Usage and Management of Collections of References

Erik Wilde
ETH Zürich (Swiss Federal Institute of Technology)

TIK Report 194 (June 2004)

Available at http://dret.net/netdret/publications#wil04h

Abstract

Collections of references are an important part of scientific and scholarly work. For many people, collections of references are the most advanced form of formal knowledge representation they are using. However, today's standards and tools for collections of references are pretty poor, providing closed environments with no or little extension mechanisms. In this paper, we describe our goal to design and implement an advanced system for collections of references. The primary goal of this system is to provide users with a tool that matches their requirements of semantic richness vs. usability, which are competing goals. As a first step towards this goal, we designed and conducted a survey among the employees of a large university, trying to find out how people are managing their references today, and what their expectations are if a new tool became available. The results of the survey are presented, followed by conclusions that are the guiding principles for the upcoming ShaRef project. The goal of this project is to design and implement a system for reference management that runs Web-based as well as stand-alone, is easy to use, supports collaborative collections of references and collection sharing, has an open and extensible data model, covers the majority of user requirements according to the 80/20 principle, and thus provides scientists and scholars with a better way to manage their collections of references.

1 Introduction

Digital access to libraries and digital libraries are indispensable tools for scientists and people seeking access to scientific and other information. However, an often overlooked but nevertheless important tool in the scientific community are personal bibliographies, often accumulated over the duration of many years, which reflect a scientist's interests, the scientific literature read so far, and maybe additional information, such as comments or keywords in the bibliography. Consequently, bibliographies do not in any way compete with digital technologies for libraries, but instead are an important complementary tool for the personal information management of scientists. Since bibliographies are often personal, they receive much less attention than digital technologies for libraries, which are easily visible for many users and are located at a central location in the process of digital library access.

In this paper, we describe the results of a survey at the *Swiss Federal Institute of Technology Zürich (ETHZ)*, which has been conducted as the initial step for a project targeting the design and implementation of a tool for generic reference management. In Section 2 we describe how references are handled today. Since the final goal of the project is to provide a

tool that is based on actual user requirements, we conducted a university-wide survey, which is described in Section 3. Based on the survey results, we then discuss the two most interesting issues in more detail, the nature and scope of references in Section 4, and the questions of schemas for references in Section 5. Some remarks about our focal points for future work conclude the paper.

2 Collections of References

Collections of references are an important tool for many scientists, and for the majority of scientists, they are the most advanced step they take in the general direction of knowledge representation. In the most generic sense of the word, a reference is "something (as a sign or indication) that refers a reader or consulter to another source of information (as a book or passage)" (Merriam-Webster). Our culture is based on very many different types of references, and if we want to provide tools for generic reference management, we must make some assumptions about the references and reference management that we want to support:

- *Referable Resources:* Collections of references are most useful when the resource being referred can be uniquely identified through some sort of standardized reference mechanism. Reference mechanisms are often introduced when experience has shown that a well-defined reference mechanism would be of value. Popular examples for established referencing schemes are the *ISBN* [3] system, the Web's *Universal Resource Identifier (URI)* [2] concept, and the *Digital Object Identifier (DOI)* [1] of the library world.

  Given the broad range of many reference specifications, many things may be referenced. For example, the current list of assigned URI schemes maintained by the IETF[1] lists almost fifty schemes, ranging from well-known schemes such as `http` and `ftp`, to more exotic schemes such as `tel` and `fax` for telephone and fax numbers. Basically, everything that has been assigned a URI scheme can be referenced by a URI. DOIs also cover a very broad area, with the standard stating that a DOI is "used to identify intellectual property in the digital environment".

  However, since references often are managed in a way specific to the resources being referenced, it would not be a useful approach to manage all references using a single model and tool. Nobody (or almost nobody) manages an address book (references to people and metadata associated with these references) and bibliographic references in a single tool, because for most users these two issues are too far apart. Even though one might argue that it could be beneficial to have a unified data model and supporting tools for people and bibliographic resources (after all, authors are people), it is not our goal to create and support a unified data model for all types of references. However, in Section 4 we argue that the references to Web and bibliographic resources should be managed in a unified way, because these two worlds are clearly converging.

- *Organization of References:* In many cases, references must be annotated to be of any use. In principle, two ways of annotations can be distinguished. *Unary* annotations are metadata associated with exactly one reference, describing interesting facets of the resource being referenced (for example, keywords) or the reference itself (for example, the creation date of the reference). On the other hand, *n-ary* annotations connect various

[1] `http://www.iana.org/assignments/uri-schemes`

references by describing relations between resources (for example, a theory presented in one article has been criticized in another article) or references (for example, two references refer to two copies of the same resource).

The question of how references should be embedded into a framework providing annotation features covers a very broad area, ranging from methods as simple as the folder metaphor for bookmarks implemented by most browsers, to very complex semantic schemes including reasoning and other features known from knowledge representation and artificial intelligence. Again, our goal is not to identify the most comprehensive way of how references can possibly be organized, but instead identifying a useful subset for our application area. The most interesting question therefore is the unification of Web bookmarks and bibliographic references (discussed in more detail in Section 4), which requires a reference annotation scheme that is flexible enough to cover these two areas of references and their different characteristics.

Since references and the organization of references in their most generic way are rather abstract concepts, it is important to focus on certain application areas and provide models and tools that provide good support for these areas. Most importantly, the tools must be powerful and simple enough to enable users to create and maintain collections of references. If the model is powerful but requires cumbersome handling of obscure data formats and tools, then the resulting system may have all the properties that were required by the requirements analysis, but there will be almost no users.

Thus, our goal is to design a system that is easy to understand and easy to use. It must cover the most important aspects of references required by scientific and scholarly work, and support these aspects through a well-designed interface. It must be extensible to provide users from certain application areas and with more advanced requirements to extend the system according to their requirements, but still being able to benefit from the basic support the system provides.

In order to better understand the real users' actual behavior and their requirements, we designed and conducted a survey to find out how personal reference management is happening today among the members of a large university.

3 Survey Design, Implementation, and Results

ETHZ is a large university with more than 12'000 students (including 2'500 Ph.D. students) and 8'000 employees. Since most students are at the beginning of their scientific activities and often do not maintain their own bibliographies, the survey was conducted among ETHZ's employees only (most Ph.D. students are also employed at ETHZ as full- or part-time assistants). Not all of the 8'000 employees are working scientifically, but for the email address administration of ETHZ it would have required some effort to create a email list containing only the scientifically working employees, so the email asking to answer the survey about "Bibliographic and Reference Information Management" was sent to 8'000 people. In the first ten days after the email had been sent, we received 1'016 answers, which were used as the foundation for our survey results. The survey's Web page is http://dret.net/projects/sharef/survey1/, containing the English and German questionnaires, the survey's summarized results, the individual textual replies, and an anonymized XML document containing all individual answers.

The survey contains three parts, dealing with the different aspects of the reference management as it is today, and expectations which users may have for a new tool for reference

1.1 Do you use and maintain bibliographic information?	No: 26.4%	
	Yes: 73.6%	EndNote: 40.5% BibTeX: 20.6% Other: 12.6%
1.2 Do you share your bibliographic information with other users?	No: 41.9%	
	Yes: 33.8%	Informally: 30.1% Organized: 3.7%
1.3 How do you enter your bibliographic information?	Manually: 32.1% Copy/Paste: 18.2% Import: 23.9%	
1.5 Do you use any additional classification or information in your bibliography?	No: 34.8% Yes: 38.5%	

Table 1: Results: Bibliographic Information

management. All parts of the survey are very short, because the goal was to design a survey that should take at most 10 minutes to fill out. Otherwise, many users would have considered it to be too time-consuming to answer the survey. These were the individual parts of the survey:

1. *Bibliographic Information:* This part of the survey deals with the current management of bibliographic information, containing questions about the tool being used, private vs. shared bibliographies, the working style, the size of the bibliography, and any additional annotations being used.

2. *Bookmarks:* This part is very similar to the first part, but covers Web bookmarks rather than bibliographic references. Not surprisingly, many users consider bibliographic references and Web bookmarks to be entirely different things.

3. *User Expectations:* This final parts asks users about their expectations for a new tool for reference management. It contains questions about reference sharing, handling bibliographic references and bookmarks with a single tool, Web access to collections of references, and the willingness to invest some time to get used to a new tool.

The results of the individual parts are shown in Tables 1 through 3. All percentages in these tables are relative to the total number of replies received. Consequently, in all questions with a finer differentiation of the "Yes" part (such as question 1.1, asking for the tool being used for bibliographic information management), these numbers do not add up to 100%, but to the total percentage of the "Yes" part. In cases where the "No" and "Yes" parts do not add up to 100%, this question has not been answered by all users.

The answers to the first part of the questionnaire are shown in Table 1. These answers show that many users are collecting bibliographic information (it should be kept in mind that not all of the 8'000 people being polled are working scientifically), that reference sharing is

2.1 Do you use and maintain Web bookmarks?	No: 9.4%	
	Yes: 90.6%	IE: 46.7% Mozilla: 30.1% Other: 14.0%
2.3 Do you use additional information for managing your Web bookmarks?	No: 26.3%	
	Yes: 64.3%	Folders: 62.2% Other: 2.1%

Table 2: Results: Bookmarks

common but mostly on an informal basis (most probably file sharing), and that many users use additional semantic annotations for their bibliographic references[2]. Another question in this first part was question 1.4 asking for the total number of bibliographic references contained in the collection. The answers to this question are shown in Figure 1.

Table 2 shows the answers to the second part of the questionnaire. It shows that the vast majority of users are collecting Web bookmarks, and that most of them do so by using the browsers' bookmark facilities. Even though the browsers offer only very poor bookmark management features, the integration of bookmark management into the browser makes it a very convenient way of managing bookmarks. The majority of users are organizing their bookmarks into folders (the classification metaphor implemented by browsers), and only very few users are using additional methods of classifying bookmarks[3]. A possible interpretation is that even though the folder metaphor clearly is a weak classification mechanism, many users consider Web bookmarks to be of lesser importance than bibliographic reference, and thus invest less effort to manage them. Another question in this second part was question 2.2 asking for the total number of Web bookmarks contained in the collection. The answers to this question are shown in Figure 1.

Since the survey was designed to find out how reference management could be improved to help users managing their bibliographic and Web references, the answers to the third part of the questionnaire, shown in Table 3, are very interesting for our future work. Sharing of references is considered as being useful by 50% of the users, and roughly half of the people willing to share would like to do so collaboratively, the other half in a read-only variant.

Surprisingly for us, managing bibliographic references and Web bookmarks together is only considered useful by 30% of the users. A possible explanation for this is that many users do not yet see the value of a semantic annotations connecting bibliographic and Web references (for example, a certain concept is discussed in a book as well as on a Web page), because their current conceptual model (which most likely is only informally defined through their working style and tools being used) treats both types of references as something completely different. Since this issue is rather important to the project, it is discussed in more detail in Section 4.

Being able to access references over the Web is generally regarded as a useful feature, however many of the proponents of Web-based access also want to have an offline way of accessing their references, so providing a Web-only tool is not an option. Finally, the majority

[2] http://dret.net/projects/sharef/survey1/comments#q-1-5-2 lists the answers to this question.
[3] http://dret.net/projects/sharef/survey1/comments#q-2-3-3 lists the answers to this question.

3.1 Would you like to share your bibliography and/or bookmarks?	No: 49.9%	
	Yes: 50.1%	Read-only: 23.6% Collaboratively: 26.5%
3.2 Would you like to manage your bibliography and bookmarks using a single tool?	No: 70.9% Yes: 29.1%	
3.3 Would you like to be able to access your bibliographies and bookmarks over the Web?	No: 31.9%	
	Yes: 68.1%	Offline & Online: 53.6% Online only: 14.5%
3.4 If ETHZ provided a tool for managing bibliographies and Web bookmarks online, would you consider using it?	No: 22.2%	
	Yes: 77.8%	Must be easy: 52.5% May be complex: 25.3%

Table 3: Results: User Expectations

of users would be willing to try out a new tool for reference management, but the main focus is an easy tool with a low barrier-to-entry, rather than a complex and powerful tool that is hard to master.

Finally, the answers to questions 1.4 and 2.2 are summarized in Figure 1. It can be seen that on average, users have more bibliographic references than they have Web bookmarks, probably due to the fact that bookmarks are considered less important and persistent, so that it does not make a lot of sense to collect many of them. However, we are sure that a large percentage users also maintain "bookmarks" for many of their bibliographic references (for example, URIs of papers and articles, which in many cases will be copies within a local file system), and that these "bookmarks" have not been included in the bookmark figures.

The results of the survey were used to back a project proposal (described in Section 6) for a university-wide platform for the management of references. In addition to the statistics presented here, the survey also generated a lot of useful and interesting feedback from various users at ETHZ[4]. Some of these answers prompted us to include new directions in our original design goals (for example, we now consider adding a Web Service API to the system for making it accessible from within applications), and we expect to gain even more insight and new ideas from the still ongoing discussions with various participants of the survey.

4 Libraries and the Web

One of the main problems of the survey is that we designed it with the idea in mind that bibliographic references and Web bookmarks are very similar in nature, and that a well-designed platform for reference management should treat them uniformly. However, many users obviously did not share our perspective. This is partly due to the fact that traditionally,

[4] http://dret.net/projects/sharef/survey1/comments#q-comments lists the answers to this question (many of them are in German).

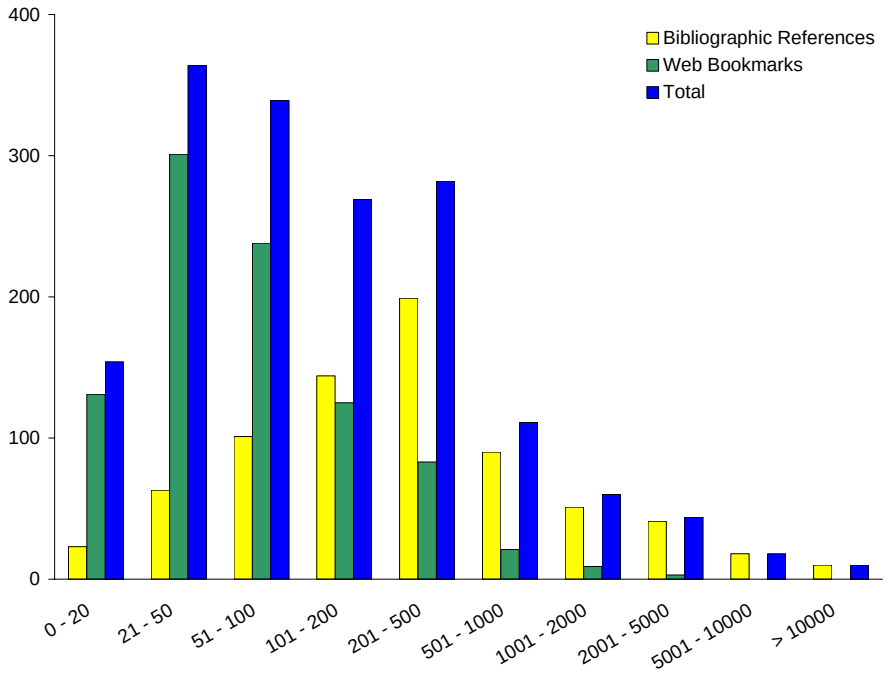

Figure 1: Results: Number of References per User

users assume that libraries and the Web are two entirely different things. This distinction is slowly disappearing, and the best way to deal with this change is to design a system that is able to cope with this convergence of the two worlds. Depending on the subject area, this convergence is happening slower or faster, but it is safe to assume that the Web as a source of information will remain relevant for a long time to come, and that for an increasing number of disciplines the boundaries between "traditional publishing" and "putting things on the Web" will become increasingly fuzzy.

Section 5 describes formats for bibliographic references, and these formats are increasingly used for references to Web resources as well. Conventions have been developed how to cite Web resources [7, 12], and the current situation thus leaves us with the bizarre effect that some Web references are kept in bibliographic reference formats (those that are considered to be useful as bibliographic references), while others are managed as bookmarks (those that are considered to be of no bibliographic value). Lawrence et al. [4] describe that the persistence of Web resources is a problem and probably will remain a problem (after all, the openness and flexibility of the Web is one of the key factors that made it a huge success), but this does not mean that Web documents should generally be excluded from reference management. It simply means that the platform providing Web reference management functions must be able to deal with this appropriately, for example by providing an automated link checking service.

We believe that the benefits of a unified reference model for bibliographic references and Web bookmarks vastly outweigh the organizational problems of Web reference persistence. The ability to maintain references to a diverse range of resources in an environment providing rich semantics for unary and n-ary reference annotations will provide more benefit to users than the inconvenience of having to deal with broken links.

EndNote Abstract Field Text:

The structure, with the chemical formula $[Ga_{20}P_{16}O_{64}F_8(OH)_4] \cdot 4C_{10}H_{24}N_4$ per unit cell, consists of gallophosphate sheets of double 4-ring (D4R) units connected to one another to form 10-rings.

EndNote XML:

```
<ABSTRACT>
    <styles>
        <style face='64' start='44'/><style start='46'/>
        <style face='64' start='47'/><style start='49'/>
        <style face='64' start='50'/><style start='52'/>
        <style face='64' start='53'/><style start='54'/>
        <style face='64' start='58'/><style start='59'/>
        <style font='23' start='60'/><style start='61'/>
        <style face='64' start='63'/><style start='65'/>
        <style face='64' start='66'/><style start='68'/>
        <style face='64' start='69'/><style start='70'/>
    </styles>The structure, with the chemical formula
    [Ga20P16O64F8(OH)4]&#x25CA;4C10H24N4 per unit cell, consists of
    gallophosphate sheets of double 4-ring (D4R) units connected to
    one another to form 10-rings.
</ABSTRACT>
```

Figure 2: EndNote XML Format

5 Bibliographic Reference Formats

There are many established formats for bibliographic references, but the most important formats for individual users clearly are BibTeX and Endnote (as indicated by the answers to question 1.1 in Table 1). In the following sections, we shortly describe these formats. These descriptions are primarily from the point of view of how easily they could be used as a starting point to migrate to a new platform, because it clearly is a requirement for users to be able to import their existing information without any loss of information.

5.1 BibTeX

BibTeX [8] is a program for the LaTeX formatting system and is very popular in the area of math-oriented sciences, because LaTeX is a very powerful typesetting system for math formulas. BibTeX uses a text-based format, and most users maintain their bibliographies using simple text editors. For users from other disciplines, this rather arcane way of working with a computer is sufficient to drive them away from BibTeX.

The data model of BibTeX was never really defined and employs some rather pragmatic solutions, such as the `crossref` field, which carries inheritance semantics when linking entries. Exporting BibTeX to other formats looks easy at first sight, but in fact proves to be very

cumbersome when all features of BibTeX (such as text macros and the built-in rules for parsing possibly multiple names) have to be fully supported. Furthermore, the poor implementation of many BibTeX styles has forced many users to include work-arounds in their bibliography, the most popular example being the forced lowercasing in many styles, which led many BibTeX users to routinely protect their content from this type of processing. Another example are unreasonable constraints, such as the inability to use certain entry types with author and editor fields.

5.2 EndNote

EndNote is the most popular product for managing bibliographic references. The two main advantages of EndNote over BibTeX are a modern GUI, and the ability to connect to online catalogues through the Z39.50 protocol, which enables user to download references instead of typing them in. EndNote also provides tight integration with Microsoft Office products, which makes it easy for users to include EndNote references in their Office documents.

EndNote stores references locally in a proprietary format, but also provides export facilities. Unfortunately, all export formats seem to lose information, and even the recently introduced XML export option does not work reliably. An example is shown in Figure 2, which shows an abstract field (as it appears in EndNote as formatted text), and its XML export format (exported on a Macintosh). It can be seen that the centered dot is exported as the Unicode character U+25CA, which is the lozenge sign (usually rendered as ◊). Through the `font` element, it is then mapped onto the glyph of a centered dot, but a reasonable XML export would have exported the centered dot as Unicode character U+00B7 without any font mapping. None of the XML export is documented, and in addition to this exotic behavior on a Macintosh platform, the Windows version of EndNote exports `&#x0;` for the centered dot (the Unicode NUL character), which is not even well-formed XML.

5.3 Comprehensive Formats

BibTeX and EndNote are rather simple reference formats, they are based on end-user needs rather than libraries and their complex cataloguing requirements. Even with a focus on XML, many complex formats such as *MARC XML* and *XMLMARC*, or MARC subsets such as the *Metadata Object Description Schema (MODS)* are available, but the differences between these (and many others) are often subtle and only meaningful in a certain application context. Instead of selecting a concrete format for representing references in our user-centered architecture, the following list of requirements has been identified as relevant for the reference format:

- *Expressiveness:* Since most users will be migrating from existing collections of references, the format should be a semantic superset of the two most popular formats BibTeX and EndNote. However, the expressiveness should be targeted to hit the 80/20 point, since the openness and extensibility (see next item) of the format makes it easy for users to develop and use their own format extensions.
- *Openness and Extensibility:* The reference format should be open and extensible, so that additional user requirements can be handled. This means that users may use the format as container for their more advanced data models, which they can built on top of the integrated basic model.

- *Reference Linking:* Since the main focus are end users, the format must support the interconnection of references, thus enabling users to link references together in semantically meaningful ways. Collections of references for many users are an important way of personal knowledge management, and being able to connect things is essential for even the most modest knowledge representation efforts.

We have not yet decided which model and format to use, but we are sure that it will be based on XML. In all the dimensions that we have mentioned, there are more ambitious approaches documented, such as the *ClaiMaker* [6] system for semantic linking of scholarly documents, or the *CiteSeer* [5] approach for autonomous citation indexing. However, rather than concentrating on one area, it is our goal to find a combination which is easy to use, appeals to the majority of users, and re-uses existing and proven concepts and technologies.

6 Conclusions and Future Work

The survey presented in this paper has been conducted as the starting point for the *ShaRef Project* (for "Shared References"), which runs from mid 2004 until the end of 2005. Currently, the project has not yet started, but the project's home page `http://dret.net/projects/sharef/` will be updated regularly to give up-to-date information about the project's status and publications. The project goals are as follows:

- *Information Sharing:* Information can be shared among users through published information (read-only) or collaborative management of collections of references (based on authentication and authorization features).
- *Open and Extensible Reference Management:* An open data model that can be extended by users to fit their personal reference management requirements. The platform provides basic, but extensible support (for example, through the use of XSLT plugins for displaying data).
- *Flexible Import and Export:* Popular formats such as the formats described in Section 5 should be supported as import and export formats. Import and export are lossless operations.
- *Web-based and Stand-alone:* The platform can be used Web-based (through a GUI or a Web Service API) or as a stand-alone application.

To reach these goals, it is planned to conduct a second, more detailed and more targeted survey among potential users of the ShaRef system from different scientific disciplines. These users will also serve as pilot users to test early prototypes. As starting point, we can use existing BibTeX converter tools [10] and our existing Web-based prototype [13], which has been implemented as a PHP/MySQL application.

Related work that we are aware of are the *RefDB*, *BibWord/BibShare*, and *RefWorks* projects, with the first one being an open source initiative, the second one a Microsoft-funded research effort, and the third one a commercial product. However, the combination of openness and extensibility, Web-based and stand-alone operation, and the ability to share information are unique properties of the ShaRef project.

We have not yet examined the possibilities of integrating the ShaRef architecture with technologies and platforms such as OpenURL [11] and CrossRef [9], but one of the main issues of ShaRef is the integration with existing information resources. Therefore, interfaces to important technologies or systems will either be integrated into the ShaRef system itself, or using the XML-based data model of ShaRef and the openly accessible API, it will be easy to integrate additional systems as add-ons to ShaRef's basic services.

7 Acknowledgements

I would like to thank Marc Petitmermet for his help regarding EndNote's idiosyncratic XML format. His complete notes regarding EndNote's XML can be found at http://www.mat.ethz.ch/research/publications/xml_format.

References

[1] American National Standards Institute. Syntax for the Digital Object Identifier. ANSI/NISO Z39.84-2000, May 2000.

[2] Tim Berners-Lee, Roy T. Fielding, and Larry Masinter. Uniform Resource Identifiers (URI): Generic Syntax. Internet draft standard RFC 2396, August 1998.

[3] International Organization for Standardization. Information and documentation — International Standard Book Numbering (ISBN). ISO 2108, 1992.

[4] Steve Lawrence, Frans Coetzee, Eric Glover, David Pennock, Gary Flake, Finn Nielsen, Robert Krovetz, Andries Kruger, and C. Lee Giles. Persistence of Web References in Scientific Research. *IEEE Computer*, 34(2):26–31, 2001.

[5] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.

[6] Gangmin Li, Victoria Uren, Enrico Motta, Simon Buckingham Shum, and John Domingue. ClaiMaker: Weaving a Semantic Web of Research Papers. In Ian Horrocks and James A. Hendler, editors, *Proceedings of the First International Semantic Web Conference*, volume 2342 of *Lecture Notes in Computer Science*, pages 436–441, Sardinia, Italy, June 2002. Springer-Verlag.

[7] Xia Li and Nancy Crane. *Electronic Styles: A Handbook for Citing Electronic Information*. Information Today, Inc., Medford, New Jersey, September 1996.

[8] Oren Patashnik. BibTeXing. Technical report, February 1988.

[9] Ed Pentz. CrossRef: A Collaborative Linking Network. *Issues in Science & Technology Librarianship*, (29), Winter 2001.

[10] Luca Previtali, Brenno Lurati, and Erik Wilde. BibTeXML: An XML Representation of BibTeX. In *Poster Proceedings of the Tenth International World Wide Web Conference*, pages 64–65, Hong Kong, May 2001. ACM Press.

[11] Herbert Van de Sompel and Oren Beit-Arie. Open Linking in the Scholarly Information Environment Using the OpenURL Framework. *D-Lib Magazine*, 7(3), March 2001.

[12] Janice Walker and Todd Taylor. *The Columbia Guide to Online Style*. Columbia University Press, New York, October 1998.

[13] Erik Wilde. Towards Federated Referatories. In *Proceedings of the SINN03 Conference on Worldwide Coherent Workforce and Satisfied Users*, Oldenburg, Germany, September 2003.