OPEN, SELF ORGANISING REPOSITORY FOR SCIENTIFIC INFORMATION EXCHANGE

Bob Martens (*); Bo-Christer Björk (**); Ziga Turk (***)
*Vienna University of Technology (Austria) / **Swedish school of Economics and Business Administration - Helsinki (Finland) / ***University of Ljubljana (Slovenia)
b.martens@tuwien.ac.at; bo-christer.bjork@shh.fi; ziga.turk@itc.fgg.uni-lj.si

Keywords: Scientific Knowledge Management, Retrospective CAAD Research, CAAD-related Publications, Web-based Bibliographic Database, Machine Learning

Abstract
In the paper-based world, CAAD-associations, such as SIGRADI, and scientific publishers aim at getting the right people together and for making sure their work gets distributed to their peers. Electronic networks, such as the Internet, are providing scientists with the means to pursue those activities on their own. In this paper we present the goals of an EU project called SciX (Scientific information eXchange). The goal of is this project is to analyze the business processes of scientific publishing, to invent new publication models and through a series of pilots to demonstrate how this should work. In the envisioned scenarios, professional associations such as SIGRADI play an important role.

Resumen
En el mundo basado en el papel impreso, CAAD asociaciones, tal como SIGRADI y editores científicos están dirigidos a conseguir que las personas apropiadas se unan para asegurar que sus trabajos son distribuidos a sus iguales. Las redes electrónicas, como Internet, están proporcionando a los científicos el medio para proseguir esas actividades por sí mismos. En este documento presentamos los objetivos de un proyecto europeo llamado SciX (Intercambio de Información Científica). La finalidad de este proyecto es analizar los procesos de negocio de la publicación científica, inventar nuevos modelos de publicación y a través de una serie de pilotos, demostrar cómo ésto debería funcionar. En los escenarios visualizados, asociaciones profesionales tales como SIGRADI, juegan un importante papel.

1. Introduction
The history of the scientific publishing starts in the 17th century when the Royal Society of London created the ***Philosophical Transactions of the Royal Society of London*** (Gudeon, 2001). The intention was to create a public registry of ideas - a logbook or journal of the "present undertakings, studies and labours of the ingenious" - who thought of what first - to protect intellectual property and ensure the rapid evolution of scientific knowledge

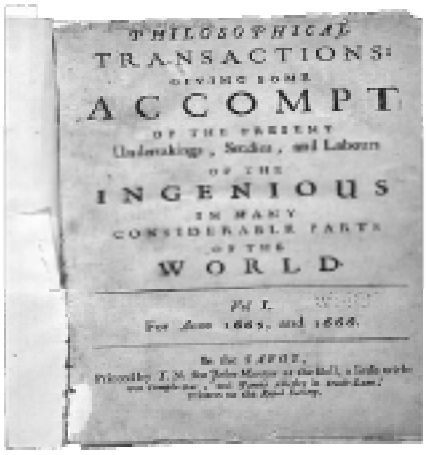

PHILOSOPHICAL
TRANSACTIONS:
ACCOMPT
INGENIOUS
WORLD

Fig 1 - Cover page of the Philosophical Transactions

For a long time, scientific publishing remained largely in the hands of learned societies and similar, scientist-driven institutions. Publishers have been entering the market since the mid 19th century, but their role has been marginal and profits negligible until the 1960s, when the Science Citation Index (**http://www.isinet.com/**) was introduced and the number of universities throughout the developed world grew quickly.

The business model of the publishers is rather a fascinating one. Scientists do the research, they write papers, they review their peers' work and they edit scientific journals. They give away the copyright to their work, for free, to a party that has not been taking part in the value-chain before. They then subscribe to usually rather expensive journals, so that they can learn about the work of their peers. In the SciX-project we believe that giving away the right to copy (copyright) and distribute results of scientific work to commercial publishers hinders the efficient exchange of this information and makes scientific results harder and more expensive to get.

1.1 Previous work
The SciX-partners have been active in the field of electronic publishing since the mid 1990s. Bo-Christer Björk and Ziga Turk have been the editor and one of the co-editors of the Electronic Journal of Information Technology in Construction (Itcon). The average time from submission of

a paper to its publication has been less than 6 months. Each published paper had an average of about 1.000 readers viewing the abstract and about 1.400 downloading the full text.
Since 1998, Bob Martens and and Ziga Turk have been managing CUMINCAD - Cumulative index of CAD (http://www.scix.net/cumincad) - the largest freely available database of papers related to computer-aided architectural design, particularly related to the education in this area. In the framework of annual conferences organized by regional CAAD-Associations (SIGRADI in South America, ACADIA in North America, eCAADe in Europe and CAADRIA in Australasia) thousands of papers have been published. Rarely were the proceedings published by a professional publisher, therefore, the texts were neither entered into commercial indexes, nor were they sold commercially. The full texts were not broadly available; only conference attendees had copies. On the other hand, the associations retained in most cases the copyright to this work and could therefore allow its publication/archiving in CUMINCAD. Thus this work is available on the net and rescued from oblivion. At the time of writing, CUMINCAD includes 4.150 papers with abstracts. 883 papers are available in full text as well.
Within the SIGRADI-Association not only Conference Papers in the English language are published. CUMINCAD primarily concentrates on English publications. Preferably the full-text should be in English, contributions from the CAAD-conferences issuing an English Abstract are also included. This criteria, however, does not apply to some SIGRADI-papers and therefore, CUMINCADES was founded, in order to provide all SIGRADI-Papers (see direct link on cumincad.scix.net.)

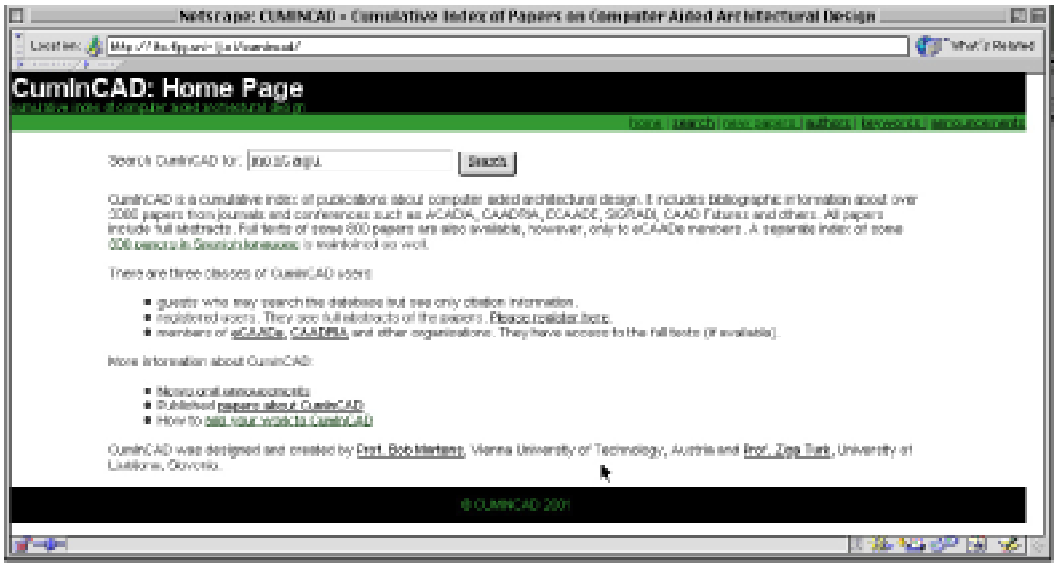


Fig 2a-b - User interface of the CUMINCAD database

1.2 Goals of this paper

The goal of this paper is to engage the SIGRADI-community in the SciX project. Since 1997, about 440 papers have been published in the SIGRADI-proceedings. Most of these proceedings are the so-called gray literature - published by the conference organizers - not generally available to a broader audience. And yet in this community valuable contributions have been made, particularly in relation to computer-integrated construction and product modeling. According to a study (Umich, 2001), about 50% of the costs related to making some literature electronically available is related to scanning and further 20% to the digitalization of the material. By working closely with the scientific community and with the scientists who authored the material these costs can be saved.

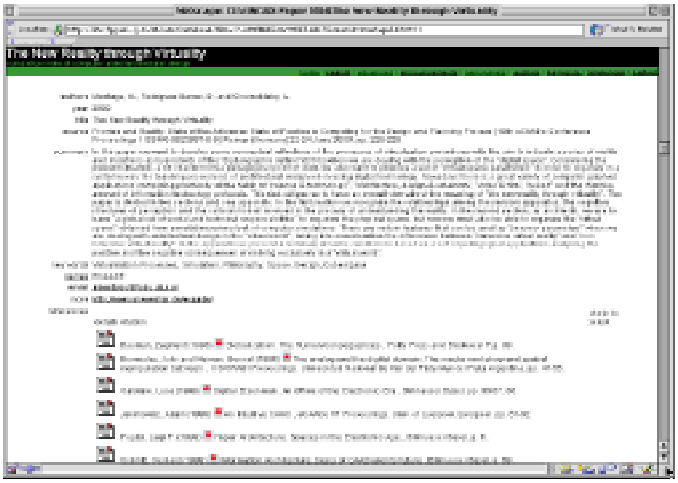
In the SciX project we envisage the setup of a target user-group with representatives of the main professional organizations creating the scientific publications. The role of this group is to comment on the work so that the results are relevant to the community. On the other hand, within SciX, services and tools will be created and placed on the open source license, which could be useful for a community like SIGRADI.

1.3 SIGRADI and SciX

SIGRADI-members are the potential users of SciX's platforms, authors and readers of the papers. The objectives of this contribution focus on involving SIGRADI-community in the developments in SciX, on fine-shaping the goals as well as on defining the requirements and monitoring the usability of the pilots. CUMINCAD serves as a pilot with extended content in the field of CAAD.

As all SIGRADI-Conferences up to now have been included in CUMINCAD (bibliographic particulars with summary, etc.) the issues of 1998, 2000 and 2001 have been made available in form of full papers in pdf-format. At present, the years 1997 and 1999, not completely available in digital form, are being converted accordingly and will be provided online by the end of 2002. Furthermore, all Conference Papers since 1997 surely could be made available on CD-ROM without going to any extraordinary efforts.

These activities focus on significantly increasing potential availability as well as on developing and enhancing a "knowledge management". As a rule conference papers contain references specifying their positioning within the particular context of topics. With a few exceptions this applies to most SIGRADI-papers. These references can be taken from the pdf-files and are to be broken down into four components: author(s), year, title and source. Even though this splitting up can be performed by means of program routine manual re-editing will be necessary as the uniform quotation manner is not available in all cases. The implementation of a collection of references from previous SIGRADI-proceedings is also scheduled for the end of 2002 (see prototype in fig. 3a-b). A first approximation as interlinking of references with each other as well as cross-linking with the CUMINCAD-records might lead to a novel way of consideration. Thus future interest may also be directed to the connection of the most quoted references and their correlating papers. In the end both the selections of the references as well as the sessions etc. are a result of human structuring ability and could be useful for machine learning efforts.



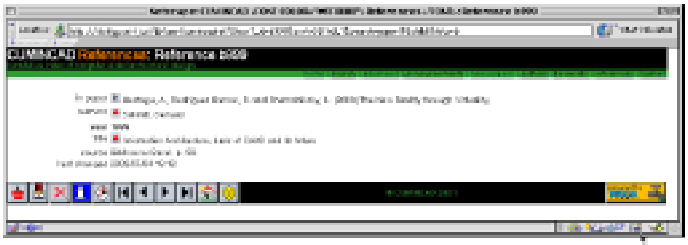


Fig 3a-b - CUMINCAD database: extension with references (citations)

2. Related Work

Both professional organizations, groups of publishers as well as specialized companies are providing added value services related to scientific publishing. Several bibliographical databases are providing sophisticated search engines on bibliographic information about publications (such as titles and abstracts). Full texts are, as a rule, not available.

Table 1 - Commercial indexes and bibliographic databases

Number of abstracts	8.000.000	500.000	575.000	4.000	2.500.000
Availability	$	$	$	Free	Free

The Internet represents a threat to traditional publishers. While some years ago, the Internet was a first resource for getting scientific information (Bjoerk and Turk, 2000), it is today becoming the only resource, particularly with the young researchers. Traditional publishers are responding with services such as ScienceDirect that allows pay-on-demand access to the full texts of published papers.

Another strategy of publishers is to avoid dealing directly with the readers of the journals and attempting to close direct, longterm deals with either whole universities (Landesman - Van Reenen, 2000) or whole countries (http://www.lib.helsinki.fi/finelib/). Although discounts are offered if an institution subscribes to a full spectrum of journals the economies of such deals for the funding bodies and the researchers are not necessarily positive.

2.1 Free Publishing Model

The idea to use the Internet for scientific publication is not new. Existing solutions are of the following types:

- ***Preprint archives*** offer drafts of papers that have been submitted to publication in paper-based journals. No quality control is provided. Often, the papers are quite similar to the final works published. Perhaps the best known such archive is the Los Alamos or arXiv preprints archive (http://www.arxiv.org/).
- ***Electronic journals (eJournals) and magazines (eZines)***. Similar to ITcon they provide similar quality control mechanisms as paper-based publications. 400 such journals supposedly existed in 1999, including a Journal on Electronic Publishing. Today this number is estimated at over 1.000.
- ***On-line bibliographies*** are collections of papers (usually without full text) from a certain discipline. After having been published as a booklet for a number of years the abstracts are currently freely available through a database on the web. A well known example is the CiteSeer service offering full texts of some 2.5million papers related to computer science. CiteSeer is accumulating the papers from the Web and copying them from authors' websites to one central location where they are, classified, index and cross-referenced.

The problems of all kinds of services include:

- ***Sustainability*** - Although the funds required to run such services are rather small, after the initial work done by the enthusiasts, a stable funding is required. The mortality rate of the electronic journals was 25% over two years (Wells, 1999).
- ***Copyright*** - Many services include material that has been previously published in a way that required the transfer of the copyright.
- ***Prestige*** - An important factor in deciding where to publish is the prestige of a journal (Bjoerk and Turk, 2000), as perceived by the universities' or national research review processes. It is not uncommon, that a publication in a fully reviewed electronic journal is less valuable than publication at a conference where the author actually paid a fee to get the work published in impressively hard-bound proceedings.

2.2 Examples in the Field of Software

The policy of the ARPA and the NSF in the United States was that all research supported through public funding should make the results available free of charge. This has not been entirely true for published papers, but has worked excellently with software. Programs written in the context of research projects were made available - for free, usually including source code - on the Internet. In fact, the software to run the Internet in the first place was available for free. This created the critical mass for the so-called open-source initiative (http://www.opensource.org/). An increasing number of operating systems, application programs and tools are available for free. The market share of those systems is growing and they are being used as a platform for vertical applications by companies such as IBM.

On the other hand, the European funded research projects (such as the 4th and 5th Framework Projects) never made a requirement for making the results publicly available. The excuse used was that commercial companies are co-funding this work and that they are not interested in making available what could be their competitive advantage. We are not aware of the scientific community challenging this system. Labeling most of the reports "restricted" actually restricted the readership to the project officers and the reviewers.

3. Goals of SciX

2-4% of the European GDP is spent on research and development - on creating new knowledge. While several projects deal with the management of knowledge created within the industry, little has changed in the past hundred years in the ways knowledge, created by scientific research and published in scientific journals, is handled. The current mainstream scientific publication process has so far been only marginally affected by the possibilities offered by the Internet, despite some pioneering endeavors. This does not result from lack of enthusiasm, but rather from a lack of sound business models and pilots to demonstrate the benefits of totally free scientific publication archives to the organizations to ultimately fund the development and maintenance of such.

The objectives of this project are:

- to enable scientists time- and cost-efficient access to their peers' work by creating a repository of electronic publications;
- to make the scientific materials in the repository also available to non-scientists - engineers, architects from the industry and explore new business scenarios;
- to support building a virtual on-line community of authors and readers.

To accomplish above in SciX it is intended to:

- create the necessary services infrastructure and populate it with at least 5.000 papers from the domain of architecture and engineering;
- strengthen the already initiated transition to new modes of scientific publishing processes so that the cheap dissemination channels of the Internet are put to efficient use; we will do so by setting up infrastructure generating an electronic journal and making it available under open-source licensing;
- perform a social-economic analysis of new business;
- investigate the legal, social and psychological obstacles to using eWork approaches in this area as well; this will include a survey amongst approx. 300 of our colleagues on their views regarding e-publishing;
- develop a method to benchmark scientific journals based on user requirements in the Internet era;
- enable efficient access to scientific results.

In this project a process reengineering view of the whole life-cycle process of scientific papers will be performed aimed at resulting in savings of 80-90% in the distribution – retrieval costs. Compared to the 10-20 % approaches often taken in development projects initiated by commercial publishers and libraries, these savings are very promising. The key issue is the paradigm shift to see scientific publications not as a commodity to be sold or archived but as an essential part in a larger scientific communication process, and to look for solutions based on the premise of globally free information on the World Wide Web, thus side-stepping some of the traditional intermediaries altogether.

3.1 Automate Repository Management through Self Organization

The amount of digitally stored technical data, both general and corporate, is growing rapidly - more rapidly than the ability of

humans to appropriately structure, classify or index it, so that it could be found and (re-) used. Typically, this information is available through different search techniques. Searching, however, implies that the user knows what to look for. Another approach to access the data is by browsing requiring a certain structure imposed over the data items. The main function of the structure is to provide user navigation through the data. The structure should tell the user what items are similar, which are different, and how they differ. The simplest structures of this kind are clusters or groups of similar data items. By using data mining techniques it is possible to create an algorithm that would create clusters of data automatically so that the clusters would be similar to the human interpretation of such data. For example, given one or a few papers related to certain topic, the machine should come up with a cluster of similar papers, which should be of interest to the reader as well. Such clustering becomes very interesting when applied to large repositories of publications, such as the one planned in this project.

3.2 Simplified Use through Intelligent Personalized Agents

Another important part of the project is a user-profiling system that would add value in combination with the automation described above. Automatic notification on new papers matching the profiles' interest and selective searches will be provided without having to create a very sophisticated profile. The user will be able to semi-automatically modify the query with assistance of the system and update his user profile.

3.3 Investigate Legal, Social and Psychological Issues

The main problem to a new vision of information exchange in science is the copyright that researchers currently give away to the commercial publishers for free, and which results in severe obstacles for potential readers to retrieve the information they need. There are also other barriers for a shift to free repositories dealing with perceived risks of Internet publishing, sluggishness of academic department to change their "rating" systems, etc. which need to be studied.

4. Conclusions

Current methods for accessing scientific results are highly inefficient in view of the technical potential offered by the Internet. This also applies to scientific research findings. From the viewpoint of the public sector financing research, they are aimed at reusing in other research and application in industry, not as a commodity to be sold per se for a profit. It would seem to prove wise for the public R&D funding bodies and for the academic community as a whole to have a completely free cyberspace of scientific information, in order to speed up the scientific research process and save costs. The objectives of the SciX project described in this paper are to explore business models and techniques which speed up the process from submission to final publication, allow a more rich content (multi-media), provide readers with more efficient mechanisms for retrieving publications of interest and increase readership through the abolition of barriers such as subscriptions.

One of the advantages of rooted communities, such as SIGRADI, is its track record and prestige; the hundreds of papers published by people, who may now be regarded as the authorities in the field. This track record, however, is remembered by a few dozens who have been attending SIGRADI conferences regularly. All others could appreciate the achievements of SIGRADI if they were electronically and freely available.

Acknowledgements

The presented work has been conducted in the context of the SciX project, funded by the European Commission under the contract IST-2001-33127. The contribution of the funding agency as well as that of the industrial partners in the project is gratefully acknowledged. The opinions expressed in this paper are that of the authors and do not necessarily represent the opinions of their employers, of the SciX Consortium or of the European Commission.

References

- Björk, Bo-Christer, Turk, Ziga (2000). "How Scientists Retrieve Publications: An Empirical Study of How the Internet Is Overtaking Paper Media", In ***Journal of Electronic Publishing***, Michigan University Press, Vol. 6/2. 2000, <http://www.press.umich.edu/jep/06-02/bjork.html>
- Guedon, J.C. (2001). "In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing", In ***Proceedings of the 138th Annual Meeting*** - Association of Research Libraries, Toronto, Ontario, May 2001, <http://www.arl.org/arl/proceedings/138/guedon.html>
- Landesman, M., Van Reenen, J. (2000). "Consortia vs. Reform - Creating Congruence", In Journal of Electronic Publishing, Vol 6/2, <http://www.press.umich.edu/jep/06-02/landesman.html>
- Martens, B., Turk, Z. and Cerovsek, T. (2001). "Digital Proceedings: Experiences regarding Creating and Using", In ***Architectural Information Management*** [eCAADe Conference Proceedings], Helsinki, pp. 25-29.
- Turk, Z., Cerovsek, T. and Martens, B. (2001). "The Topics of CAAD – A Machine's Perspective", In ***CAAD futures 2001*** [Conference Proceedings], Eindhoven, pp. 547-560.
- University of Michigan Library Services (2001). "Assessing the Cost of Conversion", In <http://www.umdl.umich.edu/pubs/moa4_costs.pdf>
- Wells, A. (1999). "Exploring the Development of the Independent, Electronic, Scholarly Journal" [M.Sc. in Information Management], In ***Electronic Dissertations Library***, Department of Information Studies, University of Sheffield, UK, http://panizzi.shef.ac.uk/elecdiss/edl0001/>

All URLs mentioned in this paper were checked in August 2002.