IST-2001-33127

SciX

Open, self organising repository for scientific information exchange

D5: Content Sources and Acquisition Techniques

Responsible author: Bob Martens (TUW)
Co-authors: Ziga Turk (LJU), Grahame Cooper(USAL)

Type: report
Access: public
Version: 1.0
Date: October 31th 2002

EXECUTIVE SUMMARY:

This document is Deliverable 05 of the SciX project and details the results from a period of investigation into the identification of content sources. Its purpose is to investigate the procedure concerning the harvesting of content sources, to be made available as initial content for a demonstration repository. 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 workpackage 2 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.

The identification of content sources for feeding a repository lead to a closer examination as well as selection of relevant and possible available materials, such as conference papers, theses and journal papers in the field of CAAD and Construction IT. It was furthermore investigated, what types of information are in principle available (bibliographical data, summaries, full text, etc.) A key issue was also concerned with the question, whether the information is already fully digitally stored. Was the whole production process executed on a digital track or has the material been digitised? Another important factor is concerned with the handling. Can a bunch of materials be handled at once or is a retrieval step by step or even one by one necessary?

Acquisition was primarily focussing on setting up contacts and liaisons with a view starting negotiations with several (international) associations. The analysis on content sources in detail has some overlap with the acquisition for the repository. In the stage of analysis not all necessary information may be accessible to get a clear picture. Preliminary contacts have to be established, in order to fill in information gaps and to get a clear overall view. On the basis of a completed analysis – as presented in this report – steps towards concrete acquisition can be made. In particular, copyright issues will in many cases be a difficult topic. The use of smart acquisition techniques ensures that manual work in the entering of the initial collection of papers will be minimised. Also, priorities will have to be set as some materials can be made easy or even immediately available and others require more effort. For this reason the workflow concerning digitisation and conversion has been studied, as different levels of output are feasible.

RELEASE HISTORY

date	changes
4.10.2002	draft outline
25.10.2002	final draft
31.10.2002	version 1.0 released

TABLE OF CONTENTS:

EXECUTIVE SUMMARY: 2
RELEASE HISTORY 3
TABLE OF CONTENTS: 4

1. CONTENT SOURCES: PROCEEDINGS, JOURNALS AND THESES 6

1.1 INTRODUCTION 6
1.2 CONFERENCE PROCEEDINGS 6
1.2.1 CIB – W78 (Global) 7
1.2.2 ECAADE (Europe) 7
1.2.3 ACADIA (North America) 7
1.2.4 SIGRADI (South America) 8
1.2.5 CAADRIA (South East Asia) 8
1.2.6 CAAD Futures foundation (Global) 9
1.3 THESES AND DISSERTATIONS 9
1.3.1 University Microfilms (UMI) 10
1.3.2 Dissertation.com 10
1.3.3 Access to Theses in University Libraries 11
1.4 CAAD-RELATED JOURNALS 14
1.4.1 ASCE / Journal of Computing in Civil Engineering 14
1.4.2 CHI: Conference on Human Factors and Computing Systems 14
1.4.3 CIDAC / Computer-Integrated Design and Construction 14
1.4.4 Computer-Aided Design 15
1.4.5 Computers & Graphics 16
1.4.6 Design Methods 16
1.4.7 Design Studies 16
1.4.8 IJCIT / The Int. Journal of Construction IT 17
1.4.9 International Journal of Human Computer Studies 17
1.4.10 ITCon 18
1.4.11 Journal of Architectural Engineering 18
1.4.12 Journal of Collaborative Computing 18
1.5 REFERENCES AS INPUT SOURCE 19
1.6 CONCLUSIONS ON CONTENT SOURCES 20

2. ACQUISITION AND SUBMISSION 22

2.1 INTRODUCTION 22
2.2 STRATEGY FOR COLLECTION: LIAISONS WITH ASSOCIATIONS 22
2.3 ASPECTS OF INDIVIDUAL SUBMISSION 23
2.4 WEB HARVESTING AND MINING 23
2.5 COPYRIGHT ISSUES 24
2.6 CONCLUSIONS ON ACQUISITION 25

3. WORKFLOW: DIGITISATION AND CONVERSION 26

3.1 INTRODUCTION 26
3.2 PDF-FORMAT 26
3.3 EXAMPLES OF DIGITISATION PROJECTS 27

3.4 FINANCIAL IMPLICATIONS AND COST/BENEFIT ANALYSIS 29
3.5 CONCLUSIONS ON VARIATIONS IN WORKFLOW 31

4. CONCLUDING REMARKS 33

1. CONTENT SOURCES: PROCEEDINGS, JOURNALS AND THESES

1.1 INTRODUCTION

First of all it has to be stated that the scope of content sources in this framework is being focussed on the field of *CAAD* (Computer Aided Architectural Design) and *Construction IT* (Information Technology). The wider context of architecture and building engineering – to which CAAD and Construction IT are related - would have been too general. CAAD can furthermore be regarded as a partial set within the broader CAD (Computer Aided Design). If not explicitly otherwise stated, *CAAD and Construction IT* are being dealt with in this report.

Conferences serve as an important starting point in this respect. It could be observed, that the corresponding proceedings were rarely published by a professional publisher. Therefore, the information was neither entered into commercial indexes, nor was this sold commercially. Furthermore full texts were not broadly available as usually only conference attendees had copies. The situation regarding the serving associations will first of all be more closely examined and different regional situations in this context will be presented.

The situation concerning the dissemination and retrieval of dissertations and theses is a similar one, except that commercial services to redistribute the original work are available. 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".

Furthermore, the situation concerning journals in which researchers publish their work will be more closely examined. Also the concept of individual submission 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 aim of all these efforts should not stop at the point of collecting and distributing, but also serves as a starting point for further research directed towards content analysis. A useful step in this direction will be described in the framework of gathering references as content source for a repository.

1.2 CONFERENCE PROCEEDINGS

In the 1980's first considerations regarding the implementation of CAAD as a research and teaching area occurred at many architecture schools, soon resulting in a need for exchange of ideas and experiences. Platforms such as ACADIA, CAADRIA, eCAADe 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. What happens to smaller editions is that they are safely stored away in the studies of participants of the conferences and university members, but rarely become available to the public. Library networks have so far not been systematically supplied with copies, as far as could be traced back.

1.2.1 CIB - W78 (GLOBAL)

CIB stands for "International Council for Building Research Studies and Documentation". Since 1953 CIB has been a forum for cooperation and a unifying force in construction worldwide, fostering innovation and the creation of workable solutions to technical, economic and social problems. It is organized into 41 active working commissions and 13 task groups. The scope of the Working commission 78 (W78 - http://w78.civil.auc.dk/) is to foster, encourage and promote research and development in the application of integrated IT throughout the life-cycle buildings and related facilities. It primarily relates to the integration and communication of data, information and knowledge in the building's life cycle.

CIB W78 has been providing a forum for research related to construction informatics. Since 1983 it has been hosting yearly workshops. Proceedings started to get published from the later 1980s onwards. It is estimated that overall about 900 papers were published. Digital versions of papers are available since 1996. Earlier works require scanning. Except for the 1993 workshop, the W78 holds the copyrights of the works, therefore no legal problems are expected in putting all that content on line.

SciX coordinator is a member of the W78 board. The project was introduced to the W78 during the 2002 meeting in Aarhus. Full support of the W78 to get the full history of the proceedings on-line was promised.

1.2.2 ECAADE (EUROPE)

ECAADE (Education and research in Computer Aided Architectural Design in Europe - http://www.ecaade.org) is a non-profit making association of institutions and individuals with a common interest in promoting good practice and sharing information in relation to the use of computers in research and education in architecture and related professions. ECAADE was founded in 1983. The association has an annual conference which is hosted by a different University each year.

From the 1997 conference onwards, digital data were made fully available to the community. Furthermore the proceedings in the period 1994-1996 were partly digitally archived and were supplied by previous conference chairs. The council of the association took the decision to produce a CD-Rom with digital proceedings starting with the first conference in 1983. The annual conference is for ECAADE the no1. activity and, after nearly two deceades of development, has become widely recognized. Proceedings in the period 1983-1993 have been fully scanned with OCR and were converted to pdf. This work has been financed by the association and was made available to SciX. The investments made for digitisation etc. were covered by the revenue of selling a CD-Rom. The inclusion into a SciX repository is focussing on retrieval of individual papers, which is another direction of use compared with the offline CD-Rom (collection of proceedings – the annual production defines the order).

1.2.3 ACADIA (NORTH AMERICA)

ACADIA (Association for CAD in Architecture - http://www.acadia.org) was formed in the early 1980's for the purpose of facilitating communication and critical thinking regarding the use of computers in architecture, planning and building science. A particular focus is education

and the software, hardware and pedagogy involved in education. The organization is also committed to the research and development of computer aides that enhance design creativity, rather than simply production, and that aim at contributing to the construction of humane physical environments.

The steering committee supports the inclusion of ACADIA-papers into a repository and a request for data showed that the 1989 and 1995 were archived as MSword- or Xpress-documents. These sets of data could be converted easily into pdf. The reason why other proceedings were not archived has to do with the fact that storage space was in earlier days much more expensive than it is nowadays. With the release of the book of proceedings the job was finished and possible republication was not considered. From 1998 on, the ACADIA-proceedings have been created as pdf-files. However, in the case of ACADIA-papers, the publication consent before the year 2000 did not include electronic publication (i.e. storage in a repository) and the association is facing a difficult job in terms of asking for countersigning consents for this from individual authors. The steering committee, represented by the president, will contact all individual authors to obtain new consents.

1.2.4 SIGRADI (SOUTH AMERICA)

SIGRADI (Sociedad Iberoamericana de Grafica Digital - http://www.sigradi.org) is the Iberoamerican Society of Digital Graphics and organizes an annual conference to discuss the last applications and possibilities of graphic technologies, with the participation of international specialists. SIGRADI conferences have been organized since 1997 in Buenos Aires, Mar del Plata (1998), Montevideo (1999), Rio de Janeiro (2000) and Concepcion (2001).

As SiGraDi is a “younger” association, the situation concerning complete archival of data seems to have a better chance. However, a search learned that the first proceedings (1997) were still partly available as html-files and the remaining papers are asking for a scan. The situation for 1999 is somewhat different, as only plain text files could be retrieved from the archive (and some contributions are missing) – a branch of loose, non-formatted files, which had to be linked to the corresponding pages. The 1998 proceedings were converted by the association and presented according to the original layout. From 2000 on, a CD-rom with conference proceedings in pdf-format was produced and part of the conference package. Although there is a multilingual approach (English, Spanish, and Portuguese), the aim is to have an English summary and title as a part of the final paper. In this respect a number of titles will be translated by the Association. Furthermore the difficult economic circumstances in South America do not allow for big investments and the participation in a repository is therefore positively seen.

1.2.5 CAADRIA (SOUTH EAST ASIA)

CAADRIA (Computer Aided Architectural Design Research in Architecture - http://www.caadria.org) is an association of those who teach and conduct research in computer-aided architectural design in schools of architecture throughout Asia. It was established on 26 April 1996. The first conference was held in Hong Kong in 1996, the second one in Hsinchu, Taiwan, 1997, the third one in Osaka, Japan, 1998, the fourth one in Shanghai, China, 1999, the fifth in Singapore in 2000, the sixth in Sydney 2001 and the seventh in Cyberjaya, Malaysia 2002.

Concerning their track record, CAADRIA has a similar background as SiGraDi. The board of CAADRIA made all previous proceedings available as pdf-files and welcomes the participation in a repository. For the moment the 1996 and 1997 proceedings are non-text-pdf-files (only scanned as image) and therefore not searchable in full text mode. However, this may be adjusted in the near future.

1.2.6 CAAD FUTURES FOUNDATION (GLOBAL)

CAAD Futures (http://www.caadfutures.arch.tue.nl/) was set up under Dutch law in 1985 with three founding members; Tom Maver, Rik Schijf, and Harry Wagter with the purpose of promoting, through international conferences and publications, the advancement of Computer Aided Architectural Design in the service of those concerned with the quality of the built environment. CAADfutures runs an international conference biennially (Delft - 1985, Eindhoven - 1987, Boston - 1989, Zurich - 1991, Pittsburgh - 1993, Singapore - 1995, Munich - 1997, Atlanta - 1999, Eindhoven - 2001).

The case of CAAD futures is somewhat different as this foundation released its proceedings in cooperation with a publisher. However, for the period 1986-1997, six different publishing houses hold the copyright; from 1999 on an agreement was made with the same publishing house (the 2001 conference provided delegates with a CD-Rom for personal use). Negotiations concerning electronic distribution by the foundation with all of these publishers were rather time-consuming, but in the end successful, as the copyright issue did not block digitisation. In fact the “out of stock”-situation made it easier to reach a positive agreement. Currently investigations are performed concerning availability of digital data (good chances for 1997 and 1999) in order to avoid investments for scanning; the quality of pdf-files which are created from the original data is better, as no analogue step is involved. Furthermore the pdf-files are reasonably smaller. The matter of financing the efforts of digitisation is pending. However, refinancing by means of selling a CD-Rom with a collection of digital proceedings might serve as a working solution (see experiences of eCAADe).

1.3 THESES AND DISSERTATIONS

A thesis or a dissertation is by its very nature produced in very limited quantities and in many cases the only copy available is the archival copy deposited in the library. In the course of a request, cost and delay factors are a significant deterrent. The lack of usage is attributed to a number of factors, such as knowledge that the thesis exists, the contents of the thesis or 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 web would greatly increase knowledge, access and availability of such a significant resource. Most students 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 will 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.

For instance on one hand enterprises like University Microfilms provide of a large collection of dissertations and theses, but work on a commercial basis. On the other hand individual libraries have different ways to make these academic “products” visible, but the outcome is like a widespread, mosaic landscape of knowledge. The collection of conference proceedings (see chapter 2), however, points by means of the references (see chapter 6) to a number of university sites with a remarkable output. In this respect the support of the CAAD-community - both writers and supervisors - is feasible. 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.

A worldwide variety in educational systems has to be observed and e.g. the “product” thesis can represent at two different university sites in fact a different level of output. In this framework the gathering of theses and dissertations is aimed at, which lead to a doctoral degree. Diploma work etc. – even named “thesis” - is not to be included.

1.3.1 UNIVERSITY MICROFILMS (UMI)

University Microfilms (http://wwwlib.umi.com) first opened its doors in 1938. Soon afterwards began gathering, indexing, filming, and republishing doctoral dissertations in microform and print. By 2000, the UMI Dissertation Abstracts database archived over 1.6 million dissertations and master's theses. Some one million of them are available in full text in print, microform, and digital format. The database includes citations for materials ranging from the first U.S. dissertation, accepted in 1861, to those accepted as recently as last semester. Graduate students customarily consult the database to make sure their proposed thesis or dissertation topics have not already been written about. Students, faculty, and other researchers search it for titles related to their scholarly interests.

Although this service is heavily focussing on the US, also materials from outside this area are recorded. The web interface is easy to use and bibliographical data from 1997 on is freely available. Also summaries are included and a sample of 24 pages can be downloaded in pdf-format. This preview supports the personal selection mechanism. Having made a choice, the thesis can be ordered as a printed version or retrieve a full pdf-file. The download of the electronic version does not cause delay. The subject « architecture » delivers at the time of writing 304 entries.

1.3.2 DISSERTATION.COM

Academic Dissertation Publishers (http://www.dissertation.com) – the company behind dissertation.com - can also be regarded as a commercial enterprise in the area of providing students, researchers, and the general public with low cost access to academic work. Publications are made easily accessible on-line and through thousands of booksellers, reducing the cost of acquisition and speeding delivery to those interested. The ISBN number found on the reverse of every book is the key to commercial book distribution. Without an ISBN number, most bookstores do not have access to it. Remarkably, theses and dissertations are rarely assigned this important identification number. Consequently, they are completely isolated from the mainstream book trade. The dissertations within « dissertation.com » are published in the

form of of an e-book. and can be printed on demand as paperbacks. The corresponding internet-site provides bibliographical data, summaries and samples of 25 pages for free review. Besides this, also an email-adress or url of the author is displayed. However, the search options are rather limited (author, title) and an advanced search option is missing. Unfortunately no indication appears concerning the total number of stored e-books. A search for « architecture » only delivered three entries. Download of an electronic file is the cheapest (and quickest) solution. Authors pay a fee for getting their work entered in dissertation.com, but are given royalties for each copy sold (between 20 and 40% from the revenues).

1.3.3 ACCESS TO THESES IN UNIVERSITY LIBRARIES

The concept of Electronic Theses and Dissertations (ETD's) was first openly discussed at a 1987 meeting arranged by the University of Michigan. As a follow-up, Virginia Tech funded development of the first SGML Document Type Definition (DTD) for this purpose. Since 1994, the short term solution at Virginia Tech has been for students to submit their documents as pdf-files. If the ETD passes published quality requirements, the library catalogs the ETD and places it on the electronic bookshelf for ETD's, which supports flexible browsing. A simple search engine facilitates access, and will be replaced by a more powerful system when the number of documents warrants it.

NDLTD, an acronym for *Networked Digital Library of Theses and Dissertations* (http://www.ndltd.org, http://www.theses.org), comprises many individual member institutions and consortia, each of which has a process in place for archiving and distribution of ETD's. The Union Catalog Project is an attempt to make these individual collections appear as one seamless digital library of ETD's to students and researchers seeking out theses and dissertations. Among the participating sites are:

- M.I.T.
- National Documentation Centre (NDC), Greece
- National Sun Yat-sen University in Taiwan
- North Carolina State University
- University of Florida
- University of Hong Kong
- University of Iowa
- University of Kentucky
- University of Michigan
- University of North Texas
- University of Stuttgart
- University of Texas at Austin
- University of Virginia
- Universitat Politecnica de Valencia
- University of Waterloo
- Uppsala University
- Virginia Tech
- West Virginia University

While ETD's are owned and maintained by the institutions at which they were produced or archived, it is possible to give searchers the appearance of a single collection by gathering all the metadata (title, author, etc.) into a central search engine. Then, when a potentially relevant document is found, the user will be redirected to the institution that contains the actual document. This approach of making metadata available to aid in discovery of resources is supported by the Open Archives Initiative (OAI). Using the OAI's Protocol for Metadata Harvesting, individual sites can make their metadata accessible to providers of search and discovery services, while still maintaining complete control over the resources.

The *Australian Digital Thesis Program* (ADT - http://adt.caul.edu.au/) is linked with NDLTD. Approximately 4,000 degrees are awarded each year in Australia. However, lack of easy access to this information means other researchers can wait months or years before papers or books describing aspects of the research are published. These publications do not always comprehensively cover the valuable information in a thesis; information which in many cases is then effectively lost. The ADT-program has two major components: digitisation of theses as part of the deposit process and the digitisation of a selected number of frequently requested existing theses. As each university is responsible for maintaining an archival copy of the theses of their own institution, each participant in the program will mount his or her own thesis on a server located in their respective institution. The participating universities use the same database configuration, standards and metadata system to ensure compatibility:

- Adelaide University
- Australian National University
- Curtin University of Technology
- Central Queensland University
- Griffith University
- Queensland University of Technology
- University of Melbourne
- University of New South Wales
- University of Queensland
- University of Sydney
- University of Wollongong
- Victoria University of Technology

The situation in Europe concerning repositories with dissertations and theses is lacks any similar (central) "Union Catalog", from which this kind of academic work can be searched. In fact the user has to visit individual university libraries one by one and to get acquainted with a variety of interfaces. This is time-consuming and makes sense only if a specific piece of work has to be retrieved (i.e. university and author are known). In this case "younger" dissertations and theses especially have a good chance to be already available in pdf-format (as digital data may be archived), and can be retrieved for free. The following selected examples – in alphabetical order - give an impression about a situation with rich diversity:

- Austrian Dissertations Database (http://www.arcs.ac.at/dissdb/welcome)
  This database is set up and maintained by the Austrian Research Centers (ARC / Seibersdorf). It comprises all Austrian dissertations from 1970 onwards. The database is built in cooperation with the university libraries, the dean´s offices and examinations

offices of the Austrian universities (from 1999 onwards by an WWW user-interface). ARC performs subject analysis and descriptive cataloguing; data are provided in the standard of international information systems. Around 80 dissertations in the subject of architecture are offered (unfortunately no full texts).

- Delft University of Technology (http://www.library.tudelft.nl/dissertations/)
  At the time of writing three dissertations in the field of architecture are being offered as full text, which can be retrieved for free.
- Dissertationen Online (http://www.educat.hu-berlin.de/diss_online/biblio.html)
  This site contains a summarization of german initiatives concerning ETD's. Unfortunately the last update took place during summer 2000 and therefore a number of broken links is the result. However, as a rough step-in this page may still be useful.
- Eindhoven University of Technology (http://www.tue.nl/bib/ftproefelders.html)
  This university offers a link-page with an overview to other university libraries with e-dissertations. Therefore, this page is a useful starting point. Furthermore this university has made all dissertations (created at Eindhoven University of Technology) available in pdf-format, and these can be downloaded for free. The faculty of architecture for example provides 90 doctoral dissertations in full text as free e-documents (from 1971 on). Dissertations can also be searched by author's name, title and subject in the "VubisWeb" library catalogue.
- ETH-Zurich (http://e-collection.ethbib.ethz.ch/diss/)
  This site offers a comprehensive collection of dissertations, which were defended at ETH Zurich. Currently nearly 80 e-dissertations are available in pdf-format and can be downloaded for free.
- Lund University Dissertation Abstracts (http://www.lub.lu.se/cgi-bin/search_diss.pl?faculty=Technology)
  The interface consists of a listing, in which it is not possible to enter a search term. This means that the user has to know exactly what he is looking for. However, this may also be practical for internal use.
- RWTH Aachen (http://www.bth.rwth-aachen.de/ediss/ediss.html)
  Within the faculty of architecture eight e-dissertations can be retrieved at the moment (period: 1998-2002). The full texts are made available in pdf-format for free.
- Theses in GB and Ireland (http://www.theses.com/)
  This database covers accepted theses from 1970 to 2001, covering all of volumes 21 to 50 and parts 1 to 5 of volume 51 of the equivalent print publication "Index to Theses". The interface does not offer full texts and registration is obligatory.

These examples show, that a unifying catalog with metadata, in which a large number of dissertations and theses is being recorded (at least by means of bibliographical data extended with a summary) would make sense. The availability of full text ("e-diss") at the local university is also preferable, but as soon as the user knows what he is looking for, a contact to a specific author could probably be easily established with the support of the university.

Looking in the direction of the middle-east region database-portals could not be traced back. In the far-eastern region the barrier of non-english-language (resulting in missing partial translation) stops further research at this moment.

1.4 CAAD-RELATED JOURNALS

The papers in the conference proceedings contain citations from a number of journals, which are to be closer examined in this chapter. In the framework of the a meeting of the SciX-consortium suggestions for a possible selection of journals were made. As all journals display themselves on the web, the browsing was relatively easy. The observations will be presented in alphabetical order.

1.4.1 ASCE / JOURNAL OF COMPUTING IN CIVIL ENGINEERING

http://ojps.aip.org/journal_cgi

Advances in computing continue to strongly influence our profession. New methods and tools are emerging that enable civil engineers to use computing in creative and imaginative ways. To realize the full potential of these advances, civil engineers must understand fundamental issues in computing education, research, and professional practice. The goal of this journal is to serve as a resource for emerging ideas in civil engineering computing. Subject areas include software such as new programming languages, database management systems, computer-aided design systems, and knowledge-based expert systems; hardware for robotics, bar coding, remote sensing, and data and knowledge acquisition; and strategic issues such as the management of computing resources, implementation strategies, and organizational impacts. The journal is intended to be of interest to students, researchers, and professionals in all disciplines of civil engineering.

Period of examination 1995 - 2002

1.4.2 CHI: CONFERENCE ON HUMAN FACTORS AND COMPUTING SYSTEMS

http://portal.acm.org/results.cfm?coll=Portal&dl=Portal&CFID=3585679&CFTOKEN=80461420 - http://www.sigchi.org/

The annual CHI conference is the leading international forum for the exchange of ideas and information about computer-human interaction (CHI), also known as human-computer interaction (HCI).

Period of examination 1995 - 2001

1.4.3 CIDAC / COMPUTER-INTEGRATED DESIGN AND CONSTRUCTION

http://www.lboro.ac.uk/cidac/papercontributions.htm

The International Journal of Computer-Integrated Design and Construction (CIDAC) is intended to provide a forum for the dissemination of information related to the use of computers and associated technologies in the integration of the design and construction processes. The journal publishes both original research papers as well as practical papers on aspects of computer-integrated design and construction. Papers on theoretical, industrial and computing developments which have a bearing on computer-integrated design and construction will also be published. The scope of the journal is wide and includes the following and related topics:

- computer-aided design

- computer-integrated construction
- concurrent engineering in construction
- computer-integration of design activities
- computer-aided cost planning and control
- computer-aided construction process planning and scheduling
- information management in integrated design and construction
- organizational/human issues in integrated design and construction
- intelligent systems in integrated design and construction
- life-cycle design of facilities
- computer-integrated facilities management
- computer-aided construction site layout design
- computer-aided construction safety management
- communication issues in integrated design and construction
- product and process modelling

All citations have been examined (1999-2001)

1.4.4 COMPUTER-AIDED DESIGN

http://www.elsevier.nl/gej-ng/10/15/39/show/toc.htt?year=1996

Computer-Aided Design is an established international journal that provides engineers, designers and computer scientists in academia and industry with key papers on research and developments in the application of computers to the design process. Computer-Aided Design invites papers reporting new research and novel or particularly significant applications within a wide range of topics, including:

- CAD in conceptual design
- Design automation and optimization
- AI in design
- Geometric methods and applied computational geometry
- Surface and solid modelling
- Parametric, constraint-based, and feature modelling
- CAD interfaces to testing and analysis, including finite-element methods
- Design and planning for manufacturing, including numerical control, rapid prototyping and robotics
- Design and planning for assembly, maintainability, recycling etc
- Engineering data management and exchange, including design databases, component selection, product models, and life-cycle modelling
- Space and facilities planning and layout
- CAD user interfaces, including computer graphics, virtual and augmented reality
- Significant benchmarks, APIs, formats and standards in CAD

Extended information is available, partly with summary and email-contact (1996-2002)

1.4.5 COMPUTERS & GRAPHICS

(International Journal of Systems & Applications in Computer Graphics)

http://www.elsevier.com/inca/publications/store/3/7/1/

Computers & Graphics is dedicated to the dissemination of information on the application and use of computer graphics techniques. The journal encourages articles on: 1. Research and applications of computer graphics. Emphasis will be placed on graphical man/machine interaction and the application of graphics to problem solving. 2. Tutorial papers in the area of computer graphics. 3. State-of-the-art papers on various aspects of computer graphics. 4. Information on innovative uses of various graphics devices and systems.

Computers & Graphics provides a medium for the communication of information concerning graphical man/machine interaction and the applications of computer graphics. The emphasis of the journal is on interactive computer graphics using CRT-type consoles and manual input devices such as light-pens, tablets, and function keyboards, and, within this scope, on graphical models, data structures, attention-handling languages, picture manipulation algorithms and related software. The Editor welcomes papers dealing with applications of current interest, including, but not limited to, computer-aided design, management information systems, simulation, process control, computer-aided education, pattern recognition, graphic arts, computer generated movies, medical research, architectural design, transportation systems, the design of integrated circuits, graphic operating systems, display techniques, graphic system design and evaluation (including hardware), graphic programming languages, interactive languages, man-machine communication techniques, and mathematical problem solving.

No digital information is available on the website.

1.4.6 DESIGN METHODS

http://www.ericae.net

Links to some of the best full-text books, reports, journal articles, newsletter articles and papers on the Internet that address educational measurement, evaluation and learning theory are available on this website. The selection of documents is based upon criteria that are widely accepted in the library and information science community and provides a framework to browse these resources.

More than 200 entries could be retrieved and were examined, but there is no relation towards CAAD.

1.4.7 DESIGN STUDIES

http://www.elsevier.nl/inca/publications/store/3/0/4/0/9/

Design Studies provides a unique forum for the discussion and development of the theoretical aspects of design, including its methodology and values. It is the only journal to approach an understanding of design from comparisons of its applications in all areas, including engineering, architecture, planning and industrial design. As the concept of design becomes increasingly important, it is vital for researchers, educators and practising designers to stay abreast of the

latest research and new ideas in this rapidly growing field; with its truly interdisciplinary coverage, Design Studies meets these needs with maximum effect. The journal reports on new developments, techniques, knowledge and applications in the practice of design, as well as design education: how design techniques may be taught, the approach to ill-defined problems and the impact of new technologies. Coverage includes design management, design methods, participation in planning and design, design education, AI and computer aids in design, design in engineering, theoretical aspects of design, design in architecture, design and manufacturing, innovation in industry and design and society. Design Studies is published in co-operation with the Design Research Society (http://www.drs.org.uk/)

Contains a large number of interesting contributions. However, abstracts are not available online (1995-2001).

1.4.8 IJCIT / THE INT. JOURNAL OF CONSTRUCTION IT

http://www.scpm.salford.ac.uk/ijcit/homepage.html

The Journal aims to provide researchers in academic institutions, as well as professionals in private practice and commercial organisations, with high quality technical papers on current developments in the rapidly changing environment of Information Technology in the Construction Industry on a Worldwide basis. It disseminates research results and communicates new practical ideas, applications and developments to construction professionals to maintain their competitive edge. The journal covers basic and applied research, practical developments and case studies in areas related to design and construction.

Available records were examined (1999-2001) - This journal does not exist any longer.

1.4.9 INTERNATIONAL JOURNAL OF HUMAN COMPUTER STUDIES

http://www.hcibib.org

The International Journal of Human-Computer Studies publishes original research over the whole spectrum of work on both the theory and practice of human-computer interaction and the human-machine interface. The journal covers the boundaries between computing and artificial intelligence, psychology, linguistics, mathematics, engineering, and social organization. Research Areas include:

- Intelligent user interfaces
- Natural language interaction
- Speech interaction
- Graphical interaction
- Innovative interaction techniques
- Expert systems
- User models and models of users
- Empirical studies of user behaviour
- The psychology of programming and programmer performance
- Program comprehension and debugging
- User interface prototyping and management systems
- Interface design and evaluation methodologies

- Intelligent tutoring and coaching systems
- Problem-solving, organization, and communication
- Systems theory
- Information and decision support systems
- Innovative designs and applications of interactive systems with critical evaluation
- Relevant theoretical and practical advances in supporting disciplines
- Virtual reality
- Multimedia

Contains a large number of entries (23.500), from which a smaller number was filtered with keywords, such as Architecture, CAAD, Design und Education. All in all 518 records were closer examined.

1.4.10 ITCON

http://www.itcon.org

Founded in 1995, the Electronic Journal of Information Technology in Construction is a peer-reviewed scholarly journal on the use of IT in construction. Articles are submitted and published electronically. Biannually, a limited number of copies is printed as well. The Journal is committed to minimizing publication delays, and to promoting maximum flexibility in the ways that readers use the journal for teaching, research, and scholarship. Readers' license is limited only as required to insure fair attribution to authors and the journal, and to prohibit use in a competing commercial publication

All citations have been examined. Full papers in pdf- and (some) in html-format available.

1.4.11 JOURNAL OF ARCHITECTURAL ENGINEERING

http://ojps.aip.org/journal_cgi/dbt?KEY=JAEIED

The Journal of Architectural Engineering will provide a multidisciplinary forum for dissemination of practice-based information on the engineering and technical issues concerning all aspects of building design. Peer-reviewed papers and case studies will address issues and topics related to buildings such as planning and financing, analysis and design, construction and maintenance, codes applications and interpretations, conversion and renovation, and preservation.

Period of examination 1995-2002

1.4.12 JOURNAL OF COLLABORATIVE COMPUTING

http://www.kluweronline.com/issn/0925-9724

The Journal of Collaborative Computing is devoted to innovative research in Computer Supported Cooperative Work (CSCW). It provides an interdisciplinary forum for the debate and exchange of ideas concerning theoretical, practical, technical, and social issues in CSCW. The journal arose as a response to the growing interest in the design, implementation and use of

technical systems (including computing, information, and communications technologies) which support people working cooperatively. The scope of the CSCW journal remains to encompass the multifarious aspects of research within CSCW and related areas – from ethnographic studies of cooperative work to reports of the development of CSCW systems and their technological foundations.

Website contains valuable information and has been examined (1997-2002). Pdf-files of full papers are also available.

1.5 REFERENCES AS INPUT SOURCE

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 the 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 define an added value. An index of citations would open up for rankings: Who are, for example, the most influential authors? Or: What are the most cited publications? Individual authors would also be able to trace back, who is "citing" - resp. in which context - their achievements explicitly.

As soon as papers in a digital format are being collected, references can more easily be extracted. The relationship to the paper in which the citation was mentioned, should be represented by means of the original paper_id. This link is of crucial importance. Furthermore a splitting up of each line has to be performed into different fields: author(s), year, title and source. However manual rearrangement is unavoidable. The easiest field is doubtless the field "year", which consists of 4 characters. The "authors" field requires consistence in terms of a coherent order and syntax. Concerning title and source, the citation is to be taken as it is and unnecessary characters may be deleted.

In the case of the CUMINCAD-repository, an acronym for Cumulative Index of CAD (http://cumincad.scix.net), references can be 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" record 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 duplicates can be filtered easily, by means of a search in the field "source". Also a search on "thesis" would deliver a number of entries for a cumincadTHESIS database (see also chapter 2.2). The remaining entries will have to be sorted and looked through, taking into account that duplicates etc. will be given. This procedure, furthermore requires appropriate utilities within the database-system, which preferably supports recognition of possible duplicates.

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 sentences 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 browsing 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. The full source code of ResearchIndex is available at no cost for non-commercial use.

1.6 CONCLUSIONS ON CONTENT SOURCES

The creation of CUMINCAD in 1998 based on the insight, that no (web-based) repository existed to support further dissemination of research work in the field of CAAD and Construction IT. The main channel of exchange has been in the format of annual conference meetings with proceedings as a tangible result. More than 70 conferences were organized in the past. In total about 3.200 conference papers could be traced back in the period 1981-2001 by the associations mentioned in this report. It has to be said, that the level of professionalism has significantly upgraded in the last five years. Also the annual production of papers has been raised substantially and this together with a more rigorous blind reviewing procedure. In 2001, 410 papers were published in conference proceedings. However, this seems to be a number which will not grow in the near future. The current extension on a annual basis - without the biannual CAAD futures - is approx. 350 papers. CUMINCAD, which developed on a shoestring budget, supports the important task of information management, as no other, similar initiative could be identified so far for this field of research. The associations in charge are established on a non-profit base and have no commercial interest in predominantly “making money” from the conference activities.

Concerning the “production” of CAAD- and Construction-IT-related dissertations, no exact figures can be delivered as is the case, for example, with conference proceedings. A closer look at the full collection of the Faculty of Architecture at Eindhoven University of Technology – which provides of a complete set of doctoral e-dissertations (since 1974: 90) - shows that a smaller number (4) focus on topics related to CAAD and Construction IT. One could use the number of architectural education sites in europe (approximately 250) as a indicator. However, not every location has for various reasons the same research output in terms of defended doctoral dissertations. The total number architectural education sites does in a worldwide perspective probably not exceed the number of 1.000. The same number may possibly apply to individual researchers, who are explicitly dedicating their work too CAAD and Construction IT.

Especially in Asia a remarkable number of PhD-students are currently working on dissertations (this can be observed from the submission entries for review in annual conferences). An estimation for the Nordic countries in Europe concludes that around 15 PhD theses in Construction IT and CAAD were defended since 1995. A worldwide figure of 50 dissertations per year and 500 all together since around 1990 seems to be reasonable. Further analysis on

both the dissertation indexes as well as references from conference papers may assist the collection of secured data regarding dissertations. Filtering a listing of references as indicated in chapter 1.5, in term of a search for dissertations and theses, could lead to an identification of university sites with “dissertation activity”.

The listing of journals in chapter 1.4 covers an important part of complementary publication output. However, access depends on subscription and databases will seldom present the full-paper. The bibliographical data may be extended with a summary and keywords. 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. Web-based databases avoid a retype, but attention has to be paid to the re-entry of information into fields for data import. An exact number of possible entries cannot be stated here, but the journals as such are relatively easy to find (no “grey” literature). The recording of at least metadata in a repository frees individual researchers from the need to visit corresponding individual journal sites on the web.

Taking into account that a conference paper has, on average, up to ten citations, the accumulation of these would allow for a remarkable extension of the repository. And also the new entries may potentially be extendable with full-texts, and thus allow for another batch by performing the same procedure. If, for example 2.000 conference papers are processed, than the rough listing will probably contain 20.000 references. However, a couple of filtering processes must take place (duplicates with main database records, multiple citations within the newly created listing, relevance towards CAAD and Construction IT, incomplete citation data, non-english reference, etc.). The final number will after these steps of filtering be reduced to less than 10 %.

The accumulation of both secured as well as roughly estimated numbers of all these sources may lead to a repository consisting of 5.000 to 10.000 records. Annual growth can be considered nowadays to be around 500 publication entries.

2. ACQUISITION AND SUBMISSION

2.1 INTRODUCTION

In chapter 1 the framework concerning content sources was elaborated, both in quantitative as well as in qualitative terms. Filling a repository with bibliographical data and probably extended information such as summary, keywords and email address of the authors may be of use. However, a substantial part of the sources (especially conference papers and dissertations) can be characterized as so "grey literature", i.e. difficult to acquire as full text for instance by means of interlibrary loan. Furthermore this procedure can take weeks or even months, as the request is being sent from one library to the other. Nowadays users tend to switch their search in such a case towards other sources – which are probably immediately available – and therefore the online repository should try to realize this as well.

2.2 STRATEGY FOR COLLECTION: LIAISONS WITH ASSOCIATIONS

The research for taking stock of availability of digital data showed clearly that the annual proceedings from the year 1997 on are, in most cases, 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.

The managers of a repository will have to work on the setting up of a stable liaison with the executive committees of the associations in charge, as the right to enter full text versions of conference papers has to be granted. Of course making this available to the membership of an association is providing service and also important for the 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. 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 case of the CUMINCAD-repository, it could be observed that finally all associations as described in chapter 1.2 took formal decisions to participate.

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 the conferences –provides an ideal environment for “marketing” the repository and to present an outlook toward further planned extensions.

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, the other associations followed soon as a careful description of the framework was communicated during the conferences.

2.3 ASPECTS OF INDIVIDUAL SUBMISSION

For the extended use of the *Cumulative Index of CAD* (CUMINCAD – http://cumincad.scix.net) approx. 800 registered users are recorded (so called "friends"). This is a substantial number and in order to offer new extensions this user-group could first of all be requested to input "human expertise" by means of submitting CAAD-related-papers (resp. personal bibliographies). The condition for every submittal would be the delivery of the corresponding full-papers in form of pdf-files as well as english summaries. A similar request could be sent out to the resp. CAAD-Listservers, which are being served by CAAD-Associations, such as eCAADe, ACADIA, CAADRIA and SiGraDi. This procedure would possible enlarge the user group.

As a repository is steadily growing, individual users are interested 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 and this independently from a specific computer. Furthermore such a procedure could allow for the identification of misinterpretations in existing records. A repository which contains a critical mass of initial content would allow for creating an index of authors (with email addresses) and thus support the establishment of direct contacts with authors for further (relevant) submissions.

2.4 WEB HARVESTING AND MINING

An important source of the raw materials of the papers are the author's Web pages and institutional archives. The firsts 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 be therefore very beneficial if such works could be brought into the databases such as the ones planned by SciX.

There are basically two possible methods to do so:

- The author may submit a URL of a list of his works and that 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.

- The author may be given some easy to use tools, so that his/hers personal archive would comply to some standard, such as the OAI. 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, we intend to address these issues in these ways:

- Provide a low barrier tool for self archiving of works, that will be OAI compliant and would be harvested into the central SciX index.
- Provide hosting of works on the SciX servers in a way that could be easily incorporated in the author's web pages so that the authors would have satisfaction of running their own digital archive and have full control over it.

2.5 COPYRIGHT ISSUES

Concerning copyright issues it has to be said that the associations do not have a strong commercial interest in terms of selling conference proceedings (non-profit status, charity). On the contrary there is a belief that indexing and availability can be regarded as an add-on for the association. The decision-making line is relatively short as long as the association holds the copyright. This is not the case with CAAD futures, where traditionally well-known publishing houses took over the publication rights. Recent negotiations showed that in case, where conference proceedings are out of stock, electronic re-publication is not blocked by the copyright. Furthermore a large part of the stock has been sold to the CAAD futures association on the occasion of the corresponding conference. This means that the basic costs have been recovered and no investments had to be made by the publishing house. In the case of ACADIA the use of publication consent did not include electronic publication in the past and therefore, agreement needs to be received by the individual authors in charge. This is time consuming and an announcement in internal channels (listserver, quarterly, etc.) may be helpful. So far in SciX we cannot report on the outcome of the negotiations with the commercial publishers.

Copyright is a major issue when the works that should be digitally archives have been published by a commercial published and not directly by the professional organisation. One can clearly feel that in the first case there is a primary interest in selling the "book" while in the latter is interested in broad and wide dissemination. Unfortunately, through their contacts with the libraries, the commercial publishers are in a much better position to put the books into the libraries than independent publishers or the professional societies.

The business process analysis should address these issues and compare the benefits and down-sides of independent vs. professional publishing and provide a summary for conference publishers on how to make the best deal out of it. Somewhere in the middle are the publication services that can do the technical work on the publication of the proceedings but do not claim the copyright. Another thing worth addressing are the copyright policies of the publishers. Since some organisations, for example US and Canadian government organisations, do not, in principle, give away the copyright to scientific work, a special copyright agreement is available for them. This agreement is hardly available for universities at large. This is what a general publication agreement says:

*"The Author hereby **assigns to the Publisher the copyright** to the Contribution named above whereby the Publisher shall have the exclusive right to publish the said Contribution in print*

and in electronic form and translations of it wholly or in part throughout the World during the full term of copyright including renewals and extensions and all subsidiary rights."

The "under the table" agreement available to some only writes:

*"The Author **hereby retains the copyright** to the Contribution named above whereby he shall have the exclusive right to publish the said Contribution in print and in electronic form and translations of it wholly or in part throughout the World during the full term of copyright including renewals and extensions and all subsidiary rights."*

Perhaps the governments in the EU should pass a similar directive and therefore enable these other kinds of copyright agreements to be generally used. They would allow for electronic publication of works. Some limitations could be agreed upon, for example some lag between electronic and paper publication, downgraded image quality or a mandatory pointer to a page where the book could be purchased.

2.6 CONCLUSIONS ON ACQUISITION

In chapter 2.2, it was stated that eCAADe took the lead in the support of a repository and indeed nearly 900 full text papers were made available. Also CAADRIA and SiGraDi each contributed more than 300 papers. ACADIA has so far provided 80 papers and another 300 will follow as soon as the publication consents for electronic republication are collected. In the same way the contribution from CAAD futures can be characterized (60 papers), but 320 papers are waiting for digitisation (copyright issue is cleared here already). This means that 1.700 full text papers are collected and 600 are in the queue for processing (awaiting copyright or digitalisation). The bibliographical data or summaries are available as database entries for all of these publications. From the year 2002 onwards around 300 to 350 full text papers will be contributed, which is promising as this is about 2/3 of the whole estimated annual production. In the initial stage a repository will probably not provide more than a third of the records with an attached full text, but taking into account these figures of growth, in a couple of years this ratio will improve. Having a high percentage of recent papers available as full texts is also an important issue.

Concerning the cumincadTHESES- database, it has to be said, that the search for e-dissertations requires a closer and more comprehensive examination of university library sites. More and more e-dissertations are made available for free by universities, but retrospective digitalisation is rare. A contact with individual authors of dissertations may be established with the help of the associations. In this respect, an initial collection of 250 dissertations should be feasible. The recording in a repository might possibly encourage authors to invest in a digitisation of their work (for costs and benefits see chapter 3.4)

As a matter of principle authors have to waive the copyright to publishers when publishing a paper into a journal. However, as these journals do not belong to the category “grey literature” the dissemination is not blocked. Commercial interests on this side are, however, not in line with free electronic publishing.

3. WORKFLOW: DIGITISATION AND CONVERSION

3.1 INTRODUCTION

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 digitalisation of collections may serve a larger audience, more independently from time and space limitations. Also academic associations and other organizations have achieved remarkable results so far.

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. Analogue materials (paper-based media), have to 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 mistakes. Depending on the result, elimination of these mistakes 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 out this niche.

In this chapter a calculation of data storage and/or presentation costs (e.g. in databases) is not considered. Possibly there will be no crucial cost problem on a long-term basis, because those costs may sink substantially in the future. In the present situation it must be said, however, that the maintenance of a book shelf may – just in a small number of cases - still be regarded as the cheapest form of storage. Total digitisation (in terms of scanning, OCR and conversion to, e.g. PDF) may be uneconomic or not eligible for financing and requires selection and setting of priorities. The digital version should, however, deliver an increase in value in relation to the original format.

3.2 PDF-FORMAT

For the time being the pdf-format serves as a solution to provide the community with information. The abbreviation "pdf" stands for "portable document format" which can be implemented on practically all current platforms (http://www.adobe.com). Pdf-files have become popular and aim mainly at producing a printing format in its original form on any available computer by means of a reader (free of charge) or a plug-in, which means that complex software-installations become obsolete and usage is available without any "ifs and buts". The html-format seems not to a working solution as, for example pagination is difficult to control. Also, conversion from layout-files to html may require intensive labour. However, pdf-files can be created from any type of software used, as long as a printing is provided. In fact the work is nearly the same as performing a printing job. The result will not appear on a piece of paper, but on the screen. Independent of how the computer is configured or, for example, which fonts are installed, the user of the pdf-file can printout and the result will appear the same as it

was on the editor's machine. It has to be said, that pdf-technology allows for a number of restrictions, such as viewing only (disable of printing) or, for example disabling the selection of text, in which case the document would principally allow for full-text search.

3.3 EXAMPLES OF DIGITISATION PROJECTS

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.

The following selected examples – in alphabetical order – presents various kinds of experiences and considerations:

- Center for Retrospective Digitization, Göttingen State and University Library (http://gdz.sub.uni-goettingen.de/en/)
  Text material is scanned from microfilm (with new production date) and from the original. Scanning from microfilm (35mm) is contracted to a vendor; scanning from the original is undertaken in-house using the library's own equipment (at 600 dpi). The Zeutschel scanners run under the scan software "SRZ ProScan Book", developed for the GDZ by the Satz-Rechen-Zentrum Company in Berlin to meet special production scanning of older books (e.g. with TIFF-header editing, production control window with tree view over scanned pages, masking and cropping of pages during the scanning process) The scanning process produces a high quality digital master in TIFF-format. Derivatives (GIF, JPEG) will be created on-the-fly for online delivery. A PDF version can be obtained for downloading or for offline delivery on CD-R. Core bibliographic information is added to the TIFF header of the digital master to reference each image to a bibliographic record in our online library catalog. The digital masters are stored offline on CD-R, an ISO-standard storage medium.
  A well-known problem is the recognition of textbooks in "fraktur" (gothic). As the broad evaluation of OCR-programs by the GDZ pointed out, standard programs as well as sophisticated trainable programs (Prime Recognition, ProLector, Optopus, FineReader) do not provide solutions for an economically automatic recognition of this kind of texts. Even the creation of a raw and dirty text version can not be reached with this tools. The GDZ in Göttingen applies the Russian program "FineReader" (vers. 4.0). Non gothic texts, even from older books and difficult image quality can be recognized with a low failure rate, allowing the creation of a text index for background search. Other institutions in Germany also discovered this problem as a chance to fill a special segment of the market for text recognition programs.
  For an intervening period - until the issue of Gothic text processing can be resolved - the minimum for accessing digitized text will be the availability of navigational tools given

by the original itself, like tables of contents, indexes, list of illustrations, hyperlinked to the referenced portions (e.g. chapters, image pages) of the documents.

- Digital eCAADe Proceedings (http://www.ecaade.org/)
  The council of this CAAD-association was confronted with a situation of nearly two decades of annual conferences and a corresponding publication output, to be characterized as “grey literature”. Therefore the decision was taken to carry out retrospective digitisation 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).
- Digital preservation and METAe (http://meta-e.uibk.ac.at/)
  The project will ease the automated creation of (technical, descriptive, and structural) metadata during the image capturing and digitisation process. In other words: METAe will integrate metadata capturing into application software and it will pick-up recent developments and emerging standards in order to make the output highly compatible with existing digital library systems.
  The METAe engine itself is designed as a comprehensive software package where all steps necessary for the digital conversion (re-formatting) of printed material (books, journals) can be conducted by a well trained end-user. The input will be scanned (still) page images, the output will be an "archival information package".
  The greatest progress will be made by introducing layout and document analysis as key technologies for structural and partly descriptive metadata capturing. A high amount of the typical "keying work", e.g. the ordering of text divisions such as chapters, sub-chapters or the linking of articles within a periodical will be taken over by the METAe engine. Page numbers, headlines, footnotes, graphs and caption lines are promising candidates for automatic processing as well.
- Forum Bestandserhaltung (http://www.uni-muenster.de/Forum-Bestandserhaltung/)
  The “commercial value” of the collections of the Bavarian State Library might lay - with all caution, which is required with such statements –in an order of magnitude of 5 billion Euro. Seven million volumes correspond –with an average of 300 pages for each volume – to approximately two billion pages. Full text recognition with a high degree of accuracy may cost up to 5 Euro per page. Despite all kinds of progress by means of machine labour, this value may not drastically change in the future. Labor costs will not, however, under any circumstances sink. Thus the digitisation of the collections of the Bavarian state library would cost about double as much, as it is estimated value.
  The value of digitisation has to be related to innovative forms of content management. For instance the integration into library systems as well as the use of global search machines illuminates equally, as bad retrieval results will be with neglect of this aspect. Selection decisions are to be taken under the aspect of cost user weighings with a relatively broad spectrum of technical possibilities.
- Making of America - MoA (http://moa.umdl.umich.edu/)

MoA is an electronic research repository serving historical scholars, and as such is part of the academic library context out of which it grew and seeks to develop protocols for the selection, conversion, storage, retrieval, and use of digitized materials on a large, distributed scale. MoA4, begun in 1998, added almost 8,000 volumes to the MoA collection – over 2,500,000 pages of monographic content. MoA4 converted the vast majority of the 1850-1876 U.S.-imprint.
Pages were scanned as preservation-quality (600dpi) image files, and the Preservation Department undertook a quality control process for the images, using a roughly 5% sample. The automatic generation of the OCR makes it possible for the project to take advantage of advances in OCR technology as they become available: when better recognition becomes possible, the OCR can be automatically regenerated at very little cost.
In order to explain costs and thus make them meaningful in budget calculations for future projects, it is important to understand the steps in the conversion process, the tools used and the human labor involved. The page is used as the unit for representing the costs. An outside vendor was contracted to perform scanning and burning to CD of the page images. Fourteen bids were received, ranging from two Euro to ten cents per page. Final calculations showed an expenditure of thirteen cent per page.
The cost of capturing character-level information from page images is variable, depending on tools and methods. It is important to note that in some cases, only manual "keyboarding" can provide reasonably adequate capture and retrieval, which then of course creates a much higher cost for performing retrieval. Foremost among the variables in OCR cost is the choice of an OCR software package. The OCR software used at the University of Michigan in the Making of America IV project is PrimeOCR from PrimeRecognition. PrimeOCR was chosen for production operations because it provides demonstrably more accurate output in the OCR process. It is, however, many times more expensive than an off-the-shelf OCR package. TextBridge, for example, can be licensed for a few hundred dollars per machine, while PrimeOCR may cost more than ten thousand dollars for a multi-machine implementation. A small project would be well advised to outsource or to choose inexpensive software and to use a fraction of an existing staff member, while a project that performed more OCR might wish to pay the slightly higher cost to achieve markedly higher accuracy and sustain higher throughput.

A growing number of vendors are offering services concerning the creation of full text versions. For instance *Gutenberg Neue Medien* (http://www.fortunecity.de/lindenpark/barock/198) promises to deliver OCR-work with an accuracy from 2-4 characters per 5.000 characters. Prices are starting from 0,29 Euro for 1.000 characters and depends on a representative sample delivered. Layout and formatting are not included, which has to be taken into account for a specific field as architecture, where images play a dominant role.

Companies in China rely on cheap labour costs and offer a "retype" at 0.7-0.8 Euro per 1.000 characters with 99.998% accuracy. Please note that an average page in this report has nearly 2.000 characters.

3.4 FINANCIAL IMPLICATIONS AND COST/BENEFIT ANALYSIS

The presentation of a number of examples (chapter 3.3) provides a rough idea of the workflow involved in digitisation. However, a comparison with a cost/benefit analysis should be carried

out and will be related to the content sources described in this report. The publications involved charge contgain a mix between text and accompanying images/figures. This is not surprising as the field of architecture, or the transmission of entities thereof, is heavily depending on visual communication. A combination of text and image is therefore in many cases crucial. Furthermore the materials are not unique the sense that they only exist exactly one time (as is the case for example with medieval manuscripts etc.). The longest time frame which can be discovered for *CAAD* and *Construction IT* here is less than fifty years. Quantities of less than 50.000 pages (and not millions!) would potentially be required for digitisation in this area. However, a certain minimum quantity need to be delivered in packages as the basic overhead for handling etc. plays a role. In this respect an estimation for 1.000 pages with different types will be displayed (this is regarded to be equivalent to 125 papers), but overall measures such as further handling (entry in a database, production of a CD-rom etc.) is excluded:

A. *Source in a digital format is available – Conversion to full text pdf*

   As a matter of 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.

- *Estimation: 4 Working hours – 200 Euro / = 0,20 Euro per page*

B. *Source in a paperbased format is available – Conversion to "image"-pdf*

   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".

- *Estimation: 300 Euro / = 0,30 Euro per page*

C. *Source in a paperbased format is available – Conversion to full text pdf*

   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.

- *Estimation: 1.900 Euro / = 1,90 Euro per page*

D. *Source in a paperbased format is available – Retype and Conversion to full text pdf*

   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.

- *Estimation: 1.600 Euro / = 1,60 Euro per page*

E. *Source in a paperbased format is available – Selected conversion to full text pdf*

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.

- *Estimation: 500 Euro / = 0,50 Euro per page*

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 (file management: for example splitting up of a single pdf-file into individual pdf-files. In case a splitting of individual files (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 and would thus introduce a cost of 0,40 Euro per page.

3.5 CONCLUSIONS ON VARIATIONS IN WORKFLOW

The description of variations of chapter 3.4 aimed 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 digitisation 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 (enrichment with “intelligence”).

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 with 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 due to retype is missing, 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 mistakes 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.

4. CONCLUDING REMARKS

In the SciX-project the CUMINCAD-repository, which was so far developed on a shoestring budget, is used as an object of study and research. The sustainable liaison with the different associations in the area of *CAAD* and *Construction IT* allows us to catch up with the publication output from the last 15 to 20 years. Indeed the “pioneers” from the early days may still be active, but one retirement follows another and the inclusion of materials should be secured.

However, in terms of collecting “coherent” initial content, first of all the gathering of conference proceedings has to take place. The realization of this has the full support of the associations in charge. The procedure requires prioritisation and points therefore to actions that will quickly lead to tangible results. Furthermore higher priority will be given to recent publications, which are available in a digital format and cause minimal costs for handling, as indicated in variation A. (chapter 3.4). For example, 4.000 pages – which is equivalent to 500 papers – would lead to an investment of 800 Euro for handling (or 20 working hours). Furthermore, high priority should be given to these materials, for which the copyright issue has been cleared. Digitisation of older, non-electronic materials should be performed at a later stage and by third parties, which are, for example, to be commissioned by the associations (“outsourcing”). However, such an effort can be regarded as an added-value for the repository, but is not decisive for the development of a repository.

In terms of sustainability, the extension of the repository is not in danger on completion of the SciX-Project. The annual entry of about 500 publications is feasible and would be accompanied for a large part of the records by full text documents. Further research and analysis on the stored content may follow and would therefore in principle allow for an innovative handling of information. This is not possible by means of mentally driven abilities of individuals. Even if someone retrieves every day 100 pages and studies them carefully, soon the limitation of personal memory would appear. Sustainable acquisition of content is tangible by means of a well equipped repository. CUMINCAD is in this respect a unique repository and serves a growing user-community, in which especially doctoral degree students benefit very much from a collection of structured information.

The observation of developments concerning digitisation of materials should take place frequently in the future. Also tests with, for example, the ResearchIndex-software “CiteSeer” makes sense, as the source is for free for non-commercial environments. The use of other engines, such as METAe (currently under development in the framework of an other IST-project), should be studied as well.

The actual status of the repository will be described in delivery D06, which provides actual figures about the achievements.

We can conclude that the best strategy to create complete and relevant topical digital archives lies with the good relationships with the societies whose members did the research and wrote the materials. It is of utmost importance these societies take such a archive to be "their own" and that are represented in the steering of the archive. Such an approach is also contributing to

the reversal of the process that took scientific publishing away from the professional societies in the early 20th century.