A STUDY ON CONTENT SOURCES AND ACQUISITION TECHNIQUES OF CAAD-RELATED PUBLICATIONS

BOB MARTENS
Vienna University of Technology (Austria)
b.martens@tuwien.ac.at

ZIGA TURK
University of Ljubljana (Slovenia)
ziga.turk@itc.fgg.uni-l.si

AND

GRAHAME COOPER
University of Salford (U.K.)
g.s.cooper@salford.ac.uk

Abstract. The scientific publication process has been so far only marginally affected by the possibilities of the Internet. This may be attributed to a lack of sound business models and pilots to demonstrate the ultimate benefits of free scientific publication. A team of universities, Internet publishers and applied research institutes proposes in the framework of the SciX-project (open, self-organizing repository for scientific information exchange) to demonstrate these benefits and re-engineer parts of the scientific publication process. This contribution focuses on the findings from investigations into the identification and acquisition of content sources - such as conference papers, theses and journal papers - related to the field of CAAD. Special attention has been paid to the different types of available information, such as bibliographical data, summaries, full texts, etc. In this paper, an overall estimation concerning the expected scientific output in the field of CAAD (within a midterm perspective), and the dissemination of already fully digitally stored publications as well as procedures (with financial figures) for retrospective digitalization of paper-based publications will be presented. The workflow concerning digitalization and conversion was studied, as different levels of output are feasible.

Keywords. Retrospective CAAD Research, Web-based Bibliographic Database, Electronic Publishing

1. Repositories and Sources of Content

The CUMINCAD-repository is used as an object of study and research within the SciX-project. CUMINCAD, an acronym for Cumulative Index on CAD (http://cumincad.scix.net), results from the insight, that no (web-based) repository existed to support further dissemination of academic teaching and research work in the field of CAAD. A similar initiative called i-CAADRIA, however, provides a service to the CAADRIA-membership by offering a database that focuses exclusively on CAADRIA-publications.

Web services are often trapped in a vicious circle. If a repository is empty, users are not motivated to enter their own data into the repository. Therefore, the repository remains empty. The objective of "SciX-Workpackage 2" on repository content of the SciX-project is to cut this vicious circle by providing an initial set of content that would establish the resource as a relevant one. Then the snowball effect would take care of individual scientists being personally interested in offering their work for entry. This paper is closely related to "SciX-Deliverable 5" on content source and acquisition techniques of the SciX-project and details the results from a period of investigation into the identification of content sources: proceedings, journals, theses and citations. Its purpose is to investigate the procedure concerning the harvesting of content sources, to be made available as initial content for a demonstration repository such as CUMINCAD.

1.1. CONFERENCE PROCEEDINGS

In the 1980's first considerations regarding the implementation of CAAD as a research and teaching area occurred at many architecture schools, quickly resulting in a need for exchange of ideas and experiences. Platforms such as ACADIA, CAADRIA, eCAADe, SiGraDi and CAAD Futures were established to fill that gap. The (bi-) annual conference can be regarded as the major event and various other events were to follow. Up to the second half of the 1990's paper-based proceedings were published, generally with a small circulation only. Smaller editions such as these are safely stored away in the studies of participants of the conferences and university members, but rarely become available to the wider public. Library networks have so far not been systematically supplied with copies.

A total of 2.172 conference papers were released in the period 1981-2001 (worldwide, 53 conferences). The current publication output on a annual base - without the biannual CAAD futures - is approx. 300 papers. This, however, seems to be a number which will not grow substantially in the near future.

1.2. THESES AND DISSERTATIONS

A thesis or a dissertation is by its very nature produced in limited quantities and in many cases the only copy available is the archival copy deposited in the library of the awarding institution. In the course of a request, cost and delay factors are a significant deterrent. The poor level of usage is attributed to a number of factors, such as lack of knowledge that the thesis exists or of its contents, or lack of ready availability. The relatively restricted access to print theses is the predominant reason for their under-utilisation. Making the full-text available from any computer desktop across the internet would greatly increase knowledge, access and availability of such a significant resource. Most authors write their theses nowadays in electronic format using standard word-processing and desktop publishing as well as graphics software. These tools also provide them with the opportunity to include multimedia components. However, use of these technologies is limited by the requirements for theses to be submitted in paper format. Changing the means for submitting theses from paper to electronic format would result in a more efficient and less costly process for the student in terms of the cost and time involved in making multiple paperbound copies.

Unlike, for example, CAAD conference proceedings, exact figures cannot easily be quoted for the "production" of CAAD-related theses and dissertations. A closer look at for example the full collection of the Faculty of Architecture at Eindhoven University of Technology – which provides of a complete set of doctoral e-dissertations (90 full texts since 1974) - shows that a smaller number is focusing on or is related to CAAD and Construction IT. Due to various reasons, however, not every location has the same research output in terms of acknowledged doctoral dissertations. All in all, the total number of architectural education sites probably does not exceed 1.000, possibly corresponding to the number of individual researchers explicitly dedicating their work to CAAD. The worldwide number of 50 dissertations annually making up a total of 500 since approx 1990 seems reasonable.

1.3 JOURNAL PAPERS

There is a wide variety of academic journals in which CAAD-scientists publish their work. A search on a number of journals - with the exception of e-journals - clearly showed that databases rarely issue an eprint. However, the displayed bibliographical data may be enriched with a summary and keywords. An exact number of possible entries is not available here, but the journals as such are relatively easy to find (no "grey literature"). Entries will have to be selected manually (volume by volume), as not all papers will be related to the field of CAAD and construction IT. Recording of at least metadata in a repository reduces the need for researchers to visit corresponding individual journal sites on the web.

1.4. CITATIONS AND REFERENCES

The function of citations can be regarded as a way to describe the context of a specific publication. Some references are more influential than others and will therefore appear more often. References could be collected from recorded full papers in a repository and would, for example, allow for cross-referencing. Having information available in full text thus allows for doing so, and such an effort would support content analysis and in this way provide an added value. An index of citations would open up for rankings: Who are, for example, the most influential authors? Or: Which are the most cited publications? Individual authors would also be able to trace explicitly who is "citing" their achievements and in what context.

In the case of the CUMINCAD-repository (http://cumincad.scix.net), references are collected for storage in a cumincadREFS-database. These records are linked to the original record-ids in CUMINCAD. Not all references are directly related to CAAD, but some are and it is necessary to consider entering these as "new" records in CUMINCAD. However, a certain number will reference other papers in the series of the CAAD conference proceedings, which are recorded already in CUMINCAD. The remaining entries will have to be sorted and looked through, taking into account that duplicates etc. will be given.

ResearchIndex (also known as *CiteSeer* / http://citeseer.org) is a scientific literature digital library that aims to improve the dissemination and feedback of scientific literature, and to provide improvements in functionality, usability, availability, cost, comprehensiveness, efficiency, and timeliness. ResearchIndex indexes pdf-papers on the Web, and provides a number of features:

- Similar documents: ResearchIndex shows the percentage of matching word indexes between documents.
- Full-text indexing: ResearchIndex indexes the full-text of the entire articles and citations. Full boolean, phrase and proximity search is supported.
- All cited documents: ResearchIndex computes citation statistics and related documents for all articles cited in the database, not just the indexed articles.
- Reference linking: As with many online publishers, ResearchIndex allows the user to browse the database using citation links.

Rather than creating just another digital library, ResearchIndex provides algorithms, techniques, and software that can be used in other digital libraries.

2. Strategies for Collection

The research aimed at taking stock of availability of digital data showed clearly that the annual CAAD conference proceedings from the year 1997 onwards are amongst the first ones that have been published in a digital format. Printed proceedings were still produced, and a parallel CD-Rom with the digital counterpart could easily be created. Although the starting point for some of these associations is marked in the early 1980's, associations founded later did not provide a set with complete digital data. Negotiations with these associations, represented by their councils or steering committees, showed that there is a strong interest and also some partial financial commitment to make this retrospectively available. Of course making this available to the membership of an association is providing a service and is also important for prestige. As soon as a critical mass of relevant records is being covered, the attractiveness of a repository is determined by the membership. It needs also to be remembered that this kind of non-profit-association depends very much on volunteers and the organization of the annual conference (as the no. 1 activity) takes a lot of the capacity, although much can be done to automate the processes and workflows of such activities on the back of an electronic repository. Participation of more (regional) associations in terms of feeding a repository with full text papers can lead to access on a mutual basis (contribution or distribution). In terms of making a repository visible and attractive, an ongoing liaison with these associations is of crucial interest as both materials (papers, etc.) as well as the users (target group) can be found in a repository. Although eCAADe took the lead in this matter, other CAAD associations followed soon as a careful description of the framework was communicated during the conferences.

Submission, archiving and distribution of electronic versions of theses and dissertations – so called ETD's: Electronic Thesis or Dissertation; "e-Dissertation" or "e-Thesis" - can all in all be regarded as a fruitful option of extension within a repository. Concerning the collection of e-theses, many institutional archives have been started with e-collections of dissertations and theses. For instance on one hand individual libraries have different ways to make these academic "products" visible, but the outcome is like a widespread, mosaic landscape of knowledge. On the other hand enterprises like *University Microfilms* provide of a large collection of dissertations and theses, but work on a commercial basis. In this respect the support of the CAAD-community - both writers and supervisors - is feasible. A direct contact with the author (or supervisor) is useful, but sometimes difficult to establish due to changes in a professional career.

The repository will grow in terms of credibility as support by a number of associations is given. This makes it attractive for individuals to have materials

recorded. In the course of an academic career, a number of publications will be created by individuals. However, publication channels, in which these academic "products" appear can be rather different and an individual researcher must take care of his personal bibliography. In this context, contributions from individual CAAD-users can be regarded as a potential input source for a repository at a later stage. With the support of the associations, contact to individuals can more easily be established, to input their publications, other than conference papers (which have been covered already). The attendance by scientists to conferences provides an ideal environment for "marketing" the repository and presenting an overview of further planned extensions.

2.1. ASPECTS OF INDIVIDUAL SUBMISSION

Cumulative Index of CAD (CUMINCAD – http://cumincad.scix.net) has approximately 1.100 registered users. This is a substantial number, and in order to offer new extensions this user-group could first of all be requested to input their individual expertise by submitting CAAD-related-papers (drawn from personal bibliographies). An important condition for every submission is the delivery of the corresponding full-papers in the form of pdf-files as well as English summaries.

As a repository is steadily growing, individual users will wish to be recorded with their bibliographies in the corresponding area. The inclusion of e-prints and preprints makes sense and the storage has to be arranged in different categories, so that users of these information packages are well informed about the status of a publication. In fact an individual user could use this repository environment all over the world, independently of any specific computer. Furthermore such a procedure could allow for the identification of misinterpretations in existing records. A repository that contains a critical mass of initial content would allow the creation of an index of authors (with email addresses) and thus support the establishment of direct contacts with authors for further (relevant) submissions.

2.2. WEB HARVESTING AND MINING

An important source of the "raw" paper materials are the author's web pages and institutional archives. The first are usually an excellent source for topic based archives like the ones addressed in SciX. A person usually works in one or a few related and therefore relevant fields. Internet searches that would not explicitly name the author of the paper would most likely rank such works low on the order of the results. It would therefore be very beneficial if such works could be brought into databases such as the ones planned by SciX. There are basically two possible methods to do so: (1) The author may submit a URL of a list of his works and these would then be copied

automatically into a standard archive. *CiteSEER*, for example, is using this strategy. The main problem with this approach is that it is difficult to correctly extract the metadata because the works are not presented in a standards compliant format. Several tools for harvesting such information are available. Perhaps the most famous is the Harvest system developed at the University of Arizona; (2) The author may be given some easy to use tools, so that his/her personal archive would comply to some standard, such as the OAI (Open Archives Initiative, www.openarchives.org). The main problem of this approach is the added overhead on the author's side and the need to use server side programming which is a much more complex task than simply placing works on-line.

In the SciX project, it is intended to address these issues in such a way, such that a low barrier tool for self archiving of works will be created. It will be OAI compliant and would be harvested into the central SciX index. Furthermore, works will be hosted on the SciX servers, and these could be easily incorporated in the author's web pages so that the authors would have satisfaction of running their own digital archive and having full control over it.

3. Workflow: Digitalization and Conversion of Metadata

The type of published material, in terms of quantity and quality, defines the starting point for considerations. In many cases only publications which were produced up to five years ago are available in a digital format. Paper publications have first of all to be scanned, unless a retype is envisaged. This step is characterized by the choice of resolution and possible editing of meta-information. The conversion to a general readable format (such as pdf) is not too labour intensive and can be handled in a batch. Although scanned text is interpretable to humans on the screen, for the machine it is just an image. Creation of searchable full text requires *Optical Character Recognition* (OCR), which leads to a certain percentage of interpretation errors. Depending on the result, elimination of these errors may be labour-intensive, and requires qualified personnel. Parts of the workflow are eligible for outsourcing, such as scanning. Regarding the growth of the DL-Market, private vendors will probably try to fill this niche.

Published materials, from which no data have been archived, can be digitised, but this first of all requires scanning to be done page by page. Presentation of the original layout, traced back to the original proceedings, may be considered. Already at this point of the working process it is possible to create a pdf-file of the scan. However, the information displayed is just an image. The additional step of performing Optical Character Recognition (OCR), as well as conversion to, for example, a word processor, leads to a situation in which – depending on the quality of the printed material – interpretation mistakes have to be corrected manually. Again, after this step,

pdf-files can be created, which consist of "text" and therefore a full-text-search is then possible. This can be regarded as a high added value and opens up various forms of content analysis. Search engines would be able to use the expressions found in the full texts and refer to this.

3.1. FINANCIAL IMPLICATIONS OF DIGITALIZATION OPTIONS

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 created (Table 1). 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 1.000 pages are regarded to be equivalent to 125 papers.

In case where individual files are split (for example corresponding with papers) according to a necessary file naming convention, up to two working hours have to be added (i.e. update calculation with another 0,10 Euro) to the calculated prices. Entry of 125 full text summaries into an excel-sheet for import into a database (variation A., C., D, E.) also requires at least another two working hours. More time consuming is a similar measure concerning the gathering and splitting up of references into four separate fields (authors, title, year, source), as the number of lines will easily exceed 1.000 (in average up to 10 citations per paper). Eight working hours must be estimated which would correspond to a cost of 0,40 Euro per page.

TABLE 1. Overview on variations in digitalization (assumed quantity of 1.000 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** paperbased **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 orginal 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** paperbased **format – Conversion to full text pdf**
	Estimation: 1.900 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** paperbased **format – Retype and conversion to full text pdf**
	Estimation: 1.600 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 2.000 characters per pages define the basis for estimation. A recreation of the original layout is not included here as this would require scanning and layouting.	
E.	**Source in** paperbased **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.	

In case the budget does not force a reduction of costs, the most complete version and "re-use" of digital data would require an investment of 1,9 Euro (Variation B.) plus content-based manipulations of the content (extraction of summaries and references with preparation of input files for a database / 0,50 Euro) at a total expenditure of 2,4 Euro per page.

In the case where a large number of papers must be made available and also a tight budget is provided, decision-makers may opt in the first stage for variation B, which includes scanning. Unless variation D is chosen, where the preliminary step of scanning is missing due to retyping, the expensive step of OCR – which leads to searchable full text - may be feasible at a later stage. Technological developments would provide improvements in the meanwhile.

Further combinations are imaginable such as, 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 (1.667 full texts, from which 559 scanned and 1.108 converted from digital source), the total investment is 8.200 Euro. However, nearly 90% of these costs are related to full digitalization (Variation C.) of 3.500 pages from eCAADe-proceedings.

3.2. CASE-STUDY: DIGITAL ECAADE-PROCEEDINGS

The council of this CAAD-association (http://www.ecaade.org) was confronted with a situation in which nearly two decades of annual conferences and a corresponding publication output were characterized as "grey literature". Therefore the decision was taken to carry out retrospective digitalization with full text, in the original layout. The Viennese company "Mediatecture" made an offer, which was based on a test scan of a sample book of proceedings. Around 3.500 pages were scanned, recognized and mistakes were manually eliminated. Finally, pdf-files were created. The price for this job is related to the quantity (3.500 pages) and breaks down at 1.9 Euro per page for the whole procedure. The results of the work are twofold: entry of the papers in an online-repository and creation of a CD-Rom with a collection of Digital Proceedings. The revenues from the CD-sales financed the whole project, but it has to be said, that a return of investments may take some time (in this case 2 years).

4. Conclusions

The research topic of *Digital Libraries* (DL) is currently of high interest and numerous projects are being conducted in this area. It is no surprise that libraries have been active in this field, as a retrospective digitalization of collections may serve a larger audience, more independently of time and space limitations. Also academic associations and other organizations have achieved remarkable results so far.

It has been observed, that CAAD conference proceedings - with the exception of CAAD futures - are rarely published by a professional publisher. Therefore, the information is neither entered into commercial indexes, nor is this sold commercially. Furthermore full texts are not broadly available as usually only conference attendees have copies. While this used to be negative in the past, the fact that nobody has strong interest in the copyright of these papers means that they can now become part of free on line libraries and are reaching a much wider audience than professionally published proceedings. This is a fact that needs to be increasingly taken into account by the conference organisers.

The situation concerning the dissemination and retrieval of dissertations and theses is similar to that of conference proceedings, except that commercial services to redistribute the original work are available (for example *University Microfilms)*. However, neither full access is given nor is the redistribution free of charge. In the sense of free electronic publishing, a dissemination of this type of "grey" research work requires support in terms of a "seamless digital repository". Journal papers cover an important part of complementary publication output and are relatively easy to find (no "grey" literature). However, access depends on subscription, and databases will seldom present the full-paper. The concept of individual submission of metadata by authors into a repository could work out well, as the community is already well organized (networked) and can look back at a "history" of longer than two decades.

The accumulation of both secured as well as roughly estimated numbers of all the sources described in this paper may lead to a repository (related to CAAD) consisting of 5.000 to 10.000 records. Annual growth presently amounts to approx. 500 publication entries. Taking into account that, for example, a conference paper provides an average of ten citations, the accumulation of these would result in a remarkable extension of the CUMINCAD-repository. And also the new entries may potentially be supplemented with full-texts and thus allow for another batch by performing the same procedure. If, for example, 2.000 conference papers are processed, then the rough listing will probably contain 20.000 references. A number of filtering processes, however, are required (doubles with main database records, multiple citations within the newly created listing, relevance

concerning CAAD, incomplete citation data, non-English reference, etc.). After these filtering steps the final number will be reduced to less than 20%. Filtering a listing of references as indicated, in terms of a search for dissertations and theses, could lead to an identification of university sites with "dissertation activity".

A description of variations aims at making the relationship between *costs* and *benefits* visible. Upon availability of budget, decisions can be made accordingly leading 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".

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 homepage of the SciX project is at http://www.scix.net/. 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, Proceedings, Association of Research Libraries. In *Proceedings of the 138th Annual Meeting*, Toronto, Ontario, May 2001 [http://www.arl.org/arl/proceedings/138/guedon.html]

Martens, Bob, Turk, Ziga. (2002). Digital CAADRIA-Proceedings: Retrospective Analysis of Content, in CAADRIA 2002 Conference Proceedings, pp. 23-30

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

Van Rijsbergen, C.J :1979, Information Retrieval, 2nd ed. London, Butterworth, 1979.