Re-engineering the scientific knowledge management process: the SciX project

Bob Martens[a,*], Ziga Turk[b], Bo-Christer Björk[c], Grahame Cooper[d]

[a] Vienna University of Technology, Austria
[b] University of Ljubljana, Ljubljana, Slovenia
[c] Swedish School of Economics and Business Administration, Finland
[d] Salford University, United Kingdom

Abstract

In the past, paper-based publications were both the medium for the information exchange among the scientists as well as the measure of their scientific quality. Recently electronic publication is increasingly important for the dissemination of scientific work, while evaluation of scientists and institutions remains largely based on the printed publications. In the "material world", associations, such as eCAADe have been providing a stage for getting the right people together and through annual conferences and seminars made sure that research results were heard and read. The Internet and the Web are allowing the scientists to pursue those activities on their own or through flexibly organized "virtual" associations. Electronic publications also provide a basis for efficient management of scientific knowledge—discovering related work, reduction of the duplicate efforts, establishment of virtual research teams etc. In this paper, we present the goals and initial results of an EU-project called SciX: "Open, Self Organising Repository for Scientific Information Exchange". The goal of SciX is to analyze the business processes of scientific publishing, to invent new publication models and through a series of pilots to demonstrate how this will work. In the envisioned scenarios, professional associations such as eCAADe play an important role. Their members are the potential users of SciX's platforms, authors and readers of the papers. Associations could also become the publishers and archivists of the knowledge created within their respective community. The objectives of this contribution focus on involving the eCAADe-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. © 2003 Elsevier B.V. All rights reserved.

Keywords: Scientific knowledge management; Retrospective CAAD research; CAAD-related publications; Web-based bibliographic database

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* [1]. The intention was to create a public registry of ideas—a logbook or journal of the "present undertakings, studies and labours of the ingenious" to protect intellectual property and ensure the rapid evolution of scientific knowledge (Fig. 1). For a long time, scientific publishing remained largely in the hands of learned societies and similar, scientist-driven institutions. Publishers have been

* Corresponding author.
 E-mail addresses: b.martens@tuwien.ac.at (B. Martens), ziga.turk@itc.fgg.uni-lj.si (Z. Turk), bo-christer.bjork@shh.fi (B.-C. Björk).
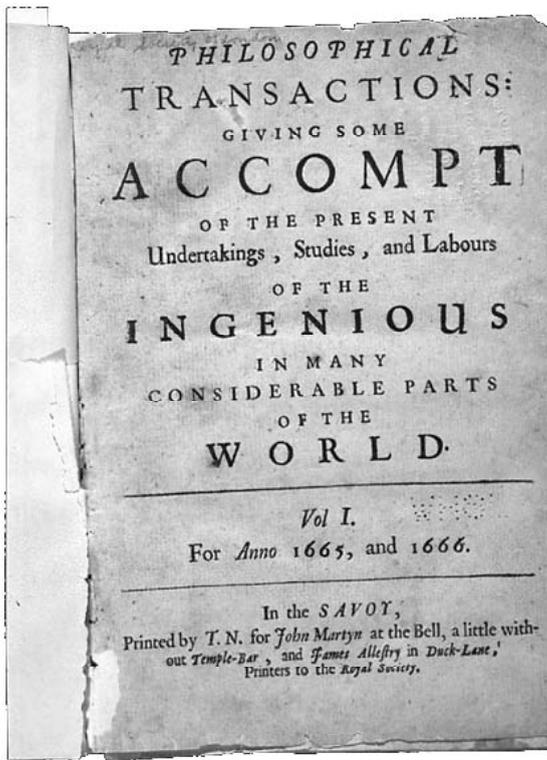
Fig. 1. Cover page of the Philosophical Transactions.

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.

An increasingly important client of the publishers and the indexing businesses are the government agencies funding the scientific research. They use these publications and corresponding indexes to evaluate the performance of scientists and institutions. While the relevance of these publications for the exchange of knowledge is decreasing, their importance for the evaluation of the scientific community is increasing, which is leading in the end to a rather strange situation.

The scientific community is losing its independence [1] because of a decisive role of businesses outside the scientific community in the evaluation process that has a direct consequence in the funding. By providing journals and indexing them, these

businesses are creating an opportunity for the researchers to "score" and when they do so, more funding for that research is available, which means more subscribers for that particular family of journals.

The "publish or perish" maxim explains why the scientists are willing to accept almost any conditions from the publishers: they do the research, they write papers, they review their peers' work and they edit scientific journals. They give away the copyright to their work to a party that has not been taking part in the value-chain before. They do all of this for free and finally 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 hinders the efficient exchange of this information and makes scientific results harder and more expensive to acquire. We argue that the control of the scientific information exchange should be returned to the scientific community, while maintaining the strictest possible quality control that the peer-review process enables.

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 1000 readers viewing the abstract and about 1400 downloading the full text.

Since 1998, Bob Martens and Ziga Turk have been managing CUMINCAD-Cumulative index of CAD (http://cumincad.scix.net)—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 (ACADIA in North America, eCAADe in Europe, Sigradi in South America 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 the CUMINCAD. Thus this work is available on the net and rescued from oblivion. At the time of writing, CUMINCAD includes 4618 papers with abstracts. 1667 papers are available in full text as well.

1.2. Goals of this paper

The goal of this paper is to engage the CAAD-communities in electronic publishing. The tools,

Fig. 2. (a–b) User interface of the CUMINCAD database.

services, guidelines and business models developed by SciX may provide the necessary infrastructure. This will also include scripts to support web-based digital archives ranging from global digital libraries such as CUMINCAD to institutional or personal archives. They will go beyond simple repositories of papers and include community building facilities, quality control (reviews), discussions as well as advanced knowledge management. Web services are being built that allow for some of that to be rented on the Web, without the need for one's own server or programming expertise. A set of guidelines discusses how to approach the problem of getting the legacy content into the library, both from the technical and organizational perspective. The business model study addresses the need to make such community-based efforts sustainable over longer periods of time (Fig. 2).

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 for free (Table 1).

The Internet represents a threat to traditional publishers. While some years ago, the Internet was a first resource for getting scientific information [2], it is today becoming the only resource, particularly for 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 attempt-

ing to close direct, long-term deals with either whole universities [3] 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. Four hundred such journals supposedly existed in 1999, including a Journal on Electronic Publishing. Today this number is estimated at over 1000.
- *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 2.5 million 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.

Table 1
Commercial indexes and bibliographic databases

	Ei compendex	ICONDA	RSWB	CUMINCAD	CiteSeer
Number of records	6,000,000	500,000	575,000	5000	2,500,000
Availability	$	$	$	Free	Free

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, stable funding is required. The mortality rate of the electronic journals was 25% over 2 years [4].
- *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 [2], 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.

2.3. The open archives initiative

In the framework of paper-based publishing, a few dozen publishers control most of the scientific publications and making a rather complete index involves including the publications of the few major ones. If, however, thousands will be creating digital archives on the Internet, indexing that information could be quite challenging. Web search engines, such as Google or Altavista are an appropriate tool to search for scientific information.

The Open Archives Initiative (http://www.openarchives.org/) is standardizing the metadata structure and the API of an archive so that the archive can be indexed and several archives can be searched by the users at once. Moreover, Open Archives Initiative is developing standards that aim to facilitate the efficient dissemination of content.

3. Goals of SciX

Two to four percent 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 way 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 endeavours. This does not result from lack of enthusiasm, but rather from a lack of sound business models and pilots to demonstrate the benefits of free scientific publication archives to organizations who 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, archi-

tects 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 5000 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 approximately 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 re-engineering view of the whole life-cycle process of scientific papers will be performed, aimed at results with savings of 80–90% in the distribution, which are 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, which requires 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 semiautomatically 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.
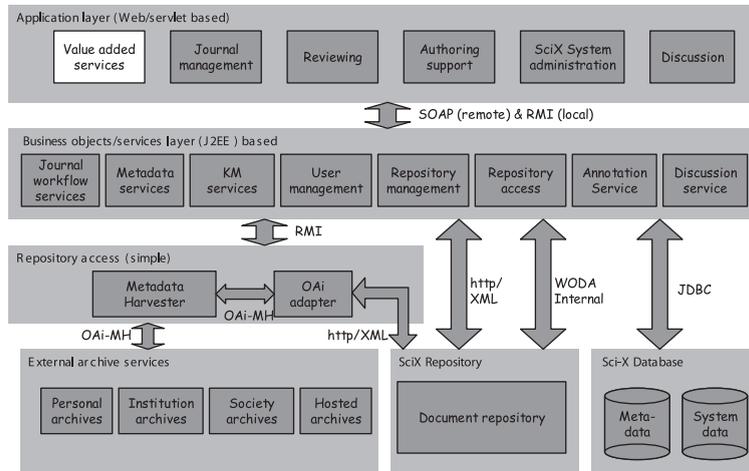
Fig. 3. Separation of various SciX-services with interoperability.

3.4. Develop benchmarking methods for scientific journals

Typically scientific journals have been rated by prestige, often based on subjective evaluations or to some extent on the use of citation indexes. Ratings have been done implicitly through university departments, for instance in shortlists of accepted publications for promotion. Little attention has been paid to questions of how quickly and efficiently the informa-tion passes to experts for whom the information could be useful. Thus it would be very meaningful to develop methods for benchmarking journals including factors other than the scientific quality of the papers (turnover time from submission to publication, avail-ability, readership etc.). Such a benchmarking tool will be developed in the project and tested with a number of journals of different categories. The main value of such a tool could increase awareness within scientific communities of the deficiencies of their
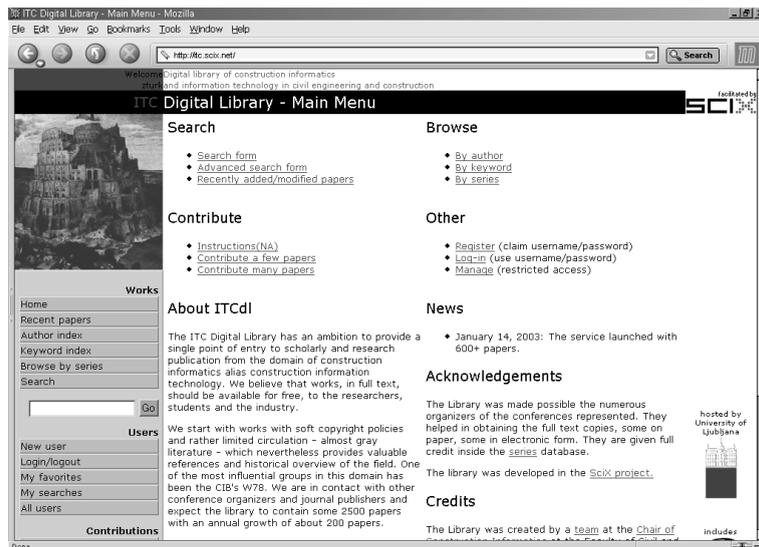

Fig. 4. User interface of the ITC database.

current communication process, hopefully triggering activities changing the process.

4. Mid-term results

At the time of writing, SciX is at about one half of its duration. This section presents the draft architecture of the SciX services and initial prototypes.

4.1. The SciX architecture

The electronic publishing area is extremely fast moving, developing at a significant pace, particularly

in the areas of inter-working. Therefore, a modular architecture is proposed for the SciX system, allowing functions to be included, left out, added, or replaced relatively easily in any particular implementation. A widely accepted principle is the separation of data storage from service provision.

The SciX architecture is identifying a number of business objects such as article, user, annotation, review, recommendation, discussion, rating, reading etc., that can provide various kinds of applications on top of a digital library. The objects are being implemented as Web services that can be fairly self-sufficient, however, combined they can provide applications such as a self-archive, institutional archive,
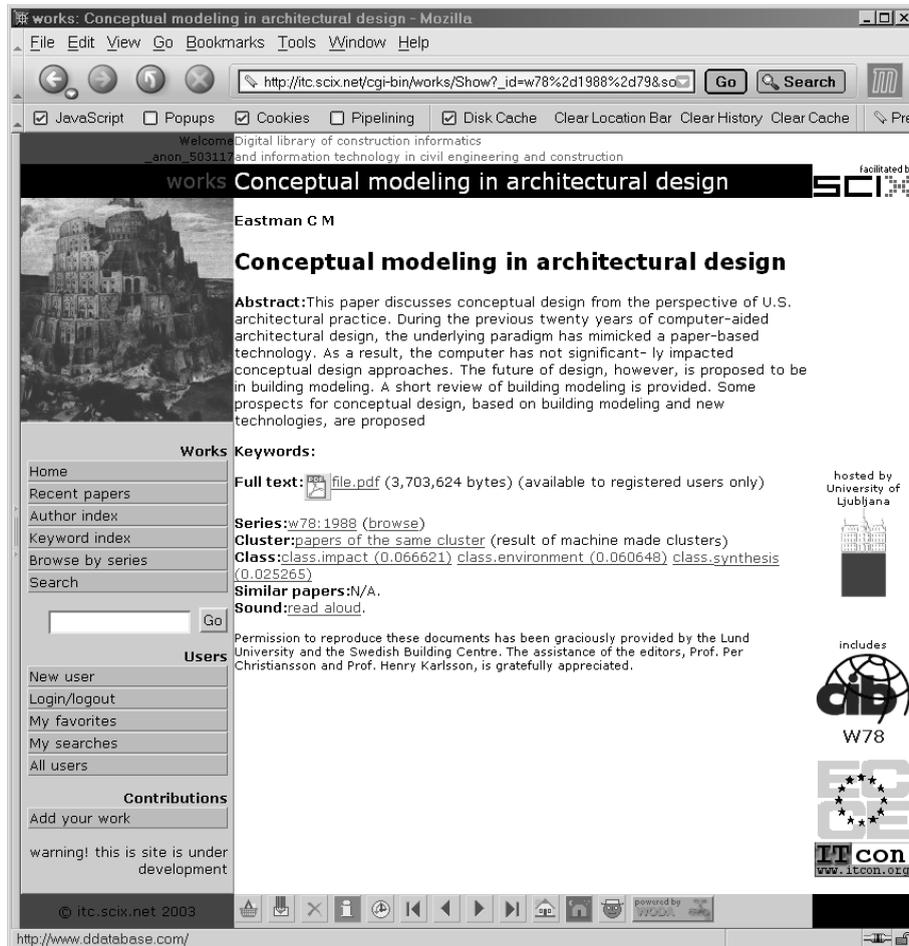

Fig. 5. A display of a single paper.

topic archive, workshop or conference and a journal. Among others, the differences among these applications are different levels of quality control and differences in the workflow and timeline.

Additional "knowledge management" (KM) services will be provided on top of a repository. This may include such things as comparison of articles, searching for similar articles, summarisation, and linking between articles. These services may be provided as additional modules, separate from the repository. It will be important that such services are able to access metadata and articles from other repositories in addition to (or instead of) a local one if they are adequately to support activities such as authoring.

While services may often be accessed through a simple, web-based, thin-client interface, it is important to allow for access by other means, including more closely integrated approaches. For example, authoring tools will need to integrate quite closely with KM services and metadata services in order to provide the kinds of benefits identified in the requirements analysis (Fig. 3).

A modular architecture is proposed for the SciX system, allowing objects to be included, left out, added, or replaced relatively easily in any particular implementation or application. This is made possible because the object-oriented schema is not implemented in a monolithic relational database application but rather by a number of services; the collaboration among them can be established at runtime.

4.2. SciX pilot

Many members of the SciX consortium have been deeply involved with the CIB's Working Commission 78 "Information Technology in Construction". CIB is the International Council for Building Research and Information and perhaps the most important world-wide acting association in the filed of engineering and construction. SciX technology is used to implement the digital library of the working commission 78. Over the last 15 years, it published about 800 papers that would have remained a well-kept secret of a few that possessed proceedings. Now these papers are available for free on the Internet. Add on services, such as the search for similar papers, automatic clustering, discussion forums, reading lists etc., are available as well (Figs. 4 and 5) [6].

Table 2
Overview on variations in digitalization (assumed quantity of 1000 pages—corresponds to around 125 papers)

(A) Source in digital format—conversion to full text pdf
 Estimation: 4 working hours—200 Euro/=0.20 Euro per page
In principle, any electronic document type can be converted into pdf. In case digital data is properly archived and so far no pdf-files were created, the work involved in order to create a pdf-file is not extremely time-consuming. Experience is necessary in order to have the resolution as well as compression set in an appropriate way, so that the resulting pdf-output is not too large (internet-download etc.). The content provider may decide not to outsource this job, in order to avoid direct access to the original source.

(B) Source in paper-based format—conversion to "image"-pdf
 Estimation: 300 Euro/=0.30 Euro per page
Taking into account that the material is not unique (i.e. not the only existing copy)—the original may be scanned in the same way as photocopies are produced, with a certain damage of the book. The output would be a single pdf-file, which does not support a full text search, as the basis is still an "image".

(C) Source in paper-based format—conversion to full text pdf
 Estimation: 1900 Euro/=1.90 Euro per page
The procedure as described under (B) is extended with an OCR-conversion of the scanned page. The elimination of mistakes after this step has to be performed manually and is time-consuming. Furthermore, the efforts depend rather much on the printing quality and the font used. Therefore, the price mentioned below has to be seen for an average printing quality as reachable since about 10–15 years. Finally a conversion to a single pdf-file will be performed.

(D) Source in paper-based format—retype and conversion to full text pdf
 Estimation: 1600 Euro/=1.60 Euro per page
The step of scanning is missing here, and the outcome is a plain text file, which can be easily converted to pdf. The average number of 2000 characters per pages define the basis for estimation. A recreation of the original layout is not included here as this would require scanning and lay-outing.

(E) Source in paper-based format—selected conversion to full text pdf
 Estimation: 500 Euro/=0.50 Euro per page
The procedure as described under (C) is extended with an OCR-conversion of selected pages, which contain the summary and the references. For 125 contributions, half a page each is counted for the summary and/or references. Therefore, the calculation is a mix of (B) for 875 pages and (C) for 125 pages.

4.3. Variations of digitalization

A major cost and effort in creating digital libraries is in obtaining the legacy material, which usually only exists on paper. Current material is available in digital format. In order to get a clearer view on the impact of different options in the range of digitalization, an overview with estimated financial figures was developed. It has to be noted that the output of these variations is a single pdf-file. Depending on the final organization of the file information, some rearrangements have to be made as 1000 pages are equivalent to 125 papers (Table 2).

A description of variations aims at making the relationship between *costs* and *benefits* visible [7]. Upon availability of budget, decisions can be made accordingly, which lead to a selection. This applies especially to retrospective digitalization projects. Decision makers can opt between the alternative of having as many as possible paper-based pages converted into a digital format, without further "intelligence", or they can also focus on a selection and therefore choose the full text option, which is far more expensive than the electronic content but allows for enrichment with "intelligence".

Further combinations are imaginable. For example, the elimination of errors in variation C, which could be replaced by a retype as described in variation D, however, the estimation for this new variation procedure would end somewhere around 2 Euro per page. Vendors who offer services in this area will ask for a representative sample in order to produce a realistic offer.

Based on the above figures, in the case of the CUMINCAD database (1667 full texts, from which 559 scanned and 1108 converted from digital source), the total investment is 8200 Euro. However, nearly 90% of these costs are related to full digitalization (Variation C.) of 3500 pages from eCAADe-proceedings [5].

5. 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 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 and 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.

Acknowledgements

References

[1] J.C. Guedon, In Oldenburg's long shadow: librarians, research scientists, publishers, and the control of scientific publishing, proceedings, association of research libraries, Proceedings of the 138th Annual Meeting, Toronto, Ontario, 2001 May, http://www.arl.org/arl/proceedings/138/guedon.html.

[2] B.-C. Björk, Z. Turk, How scientists retrieve publications: an empirical study of how the internet is overtaking paper media, Journal of Electronic Publishing, vol. 6/2, Michigan Univ. Press, 2000, http://www.press.umich.edu/jep/06-02/bjork.html.

[3] M. Landesman, J. Van Reenen, Consortia vs. reform—creating congruence, Journal of Electronic Publishing, vol. 6/2, http://www.press.umich.edu/jep/06-02/landesman.html.

[4] A. Wells, Exploring the Development of the Independent, Electronic, Scholarly Journal [MSc in Information Management], in Electronic Dissertations Library, Department of Information Studies, University of Sheffield, UK, 1999, http://panizzi.shef.ac.uk/elecdiss/edl0001/.

[5] B. Martens, Z. Turk, T. Cerovsek, Digital proceedings: experi-

ences regarding creating and using, Architectural Information Management [eCAADe Conference Proceedings], eCAADe Association, Glasgow, Helsinki, 2001, pp. 25–29.

[6] Z. Turk, T. Cerovsek, B. Martens, The Topics of CAAD—A Machine's Perspective, CAAD futures 2001 [Conference Proceedings], Kluwers, Eindhoven, 2001, pp. 547–560.

[7] University of Michigan Library Services, Assessing the Cost of Conversion, 2001, http://www.umdl.umich.edu/pubs/moa4_costs.pdf.