From Digital Libraries to RIAL Electronic Publishing

Dror G. Feitelson
School of Computer Science and Engineering
The Hebrew University of Jerusalem
91904 Jerusalem, Israel

Abstract

The linked nature of the World-Wide-Web allows bibliographical repositories to contain links providing access to full text. This leads to the creation of topical digital libraries that index a certain field, without concern for archival storage. Such digital libraries serve focused Internet communities as portals into the relevant scientific knowledge. Archival storage may be provided by a small number of worldwide repositories, which are accessed via the indexing services provided by the libraries. In the long run, this can lead to RIAL electronic publishing, in which literature is deposited directly in the archive, and Rated, Indexed, Annotated, and Linked by users and libraries. This blurs the traditional distinction between officially published and yet unpublished works, and creates a single integrated environment in which all works are instantly accessible. This development is being lead by Computer Science, perhaps because this field trails others in the timeliness of its official print journals.

1 Introduction

In the last decade a great change has occurred in the working environment of most scientists. The Internet, which was previously used mainly for e-mail, now provides the infrastructure for the world-wide-web, with its abilities for on-line publishing and global search. This has started to have an impact on how scientific knowledge is distributed, with many papers now being available from personal home pages even before their official publication in a scientific journal. In fact, this is so pervasive that it is now possible to create digital libraries that rely exclusively on such "personal" publication, and out-perform those that depend on traditional publishing venues [20].

Publishers and libraries have naturally not remained indifferent to these developments. Practically all publishers use the Internet to streamline the publication process, as well as for the distribution of their products at a single-paper granularity (as opposed to the volume-based structure of traditional print products). Numerous digital library projects have been initiated, some on a grand scale and with high-volume funding (e.g. [7]). But, as we show in the next section, scientific publishing has not changed much. The Internet is being used in a

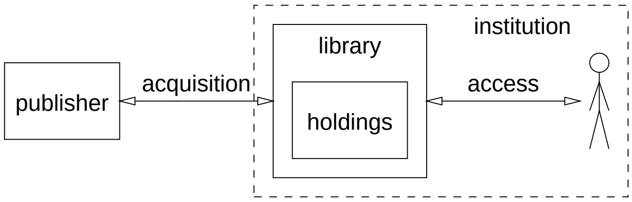

Figure 1: *The traditional model of access to scientific literature.*

rather limited way to improve communication between the parties involved in the publishing process, without changing the process itself.

While the technological progress involved in digital libraries is impressive, it barely starts to exploit the potential of connectivity brought about by the Internet. Still missing is the integration of the different services into one cohesive whole, which provides scientists with a comprehensive solution to their information gathering and dissemination needs. This includes not only the publication of papers and keyword-based search, but also citation indexing, collaborative ranking, and long-term archival storage. Taken together, the integration of these services can lead to a real paradigm shift in scientific publishing.

2 Scientific Publishing Today

Scientific publishing is carried out largely through the interaction of two types of players: publishers and libraries.

The role of publishers is to create new literature, focusing on the technical and business aspects as opposed to the creative ones. This includes overseeing the peer-review process, printing of the accepted papers, production of journals, and distribution to clients, many of which are libraries. The role of libraries is to make the published literature available to their patrons. Traditionally this has been based on two inter-dependent undertakings. First, the library needs to maintain a large collection of literature on-site, so as to have it on-hand when requested. Second, the library needs to provide an indexing and cataloging service, to enable the available literature to be found. However, this has typically been limited to books. Patrons are expected to identify desired scientific papers on their own, probably by tracking citations. The library is responsible only for the delivery of the requested ones (Fig. 1).

The use of electronic technology has typically stayed within this framework. Publishers use it to streamline the review process, providing tools for communication between editors and referees, for on-line status checking, and for delivery of the finished product. Many also acknowledge the ubiquity of desktop publishing, and accept manuscripts in electronic form rather than re-typing them. Libraries provide on-line access to their computerized catalogs, and subscribe to electronic journals (whose articles can be delivered individually to the desktop) on behalf of their patrons.

An important departure from common practice is the creation of internal digital libraries by the publishers. Thus institutional libraries are tempted to purchase access rights to all of a publisher's products, rather than to a selected subset. Together with access to the papers

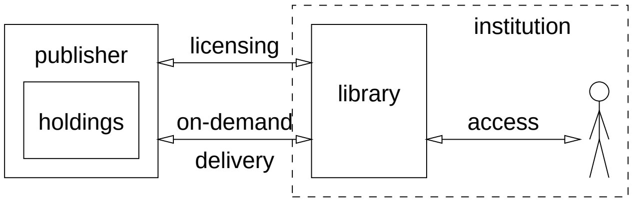

Figure 2: *On-demand access to articles provided electronically by the publisher.*

themselves, the publisher often provides a keyword-based notification service that alerts customers to the publication of new works in which they may be interested, and a keyword-based search facility for the identification of articles that have already been published in the past. Thus the publishers take upon themselves the provision of an index to scientific papers, similar to the index that libraries maintain for books. With such services, the actual libraries in the client institutions may no longer have local holdings — instead, they serve as a channel to access the digital library services provided by various publishers. This rather dramatic shift is the subject of the next section.

An outstanding problem is that of commitment for archival access, e.g. when a commercial publisher ceases to publish a journal (this is probably less of a concern with publications from non-profit professional organizations). There is a danger that long term storage in the publisher's digital library will be controlled by simple economic considerations: when cost exceeds income it is no longer economical to keep papers on-line. Thus publishers may not share the commitment for archival storage that libraries would expect. This may ultimately lead to the establishment of separate archiving services, as described below.

3 The Shifting Role of Libraries

A downloaded copy of a scientific paper printed on a laser printer and stapled together is just as good as an original print copy. And it is typically available world-wide from a single server operated by the publisher. Hence there is *no need for libraries to maintain a large in-house collection*, at least not for scientific journals.

The responsibility of the institutional library therefore shifts to securing access to the servers operated by various publishers [6]. The goal remains the same — reduce the pains experienced by patrons as they try to access desired literature. But the means change: instead of trying to acquire the literature in advance and storing it locally, libraries must enter into licensing deals that allow for direct delivery upon request. The transactions involved in the ordering, delivery, and payment for each article will be done automatically behind the scenes, without bothering the client who made the request (Fig. 2).

And what about digital libraries? This term is loosely applied to any establishment that serves information in a digital form. Some institutional libraries are therefore labeled as digital libraries, by virtue of having digitized part of their holdings and making it available over the Internet. However, in the context of scientific literature, institutional libraries typically only serve to set up the relationship with the real digital libraries, which are operated by the

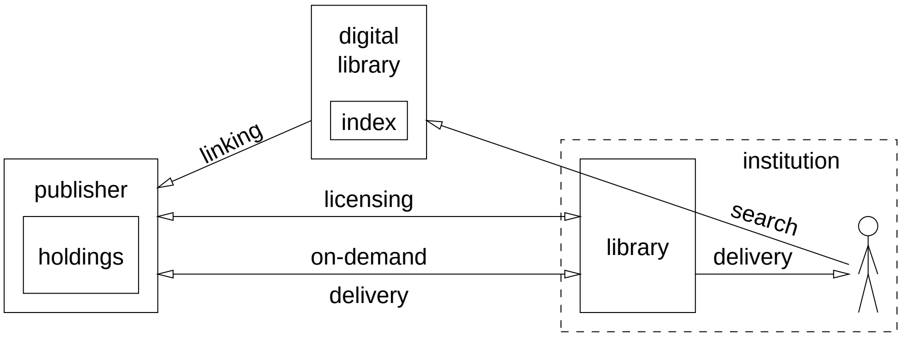

Figure 3: *Independent digital libraries serve as portals to scientific knowledge by indexing and linking to the wares of many publishers.*

publishers. These digital libraries serve the requested documents, and also provide keyword search and notification services.

The most interesting recent development, however, is the creation of independent digital libraries. The defining attribute of these libraries is that they deal with document *surrogates*, rather than with the documents themselves [19]. Their function is to provide *added-value services* such as improved indexing, filtering, and translation, thus allowing clients to find what they seek [4]. They are distinguished from the digital libraries operated by publishers in that they are not constrained to a pre-defined and limited set of sources: any published document, in any language, can be indexed. They are distinguished from traditional indexing and bibliographical services in that they provide an extended service, and also *affect the delivery* of the found articles from external sources (e.g. by using digital object identifiers [29] to access the digital libraries operated by publishers). They are *portals to scientific knowledge* (Fig. 3).

Another important attribute of these new digital libraries is the shift from the physical to the virtual realm. Traditional institutional libraries are grounded in the physical institutions to which they belong, and serve a community whose main unifying attribute is physical proximity. The new digital libraries serve virtual Internet communities, which are based on common interests. Examples include CiteSeer [25], which caters for the whole of computer science, The WEBBIB project [31], which is focused on distributed computing on the web, and the HCI bibliography [14].

Several technologies are being developed to improve the indexing. The first is improved keyword-based classification using various machine-learning approaches such as boosting [27]. A promising alternative is to shift to *semantic* indexing. One form of semantic indexing is based on indexing citations and analyzing the network of inter-citations among articles. This already exists in several digital libraries, based on automatic examination of the full text of on-line documents. Another approach is to engage users in the indexing process, based on new electronic technology for collaborative work, as in the BoW project [8]. This includes a hierarchical concept index reminiscent of a book's table of contents, rather than the conventional flat space of keywords.

Given that indexing is so important, it is natural that some companies try to make money off it. Traditional indexing services such as INSPEC use keywords to index documents and match them with queries. Relying on human expertise to create indexes and unravel connections among documents is typically too expensive for basic science, but exists in other fields such as Law [22].

4 Long-Term Archiving

Professional societies, such as the ACM, consider long-term archiving of their publications to be part of their mission. As noted above, commercial publishers may not share this commitment. In fact, contemporary licensing agreements typically cover only the term of the contract, and do not make any promises regarding future availability. Thus libraries run the risk of one day losing their access to certain parts of the literature they wish to provide to their patrons. This can happen even despite the good will of everyone involved, e.g. if a publisher goes out of business.

So if libraries don't maintain in-house collections, and publishers may not provide their publications indefinitely, who does the long-term archiving? A possible solution is to use a separate archiving service, either commercial or created by consortia representing many libraries. The idea is to store everything in this archive, and use it as a backup for the digital libraries provided by publishers. Thus articles are initially requested (and purchased) directly from the publisher, but if they are unavailable (temporarily or permanently) then they can be retrieved from the archive. The archive collects a fee, and may pass part of it to the original publisher. This scheme is actually an extension of the common practice of using the Copyright Clearance Center to handle copyright payments — the extension being that the center also provides the copies.

Once we postulate the existence of such an archive, we note that it can be used as more than just a backup for publishers. In fact, the archive can turn into a "digital deposit" library, where all sorts of items can be stored: not only journal articles, but also conference papers, technical reports, and unpublished manuscripts. This provides for simplified one-stop shopping, where all literature needs can be met. Thus it has an advantage over projects that specifically target only gray literature [13, 11], which by definition limit themselves to the lesser types of literature.

In this context it should be noted again that we are only referring to services for the provision of access to scientific articles, and in particular new publications that are "born digital". This does not encompass all that has been bundled under the label of digital libraries. Specifically, institutional libraries can have unique digital records that turn them into archival storage sites as well. Examples include digitized medieval manuscripts, political ads, and other unique holdings.

While the concept of a long-term archival facility is simple, it involves many non-trivial technical issues. One obvious issue is reliability, in order to guarantee the persistence of the stored information across power outages, earthquakes, and political upheaval. This can be achieved by redundant storage at multiple sites around the globe. An extreme example is the Eternity Service, which distributes data with enough redundancy to guarantee that it cannot be deleted even if so desired [1].

Another issue that has received a lot of attention is how to keep the information accessible in the face of changing technology [23]. The problem is that both storage media and data

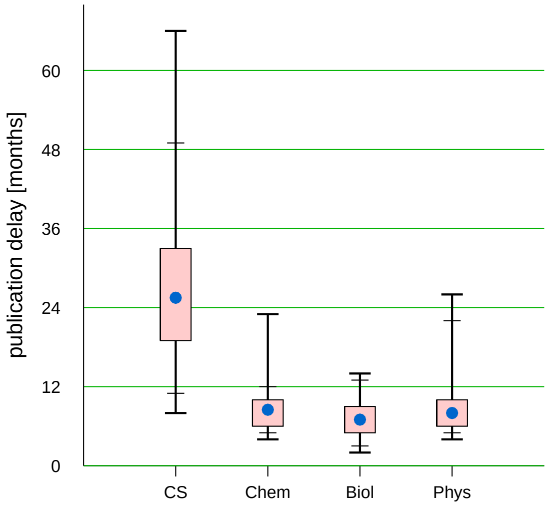

Figure 4: *Publication delays in scientific journals in different fields. [Data sample is the first 10 papers in 2000 from each of 5 random journals in each field. Box indicates 25 and 75 percentiles of distribution, with circle denoting median. Whiskers show minimal and maximal values sampled, with ticks at 5 and 95 percentiles.]*

formats keep changing, making it impossible to access and interpret information that was stored only a relatively short time ago (as opposed to ink on paper, which can be accessed even after thousands of years). Recognizing that the information is more important than the format leads to the use of formats based on standard markup languages, typically SGML or its subset XML [15]. A more extreme approach is to translate all proprietary formats into a simplified standard one, maintaining the content at the possible expense of form [21].

5 Towards RIAL Electronic Publishing

The combination of digital libraries and archival services can provide reasonable access to the scientific literature, based on extensive indexing, and at a reasonable cost. However, we still need to address one last problem that exists in the current scientific publishing system: that of the refereeing process and its associated publication delay.

Peer review is a basic staple of scientific publishing, and is considered the most important factor in the control of quality. Its importance arises from the fact that it is applied a-priori, and prevents sub-standard articles from entering the "official" literature. However, it also has acknowledged limitations, exemplified by the fact that citation counts are preferred as a posteriori measure of impact. Moreover, the peer review process has several drawbacks from a publication standpoint. The first and foremost is the need to find qualified referees, and get them to read and judge the submitted papers. This is time consuming for both the editors and the referees, and may lead to substantial publication delays (the field of

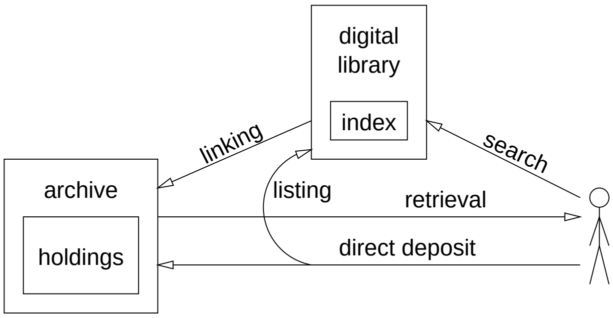

Figure 5: *With RIAL electronic publishing, documents are deposited directly in the archive, and accessed using the index of a digital library.*

Computer Science is especially notorious in this respect, with a median delay of nearly 2 years! see Fig. 4).

With digital libraries and archiving services as described above, the whole peer-review process can be circumvented. Any unpublished manuscript can be stored in the archive, and it can also be listed and indexed in the relevant topical library. In fact, the manuscript is actually published *by being listed in the digital library!* (Fig. 5). In addition to avoiding the overhead of peer review, this has the benefit of avoiding other overheads associated with official publishing, such as conformance to arbitrary formatting requirements.

But what about the quality control? It is true that by allowing un-reviewed manuscripts to be published we ditch the a-priori quality control in favor of immediate accessibility. However, this does not mean that manuscripts will not be subject to quality control once they are part of the digital library. In fact, they may be subjected to a much more stringent form of quality control, in which the review process never ends.

The quality control actually comes in two forms: automatic and user-based. The automatic procedures are largely based on the citations accrued by the article over time. A simple measure of standing is to just count such citations. More sophisticated measures can also be used, such as Kleinberg's measures of authorities and hubs [18].

User based mechanisms include various forms of feedback, like those used in the BoW project (Fig. 6) [8]. The first such measure is again a popularity count: how many times was the article surrogate viewed? And of these times, how many lead to a request to export the bibliographic data or to retrieve the full text? Other measures involve more explicit action by users. For example, users may be provided with simple facilities to express feedback regarding papers which they visit, e.g. by noting a positive or negative evaluation. The amount of detail solicited should be carefully considered, so as not to alienate users and cause them to disregard the feedback mechanism altogether. Alternatively, a simple feedback mechanism can be provided by default, with more detailed facilities available for those who request them.

Considering all of the above leads to the concept of integrated RIAL electronic publishing:

- Articles are **R**anked by users of the digital library, and by automatic means such as citation analysis, as described above. This ranking is used when displaying search results. In particular, articles that receive significant negative feedback will never be

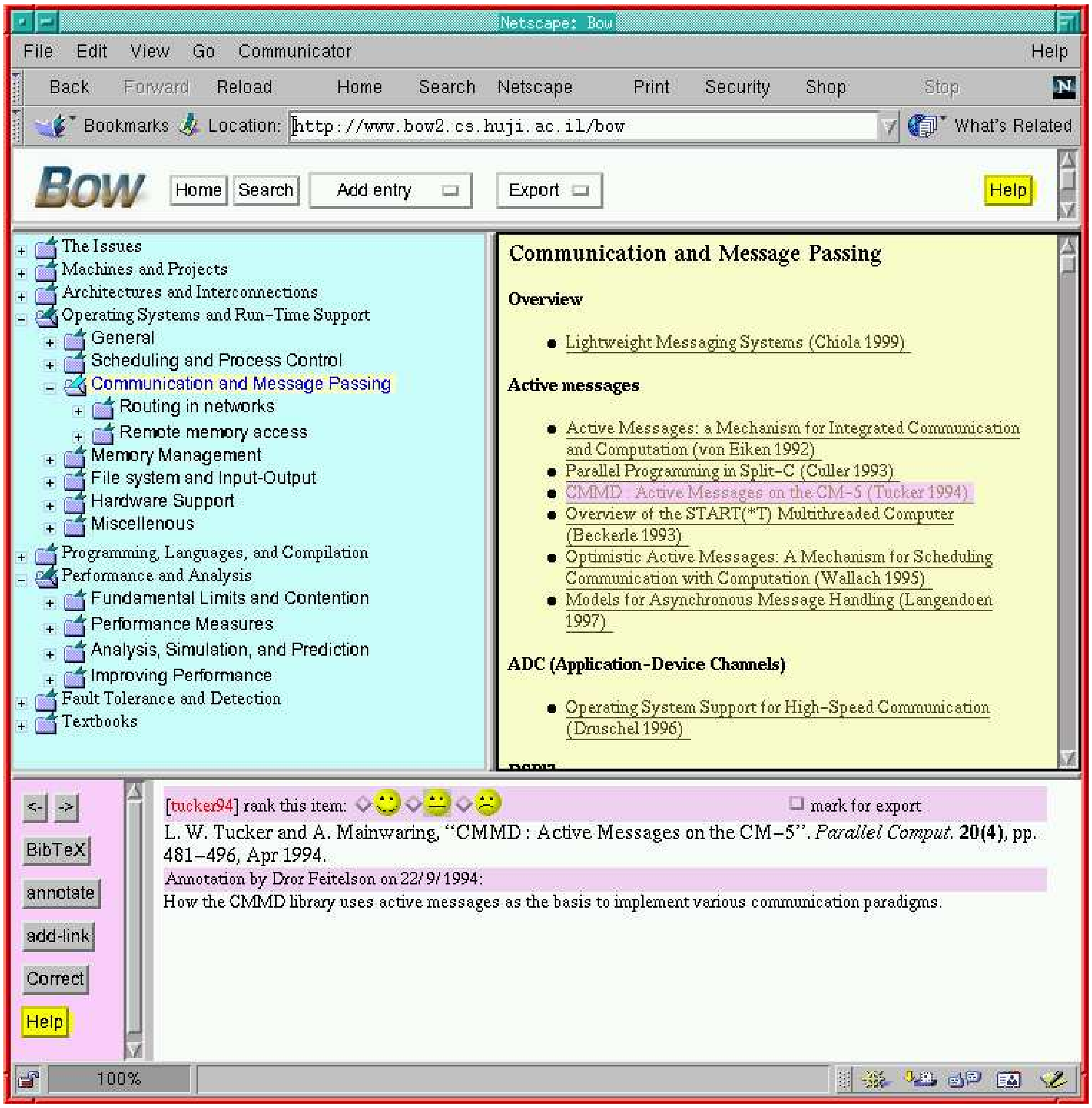

Figure 6: *The BoW user interface is composed of three main frames: the hierarchical concept index (top left), a list of entries in the currently selected topic (top right), and a representation of the currently selected entry (bottom), which includes a set of three radio buttons for ranking and a checkbutton for export in BibTeX format. At bottom left are buttons that invoke operations on this entry, including the addition of links and annotations.*

displayed to the casual browser by default, but will be kept in the background for use by the initiated. The ranking process is continuous and never stops; in particular, it may change the standing of a paper as the field as a whole evolves.

- Articles are **I**ndexed in a hierarchical topical index. This is initially done automatically based on lexicographic analysis of keywords [10]. The results of the automatic indexing are used as the basis for explicit indexing by the users of the digital library. Such explicit indexing provides much needed semantic indexing, based on first-hand understanding of the articles by expert human readers. In particular, it can identify articles that are related to certain concepts despite the use of different vocabulary.

- Articles are **A**nnotated by users, providing for updates, the identification of relationships among articles and concepts, and assessment of impact. This can be viewed as a form of on-line discussion, which continues as long as the article generates interest. Annotations can be displayed anonymously, but the system keeps track of their authors in order to enable editors to resolve disputes.

- Articles are **L**inked to related resources (e.g. the full text, an updated version, or a web site containing software that was used in the reported research). In addition, links across articles are possible. This includes both links among articles that cite each other, and among versions of the same article and followup work. User input may be used to distinguish between important connections and inconsequential ones (e.g. citations to seven other articles just to say that this article uses a different approach).

Naturally, all of this should also be stored for posterity, whether in the digital library itself or in the long-term archive.

Of particular importance is the observation that ranking by users of the digital library can also include formal reviewing, at least for papers that attract reviewers. An assigned editor can judge the submitted reviews, and decide whether they justify conferring the rank of "accepted" on the paper in question. This decision will take the standing of the volunteer reviewers into account, as well as their possible relationship with the paper's authors. Reviews can also be forwarded anonymously to the author, in order to encourage revisions that will improve the paper. The default full text version accessible via the digital library will always be the most up-to-date, with a possible link to a previous "authoritative" version in case of journal papers.

The end result is that the digital libraries will be able to distinguish between three classes of papers:

- "Officially" published in some refereed journal or conference.
- On-line "accepted" based on reviews volunteered by users.
- Other manuscripts that have not passed a quality check.

The ranking of papers can then be calculated by a weighted consideration of their class and their popularity as measured by references from other papers.

To summarize, RIAL electronic publishing is based on digital libraries that act as portals, and archives that provide long-term storage. Features include the following:

- One-stop shopping: a topical digital library should provide access to all the relevant literature in the field. This includes for both official and gray literature, from all publishers and sources.

- Deep indexing is used to classify the literature and to support focused search and browsing.

- Quality indicators help users zero in on the highest quality references available. These indicators include acknowledgement of official publication in conventional journals, citation counting and indexing, and on-line evaluation and discussion provided by other users.

6 Economic and Social Factors

The discussion so far has concentrated on technical issues. But the realization of RIAL electronic publishing will depend even more on economic and social issues.

The advent of the Internet as a distribution medium and the creation of electronic journals has prompted much discussion about the business models and pricing of electronic publishing. The opinions are mixed. While it seems that there is a potential for savings in the production and distribution of electronic literature, this may be more than offset by additional expenditures on infrastructure (e.g. hardware for servers) and the overhead of dealing with a new technology. An important observation is that the price — regardless of the medium being used — must cover the "first-copy cost", which is dominated by the cost of staff and composition [26].

Independent digital libraries, and the possible move towards RIAL electronic publishing, are part of an attempt to provide services for free, thus avoiding exorbitant publishing fees. In the long run completely free publication is unrealistic — at least the storage facilities of the long-term archive must be paid for in some way. But by using electronic copy created by the authors, much of the first-copy cost incurred by publishers is avoided. When authors routinely use some standard markup format for their works, the cost of refreshing the documents as technology changes will also be avoided.

Part of the problem with high journal prices is that academics don't get paid royalties for publishing – they get prestige. Thus, in contradistinction with the common case in which authors and publishers have common interests in the preservation of copyright, authors and publishers are at ends when it comes to publishing journal articles [23]. The reason that high journal prices are tolerated is that publication in such journals (after surviving the peer review process) is widely considered the best metric of academic success. Electronic journals, on the other hand, still do not enjoy the same stature [28]. It stands to reason that the situation with RIAL electronic publishing will be even worse, due to the total lack of a-priori refereeing. However, this approach has great potential. First, ensuring that all papers are uniquely identified by an on-line access code provides for much simplified tabulation of citations. Integrated citation indexing can then be used for very precise impact evaluation at the level of individual papers, rather than relying on general journal impact factors. Second, posteriori refereeing as possible with RIAL electronic publishing may be considered more meaningful than the limited a-priori refereeing common today.

In this context it is interesting to note that in terms of actual dissemination of information, alternatives to journal publication are widely used already (at least in Computer Science) (Fig. 7). One reason is probably the relatively long publication delay of journals in this field (recall Fig. 4). There is therefore reason to hope that efficient venues such as RIAL electronic publishing will be used for pragmatic reasons, and subsequently, will also

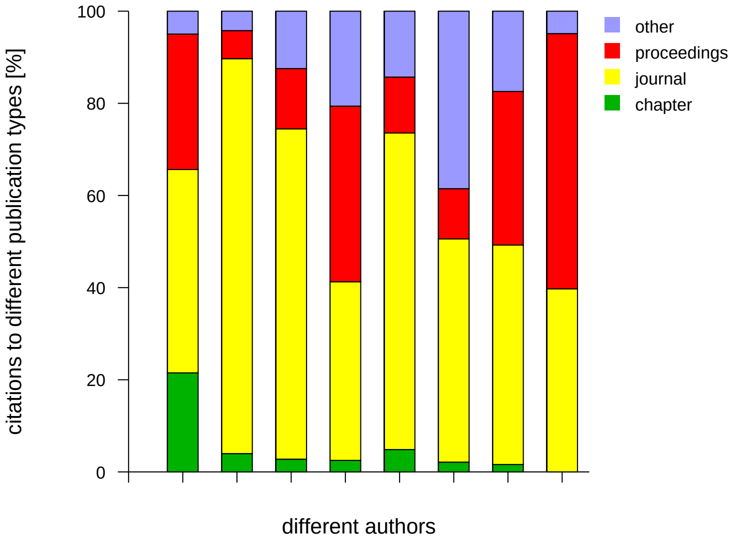

Figure 7: *Distribution of citations to works of different types by the same author indicate acceptance for dissemination of information by conference proceedings and other means, including technical reports, manuals, web pages, and unpublished manuscripts. [Data from CiteSeer. Authors ranked 15, 115, 215, etc. were selected, and citations to the top 100 publications of each one tabulated. Citations to books were excluded.]*

acquire recognition. Acceptance and experience with similar projects, such as the computing research repository [13] and patron-augmented digital libraries [12] will also help.

7 The BoW Repository: A First Step

BoW (Bibliography on the Web) is designed as a first step in the directions described above. The technology is based on a database to store the data, Java and SQL to manipulate it, and XML and XSLT to ship and present it. The user interface is based on conventional web browsers (see Fig. 6), and employs Javascript for certain client-side operations [9].

The structure shown to users is a *deep hierarchical index*. By this we mean that the number of branches is perhaps an order of magnitude smaller than the number of entries in the database, but not smaller; contrast this with conventional classification schemes, which typically employ a fixed set of up to a few dozen topics even when the number of entries grows to tens of thousands. The index therefore supports focused browsing that can lead o serendipitous discoveries of related documents, as occurs at physical library shelves today. Users can also view the bibliographical information and export it. The current version does not automatically provide links to the full text, although users can add such links.

Apart from the hierarchical index, the main distinctive feature of BoW is the support for user input, with the goal that each user will be able to benefit from the experience of others [8]. To begin with, users can add new entries to the repository. The system then automatically indexes the new entry, and forwards the suggested indexing to the user for verification [10]. Users can also add annotations and links to existing entries.

The above features were carefully selected, with the overriding goal of keeping the system simple. Users are expected to be able to use the system with no prior training. No bold new concepts are introduced — only new ways of exploiting Internet technology and the interconnected nature of the web. Everything is geared towards support of users needs in terms of finding relevant high-quality scientific references. User input is either implicit or low-key, as part of normal usage.

Small scale demonstrations of the system have elicited very positive feedback. The true test will be the opening to the public of a repository on the topic of parallel systems. This repository is primed with over 3000 entries collected over more than 10 years, and has an index of about 140 nodes. Thus there is ground for hope that it will be sufficiently interesting for the community of scientists working in this field to overcome the "new technology" barrier. The site is expected to be opened in the first quarter of 2001.

Comparing BoW with related projects, all in Computer Science, leads to the following differences. The ACM [3] and IEEE-CS [16] digital libraries are operated by publishers, and therefore limited in scope according to *source* rather than *topic*. The Hypertext Bibliography Project [17] spans multiple publishers, but still does not support literature from arbitrary sources (including gray). Its use of the net is limited to "live" citation links among papers. The WEBBIB project [31] is similar in this respect. CiteSeer [25] provides rather comprehensive citation indexing and lists of related papers, but is rather poor at providing conventional bibliographical information. CoRR, the Computing Research Repository [2], is perhaps the closest to BoW, but stems from being a repository for gray literature, and is therefore limited to literature that is deposited in it and survives copyright claims by publishers. BoW, on the other hand, starts from indexing and user collaboration. The patron-augmented digital library approach [12] puts emphasis on support for the full cycle of scholarly work, including authoring, whereas BoW is geared towards being just a digital library, allowing users to retain their regular work environment. Finally, the option to add annotations is subsumed by the notions of multivalent documents [24] and semantic networks [5]. Again, in BoW we prefer to make a much smaller departure from the traditional use of literature.

The focus of the BoW project is on the user-collaboration aspects of digital libraries and RIAL electronic publishing. Thus the main developments are in the area of user-interfaces that make it easy to contribute to the system, without annoying and burdening the users. Missing features that need to be added in order to proceed to full support of RIAL electronic publishing include integration with automatic identification of full text on the web, and support for citation indexing both for ranking and for linking related articles. In addition, tools for editorial support and refereeing are needed.

8 Conclusions

RIAL electronic publishing is based on topical digital libraries which become portals to the scientific literature pertaining to the topic in question, and on the creation of long-term archives that serve as curators for this literature. The integration of various technologies into the library portal, including collaborative work and autonomous citation indexing, can lead to a system that is just as valued as the traditional system of scientific journals, and has many advantages in terms of timeliness and flexibility.

Compared with endeavors to hyperlink everything into one big associative web of intelligence (perhaps exemplified by the open hyperdocument system [30]), these ideas may seem low-key and mundane. And indeed, they are largely based on existing technology, and are limited to the realm of scientific literature. But on the other hand they are also rather radical: this is scientific publishing with practically no role for commercial publishers and institutional libraries as we know them today! It remains to be seen whether this vision will live up to its promise. To help it along, we are starting with the infrastructure required for the digital libraries. If users indeed pick up this technology, the path to RIAL electronic publishing is relatively straightforward.

Acknowledgement

This research is supported by the Ministry of Science Culture & Sport under the program to develop scientific and technological infrastructure.

References

[1] R. J. Anderson, "*The eternity service*". In *Pragocrypt '96*, 1996. Available online at URL http://www.cl.cam.ac.uk/users/rja14/eternity/ eternity.html.

[2] Association for Computing Machinery, "*ACM computing research repository*". URL http://www.acm.org/repository/.

[3] Association for Computing Machinery, "*ACM digital library*". URL http://www.acm.org/dl/.

[4] A. Brewer, W. Ding, K. Hahn, and A. Komlodi, "*The role of intermediary services in emerging digital libraries*". In 1st *ACM Intl. Conf. Digital Libraries*, pp. 29–35, Mar 1996.

[5] S. Buckingham Shum, E. Motta, and J. Doningue, "*ScholOnto: an ontology-based digital library server for research documents and disclosure*". *Intl. J. Digital Libraries* **3(3)**, pp. 237–248, Oct 2000.

[6] D. M. Choy et al., "*The Almaden distributed digital library system*". In *Digital Libraries: Research and Technology Advances*, N. R. Adams, B. K. Bhargara, M. Halem, and Y. Yesha (eds.), pp. 203–220, Springer-Verlag, May 1995. Lect. Notes Comput. Sci. vol. 1082.

[7] "*Digital libraries initiative*". URL http://www.dli2.nsf.gov/.

[8] D. G. Feitelson, "*Cooperative indexing, classification, and evaluation in BoW*". In *7th IFCIS Intl. Conf. Cooperative Information Syst.*, O. Etzion and P. Scheuermann (eds.), pp. 66–77, Springer-Verlag, Sep 2000. Lect. Notes Comput. Sci. vol. 1901.

[9] D. G. Feitelson, A. Kremer, D. Er-El, and M. Geffet, *Design and Implementation of the BoW User Interface.* Technical Report 2000-40, Hebrew University, Oct 2000.

[10] M. Geffet, *The Automatic Hierarchical Indexing and Document Matching in BoW.* Master's thesis, Hebrew University, Nov 2000.

[11] J. Ginsburg, "*Licensing, copyright, and fair use: the Thesauron project (towards an ASCAP for academics)*". In *Technology and Scholarly Communication*, R. Ekman and R. E. Quandt (eds.), pp. 337–346, University of California Press, 1999.

[12] D. Goh and J. Leggett, "*Patron-augmented digital libraries*". In 5th *ACM Conf. Digital Libraries*, pp. 153–163, Jun 2000.

[13] J. Y. Halpern and C. Lagoze, "*The computing research repository: promoting the rapid dissemination and archiving of computer science research*". In 4th *ACM Conf. Digital Libraries*, pp. 3–11, Aug 1999.

[14] "*HCI bibliography: human-computer interaction resources*". URL http://www.hcibib.org/.

[15] S. Hockey, "*Making technology work for scholarship: investing in the data*". In *Technology and Scholarly Communication*, R. Ekman and R. E. Quandt (eds.), pp. 17–36, University of California Press, 1999.

[16] IEEE Computer Society, "*Computer society digital library*". URL http://www.computer.org/epub/.

[17] D. M. Jones, "*The hypertext bibliography project*". URL http://theory.lcs.mit.edu/~dmjones/hbp/.

[18] J. M. Kleinberg, "*Authoritative sources in a hyperlinked environment*". In 9th *ACM-SIAM Symp. Discrete Alg.*, pp. 668–677, Jan 1998.

[19] B. Lang, "*Bricks and bytes: libraries in flux*". In *Books, Bricks & Bytes: Libraries in the Twenty-First Century*, S. R. Graubard and P. LeClerc (eds.), pp. 221–234, Transaction Publishers, New Brunswick, NJ, 1999.

[20] S. Lawrence, C. L. Giles, and K. Bollacker, "*Digital libraries and autonomous citation indexing*". *Computer* **32(6)**, pp. 67–71, Jun 1999.

[21] K-D. Lehmann, "*Making the transitory permanent: the intellectual heritage in a digitized world of knowledge*". In *Books, Bricks & Bytes: Libraries in the Twenty-First Century*, S. R. Graubard and P. LeClerc (eds.), pp. 307–329, Transaction Publishers, New Brunswick, NJ, 1999.

[22] J. F. Metzl, "*Searching for the catalog of catalogs*". In *Books, Bricks & Bytes: Libraries in the Twenty-First Century*, S. R. Graubard and P. LeClerc (eds.), pp. 147–160, Transaction Publishers, New Brunswick, NJ, 1999.

[23] A. S. Okerson, "*Buy or lease? two models for scholarly information at the end (or the beginning) of an era*". In *Books, Bricks & Bytes: Libraries in the Twenty-First Century*, S. R. Graubard and P. LeClerc (eds.), pp. 55–76, Transaction Publishers, New Brunswick, NJ, 1999.

[24] T. A. Phelps and R. Wilensky, "*Toward active, extensible, networked documents: multivalent architecture and applications*". In 1st *ACM Intl. Conf. Digital Libraries*, pp. 100–108, Mar 1996.

[25] "*ResearchIndex: the NECI scientific literature digital library*". URL http://csindex.com/.

[26] S. Rosenblatt and S. Whisler, "*The library and the university press: two views of costs and problems in scholarly publishing*". In *Technology and Scholarly Communication*, R. Ekman and R. E. Quandt (eds.), pp. 309–334, University of California Press, 1999.

[27] R. E. Schapire and Y. E. Singer, "*BoosTexter: a boosting-based system for text categorization*". *Machine Learning* **39(2/3)**, pp. 135–168, May 2000.

[28] C. Speier, J. Palmer, D. Wren, and S. Hahn, "*Faculty perception of electronic journals as scholarly communication: a question of prestige and legitimacy*". *J. Am. Soc. Inf. Sci.* **50(6)**, pp. 537–543, 1999.

[29] "*The digital object identifier system*". URL http://www.doi.org/.

[30] "*The open hyperdocument system*". URL http://www.bootstrap.org/ohs2/.

[31] "*The WEBBIB project*". URL http://www.cs.wpi.edu/˜webbib/.