Integration of Digital Services for Libraries

R. Bayer, R. Kallenborn, H. Haddouti, W. Wohner, A. Mödl, R. Heinrich, D. Nitsche
Technische Universität München

Abstract

A library is a collection of books and journals augmented by a set of services like acquisition, cataloguing, lending, etc. In the digital age the role of libraries, in particular the nature of services they offer, is changing dramatically. A "Digital Library" is considered naively as a collection of documents published in digital or electronic form. The essential property of a digital library is, however, that it offers all its services in digital form, completely independent of the original nature of the documents that make up the library.

In this paper we consider the essential services of a library and discuss, how these services will change and how they should be offered in the digital age. We will describe technical, architectural and organizational alternatives and solutions and augment these deliberations by practical examples and experiences gathered in several digital library projects like OMNIS, VD17, Elektra, and DIBWIN at the Technical University in Munich and in cooperation with various partner libraries. These projects range from the digitization of old books (VD17) to an electronic delivery service for scientific articles (Elektra).

1 Introduction and present Status

According to the Oxford Dictionary, LIBRARY has two meanings:

1. a place set apart to contain books for reading, study or reference
2. the books contained in a library, a large collection of books, public or private.

Both definitions of LIBRARY have almost nothing to do with the concept of a DIGITAL LIBRARY as envisioned in this paper. A digital library is characterized by a set of digital services which it offers online, it does not necessarily have a dedicated building (a place set apart), nor is it a large collection of books contained in a library.

In this information age, however, every classical library in the sense of the Oxford Dictionary must strife to also become a digital library by offering a set of attractive digital services. This minimal set of digital services which a library must offer to its readers in the future are:

- find books and scholarly articles according to the specifications of the reader, via a comprehensive online catalogue or via a broker
- enable access to or deliver (parts of) books and articles in digital form online,
- collect money for providing these services on the basis of selling, subscription, or pay per view.

Superficially, these services look more like a bookstore than like a library. This is true and I am convinced that libraries will get difficulties to justify their existence unless they adopt their new role and offer very advanced digital services quickly and effectively.

In order to provide these digital services to its readers, a library must construct a "virtual collection" of documents which consist of a variety of real documents: its own conventional paper based documents, the paper documents of cooperating libraries, its own digital documents as well as the digital documents available via Internet. This virtual collection must have a comprehensive and detailed catalogue to act as a finding aid, it is the repository from which documents can be delivered quickly and to which access can be granted for money or free of charge.

So, what will differentiate a library from a bookstore in the future? In the US Amazon has captured more than 3 million customers on the book retail market in just two years and its sales are approaching a billion dollars per year, similar developments are likely to happen in Europe soon. Will this trend also impact or endanger libraries? Traditional libraries have a better chance of survival than bookstores, simply because they own the treasures of the past and because they have a powerful, professional organization which can offer its services better, cheaper and more efficiently. In particular, a digital library will mainly be a meta-library, which manages and offers information about books and articles rather than housing them and providing shelf space for books and desk space in the reading room.

2 Guarding the Treasures of the Past

One of the most important tasks of a library is to preserve materials that are out of print. Libraries are doing an admirable job in honoring this duty, but they have to go one

First Russian National Conference on
DIGITAL LIBRARIES:
ADVANCED METHODS AND TECHNOLOGIES,
DIGITAL COLLECTIONS
October 19 - 21, 1999, Saint-Petersburg, Russia

step further. In many cases these treasures are too valuable or too rare to make them accessible to the public as originals. Retroconversion to digital form for preservation as well as availability over the WEB for public accessibility are the technical solution to the problem, but this solution is tedious and expensive.

In the projects Oettingen-Wallerstein [1] and VD17 [2] (both funded by the Deutsche Forschungsgemeinschaft – DFG) and OMNIS [3] we provided technical solutions and gained experiences in building a combination of a conventional MAB (the German standard for the exchange of bibliographic records, similar to MARC) catalogue extended by the digital images of key pages like title page, owner and dedication pages, outstanding selected prints etc. From these projects we learned two lessons:

a) The technical solutions - although simple in principle - are complicated in reality. All materials like bibliographic records and pixel images are stored in standard relational databases. But a group of about 25 librarians in 6 very important German libraries are cataloguing all day online. The long delays and the still limited effective bandwidth of the Internet forbid the standard solutions of classical online transaction processing (OLTP). Therefore, specialized techniques like fingerprinting of documents were adapted to ease synchronization, the databases are partially partitioned and replicated to guarantee acceptable performance and certain operations have to be batched. The interested reader is referred to [7] and [10].

b) The cost of retroconversion is quite high, at least if the high standard prescribed by the DFG library advisory board is observed. It had been projected by experts and experience over two years of operation has confirmed that a librarian can catalogue about 12 titles per day, or 2400 titles per year. This amounts to roughly 40 DM pure labor cost per title, or 16 Million DM for the estimated 400.000 titles to be catalogued. This project is scheduled to last for 10 to 12 years and will produce just a multimedia catalogue, not even full conversion of books.

Full conversion would be possible with the present technology and could be offered selectively for special collections, or on demand for single volumes if funding of the cost would be available. I estimate that complete conversion of a modern book could be achieved at about the same price as producing the catalogue entry, once the book has been catalogued.

For the conversion of old books, however, special equipment, special training and care are necessary. Special and quite expensive book scanners or digital cameras are required to scan books that can only be partially opened, software is needed to correct an image that is taken from the bent page of a partially opened book, people must be trained to handle books and to adjust equipment properly. Therefore, the obstacle to full retroconversion of old books is the labor cost, but neither computing power nor storage capacity. In other words, anything the market would pay for could be retroconverted with technology available today.

3 Archiving of Digital Materials

Libraries are doing an excellent job and are spending a lot of money (including money for buildings) to preserve paper books, but they seem nearly helpless to archive the digital materials of today.

In the past, publishers, even University departments were obliged to deliver one reference copy of a book or even a doctoral dissertation to one or several libraries charged with the responsibility to preserve these materials forever.

German publishers shed this responsibility by simply sending a reference copy to the Deutsche Bibliothek in Frankfurt, typically a CD-ROM for digital publications. It is most questionable, whether this material will still be usable (readable) a few years from now. However, the libraries are still in an enviable position compared to the Bavarian National Archive, which receives the material not when it is published, but about 30 years later, when the various departments of the Bavarian State try to get rid of old material. Then some of it has to be archived; nobody knows, however, how to do this. Simply shelving magnetic tapes or floppy disks is certainly not an acceptable solution.

Note, however, that the sheer possibility to read 30 year old media is not even enough. Much more critical is the capability to run the old software to process and to interpret this information. Presently, there is no chance to solve this problem of the Bavarian State Archive and probably of many other archives.

In my opinion, "preserving the past" is one of the key roles of a library that set it apart from publishers, bookstores and even professional organizations like ACM or IEEE, and libraries must solve this problem somehow.

The obvious solution is: Collect upon publication and convert the whole material to the next generation of technology and products as they become available on the market. This, however, would be a continuous monumental task and would still not solve the problem of migrating the software needed to process the documents. To give an example: the DEC PDP 11 once was a very popular computer in the seventies. Documents prepared then and stored on tapes are probably extremely difficult to revive just for the purpose of reading them, which is far from the problem of processing them.

Here librarians are confronted with a fundamental shift in paradigm: the new digital documents are not simply *read*, they must be *presented dynamically* in order to be fully perceived. More specifically: just storing a PowerPoint presentation for reading is far from sufficient since much additional information may be conveyed by the dynamic process of presenting it. To use a picture from object oriented software: documents are not just printed materials, they are objects encapsulated with the software methods to present the information to the perceiver, and therefore, they must also be archived as such complete objects.

4 The new Publishing Process

To understand the role of a digital library, let us briefly consider the publishing process. The observation that literature is published via different media and in digital format is very superficial, the impact is reaching much further in at least two aspects:

4.1 Distribution Channels

Today's distribution system for books looks as follows:

Author ⇒ Publisher ⇒ Wholesaler ⇒ Bookstore.

¿From the bookstore the book goes to the reader directly or indirectly via the library.

What is the future role of the library in this supply chain? In Germany there are about 2.100 publishers, only 53 wholesalers, about 4.800 bookstores, about 700 scientific libraries, and hopefully about 75 million readers. It is likely that this complete production and delivery chain will be replaced by a direct digital link from the publisher to the reader. Amazon is only the tip of the iceberg, since it only replaces the bookstore and still ships physical books instead of digital documents, so there is much more change to come.

4.2 Hypermedia, the Enemy of Books

Electronic media and digital formats will have a much deeper impact than just changing the distribution channels: they free text from the bondage of linearity and turn it into Hypertext. But hypertext cannot be published on paper, it demands the new media. More and more text books are published in or at least accompanied by a hypermedia format [11]. It is not clear yet, how this will ease the process of reading and learning, deviating from linear reading by skipping or going into more depth and detail with interactive support. Certainly, hopes are high, and if they bear fruit this may be the end of the classical text book and of the scientific library as we know both of them today. Just consider this paper, it contains 8 URL links and in combination with a computer on the Internet turns into a hypertext document which is much more valuable and informative than just the plain paper version in the conference proceedings.

So the role of the library as a repository of physical books may diminish and will become much more demanding as it must mutate into a meta-library. The speed of change is stunning: looking at my own published papers, the first single URL showed up in [13], in recent dissertations in my research group about 20% of the citations are URLs, in this paper it is 44%.

Probably quite soon citations in scholarly papers will only be read, if they are available on the WEB. Therefore in the next step they must be on the WEB in order to be read and cited. This effect will probably create a tremendous pressure to publish digitally and may lead to a sudden revolution comparable to the introduction of fax, email and mobile phones within a very short time.

Hypertext documents offer a lot of flexibility for interaction with the reader, but still they are static and closed documents. There is, however, a step beyond hypertext, namely dynamic and evolving documents. Consider a research community like historians and sociologists working with the VD17 materials who would like to annotate texts by comments, translations, interpretations, links to other texts etc. To manage such dynamic complexes of documents is an entirely new challenge for libraries and librarians.

5 The digital Meta-Library

It should be clear by now that the main task of a digital library will no longer be to buy books, to subscribe to journals and to store them in shelves; instead it must help the reader to find and to identify information, to deliver it instantly to him, and to make sure that the proper presentation tools are available. The task will shift from information storage to information management and mediation. To solve this problem the crucial asset is meta-information: the knowledge how and where to find required information, how to get it, how to deliver and how to present it.

The classical tool to solve this problem is the library catalogue. This too will change dramatically in the digital age. Multimedia catalogues like VD17 are just simple beginnings. What should be the functionality of future catalogues, how can they be built? Let us look at past and present:

Library materials were originally organized according to card catalogues sorted by authors or subjects. Therefore, most libraries had two card catalogues, one for authors, one for subjects. These catalogues were highly redundant, since separate placeholder cards were needed for multiple authors and for multiple subjects covered by a book or an article.

Computer technology made it possible to merge both card catalogues into one, to eleminate most of the redundancy, to build the catalogue cooperatively by several libraries, thereby saving cost, and to support rather flexible search capabilities. These electronic versions of catalogues are available to the public via computer terminals and therefore are called OPAC (online public access catalogue).

Building good catalogues is still a tedious and very expensive process. The cost of a simple catalogue entry mentioned for the VD17 system is typical and quite a surprise for the layman, 30-50 DM per title.

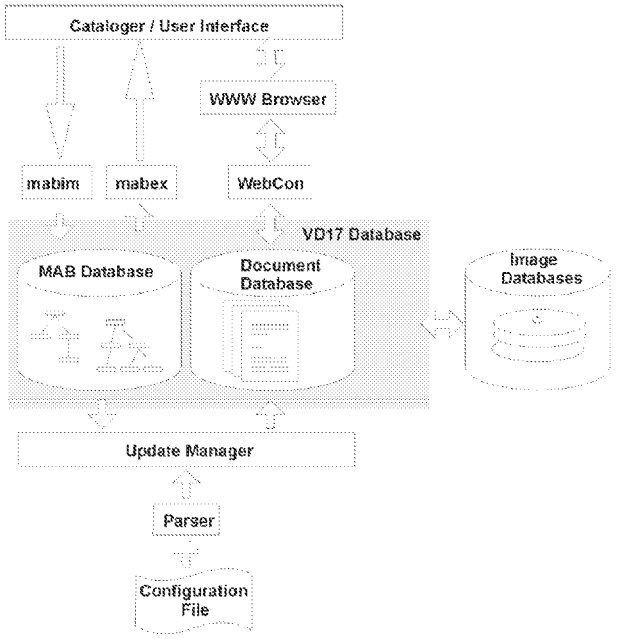

Fig. 1: VD17

Because of the limited functionality and coverage of library catalogues scientists rely more and more on the search engines of the Internet. Although this is a poor and insufficient substitute, libraries will have to work hard to offer a substantially better alternative.

The catalogue is the meta information and will probably be the most valuable asset of the future digital library. What functionality must it offer and how can it be built efficiently and economically?

State of the Art Techniques

Typing a bibliographic record into the computer according to a precisely prescribed format, checking standard spelling of authors, publishers, etc. is too expensive, even if a certain

percentage of records can be copied from the centralized catalogue, if they had already been entered by somebody else.

Therefore, in the OMNIS and Elektra systems we switched to a semiautomatic process:

- scan key pages like cover and table of contents of books, first pages of journal articles
- convert the scanned pages into plain text by OCR (optical character recognition)
- fill information for the bibliographic record partially from the database, e.g. publisher of a journal, year and number of next journal issue, etc.
- fill the rest by cut and paste from the OCR text.

The main problem of this technique is the low quality of the OCR text. Our technique has two main advantages:

- lower cost, namely about 4.5 DM per entry for a journal issue instead of about 40 DM per entry.
- much more detailed information about documents is available via this multimedia catalogue, in particular: table of contents for books and abstracts on the first page of articles as a facsimile of the original form, fonts and layout.

Importing other Catalogues

We could achieve additional considerable streamlining in building the catalogue of the document delivery system Elektra (elektronischer Aufsatzdienst) by subscribing to the SWETS service for bibliographic records and abstracts of the articles of about 14 000 scientific journals worldwide. This covers many fields, but in particular 419 journals of the total of 534 journals in Mathematics and Informatics, to which our library subscribes: The technique described before is used to add articles in those journals not covered by SWETS. In addition the catalogue is enhanced by scanning the first pages of all articles. With this technique the labor cost per journal article came down to less than 1.- DM. In addition we are sharing this cost between several participating libraries. Presently the Elektra Catalogue has 65.000 entries and a total of 5 GB. The catalogue is growing by about 3000 entries per month.

Semantic Knowledge via Domain Catalogues

The Elektra Catalogue covers all of Computer Science and Mathematics and is rather unstructured, just a set of entries searchable by any mixture of attribute based search and full text search with Boolean retrieval. It would probably be more useful to have *knowledge complexes* covering certain subjects, like galaxies in the chaos of the Internet. Such knowledge complexes can probably be built by collecting references, which are semantically connected, and preserving this semantic connection. Following are just some ideas, which we have not tried yet:

Citation Complexes

Conferences or survey papers have a certain topic like "deductive databases" or "multidimensional indexes". The papers cited in a survey paper have been carefully selected by the authors who should be experts in the field, these papers should have been read and judged relevant. For electronically published material it is fairly easy to parse the text, to find and to collect citations, and to record in addition the information that document B has been cited in document A. If B is also published electronically, this process can be iterated to some depth or back to some threshold in time etc. In addition one can do a reference count to automatically build a citation index ranking the most frequently cited and therefore probably the most important papers. Such semantically enriched catalogues would be much more useful than today's collections of isolated document references.

Organization Complexes

Other possibilities to construct Domain Catalogues would be to start from the homepages of well known researchers, research projects or visitors of conferences, and to follow the citations and hypertext links to papers and technical reports found there. Probably rather simple syntactical analysis suffices to identify relevant links and to compile comprehensive catalogues at least semiautomatically. Building such semantic catalogues can be done only semi-automatically, if a certain level of quality should be achieved, which can be guaranteed by adding human intelligence to the process. Building such catalogues is a challenging task and should be one of the prime value added services of a digital library. Today libraries boast with the number of books they have on their shelves or even with running meters of filled and still free bookshelves, the most important argument, to get a new building. The quality measure in the future will be the size of the catalogue or even better, the cumulative size of catalogues and virtual collections that are accessible and made available by the broker system of the library.

Backsourcing Responsibility to Publishers or Authors

It would seem most natural and effective to shift the responsibility for constructing a catalogue record in a standardized format back to the publisher or the author. A literary work would be considered "published" only, if this standardized record were available. Although various efforts have been made, this method has not been completely successful and is limited to issueing the ISSN and ISBN, which are not of much help. One reason that this approach is not successful is that the pure bibliographic record is extended by a lot of local information like date of acquisition, signature, state of lending, reservation etc. Note that most of this information will disappear in an online digital library. Therefore, building catalogues will become simpler at least in some respects.

6 Retrieval

With the WEB, OPACs were made available via Internet and Browsers, but their functionality did not change. In particular, information retrieval via free text search and classical OPACs were not integrated. In the OMNIS system [14] we combined both retrieval techniques, covering tables of contents of books and journals by free text search (via the SQL interface of a relational DBMS) and the bibliographic records via the structured search capabilities over the attributes of relations in a relational DBMS.

It is a general observation, that library users have considerable difficulty in finding documents via structured interfaces. Therefore, in OMNIS we combined both techniques by partitioning the attributes according to the conventional

categories in catalogues, but doing a free search over attribute values. Yet, much more advanced retrieval techniques should be offered in digital catalogues:

- thesaurus with the specific semantics of a subject like Informatics or Physics
- reduction of words to their stems both in the catalogue and in the queries
- multilinguality

Basically these are well researched methods of computer linguisitics and should be introduced into advanced catalogues.

User Profiles

Scientists and researchers are interested in certain, usually highly specialized subjects, which they can specify by so called "profiles of interest". They would like to be notified automatically if a new work is published matching their profile. Therefore, a library should offer such a profiling service. This, however, is a challenging problem from a database point of view, see [12].

Relevance Judging

Information retrieval research has developed elaborate methods for relevance judging and ranking of documents, like word counts for queries weighted by the inverse length of a document. The most satisfying and reliable way to judge relevance is to show the reader part of the document, like table of contents or abstract, therefore they must be part of the catalogue like in OMNIS and Elektra.

Another possibility is to compute textual similarity between documents and to retrieve clusters of similar documents. Also papers cited in a document could be relevant and become part of the retrieved hits, which would require an extension of the catalogue (as discussed in chapter 5), but also of the query language.

Before ordering an article, the reader might want to see the chapter headings to improve relevance judging. For documents written in markup languages like SGML or HTML such information could be extracted with a modest amount of effort and integrated into the catalogue and into the retrieval process.

Brokers

An alternative to "Importing other Catalogues" (see section 5) is building a broker, which just accesses other catalogues for retrieval without actually importing them. Various such brokers have been built [15], [16]. The difficulty is that brokers have to deal with various catalogues with different query interfaces and formats, and they have to combine redundant results in various formats.

7 Location and Availability

Libraries have an elaborate signature system to physically place and locate books. A book returned to the wrong shelf is usually lost for a long time or even forever.

In addition, a reservation and recall system is needed for books that have been reserved for borrowing. For journal articles such a system does not even work, since usually many articles are contained in a single journal issue. As soon as single issues are bound into yearly volumes the problem becomes even worse. Therefore, many libraries take the brute force action not to lent journal material, maybe a necessary, but very unfriendly gesture towards readers.

In the digital age, this problem of lending must disappear altogether. A library will not keep physical volumes around, but fetch documents via their URL or URI and the Internet directly from the server and deliver it to the user. There is no need to monitor, who borrowed what and when and how long, since new digital copies are simply pulled from the original.

Of course, this touches copy right, unauthorized proliferation, misusage, payments in electronic or conventional form, authentication methods for documents, etc. What is still needed, however, is user management, rights management, access control, identification, authentication and authorization of users in order to control the use of the system services and to secure payments via electronic payment systems.

8 Document Delivery

Cataloguing and retrieval deal with meta-information and just help to identify relevant documents. Getting a copy of the document is an entirely different matter, this problem is solved in various, largely unsatisfactory ways:

- copying or reading journal articles in the reading room of the library. This is unsatisfactory in a highly dislocated environment like our CS department, which until recently had been spread over 14 buildings all over Munich
- borrowing a book from the library, which often causes considerable delays
- borrowing a book via interlibrary loan, which is complicated, costly, and causes even more delays
- ordering an article via a document delivery service with fax or paper mail delivery.
- ordering an article via Subito, a national effort in Germany to set up a cooperative document delivery system with acceptable cost (5 DM per article), but rather slow delivery times (24 hours for urgent to 72 hours for normal delivery). Presently there are 20 libraries active as Subito delivery centers. Subito is just a document ordering (via specialized email form) system, it lacks a comprehensive catalogue, the delivery is not really integrated into the system and payments are collected conventionally by paper bills sent out by mail. At least Subito is on the right track by replacing buying or subscription by "pay per view".
- ordering via ELEKTRA, our more advanced document delivery system, in which multimedia cataloguing, previewing and inspection, ordering, delivery and electronic payment are fully integrated and which delivers in a very short time (less than two hours) [18].

9 Payments

Delivery of documents in electronic form should be very cheap, for such transactions conventional payments are not feasible. Therefore we combined Elektra with the Chablis [17] system for electronic payments and we are experimenting with several forms of digital payments via Internet.

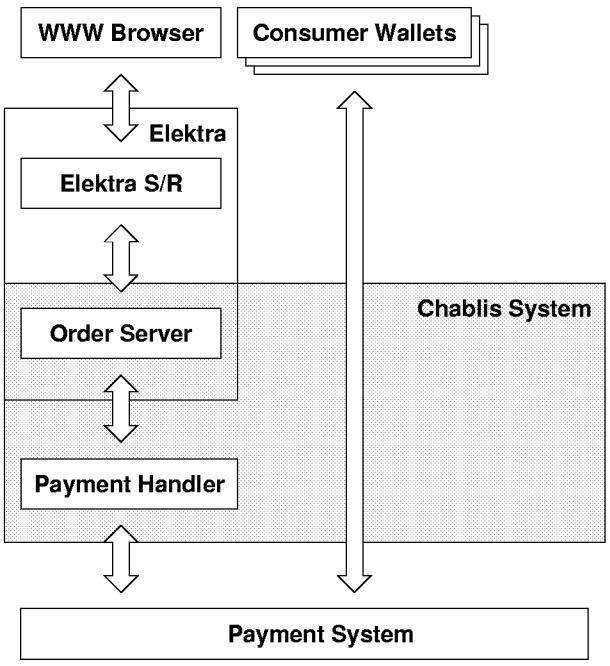

Fig. 2: Elektra Chablis

Although libraries are publicly funded and are not used to act commercially, they will probably be forced in the future to deal more with money, hopefully not with cash, but they will have to acquire considerable competence with electronic payment systems, probably for small amounts.

10 10. Summary and Outlook

We have argued that the future digital library will be completely different from today's libraries. Its prime task will be the careful and competent combination of many digital services into a well integrated, smooth, effective and reader friendly system.

The library will have to go through a metamorphosis into a meta-library justifying its existence with the added value of integrated digital services.

References

[1] *FORWISS–Projekt Oettingen–Wallerstein* [http://www.forwiss.tu-muenchen.de/~oewal]

[2] *FORWISS–Projekt VD 17* [http://www.forwiss.tu-muenchen.de/~vd17]

[3] *OMNIS Hauptseite* [http://omnis.informatik.tu-muenchen.de/]

[4] Kallenborn, R. *OMNIS/Myriad – Einsatz und Perspektiven eines multimedialen Systems* Bibliotheksforum Bayern.:1995.–Jahrgang 23, Heft 1

[5] R. Bayer, R. Heinrich, R. Kallenborn, A. Mödl, D. Nitsche *ELEKTRA - Tor zur elektronischen Fachinformation* Mitteilungen der Technischen Universität München fr Studierende, Mitarbeiter, Freunde 1997/1998.–Heft 6

[6] H. Haddouti, W. Wohner, R. Bayer *Towards a Scalable System – Architecture in Digital Libraries* IDEXA'99, Florence 30.8 - 3.9, Springer Verlag, Berlin, 1999

[7] M. Dörr, H. Haddouti, S. Wiesener *The German National Bibliography 1601 - 1700: Digital Images in a Cooperative Cataloging Project.* Proceedings of IEEE ADL'97, Washington, DC, May 7-9, 1997, IEEE Computer Society Press, Los Alamitos, 1997 [http://www.forwiss.tu-muenchen.de/~haddouti/adl97.ps]

[8] H. Haddouti *Issues of Libraries in the Digital Era* Proceedings of NEW MISSIONS OF ACADEMIC LIBRARIES IN THE 21ST CENTURY: AN INTERNATIONAL CONFERENCE, Beijing, October 25-28, 1998 [http://www.lib.pku.edu.cn/98conf/paper/d/Hachim_Haddouti.htm]

[9] H. Haddouti *Multilinguality Issues in Digital Libraries* Proceedings of the EuroMed Net'98 Conference, Nicosia, March 3-7, 1998

[10] M. Dörr, H. Haddouti, S. Wiesener: *Das 17 Jahrhundert im Netz* DFN Mitteilungen 41, Juni 1996, pp 4-6

[11] W. Kießling, G. Köstler: *Multimedia-Kurs Datenbanksysteme* Springer Verlag 1998

[12] C. Nippl, M. Jaedicke, B. Mitschang: *Accelerating Profiling Services by Parallel Database Technology* Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), Las Vegas, Nevada, July 1997

[13] R. Bayer: *The Universal B-tree for multidimensional Indexing: General Concepts* WWAC'97, March 1997, Tsukuba, Japan

[14] R. Bayer: *OMNIS/Myriad: Electronic Administration and Publication of Multimedia Documents* Informatik, Wirtschaft und Gesellschaft, 23. GI Jahrestagung, Springer, Dresden, 1993

[15] *XRCE Knowledge Broker* [http://www.rxrc.xerox.com/ats/xtrim]

[16] *Karlsruher Virtueller Katalog* [http://www.ubka.uni-karlsruhe.de/kvk.html]

[17] O. Gentz: *Integration der digitalen Zahlungssysteme CyberCash und Ecash in den elektronischen Aufsatzdienst Elektra* Technische Universität München, Diplomarbeit, February 1999

[18] *DIBWIN Hompage* [http://elektra.informatik.tu-muenchen.de]