Prints of the 17th Century in a Distributed Digital Library System

C. Böhm[1], A. Oppitz, P. Vogel[2], S. Wiesener[1]

[1]Bayerisches Forschungszentrum für Wissensbasierte Systeme (FORWISS)
[2]Technical University of Munich, Department of Computer Science

email: [boehmc, vogel, wiesener]@informatik.tu-muenchen.de

Abstract: In recent years, library systems have evolved from simple electronic catalogues to extensive document sources. Progress in storage media, database and wide area networking technologies allow access to documents in their original form. Besides "modern" documents, which are often created and distributed in electronic exchange formats, huge amounts of ancient, paper based literature which are valuable for scientific research exist. In the long-term project VD17[1] all prints of the 17th Century published in the German-speaking area will be registered, scanned and stored in a distributed digital library system. Historians, literary scholars, and other scientists in the field of humanities can later access catalogue data of the prints together with scanned key-pages via Internet. About 250,000 titles with more than a million pixel images will be held on magnetic hard disc devices and optical storage media in jukeboxes.

This paper gives an overview of the project's goals. It describes the system architecture and dataflow of the enhanced OMNIS digital library system which is used as technological basis for the project. The solution is based on a previous pilot project.

1 Introduction

Today, digital libraries support research activities in many fields. Not only researchers of natural or engineering sciences, but also historians or arts scholars use network-based information systems and profit from the latest technological developments. Multimedia aspects, especially image access, drastically enhance the practical value of digital libraries. In particular, this applies to historical prints which cannot be transcribed into character sequences without loss (see Figure 2).

Scientists of various disciplines have stated a serious deficit in access facilities to bibliographic data about German literature of the early modern times. Especially the 17th Century, the classical epoch of art in France, Britain and all over Europe, is not very well registered in German literature catalogues. On the one hand, this is a result of historical developments. The thirty years war and some epidemics have led to a cultural breakdown and the loss of catalogue material. On the other hand, there are obstacles, resulting from the way how librarians are treating historical prints in their archives

[1] The VD17 project is promoted by DFG (Deutsche Forschungsgemeinschaft)

today. Documents with an age of a few hundred years are considered as too valuable and often too fragile to be handed out to a broad community of interested persons. Thus researchers have only a limited access to historical library archives of the 17th Century. Currently available are some isolated electronic catalogues of several libraries or library networks which have enormous differences in quality and coverage. In addition, they only offer textual catalogue information by uncomfortable query interfaces.

The goal of the long-term project VD17 (Registration of 17th Century Prints Published in the German-speaking Area) is a digital library system which not only contains a rather complete bibliographic catalogue of 17th Century prints but also pixel images of scanned key-pages. These images will offer a new dimension of quality to scientific users, as the prints include much relevant information in drawings, seals, handwritten annotations, and layout. Catalogue data can be retrieved via a comfortable fulltext query interface and is linked to the relevant image sequences. The client/server architecture allows worldwide access. Optical storage media in jukeboxes cope with the huge amounts of image data: more than 250,000 catalogue entries will be registered, accompanied by more than a million pixel images (color and b/w).

1.1 Project Stages

As a pilot project to VD17, the registration of the historical Oettingen-Wallerstein Library in Augsburg is in progress. In this project, the technological basis was created and verified. By now, 79,000 catalogue entries with an amount of 14,498 pixel images have been stored in the distributed digital library system OMNIS [2], [3] and are accessible via Internet today. In this pilot project many experiences how to cope with image databases on optical storage media were gained.

The VD17 project has started with an installation phase of 18 months where OMNIS is to be adapted to all project demands and installed at participating libraries. To allow concurrent work at several locations, synchronized cooperative registration is supported. The following main phase will exceed ten years, in which the libraries will do their complete registration and scanning of the prints in the final OMNIS digital archives.

1.2 Infrastructure and Workflow

All public German libraries with stocks of the 17th Century take part in the VD 17 project. The main parts of the relevant literature are concentrated in three large libraries in Berlin[1], Munich[2] and Wolfenbüttel[3]. These libraries have been chosen as processing centers. They are equipped with facilities for catalogue registration and image entry and also process the prints from the smaller libraries in their area. The catalogue and image data are then transfered via Internet to a central Catalogue Database and several Image Databases in Munich where they are available for world-wide retrieval. In the main phase of the project high-end scanning equipment is required, as many prints are too

[1] Staatsbibliothek zu Berlin

[2] Bayerische Staatsbibliothek München

[3] Herzog August Bibliothek Wolfenbüttel

fragile to be digitized with conventional scanners. If available, microfilms are digitized instead of the original prints.

The following section shows how the VD17 project will be realized with the OMNIS library system. Especially catalogue and image server are described in more detail. OMNIS retrieval clients are presented with their retrieval facilities.

2 VD17 and the OMNIS Digital Library System

OMNIS is an information retrieval system for administration of documents in libraries and office organizations [3]. It has been developed at the Computer Science Department of the Technische Universität München and FORWISS (Bavarian Research Centre for Knowledge-Based Systems). It is supported by Myriad, a fulltext database system [12] with an additional SQL interface and TransBase [13], a pure relational SQL database system. OMNIS runs in a client/server environment and offers a number of economic and functional advantages by consequent utilization of latest achievements of data technology.

2.1 Documents

The heart of the whole OMNIS philosophy is the "document" which is the unit for the archiving process and retrieval results.

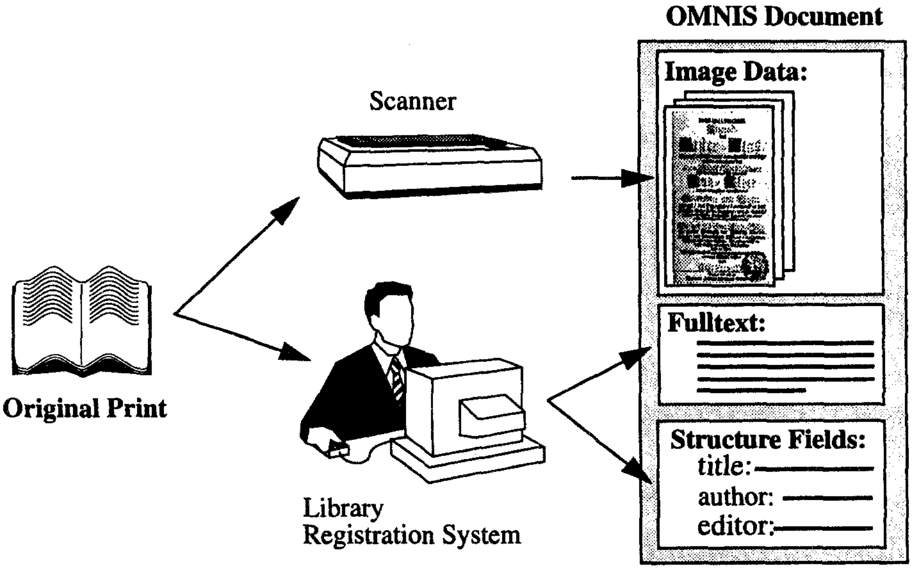


Fig. 1. 17th Century print as logical OMNIS document

Each OMNIS document represents a VD17 catalogue entry and contains information concerning historical prints in three different sorts of attributes:

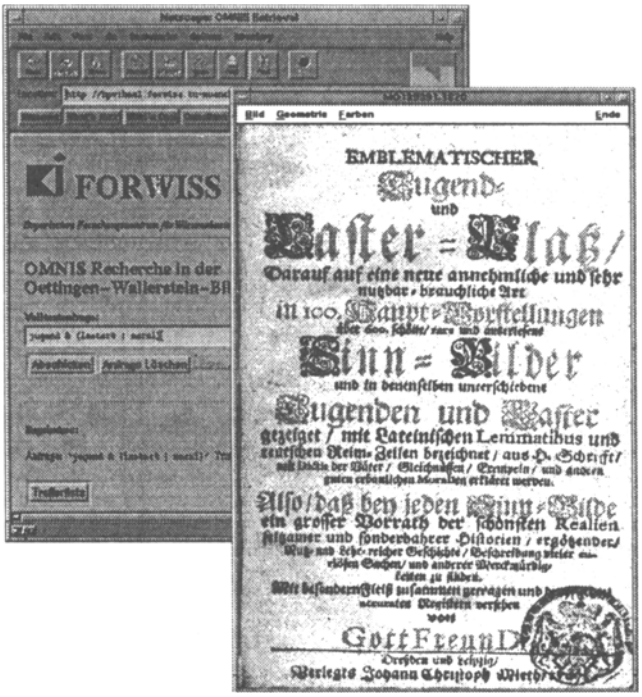
- **Structure Fields** describe the catalogue entry in a relational way. Fields like *author, title, etc.,* provide structured information as known from traditional literature retrieval systems. Structure fields are the basis for relational queries and can be accessed by users. Only a small part of each catalogue entry is stored as OMNIS structure fields in Myriad databases.

- **Fulltext** contains the whole catalogue entry as a text body. It is the basis for comfortable fulltext queries and can be accessed by retrieving users. A document's fulltext attribute is an unstructured sequence of words stored in Myriad databases and represents a superset of the document's structure fields.
- **Image Data** in some pixel format may be attached to each document. These images are stored as BLOBs (Binary Large OBjects) [10] in TransBase databases and can be shown to the user.

Note that OMNIS documents are logical units only. Physically, each document is distributed over a Catalogue Database and an Image Database. Figure 1 depicts the data classes which constitute a document.

2.2 Retrieval Interface

The main target group for the VD17 project are students and scientists of literature. For this group, a new dimension of literature search and access opens: a simple fulltext retrieval language allows effective on-line catalogue search.



EMBLEMATISCHER
Tugend-
und
Laſter = Glaß/
Darauf auf eine neue annehmliche und ſehr
nutzbar- brauchliche Art
in 100. Haupt-Vorſtellungen
über 600. ſchöne/ rare und auserleſene
Sinn = Bilder
und in denenſelben unterſchiedene
Tugenden und Laſter
gezeiget / mit Lateiniſchen Lemmatibus und
teutſchen Reim-Zeilen bezeichnet / aus H. Schrifft/
mit Sprüchen der Väter / Gleichnüſſen / Exempeln / und andern
guten erbaulichen Moralien erkläret werden.
Alſo/ daß bey jeden Sinn-Bilde
ein groſſer Vorrath der ſchönſten Realien
ſeltzamer und ſonderbahrer Hiſtorien / ergötzender/
Nutz- und Lehr-reicher Geſchichte / Beſchreibung vieler cu-
riöſen Sachen/ und anderer Merckwürdig-
keiten zu finden.
Mit beſondern Fleiß zuſammen getragen und [illegible]
accuraten Regiſtern verſehen
von
GottFreund
Dreßden und Leipzig/
Verlegts Johann Chriſtoph Mieth

Fig. 2. World-Wide Web retrieval interface

Scanned key pages can be requested and become available as images at the clients' desktop in a few seconds. They are open to users without special access rights as the valuable originals need not be touched.

The manual registration procedure distributes the produced data over a large number of fields. In this way, a very detailed classification of document attributes is achieved, which is needed for the management of prints by professional library personnel. For each VD17 print nearly a hundred fields may be filled with bibliographic data but usually only a subset of them is really used.

Offering this broad variety of fields to the often inexperienced retrieval users would make it difficult to decide what fields should be used for querying. That's why OMNIS groups all fields describing a VD17 print into a fulltext body, where field names are no longer distinguished. Only a few very common fields can be queried by OMNIS structure fields (see Chap. 2.1). The OMNIS retrieval interface allows combined fulltext and structure field queries. Wildcards and boolean operators, as well as phrases (sequences of words), support the formulation of complex queries.

Processing the query

```
goetze georg & (laster| verfall) & ?Erscheinungsjahr=1707
```

in the Oettingen-Wallerstein library would deliver the print (bibliographic data and pixel images) shown in figure 2, because it contains the phrase "Goetze, Georg H.", one of the words "Laster" (vice) or "Verfall" (corruption) and its structure field `Erscheinungsjahr` (year of publishing) contains "1707".

Additionally, wildcards may be used to replace single characters, arbitrary character sequences or words (distance search). Two different query input interfaces allow easy fulltext and structure field query formulation, partly supported by input masks.

In order to reach many users, several retrieval clients were developed. A special OMNIS client is available for Unix/X-Windows platforms. A VT100 client offers access to users with ASCII terminals. It can be used to install a Telnet gateway into OMNIS libraries. Of course, no pixel images are transferred. Today most users access OMNIS libraries via WWW. An OMNIS-WWW gateway [4] offers full OMNIS retrieval facilities to WWW clients.

Try the preliminary URL

```
http://hpwibas1.forwiss.tu-muenchen.de:8001/
```

which is a gateway to the historical library Oettingen-Wallerstein, built in the pilot project of VD17. It contains 79,063 documents from the Middle Ages to modern times with 14,989 scanned pixel images.

2.3 Entire System Architecture

The whole OMNIS system has a distributed client/server architecture. Two usergroups can be distinguished: the first group, the library and administration personnel enters data, registers prints, stores catalogue entries in the central catalogue database, scans prints and archives pixel images into Image Databases. The other group, the retrieval users access catalogue and image data for browsing and studying purposes.

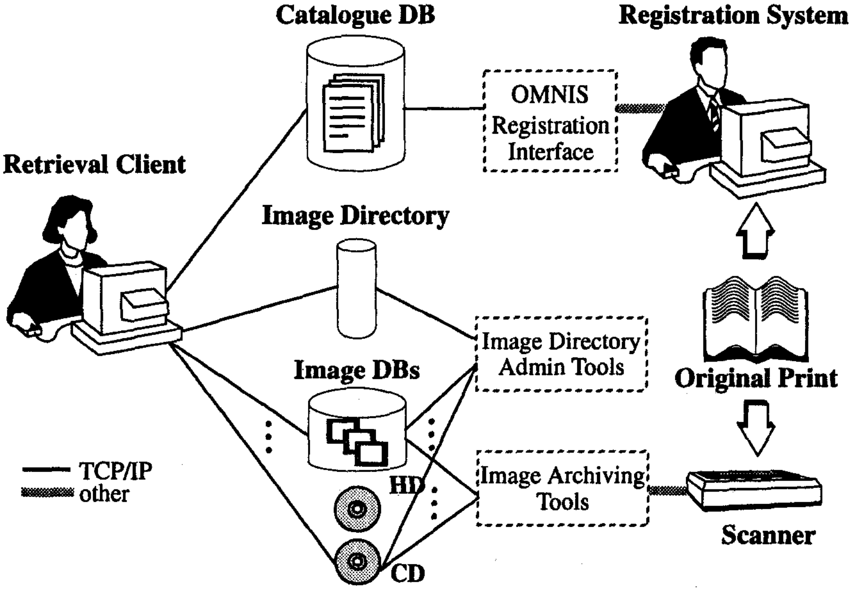


Fig. 3. VD17 system architecture

One feature of the VD17 project is the separation of image data from textual catalogue data. Both are managed from completely independent database servers (see Figure 3). This is a direct consequence of the extremely different characteristics of image data compared with textual data:

- To ensure efficient fulltext and structure field retrieval for catalogue entries, special index structures are used by the Myriad DBMS. Storage media devices should allow fast random access with short track-to-track seek times. This is provided by a fast hard disc. Image data, however, is not used in the searching process and would overload the disk. Therefore it is just referred to in the catalogue by simple IDs and stored in separate databases (see section 2.5).

- Pixel images consume much more memory than textual catalogue data. Special high capacity storage media are needed to cope with about a million color and black-and-white images which again are usually not suitable for fast search. In the VD17 project CD-ROMs in jukeboxes contain images mainly. Here, the pure data transfer rate is more significant, because once located, images are read as linear successions of blocks.

- Catalogue and image data can be used independently. Catalogue data can be offered without picture access and pixel images can be combined with other catalogue databases in future projects. To increase catalogue retrieval speed, Catalogue Databases can be replicated locally, while image data can still be delivered from remote Image Database servers.

- Image Databases require a different maintenance. While Catalogue Databases should be backed-up regularly, because they are located on magnetic devices, image databases on CD-ROM devices are robust. In addition, a backup of all images is too expensive to be run daily.

The main server types of the VD17 project are Catalogue Database server, several Image Database server and an Image Directory server. Catalogue Databases manage textual catalogue entries. Besides a pure retrieval facility, they provide synchronized catalogue entry updates for a cooperative registration in the three registration centres Berlin, Munich and Wolfenbüttel. Image Database servers manage picture data and picture attributes of the corresponding pixel images. They may be located at different sites. A registration ID connects catalogue data to picture data. The Image Directory manages the locations of Image Databases. By requesting the Image Directory, retrieval clients determine which image server contains the desired picture sequence referring to a catalogue entry.

Special OMNIS retrieval clients (see Chap. 2.2) directly communicate with a Catalogue Database, Image Databases and a designated Image Directory. The World-Wide Web gateway [4] in fact is an OMNIS client which is connected to a HTTP server [11]. It generates HTML pages for user interaction (e.g., query input and result output). The WWW clients' HTTP commands are translated into OMNIS Catalogue Database, Image Database and Image Directory requests. Resulting data like catalogue entries, hit lists and image data are transformed back into a WWW client readable format.

2.4 Catalogue Management System

The main function of the central Catalogue Database is the storage and management of catalogue entries. On the one hand it synchronizes cooperative print registration at three distributed registration centres (see Chap. 1.2) and, on the other hand, it offers powerful retrieval facilities to end users.

One challenge of the VD17 project is the integration of legacy catalogue registration systems with modern OMNIS document management. In the participating libraries, catalogue data is still entered in hierarchical structures, which are a burden of hierarchical database management systems. Of course, these structures somehow meet the demands of efficient catalogue management by professional library personnel, because they, e.g., allow the clustering of copies to titles, titles to supertitles etc. (see Figure 4). In contrast to OMNIS fulltext document retrieval, where all relevant catalogue information of a print is stored in one flat document, access to hierarchical catalogue structures requires awkward entry point localization and navigation. This is not acceptable for the broad community of users. Nevertheless, both philosophies have to be integrated: user retrieval shall be document oriented, while catalogue registration and management shall remain hierarchical to cope with older registration systems and ensure catalogue data exchange in standard formats.

The catalogue server redundantly manages catalogue entries in two logically independent databases. One stores and synchronizes hierarchical catalogue trees in a relational data model, while the other holds the corresponding OMNIS documents for retrieval.

Figure 4 depicts how OMNIS documents are built from a catalogue tree. Coming from a title record, all contributing copy and supertitle records are collected, formatted and supply fulltext form and structure field entries of a single OMNIS document. Note that there will arise more than a hundred thousand of such catalogue trees in VD17 and each tree contributes to several documents. A catalogue record contains numerous attributes which describe parts of a VD17 print.

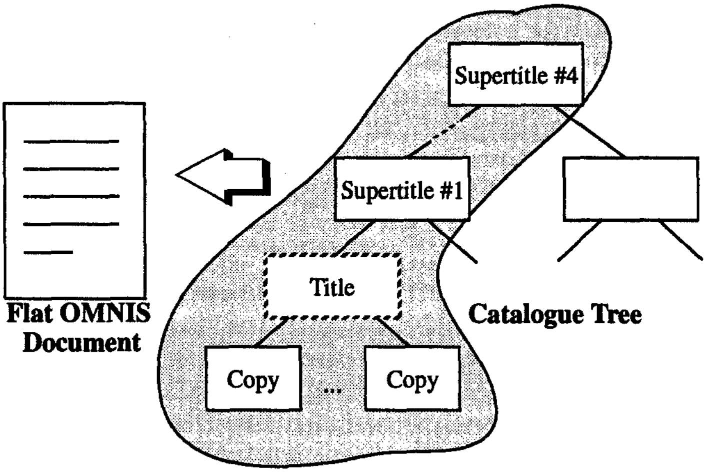


Fig. 4. Paths in catalogue trees form OMNIS documents

If a record in a catalogue tree is modified by the registration personnel, one or more OMNIS documents have to be re-generated. Responsible for this task is a component, called the Update Manager. Whenever the Catalogue Database changes state (e.g., after updating, inserting or deleting a catalogue record), the Update Manager evaluates which records have been modified. All OMNIS documents, built of these records are cancelled and rebuilt. The Update Manager's algorithm ensures that only a minimum set of documents has to be reprocessed after an update.

For registration purpose catalogue trees are searched as OMNIS documents via fulltext/structure field queries. On demand, the corresponding catalogue tree is checked-out to the legacy registration system. During check-out time, the whole tree is locked. At check-in of the modified tree (e.g., newly registered copies were added), the Update Manager rebuilds OMNIS documents and releases locks. Note that check-in/ check-out procedures are needed, as the registration process of a historical print is a demanding and highbrow task which may last up to several hours.

2.5 Image Management System

Due to the chracteristics of the VD17 project, the image data management system's architecture has to satisfy the following practical demands:

- It has to manage enormous amounts of image data but still allow efficient access to single images.
- Because of the long term of the project the system architecture has to follow an

open design technology. This allows future improvements as a result of technological progress (e.g., adapting image quality to further developments in storage media and network technologies).

- The system has to be rather independent of a special storage media type. The administration personnel can choose convenient storage media (e.g., hard discs, MOD, CD-ROM, tapes). This demand is met in a completely transparent way for the retrieval user.
- The system has to be designed as a distributed database in order to allow a decentralization of the image management.

In order to satisfy these demands a design concept as described below has been developed. It's central idea is to structure the Image Management System into three parts: The Image Databases, the Image Directory and the Image Administration Database.

The Image Databases (see Figure 3) contain pure image data, stored in BLOBs (Binary Large OBjects) [10], as well as registration IDs and image attributes like resolution, color depth, size, compression method and image format. The registration ID represents the connection to the catalogue data, i.e., refers to hierarchical catalogue entries and thus to OMNIS documents (see Chap. 2.4). The latter attributes are used to display an image with an image viewer at the retrieval client's system. They may be accessed independently from the image content and allow future retrieval clients to make decisions about image access (e.g., considering transfer costs or client features). Image quality, storage medium and storage locality of images can be freely chosen.

The interface to image retrieval is provided by the Image Directory (see Figure 3). It contains references to all Image Databases and their images. Image access queries are at first addressed to an Image Directory. Retrieval clients assign each Catatalog Database to a designated Image Directory.

The Image Administration Database is used for administration purposes only. It stores detailed information about each archived image and helps to plan and fill Image Databases. The name of the source (i.e., location of producer), the name of the corresponding Image Database, it's original size, but also color depth, resolution, type of the used compression and format are stored in it. The Image Administration Database is provided to facilitate statistical analysis, but it is also necessary to control and synchronize the distributed image archiving process. It is not used for image retrieval.

Image Archiving

When storing images it is necessary to find a compromise for competing objectives as high image quality, low storage costs, quick network transfer and fast display at client systems. Images with a size of about 11.7." x 8.3", scanned with a color depth of 24 bits and a resolution of 300 dpi occupy about 25 MB of storage. This, of course, implies unacceptable transfer rates and storage space requirements and illustrates the necessity of using adequate techniques for image data compression. In the pilot project "Oettingen-Wallerstein", valuable experience could be gained regarding methods with optimal data compression and acceptable quality loss. In this project about 15,000 images of historical documents were processed and stored in two Image Databases on 2 CD-ROMs. The total images source size was reduced from 23,5 GB to 1 GB.

During the main phase of the VD17 project, images will be scanned with a resolution of 300 dpi and three different color depths:

- **Black-and-White** (1 bit) images will result from prints which originally did not contain color information and are in good condition. They will be compressed without loss using CCITT Group 4 compression [1]. An average compression rate of 10 was achieved in "Oettingen-Wallerstein".
- **Greyscale** (4 bit) images will result from prints which originally did not contain color information but suffer from dirtying and yellowing. They will be scaled down to 150 dpi and then compressed without loss using LZW compression [1]. An average total compression rate of 17 was achieved in "Oettingen-Wallerstein".
- **Color** (24 bit) images will result from originally colored prints or prints which received color information in later annotations (e.g., seals, handwritings). They will be scaled down to 150 dpi and lossy compressed using JPEG compression [14]. An average total compression rate of 54 was achieved in "Oettingen-Wallerstein". The average color image size was 180 KB which is suitable for available storage media and network qualities.

After scanning onto a local hard disc at one of the registration centers, information of each pixel image is stored in the central Image Administration Database. These data are the basis for planning the distribution of images to databases, which is done by administration personnel. At last, images are loaded into their designated hard disc Image Database. All those transfers use TCP/IP. If such a database reaches 650 MB size, it is written onto a CD-ROM which represents an independent Image Database from now on and is typically managed by a jukebox. Other storage media types can be handled similarly.

To make an Image Database accessible to retrieval clients, it has to be announced to an Image Directory, which maps registration IDs from OMNIS documents to Image Database names. Note that each Image Database may be announced to many Image Directories, i.e., images may be used in different catalogues. Think of local catalogue replications for fast retrieval, whereas images are stored remote. Image Databases can also be decentralized, i.e. each contributing library manages its own images locally but also contributes to remote libraries via network.

Image Retrieval

For each image request the client at first accesses the Image Directory, selects the address of the corresponding Image Database and finally requests the image content from there (see Figure 3). For developing OMNIS retrieval clients easily, an API (Application Programmers' Interface) is available which handles Image Directory and Image Databases transparently.

3 Conclusion

Scientists and students are awaiting an appropriate access path to historical material which is spread over many library archives. There have been several hundreds of thousand of books of the 17th Century with extraordinary value, that have not been used satisfactorily, simply resulting from an unsufficient registration or old-fashioned catalogue retrieval systems.

With fulltext retrieval and the possibility of downloading scanned documents' key-pages, the project VD 17 will open up new dimensions in researchers' methods of work. Effective retrieval techniques will lead to drastically reduced costs of research undertakings. Interested people will have access to information which often was locked behind closed doors for more than three hundred years.

The VD17 project is an example how other disciplines may benefit from the consequent utilisation of newest data and network technology. It demonstrates that the OMNIS digital library system is a feasible technological basis for such applications. Apart from the registration and archivation of historical prints, other document types can be handled similarly without much modification of the systems architecture.

4 References

[1] Aldus Cooperation: TIFF Revision 6.0, Aldus Developers Desk, Seattle, June 1992

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

[3] R. Bayer, P. Vogel, S. Wiesener: OMNIS/Myriad Document Retrieval and Its Database Requirements, DEXA 94 (Database and Expert Systems Applications), Proceedings, Springer Verlag, Berlin, 1994

[4] A. Clausnitzer, P. Vogel, S. Wiesener: A WWW Interface to the OMNIS/Myriad Literature Retrieval Engine, Third International World-Wide Web Conference, to be published in COMPUTER NETWORKS AND ISDN SYSTEMS, Elsevier/North Holland, Amsterdam

[5] E. A. Fox: Digital Libraries of the Future, Proceedings of the First ACM International Conference on Multimedia, Anaheim California, August 1993

[6] E. A. Fox (ed.): Source Book on Digital Libraries, Technical Report # 93-35, Dept of Computer Science, Virginia Tech, 1993, also on ftp://fox.cs.vt.edu/pub/DigitalLibrary

[7] IBM Corporation: The Vatican Library, Press Release, April 11, 1994, on http://www.ibmlink.ibm.com/oi/press/pressrel/94041103.html

[8] R. R. Larson: Design and Development of a Network-Based Electronic Library, Sequoia 2000 Technical Report 94/54, University of California, Berkeley, 1993

[9] B. Lincoln: WAIS Bibliography; Technical Report, Thinking Machines, August 1991

[10] K. Meyer-Wegener: Multimedia Databases, Teubner, Stuttgart, 1991

[11] NCSA: NCSA httpd, http://hoohoo.ncsa.uiuc.edu/docs/, 1995

[12] TransAction Software GmbH: Myriad System and Administration Guide, TransAction Software GmbH, Munich, 1992

[13] TransAction Software GmbH: TransBase Relational Database System System and Administration Guide, TransAction Software GmbH, Munich, 1993

[14] G. K. Wallace: The JPEG Still Image Compression Standard, Communications of the ACM, April 1991