departments, or even complete universities.

In this paper, we concentrate on the aspects of ShaRef which are central for supporting communities of collaborating researchers, and the way they can contribute their information and knowledge to others. The relevant aspects are on the one hand ShaRef's sharing and collaboration facilities, which are described in Section 4.1. These facilities allow researchers to jointly work with bibliographic data, thus aggregating information and knowledge about their particular field of expertise. In addition to this, Section 4.2 describes the ways in which the information can be published and exported, thus allowing people outside of the research community to benefit.

Before we describe these features of the ShaRef system, we first briefly describe the system itself by describing its data model in Section 2, and a possible application scenario in Section 3. Throughout this paper, we refer to ShaRef as a system which can be used by almost any user. ShaRef is based on a server/client-architecture and implements two clients, one being a Java-based rich clients which requires a Java Runtime being installed on the client, and the other being a Web-based client which can be used by any client using a standard Web browser.

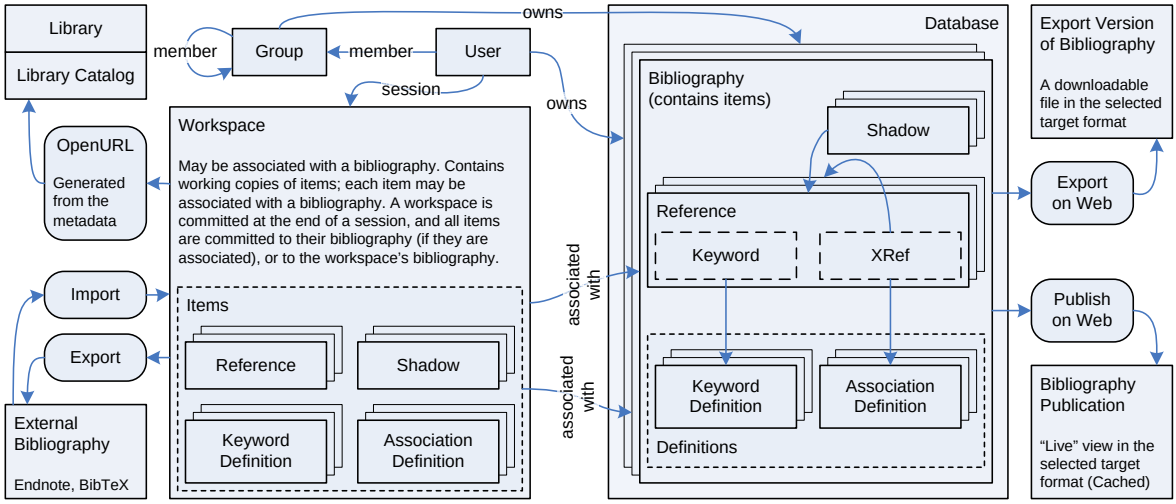

Figure 1: ShaRef Data Model

2 Data Model

ShaRef uses a hypermedia-inspired data model [12] shown in Figure 1. While bibliographic data is represented using standard fields comparable to the *Dublin Core Metadata Element Set* [5], ShaRef also supports different kinds of relationships between references, and different ways to describe references.

Shadows and *Cross-References* are two different ways to reuse existing references. *Shadows* are a way to create a live link to an existing reference (comparable to Unix *Symbolic Links*), any changes in the original reference will be reflected in the the shadow. A shadow can have additional annotations, but no other additional fields. *Cross-References* are a way to reuse parts of a reference, they are inspired by BibTeX's `crossref` field, which is an easy way to reuse parts of a reference, for example a volume of proceedings or a book with articles by different authors. The volume then is described by a reference, and the cross-references are using the volume's reference and only add the additional information, for example the chapter title, number, author, and pages, while the volume information (for example the volume title, publisher, date, and ISBN) is reused.

Associations describe semantic relationships between publications (i.e., the resources described by references), for example the fact that one publication is an updated version of the other. There is no built-in set of association semantics in ShaRef, but different user communities can define and reuse association types by using ShaRef concept of shared bibliographies.

Keywords are a way to describe concepts, and keyword definition can be referenced from within references, or from within other keyword definitions. As with associations, there also is no set of predefined keywords in ShaRef, but again these can agreed upon in user communities, and can then be reused by means of bibliography reuse.

A *Bibliography* is a set of entries (which may be references, shadows, and association or keyword definitions), this concept is resembling a file in BibTeX or a database in EndNote. The concept of a bibliography is important in ShaRef because the bibliography is the unit of access management and control. Bibliographies have owners, writers, and readers, and access rights of users for a bibliography and the entries within it are enforced using the identity of a user and the access control information for this bibliography.

Finally, all bibliographies of a ShaRef installation constitute the so-called *Database*, and again this concept is important because it defines the limits of what references between ShaRef entries can address. Thus, it is only possible to reuse references and association and keyword definitions within the realm of one ShaRef database, everything outside of the database may be addressed as a Web resource with a URI, but is no longer part of ShaRef's built-in data model.

Since ShaRef supports access management and control, it needs a way to identify and authenticate users. ShaRef uses a user and group management concept with a simple model where groups may contain groups, and this way it is easily possible to create any kind of structure which is required to control access to ShaRef data. Users may be identified and authenticated using a built-in mechanism, or by using an external service, such as an existing authentication service where people are already registered.

3 Application Scenario

A typical application scenario (and the setting the design use cases were based on) is a research group working on a particular topic. ShaRef in this case serves as the central repository for collecting publication data about this topic (both related work and the own publications). This publication data can be either traditional bibliographic data (books, journal articles, conference papers, and similar forms of publication), but it may also be data about Web resources, which in most cases simply have a URI and a title. In this case, ShaRef also serves as a "centralized bookmark" repository, and the advantage in comparison to other bookmark management software is that all ShaRef concepts such as associations and keywords can also be applied to describing bookmarks, and that the bookmarks can be related to traditional bibliographic data (for example, they can be associated with a book).

As a starting point, researchers can import their existing bibliographies into ShaRef, using its import filters. They can freely decide to either put everything in a common project bibliography, or to keep personal bibliographies and make these available for reading and/or writing to other project members. Additionally, freely available publication data may be imported into the system, for a computer networking research group for example the Internet RFC database, which is available as an XML document at `http://www.rfc-editor.org/rfc.html`. Using ShaRef's open XML data model, it is easy to write an import filter transforming this information into ShaRef data.[1] After this transformation, this publication data can be imported into the system as well, and it is available to be used in bibliographies or associations from other references.

After this initial setup phase, ShaRef serves as the central repository for publication information, and if users wish so, they can always keep "their bibliography" as a separate bibliography, but they can still make it available to others (an initial study [11] conducted before the start of the project showed that the majority of respondents found it essential to be able to control "their personal information", even though they would find it useful to share it with others). Depending on the setup, the publication data may thus be an anonymous set of references in a central bibliography, or it may be a personalized collection of bibliographies where entries maybe associated across bibliographies, but there always is a well-defined owner for each entry because all bibliographies are owned by individual users.

Regardless of how the system is being used, the information and knowledge inside the ShaRef system can be used for working directly with the system (browsing reference data, updating it, or inserting new entries), but it should also be made available to others, such as in the following examples:

- *Reading Lists:* For new project members, people interested in the research group's field, or scholarly activities, it may be useful to be able to create targeted reading lists. By creating a separate bibliography and populating it with shadows of all entries that should be part of the list, it is possible to create a list which is not a snapshot, but a live view of the research group's bibliography as it evolves. New shadows can be added to the reading list whenever new references are found to be a useful addition to the list.

- *Document Preparation:* When creating documents, project group members may use different document preparation systems, and they will thus require different bibliography formats. These formats can be exported and then used for document preparation. ShaRef also allows to publish these bibliography formats, which means that it is possible to download the most recent version of a set of references at any time.

[1] `http://dret.net/rfc-index/` shows (as a ShaRef HTML export) how such a transformation of an existing publication database into the ShaRef system can create an highly interlinked information repository.

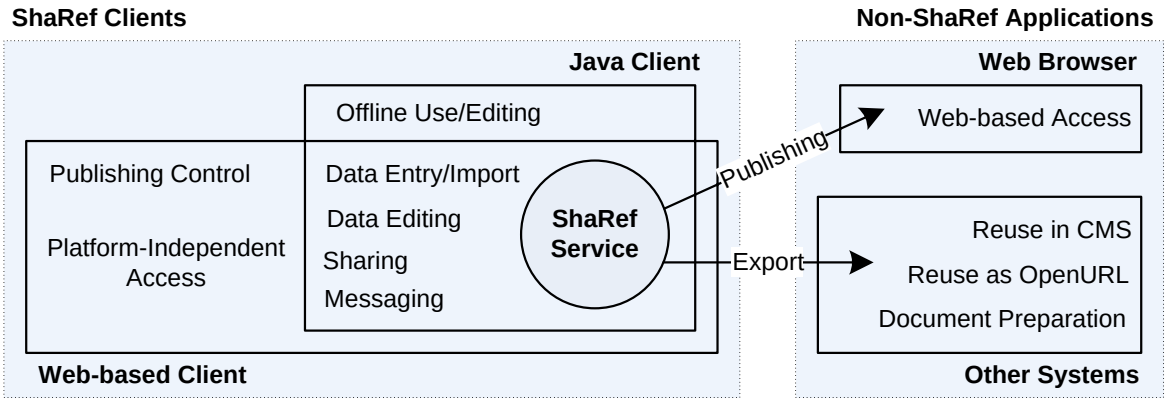

Figure 2: ShaRef Use Cases

- *Publication Lists:* For milestone reports or the final report, it may be necessary to create publication lists. These lists may be necessary at different levels, for example for compiling the project publication list, the research institute's publication list, and finally the university's publication list. Through a messaging mechanism it is possible to send and receive bibliographic data, and if this data contains shadows, the recipient may keep it this way (in which case it may change when the original is changed), or may instantiate the shadows, thus creating a persistent snapshot of the publication list data.

The reference information is stored in a central database and can be retrieved and used using one of the available clients (Java or Web-based are the current implementations), but it can also be made available through configuring publishing channels, which publish the information through a URI. This makes it possible for any user with a Web browser to access this information, and it can also be reused by applications retrieving the information through a simple HTTP `GET` (this design is often referred to as an *HTTP Web Service*).

The latter possibility (the HTTP Web Service) can be used to re-use ShaRef data inside other systems, such as Web site which should contain publication lists. Rather than exporting the data and copying it into the Web site (which creates redundancy and thus update problems), it is possible to dynamically include the data from within ShaRef through the HTTP Web Service, which can then be rendered as an integral part of the Web site.

4 Supporting Communities

The general design goal of the ShaRef system has been to support communities of collaborating researchers in their activity of gathering, reviewing, and modifying information about scholarly resources. Since a new tool cannot be expected to be used by everybody, the design has been made as open as possible, allowing to include other users in the community as well.

This general design guideline is shown in Figure 2, which shows the two different ShaRef clients on the left side (a Java-based and a Web-based client), and the two most important other scenarios where the publication data can be used, which are Web browsers, and any applications processing publication data.

While the design has been centered around the idea of being non-exclusive, there are a number of activities which can only carried out while staying inside of the system (i.e., on the left side of the figure), and these are described in Section 4.1. Other activities, however, are possible without using the system itself (i.e., on the right side of the figure) but by only using data which has been produced by the system, and these activities are described in Section 4.2.

4.1 Sharing and Collaboration

Because the system should support well-defined user groups, it has not been designed to be an anonymous system. In order to assign access rights to data and to control the authorization of users to access data, it is necessary to introduce the concepts of *access control* and *authentication* as described by Tolone et al. [8]. To better support larger communities and groups of users, a user and group management and access control system has been designed and integrated into the system [6]. Using this component, it is possible to manage users, create groups of users, and assign access rights to bibliographies.

The unit of access control in ShaRef is the bibliography, it is not possible to use access control on a finer level (i.e., on bibliography entries). This makes the system easier to understand and easier to manage, but in some cases it may force users to split bibliographies for access control reasons only. The design decision was to accept these cases with the goal to create an easier-to-use system.

Since in many research setting users are already registered through some registration service, the authentication task can either be processed within the system, or by an external service. In this case, a user logging in to the system supplies the user name and the password, which are then forwarded to the external service. If the external service authenticates the user, this external authentication is accepted by ShaRef and no internal authentication in performed.

After a user has logged in, the identity is known throughout the session and can be used for access control. A user can be a member of one or more groups. Users and/or groups can be members of groups (cyclic structures are prohibited). The idea is that hierarchical structures of organizations can thus be easily reflected in the organization of the users and groups. Determining whether a user is authorized for some action is thus equivalent to testing whether this user is directly or indirectly (through group membership) authorized for this action.

Sharing of data is accomplished by assigning access rights to a bibliography which then can be accessed (in read or write modes) by the users and/or groups which have been granted access. Thus, it is easily possible to create personal bibliographies (no read access for anybody) or fully public bibliographies (read or even write access for everybody). anything is between can be accomplished by using the user and group management features, and this usually involves some kind of coordinated setup of the group structure so that the required access structure can be implemented. By combining the user and group management facilities with the data model aspects of bibliographies and shadows, it is possible to setup many different scenarios.

Collaboration either takes place indirectly and asynchronously by working with the same bibliographies, and in this case the collaboration's goal is to evolve and improve the bibliography data. There also is a messaging feature which allows a more direct form of collaboration, allowing users to send messages to other users or groups. These messages may either be text messages, or they may contain bibliographic data, which can then be reused by the recipient in every way supported by the system, for example by exporting it.

To better achieve the design goal of being non-exclusive, there are two clients available, they are shown in Figure 3. The Java client supports all features that a rich client can support and which are impossible to implement in a Web-based client. However, the Java client must be installed on a system before it can be used, and it has some installation requirements (JRE 1.5) which may not be met by all systems. Therefore, a Web-based client [2] is available which supports most of the functionality of the Java-based client.

Naturally, the Web-based client does not support some of the functionality which cannot be implemented in a standards-compliant Web-based client, such as complex menu structures and multi-window interfaces. The most important restriction of the Web-based client, however, is the fact that it depends on the server to be available. This is quite natural for a Web-based client, but limits the client to online scenarios.

The Java client, on the other hand, supports a so-called "offline mode", where it download the contents of one or more bibliographies to a local database, and then makes these available offline. When going back online, the contents of the offline database are synchronized with the online data. This way, users can still use their bibliographic data inside of the system, even when being offline.

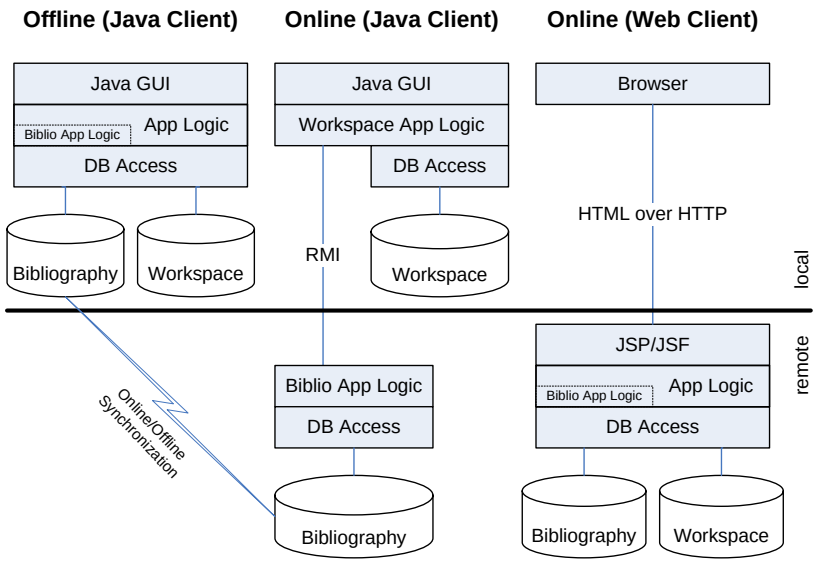

Figure 3: ShaRef Java and Web-based Clients and Bibliography Access

4.2 Publishing and Exporting

The Web-based client makes it easy for users to use the system, but still requires users to log on and authenticate themselves, and then it provides the interface for managing bibliographic data. For some users, this may already be more than they want, or it may simply be too complicated to log on to the system in order to simply browse through some bibliographic data.

Through the publishing mechanism, data can be made available through a publishing channel, which is a dedicated URI where the publishing data can be retrieved. This publishing data can be retrieved at any time and is a live view of the bibliographic data inside the system. The publishing mechanism can be described as a sequence of three steps:

1. *What?* As a first step, it must be decided which data to publish. This is like a find function, it allows users to define the data source in terms of a bibliography and an optional filter.

2. *How?* After selecting what to publish, it must be decided how to publish it. This involves selecting the publishing format (XML, HTML, BibTeX, EndNote), and the options for this format.

3. *Where?* As the final step, the selected publishing format must be made available at a known URI. It is possible to either specify a URI (or rather a part of it, because the prefix is determined by the server name of the installation), or to let the system specify a URI.

Publishing of contents is a convenient way to let collaboration partners share bibliography data in the format of their choice. The only restriction is that this data is read-only data, because it is published through the system in the required target format, but any modifications of the data have to made within the system (by using one of the two clients).

Another possibility for linking the bibliography data inside the system with the outside world are *OpenURL* [1] URIs. These URIs allow users to follow links into their preferred library catalog. They have to configure the prefix of their library's OpenURL server, and then can follow links from their ShaRef bibliography into this library catalog, for example inquiring whether a particular publication is available. On the implementation side, this can be achieved by encoding the publication metadata in the OpenURL format, which is then received and analyzed by the library server. From that point on, users are inside the library catalog system as if they had manually entered the publication metadata

to search in the library catalog. This makes it easier for users to use their personal data to find data in other systems, and in the same spirit we generate links to Google, which make it easy to search the Web for resources related to a publication's metadata.

While some target formats are given (such as OpenURL, HTML, BibTeX, and EndNote), others are more proprietary and impossible to implement in advance. One such example is the XML format for the *Content Management System (CMS) Silva*. In order to publish data that can be incorporated into Silva, it must be transformed into the right XML. This XML must on the one hand conform to Silva's XML schema, but it must also conform to the guidelines of a specific CMS installation, which in most cases are defined by Web designers. While this requires some programming on the system side, it can be easily done by using *XSLT 2.0*, which is used as the transformation language inside the system.

Generally, the whole system is XML-based and has been designed for easy migration from and to ShaRef. This way, transformations for import or export/publishing filters can be implemented with a minimum of effort, and it is thus easy to integrate new sources or targets into the system.

5 Discussion

In the following sections, we discuss related work, the contributions which make ShaRef an interesting candidate for cooperation in communities of collaborating researchers, and future work that we have in mind for future developments of the system.

5.1 Related Work

The ShaRef system as it has been presented in this paper is unique in its combination of features for bibliography management, sharing, and import and export/publishing.

Since the most popular tools for bibliography management are single-user tools, a number of Web-based platforms such as *CiteULike* or free software such as *JabRef* or *BibShare* [3] have been developed. These solutions lack the authentication and access control features of ShaRef, which are essential in research settings where collaboration should be supported, but must be controlled.

Targeting the same or similar ideas as ShaRef, a number of commercial products such as *RefWorks* or *Net Snippets* are available. These solutions lack the open and non-exclusive approach of ShaRef, which is based on an open and XML-based data model and has been designed to work in an open environment.

Apart from Web-based and free or commercial products, there are also a number of research projects which are targeting similar ideas. The *ClaiMaker* [9] system is more advanced in the area of how to associate publications (it has a built-in ontology which allows reasoning with these associations), but lacks the general usability issues such as easy integration into a heterogeneous environment of existing bibliographies. The *Bibster* [4] project focuses on formalizing semantics and a peer-to-peer architecture, but also lacks the features which would make it usable in a heterogeneous environment. The *Hunter Gatherer* [7] approach is not targeted at publication data but at excerpts of Web pages, but it focuses on the question of how people work when interaction is introduced in the information gathering process.

5.2 Contributions

The main contribution of ShaRef is the combination of features which make it a useful addition to a general infrastructure of knowledge tools within organizations. The user and group management concept can be used to reflect organizational as well as task-oriented groups, and thus enables users to quickly form communities that they find useful for information sharing. Deploying ShaRef inside an organization in most cases requires some form of configuration and customization, and the open XML-based design is ideally suited to accomplish this by allowing new transformations of import and/or export/publishing formats to be supported by the system.

5.3 Future Work

One area of improvement would be a fully functional API through a Web Service interface such as WSDL/SOAP, which would fully expose the publishing interface. Currently, publishing has to be configured through the Web-based client, and the HTTP-based Web Service can only be used to retrieve the data being published. Depending on requirements from other applications, such an interface may be added in the future, improving the programmatic access to the system.

6 Conclusions

The ShaRef system has been designed to support groups of collaborating researchers. Its primary task is the shared management of publication data, and this is accomplished through a data model of access-controlled bibliographies. The goal of the system design is to be non-exclusive, which is implemented through an open XML-based data model, supporting import and export of other formats. Furthermore, the export and publishing features allow users which are not using the system itself to reuse the system's data in other applications, for example in document preparation systems.

Apart from the above features, the system also supports scenarios where data need to be reused in different contexts, such as the management of publication lists in organizations, and the reuse of publication data for different activities, for example reading lists for lectures or courses.

References

[1] American National Standards Institute. The OpenURL Framework for Context-Sensitive Services. ANSI/NISO Z39.88-2004, April 2005.

[2] Thierry Bücheler. ShaRefWeb: A Web Interface for the ShaRef Service. Master's thesis, Computer Engineering and Networks Laboratory, Swiss Federal Institute of Technology, Zürich, Switzerland, October 2005.

[3] José H. Canós and Eduardo Mena. BibShare: An Interoperable System to Access and Maintain Bibliographic References. In *III Jornadas de Trabajo DOLMEN*, Madrid, Spain, November 2002.

[4] Peter Haase, Björn Schnizler, Jeen Broekstra, Marc Ehrig, Frank van Harmelen, Maarten Menken, Peter Mika, Michal Plechawski, Pawel Pyszlakand Ronny Siebes, Steffen Staab, and Christoph Tempich. Bibster — A Semantics-Based Bibliographic Peer-to-Peer System. *Journal of Web Semantics*, 2(1), 2005.

[5] International Organization for Standardization. Information and Documentation — The Dublin Core Metadata Element Set. ISO 15836, November 2003.

[6] Nick Nabholz. Ein Benutzerkonzept für kollaborative Applikationen am Beispiel von ShaRef. Master's thesis, Hochschule für Technik, Wirtschaft und Verwaltung Zürich, Zürich, Switzerland, June 2005.

[7] Monica C. Schraefel, Yuxiang Zhu, David Modjeska, Daniel Wigdor, and Shengdong Zhao. Hunter Gatherer: Interaction Support for the Creation and Management of Within-Web-Page Collections. In *Proceedings of the Eleventh International World Wide Web Conference*, pages 172–181, Honolulu, Hawaii, May 2002. ACM Press.

[8] William Tolone, Gail-Joon Ahn, Tanusree Pai, and Seng-Phil Hong. Access Control in Collaborative Systems. *ACM Computing Surveys*, 37(1):29–41, March 2005.

[9] Victoria Uren, Simon Buckingham Shum, Gangmin Li, John Domingue, and Enrico Motta. Scholarly Publishing and Argument in Hyperspace. In *Proceedings of the Twelfth International World Wide Web Conference*, pages 244–250, Budapest, Hungary, May 2003. ACM Press.

[10] Erik Wilde. References as Knowledge Management. *Issues in Science & Technology Librarianship*, (41), Fall 2004.

[11] Erik Wilde. Usage and Management of Collections of References. Technical Report TIK Report No. 194, Computer Engineering and Networks Laboratory, Swiss Federal Institute of Technology, Zürich, Switzerland, June 2004.

[12] Erik Wilde. Shared Bibliographies as Hypertext. Technical Report TIK Report No. 224, Computer Engineering and Networks Laboratory, Swiss Federal Institute of Technology, Zürich, Switzerland, May 2005.