

# Intelligent bibliography creation and markup for authors: A step towards interoperable Digital Libraries (*extended version*)

Bettina Berendt, Kai Dingel, and Christoph Hanser

Institute of Information Systems, Humboldt University Berlin,  
D-10178 Berlin, Germany,  
<http://www.wiwi.hu-berlin.de/~berendt>

**Abstract.** The move towards integrated international Digital Libraries offers the opportunity of creating comprehensive data on citation networks. These data are not only invaluable pointers to related research, but also the basis for a evaluations such as impact factors, and the foundation of smart search engines. However, creating correct citation-network data remains a hard problem, and data are often incomplete and noisy. The only viable solution appear to be systems that help authors create correct, complete, and annotated bibliographies, thus enabling autonomous citation indexing to create correct and complete citation networks. In this paper, we describe a general system architecture and two concrete components for supporting authors in this task. The system takes the author from literature search through domain-model creation and bibliography construction, to the semantic markup of bibliographic metadata. The system rests on a modular and extensible architecture: VBA Macros that integrate seamlessly into the user's familiar working environment, the use of existing databases and information-retrieval tools, and a Web Service layer that connects them. We close with an outlook on two possible futures: one in which ubiquitous URNs/DOIs will solve the document-identification problem and thus ensure correct citation-network data, and one in which Semantic Web technology will allow for heterogeneous and flexible scientific communities and complex similarity and citation relation between documents.

**Keywords:** User interfaces for Digital Libraries, Collection building, management and integration, System architectures, integration and interoperability.<sup>1</sup>

## 1 Introduction

A digital library is more than just a collection of documents: It is an interconnected collection. The citation networks emerging from documents referencing

---

<sup>1</sup> This is an extended version of the paper with the same name to appear in *Proceedings of ECDL 2006*. URL: [www.wiwi.hu-berlin.de/~berendt/DL/](http://www.wiwi.hu-berlin.de/~berendt/DL/)

one another bear a significant part of science’s semantics. This well-known observation makes the integration of different digital libraries one of the key tasks for the future of DL – on a European and ultimately on a worldwide scale.

Citation data are invaluable pointers to related research, citation networks give invaluable cues to the structure and development of science and the flow of ideas through human knowledge. Last but not least, citation data are for ranking journals, authors, and Web resources.

The importance of these networks is reflected in a large number of commercial and in particular non-commercial services that archive bibliographic metadata including its links to authors and institutions (e.g., [www.informatik.uni-trier.de/~ley/db/](http://www.informatik.uni-trier.de/~ley/db/) (DBLP), [repec.org](http://repec.org)), bibliographic metadata including citation metadata (e.g., [scientific.thomson.com/products/{sci|ssci}](http://scientific.thomson.com/products/{sci|ssci}), [portal.acm.org/guide.cfm](http://portal.acm.org/guide.cfm)), citation metadata and links to third-party metadata and/or full text ([scholar.google.com](http://scholar.google.com)), full text and bibliographic metadata (e.g., [www.arxiv.org](http://www.arxiv.org)), or full text, bibliographic metadata, and citation metadata (e.g., [citeseer.ist.psu.edu](http://citeseer.ist.psu.edu), [www.citebase.org](http://www.citebase.org), [www.slac.stanford.edu/spires/hep/](http://www.slac.stanford.edu/spires/hep/), [portal.acm.org/dl.cfm](http://portal.acm.org/dl.cfm)). These services are heavily interlinked, and some of them use the other’s code, thus extending their core functionality; for example, arXiv links to citation metadata in SPIRES-HP or Citebase, RePEc has full texts of some of its content and uses the Citeseer code for citation indexing.

The creation of citation-network data involves the identification of the publications represented in documents and of the links between them. However, this seemingly trivial task has remained a hard problem. Manual indexing is still performed in some services (e.g., DBLP, SCI/SSCI), but is too costly to maintain up-to-date comprehensive information, and it is subject to human error. Autonomous citation parsing and indexing ([13] and the methods used by the other services named above) is limited by the visibility of documents, the heterogeneity of citation styles, and the recognition rates of parsing algorithms, and therefore create data that may be incomplete and noisy.

Errors in citation networks propagate and may have serious consequences on citation indexes, authors reputation, and the memory of science as a whole [7]. It is therefore increasingly being recognised that within the complex process of scientific reading and writing, there is exactly one point where correct and high-quality (meta)data can and must be created: When the author of a scientific work writes a document’s reference list.

Authors can do this in one of three ways: (i) by writing plain text (MS Word or, in Latex, `bibitem` entries) but attempting to cite in a structured way according to conventions like Harvard or APA, (ii) by supplying structured metadata (e.g., in BibTex or EndNote format, or by using templates such as <http://edoc.hu-berlin.de/diml/>), or (iii) by referring to a persistent, globally unique identifier such as a URN or DOI. The first route is the one taken by most authors and the one exploited by autonomous citation indexing; the error rates show that it is not sufficient. Our prior research [3, 1] has shown that the second route is generally, like other elements of structured writing (e.g., assigning standardised keywords and performing standardised markup), little known,

unpopular, and/or, with the existing tools, performed inconsistently, resulting in error levels like those of the first route. The third route is, at least currently, an elusive goal; most documents either have no URI or appear in various instances (cf. the “groups” of occurrences shown in Google Scholar). In this situation, repository-dependent identifiers which we term *virtual document URNs* are the best proxy, but they do not guarantee persistence and can therefore not replace the standard bibliographic metadata.

Therefore, our approach is to support the second route but (a) to make it look and feel like the first route, (b) employ machine intelligence and interactivity to improve quality, and (c) to motivate authors by showing them *in the same environment* what they can gain from citation metadata. The goal is to create a positive-feedback loop in which more and more correct metadata are created and used for scholarly progress. Our contribution lies in this integrated approach to combining existing methods (bibliometric analyses, information extraction, and interface design) – to the best of our knowledge, no other, comparable tool or service exists that is modelled on the whole process of scientific writing, that accompanies authors in their standard environment, and that supports them in their learning about science and authoring.

The second contribution of this paper is a discussion of two possible futures for citation metadata: We discuss why the standardisation of bibliographic metadata is only a first step in the creation of correct citation networks, and we propose two solution paths. Specifically, multiple *documents* (the physical entity which contains recorded information: a book, a file on the author’s home-page, an article in a printed or online journal, ...) may effectively represent the same *publication* (a work issued to the public in the form of a document). These document instances must be integrated to create correct citation networks.

The paper is structured as follows: Section 2 contains a brief overview of related research. Section 3 gives an overview of system requirements and architecture, and then describes the two tools for bibliography creation and markup. Section 4 sketches and discusses two possible solution paths for assigning documents to publications (a “DOI path” and a “Semantic Web path”). Section 5 concludes with an outlook on future research.

## 2 Related Research

Although there is a broad consensus among librarians what **metadata** are needed to identify a publication, different scientific disciplines regard different entries as essential. An example is the issue number of journals – this is not required in the APA citation guidelines as long as page numbers within a volume are consecutive, but it is required in the Harvard citation style. Thus, economists, but not psychologists, might consider the issue number an essential item of interoperable bibliographic metadata. Thus, digital metadata standards such as OAI or MARC generally keep field requirements to a minimum in order to ensure compatibility. In particular, citation-network data are not required elements.

However, metadata standards can be extended by such information. For example, the CiteSeer metadata (<http://citeseer.ist.psu.edu/oai.html>) contain a tag `<oai_citeseer:relation type="References">` that states that the current document cites another document. Both the citing and the cited documents are identified by their CiteSeer-internal URI, an OAI identifier, which we will call *virtual publication URN*. In general, many repositories will provide URNs that are not publication URNs, but virtual publication URNs: First, because an OAI identifier is generally a URN of the metadata and not of the contents behind it, such that different repositories may index the same publication. Second, because even within one repository, parsing errors etc. may lead to the indexing of two documents as distinct that fact represent the same publication (see Section 4).

When cited publications have no URNs, the information in the bibliography section of documents may be marked up by field, such that the cited publication is identified by a virtual ID (“the fifteenth reference of document X”) and its “author”, “title”, “year” etc. constituents. An example of *standardised bibliographic metadata* is the `<BIBLIOGRAPHY>` section of the Dissertation Markup Language DiML (<http://edoc.hu-berlin.de/diml/>). It contains the usual fields, all of which are optional to allow for a wide variety of citation styles. The markup of fields can be done in WYSIWYG fashion in the document itself, or it can be generated from tools supporting structured bibliography maintenance.

**Structured bibliography maintenance** is supported by many tools including Endnote (for MS Word users) and Bibtex (for Latex users), but is often not used or used incorrectly. Controlled vocabularies like the ACM classification or metadata standards like Dublin Core are not an integral part of writing. Individualised annotation of bibliography database entries with personalised fields is possible, e.g., in Bibtex, but that information is not processed further.

In our research on authors and documents in an institutional repository [3, 1], we found that metadata generation is generally done in WYSIWYG fashion (marking individual elements with a mouse and assigning an MS Word template), is done as a post-processing step, is perceived as an annoying add-on to writing and may even deter authors from contributing. Also, many errors remain in spite of the careful postprocessing of documents by the repository’s staff.

Network citation data are invaluable pointers to related research and the basis for a rich set of insights into the structure and evolution of domains like science or patents [8]. **Citation analysis**, in particular co-citation analysis [15], serves to analyze and visualize generic vs. specialised authors and topics, “specialty narratives”, the changing “frontiers of science”, and changes in scientific paradigms, see [8] for a recent overview.

Citation analysis is also the basis for a wide range of evaluations. For example, linkage is interpreted as a (one-dimensional) measure of the “importance” of a journal (impact factor [11]), Web page / Web site (PageRank [6]), or author. Also, citations are increasingly recognised as bearing rich information beyond a mere ranking. For example, citation context in scientific documents and anchor texts of Web hyperlinks have been shown to be good sources for classifying the citation targets (better than those targets’ texts themselves) [5, 19].

**Citation parsing** is an essential component of automatic citation analysis [13]. Citation parsing operates on ASCII data extracted from documents (often, PDFs), and it searches for the usual bibliographic metadata like author, title, etc. It is therefore an example of information extraction: information retrieval from unstructured or semi-structured documents into a known structure of fields. Two main approaches can be distinguished (see [20] for a recent survey). The first uses machine-learning to infer the sequence(s) of fields and represents them, for example, as a Markov chain. The second uses wrappers, regular-expression templates that rely on the fact (or requirement) that bibliography entries follow a limited and well-known set of conventions as defined in, for example, the Harvard, Chicago, or APA styles. While machine-learning approaches have been shown to be more accurate, they need large, hand-labelled training sets, which is a problem in an institutional repository with many very different citation styles, even within one discipline. In contrast, the template-based approach can analyse a repository without a learning phase, and it scales much better to large repositories. This approach is therefore used in major autonomous citation indexing repositories like CiteSeer or Citebase.

There are many **tools** that support the use of citation networks. For example, archives like CiteSeer, Citebase, or DBLP provide topical search for scientific publications, and Google Scholar ([scholar.google.com](http://scholar.google.com)) employs Google's search technology to determine whether a Web page is likely to be a scientific article or not. In addition, many of these tools offer a wide range of citation-analysis measures such as active bibliographies, number of citations, co-citation, and various similarity measures. Visualisations show co-citations and other connections between individual documents (e.g., [sourceforge.net/projects/dbl-browser/](http://sourceforge.net/projects/dbl-browser/), [www.pmbrowser.info/CiteSeer.html](http://www.pmbrowser.info/CiteSeer.html)). Visualisations of whole citation networks use dimensionality-reduction techniques like principal components analysis or Pathfinder networks [16, 9, 8].

However, to the best of our knowledge, these efforts have not been integrated into an authoring environment that is modelled on the whole process of scientific writing, that accompanies authors in their standard environment, and that supports them in their learning about science and authoring.

### 3 Using and extending citation networks during authoring

The architecture and the tools described in this section help authors create correct metadata in order to enable autonomous-citation-indexing repositories to extract more, and more correct metadata and thus create more correct systems of virtual publication URNs. In addition, authors are encouraged to use existing virtual publication URNs in their bibliographies, to further simplify the task of autonomous citation indexing.

### 3.1 Requirements and system architecture

An intelligent author-support system should support the main elements and processes of scientific writing. Bibliography maintenance and the use and extension (by one's own work) of citation networks are key elements of scientific writing. Other important elements are community-related activities like identifying potential collaborators and discussing with them. (Bibliography maintenance may in turn be an important element of collaboration, as bibliography-exchange portals like [www.bibserv.org](http://www.bibserv.org) [14] or [bibster.semanticweb.org](http://bibster.semanticweb.org) [12] show.) Towards this end, we have created further components in the same system framework that we describe elsewhere [2].

Furthermore, the system should (i) integrate itself seamlessly into the user's everyday working environment, (ii) be modular, easily maintainable, extensible, and updatable, and (iii) provide for the analysis of its usage behaviour as a basis for system improvements.

As our studies (and observations elsewhere) show, the vast majority of authors use Microsoft Word for producing their texts, and most use low-end to medium PCs. To produce a solution that requires minimal installation activities, but gives access to rich and up-to-date functionality, as well as to the huge and distributed literature databases available online, we employ a combination of VBA Macros, Web Services, and Web-independent backbone intelligence. This architecture has the added advantage that the intelligent services reached via Web Services can also be called from non-MS-Word user interfaces (see [2]).

The next two sections describe the two components of our authoring system that (i) show the user the advantages of existing and electronically available citation networks and (ii) support her in her creation of high-quality input for citation networks. Because "reading" usually precedes "writing", we describe the literature search and bibliography creation tool first.

### 3.2 Literature search, domain structuring, bibliography creation

**Motivation.** Many tools support searching and also employ bibliometric methods to provide additional information (see Section 2). However, while these tools support reception of information and interaction with structured data, they do not support the active construction of a personalised domain structure.

**Functionality and user interface.** The bibliography-creation component supports bibliography creation with domain structuring. The input is a search term. The output is compiled in three stages: First, a bibliographic database is searched, the matching items are returned, and they are clustered. Second, the user is encouraged to label the clusters, and to modify the automatically-derived grouping to both reflect and develop his perception of the scientific domain in terms of a topic structure. Third, he can include the results in personal documents and also make them available also to others. Availability should ideally be global, so the Web was chosen as publishing medium. XML was chosen as the representation format to retain as much semantic structure as possible.

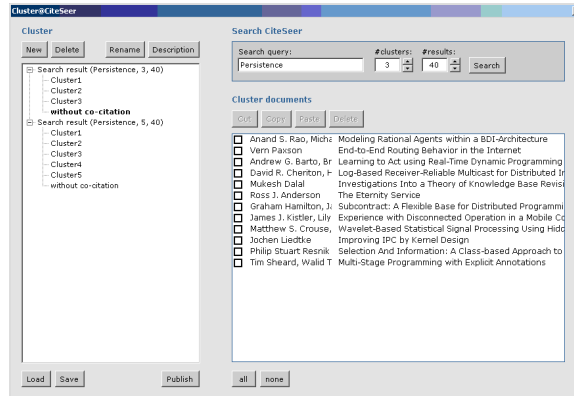


Fig. 1. Bibliography/domain structure creation interface.

Results can be saved and re-loaded for further processing, and hyperlinks in the macro output enable the user to directly retrieve the full text of a document from CiteSeer. Example screenshots are shown in Figs. 1 and 2.

**Data sources and computational intelligence.** We use the CiteSeer database because of its broad coverage<sup>2</sup> and rich structure, and also because it offers an OAI interface.

CiteSeer offers a localised co-citation search that starts from a given document and returns those documents that are co-cited with it. Our tool extends this by a more global view (the context of the immediately-relevant documents), a context-aware similarity measure (the Jaccard coefficient rather than the absolute number of co-citations), and the support of domain-model construction.

We focus on co-citation as a basis for similarity because this has been found to be an excellent indicator of publication similarity and also of global changes in an academic field [15].

**Implementation notes.** A VBA macro interacts with a php Web service, which accesses further information sources.

Processing has four stages: (1) The search term is transformed into an HTTP request to CiteSeer. This guarantees access to the current database, and it returns an HTML page from which the bibliography IDs (`oai:CiteSeerPSU`) can be extracted. This is done by a wrapper that extracts the ID from the hyperlink `http://citeseer.ist.psu.edu/correct/ID`, which is associated with the hyper-text anchor “(Correct)”. The output of this step is a list of document IDs  $D$  that are relevant to the search term.

(2) For each document  $d \in D$ , the list of IDs all documents from the CiteSeer database ( $DB \supseteq D$ ) that cite  $d$  is compiled. The information is obtained by a search in the OAI metadata offered by CiteSeer (tag `<oai_CiteSeer:relation`

<sup>2</sup> Sources for fields other than computer science should be explored in future work.

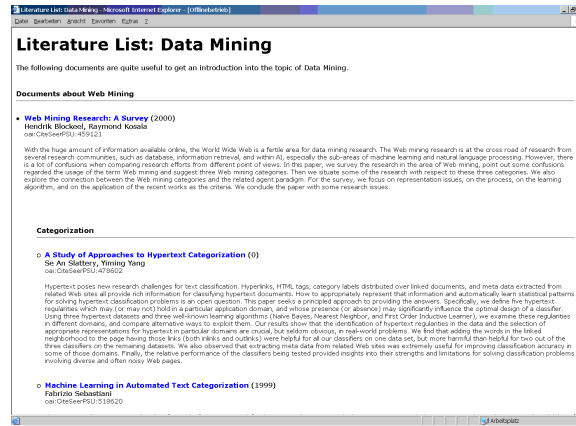


Fig. 2. Domain structure publishing output.

`type="References">`, see Section 2).<sup>3</sup> This information creates the (transposed) citation matrix that is used for clustering documents in step 4.

(3) Bibliographic metadata for result presentation (author, title, etc.) are retrieved via CiteSeer’s OAI interface. This guarantees current data, and it limits the number of requests to the necessary minimum. These metadata include the CiteSeer virtual publication URN and the CiteSeer URL. The latter may be regarded as a proxy of the virtual publication URN, and it is much more obviously useful for the user to record in her private bibliography. In this way, this component supports the re-use of existing citation-network data.

(4) The documents  $d_i \in D$  are clustered using the toolkit CLUTO ([www.cs.umu.edu/~karypis/cluto](http://www.cs.umu.edu/~karypis/cluto)). We employ hierarchical single-linkage clustering [18] and use the Jaccard coefficient as similarity measure. This similarity measure has the advantage that non-citing documents cannot induce similarity. The Jaccard coefficient has first been used in co-citation analysis by [17]. The number of clusters is set to the minimum of (a) the number desired by the user and (b) the number of documents minus 1 (to allow at least one two-element cluster). If present, isolated documents [18] are first put into an additional cluster called “without co-citation” to avoid arbitrary assignments and to respect the observation that co-citation clusters do not represent the entire relevant literature that covers a topic [4].

<sup>3</sup> This database query represents a typical usage of a harvested and then queried local OAI mirror, and we plan to investigate the possibility of creating this kind of mirror in the future. At present, we employ a smaller local database built from the metadata dump offered for download by CiteSeer, and projected onto the “references” relation. This may lead to incomplete data, but informal experiments have shown that the effect is negligible.

### 3.3 Structured writing: High-quality bibliography markup

**Motivation.** Authors (even authors with extensive experience in writing reference lists) make errors, cite erroneously, etc. In addition, parsing tools are not perfect. Results on the merits of re-representation for learning (e.g., [10]) suggests that re-representation (here: of a reference-list entry as a structured entity in a VBA form) may help to avoid author errors. At the same time, it can be the basis for a structured, machine-readable, and semantic representation that obviates the need for parsing in later processing stages and thus reduces or eliminates parsing errors.

**Functionality and user interface.** The author can mark the whole reference list with the mouse to receive a series of formatted bibliography entries as proposals. Errors in automatic recognition can easily be spotted and corrected (see Fig. 3). Once the system proposal has been accepted, the macro writes a surface text into the Word document that is formatted according to the chosen citation style (here, Harvard or APA) and a metadata markup that contains the correct field entries. Our system was created for users of a repository that relies on DiML (see Section 2); a generalisation to other markup schemes is straightforward.

Fig. 3. Bibliography markup interface (example reference type).

**Implementation notes.** The VBA macro calls a php Web service, which issues system calls to perl scripts of two programs: CiteSeer and ParaTools.

**Data sources and computational intelligence.** We start information extraction with the CiteSeer code. As an inspection of the CiteSeer Web site shows,

the regular expressions used in this code are very effective at extracting author and year information and fairly effective at extracting title information. However, at present the CiteSeer system does not extract further bibliographic information, probably because this information suffices for the tasks at hand: author/year/title are used to build the citation matrix, and the full text is used for keyword search. On the CiteSeer Web site, the extracted bibliographic information is shown as a sparse Bibtex entry; it is up to the community to add more information manually.

We then attempt to fill missing slots by using ParaTools (<http://paracite.eprints.org>), which are the basis of the software behind Citebase. In the ParaTools, several templates are offered that describe possible formats of a citation (e.g., *lastname, initial. (year). title ...*, but also *firstname lastname: title ... year*). Template elements are described as regular expressions (e.g., a *year* consists of 4 digits). Templates have two weights: reliability (e.g., a URL is more standardised than an author name) and concreteness (e.g., a template containing the constant “in press” is more concrete than one in which the year still has to be resolved). The system identifies all templates matching a bibliography entry and chooses the one with the highest weight. The template library can be extended.

### 3.4 Evaluation

To test bibliography parsing, we evaluated the results of parsing 82+90 references that were randomly sampled from documents on [edoc.hu-berlin.de](http://edoc.hu-berlin.de) (as representative of institutional repositories with careful document choice, post-processing, and manually generated metadata) and from a random selection of computer science articles gathered via the CiteSeer Web site (as representative of well-known Web-crawling repositories with autonomous citation indexing). CiteSeer recognized 89% of the references’ author/year/title information correctly in the non-EDOC sample, and 61% in the EDOC sample (since author recognition is partly lexicon-based, German names presented a problem). In addition, for literature published by an organisation, a virtual person was created as author, e.g. “Asian Development Bank” was transformed into “A. Bank”. The ParaTools themselves discovered a number of isolated elements correctly (e.g., the author of entry 1, and the title of entry 2), but only seldom parsed a whole reference correctly (10% of our sample). When executed after the CiteSeer parse, however, further bibliographic information was correctly identified by ParaTools for 17% of the sample. It should be stressed that this is likely to be a lower bound since we restricted our tests to the template database shipped with ParaTools. For further tests, we attempt to extend the template database by typical citation styles found among our users. On the other hand, it can also be considered an upper bound, since the analyzed documents were already final versions.

In the design of clustering methodology, we relied on the earlier work (see Section 3.2) that demonstrated the adequacy of co-citation for feature selection and the Jaccard coefficient as similarity measure. In principle, a number of objective quality measures of intra- and inter-cluster similarity could be measured, and the results compared against other settings (average linkage instead of single

linkage, partitional clustering instead of hierarchical clustering, etc.). Based on preliminary user tests, however, we believe that subjective measures of utility and of “correctness” are much more adequate for evaluating this component.

Ultimately, success will rest on the users’ satisfaction with the system. This requires large-scale user testing that we have recently begun.

## 4 From standardised metadata to citation networks

The standardisation of bibliographic metadata is but a first step in the creation of correct citation networks: It is still possible (and likely) that multiple document-instances of the same publication exist on the Web, and that different citations of this publication are assigned to different documents. Two principal solutions appear possible: a centralised and a decentralised approach.

The centralised approach may also be termed the “DOI solution path”. Assuming that in the foreseeable future, each publication will have a DOI (or some other, persistent and universally agreed-upon URN), and that each document-instance of a publication will be identified with this URN, one could extend the bibliography-markup tool by a subsequent call to a Web service that in return to the user-corrected metadata searches for the publication URN. If no result can be found, this is an indication of a remaining error in the form entries (e.g., a typo in the author’s name which cannot be detected by the parsing algorithms), and the user is asked again to correct the fields.

Components for this solution path already exist: Web-based forms for retrieving the DOI given bibliographic metadata ([www.crossref.org/guestquery/](http://www.crossref.org/guestquery/)) and for retrieving the bibliographic metadata for a given DOI ([dx.doi.org](http://dx.doi.org)). However, there seem to be some problems in coverage (informal tests yielded empty results for several searches for existing, DOI-annotated publications). Also, the returned metadata are not in a standard format and therefore must be processed again by the user in order to integrate them into his document. These problems could easily be solved, as the Bibtex or Endnote exports of a number of databases (e.g., the ACM Digital Library) show. To support program access, Web-based forms should and will be supplanted by Web services. Thus, for example, CrossRef has announced that it will offer a Web Service for DOI search (see [www.crossref.org/01company/pr/press113005.htm](http://www.crossref.org/01company/pr/press113005.htm)).

However, the current development of the Web and also of scientific publishing suggests a different possible future, which may be termed the “Semantic Web solution path”. Centralised and cost-inducing identification schemes like DOI will find it difficult to attain complete coverage in a world in which a sizeable amount of informal (“grey”) literature gets created and is cited. Also, documents that are “the same” to different extents (instances or versions: multiple copies of the same file at different Web addresses, or different stages of the same publication, like conference paper, preprint, journal paper) will continue to co-exist and be referenced in different ways by different authors. Thus, identifiers will in general be unique within their respective contexts (the Semantic Web itself is based on the notion of URIs!) and may become persistent and universal iden-

tifiers of *a document in a special context*. However, they will in general not be globally unique and thus not identify *a document*. Therefore, they cannot be automatically retrieved given a set of bibliographic metadata. If retrieval is not fully automatic, it is unlikely that authors will employ URNs in a consistent and complete way.

Authors and readers will continue to congregate around archives like CiteSeer, Citebase or arXiv ([www.arXiv.org](http://www.arXiv.org)), and within these “communities”, virtual publication URNs will continue to exist and make sense. Many of these archives are already interlinked today (e.g., pointers from CiteSeer to DBLP and the ACM Digital Library). This linkage is a mapping from one system’s virtual publication URN to another system’s virtual publication URN. Google Scholar displays the multiplicity of instances as “groups” of occurrences of one document. In the future, the semantic equality that is implicit in these links (“see *the same* article in DBLP”) should be augmented by a graded notion of similarity that can also reflect uncertainty (“this may be the same publication, or it may be a different one”), versioning (“this is likely to be a different version of the same publication”), or other relations. The set of documents that are “similar enough” to one another could then constitute a node in an aggregate citation network. Systems that encourage scientists to share their metadata, such that different metadata sets co-exist for what appears to be the same publication, may be a first step in this direction (e.g., [www.bibserv.org](http://www.bibserv.org)).

## 5 Conclusions and outlook

In this paper, we have proposed a general system architecture and two concrete components for helping authors contribute to a “better web of science”. The aim is to demonstrate to authors the merits of rich and correct citation-network data, and to enable them to provide correct bibliographic metadata in their own documents that will extend the existing citation networks.

In future work, we plan to extend the system’s functionality. Text mining could be used to identify further sources of publication similarity such as closeness in the document or similarity of context [13, 17], and for proposing cluster labels [4]. Also, machine learning will be needed to establish the complex mappings between documents, and virtual publication URNs, that have been sketched in the “Semantic Web solution” to the problem of the n:1 relation between documents and publications.

Another limitation of the current system is that it does not learn from user interactions. In future work, the system will be personalised to adapt to the user’s citation styles, modes of searching, and modes of grouping.

*Acknowledgements.* We thank Derya Saki and Mert Sengüner for creating the bibliography markup macro, Daniel Trümper for providing the ParaTools interface, Sebastian Kolbe for help in setting up the system, and Lee Giles and Isaac Council for providing us with the CiteSeer code and many answers to our questions.

## References

1. Berendt, B. (2005). Understanding and Supporting Volunteer Contributors: The Case of Metadata and Document Servers. In *Knowledge Collection from Volunteer Contributors. Papers from the AAAI 2005 Symposium* (pp. 106-109). Technical Report SS-05-03. Menlo Park, CA: AAAI Press. <http://www.wiwi.hu-berlin.de/~berendt/Papers/SS505BerendtB.pdf> [access date of all references: 11 March 2006]
2. Berendt, B., Brekenfeld, H., Dingel, K., Hanser, Ch., Krause, B., & Lohde, T. (submitted). IR-THESIS: An intelligent authoring environment for learning scientific writing.
3. Berendt, B., Brenstein, E., Li, Y., & Wendland, B. (2003). Marketing for participation: How can Electronic Dissertation Services win authors? In *Proceedings of ETD 2003*. <http://edoc.hu-berlin.de/etd2003/berendt-bettina/>
4. Braam, R.R., Moed, H.F. & van Raan, A.F.J. (1991). Mapping of Science by Combined Co-Citation and Word Analysis. (I & II) *J. of the American Soc. for Inform. Science*, 42(4), 233–266.
5. Bradshaw, S. (2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *Proc. of the 7th ECDL*.
6. Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *WWW7 / Computer Networks*, 30, 107–117.
7. Cardona, M., & Marx, W. (2004). Verwechselt, vergessen, wiedergefunden. Referenzen – das fehlerhafte Gedächtnis der Wissenschaft(ler). *Physik Journal*, 3 (11), 27–29. <http://www.pro-physik.de/Phy/pdfs/ISSART21255DE.PDF>
8. Chen, C. (2003). *Mapping Scientific Frontiers*. London: Springer.
9. Chen, C., & Carr, L. (1999). Visualizing the Evolution of a Subject Domain: A Case Study. *IEEE Visualization 1999* (pp. 449–452).
10. Cox, R. (1996). *Analytical reasoning with multiple external representations*. PhD thesis, Department of Artificial Intelligence, University of Edinburgh.
11. Garfield, E. (2004). *Essays / Papers on Impact Factor*. <http://www.garfield.library.upenn.edu/impactfactor.html>
12. Haase, P. et al. (2005). Bibster – A Semantics-Based Bibliographic Peer-to-Peer System *Journal of Web Semantics*, 2(1). <http://www.websemanticsjournal.org/ps/pub/2005-8>
13. Lawrence, S., Giles, C.L. & Bollacker, K.D. (1999). Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, 32, 67–71.
14. Richardson, M., & Domingos, P. (2003). Building Large Knowledge Bases by Mass Collaboration. In *Proc. of the 2nd Int. Conf. on Knowledge Capture* (pp. 129-137).
15. Small, H. (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *J. of the American Soc. for Inform. Science*, 24(4), 265–270.
16. Small, H. (1994). A SCI-MAP case study: building a map of AIDS research. *Scientometrics*, 30, 229–241.
17. Small, H. & Greenlee, E. (1980). Citation Context Analysis of a Co-citation Cluster: Recombinant-DNA. *Scientometrics*, 2(4), 277–301.
18. Small, H. & Griffith, B.C. (1974). The Structure of Scientific Literatures, I: Identifying and Graphing Specialities. *Science Studies*, 4(1), 17–40.
19. Utard, H., & Fürnkranz, J. (in press). Link-Local Features for Hypertext Classification. To appear in M. Ackermann, B. Berendt, M. Grobelnik, A. Hotho, D. Mladenic, G. Semeraro, M. Spiliopoulou, G. Stumme, v. Svatek, & M. van Someren (Eds.). *Semantics, Web, and Mining - extended and revised papers from the EWMF and KDO workshops at ECML/PKDD 2005*. Springer, LNCS/LNAI.
20. Yong, K.N. (2005). *Citation parsing using maximum entropy and repairs*. Honours Year Project Report, Department of Computer Science, National University of Singapore. <http://wing.comp.nus.edu.sg/publications/theses/yongKiatNgThesis.pdf>