



Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material

COLLATE

IST-1999-20882



Final Project Report

– Deliverable No D11.1 –

Last update: 3 February 2004

Reporting period	Runtime of the COLLATE project		
Start date	1 September 2000	End date	30 November 2003
Project duration	39 months		
Project partners	Coordinator: Fraunhofer IPSI Contractors: IPSI, DIF, FAA, NFA, Uniba, Sword, Risoe		
Document name	D11.1_V1_COLLATE.doc	Version V 1	draft <input type="checkbox"/> final <input checked="" type="checkbox"/>
Authors	Adelheit Stein & all COLLATE partners		
Contact person	Dr. Adelheit Stein Fraunhofer IPSI Dolivostrasse 15 D-64293 Darmstadt	Phone: +49 6151 869 841 Fax: +49 6151 869 989 Email: stein@ipsi.fraunhofer.de	

Table of Contents

1	Project Overview.....	3
1.1	Overall Goals and Achievements.....	3
1.2	Consortium and Role of Partners.....	4
2	Project Objectives.....	6
3	Methodologies	10
3.1	COLLATE System Architecture	10
3.2	Integration Methodology	11
3.3	Knowledge Processing and Information Modeling using RDF	11
4	Project Results and Achievements	13
4.1	Digital Collection and Domain Data Bases	13
4.1.1	Preparation and Digitization of Source Material	13
4.1.2	Document Representation and Indexing Schemes	18
4.2	Document Pre-Processing Modules	24
4.2.1	Automatic, Intelligent Document Processing (WISDOM++).....	24
4.2.2	Image and Video Analyses Tools.....	32
4.2.3	Digital Watermarking of Digitized Documents.....	40
4.3	Integrated COLLATE System.....	47
4.3.1	XML Content Management and Retrieval Services	47
4.3.2	User Interfaces for Indexing, Retrieval and Collaboration Support.....	49
4.4	Working with the COLLATE System.....	56
4.4.1	Results from the Users' Work with COLLATE	56
4.4.2	Empirical Evaluation of User Experiences	64
5	European Added Value.....	72
6	Outlook.....	73
7	Conclusions.....	73
8	References.....	75
9	Annex.....	77
9.1	COLLATE Deliverables.....	77
9.2	COLLATE Publications and Conference Presentations	78
10	List of Figures	87

1 Project Overview

1.1 Overall Goals and Achievements

In September 2000 an international team of technology developers, content providers and a designated evaluation partner started out to develop and put into practice a new type of *collaboratory* in the domain of cultural heritage. The overall goals of the COLLATE project were to implement and evaluate in real life a “collaboratory in use” that offers new ways of document-centered knowledge work to distributed user groups. In an interdisciplinary approach, research and development within COLLATE focused mainly on two target areas, i.e. establishment of

- an innovative, comfortable working environment for domain experts and information consumers providing task-based, context-aware support of the users’ collaborative work with digitized data, knowledge and metadata included in COLLATE’s digital repository;
- a comprehensive digital library/archive for European historic film documentation that has thoroughly been analyzed, interpreted, indexed and annotated by a multi-national team of film experts.

The collaboratory was designed to account for requirements and work processes of user groups from the Humanities, particularly those working with historic material and cultural heritage. For a prototypical application domain we chose the heritage of European film making in the 20ies and 30ies. We incorporated a large body of multiformat, multimedia documents in the digital collection. The COLLATE system technologies, however, were designed to be easily adaptable to other – similarly information-intensive – application domains and usage contexts.

COLLATE’s digital collection of rare historic documents was provided by three major film institutes/archives from Germany, Austria and the Czech Republic. It consists of about 18 000 digitized document pages describing film censorship procedures related to historic films and enriched context documentation such as press material, digitized photos and film fragments. Members of these institutions – film historians and archivists – worked as pilot users, employing the COLLATE system for detailed cataloguing of the document collection and for in-depth content indexing and annotation of relevant sub-collections.

The developed system technology incorporates cutting-edge document pre-processing facilities, XML-based content management and advanced content-based retrieval functionalities. The document pre-processing modules employ innovative technologies for digital watermarking of documents, automatic intelligent document analysis/recognition and automatic content-indexing of image and video material. The COLLATE core system modules for data management and access/retrieval are complemented by a multi-agent-based collaboration layer and user interface system that offers situation-dependent, proactive assistance to the COLLATE users in their collaborative indexing and annotation work.

As a virtual knowledge and working environment for distributed user groups, the COLLATE system provides content-based access to the repository and appropriate task-based interfaces for analyzing, comparing, indexing and annotating the material. It supports the users’ individual and collaborative work with the sources, continuously integrating the user-created knowledge (metadata, annotations, etc.) into the system. This growing body of metadata is exploited by the COLLATE system using intelligent document processing and advanced XML-based content management and retrieval functionalities.

The COLLATE project documents the experiences of the domain experts’ real-life work using the collaboratory for in-depth indexing and annotation of the document collection as well as for detailed case studies on film censorship (see examples at www.deutsches-filminstitut.de/collate/index.html). Our approach featured iterative system development, where evaluation steps were explicitly built in and the users themselves took on a central role in the overall system.

At the end of the project – in November 2003 – we are able to offer to future content suppliers and expert users the following COLLATE system components and functionalities.

First, we have available a full-fledged, *complex indexing and retrieval system* whose main components – from a user's point of view – are:

- **A working environment for user inputs** which supports easy access to and inspection of digitized documents (text and pictorial material) and provides expert users with document- and task-specific input forms for detailed cataloguing, indexing and annotation of these documents.
- **An advanced search and retrieval engine** which allows content- and context-based access to documents and/or annotations of documents, integrating various types of customizable search functions (e.g., direct database access, attribute-value searches in metadata and full text searches in annotations, abstracts and transcriptions) on user-created as well as on automatically generated metadata to satisfy the diverse information needs of the end users.
- **An annotation-based collaboration facility** which mediates implicit and explicit communications between the users (e.g., mutual requests and task assignments), allowing to create and maintain a discourse in the form of nested annotation structures and notifying the virtual collaboration team with certain information about the state of their collaborative work.

The COLLATE Web client is implemented in JAVA and can thus easily be installed on any platform running the current version of the JAVA Runtime Environment (v1.4.x). After downloading the bundled zip file from the COLLATE Website www.collate.de (in the restricted area, currently password-protected), the client has to be extracted to some folder and can be started by running the corresponding batch file.

Additionally, we can offer the functionalities of the three *document pre-processing modules* developed and put forward within the COLLATE project:

- **Document structure recognition system** (WISDOM++): supports the workflow from digitized paper/text documents to the recognition of semantic text structures, which are represented in XML, can thus be analyzed by OCR and then searched by full text retrieval mechanisms.
- **Digital watermarking system**: supports the protection of property rights of the content suppliers by offering advanced mechanisms (not commercially available) to mark sensible documents with invisible but highly protective and robust “copyright” and “integrity” watermarks.
- **Automatic image classification system**: supports automated, rule-based indexing of pictorial material (digitized photos, posters, video frames, etc.), which then allows content-based search/retrieval of images not indexed manually/intellectually.

All implemented COLLATE system modules and the users' actual work experiences with the COLLATE indexing, annotation and retrieval user interfaces are described in more detail in *Section 4* of this report.

1.2 Consortium and Role of Partners

The COLLATE consortium consisted of three *technology developers*, three *content providers/pilot users* of the COLLATE system, and a designated *evaluation partner* for empirical user studies (see table below).

The partners' roles and activities in the COLLATE project were:

- **Fraunhofer IPSI** (formerly GMD-IPSI) was both the leader/coordinator of the scientific and technological developments and major technology provider of the project. IPSI also closely worked together with the DIF in order to coordinate the content-related activities of the three archives, particularly their work with the COLLATE system, and with RISOE who conducted the empirical evaluation studies.
- **UNIBA** was an IT research partner and technology developer, concerned with the automatic, intelligent document preprocessing facilities in COLLATE. In the system integration phase Uniba cooperated closely with the other technology providers, in particular with SWORD, in order to integrate outputs of WISDOM++ via the XML Content Manager into the COLLATE database.

- **SWORD** – the only commercial partner in the consortium – worked as a technology developer and was mainly responsible for the implementation and integration of the XML Content Manager into the COLLATE system. SWORD also coordinated the discussions about commercial exploitation of the COLLATE system and was responsible for preparing a preliminary version of the technology implementation plan (TIP).
- **DIF, FAA, NFA** had altogether very similar tasks as far as the content provision and indexing/annotation activities were concerned. Their main tasks were the identification of the test collection and preparation of source material for digitization (within Workpackage 2), and the “Preservation case studies” (Workpackage 8) which included as the major part the cataloguing, indexing and annotation of the digital COLLATE collection. **DIF** coordinated all activities of the archives, suggesting appropriate work plans and supervising the work progress. It was always in close contact with the general coordinator, IPSI, both for management issues and for discussion of user requirements and system design issues.
- **RISOE** was responsible for the systematic evaluations at various stages of the project. They conducted several empirical studies involving the archive users, in order to provide the basis for a participatory design where the users’ requirements are of central concern. Through a close cooperation with IPSI results of these analyses served as input for improving and redesigning system features as far as necessary.

Short Name	Institution	Location	Organization Type	Role in COLLATE
IPSI	Fraunhofer IPSI (Integrated Publication and Information Institute, Fraunhofer Gesellschaft e.V., formerly GMD)	Darmstadt, Germany	National IT research and development institution	Technology provider (project coordinator)
UNIBA	University of Bari (Dept. of Computer Science, LACAM Lab)	Bari, Italy	University Dept. for Computer Science	Technology provider
SWORD	Sword Information and Communication Technology S.r.l.	Bari, Italy	SME software house	Technology development partner
DIF	Deutsches Filminstitut – DIF	Frankfurt, Germany	Film institute/archive	Content provider & pilot user (contents coordinator)
FAA	Filmarchiv Austria	Vienna, Austria	Film institute/archive	Content provider & pilot user
NFA	Národní Filmový Archiv	Prague, Czech Republic	Film institute/archive	Content provider & pilot user
RISOE	Risø National Laboratory (Systems Analysis Department, HCI Lab)	Roskilde, Denmark	National research institution, Dept. for system analysis	Evaluation partner

2 Project Objectives

As pointed out in the Technical Annex to the project contract, COLLATE aimed at two complementary **strategic overall goals** (see also the goal hierarchy in Figure 1):

- To ensure content-based/semantic **accessibility** of the target contents (cultural material and historic documents in the current application domain).
- To establish empirically proved evidence for the **acceptability** of the collaboratory approach in our present example domain of historic film documentation and censorship studies.

In this approach, technology development and empirical evaluation of the developed system in a real-life environment were closely intertwined. Outputs from both areas of project work strongly influenced each other to allow an **iterative, dynamic system development**. Evaluation steps were explicitly built in, and the users of the system (collection administrators, archivists, film scholars, etc.) were actively involved throughout the various development cycles.

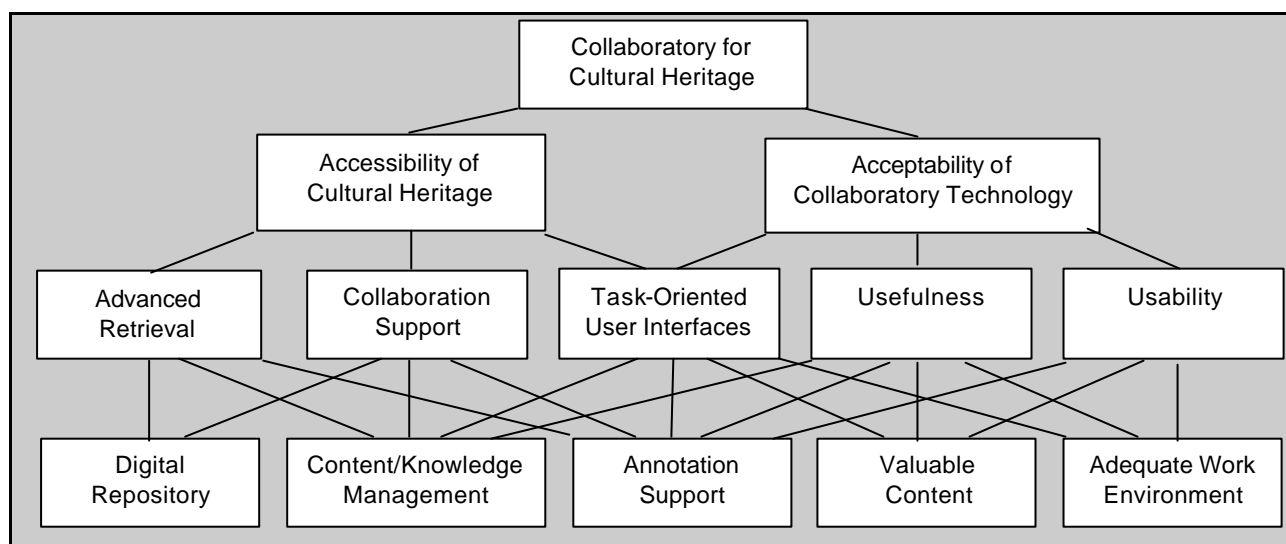


Figure 1 Hierarchy of COLLATE goals

More specifically, we aimed to support the entire workflow in question: This ranges from the first steps of creating a large digital repository by digitization of valuable conventional archive sources (paper documents and video fragments, e.g., text documents, photographs, posters, etc.) until the last steps of reception and “consumption”, i.e. the end-users’ retrieval and inspection of content- and context-relevant items from the underlying information system/database. These processes involve both expert users working with the digital material and creating new contents (metadata) and end-users who may merely be interested in viewing the contents (source documents and experts’ annotations).

Establish a collaboratory in new domains: Culture and the Humanities

The term “collaboratory” (a merger of *collaboration* and *laboratory*) has been defined as a virtual center in the Web, where professionals and lay persons are provided with means for interacting with colleagues, accessing instrumentation, sharing data and computational resources, and accessing information stored in digital libraries and archives (cf. Kouzes et al. 1996, Wulf 1989).

Various collaboratories have been employed since the early 90s, mainly in the Natural Sciences, but so far we have found – aside from some systems with very limited functionality – only a few similar efforts in the Humanities. Whereas the organization and preservation of historic knowledge in the Humanities are

still comparable to those of other disciplines, some of the work processes in the more interpreting sciences are different and need to be supported by special system functionalities.

There exist many – but mostly informal and non-institutional – contacts between cultural archives constituting specific professional communities. However, effective and efficient technological support for collaborative knowledge working is still missing. Technologically, the World Wide Web can serve both as a standard communication platform for such communities and as a gateway for document-centered digital library applications. Considering this, the COLLATE project aimed to establish a new type of collaboratory for a specific example domain – currently: film history research – which can easily be accessed via a Web interface by distributed user groups.

The COLLATE system does not only provide the functions of a traditional digital library, but on top of that it employs a generic approach to collaborative knowledge working with historic sources, i.e. supporting users in their analysis, interpretation, evaluation of the sources and the creation of new knowledge as a result of this work with the material.

Ensure access to distributed digital collections and expert knowledge

To date, a huge amount of valuable historic and cultural sources – a major part of our cultural heritage – is imperiled and buried in various national archives. Thus, accessibility, usage and full knowledge of this material are severely impeded. During the last decade many efforts and initiatives to improve this situation emerged at national and international levels. New and innovative information technologies were employed, e.g., by research programs and other funding for the preservation and improvement of access to cultural heritage artifacts and other rare sources such as historical documents. This growing awareness has brought forward a large number of specific projects, e.g., for electronic rebuilding and restoration of lost physical artifacts, or systems offering access to virtual museums. Much less efforts have been spent to build up digital archives and libraries which offer access to rare and fragile sources like historic paper documents, pictures, films, etc., especially in the Arts and Humanities.

Two major problems contribute to this highly unsatisfactory situation:

- *Immediate access* to the numerous, rich collections of existing historical archive material is impeded due to (1) difficult-to-use or (electronically) unavailable sources, both documents and formal reference systems, and (2) the lack of appropriate content-based search and retrieval aids that help users find what they really need.
- *Expert knowledge* of providers and users of such collections can so far not sufficiently be exploited for the organization, evaluation and provision of contents. Many informal and non-institutional contacts between cultural archives constitute specific professional communities, which today, however, still lack effective and efficient technological support for collaborative knowledge working.

Figure 2 gives an overview of the user requirements we elicited at the beginning of the project. The overall concern of the film institutes/archives is the offering of high quality information services in the film domain. For this service, archive people adopt the roles of scientific authors, cataloguers, indexers, information brokers and consultants of colleagues. These roles can be mapped onto different work tasks with specific needs and requirements. The main requirements we deal within COLLATE are: access to integrated archive collections, knowledge sharing, and advanced search and retrieval functionalities.

The integration of collections demands as a first step the digitization of documents. In COLLATE, the digitization of historical documents was indeed preparatory work for setting up the target collection, but it considers the current quality standards of digital long-term preservation. All digitized documents are managed in a database system that guarantees an integrated access to the diverse collections.

Digitization and system integration assure the desired accessibility of the documents, but do not yet achieve effective and precise retrieval. Content-based retrieval, for instance, demands as a prerequisite, content analysis and documentary indexing (abstracting, indexing, classifying) of the document collection.

In addition to the intellectual document examination and indexing/annotation by the users, the COLLATE system employs innovative document preprocessing technologies, e.g., for digital watermarking, (semi-

)automatic document structure analysis of digitized text documents and automatic indexing of pictorial material (photos and other images).

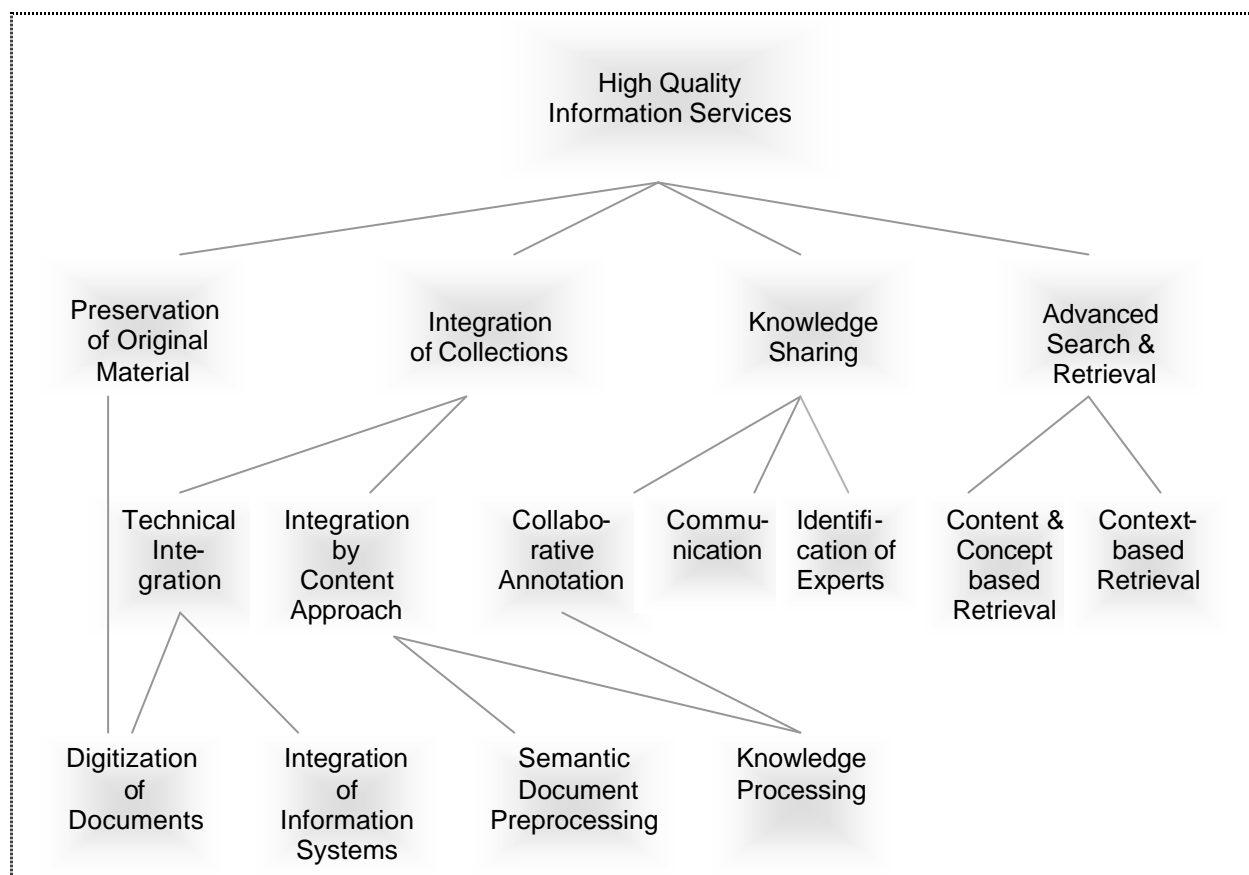


Figure 2 User requirements and tasks

Elicit proved evidence for the acceptability of the COLLATE collaboratory

A major concern of the COLLATE project was – unlike in many other EU-funded projects – the systematic and methodological sound empirical evaluation of the user experiences and acceptance of the COLLATE system. We have performed a lot of focused empirical evaluation studies in order to elicit authorized evidence about the users' traditional work environments (in the film archives) and their new work procedures in COLLATE. The evaluation activities followed two main strands:

- Archivists, film scientists and students from the three archives (15-20 persons) have tested and worked with the various versions of the COLLATE prototypes since December 01 (deployment of the first working prototype). The archive users were continuously in contact with IPSI as the technology coordinator, reporting their experiences and suggestions for improvements of specific system features as well as upcoming new requirements resulting from the daily work with the system. Immediate bugs or missing features could mostly be fixed shortly, but it was also quite important to keep track of the often changing user needs and work strategies. These requirements were collated and checked against design suggestions made by the evaluation partner Risoe and discussed in various meetings with DIF as the archives' coordinator.
- On a more systematic basis Risoe conducted several, designated studies of the user needs and requirements, as well as detailed empirical and analytical evaluations of the various COLLATE prototypes (for details see *Section 4.4.2*). Results of this work were thoroughly documented, reported in time to the technology developers (first of all to IPSI), and jointly discussed at various

meetings in order to adapt the subsequent system design accordingly. Comprehensive reports of the empirical analyses results are included in the deliverables D1, D9.1 and D9.2.

3 Methodologies

3.1 COLLATE System Architecture

The COLLATE system architecture (see *Figure 3*) is based on the reference model for an Open Archival Information System (OAIS).

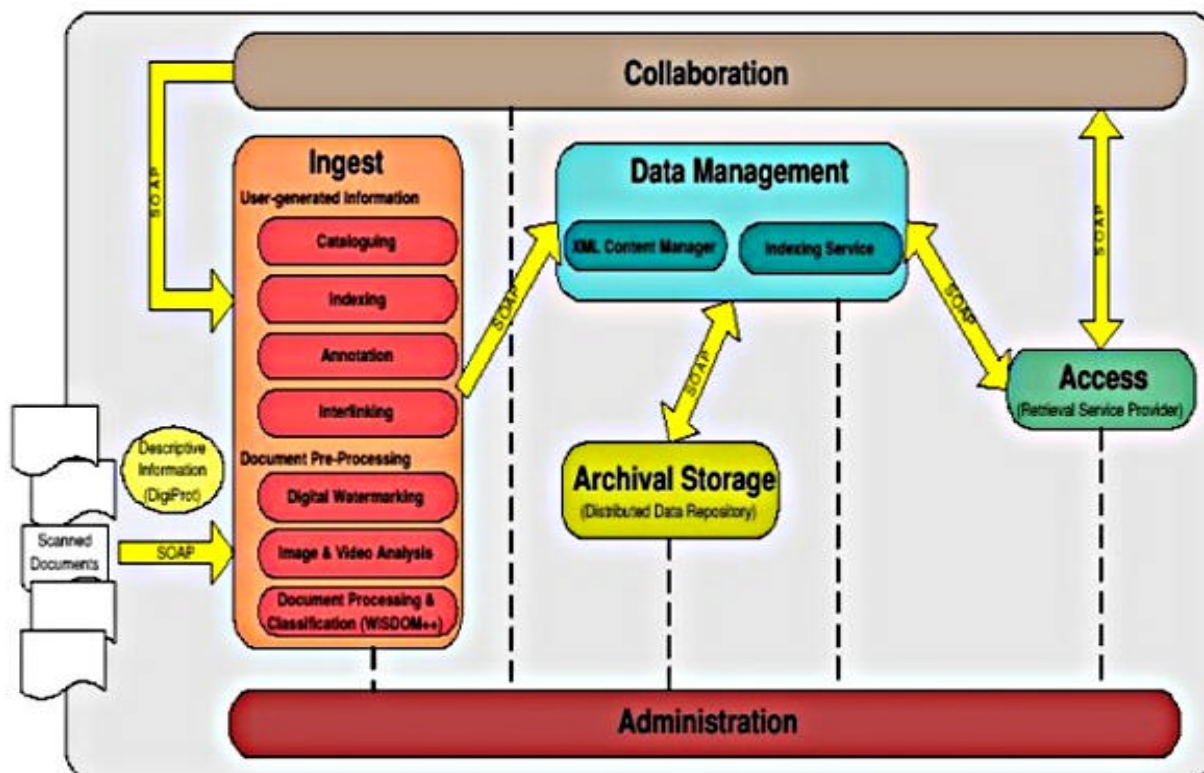


Figure 3 COLLATE system architecture

According to the definition in (CCSDS, 2002), “an OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available to a designate community”. As the OAIS approach explicitly addresses organizational needs, it is more focused on our application domain than the framework defined by the Open Archives Initiative (<http://www.openarchives.org/>), which has been founded in the area of eprint archives for enhancing communication among scholars. An OAIS consists of several modules, which are Ingest, Data Management, Archival Storage, Access and Administration. We slightly modified this model by introducing an additional collaboration layer and neglecting the preservation planning layer described in (CCSDS, 2002). An OAIS is surrounded by a *producer-management-consumer* environment; producers are those actors providing the content or the information to be preserved; *managers* define the overall policy of an OAIS (and thus do not perform day-to-day archive operations); *consumers* are using the services provided by an OAIS in order to find the information they are interested in.

Consumers access the system by invoking services of the *Access* component. Advanced retrieval functionality, e.g., based on scientific discourses, is provided here, which calls Data Management services. The *Administration* component is used, e.g., to monitor and improve archive operations, as well as to manage the configuration of the system. COLLATE also introduces an additional *Collaboration* component, which is responsible for the collaborative process.

Ensuring scientific collaboration with other experts in the cultural domain is one of the most crucial challenges in COLLATE and thus has to be reflected in the architecture. Producers (i.e., film scientists or archivists) submit scanned material to the system, using the Ingest component, which is being pre-processed and sent to the Data Management. Once the document is storing, user-generated metadata (cataloguing, indexing and annotating) is created collaboratively. If, for instance, a user retrieves a specific document and the metadata already associated to it, she might be willing to contribute additional knowledge, e.g., comment upon an annotation by an other user or complete missing cataloguing information.

In COLLATE we focus on active, system-internal support for collaboration, in particular proactive notification about, e.g., newly submitted documents and requests for comments broadcasted to relevant domain experts. It should be possible to bring together experts working in similar contexts who do not know of each other by now.

3.2 Integration Methodology

In order to bind the different sub-systems putting together *Data Management* component, a middleware-oriented approach (SOAP) was adopted.

Communications among modules of a software system must follow the set of rules defined in a communication protocol. The design process of a communication protocol is not ruled by a formal method. The only requirement is that both the sender and the receiver must be able to send messages and to understand messages.

Today's IT market offers good solutions (frameworks) to build information exchange protocols. In order to avoid the definition of a new owner model to use in the COLLATE Project, we selected SOAP as a specification of a framework solution for communication among system components. SOAP provides a simple and lightweight mechanism for exchanging structured and typed information between peers in a decentralized, distributed environment using XML. SOAP does not itself define any application semantics such as a programming model or implementation-specific semantics; rather it defines a simple mechanism for expressing application semantics by providing a modular packaging model and encoding mechanisms for encoding data within modules. This allows SOAP to be used in a large variety of systems, ranging from messaging systems to RPC.

SOAP consists of three parts:

- The SOAP envelope construct defines an overall framework for expressing *what* is in a message; *who* should deal with it, and *whether* it is optional or mandatory.
- The SOAP encoding rules defines a serialization mechanism that can be used to exchange instances of application-defined data types.
- The SOAP RPC representation defines a convention that can be used to represent remote procedure calls and responses.

Although these parts are described together as part of SOAP, they are functionally orthogonal. In particular, the envelope and the encoding rules are defined in different name spaces in order to promote simplicity through modularity. The SOAP protocol is based on the HTTP (a "connectionless" protocol based on TCP/IP) and it implements the Request/Response mechanism used for the Web (cf. SOAP).

3.3 Knowledge Processing and Information Modeling using RDF

During the past decade, there has been increasing consensus within the knowledge-based-systems community on appropriate conceptual components for building intelligent computer programs. Intelligent systems are now generally construed in terms of both domain ontologies and abstract problem-solving methods that operate on knowledge bases defined in terms of those ontologies. There has been less

consensus, however, regarding how to optimize the operational components and the user interfaces of tools that assist developers in the construction of knowledge-based systems. For the most part, such lack of consensus is to be expected, given the way in which domain considerations often dominate the way in which knowledge can best be entered, browsed, and updated in any computer-based tool.

From a COLLATE point of view, in order to organize the stored data in a way that supports the complex knowledge-intensive tasks users are able to perform on the repository contents suitable tools for metadata management were to be provided. The knowledge structures, which are represented by specific XML schemata, constitute the Domain Model. The extensions needed for metadata standards, have been coped by using RDF Models which enriches the structure of COLLATE domain.

Basically RDF represents information at a very level of granularity. It is a W3C recommendation for a standard representation of metadata, based on ideas with roots in knowledge representation research conducted over the past 30 years or so. The specification defines an abstract directed labeled graph model for RDF, and an XML-based serialization. The nodes of this graph are RDF resources, and the arcs are RDF properties. A RDF Schema specification describes how RDF may itself be used to define a type system based on RDF classes, and constraints on the ways in which RDF classes and properties may be combined in a description.

The importance of RDF is not that it is demonstrably than any other form of knowledge representation, but that it has a reasonable chance of becoming a widely used Internet standard, and that it is designed for use in an open Web environment. To exchange information (as opposed to raw data) between computer systems or applications requires agreement about its representation; at this aim, we have extended the core concepts of RDF, and the resulting framework is fully expressible within the graph structure of basic RDF. This adherence to basic RDF structure does not constrain the internal working of implementations.

RDF mechanism based on statements sets and contexts allow descriptions of complex systems to be constructed without necessarily having detailed knowledge of the ontological structure of the system components used. We believe this is a key enabler for the practical construction of complex system models in RDF.

4 Project Results and Achievements

4.1 Digital Collection and Domain Data Bases

4.1.1 Preparation and Digitization of Source Material

These tasks were structured by the results of discussions between the archives and IPSI. The goal was to define a workflow on a technical and conceptual level which enables the archives to work with the collaboratory. The main parts are:

- Gathering an overview about existing and accessible documents
- Definition of the digital repository
- Defining workflows for the digitization

The first step was to ***gather an overview about existing and accessible documents***. As contents were not only provided by the archives of the COLLATE consortium the collections of further archives were analyzed. The main contributors of content had been:

Archives:

- Bayrisches Hauptstadtarchiv, Munich
- Bundesarchiv Berlin –Lichterfelde, Berlin
- Bundesarchiv/Filmarchiv, Koblenz
- Geheimes Staatsarchiv Preußischer Kulturbesitz, Berlin
- Landesarchiv Niederösterreich, St. Pölten
- Österreichisches Staatsarchiv, Vienna
- Sächsisches Hauptstadtarchiv, Dresden
- Státní ústřední archiv v Praze, Prague
- Tiroler Landesarchiv, Innsbruck

Libraries:

- Deutsche Bibliothek, Frankfurt
- Národní knihovna České republiky, Prague
- Knihovna Právnické fakulty UK, Prague
- Österreichische Nationalbibliothek, Vienna
- Universitätsbibliothek, Vienna
- Wiener Stadt- und Landesbibliothek, Vienna

After exploring these collections the archives had to analyze and assess the different document types which were directly related to censorship examination and also examples of other document types on film-related matters and information relevant to COLLATE. Because of the differences between the censorship organization in the three countries there are differences between the available document types as well. Neither all document types existed in each country, nor all below listed information elements are contained in all documents. This means to gather a basic understanding of the different censorship procedures in each country as a presumption for the assessment. Finally a structure of different document types was defined with the aim to compare censorship procedures.

The background was that each country has developed its own censorship history embedded in the political history. Concerning COLLATE this meant we need to identify the different functional levels of censorship and censorship institutions. There exists a large body of literature on censorship processes in various countries, and much research had been done in this area. However, this was done mainly on

the national level, i.e. to our knowledge there are no comprehensive, comparative analyses of film censorship across countries that compare interrelations and dependencies.

For COLLATE we had therefore carried out some preliminary research in order to find out about similarities and differences concerning the three countries involved so far. These studies gave us valuable background information for various purposes: it facilitates (1) **selection** of relevant films and types of documents to be included in the repository, (2) **interpretation** of the significance of the document types and their interrelationships, (3) **analysis and evaluation** of the contents of the documents, which was quite important for the content indexing and annotation work of the film experts (*Workpackage 8*).

The **definition of the digital repository** included two main tasks. Firstly the archives developed shared selection criteria. These based upon an analytical examination including the period, the historical relevance of films, shared distribution and reception, individual censorship careers of films, film aesthetic and political considerations (like the rise of fascism, transition from the silent movies to the talkies etc.).

Secondly the structure of the digital repository was divided into two parts: a smaller **"core collection"** and a broader **"overall collection"**. Whereas the overall collection includes all censorship documents available from the archives for a specified period of time, the core collection focuses on a subset of about 100 significant films. These hundred films of the core collection represent the center for the preservation case studies (*Workpackage 8*) and are not only described by censorship documents, but include all accessible material of the archives (newspaper articles, correspondence, photos, posters, film fragments, etc.).

The third task was to **define workflows for the digitization**. In general the film archives DIF, FAA and NFA provide the documents for the digitization. Parts of the planned repository had been digitized in previous projects, e.g., the corpus of censorship decisions passed by the "Berliner Film-Oberprüfstelle" and digitized by DIF).

Prior to digitization some research and discussion between UNIBA and IPSI was done in order to define the necessary digitization parameters (image resolution, size, color depth, etc.) and to agree on common file formats and the mode of data exchange. The requirements of UNIBA for their automatic document processing and classification mechanism and those from IPSI for their tasks of automatic image classification and digital watermarking differed in some respects, and had to be harmonized. The archives, on the other hand, also discussed and defined standards for their preservation efforts. IPSI and DIF developed in the first phase – based on extended research of existing guidelines – a "digitization tutorial". This was then discussed among all archives and modified as necessary. The corresponding digitization specifications were then binding for all archives. The importance of using identical hardware and software in order to avoid compatibility conflicts was also stressed (agreement on that proved very useful in practice already).

After identifying the core collection, the archives started with the scanning of the relevant documents. Although the digitization of the documents initially appears a trivial process, great attention was paid to the task, as was indicated, among other things, by the fact that the digitization process was embedded in other work processes and functional contexts. Special small databases, called DIGIPROT and MINI-FILMOGRAPHY were made available to support this process.

Digitization Protocol – DIGIPROT

The MS Access database DIGIPROT (short for "digitization protocol") we developed fulfills the basic function of creating a record of the digitization process with all its settings and definitions (see the entry fields in *Figure 4*). Its main purpose is to record the relationship of the given digitized document and its file name, employing a special file name scheme we developed for precise identification of the documents provided by the three COLLATE archives. DIGIPROT also records the name of the operator and the digitization date, in order to be able to reconstruct responsibility later, if necessary.

A large portion of the censorship documents did not origin from the three COLLATE archives themselves but – after many formal negotiations and special agreements – could be "borrowed" from other, cooperating archives in order to be digitized and included in the COLLATE repository. Since, therefore, the COLLATE archives were granted only temporary access to most of the original paper documents, a

minimal formal classification was also necessary. In order to be able to later identify certain document features without access to the original paper document, we needed at least a simple document description and categorization (document type). An additional criterion for DIGIPROT is the inclusion of all description data which are not obvious from looking at the document itself at a later date. This includes, for example, notes on the reverse side or comments on the preservation quality, which, particularly in the case of photographs, is difficult to distinguish from poor picture quality. Scans of periodicals also include bibliographical references. Finally, a formal description not only of the original, but also of the digital image is provided, in particular, the size of the image.

Figure 4 Interface / input form of DIGIPROT

However, the most important attribute of a document is its document name, which was taken from the document title and was often identical with the name of the film it refers to. DIGIPROT simultaneously allocates a document ID. Each document thus includes three forms of referencing – a reference to its source, in its document name and the entries in DIGIPROT (provenance, etc.), a document ID, which is automatically allocated by DIGIPROT, and a file name, which refers back to its digital version. In this way conditions are created which allow the digitized documents to be further processed in a meaningful way in COLLATE.

All three COLLATE archives (DIF, NFA, FAA) created individual DIGIPROT databases to record their document collections (including the “borrowed” collections) digitized since the beginning of the project. They updated their DIGIPROT continuously as the digitization proceeded. Periodical updates were submitted to the database administrators at IPSI together with the newly digitized files for integration into the COLLATE database system.

Mini-Filmography

Using past experience of other digitization projects, a clear distinction was made between the various tasks which are generally covered by a single database. Whereas DIGIPROT only protocols the scanning process, records a minimum of formal dates and creates various references, the assignment of a document to a film, in other words its identification, is an independent procedure in COLLATE. The

rationale behind this is philological and addresses the question of the original or the reference. Since – following censorship directives – censored films are often re-presented to the Board of Censors under new titles but with only minimal changes, the paradoxical situation arises where a series of documents are stored under *different film and document titles*, although all refer to the genesis of a single film. Although this particular fact makes these censorship documents so important for film reconstruction and culture analysis, it does complicate the simple allocation of documents to one film title. In such cases a comparison with other information to be found in film databases or encyclopedias is necessary. Consequently, this was made into a separate procedure, both with regard to database conception and from the viewpoint of time. Nevertheless, both databases are integrated as regards their input forms and can be accessed quickly. This allows users to switch quickly from one to the other both when inputting and when conducting research.

The central function of the Mini-Filmography (see *Figure 5*) is thus to identify the corresponding film title and to allocate the relevant document. So that the film title remains unequivocally clear on the Mini-Filmography, the major essential criteria for identifying the film are also integrated, along with its title. The second important function of the Mini-Filmography is to match the identified film to the relevant information in the databases of the appropriate institutes. This is effected by integrating the relevant film ID into the Mini-Filmography. Consequently, although the Mini-Filmography itself only contains a relatively small amount of information, it also includes complex references to other databases and files. A central interface with COLLATE's annotation and indexing interface is thus created. The last step, the allocation of the relevant national film titles to a single film, is a function that can no longer be performed by this interface and the associated local working methods. This becomes the task of the collaborating agencies.

The screenshot shows the 'filmography : Formular' window. It contains the following fields and elements:

- Film ID Collate:** 177
- id's in external databases: DIF:** 35006
- FAA:** 0
- NFA:** 0
- Original film title:** Pobocnik Jeho Vysosti
- Distribution film title:** Der Adjutant seiner Hoheit
- Director:** Fric, Martin
- Country of origin:** Czechoslovakia
- Year of production:** 1933
- Production co.:** Meissner-Film Pl
- Distribution co.:** Forum Film GmbH
- indexed by:** Volker Tuchan
- check:** ☒
- comments:** from this film language versions had been produced: french, german, czech
- Document ID list:**

document
1125
1126
1127
1128
1345
1346
1347
1348
1551
1991
1992
1993
- DATENSATZ:** 1128
- COLLATE logo**
- OPEN DIGIPROT** button with navigation arrows (<, >, <<, >>)
- Mini - Filmography V1.00**
- 22.11.2000**
- last** 1128
- DIF** button
- Datensatz:** 177 von 188

Figure 5 Interface / input form of the Mini-Filmography

Finally, *Figure 6* illustrates the references between the document, the existing databases of the film archives (Filmographic databases), the COLLATE Digital Data Repository and the COLLATE Database (metadata). DIGIPROT and the Mini-Filmography provide the data on the process of digitization and film identification and by this provide input for the COLLATE Database.

The main idea behind COLLATE was not just to make the distributed documents available, but also to utilize a collaboratory designed to support users in their work – indexing, annotation and retrieval, above all. The support involved consists of a semi-automatic page segmentation and classification (automatic assignment of concepts to document segments, or pre-segmentation to support users in finding relevant segments for further manual indexing or annotation), and at the same time to provide advanced and comfortable search possibilities. Careful classification is the prerequisite for both processes. This is the only way to functionally combine heterogeneous stocks of documents and make them available for content-based, advanced retrieval. The functional connections between the documents, which differ from country to country, also have to be made compatible and represented in the COLLATE Database.

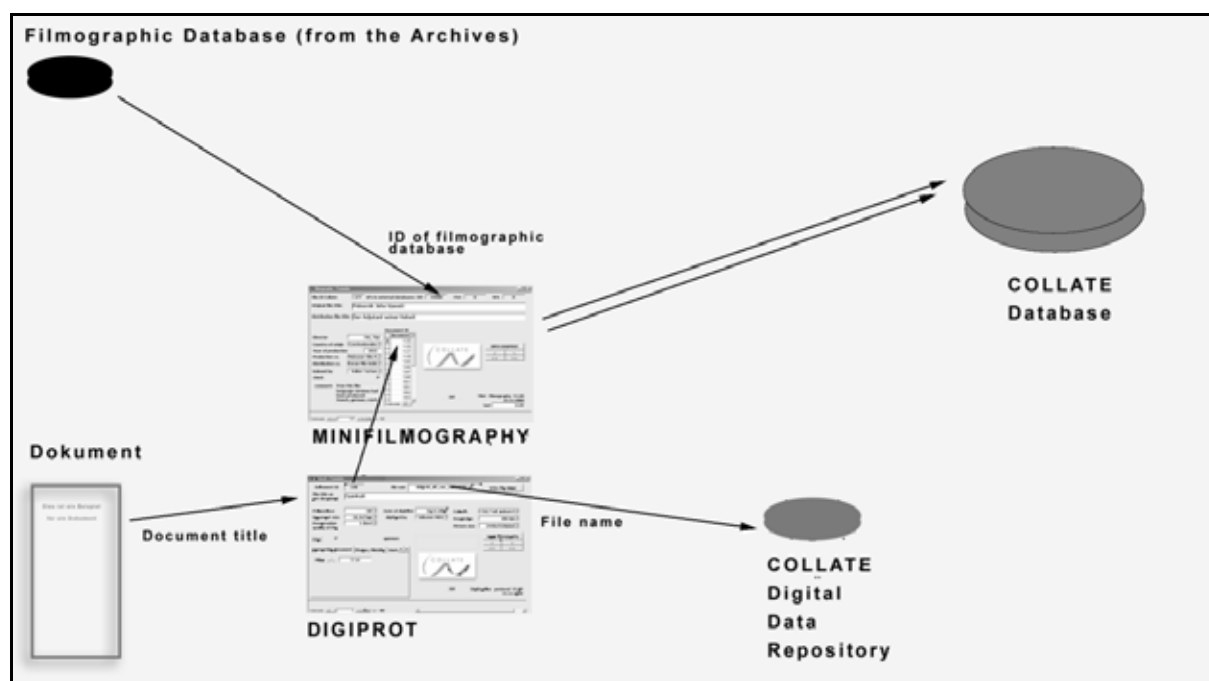


Figure 6 Organization of digitization data

A minimal formal document classification has already been undertaken, and is being represented using DIGIPROT and the Mini-Filmography. This also applies to data about document size, origin, etc. The next step was to develop a comprehensive task and domain model (*Workpackage 3*) and knowledge-based tools for content-based classification and indexing of the documents (*Workpackage 5*). Access via content is the precondition for the advanced retrieval techniques and annotation possibilities.

So far, the main focus during digitization had been on censorship documents, in particular official documents, such as censorship cards and censorship decisions. This was sufficient for the overall digital collection of COLLATE, but these document types are insufficient for the “core collection”, since a more rich corpus was needed to allow users a deeper understanding and analysis of the film censorship procedures and significance of film in that respect. For this, further text documents from contemporary magazines, correspondence, etc., had been collated so as to facilitate a reconstruction and evaluation of these complex procedures. One goal, for example, was to include production documents from the respective production companies in order to be able to present the decision-making processes themselves.

Given that censorship applied not just to the films, but also to the advertising material, i.e., photographs and posters, these too should be included for the purposes of the examination. Even if not subject to censorship, they can also serve, in general, as an additional source of information on (the reception of) a given film.

With the help of this collection of material combined with results from the film experts' and end-users' work with the material, it will finally be possible to reconstruct the institution film for a defined period of time, and make the results available to other interested scientists and the European public.

4.1.2 Document Representation and Indexing Schemes

Goals and results

The goals of WP3: *Task/Domain model and set up of the metadata base* and WP5: *Integration of knowledge tools* were to develop task-adequate multidimensional and syntactical document description schemes and to provide a comprehensive domain specific ontology as indexing terminology.

The resulting **logical document representation scheme** was elaborated. This scheme covers generic archival, librarian and documentary metadata models but respects also specific content and context related information interests of film scientists such as film life cycle information, person information and complex topical information.

The COLLATE **indexing terminology** was designed as a comprehensive cultural heritage **ontology** that integrates and harmonizes several existing standard terminologies, classifications and indexing aids for texts and graphical material as well as a proprietary film censorship vocabulary and film title concordance.

Both serve as tools for content-based indexing beyond bibliographic control and archival registration.

The logical document representation scheme was translated into adequate **data schemes and models** for document representation and implemented in the XML content manager.

Logical document representation schemes

Digitization of documents produced an uniform document format and facilitated the document saving, exchange and delivery considerably, but effaced at the same time the rudimentary document grouping and organization practiced in the film archives (and described in *Deliverable D1*, section 2.2: *Present situation of the COLLATE archives*). More, the new multilingualism of the collection and the enhancement by foreign document stocks devaluated the mental collection models and inventory knowledge the film archivists used for orientation before.

Therefore, collection and document organization had to be based on manually produced metadata representations of documents and collections. These representations provide an **integrated access to text documents as well as pictorial material on the base of assimilated document descriptions**.

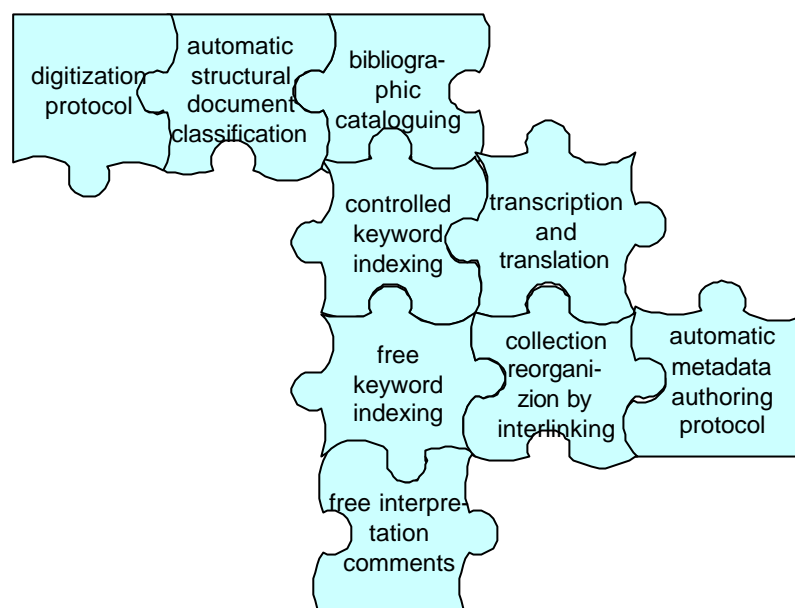


Figure 7 COLLATE's document description puzzle

Figure 7 visualizes the complex metadata puzzle for document description in the COLLATE system. Some of the indexing tools as bibliographic cataloguing schemes or the controlled keyword systematic are internally again composed from several models. The integration comprised the conceptual, workflow and interface levels as well as the integration of document information by retrieval methods.

The development of cataloguing and indexing schemes started from scratch with the elaboration of a comparative **document typology for the film censorship domain** by the three archives. The challenge of this step was the definition of document classes which were equivalent in their function in the censorship procedures in the different countries, although highly different in their formal properties.

A special problem was the identification of the specific censored film in question. Films often are referred to by different film versions, e.g., different language versions or distribution titles cited in the (censorship) documents. These alternative titles were reduced to an unequivocal key film title for all 100 films of the COLLATE "core" collection in a intellectually investigated **film title concordance**.

Cataloguing forms orientate themselves to the *Dublin Core* metadata scheme for Web publishing, but are enhanced by archive-specific data such as provenance of the original or differentiated relations between case dossier, document and document parts. For simple search, all cataloguing data are mapped to a Dublin Core conform metadata object for each COLLATE document. Furthermore, entry labels that are close to the vocabulary and concepts of the specific document type make cataloguing more intuitive for untrained indexers and more specific for information retrieval.

Integration of indexing tools and techniques was carried out by conceptual integration of terminologies, combination of techniques to an indexing workflow and visualization and implementation of indexing schemata, tools and workflow in a suite of indexing forms on the interface level.

Figure 8 sums up document description procedures und relevant tools and indexing aids in the COLLATE system.

Document description procedures	Tools
digitization protocol	DIGIPROT

automatic structural document pre-processing	WISDOM++ document processing system (see in detail COLLATE deliverable D4.1)
bibliographic cataloguing	<ul style="list-style-type: none"> • document typology • cataloguing schemata • film title concordance
keyword indexing with controlled terms	<ul style="list-style-type: none"> • controlled, integrated COLLATE terminologies
keyword indexing with free terms	dynamic vocabularies
transcription	transcription form
translation	translation form
free comments	annotation component

Figure 8 Procedures and tools for document description in the COLLATE system

The complex document indexing schemes serve for content-based and context-aware retrieval. They are **transferable** to any application that focuses on document organization and administration according to content and context, e.g. documentation applications in R&D environments, documentation for drug application or other authorization processes. Direct **reuse and exploitation** is possible for the complete cultural heritage domain.

Integrated indexing terminology

A very explicit requirement of COLLATE users (see *Deliverable D9.2: Study of end-users needs and behavior*) was the introduction of document access by subject to the archive specific document access by provenance.

Analyzing the archive's interests in content analysis doing indexing tests together with the film archivists we found a dichotomy between a general filmographic subject indexing interest and a narrower film censorship subject indexing interest.

The **filmographic view** uses film censorship documents to distillate out new information about films, like lost scenes, identification of unknown actors, the existence of language versions and others. Filmographic data help to identify non-ambiguously the film version at hand and serve to (immaterially) reconstruct a film work. The difficulty of indexing film subjects in censorship documents consists in the fact that filmographic information logically is an attribute of the film in concern and not of the censorship document. The censorship document only has a reference function for this kind of information extraction. This problem was solved by assigning a "correlation layer" to every elaborated set of indexing terms. Thus, a given set of indexing terms can refer to a censorship case, to the film in concern or to any fact of reality talked about in the censorship document.

The **censorship history view** concentrates on the argumentation and ideology of censorship acts, the clarification of censorship responsibilities, the social and political embedding of the censorship institutions or the phenomenological description of censorship practice in different countries and societies.

Subject indexing could not go back to any introduced classifications, keyword lists, thesauri or ontologies in the three film archives - apart from a Czech keyword list for image indexing. Therefore

COLLATE's approach was to combine several standard cataloguing rules, document classifications and terminologies with a application-specific start vocabulary for film censorship and a forum for further terminology development.

The **conceptual integration of terminologies** followed a layered model that specialized generic concepts such as *agent* to domain-specific concepts such as *film agent* or *film-censorship agent* and finally to application-specific concepts such as *film-censorship expert*. We understood the cultural heritage sector (museums, archives and libraries) as top-level domain, film archives as sub-domain and film-censorship document archiving as application level. Our idea was that every user should find the appropriate conceptual information access level according to her more general or more specific background knowledge.

Standardization and interoperability: COLLATE preferred the application of standardized or widely accepted knowledge tools instead of the development of autonomous and independent rules or terminologies. We developed an individually terminology only for film censorship concepts because all checked thesauri and classifications presented an evident gap for this subject (see the state-of-the-art analysis in *Deliverable D1*). Picking up well introduced knowledge tools we gain semantic and partly technical (RDF/RDFS) interoperability with other applications and benefit of the professional maintenance and updating of these knowledge tools.

Accepting the requirements of archives for complex and **multi-layered content indexing, syntactical indexing** was introduced. Syntactical indexing is much more expressive as simple coordinate indexing. In syntactical indexing, each index term is regarded as an attribute value that is qualified by the relevant attribute category, for example *agent type: film director* or *censorship argument: corrupting the youth*.

The COLLATE indexing terminology integrates and harmonizes the following pertinent standard vocabularies.

- *The Dublin Core*. Although *Dublin Core* is a cross-domain document description standard and not specific enough for cultural heritage applications, a relevant selection of cataloguing fields in COLLATE are mapped to the very familiar Dublin Core elements. They are presented in the quick search for internal and external users that guarantees easy information retrieval access.
- *FIAF's International Index to Film Periodicals: Subject Headings* provided the film-domain specific technical terminology for subject and topic indexing together with the *Library of Congress: Thesaurus for Graphic Material I: Subject Terms*.
- *The Library of Congress: Thesaurus for Graphic Material II: Genre and Physical Characteristics Terms* was exploited to define content-related and style-dependent picture genres for the picture typology used for picture cataloguing.
- The censorship specific indexing is covered by a categorized keyword list and scheme that was empirically built and continually enriched.
- The *ABC-Model and Ontology* served as background scheme for ontological analysis and structuring the vocabularies (see *Figure 9*).

The conceptual integration was effected top-down from the generic concepts over domain-specific notions to application-specific concepts. The generic *ABC Model and Ontology* presents the top level; at the same time, the *ABC-Model* provides the Cultural Heritage Sector domain specific concepts (that were originally adopted from the *CIDOC Conceptual Reference Model*). The specific concepts for the film archive sub-domain come from the *FIAF Subject Classification* and the *Library of Congress Thesaurus for Graphical Material I and II (LOC TGM I, II)*. Finally, the autonomous COLLATE vocabulary covers the film censorship concepts on the application level of the model.

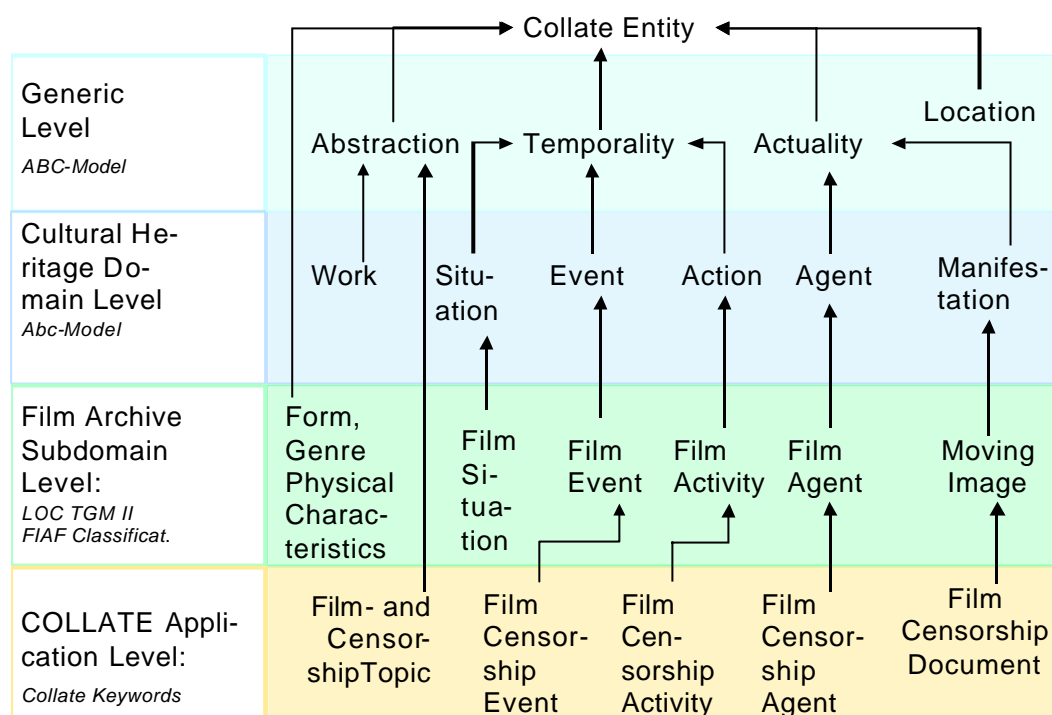


Figure 9 Conceptual integration of COLLATE indexing vocabularies

The integration process resulted in an ontology that formally implemented as RDFS-ontology (resource description framework schema) with the ontology construction tool *Protégé 2000*. This ontology was simplified and translated into syntactical indexing using a kind of indexing formula (see table below). The indexing formula is visualized in a complex keyword indexing form tab.

Agent information		Censorship subject				Film/Picture subject	Reality subject
Agent Type	Agent Name	Censored Film/Picture Part	Censorship Action	Censorship Argument	Censorship Topic	Film/Picture Topic and Subject	Reality Topic and Subject

The vocabulary for every category can be completed with free keywords since entries of free keywords can be categorized into the following groups:

The available COLLATE ontology complies to a large extent with standard vocabularies to support semantic interoperability. Due to the fact that the COLLATE ontology is designed in a multi-layer architecture that ranges from the generic top level down to an individual and specific level *the model is highly transferable* to others than the cultural heritage domain

Methods of conceptual ontology integration are basic and will be **reused** in the future in a multitude of knowledge engineering projects.

Data model for document and domain representation

For the logical representation of the various types of data and metadata for COLLATE we decided to take a hybrid approach. In cases where the data to be stored was either system-dependent (e.g., digital

watermarks, user authentication) or could be considered as rather static (imported facts: DIGIPROT and filmographic information) we decided to define relational database schemata to design their representation within the system. But this rather static structure comes to its limits when dealing with more dynamic domain metadata objects like, for instance, the cataloguing schema which had undergone several revisions. Therefore, we employed XML schema as a more flexible and scaleable alternative to static database relations. The various COLLATE domain objects (e.g., cataloguing information, keywords, annotations) are represented as instances of appropriate XML schema, uniquely identified by their URIs. One of the main tasks for the XML Content Manager is to manage these various types of metadata domain objects and map them dynamically onto the underlying relational database system. The COLLATE domain and database model is described in detail in *Deliverable 3 "Task/Domain Model and Set Up of the Metadata Base"*.

Internally, the structured, document-centered discourses are implemented using RDF descriptions, which relate the annotations to the objects they describe according to the COLLATE RDF Schema (see <http://www.collate.de/RDF/collate.rdfs>). By employing the Resource Description Framework we are able to interrelate the various data and metadata objects in a dynamic way, i.e. depending on users' information needs our system is able to generate corresponding views on the documents and their associated metadata in a distributed repository.

4.2 Document Pre-Processing Modules

4.2.1 Automatic, Intelligent Document Processing (WISDOM++)

Original objectives of *Task 4.1* "Document Processing & Classification" were the design and implementation of components for the automatic detection of the layout structure of the documents and for the automatic classification of the documents on the ground of their layout structure, to be integrated in the overall architecture. The development of the above components involved a number of related activities, such as preliminarily designing a proper representation of the document structure and successively training the component for document analysis by means of example documents. In particular, the layout structure of the documents was to be obtained by means of a component for labeling the document blocks. Then, once the layout structure had been found for a set of training documents, another learning tool was to be constructed to induce rules for the automatic classification of documents on the ground of spatial and perceptual factors.

The first step has been the choice of the database management system to be used for the main COLLATE databases residing at IPSI and UNIBA. Such a database structure had to underlie the distributed 3-tier architecture of COLLATE. While IPSI had a strong competence in using the INFORMIX DBMS in previous projects, UNIBA had gathered positive experiences with the ORACLE system. After some discussion, an agreement was found upon having a consistent scheme in order to reduce the overhead of required adoptions to the XML Content Manager by Sword. For this reason IPSI finally approved UNIBA's recommendation for ORACLE, which was set up in Darmstadt as well. Note that this decision was only based on economic reasons, i.e. COLLATE is not restricted to this database system architecture. Different architectures might be adopted in future applications.

UNIBA worked initially on its proprietary tools for automated intelligent document processing in order to evaluate the possibility of embedding facilities that help the user to manually annotate each document component according to its logical/semantic meaning. To this purpose, UNIBA worked on different documents related to the specific domain in order to understand the structure and possible solutions to submit the material to the classification systems. The major problems that immediately came up concerned the variety of formats, colors and typing organization of the original documents and, consequently, of the digitized ones. This raised the need of modifying the system WISDOM++ in order to enable the acquisition and correct visualization of documents with a format different from A4.

For a number of months, UNIBA has heavily worked on this task, obtaining good results towards a solution. Indeed, document image processing in a completely automatic way is much more complex than people could imagine. What has been done in that period, in short, is:

1. Some of the acquired images were transformed into a format suitable to WISDOM++, that is, black-and-white 300dpi images. This was necessary since processing a color 24Mb document image is much more computationally demanding than working on a 1Mb binary image. This work has not been completed, although enough document images were processed in order to get an idea of the problems to face.
2. Some document images have been processed with WISDOM++ to extract the layout. Moreover a specific COLLATE user was created in the WISDOM++ system, and some classes and some logical labels of interest have been associated to that user. In this way it has been possible to check what the system was able to recognize on binarized images.
3. In some cases it was found out that the layout had to be modified in order to extract some components with a logical meaning. WISDOM++ was extended in order to provide the user with a semiautomatic tool for the correction of the results of the global layout analysis. The extended system was made operative.
4. The possibility of learning rules for the layout analysis correction had been investigated. As a first outcome, however, a problem might be the limited number of document images already available. Some results, concerning the registration cards, are reported in a first technical report.

Let us specify the above activities in more detail. UNIBA selected a subset of the available digitized documents coming from the archives in order to obtain sample descriptions on which running the learning systems to gather preliminary directions on the automatic classification capabilities. Such a subset, consisting of 84 documents belonging to 3 different document classes, plus a reject class, included the documents showing an acceptable layout standard, and hence more suitable to be processed (in fact, the overall quality of the entire set was quite low). A number of them had previously been tagged by archive experts as regards both their class and the meaning of their significant layout components. The main features of processed documents are shown in Table 1 and some sample documents in **Error! Reference source not found.**

Table 1: Main features of processed documents

Source	Type	Size (pixel)	Resolution (dpi)	Image size (mm)
FAA	Censorship cards	4836×3408	600	204,72 × 144,27
DIF	Censorship cards	1710×1212	300	144,78 × 102,62
DIF	Censorship cards	2460×3474	300	208,28 × 294,13
DIF	Newspaper articles	Not fixed	Not fixed	Not fixed

Such a subset was preprocessed and segmented for use in the training phase. Preprocessing consisted in: 1) the transformation of original color images into black-and-white (binary) images; 2) the evaluation and removal of skew; 3) estimation of the complexity factor. Segmentation consisted in: 1) segmentation of the preprocessed image; 2) classification of image segments (blocks) in order to separate text from graphics; 3) automated analysis of the layout and manual correction of the results of the global analysis process. More specifically, the acquired images were transformed into a format suitable to WISDOM++, that is, black-and-white 300dpi images. This was necessary since processing a color 24Mb document image is much more computationally demanding than working on a 1Mb binary image. This work was not completed, although enough document images were processed to get an idea of the problems we had to face. Enlargement of the processed data set will heavily depend on the availability of other documents, provided by the archives, which have a sufficiently good layout.

To process the selected document images for layout extraction, a user was created in the WISDOM++ system, and the association of some classes and logical labels to that user was simulated. In this way it was possible to check what the system was able to recognize on binarized images. In some cases it was found out that a modification of the layout was necessary in order to extract components with a logical meaning. Hence, WISDOM++ has been extended in order to provide the user with a semiautomatic tool for the correction of the results of the global layout analysis. The extended system is now operative. Another extension of WISDOM++ aimed at making it able to perform external calls of the OCR TextBridge 2.0 which supports both Czech and German dictionaries.

Moreover, based upon the preliminary results obtained on the sample set, a suitable language has been produced in order to describe the documents according to their size and the type, size and relative position of their layout blocks. This step was also preliminary to the investigation of the possibility of automatically learning rules for the layout analysis correction. Indeed, a theory was learned for the document layout correction, in order to help WISDOM++ in improving the automatically recognized layout structure when processing new documents. Another experiment aimed at learning rules for automatically classifying documents into one of the three classes considered.

After having detected the layout structure, the logical components of the document could be identified, such as *applicant*, *registration number*, *authorization* of a censorship card provided by FAA. The logical components can be arranged in another hierarchical structure, which is called logical structure and is the result of repeatedly dividing the content of a document into increasingly smaller parts, on the basis of the human-perceptible meaning of the content. The leaves of the logical structure are the basic logical components, such as authorization and registration number. The *applicant* of an FAA censorship card encompasses the film title, length, genre and producer and is therefore an example of composite logical

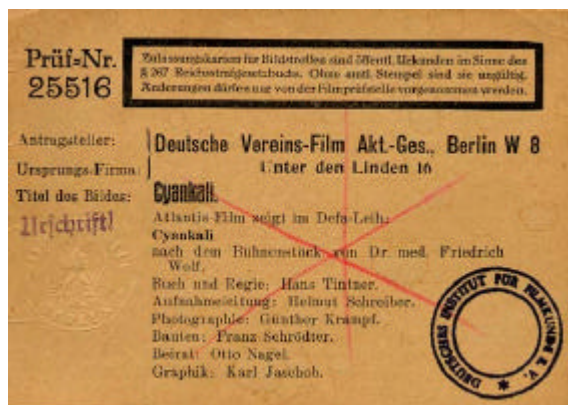
component. Composite logical components are internal nodes of the logical structure. The root of the logical structure is the document class, such as *censorship card* provided by FAA. Currently, WISDOM++ supports two-level logical structures, where the document class is the only composite logical component.



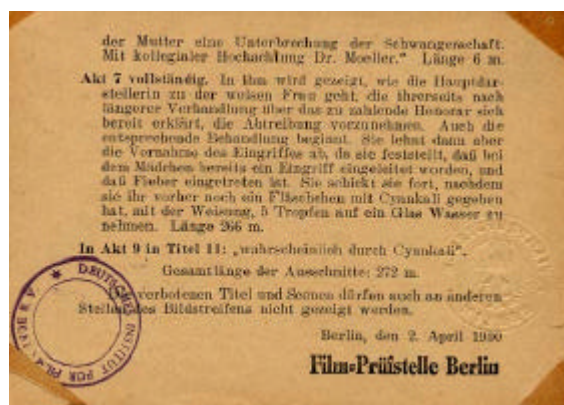
First page of a FAA censorship card



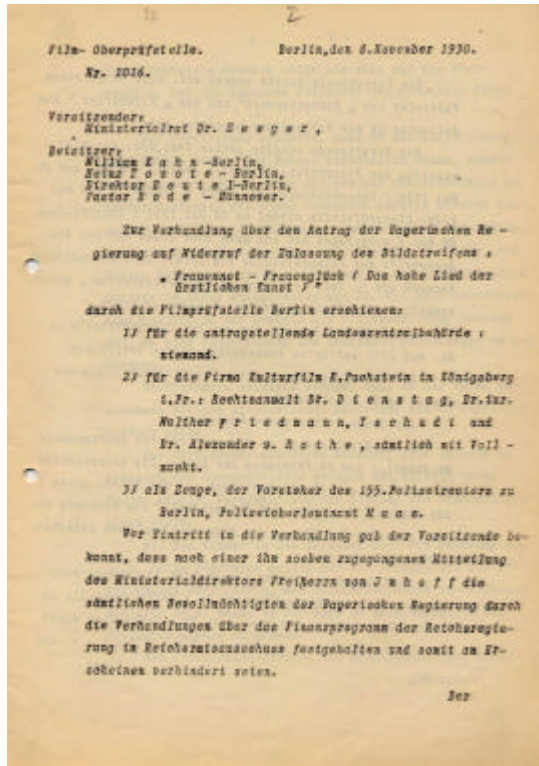
Second page of a FAA censorship card



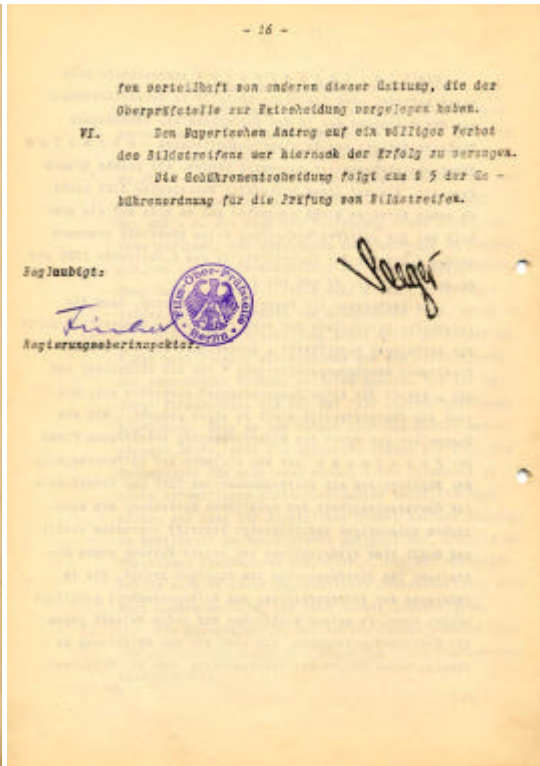
First page of a small DIF censorship card



Last page of a small DIF censorship card



First page of an A4-sized DIF censorship card



Last page of an A4 DIF censorship card

Figure 10 Examples of documents to be processed

The problem of finding the logical structure of a document can be cast as the problem of associating some layout components with a corresponding logical component. In WISDOM++ this mapping is limited to the association of a page with a document class (*document classification*) and the association of components of a logical hierarchy with second layout frames (*document understanding*) (see Figure 11). Since the kind of logical components that can be found in a document depends on the class of the document at hand, document classification precedes document understanding. By performing document image classification and understanding, WISDOM++ actually replaces the low-level image feature space (based on geometrical and textural features) with a higher-level semantic space. Query formulation can then be performed using these higher level semantics, which are much more comprehensible to the user than the low level image features. Promising results have been reached applying the learning systems available at UNIBA (ATRE and INTHELEX), although a problem might be the limited number of document images already available and the poor quality of part of them.

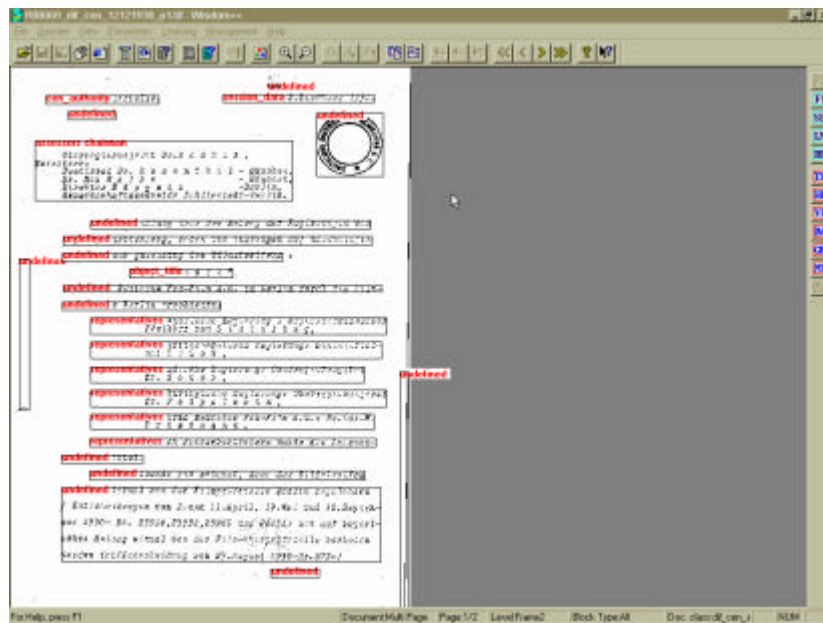


Figure 11 Automated document understanding: result of matching process against learned rules

One specific challenge for the automatic classification and understanding comes from the low layout quality and standard of such a material, which introduces a considerable amount of noise in its description. As regards the layout quality, it is often affected by manual annotations, stamps that overlap to sensible components, ink specks, etc. As to the layout standard, many documents are typewritten sheets, that consist of all equally spaced lines in Gothic type. One peculiarity in INTHELEX is the integration of multistrategy operators that may help in the solution of the theory revision problem by pre-processing the incoming information. Namely, deduction is exploited to fill observations with information that is not explicitly stated, but is implicit in their description, and hence refers to the possibility of better representing the examples and, consequently, the inferred theories. Conversely, abduction aims at completing possibly partial information in the examples (adding more details), whereas abstraction removes superfluous details from the description of both the examples and the theory. Thus, even if with opposite perspectives, both aim at reducing the computational effort required to learn a correct theory with respect to the incoming examples. Such a situation suggested the application of multistrategy reasoning capabilities in INTHELEX, such as abduction and abstraction, to the COLLATE environment. While the former can make the system more flexible in the absence of particular layout components due to the typist's style, the latter can help in focusing on layout patterns that are meaningful to the identification of the interesting ones, neglecting less interesting details.

Experiments demonstrated that application of these features actually improved the performance of the system. One characteristic in learning rules for document image understanding is that rules should reflect dependencies between logical components to enable a context-sensitive recognition. This aspect

is implemented in ATRE which can autonomously discover concept dependencies. In fact, the search space explored by ATRE is a forest of as many search-trees as the number of concepts to learn, in this case semantic labels of blocks on the base of layout information. The forest can be processed in parallel by as many concurrent tasks as the number of search-trees. Each task traverses the specialization hierarchies general-to-specific, but synchronizes its traversal with the other tasks at each level. Using this strategy ATRE can interleave the learning of the definition of a concept with the definition of another concept and find possible dependencies among them. Of course, the more the quality of document images is high and the number of examples is large, the more the system could exploit this capability.

One specific challenge for the automatic classification and understanding comes from the low layout quality and standard of such a material, which introduces a considerable amount of noise in its description. As regards the layout quality, it is often affected by manual annotations, stamps that overlap to sensible components, ink specks, etc. As to the layout standard, many documents are typewritten sheets, that consist of all equally spaced lines in Gothic type. One peculiarity in INTHELEX is the integration of multistrategy operators that may help in the solution of the theory revision problem by pre-processing the incoming information. Namely, deduction is exploited to fill observations with information that is not explicitly stated, but is implicit in their description, and hence refers to the possibility of better representing the examples and, consequently, the inferred theories. Conversely, abduction aims at completing possibly partial information in the examples (adding more details), whereas abstraction removes superfluous details from the description of both the examples and the theory. Thus, even if with opposite perspectives, both aim at reducing the computational effort required to learn a correct theory with respect to the incoming examples. Such a situation suggested the application of multistrategy reasoning capabilities in INTHELEX, such as abduction and abstraction, to the COLLATE environment. While the former can make the system more flexible in the absence of particular layout components due to the typist's style, the latter can help in focusing on layout patterns that are meaningful to the identification of the interesting ones, neglecting less interesting details.

Experiments demonstrated that application of these features actually improved the performance of the system. One characteristic in learning rules for document image understanding is that rules should reflect dependencies between logical components to enable a context-sensitive recognition. This aspect is implemented in ATRE which can autonomously discover concept dependencies. In fact, the search space explored by ATRE is a forest of as many search-trees as the number of concepts to learn, in this case semantic labels of blocks on the base of layout information. The forest can be processed in parallel by as many concurrent tasks as the number of search-trees. Each task traverses the specialization hierarchies general-to-specific, but synchronizes its traversal with the other tasks at each level. Using this strategy ATRE can interleave the learning of the definition of a concept with the definition of another concept and find possible dependencies. Of course, the higher the quality of document images and the number of examples, the more the system can exploit this capability.

Promising results obtained in preliminary experiments are reported in the Deliverable D4.1, part 1; subsequently, they have been confirmed by larger experimentations. Anyway, there were problems that still needed to be resolved. First, the documents at hand strongly suggested an extension of WISDOM++ with a segmentation algorithm that is able to handle properly color images as well as images of forms, where textual content is typically surrounded by frames. Second, it was necessary to investigate more deeply the application of a learning system devised to solve classification problems, like ATRE, to a typical planning task, such as performing a sequence of actions that lead to the desired goal. In particular, this is critical in the extraction of knowledge to support the layout analysis, where the system is asked to learn the sequence of corrective actions of the user. Third, it was important to reach a tight integration of the OCR with WISDOM++, in order to simplify user interaction and to take advantage of font information during the HTML/XML rendering process. Finally, the problem of incrementally refining the set of rules generated by ATRE when new observations are made available had to be investigated: the user could require the intervention of the incremental system INTHELEX for revising the knowledge base when a lot of reject cases appear, indicating a shrinkage in the system performance.

In order to confirm preliminary results on document pre-processing, classification and understanding, UNIBA enlarged the set of classes and documents that make up the experimental dataset, including a number of new scanned images coming from the archives. Such an enlarged dataset was pre-processed through WISDOM++ in order to obtain the first-order descriptions needed to run the learning systems.

Then UNIBA continued the process of document image understanding, by means of machine learning methods, in order to semantically label and transform relevant logical blocks into XML format. The rules have been automatically learned by the system on different "test sets". The results of this process are semantic labels in XML format representing annotations that are stored in the metadata repository by means of suitable Web Services provided by the XML Content Manager.

Problems arose from some documents due to the poor quality of some of them, the existence of noise in the document images, the need of deleting useless vertical and horizontal lines, etc. This caused some delay in starting the experimentations: new pre-processing algorithms had to be developed in order to filter the noise and eliminate the lines during the acquisition step. Then, various experiments were run on the data obtained this way. Some concerned the induction of rules for the automatic correction of the document layout structure identified by WISDOM++; others aimed at learning definitions for both the document classes and the significant semantic labels in the considered documents. The intermediate experimental results also suggested improvements that were implemented in both WISDOM++ and the learning systems. UNIBA also worked on porting WISDOM++ from an MS Access-based platform to an ORACLE-based one. Indeed, WISDOM++ was originally developed to interface a Microsoft Access2000 database, where data on both users and documents are stored. The interface was based on the ODBC standard.

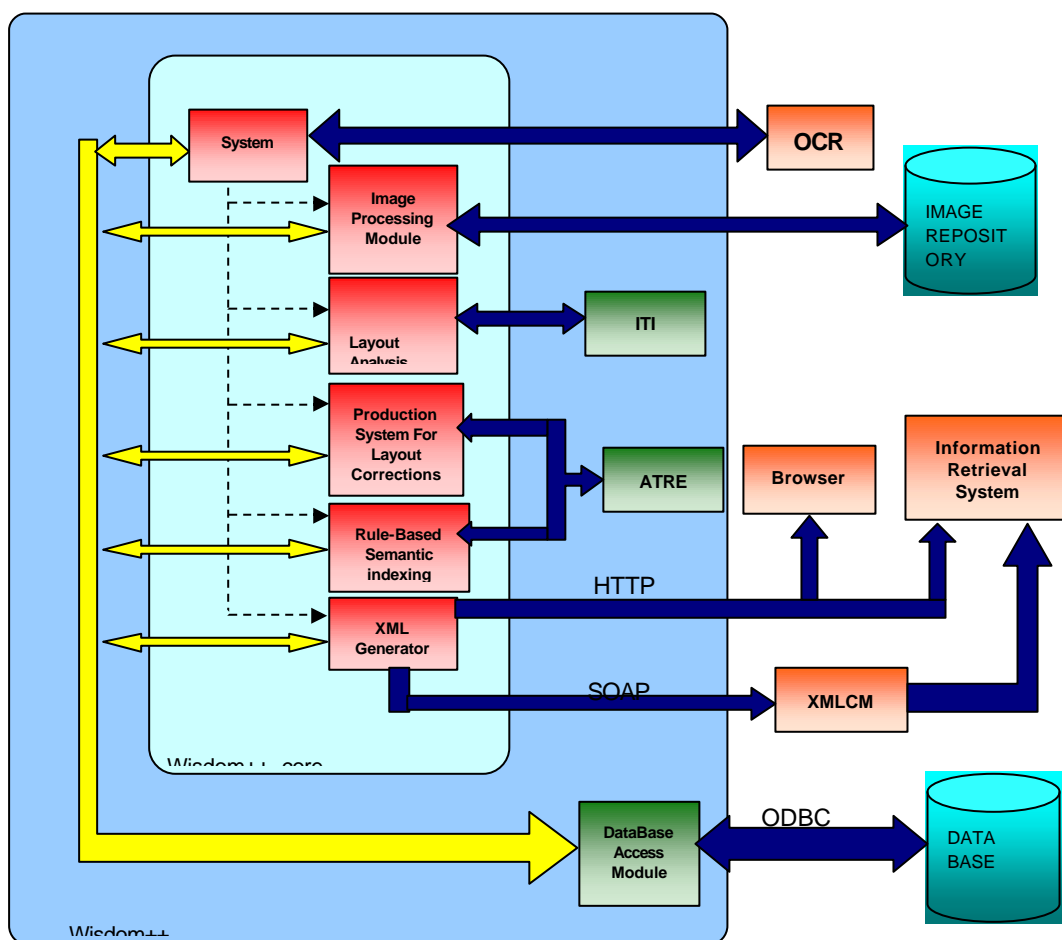


Figure 12 WISDOM++ architecture

However, the later decision of storing documents in an Oracle database lead to a considerable effort for unexpected changes in the classes that interface the database, due to the differences existing both in the data definition/manipulation languages and the query languages of the two DBMSs. For these reasons some activities were slightly delayed: above all, the "extended" test on DIF documents and the installation of the system at the archives and training of the personnel. In order to complete these activities an extension of the period initially planned was necessary (December 2002).

As a further improvement, motivated by considerations based on the results of a deep feasibility study, UNIBA acquired, customized and integrated in WISDOM++ a commercial OCR (ABBYY FineReader OCR) which supports special features for handling handwritten and poor quality documents. The learning systems used for automatically acquiring the rules to classify the documents were extended: in fact the complexity of the involved document classes, due to the low layout quality and standard, often significantly raised the runtime of the symbolic learning systems in charge of the induction of theories for classification, interpretation and layout correction. This led to focus part of the research on the development of new techniques that could speed up reasoning in first-order logic, and resulted in a new theta-subsumption and resolution procedure that is able to provide all solutions to the coverage problem with significantly improved performance with respect to classical SLD-resolution performed by Prolog. Moreover, the structure of the search space was also analyzed in order to assess the possibility of defining efficient and effective operators for theory refinement.

Other extensions concerned WISDOM++ and the DB schema (in order to include the automatically learned classification rules and to optimize access by proper indexing and querying techniques). As to WISDOM++ the extensions concerned the capability of automatically transforming input documents, in any format, into its working format and the possibility of learning rules for the layout analysis correction. In the system WISDOM++ the layout analysis is performed in two steps: firstly the global analysis determines possible areas containing sections, figures, tables etc. and, secondly, the local analysis groups together blocks that may fall within the same area. The result of the second step depends on the quality of the first step. Also, the result of the layout analysis strongly influences the result of classification and understanding accuracy. Following these considerations, WISDOM++ has been also endowed with new functionality and features that are able to support the user during the correction of the results of the global analysis, by automatically generating training examples of action selections from the sequence of user actions rules and then by learning action rules for layout correction. Moreover, a new version of the representation language for the description of layout modifications, better suited for learning correction rules, was developed and tested.

In November 2002 the integrated system WISDOM++ was installed at the DIF archive, and training of the archivists was started. UNIBA also carried out a study on the state-of-the-art algorithms for handling colors in document images, in order to assess the possibility of adding such features to WISDOM++. Consequently, on the basis of the analysis of different algorithms for handling colors in document images, UNIBA has developed a prototype module for color quantization able to work on a number of colors less than 16; such a module has been tested on the experimental dataset of documents. This feasibility study has led to a lot of open problems, like the identification of the meaningful set of colors and the investigation on how learn it, the individuation of a strategy to merge different layout structures associated to different colors and investigation on how enrich the language description with color information to take it into account in the classification and understanding phases.

The interface of WISDOM++ was changed in order to fully exploit the new release of the XML Content Manager persistence layer based on Oracle 9i, and a fault management module for correcting possible server-side errors was introduced on the client-side. UNIBA completed the COLLATE document dataset for the application of learning techniques by including two more document classes from the NFA archive, and ran on it the learning systems to produce labeling rules for the classes and the significant components. The automated annotation function (classification and understanding) was tested in WISDOM++ for documents made available by the NFA archive.

UNIBA carried out the integration between WISDOM++ and XML CM in cooperation with SWORD. With this integration WISDOM++ is able to store image pre-processing results into XML CM repositories. SWORD developed a set of Java servlets for this integration in order to allow document retrieval directly via HTTP request (besides allowing it also by SOAP). This makes possible to access the output produced by WISDOM++ via a Web browser, just as it was when WISDOM++ ran as a stand alone application (i.e., before integration).

UNIBA continued to enrich and complete the COLLATE document dataset for the application of learning techniques by including two more document classes from the NFA archive. The complete dataset was processed by WISDOM++ and by INTHELEX and ATRE in order to learn new sets of labeling rules for the classes and the significant components.

Finally, WISDOM++ was modified by integrating the error management functionality and aligning the multi-user features in order to improve its interoperability with the XML CM. Such a new version of WISDOM++ was then integrated with the XML Content Manager, and the final system architecture (see *Figure 12*) was fine-tuned, tested and monitored for assessing the correctness of the overall behavior and the improvement of the communication functionality.

4.2.2 Image and Video Analyses Tools

Rationale

The digital COLLATE collection built up consisted not only of digitized text documents (about 18 000 pages of censorship documents, legal texts, press articles, etc.) but also included 1000 digitized photographs and other pictorial material related to the core collection. Additionally, some digital video fragments were made available for our research experiments with automatic video and image analysis. All material was manually catalogued by members of the COLLATE archives, and the several thousands of documents related to the core collection were additionally indexed by keywords and annotated in detail. Within *Workpackage 6* we developed semi-automatic indexing methods for the non-textual documents to support retrieval of pictorial material that has not been indexed manually by the COLLATE users.

Manual indexing of pictorial material by subject matter and semantic content could solve an important requirement, i.e. to allow content-based retrieval of non-textual documents. The main problem with such a procedure, however, is the extremely high expenditure of time for the manual indexing and the often inconsistent, varying indexations which different persons assign to the same document. Therefore, innovative methods for automatic image indexing can offer a highly interesting alternative, especially to those digital archive providers who aim to provide very large image collections that should not only be accessible by formal/bibliographic attributes but also – at least to some extent – by semantic content.

In the following we sketch briefly our research results, and describe the methods and procedures we applied for automatic video and image indexing in COLLATE.

A three-level model for automated image indexing

A recent study (cf. Eakins & Graham 1999) suggests an image retrieval model that distinguished image retrieval systems according to three levels of query complexity, i.e. the primitive, the logical and the abstract level.

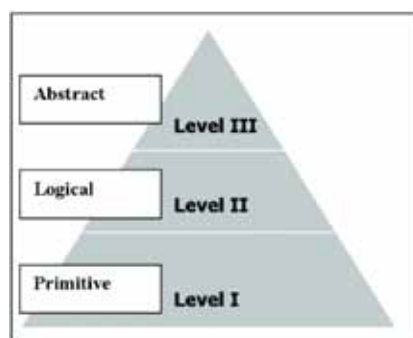


Figure 13 Eakin's three level model of image retrieval systems

- On the primitive level the query features are extracted by automatic algorithms on primitive picture characteristics and stored in a data structure or a data vector. Primitive picture characteristics, as for example the color contained in a picture, texture, or form characteristics, can be processed without additional information from a picture.

- On the logical level additionally objects and scenes within pictures are described. Objects in an image could be, for example, two human beings in a scene named “a walk on the beach”. Since the picture of the human being might have different color distributions, and could also have been taken from different perspectives, a recognition of such objects is a much more difficult topic and needs more information, which has to be trained to the system on the basis of a primitive picture algorithm.
- Abstract level queries are on a higher level of complexity. Within the complex level the images are described with respect to their symbolic meanings. The query “romantic scenes” is one possible query on this level. The picture “a walk on the beach” could be one possible resulting image, and “a candlelight dinner” another one.

Within our work for COLLATE we made use of this three level distinction to characterize the information needs of the participating archives. The model allows us to place our approach in the context of ongoing research in the area of image retrieval.

Development of the rule-based image retrieval system

Within the COLLATE project we developed a specific classification scheme for visual data in cooperation with the archive users (cf. Appendix to COLLATE *Deliverable 0*). This scheme was not intended to be used for the manual classification/indexing of the digital photo collection, though it could have been used for this purpose as well. With this scheme we mainly aimed to support the learning mechanisms of the automatic indexing tool. The classification/indexing scheme uses descriptors mainly from the logical and partly from the abstract level (e.g. Topic/human relationship/love) following the three-level model described above. Therefore, our system design included the generation of the primitive level features and the combination of those with logical and abstract descriptions of an image.

Within a previous IPSI project HERMES (Hollfelder et al 2000) we performed some first, preliminary experiments with a small classification scheme for similar purposes but with a completely different scope. In HERMES the classification was restricted to mainly visual and graphical characteristics of the images (e.g. light and contour aspects, dimension (2D or 3D), existence of light or dark objects in the foreground/background, and some very basic object typology (see Thiel et al. 1999a, Thiel et al. 1999b). As it turned out such low-level characteristics were insufficient for our present application context and potential search interests of system users. Hence a new indexing scheme was developed with a focus on more specific subject matter issues typical for the current domain.

Our approach to enabling conceptual queries on videos and still pictures is divided into the following steps: first, a video is divided into single scenes by a scene detection algorithm. The video stills indexing (analysis) system then employs a number of feature detection algorithms on selected frames. The results of these algorithms - called the feature extraction values - are used to find rules which map the values to conceptual terms.

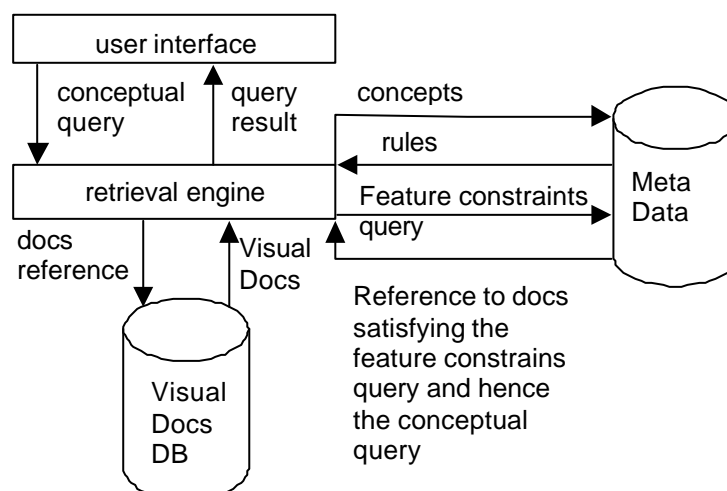


Figure 14 COLLATE image & video indexing approach

For rule generation we employ an empirical approach in which manually indexed images are used as a training set. Generated rules and extracted feature values are stored in the meta database. Note that the latter feature values are not restricted to the original sample set used for the rule generation process. Instead, feature analysis results obtained from the many times larger video collection are now used as the basis for semantic access. If the user poses a conceptual query, the retrieval engine analyses the query and maps it to a set of rules which are requested from the meta database. The rules are interpreted by an appropriate rule interpreter yielding specification of features extraction values to be searched for. If feature values matching the constraints can be found, the associated video parts are retrieved. The result is returned to the user as a ranked list of items.

The video and image analysis tools

The scene detection algorithm first divides the video scene into its single frames. Then, surrounding frames are analyzed for changes from frame to frame. If the resulting measurement value for such a change is higher than a threshold value, the system decides that this is the beginning of a new scene. To find smooth scene changes, we also analyze the changes between a frame and its sixth neighbor. In a detected scene we take the middle frame as a representative for the scene and start our image analysis tool.

In the case of conceptual image retrieval, we assume the queries to be expressions in propositional logic, e.g. "street scene \wedge daylight". Some intermediate steps may be required which employ heuristic retrieval rules, like "dimension: 3D \wedge objects: artificial \wedge source of light: natural \rightarrow street scene". Contemporary picture archives rely on manual annotations or indexing. Hence, the compilation of a rule base, although being a considerable investment, may pay off for large picture collections, which may grow or change quickly. The rules reduce high-level user concepts to combinations of simpler concepts, called "visual content descriptors", which represent abstract or class features of images, e.g., dimension (2D, 3D), type of light source, type of objects shown (e.g., natural vs. man-made) in the case of photos, but also differentiations between photos, paintings, cartoons, etc. For example, a landscape could be described by the descriptors "light source: natural light", "dimensions: 3D", "objects: natural", as well as by having a large blue area in the upper part of the image.

The content descriptors must satisfy two conditions: First, they must support the needs of the user community, i.e. it should be possible to express common user concepts in terms of content descriptors. Second, they must correspond to patterns of pictorial features which may be extracted from images in an automatic process. The translation of content descriptors into feature patterns is facilitated by a set of "indexing rules".

Being able to derive these rules (semi-) automatically whenever major changes of the collection occur – here we assume that a DL is usually supervised – enables us to cope with large and quickly growing collections, which then need only to be processed by feature analysis methods rather than be manually indexed. For this purpose, procedures using bitmap technology and statistics on elementary image features are well-known, and adopted here for our analyses.

In the image retrieval system we combined 17 algorithms based on the PBM format (portable bit map) to calculate such characteristics as the color distribution within an image, the surface texture of objects in an image, or values which express the degree of similarity in color distribution between two pictures.

Image analysis values of this type are usually stored in meta-databases, where they can be used as the basis for content-based retrieval. However, in most applications no abstraction or interpretation of these data is performed in such systems. User queries must either be formulated in terms of existing value ranges, or users must provide a sample image which is then analyzed in the same way to provide query values. Since exact matching is not very useful, often intervals are used. Compression techniques such as wavelets are sometimes used to abstract from overly detailed pixel patterns and to accelerate processing.

To enable the use of image analysis results for *more* than similarity search, they must be semantically interpreted. This can be accomplished through model-based object recognition which allows the identification of specific objects. The modeling effort is high and only justifiable in special cases, e.g., when the information needs of users can be precisely described a priori. In general, however, users of a digital library will search for various concepts, e.g., for objects, persons, events, and various motifs. In our approach, we first need to define retrieval rules --as described above--, which can be composed of appropriate content descriptors. Second, we need to derive the indexing rules which associate these content descriptors with feature patterns.

Methods for rule-based indexing and retrieval of images

Our approach to enabling conceptual queries on images is divided into the following steps: The image indexing (analysis) system employs a number of feature detection algorithms. The results of these algorithms – called the feature extraction values – are used to find rules which map the values to conceptual terms. For rule generation we employ an empirical approach in which manually indexed images are used as a training set. Generated rules and extracted feature values are stored in the metadata base. Note that the latter feature values are not restricted to the original sample set used for the rule generation process. Instead, feature analysis results obtained from the many times larger image collection are now used as the basis for semantic access. If the user poses a conceptual query, the retrieval engine analyses the query and maps it to a set of rules which are requested from the metadata base. The rules are interpreted by an appropriate rule interpreter yielding specification of feature extraction values to be searched for. If feature values matching the constraints can be found, the associated images are retrieved. The result is returned to the user as a ranked list.

As a part of our experiments, a number of feature-extraction and comparison algorithms were implemented using the PBM (portable bit map) collection of image processing software. Since texture-based classifications are very effective, the PBM texture module was selected as a promising tool.

For the rule generation, we used the following *empirical approach*: Starting point was a collection of 500 images manually classified by experts using the classification scheme developed for this purpose, i.e. the PiClasso image indexing user interface). The images were divided into two disjunct sets, a training set and a test set. We used the training set for rule generation and the test set to validate the rules.

The association between feature values and manually assigned index terms, i.e. the descriptors chosen from the classification scheme above, can now be accomplished using algorithmic statistical methods, ranging from exhaustive exploration to complex stochastic computation. Which method is applicable depends on the degree of aggregation applied to the original feature values. We start with p -dimensional vectors \vec{x} , containing the results from p different feature extraction methods, applied to the image. In the next step, the feature-extraction values may be aggregated to dynamically built constraints, e.g., ranges, or linear combinations of the feature values. If we regard ranges of scalar aggregated feature values, we can derive plausible rules to describe the content of pictures on a general level, by analyzing the feature-extraction values of the manually classified images by α -Quantile analysis. In a number of cases, however, no useful association pattern could be found due to the fact that the aggregation process was too coarse-grained. In this situation, it is useful to examine the original data, i.e. the feature vectors. In a first exploration study we employed a brute force exhaustive search (BF). The results being promising, we changed to a more efficient method.

In our next experiments, we used the Quadratic Classification Function (QCF) to calculate the probability that an image matches a classification item. The QCF gives a measure for the distance of the feature extraction values of an image and the mean values of a set of manually classified images. All of the details of this algorithm and the relevant formulas are described in COLLATE's *Deliverable 6*.

Implementation within COLLATE

Note that the inclusion of pictorial documents in the COLLATE system turned out to be problematic due to copyright problems faced by the archives. The relevant pictorial material originates mainly from the late 20s and early 30s, the archives owning copies of photographs and some posters, but the copyrights held by the film distribution companies and other institutions generally only expire after 80 years (!).

Identifying individual cases where the copyright-holding institutions do not exist any more would be extremely time-consuming and often even impossible. Therefore, the archives can not offer all digitized pictures and video films to the general public on the WWW but only to restricted user groups (e.g., other film scholars). However, they did provide IPSI with enough picture samples for COLLATE's internal research and testing purposes, i.e. for the automatic image indexing and retrieval experiments.

The archives consider pictorial material an interesting source for the description of film contents/scenes and assessment of film advertisement practices, but from the perspective of the planned censorship case studies it is secondary, i.e. offering some context information and vivid illustration of otherwise complex verbal descriptions. Manual cataloguing and indexing of pictures by the users concentrated on basic features and identification of agents/actors and events. Combined searches accessing user-generated metadata and the automatic rule-based indexing offers here an integrated approach to both content concepts and visual/graphical features of the picture documents.

Picture types and classification scheme

The main portion of pictures digitized so far for COLLATE are (posed) photographs taken by professional photographers during the film production, i.e. stills of film scenes, scenery shots from the same perspective as of the film camera, photos of the working set and person portrays of actors and film directors. Most of these pictures were part of the advertising material distributed or published for the given film, e.g. used for postcards, film posters and advertisements in film programs and journals.

Due to their age many photographs are slightly damaged with some scratches or stains, but generally the depicted objects are still quite well visible and recognizable. Photographs from this time period are mostly black and white, whereas film posters and advertisements available are in color. The latter may include photos but – typically for this time period – very often paintings and drawings and some text portions giving basic information on the film title, cast and sometimes the performance details.

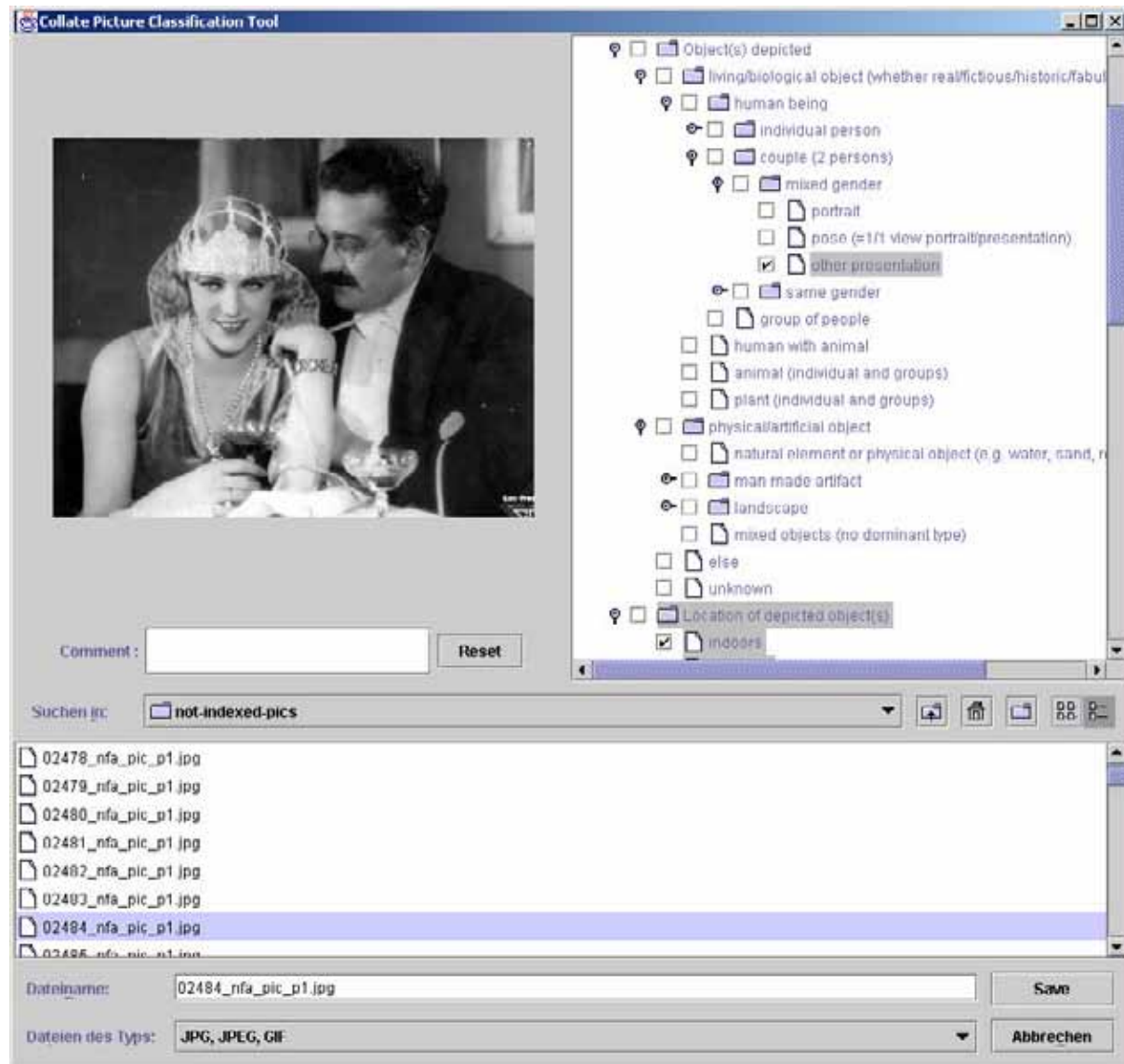


Figure 15 Screenshot of the picture classification tool with topic keywords

The pictures were digitized as high resolution TIFF or JPEG files with at least 300 dpi for the so-called “archiving versions”, and later converted and compressed to low-resolution JPEG files (72/96 dpi), which are to be used for display via the COLLATE Web interface. Details on the scan parameters, date, size and quality of the picture sources are entered in the DIGIPROT databases (see *Deliverable 2.1*, pp. 20 ff). For our feature extraction and comparison algorithms both the high- and low-resolution versions were converted to PPM (portable pixel map), and all experiments were performed using both versions in parallel.

We developed – based on analyses of 500 set of sample photographs – a classification of all available picture document types and a high-level subject matter indexing scheme that is domain-specific but extensible to a larger variety of picture types. The classification scheme was used for the manual indexing of the training set, which was then needed for the automatic image analysis and rule generation. (This should not be confused with the subject matter indexing schemes used by the archive users to manually index the photographs and other pictures in the collection).

Our COLLATE classification scheme (see *Deliverable 6*) was used within the classification and indexing tool we developed for indexing the training set as a structured list of index terms (see right-hand side of

Figure 15). For the training set we selected 500 images from the COLLATE core collection, i.e. digitized photos provided by the three archives.

Results and evaluation

In our verification, we compared the results of our retrieval engine with the outcome of manual indexing. In the first step, a set of 500 images was manually indexed, and feature vectors for these images were computed. Then the set was divided into two disjunct sets with the same number of elements. The first set was used as a training set to calculate the rules, and the second set as a testing set to estimate the quality of the discovered rules.

We used three different indexing methods to index the test set. The first set of rules were calculated using a brute force algorithm (BF), the second one using QCF with all 17 feature values (QCF) and the third one using a brute force approach to select the best set of feature values for the QCF function (BFQCF).

After generating the rules using the same training set for BF, QCF, and BFQCF we use the generated rules to index the images of the test set and store the result in log files. These log files and the manual indexing of the test set were used by the TREC evaluation software to calculate precision values for different recall steps. The resulting tables of our experiments can be found *Deliverable 6*.

The experiments show the best results with the BFQCF indexing, followed by BF which shows better results than the QCF formula. The main advantage of the BFQCF and QCF is that the calculation of the rules is faster than by the brute force algorithm. The experiments also showed that retrieval quality decreases with the higher complexity of the descriptors within Eakins' model. Level II descriptors such as "Location:Indoor" or "Object:Group" of people were retrieved with a better quality than the level III descriptor "Topic:war". These results were – though somewhat finally not as good as envisioned – nevertheless of much more practical value than initially expected. The retrieval features of COLLATE prove – in the last instance – how much of the designed features could be set into practice.

Figure 15 shows a situation where the "Pictorial Search" mode was used and a number of relevant documents were returned. This is not to be confused with the "Advanced Search" mode which also allows searching for pictorial material but is based on manually assigned metadata, such as cataloguing data and keywords assigned by professional human indexers. The pictorial search, though being less precise, is offered to account for cases where no cataloguing and indexing data have so far been entered. Even if part of the result set does not completely match all of the query restrictions (e.g., displaying an individual person or a couple) the user might still find some relevant documents browsing through the list. In any case she has always the option (at least as a expert user) to then assign precise cataloguing data and appropriate keyword terms from the structured subject classification list. Additionally, the picture can be marked up and annotated as any other document, which then can accessed also by the "Fulltext Search" tool.

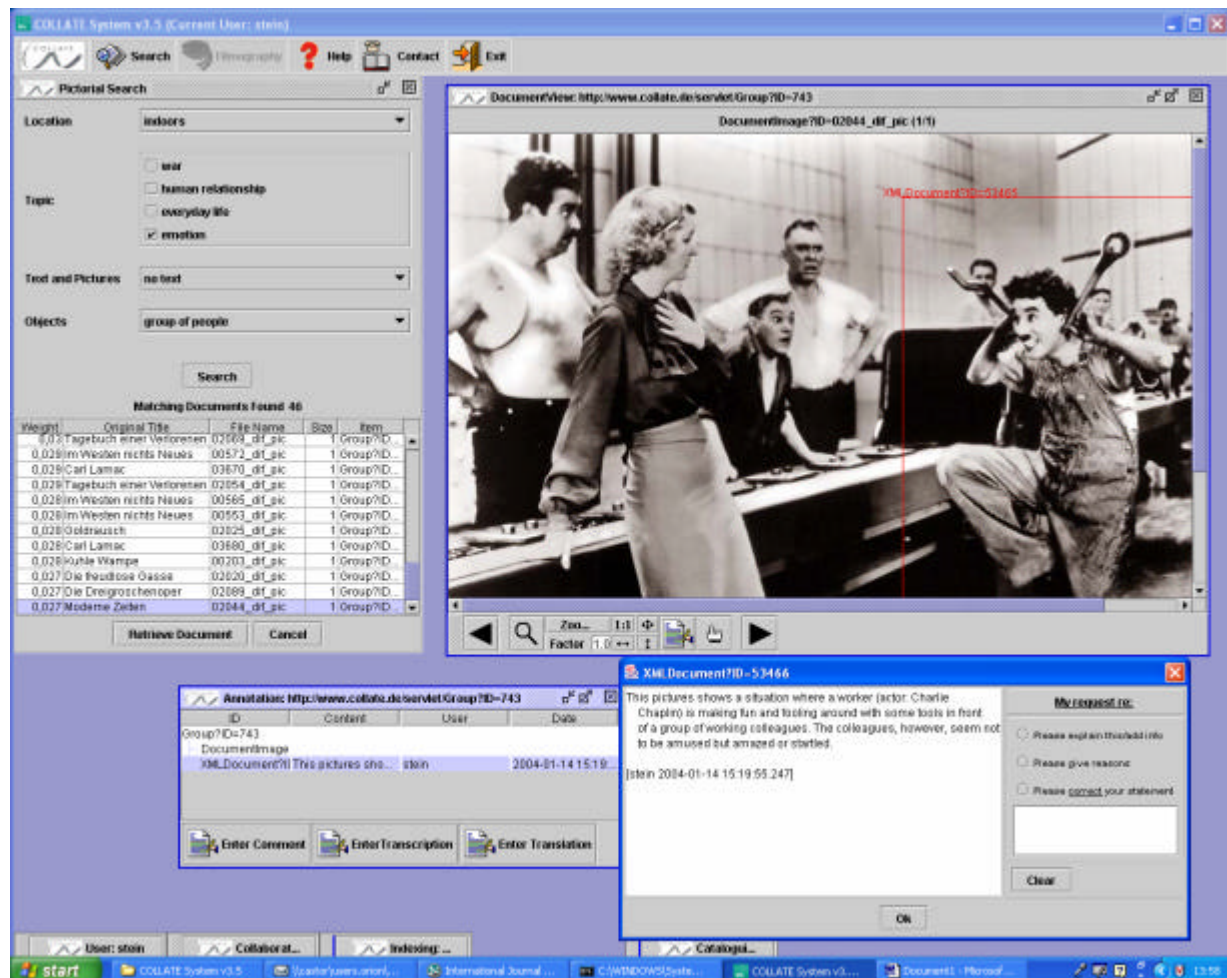


Figure 16 COLLATE screenshot of picture retrieval result (automatically indexed image)

Conclusion

In *Workpackage 6* we showed that in the case of image retrieval, the inadequacy of retrieval methods based solely on physical document content, i.e. pixel data, are particularly evident. Statistical similarity measures between pixel values may very well say little about the relevance of a given image for a given user. In order to achieve the goal of high-precision search a well thought out set of rules is needed. First experiments with rules formulated on the basis of quantile estimates led to recall and precision rates which are, in the best cases, comparable to those of probabilistic text retrieval systems. Further improvements were achieved by taking into account feature vectors. The QCF approach yielded more reliable indexing rules than the quantile analyses.

Although this approach makes it possible – to a certain degree – to index large collections of visual data, provided a relatively small manually indexed test set is available, it reaches its limits when the query complexity reaches the third level in Eakins' model. Hence, we expect to improve the retrieval quality by dividing the visual document into objects and analyzing the objects with our approach.

4.2.3 Digital Watermarking of Digitized Documents

With the technological improvements in digitalization technologies, the digital storage of a large amount of different types of traditional materials is becoming the most efficient way to preserve and protect cultural heritage, but also to improve the accessibility of rare and valuable materials by more people.

Digital preservation and electronic publishing raise exciting challenges in handling multimedia data, common to all types of digital libraries (DLs). Besides the accessibility to a great variety of information, high quality services, such as cataloguing, annotations and content-based search, are required. However, even if the system used to store, access and distribute the documents depends on their nature, this process leads inevitably to more abuses in the usage of the documents. This is often considered as an impediment to the use of electronic document distribution and it is a limit for the commercial possibility of DLs. For this reason a balance of the efforts towards improving available services, distribution and accessibility with the efforts to increase security and intellectual property rights protection is needed.

In this context it is important not only that the users pay more attention to the security problems related to the digital distribution of content, but also their awareness, that the digital content is also subjected to legally defined data protection, has to increase. This is possible only if data protection constitutes an integral part of products and services.

A security technique used to protect digital multimedia data is digital watermarking. It consists in the process of embedding data in an audio, image, video or text files, in such a way that the quality of the original data is not influenced at all by the presence of the watermarks. This technique is currently used in security applications for owner, client and data authentication (copyright, fingerprint and integrity protection).

The embedding of watermarks into the multimedia data of a digital library opens up interesting possibilities. For example, the possibility for a DL agent to locate the watermark at clients sites, check the URL against legal copies, trace the illegal distribution of material originally belonging to the DL and check the malicious manipulations eventually present in the distributed DL data.

In this section, we illustrate first a schematic model of a digital library, describing the different actors and their role, then we present a description of the security framework of the COLLATE project. We analyze the security requirements of the system, starting from the user requirements defined by the archive partners. After a brief technical description of the whole system, the digital watermarking engine for owner and data authentication is presented in detail. It is based on a protection scheme with a private server detector and symmetric copyright and integrity watermarking algorithms. In the second part of this section, the two algorithms are deeply described, and an analysis of the constraints to the algorithms due to their combination is performed.

Interconnections Model

In order to better analyze the security requirements of the COLLATE project, it is useful to describe a schematic model of the interconnections of a generic digital library, in order to illustrate the different actors and their roles (see *Figure 17*). Four main roles can be identified (see also Kohl et al., 1997):

- **End users:** individual clients requesting access to the digital library.
- **Licensing institutions:** information custodians and content owners, such as a university library or an archive. They deliver the digitized information to the end users, defining the security policy and managing the access rights. This functionality can be achieved in collaboration with a certificate authority, which authenticates the end user and the licensing institution. Some institutions provide digital distribution of their documents only through other licensing institutions and do not have directly contacts with the end users (i.e. licensing institution C in *Figure 17*).
- **Publishers:** content authors, copyright holders or agents representing them. For some documents, the licensing institution can be itself the copyright holder. The publishers are the creators of the documents (the author of a text, the photograph or the musician) or their agents;

they hold the copyright also on the digitized information, which are made available to the licensing institutions for digital distribution.

- **Certificate authority:** this can be a department internal to the licensing institution or an external certificate authority.

The illustrated model is of course just a schematic representation and versions slightly different from the described one are possible (i.e. the publisher acts also as licensing institution, or an end user is an entire community having a “site licensing”).

In the COLLATE project, there are three licensing institutions, the three document archives Deutsche Filminstitut DIF - Germany, Filmarchiv Austria, and Národní Filmový Archiv NFA - Czech Republic. Even if the digitized documents are collected in a common database, every archive defines its own security policy, independently from the other archives. For the most part of the documents they are the documents owner, but they provide access through the Web collaboratory also to documents of other archives and owners, which do not participate in the digital distribution. Since the documents are very old, the copyright holders are partially unknown. For these documents, a complicate research would be necessary to find out the real copyright holders.

The community of the end-users in the current version of the COLLATE system is a distributed but restricted network of researchers, which can access the digitized documents. Besides, the researchers can become active, since they have the possibility to make digital annotations and comments to the documents or to annotations of other researchers, already stored in the system.

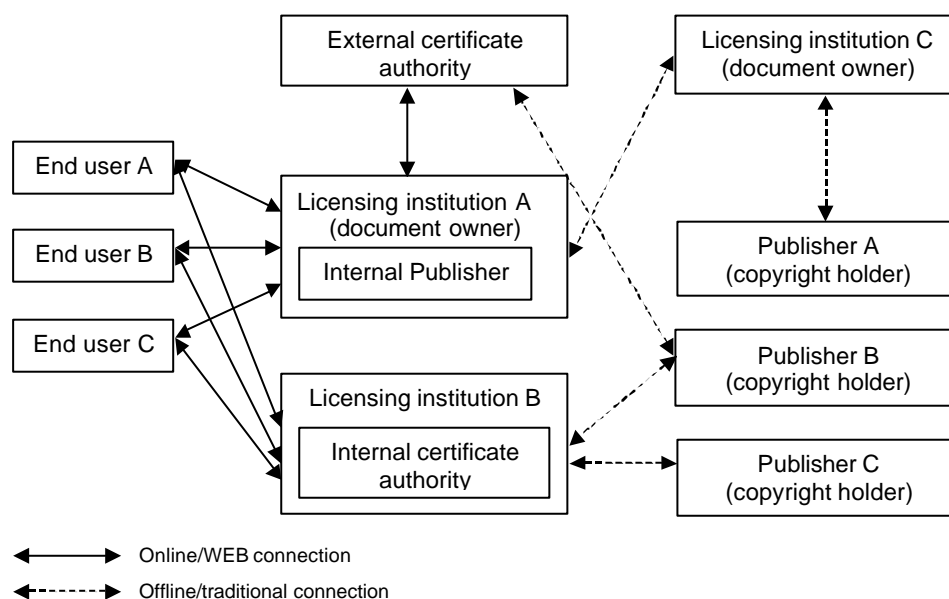


Figure 17 Model of the interconnections of a generic digital library

Security requirements

In this section we briefly analyze the main security requirements for the COLLATE system, starting from the user requirements defined by the archives.

We focus in particular on security requirements concerning digital communication. In this context, the security requirements can be classified in requirements ensuring the communication content and requirements ensuring the communication conditions.

We analyze the security requirements from the different points of view of the licensing institution and the end-user. Table 1 summarizes this classification (Wolf & Pfitzmann, 1999) from the point of view of an end-user of the COLLATE system, differencing between a passive researcher and an active researcher,

who wants to give his contribution in form of an annotation, while Table 2 classifies the security requirements from the point of view of the licensing institution:

Security Objectives	Content	Conditions
Confidentiality	Confidentiality (passive) Origin Authentication	Anonymity (passive) Server Identification and Authentication Privacy (passive)
Integrity	Integrity (received and sent)	Server Accountability
Availability	Availability Copyright (active)	Reachability Legal Enforceability

Table 1: Security requirements for the end-user

Security Objectives	Content	Conditions
Confidentiality	Origin Authentication Access Control	User Identification and Authentication Access Control
Integrity	Integrity (received and sent)	Accountability
Availability	Availability Copyright	Availability Legal Enforceability

Table 2: Security requirements for the licensing institution

Confidentiality. The end-user could require confidentiality about the accessed documents. If the end-user becomes active and makes an annotation, this requirement has of course no sense and the annotation has to be public for the COLLATE community.

Authentication: Establishment of the user and server identify and verification of the claimed identity. Origin authentication means that the origin of the data has been authenticated.

Anonymity. End-users can communicate or use a service without declaring their identity. A researcher making an annotation cannot require anonymity but has to sign the produced document.

Privacy. End-users can communicate or use a service without others can notice this. Again, researchers acting actively cannot require privacy for their actions.

Integrity. The transmitted data have to reach the end-user or the server without modifications (also called "data authentication").

Accountability. Also called "non-repudiation". Communicants cannot successively deny their communication. This implies the recording of actions performed by server and users, to permit the link of the consequences of those actions to a specific user (recording the exercising of security-related rights).

Availability. The services and resources must be available, if the end-user wants to use them.

Reachability. The user decides if a peer entity can contact him.

Legal Enforceability. The user and server have to fulfill their legal responsibilities.

Access Control. The server has to control the access rights to system information or resources.

Copyright: The server wants to protect his owner copyright for a particular copy of a document, but also the publisher copyright. The active user wants to protect his intellectual property rights for the annotations he produced.

The open issue of the protection of copyright and ownership of materials results to be an important constrain for the archive work. In fact, on the one hand, it constrains the accessibility of the archive material, since the archives want to avoid any copyright infringement and, on the other hand, it limits the commercial possibilities of the archives. Thus, controlling the problem of copyright protection could result not only in a widely accessibility to the COLLATE documents, but also provide new economical entries for the archives.

The documents of the COLLATE systems can be grouped into two sets with different resolutions and dimensions:

- **Archiving document set**, containing the documents digitized in their original size with high resolution, which are stored in the database of the digitizing archive. They are not distributed through Internet and thus the copyright protection is the most important security requirements for this set of documents.
- **Web document set**, containing the documents post-processed for the Web collaboratory environment, which are a low-resolution compressed version of the previous ones. Due to their online availability, not only copyright information is important for these documents, but also their integrity has to be protected.

Due to the major impact of the Web environment on the accessibility to the material, not only the documents digitized with very high resolution and visual quality have to be protected, but also the documents digitized with parameters particularly suitable for the Web collaboratory must be protected. The need of data authentication (integrity) is straightforward: Once the archives make available their documents in a Web collaboratory, they need to be sure that the material has not been modified. The integrity requirement is important particularly for the Web documents, since the transactions related to the archive documents can be better controlled.

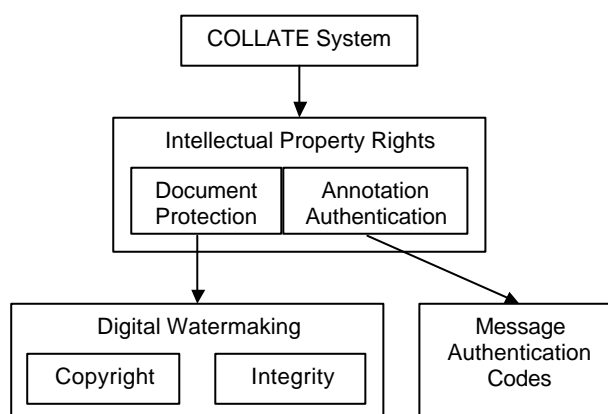


Figure 18 Intellectual property rights graph

Based on these requirements, the security aspects especially interesting for the COLLATE project are related to the content protection, in particular they address the issues of the data protection (integrity and copyright) and annotations authentication, as illustrated in Figure 18. This section deals with the data protection.

System description

From the technical point of view, the digital watermarking engine, together with the intelligent document processing and classification modules and the image and video analysis modules, is one of the main document pre-processing modules. An XML Content Manager is the interface between the watermarking

server and the end-clients, that is the historical researchers (Figure 19). The watermarking engine consists of following parts:

- Watermarking procedures of the watermarking server for the embedding, retrieval and dispute schemes.
- Inquiry request protocol for copyright information, and integrity check, performed by the client in order to check the copyright and integrity states of a specific image to use it in his research work.
- Dispute protocol, which intervenes if an integrity check fails or a discrepancy about the copyright situation of a document has been detected.
- Interfaces between the watermarking server, the XML Content Manager and the client.

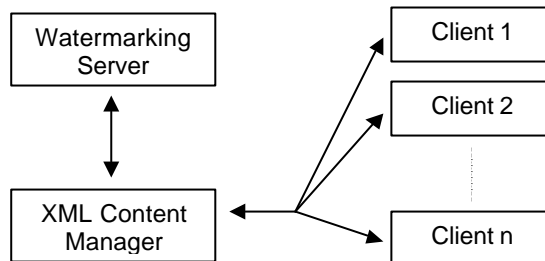


Figure 19 Relationship servers – clients

Watermarking schemes

In the following section, three different schemes providing owner and data authentication (copyright protection and integrity check) are considered with a brief description of their advantages and disadvantages:

Private server detector with symmetric algorithms: only the watermarking server performs the embedding and retrieval processes for both the algorithms and generates the secret key. The client makes a request of examination to the server and uploads the document to be checked to the server. The server retrieves the watermark and confirms/negates the integrity of the document (integrity watermark) and sends the client the retrieved copyright information (if any).

The advantages of this scheme are:

- The detector does not need to be public. This avoids that possible attackers reconstruct the embedding algorithm, analyzing the detecting algorithm.
- The secret keys do not need to be distributed, reducing the risk of threats.
- It is not necessary to authenticate clients requiring the watermarking service.
- It is usable also in an open system.

A history tracking the request of examinations coming from the clients can be compiled. It is possible to verify if an attacker tries to destroy the watermark, modifying always the same document with specific successive alterations or if an attacker tries to extract the watermarking using statistical analysis of documents possibly marked with the same secret key.

The disadvantages and risks are:

- The client cannot check in real-time the copyright or integrity situation of a specific document.
- The watermarking service could be used to provoke a denial of service of the complete COLLATE system.
- Everybody can prove his own attack. For this reason, the examination requests should be tracked and periodically controlled.

Public client detector with symmetrical algorithms: the watermarking server, which also generates the secret keys, performs the embedding process. The detector algorithm is public and the secret key is

distributed to trustworthy COLLATE clients. In this way, the client can check the integrity and copyright watermark of a document directly, without the help of the server.

Advantages of this scheme are:

- The client can check in real-time the copyright situation or integrity of a specific document.
- The disadvantages are:
- The key in the watermark detection process could be disclosed during its distribution.
- Overhead due to management of key exchange protocols and user authentication procedures.
- Lack of security: attackers can try to destroy the watermark by observing the detector for a large number of inputs (Kalker, 1998). The retrieval attempts cannot be tracked.

Public client detector with asymmetric algorithms: the embedding process is performed by the watermarking server using a private key, while a public key, produced with public key signature algorithms, is used to verify the marks' presence. The definition of public-key watermarking schemes providing copyright protection is one of the most challenging problems for the digital watermarking world. Theoretically, it is possible to define a public key watermarking scheme, providing enough information to detect the presence of the watermark but not enough to delete it. In practice, no algorithms have been proposed, that could be implemented and used in a real application.

The advantages of this scheme are:

- It is usable also in an open system.
- The scheme is suitable for Web crawling.
- The client can check in real-time the copyright situation or integrity of a specific document.

The disadvantages are:

- Lack of security: Attackers can try to destroy the watermark by observing the detector for a large number of inputs (Kalker, 1998). The retrieval attempts cannot be tracked.

The first scheme, based on a private server detector with symmetric algorithm, results to be the most suitable for the COLLATE system.

Owner authentication

The goal of the copyright watermarking process introduced in this section is to ensure IPR protection for the historical documents in the COLLATE collaboratory system. The symmetric algorithm we implemented is an extension of the algorithm proposed by Fridrich (Fridrich, 1999). It is performed in the spatial domain and the information is embedded into the luminance component of the original document. The embedding process utilizes key-dependent patterns, whose power is concentrated in low frequencies. The key dependency of the patterns enforces the algorithm security, while their low frequencies aspect enforce the robustness against compression techniques. A visual model based on a smooth-block/edge detection, adjusts the watermarking strength to the characteristics of the various cover blocks, enhancing in this way the transparency of the watermarking process. Additionally, an error detection algorithm improves the positive detection rate. The proposed extension allows the retrieval of the information also without the original cover (Stabenau & Dittmann, 1998).

For the COLLATE application the watermarking parameters particularly important are:

- *Robustness:* Due to the large size of the archive documents, copyright information can be embedded in this set with high redundancy, ensuring in this way a high watermarking robustness. The robustness is related also to the transparency parameter: to a strong watermark corresponds a smaller transparency degree but also a higher robustness degree.
- *Transparency:* It is particularly important for photos and film fragments, since censorship documents allow showing some watermarking artifacts.
- *Capacity:* This parameter depends on the size of the cover document and thus differs for the two sets of COLLATE documents.

It must be point out that the problem of rightful ownerships cannot be satisfactory resolved, using only watermarking algorithms, as a growing number of research works in the area of digital watermarking is showing (Craver; et al., 1998), (Tomsich & Katzenbeisser, 2000). Many watermarking schemes in fact provide intrinsic fragileness against protocol attacks, since they are not used into a complete security framework, providing also additional services, such as a PKI, time stamping and secure delivery of the marked copies.

Data authentication

The goal of the integrity watermarking process introduced in this section is to ensure image content authentication, detecting content-changing manipulations of the COLLATE collaboratory documents

The proposed symmetric approach uses a robust watermarking scheme combined with fragile image features (content-fragile watermarking) (Zielhofer, 2000).

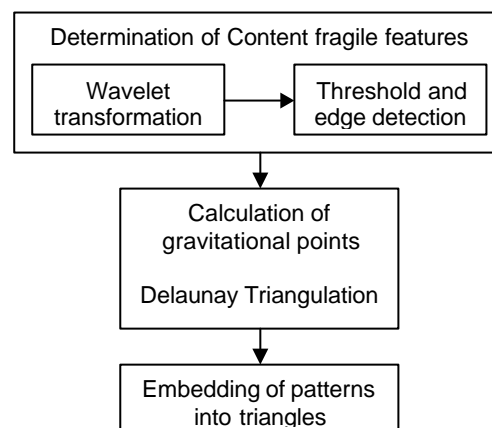


Figure 20 Embedding scheme

As fragile image features, the edges of the cover are extracted using a wavelet transformation and centers of gravity are calculated from the extracted edges with a Delaunay triangulation (O'Rourke, 1994). A pattern, depending from a specified key, is generated and embedded into the triangles, using a robust watermarking scheme. It is composed by many tiles (blocks) and its content is added or subtracted from the luminance of the original document. Figure 20 illustrates schematically the algorithm.

Combination of Copyright and Integrity watermarks

Since the security framework must provide owner and data authentication for the COLLATE collaboratory documents, the copyright and integrity watermarks must be simultaneously present into the documents. This implies that:

- The copyright watermark must be robust against the embedding of integrity watermarks
- (Small) content modifications of a marked picture have to destroy the integrity watermark, as required, but do not have to affect the copyright information.
- The visual quality of the documents could be influenced by artifacts, due to the combination of two watermarks in the same documents.

Since the size of the database of historical documents is very large, an automatic watermarking processing of the documents is demanded, which does not require the manual tuning of the algorithm parameters. But the combination of two different types of watermarks makes the automatic tuning of the parameters quite difficult. Due to the influence of the integrity watermark, it could be, for example, that for particular documents the copyright information has to be embedded a second time in a stronger way, even if the strength previously used was enough to ensure a successful retrieval of the copyright information.

In this context, an important point is the definition of sets of parameter values, which, on one hand, optimize the trade-off relationship robustness/transparency for each COLLATE document and, on the other hand, allow the combination of copyright and integrity watermarking. This definition has to be based on a possible classification of the COLLATE documents by means of visual features with respect to their reactivity to watermarks.

Conclusion

We identified the owner and data authentication of the digitalized documents as one of the challenging security issues for this project. We defined a scheme to embed copyright information into the historical documents of the COLLATE collaboratory system with a robust watermarking algorithm, while a fragile watermarking algorithm was implemented to ensure their integrity. The most important part of our work was to tune both algorithms, in such a way that they could be optimized with respect to the document characteristics and to the combination of the copyright and integrity watermarking.

These specifically developed algorithms can be tuned, in order to meet not only the most important requirements about security but also those about feasibility and complexity, without compromising the flexibility of the work of the researchers. Our future work deals with the improvement of the automatic tuning of the watermarking parameters, based on an analysis of the document visual characteristics. Furthermore, we are working to watermarking schemes, considering also time stamps and other certificating information.

4.3 Integrated COLLATE System

4.3.1 XML Content Management and Retrieval Services

Within the COLLATE System, “XML Content Manager” (XMLCM) is the denomination of the COLLATE system component which deals with a Content Management System which treats XML-based documents. As already known, CMS is the most straightforward of the various labels that attempt to describe a comprehensive solution. A CMS is simply a system to manage content. Content is the asset that needs to be managed. A system is required because several interdependent functions are required to work together within a common framework. The XML term means that the system uses XML as language to represent/manage documents, structured data, metadata and to exchange them over Internet. So, like other CMSs, the XML Content Manager is able to store, retrieve and manage heterogeneous documents and information encoded in XML format. From a COLLATE point of view, XMLCM provides an implementation of a set of functions that can be used in order to create, instantiate and manage metadata associated to multimedia documents in cultural heritage domain.

It is possible to see the XMLCM as a complete Web-enabled system for information exchange and data integration, and a technology that turns enterprise data and information into Internet objects. It provides a new type of open, robust and extensible information server technology and is possible to attach it to the Internet without programming complex server scripting and gateway administration. XML CM's objective is to establish a highly reliable, scalable and integrated open environment, while extending enterprise transaction logic to Internet objects.

Because Internet-based applications deal with complex, heterogeneous and worldwide information, the XML CM will be based on basic open communication standards for information-processing as HTTP, XML and TCP/IP.

XML has been chosen as base technology for the COLLATE XML Content Manager, because this standard promises to become the “Lingua Franca” for the Internet applications. Thus, the XMLCM is able to store XML information in a pure, native format while providing several ways of accessing the data.

By managing data in XML format, the product removes restrictions on data sources, transaction types, deployment and scalability. Users can wrap, link and run Web services, from both legacy and dynamic data sources and deliver the results via a browser, PDA or cell-phone. Besides the traditional local access XMLCM offers Web Services access that reaches scalability and interoperability levels not achievable by other paradigms. Integrated administration interface modules will permit access to external data from heterogeneous data sources, such as file systems (Persistent DOM), relational DBMSs (Oracle, DB2, ODBC) or native XML database, like Tamino.

Besides XML data objects, the XML CM is able to store a wide variety of objects, including HTML, XML or RDF pages for Internet applications and objects such as letters, documents, annotations, references to external BLOBs for other applications. Its content management functionality provides basic transactional semantics for assuring data integrity and it supports standard ACID (Atomicity, Consistency, Isolation, Durability) properties at an object level.

The major components of the XML Content Manager are the data repository, the administrator interface, the query manager, the communication protocol and output utilities. They allow actors (in our case, COLLATE system components) to work in a distributed environment, creating or updating content; to keep track of who's doing what; to produce output in a variety of configurations in a variety of ways (for example, XML documents carried out by particular XML style sheets).

The principal use of XMLCM is in the field of Integration of heterogeneous Persistence Layers (RDBMS, XML-based repositories) and XML Representation of Metadata Models.

Sword ICT is promoting its product based on XML protocol to increase the selling of XMLCM licenses by a joint venture with important software Italian companies.

Moreover, Sword ICT presented the XMLCM to an important Software Italian company to try to develop new applications based on this technology and to organize seminars to promote XMLCM in accordance with the University of Bari.

To make the ingested data managed by XMLCM ready for retrieval, it had to be indexed. For this, a special *indexing Web service* was developed. This indexing service is responsible for the maintenance of the index used for retrieval. Every time new data are ingested to the system, XMLCM stores the data in the repository and contacts the indexing service, which has to update the index accordingly. For annotations, full-text indexing is performed by calculating term weights based on the well-known $tf \times idf$ (term frequency, inverse document frequency) measure. Cataloguing and keyword indexing metadata are stored as attribute-value pairs within the system. Being designed as a SOAP-based Web service, the indexing service offers a standardized API, meaning that an indexing request is sent to the service as an XML document validating against a well-defined XML schema definition. The request is parsed and the data to be indexed is processed and written into the retrieval index, which is a relational database management system.

To allow for complex retrieval functionality, a *retrieval Web service* has been created in COLLATE. Similar to the indexing service, the retrieval service is SOAP-based as well, and queries, serialized in XML, are sent to the service in a standardized manner. The retrieval service supports full-text queries following the vector space model, as well as queries on attribute-value pairs. A retrieval weight is calculated for each document and a ranking according to descending retrieval weights is returned. This ranking is also serialized as an XML document.

Behind the scenes of the retrieval Web service, the HySpirit retrieval platform (HYpermedia System with Probabilistic Inference for the Retrieval of InformaTION, <http://qmir.dcs.qmul.ac.uk/hyspirit.html>) is working. Originally developed at the University of Dortmund. HySprit has been used as a research tool and as a commercial product in selected applications. It provides flexible abstraction layers (relational, logical and object-oriented) for modeling hypermedia and knowledge retrieval and supports retrieval strategies exploiting relationships (spatial, temporal, semantic etc.). Based on probabilistic datalog, HySpirit can access datalog facts stored in relational databases and process them applying according datalog rules. HySpirit was fully integrated into the COLLATE retrieval Web service by providing according wrapper classes to translate XML-based queries into datalog rules.

Due to the Web service-based architecture of the retrieval and indexing service, other core retrieval engines might be integrated into the system by providing according translation classes for XML queries and indexing requests. Furthermore, the Web service approach allows for easy integration into existing systems by being platform independent.

4.3.2 User Interfaces for Indexing, Retrieval and Collaboration Support

Film archives are cultural memory institutions with growing service orientation. More and more, they see their mission not only in preserving cultural heritage but in providing customized information to information consumers with a broad range of interests. Furthermore, film archives unite film scientists, linguists, historians, librarians and archivists as internal actors tracking their own specific interests, roles and tasks. The COLLATE system aimed to support this variety of users, but possibly conflicting needs and demands – for example between the demand for sophisticated content-indexing features for internal users (COLLATE archivists) and the requirement of transparency and intelligibility of the search interface for external end-users – must be pondered with caution and require a lot of adaptive background structures.

Design of the indexing/annotation interface

The COLLATE graphical user interface (GUI) prototype has been developed in very close cooperation between IPSI (the lead/responsible partner) and the internal expert users (from the film archives) – especially the DIF – additionally exploiting results from deliverable *D 9.1: Study of end-users needs and behavior*. The indexing/annotation GUI design of the COLLATE system was intensively discussed and developed in close cooperation of several contributors – mainly from IPSI and DIF – with expertise in various areas: Documentation and Information Science; Knowledge Management; Archive Databases; Film Science; Web Information Design; Intelligent Dialogue Design, Task-based User Interface Generation.

According to results from *Deliverable 9.1: Study of user needs and behavior*, the majority of **internal actors** have an academic background in history, drama, film theory, film history and literature and are especially interested in complex and scientific questions in the film domain. Against this background, they expect the COLLATE system to accelerate and ease simple tasks as known-item retrieval and overviews of available document sources and permit to concentrate on complex tasks as cataloguing, indexing, source analysis and interpretation, object assessment and publication. Since only a minority of film archive staff has a formal education in information, library or archive science, they set high importance to a self-explanatory and instructing interface for metadata input.

The archives' main requirements are according to our analyses

- Easy, efficient and high quality **content access** to all existing information sources in the film archives requiring a meta model or **meta scheme for content indexing** and cross-retrieving of all multimedia collections and **common language** and terminology for content statements.
- Need of **collaboration support** translated in the possibility to identify and consult experts, to cooperate in classifying, cataloguing, indexing and searching and to communicate via forums or email-lists.

Easiness of content indexing and content access correspond to the **usability** of the COLLATE interface. User guidance by instructing and suggestive interface features and help functions and high visual quality of the graphical user interface will be the main ergonomic criteria for the usability tests to be conducted with the first prototype.

Efficient content access depends on the **applicability and usefulness** of the implemented work flows in document indexing and document search and the handy support of single tasks.

When the COLLATE project started, the client system to be developed was expected to provide a comfortable working environment supporting highly structured archival tasks. It turned out, however, that

there were no established workflows for scientific source analysis and interpretation of censorship documents, i.e. we had to go beyond mere storage, identification and access of the material stored in the digital repository. Since working with cultural content is highly interpretative and incremental, COLLATE was designed to assist collaborative knowledge working processes where additional knowledge is based on discussions about the material at hand. In consequence, the initial static task-oriented layer concept had to be augmented by adding support for collaborative, discourse-oriented in-depth content analysis and interpretation.

Given the requirements above, content-based indexing in COLLATE is categorized into three phases, which are modeled as cooperative, not necessarily consecutive indexing subtasks (ranging from formal to content-based, see embedded box in Figure 21): **Cataloguing** corresponds to traditional bibliographic indexing familiar from the library community, taking the idiosyncrasies of the document collection into account. Keyword **indexing** is performed using an integrated, structured terminology comprising various domain-specific vocabularies. The **collaboration** task, using annotations as building blocks, models the interactions required for establishing collaborative discourses. Comments are entered in two ways, either based on explicit requests by other colleague (e.g., can you explain, provide more info, check my comment?) or without external influence (the film expert “just feels like” contributing new insights/interpretations or discuss a certain topic).

Those interactions between remote users are mediated by the COLLATE system, which might also decide to take up a more active role in that it pro-actively suggests potential collaborators (e.g., users working on documents related the same film), informs about potentially relevant documents in the repository (background retrieval activity based on working context), or simply notifies about new additions to a discourse the user registered interest in. Figure 21 illustrates an excerpt of the COLLATE Task Model, which illustrates the annotation workflow.

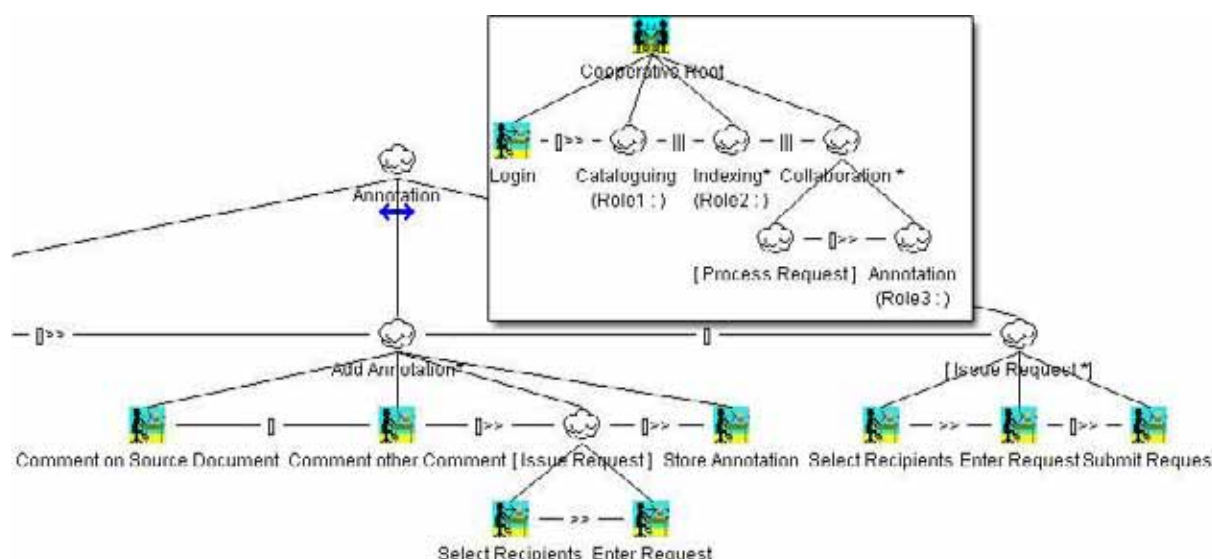


Figure 21 COLLATE task model

Collaboration support

The COLLATE system supports asynchronous collaboration in annotation/indexing for archive users. The scanned documents refer, e.g., to films, plots, or censorship cases. Together with their associated metadata objects, in particular annotations, they represent the main focus of collaborative work, i.e. collaboration is performed through annotating the various types of domain objects.

The semantic and discursive interrelations between the various domain objects can either be unspecified or they can be modeled in a more explicit way by defining specific types of admissible relations. In

addition, certain communicative acts on the meta-level (e.g., requests for clarification) are part of the COLLATE collaboration model.

This kind of system-internal collaboration can, of course, be complemented by external communication mechanisms, e.g., by email or discussion forums.

Discourse Structure Relations (DSR)

It becomes obvious that the users' individual tasks and goals have to be taken into account for modeling a collaborative system. Content-based indexing of a specific document – in this sense – can be considered as a global task, which can be decomposed into partial tasks. In the COLLATE context, the result of these partial tasks, which are to be performed by various users, is value-added information in form of metadata objects associated with the original document. But these partial tasks are only rarely performed in isolation. On the contrary, in most cases a specific annotation is part of a thematic thread, e.g., some newsgroup-like discussion about a certain topic (see Brocks et al. 2002).

To capture the coherence aspects of the discussions, we employ a model of discourse structure relations between a) binary image versions of the original document and annotations and b) discussion threads realized as annotations on annotations (the users' comments on comments).

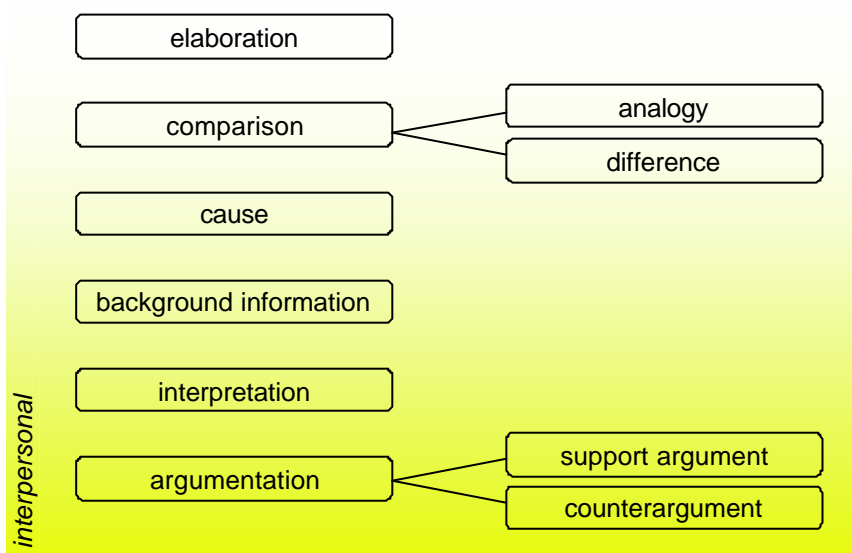


Figure 22 Discourse Structure Relations (DSR)

Our document-centered discourse model is loosely based on concepts derived from discourse theory. In particular we adopted the concept of discourse structure relations (or rhetorical relations as defined by Mann & Thompson 1987). Even though it has originally been developed for monologues in the linguistic context of text cohesion we think that discourse structure relations can be used to describe admissible relations between various data and metadata objects in the COLLATE context, especially between annotations.

Figure 22 shows the specific subset of relevant relations we employed for COLLATE, ranging from factual to more interpersonal levels (i.e. focusing on certain qualities of the participants of the discourse). They were empirically derived from analyses of some existing annotations, but can and will be further tested and verified.

For detailed reviews on other existing approaches and a meta-taxonomy of discourse structure relations see (Maier & Hovy 1993). In the following we just briefly paraphrase the discourse structure relations used in COLLATE:

- **Elaboration** – Providing additional, more detailed information (e.g., “...it's Paris in USA, and not in France...”).
- **Comparison** – Comparative relations can be further sub-structured to emphasize semantic similarities or contrasts between two elements of a discourse.
- **Cause** – To state a specific cause for a certain circumstance.
- **Background information** – Using information about the background of the author of the other annotation (e.g., “... As a lay-person the author does not take psychological aspects into account...”).
- **Interpretation** – (Subjective) interpretation of a statement being referred to (e.g., “...the author actually means...”).
- **Argumentation** – The statement or argument of the other author is either supported, or a counterargument/antithesis is formulated here.

The seamless transition from factual to interpersonal discourse structure relations depicted in *Figure 22* also corresponds to the “illocutionary aspects” of an annotation (see, e.g., Searle 1979), i.e. the specific communicative intention its author had in mind at the time of creation (e.g., from stating factual information towards active participation in a discussion thread).

Even though discourse structure relations proved adequate for modeling the interrelations between annotations it turned out, however, that there are some relevant pragmatic aspects of collaborative indexing work, which are not yet covered. In the next section we describe how discourse structure relations can be complemented by communicative acts to introduce meta-communication, i.e. explicit communication about domain objects, in a seamless way.

Communicative Acts (CA)

Scientific source analysis and interpretation can be regarded as a highly incremental process, which requires extensive domain-specific knowledge. Within a collaborative environment – as in COLLATE – a user can explicitly submit/enter “communicative acts” in order to request assistance from other users of the virtual collaboration team in order to accomplish a particular task.

Our communication model is based on a sophisticated, well-known speech-act oriented dialogue model COR (*C*onversational *R*oles). It was originally developed for modeling cooperative information-seeking dialogues between users and an information system in a domain-independent way (cf. Sitter & Stein 1992/96, Stein et al. 1999). Focusing on the pragmatics of a two-party conversation or multimodal dialogue, COR covers the basic “illocutionary” aspects and role expectations of specific user behavior associated to the user's contributions to a dialogue. In essence, it models a set of basic communicative acts that express expectations, commitments, retractions, clarifications, etc. The combination of semantic and rhetorical information – as requested in COLLATE – was examined in detail during the COLLATE project.

For the COLLATE context only a subset of COR primitives had to be adapted in order to cover the pragmatic aspects of collaborative indexing work. From the five global categories postulated by Searle (Searle 1979) and used in COR we identified only three classes of communicative acts to be of most relevance for our purposes:

- **Assertives** – Assertions represent a certain class of communicative acts that can be characterized as being either true or false. Examples are comments or annotations to a document, users of which are knowledgeable domain experts. Annotations and comments represent an important class of qualified assertives. These annotations/comments can either be requested (by other users) or they are made on their own.
- **Directives** – Attempts to get some other person to do something are classified as directives. Requests, e.g., a search for documents or specific questions (“Could you please explain...”) directed to other users represent the most significant examples of this class of communicative acts.

- **Commissives** – The illocutionary point of commissives is to commit oneself to some future course of action. A promise to provide additional information with respect to a certain object (e.g., a film) would be a good example from the COLLATE domain

The main purpose of these three classes of communicative acts is to allow for explicit communication on the meta-level in the COLLATE system, i.e. meta-communication about the various indexing tasks supported by the system.

Interrelation between Discourse Structure Relations (DSR) and Communicative Acts (CA)

On a conceptual level, the approach to combining discourse structure relations and communicative acts (the COR model) was quite new in our current domain in COLLATE. Communicative acts focus on illocutionary aspects of a specific dialogue situation, whereas discourse structure relations describe characteristic relationships between assertive acts or statements (cf. Stein & Maier 1995), e.g., annotations or comments.

At a closer inspection it becomes evident that some communicative acts might invoke certain types of discourse structure relations between the corresponding annotations. In our view, the set of discourse structure relations adopted for COLLATE are treated as relations between assertive communicative acts, i.e. annotations and comments in COLLATE (cf. Keiper et al. 2003).

From this perspective, we can regard explicit collaboration in the context of the COLLATE project as the combination of specified relation types between annotations, i.e. discourse structure relations, which are complemented by a certain set of admissible COR acts for meta-communication (on the dialogue level) referring to the various types of COLLATE domain objects (e.g., annotations, cataloguing information).

In essence, we regard discourses as results of “cooperative negotiations” where the discourse participants have certain context-related obligations as well as specific expectations.

The COLLATE system as a multi-agent environment

Intelligent information systems are characterized as interactive systems which support various tasks using knowledge about the users and their task-related actions. To support collaboration between users imposes additional requirements on the underlying technical infrastructure. The MACIS Framework (Multiple Agents for Collaborative Information Systems) is designed to support the development of complex, collaborative information systems exhibiting intelligent, pro-active behavior (see Brocks et al. 2003). The system, including its user interface, is described in terms of various classes of cooperating agents. Since users are also modeled as part of the multi-agent society, collaboration between users is seamlessly incorporated as a special form of inter-agent communication. The implementation of the COLLATE system is based on the MACIS Framework. For each task, corresponding user interface agents are provided, which allow those tasks to be performed in parallel or in a sequential order.

Figure 23 shows the annotation interface with the digitized source document to the left, the associated discussion threads in the upper right and the “Enter Comment” dialogue box in front. The dialogue box displays a comment/annotation retrieved from the system (upper left area) and below a text box for entering a new comment referring to the existing one. It also provides the sets of available discourse structure relations (pull-down menu in the “My comment” area) and explicit communication functions (the “My request ...” checkboxes on the right-hand side). Before saving a newly entered comment the user can categorize it by assigning a relation selected from the pull-down menu, e.g. “This fact/statement is different from ...” (comparison/difference) or “I support this ...” (support argument).

At the right-hand side of the “Enter Comment” dialogue box there are two areas for entering explicit communicative acts addressing the other users of the virtual team. These are mainly requests, either referring to the existing comment of the respective author in the upper part (“Please explain this ...”) or referring to the own – just entered – comment (e.g., “Please verify or correct my comment”). The latter accounts mainly for situations in which the current user is somewhat unsure whether her comment or interpretation holds true or suffices to explain the current point made, hence she requests assistance by

her expert colleagues. The system offers these communication options at any time, stores the requests and notifies the addressed user group about new and pending requests. Note, however, that the collaborating user team can (and probably always will) develop soon own strategies of how and in which situations these communication functions are to be used feasibly – which certainly depends on many factors, e.g., the specific contents and pragmatic purpose in such situations. We have some evidence that in the domain of scholarly film censorship and history research the current user team does not view these collaboration functions as some kind of conversational chat forum, but mostly use them for communication about important analysis matters that affect the whole interpretation of a given censorship case or a highly important document in question.

At the bottom of the screenshot, the current discourse situation is visualized by showing the status of incoming/outgoing requests. If the user selects a pending request, a context menu provides a set of options allowing her to directly access the document being referred to, comply with a request (e.g., providing additional information by entering another annotation), reject or postpone its processing. After login, the user is presented the most recent incoming requests as well as the document history, allowing her to directly jump to those documents being referred to in the requests or the ones she was working on during her last session respectively.

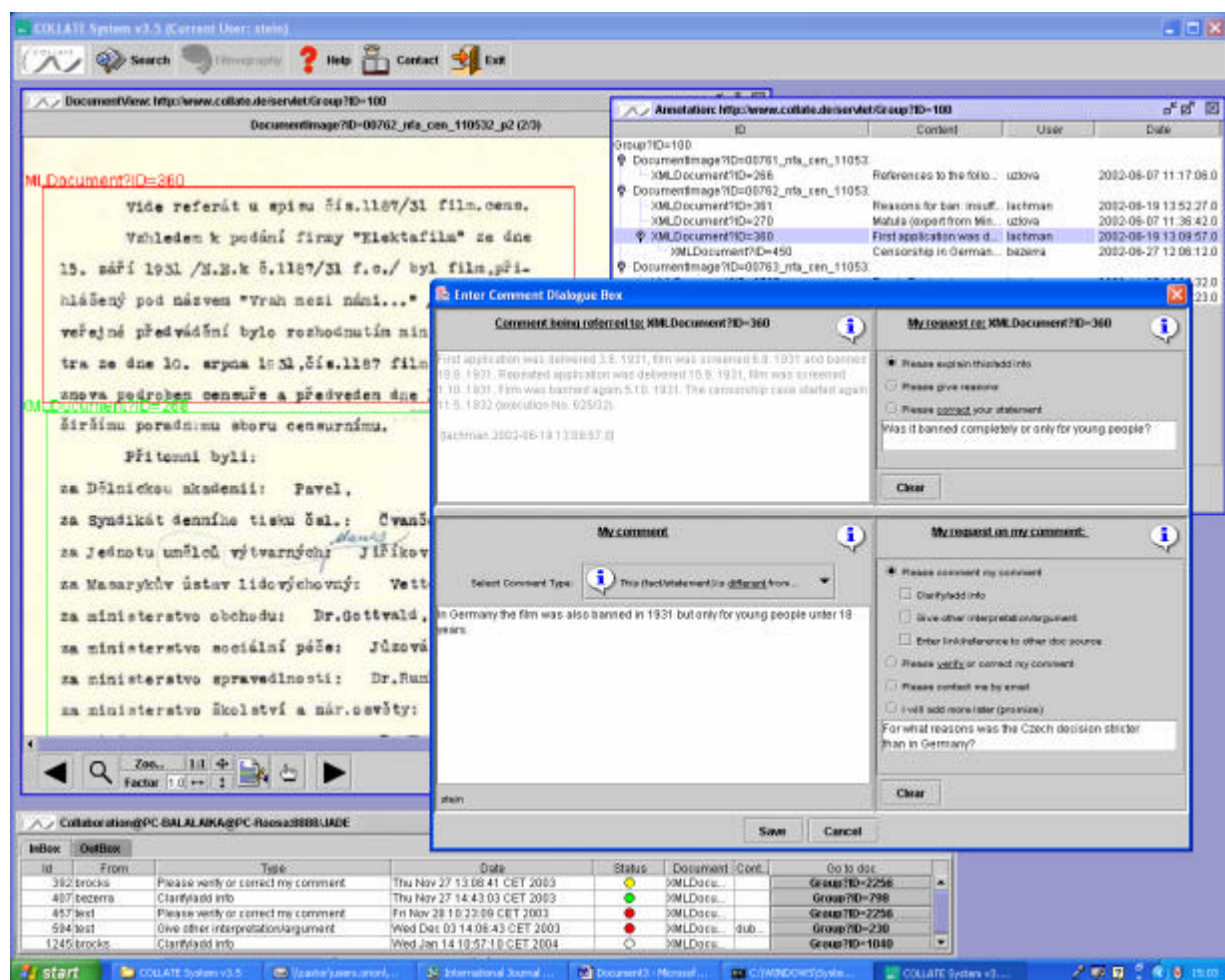


Figure 23 The COLLATE annotation interfaces

The user interface comprises dedicated user interface agents (UI agents), which serve as intelligent mediators between the users and the system. Hence, user input is no longer regarded as some method invocation, but as a communicative act expressing the current discourse goal of the user. Since users

and agents are situated within a shared context, their interdependencies and interrelations are organized according to their roles within the information system.

Hence, we have implemented the COLLATE system as a multi-agent environment, where various types of agents, i.e. users, service agents, and user interface agents, cooperate to achieve a common goal. Given some formal representation of the tasks to be performed, the overall complexity of the information system can be decomposed in terms of a) the corresponding sub-tasks, and b) the agents required supporting them.

Implementation

The MACIS Framework is based on JADE (Java Agent Development Framework, JADE, <http://sharon.cselt.it/projects/jade/>) as an underlying FIPA-compliant multi-agent platform, which provides the messaging infrastructure required for inter-agent communication. Implemented in Java, it supports the development of platform-independent multi-agent systems, which can be distributed over various machines.

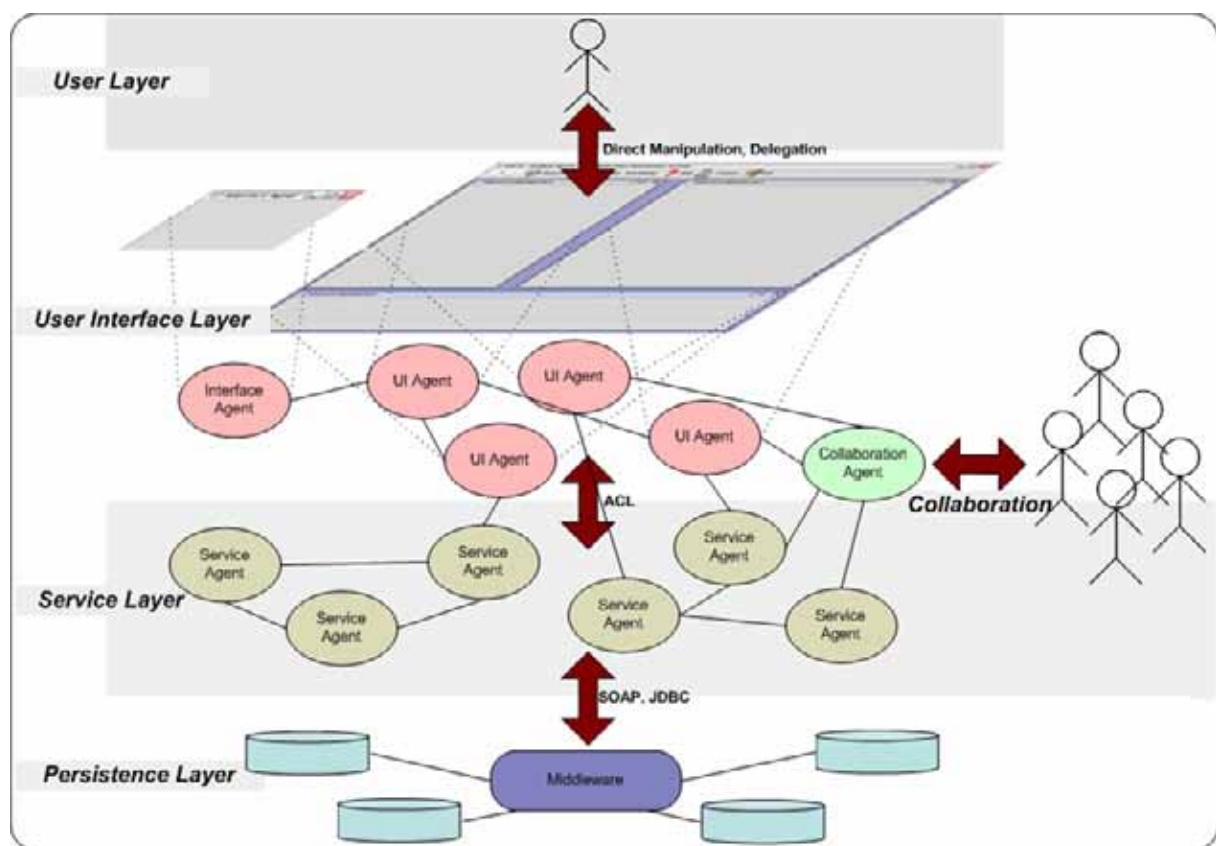


Figure 24 The MACIS framework

User Layer - The user layer provides a representation of the current user in terms of the tasks to be performed, her role in those tasks, and the communication acts required to model her interactions with the system.

User Interface Layer - The actual user interface is composed of user interface agents, which may form organizational relationships to support more complex tasks. The user interacts with user interface agents in a standard way, e.g., using direct manipulation techniques. Interface agents (as in (Maes 1994)) can be incorporated to provide appropriate task-level support (e.g., propose missing parameters) or represent personalized assistants which perform tasks on the users' behalf (delegation).

Service Layer - The service layer is comprised of those agents, which do not provide a user interface. Those service agents represent the functional core (see Buschmann et al. 1996) of the information system.

Persistence Layer - The persistence layer is responsible for the storage and retrieval of the information objects constituting the application domain.

Collaboration - The MACIS Framework incorporates asynchronous, system-mediated collaboration between users working in distributed peers by forming dynamic organizations of users, user interface agents, and service agents.

In its current version, MACIS integrates the underlying multi-agent platform with advanced support for applications following the MDI paradigm (Multiple Document Interface). Further information about the implementation of the client system can be found in *D7.1 "Design and Implementation of the indexing/annotation interfaces"*, *D7.1 Addendum "Collaboration Support in Prototype 2"* and *D7.2 "System Integration"* respectively.

4.4 Working with the COLLATE System

4.4.1 Results from the Users' Work with COLLATE

The results of the archivists work are distributed between the three tasks:

- Task 8.1: Integration of sources, indexing and annotation
- Task 8.2: Source edition
- Task 8.3: Production of surrogates for lost films.

After digitization the film experts had to catalogue the documents (see right-hand side of *Figure 25*). Here, film-specific problems come into play as we have different versions of films with the same film title, and distinguished from one another only by cuts or changed inter-titles. In addition, the film released in different countries had been labeled with various titles. The same happened with names of persons and institutions. Thus one of the first tasks is to come to an agreement which name or title will be used for reference (in the cataloguing and other metadata). Although rather trivial, this example reveals the cultural traditions because the spelling of a person's name or a title has been introduced and used since a long time. Accepting the role of these cultural traditions is an important precondition for successful work with the collaboratory.

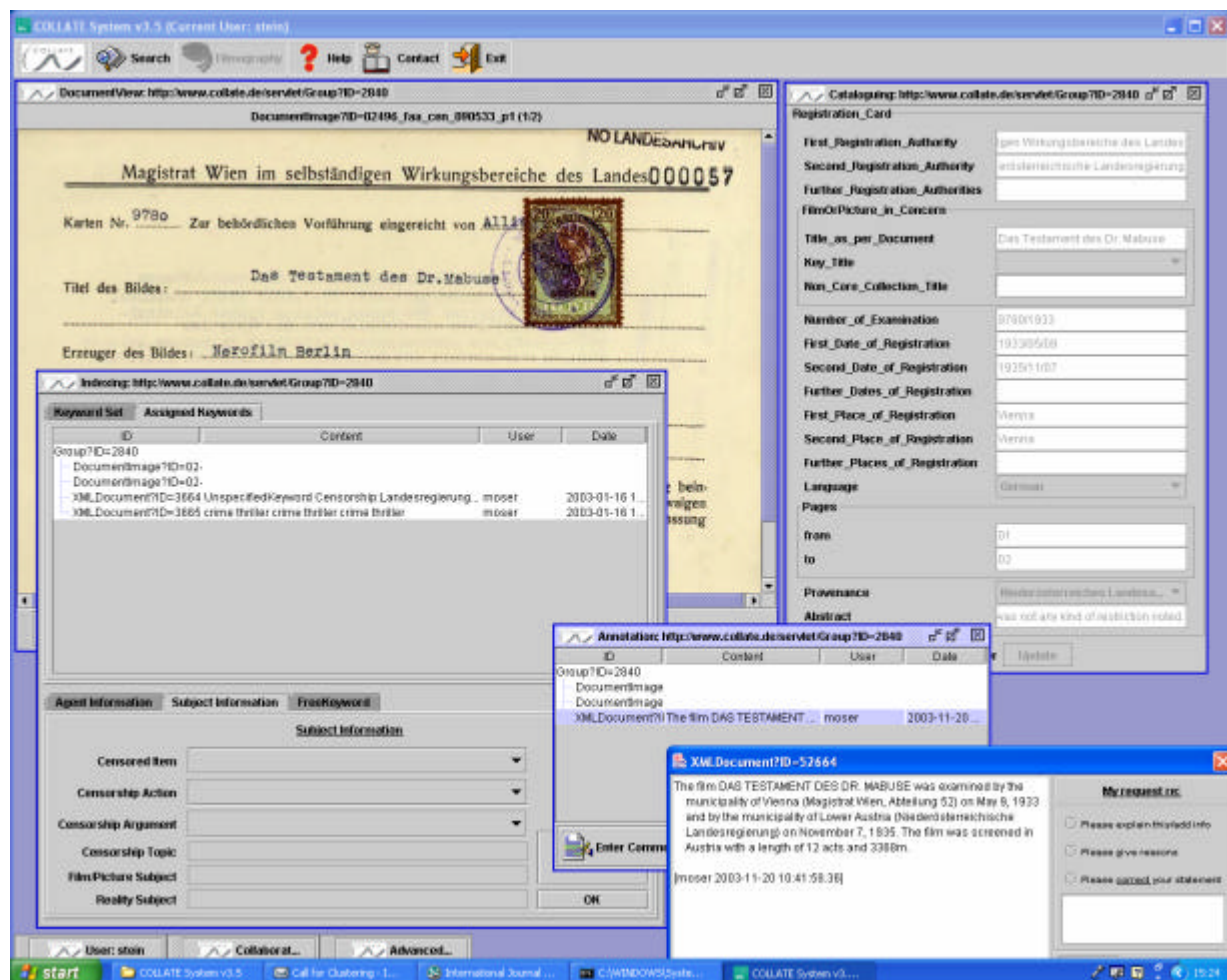


Figure 25 Cataloguing, indexing and annotation interfaces in COLLATE

Integration of sources, indexing and annotation

The main task of the COLLATE users was the cataloguing, content indexing and annotation of document sources.

We distinguish between the *overall collection* and the *core collection*. To date, members from the three COLLATE archives have digitized and catalogued all of the available censorship documents, which build the major part of the overall collection. This means more than 7000 documents with about 18000 catalogued pages (including photos). Out of this overall collection we selected 100 important films, which we were to be examined more closely. For this core collection not only censorship documents but also the corresponding press clips, photos, posters and video fragments have been digitized. Whereas documents in the overall collection are only catalogued, core collection documents are also indexed and annotated in detail. This concerns about 1600 documents with one or more keywords and annotations.

Additionally, the watermarking was tested by the archives and integrated by DIF into the documents. Also the document structure recognition with WISDOM ++ was tested with a subset of documents.

The COLLATE system allows users to enter keywords and annotations describing the whole document or to mark parts of a page and to assign an annotation only for this part (see Figure 26).

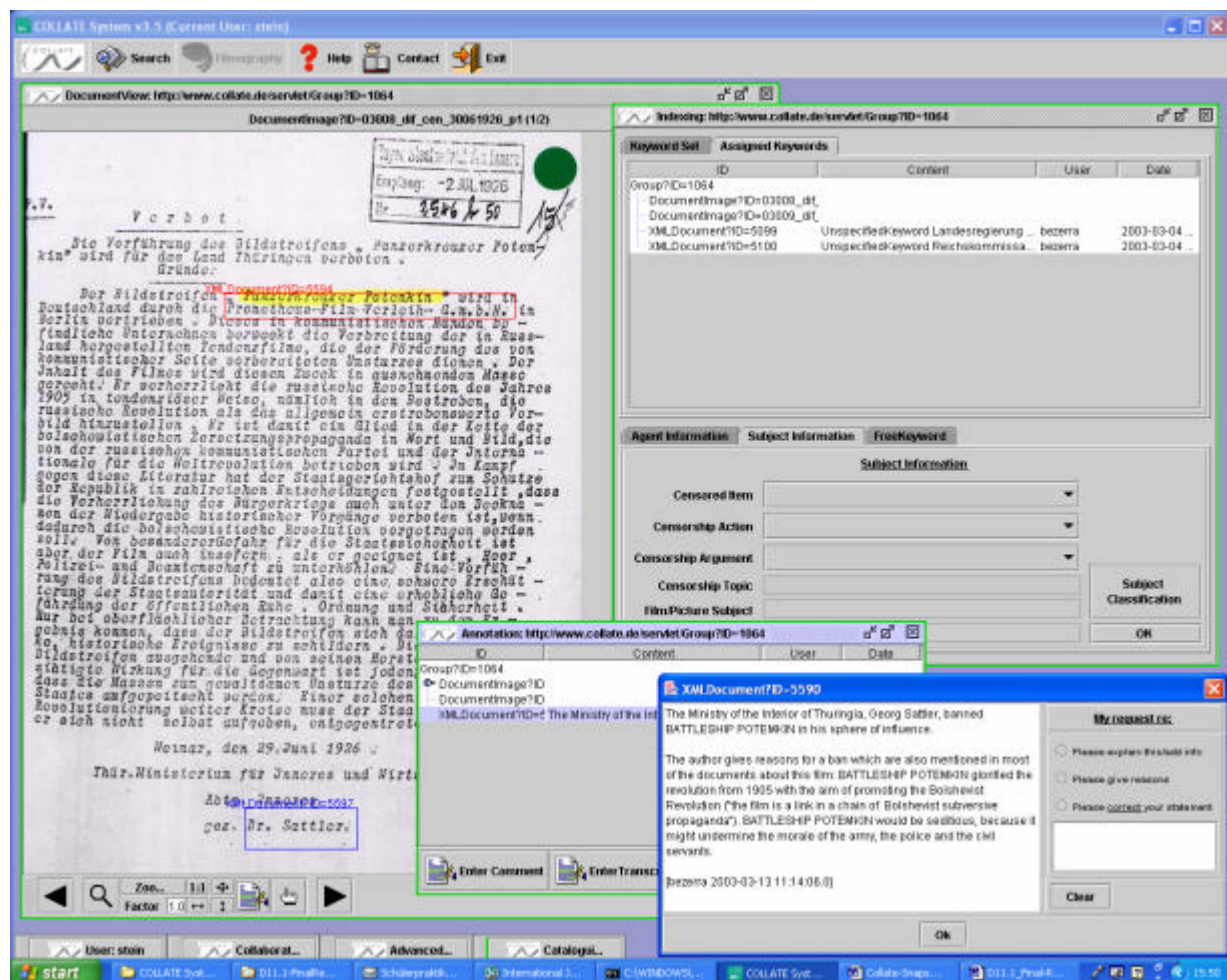


Figure 26 Annotation of a German letter: Local ban of "Battleship Potemkin"

The cataloguing of the documents includes short content descriptions in English. This enables users of all countries (Germany, Austria and the Czechia) to understand the documents or ask for further explanations if necessary. Writing annotations is not restricted to the document owner/provider or the related archives' experts. Each archive user is asked to add information to all document sources in the core collection. We decided to divide the indexing and annotation workflow into distinct work packages on so-called subsets of ten selected films from the core collection. Each set was analyzed first by members of the providing institution, and then passed to the other user groups for further analyses.

Looking back at the users' work during the project's lifetime, we can distinguish three phases (see also Figure 27), which are also characterized by which kinds of tools were used to support their collaborative work (see Keiper et al. 2003):

- Phase 1: Concept and preparation
- Phase 2: Cataloguing of all available documents
- Phase 3: Indexing, analysis and interpretation / Source edition / Production of surrogates

In the beginning (before actual implementation and usage of the COLLATE collaboratory technology) most of the work by the users was done within meetings, by email and phone. There arose an understanding about common project goals and shared interests, and social relationships started.

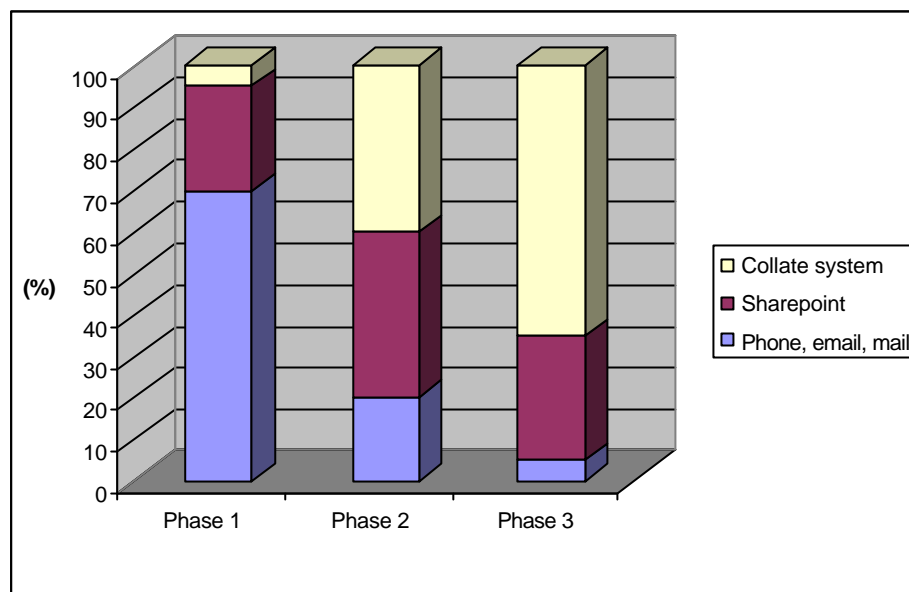


Figure 27 Work phases in the COLLATE project

For the second phase, which is characterized by discussing more precisely the goals and workflows, we looked for a tool supporting this kind of cooperation and the first work with the prototype. As the COLLATE system was just in development we decided to use a commercial product with high reliability. Sharepoint from Microsoft enables the user to communicate about a Bulletin Board with displayed threads and additionally allows simple postings, to store documents and provides access by a controlled user management. During the second phase the Sharepoint system was the most effective and acknowledged tool supporting collaborative acting.

Despite first doubts concerning the acceptance of a further software tool Sharepoint quickly was fully integrated into the workflows of the archives/COLLATE users. Due to the very positive feedback we decided to shift a great part of collaborative work from *Phase 1* (which is done by mail, phone, and email) to *Phase 2*. The success in practice is the result of a presumption discussed at the beginning. As Kouzes et al. mentioned, "the mechanisms of ritual, which moderate our interpersonal interactions, must find a place in the synthetic surroundings of a collaboratory." Therefore we started to define together a "social corset" which soon was accepted by the users, e.g. introducing:

- the division of work packages into smaller tasks (e.g. working on subsets)
- responsibilities and leadership (assigned to these tasks)
- continuous changes of these responsibilities between the participants
- rituals (such as opening of a task or declaration of its end; informing about the state of things)

For example, at the start of each subset ten films and a new discussion's moderator were chosen. Each archive then provided a list of the available documents, get a general idea about the censorship history of these films in each country and about the films' reception. This guaranteed a common basis for collaboration: Shared information before working in the COLLATE system.

The case studies: Source Edition and Surrogate Production

The acquired knowledge and experience leads to specific subtasks which present and represent this knowledge. The **Source edition** and the **Production of film surrogates** are exploiting and deepening the results done during the cataloguing, indexing and annotation work. Both case studies are supposed to check and to show the opportunities which develop when firstly a comprehensive collection is made available to the public and secondly an international group of researchers uses collaborative tools to analyze it. Corresponding to this aim of „demonstrating possibilities“ the COLLATE archives decided in favor of a wide range of tasks.

First of all we differentiate between two different analysis paths:

- the „**source edition**“ (joint editorial work on selected topics, serving as the background for an in-depth comparative analysis of the censorship practice in Austria, Czechoslovakia and Germany) and
- the „**surrogate production**“, that aims to demonstrate how censorship documents (and the analysis of the censorship practice) can be used to support film restoration.

In addition at a second level - within each task - we decided once more to adopt different approaches.

In the **surrogate production** two very different “film surrogates” (something like a “pre”-film-reconstruction) have been generated:

- the **last act** of the Austrian **sound film** CAFÉ ELECTRIK.
- the **whole Czech silent movie** ZPEV ZLATA (VETRELEC).

In the **source edition** we decided to adopt two different approaches as well: The editorial-analytical work in COLLATE concentrated

- on the one hand with a “**classical**”, **film-centered** comparative case study (BRONENOSEZ POTEMKIN)
- on the other hand with the **topic “censorship and genre”** – based on six selected horror films.

Thinking again at the aim of “demonstrating possibilities” the archives decided in favor of a comprehensive presentation on the Web. On the site <http://www.deutsches-filminstitut.de/collate/index.html> the archives DIF, FAA and NFA present the results of their wide experience of film studies using COLLATE. The efforts of the case studies - actually: four examples of the work in COLLATE have thus been made available to public.

Results of the Source Edition: **BATTLESHIP POTEMKIN**

Compared to the situation in Germany and in Czechoslovakia, **BATTLESHIP POTEMKIN** caused less sensation in Austria. Although in Vienna the film was banned for young people, it was otherwise permitted without any further restrictions. Similarities also differences in how **BATTLESHIP POTEMKIN** was dealt with in Austria, Czechoslovakia and Germany can be noticed.

The original length of **BATTLESHIP POTEMKIN** had been 1740 m. In Austria no information is left about a shortage of the film. Compared to Germany and Czechoslovakia, in Austria the longest version of the film (1600 m) was screened, while in the other two countries the screened version was between 1200 and 1480 m long. In Germany and Czechoslovakia the authorities objected to the same scenes: shots on the stairs of Odessa and the mutiny of the sailors.

The argumentation for censoring and prohibiting the film were quite similar. In all three countries the censors referred to „subversive tendencies“ of **BATTLESHIP POTEMKIN**. In Germany and Czechoslovakia the authorities were also concerned about the possibility that a public screening of the film could „endanger public order and security“. This kind of argumentation was supported by disturbances taking place during the screening in the three countries. But in Austria and Germany these disturbances were not evidence of civil commotion. In fact, they were organized by national socialists in order to offer a pretext for a ban.

Furthermore in all three countries there was a difference of censorship practice in the central district and the local regions. In the centre the film was permitted for screening (in Germany and Czechoslovakia with cuts), but rated for adults, whereas the local authorities strived for banning the film. In the case of Germany and Austria, the local governments had to find ways around the legal regulations – which made film censorship a competence of the central government – in order to achieve their goal.

At last, after the takeover of the national socialists in Germany (1933), Austria (1938) and Czechoslovakia (1939), the screening of **BATTLESHIP POTEMKIN** was completely interdicted.

Results of the Source Edition: Horror Films

The films chosen for the analysis of the topic “censorship and genre” are DRACULA (1931), FRANKENSTEIN (1931), VAMPYR (1932), FREAKS (1932), KING KONG (1933) and THE INVISIBLE MAN (1933).

All these movies can be denoted as prototypical horror films. They are characteristic productions, which shaped the conditions and meaning of the genre itself. Besides their prototypical aspects, like dialogue-parts, narrative and filmic structures or the valuing of gender, all these movies have been produced within a few years and reflect the fears and problems of two troubled decades. A new kind of verbalizing that universal fear had to be found in these films – not only for artistic aspects, but also to bypass the censorship authorities. A useful and impressive way to shift the shock and the terror into the viewer's mind and imagination was found by (re)activating well-established theatrical concepts. From this point of view an in-depth research of the horror genre and censorship during the 30ies should lead to a better understanding of the genre - and perhaps of the idea of censorship itself.

In Austria all the selected horror films granted permission by the censors and were shown to public in the 1930es. The lack of cuts or restrictions by the censorship of horror films in this country is remarkable. Unfortunately, the available registration cards did not mention any reasons for the decisions made by the censorship authorities.


Only four of the chosen films were examined (and screened) in Czechoslovakia until 1939: DRACULA, KING KONG, THE INVISIBLE MAN and FRANKENSTEIN. Relevant material illustrating details of the censorship procedure is rare. From the sporadic available material only one “regularity” of handling the horror genre can be deduced – all films were banned for young people.

The German censorship was much stricter than that of the neighboring country: only FRANKENSTEIN, KING KONG and VAMPYR were approved for public screening. FRANKENSTEIN and KING KONG, however, had been prohibited at first and were initially only permitted in appeal hearings. Even in such cases of permission the films were permitted only for adults 18 years of age or older and cuts were imposed. The verdicts of the German censorship authorities – the “censorship decisions” – were recorded in written form and explained the reasons for permission or ban, for cuts or further restrictions.

Results of the Film Surrogate Production


The reconstruction, also under good conditions like during the work on the Surrogate Production of CAFÉ ELEKTRIC, of a lost film, or parts of it, is a very complicated work. A potential success of the Surrogate Production depends on the amount of given sources. The actual state of the sources for the example ZPEV ZLATA (VETRELEC) can offer only particular and very hypothetical reconstruction of the film. Nevertheless this result could facilitate the identification of other scattered material and could help to build up a more complex image about the lost film.

The sub-project of Surrogate Production profits from the synergies of the COLLATE-project and allows to work on this task with the sources of three international archives and it triples the possibilities to succeed with the advised targets. On one hand the Surrogate Production is also a good possibility to promote the COLLATE main-idea, i.e. the necessity of a cross-border collaboration of archives to promote a better, and even more effective, analysis of given sources and information. On the other hand, there's also an economical aspect: The archives are not only keeper of cultural heritage; based on such scientific results, it can also be utilized in economical way like different reconstructions of missing film-parts or publications.



Source Edition
 Bronenosez Potemkin Horror Film

Surrogate Production
 Cafe Elektrik Zpěv zlata



Horror Film: Introduction | **Case Studies** | Freaks Worldwide | Genre and Censorship

Dracula - Frankenstein - Freaks - **King Kong** - The Invisible Man - Vampyr

The Case studies

KING KONG

USA 1933
 Director: Merian C. Cooper, Ernest B. Schoedsack

Produced by RKO Radio Pictures
 Austrian Distributor: Mondial
 Czech Distributors: PDC, Praha; Elektafilm, Praha
 German Distributor: Europa-Filmverleih AG, Berlin

Overview of the censorship case KING KONG

Austria	Czechoslovakia	Germany
24.3.1933: 9645 permitted? banned for young people? 1800 m, 4 reels	26.9.1933: 1106/33 Fc banned for young people 2750 m; 11 reels	26.07.1933: B.34168? banned
	4.12.1938: 1406/38 Fc prolongation of permission 2750 m; 11 reels	05.10.1933, O.6910 banned for young people 2266m (after cuts: 2217m)

In **Austria** KING KONG was censored in September 1933 and in October 1937. An interesting point in this case is the different length of the two film versions.

In 1933 the film was hand in for censorship by the distributor Mondial for the producer RKO at the Magistrat Wien im selbständigen Wirkungsbereiche des Landes (Magistrate Vienna). According to the [registration card](#) (no. 1016, 1.9.1933) its length was 2600 m. There was no kind of restriction noted, but according to [PFL from the 8th September 1933](#) the film was banned for young people.

In 1937 the producer/distributor RKO itself applied for censorship again now at the Besonderes Stadtamt II/3 im selbständigen Wirkungsbereiche des Landes (Special City Department of Vienna II/3). Interesting is that this re-release of the film was not mentioned in PFL and the length of the movie is different. According to the second [registration card](#) (no 1331/37, 22.10.1937) the length was only 2130 m. Again no kind of restriction was noted.

KING KONG was censored twice in **Germany**. In 26th July 1933 the film was banned outright by the Censorship Office in Berlin.

Source Edition

Surrogate Production

Introduction

Censorship Regulations

Battleship Potemkin

Horror Films

Conclusion

Bibliography

Home

Collate: Übersicht

www.collate.de

Figure 28 Source Edition on the Web: censorship and genre (case study KING KONG)

Conclusion

Although the results of the two case studies Source Edition and Surrogate Production are more visible the work with COLLATE system itself gave valuable insights.

For example NFA research provided an exemplary illustration of the myriad of censorship links between Germany, Czechoslovakia and Austria, thus demonstrating the concrete possibilities of collaboration through networking.

Sergey Eisenstein's *BATTLESHIP POTEMKIN* is considered to be one of the most important films in the past century. This silent film, which caused a massive international sensation, still reaps enthusiasm from audiences for its modern and clear style. What is less known, however, is the fact that a sound version made in Germany already existed in 1930. This version is unfortunately no longer available.

A German permit card, long considered to have been lost, was found by the NFA-Team in the archive collection "Cenzurní sbor kinematografický při ministerstvu vnitra 1919-1939 (1940)" of the State Central Archive in Prague (Státní ústřední archiv v Praze). The discovery of this permit is marked with exceptional significance in film history because it is one of the few existing documents that provides information about the dubbed version of the film. It can **support reconstruction work** on the sound version of *BATTLESHIP POTEMKIN*.

Moreover, the DIF analysis showed that it is an unusually concrete example of the extent to which the **results of censorship and self-censorship** could change fundamental aspects of a film: At the beginning of the German sound version of *BATTLESHIP POTEMKIN* from 1930 the audiences heard an explanation of the reasons for the mutiny on the "Prince Potemkin":

"...The Tsar's refusal to change the state according to the spirit of a real democracy [...] caused in Russia an atmosphere of profound animosity, which would surely detonate [...]."

This is an absolutely exceptional phrase. In the other versions of *POTEMKIN* there were no references to "democracy". Director Sergej M. Eisenstein was perfectly willing to admit *POTEMKIN*'s aim of communist agitation, and actually this statement was (and is) widely accepted. Moreover, the Moscow premiere version was introduced by a quotation of Leon Trotsky's revolutionary essay "1905"

Although the "real democracy" in the German sound version could be interpreted as "the proletariat's dictatorship" the wording of the title remains extraordinary and unexpected. This severe change in *POTEMKIN*'s synchronization can be better understood if the (unique) full permission of the film in Germany is considered: On 28th July 1926 the German censorship board had viewed the struggle on the "Prince Potemkin" as driven by pro-democratic and pro-republican values. Moreover, they had understood the audience's applause as a manifestation in favor of the constitutional order and democracy. This line of argument not only totally contradicted the previous rulings but also the film itself. All indications are that Piel Jutzi's significant change in the German sound version was "inspired" by the arguments laid down in this censorship decision.

So the work done in COLLATE clearly illustrated the potential of all participating archives, especially mutual access to material. But the concept and the implementation of the collaboratory were even more important because they made mutual, comparative studies possible. This is how new and valuable ideas in film studies evolved and at the same time existing hypotheses were revised. Furthermore, unpredicted historical findings were made possible through COLLATE, for example, a registration card for *BATTLESHIP POTEMKIN*, thought to have been lost, was found. Thus, the idea of making documents available in a collaboratory was established in practice.¹

The project basically showed that the idea of "collaboratories", which has its origins in the Natural Sciences, can also be used in the Humanities. The resounding success of our lectures and essays, enthusiastically embraced by numerous colleagues, proves our case. This also includes numerous queries and the desire for long-term, public access to film information. Typical questions or queries concern banned films, the international history of censorship for a particular film, thesis advice for university scholars, to name a few examples. Acceptance of the project has been expressed through the fact that many researchers have shown a keen interest in accessing the system.

¹ This German registration card – found in Prague – will be used to support a restoration project which is currently running in Germany.

The archives have put together a CD ROM as an interim solution. This is where the most important findings of their project work are summarized. A broad basis for discussion and general information can also be found on the CD ROM. Furthermore, plans are being made to publish a book with film researchers from other countries.

The participating archives wish to make the COLLATE system publicly accessible after the project has been completed. This is how an international platform for censorship studies and subjects in various related fields could possibly evolve. More importantly, the whole idea behind the project can be implemented: a collaboratory in use.

4.4.2 Empirical Evaluation of User Experiences

The Cognitive Systems Engineering Center at Risø National Laboratory provided a qualitative work domain analysis to determine the users' requirements to the COLLATE Collaboratory in collaboration with the three archives DIF, FFA, NFA and the Fraunhofer Institute in Darmstadt (*Deliverable 9.1*). All work functions relevant for the work tasks of retrieval, annotation, indexing and classification in each archive was covered in the work analysis. A set of these requirements from professionals were selected for implementation in the design of the content, individual and collaborative functions and the interfaces of COLLATE prototypes with the target user group being the professionals who work in archives. This work analysis also provided the foundation for the empirical evaluation of the different task facilities of COLLATE (*Deliverable 9.2*). The COLLATE collaboratory with cataloguing annotation, indexing, and retrieval of digitized, censorship materials was developed as an iterative process of design, implementation, and evaluation of several COLLATE prototypes numbered 1, 2 and 3.

Furthermore, a comparison of the users' needs in the three archives is provided together with a comparison of their work tasks, their expertise and the organizational context of their tasks. The purpose was to identify a common ground for collaboration in terms of similarities as well as the differences that might constrain the collaboration mediated through the collaboratory.

Work Analysis of the Censorship domain

In the past censorship material has been used as a means to get contextual information about an individual film. Now the film archives are increasingly interested in performing comparative studies of censorship, the institutions of censorship, and the processes involved. This task is a new activity and to a large extent only made possible by a system such as the COLLATE prototype. Such comparative studies will be concerned with a broad range of issues. The core of the task using censorship documents is the activities that involve analysis of censorship material in order to create new information and knowledge about the science of social film history. This is done through analysis of the themes of the films, their social context, and the way in which they were treated by censorship authorities in different countries at different times. On the censorship cards you will sometimes find detailed descriptions of scenes. These descriptions can be used in working out whether individual scenes were included in or excluded from the original version of the film, and sometimes the rationale for such decisions will also be available. Often, the work with censorship materials involves discussions with archivists from other archives. Such collaboration can be accomplished through email but it can be supported better and more fully by a collaboratory. This is done partly by providing shared access to the censorship documents, partly by the annotation facilities which supports indexing and more subjective free comments and discussions.

Evaluation of COLLATE prototypes

The overall objective of the evaluation of the COLLATE prototypes were to determine the degree to which the emerging prototypes and their different versions were in agreement both with the objectives of the design of the COLLATE collaboratory, and in agreement with the real life needs of its users, the archivists, as they were identified through extensive empirical studies. The evaluation of the final COLLATE prototype furnished evidence that all these goals have been met to a lesser or larger degree.

The reasons why some features have completely met these objectives while other features did meet the objectives only to some extent are manifold and very complex, and they are out of the scope of this presentation.

Evaluation methods

Evaluation of the COLLATE prototypes were carried out as an *empirical validation* of the COLLATE prototypes. The validation task aims to assess how well the three prototypes support the users' tasks, when they are working with censorship documents and what the system must provide to match the users' needs regarding support of collaboration (in two Interim reports). The empirical validation of how well the prototypes support the users' tasks was accomplished through the "Evaluation workshop" which enabled a comparison of the work analysis of archivists' use of censorship material with the content and functionality of the COLLATE prototypes. The validation also included the aspects of the prototypes that are confusing and constrain the actions and preferences of the users.

Secondly, the evaluation included an *analytical verification* of the COLLATE design, which is an assessment of the degree to which the prototypes meet the design specifications as they are derived from the empirical work analysis of users' needs and from the iterative empirical validation of the different prototypes. Verification took place as an analytical *usability* evaluation of the interface of the prototypes. In addition, analytical evaluation of the prototypes took place in order to ensure that the users' complaints about the prototypes were given properly attention during the redesigns of the different versions.

Prototype 1: Usability evaluation of the COLLATE interface

The usability evaluation of the interface of the COLLATE prototype 1 addressed:

- *The Document pane* – the upper, left pane, which contains the scanned-in documents.
- *The document representation pane* – the upper, right pane, which holds the cataloguing, keywords, free comments.
- *The Cataloguing tab* – the first tab sheet of the document representation pane containing information relating to the document type.
- *The Keywords tab* – the second tab sheet of the document representation pane containing the keywords assigned to the document.
- *The Free comments tab* – the third tab sheet of the document representation pane containing the comments assigned to the document.
- *The Filmographic info pane* – the bottom pane, which contains information about the film with which the document is concerned.

A number of issues were identified that would confuse or slow users down as well as a some issues that violated established conventions and heuristics for interface design.

Mandatory: The prototype 1 was made for a bigger monitor than some of the archives and Risø are working on, hence the bottom of the keyword window (free keywords) cannot be seen.

- I would be good if it was possible to have more than one document open at a time. When the archivists are working on a document, and they need to check something in another document, they have to close everything down, open the new document and find what they are looking for. Then they have to close everything down again, and open the first document.
- When you close a window, it is not possible to open it again. This means that you have to close everything and search for the document again. It should be possible to open and close all windows at any given time.

Redesign: These problems were accommodated and removed in the subsequent Prototypes 2 and 3. The mandatory requirements were implemented and all of the proposed changes were accommodated and removed in the subsequent prototype. Most importantly, the interface concept was changed fundamentally from a stable structure with many constraints to a more flexible and adaptive structure.

Prototype 1 and 2: Evaluation of the functionality of the collaboratory

The empirical evaluation of all aspects of these prototypes was executed by means of an “evaluation workshop” at the Vienna Archive FAA, which was led by Risø National Laboratory. The evaluation study followed a qualitative empirical research design, which implies that the design of the study is flexible in order to accommodate the addition of new topics and foci during data collection and analysis. The research design was open ended in order to encourage feedback and other contributions to the research by the participants of the study. The workshop addressed real-life collaborative work tasks amongst archivists from DIF, FAA and NFA. The four different work tasks, cataloguing, searching, indexing and annotation, were chosen as the frame for this evaluation because of their relevance to the work-domain independent of technology. For the workshop two films were chosen in collaboration between Risø National Laboratory and the three archives. In choosing to work with the film “Battleship Potemkin” in one real life task we were confronting an extremely complex censorship case that each of the archives held a considerable deal of information about. By this, the archivists’ wanted to highlight a potential case of collaboration.

The workshop also included two meetings with the participating group of archivists. The first meeting introduced the evaluation method and the second prototype to the archivists. It also contained a discussion about the use of Sharepoint. The second meeting was conducted as a focus group interview, where the participants contributed with their observations about their experience with the COLLATE prototypes. Both meetings and laboratory sessions were videotaped and the data from all sessions have been transcribed and analyzed.

Results from evaluation of Cataloguing

From the observations we were able to make during the Vienna workshop it is clear that cataloguing is the work task that COLLATE is currently best equipped to support. During the scenarios we worked with the archivists on, cataloguing of the documents that had been retrieved proceeded without difficulty as the COLLATE prototype features a separate interface for document cataloguing. This provided all of the fields that were required by the archivists in order to register a particular document. Again this should be considered as an important achievement, not least because the different archives are sometimes working with different types of documents and yet these national idiosyncrasies have apparently been incorporated into a cataloguing system that all can use. Important in itself this result is also significant for what it says about the prospects and potential of collaboration. Clearly cataloguing is, as a work task, an individual activity but at the same time it is, in this context, also a prerequisite for higher-level collaboration. As much as cataloguing is an individual task it should also be noted that it can only be undertaken on an individual basis in light of collective agreements. Such collective agreements are not always easy to broker in an international context since organizational and cultural differences play a role, so the result that has been achieved in relation to cataloguing does provide demonstrable proof that cross-cultural, inter-organizational collaboration in the film archive domain can be realized.

Results from evaluation of Searching

Searching for a document, that is a censorship decision, a registration card or an article, is an important part of the archivists’ work when using the prototype, and it is also a prerequisite for the scientific work with the different censorship documents. The COLLATE Prototype 1 offers three kinds of search functions: filmographic, DIGIPROT and filename search.

Of the three different search functions available at the time of our evaluation only one, *filename search*, can be considered unproblematic. This alludes to the fact that it is a search function that works effectively in the COLLATE prototype and supports a search strategy used by archivists in their normal task activity. It is rather limited, however, because it only retrieves one file and because the situations in which the archivists would want to retrieve a specific file are quite restricted. *Filmographic search* is unproblematic in the sense that it covers a type of search strategy that is essential for the work that the archivists undertake but its actual realization in the COLLATE prototype worked with in our evaluation is, on the other hand, far from unproblematic. In fact it was observed that the archivists tend to ignore filmographic search when retrieving documents since it is their experience that the results of such

searches are by no means exhaustive. It was also clear, however, that the archivists continue to think 'filmographically' when drawing general associations in the film domain or even looking for particular facts. Given this cognitive disposition, which has been engrained in the archivists training and professional experience, much more comprehensive support for this function is still required. Other factors mentioned, such as support for document evaluation and interruption of search, also need to be considered, but the requirement for a fully functioning filmographic search capability should be regarded as one of the most significant obstacles, potentially threatening the utility of COLLATE as a practical work-tool within the archives. In lieu of a fully functioning filmographic search function the archivists have turned to using DIGIPROT as their most effective method for retrieving documents. More by accident that design has proved to be the search function that produces the most reliable results but its effectiveness should not take away from the fact that *DIGIPROT* was conceived for entirely different purposes and is, for the most part, only effective because the archivists have spent much time learning how to work with and around it

Main Design suggestions to improve search functionality

The analysis of the evaluation data identified three major problems with the existing search functionalities in the COLLATE prototype:

1. Technical problems with the search functions. 2. Support of selection and evaluation of relevance of documents from the list of search results. 3. Inadequate support of search strategies. DIGIPROT search. Filmographic search. Advanced search. Filename search: The following search functions are desired: Search on title, specific document type, publication date and a combination of these. Content-based access to documents is very important. This requires for example search in annotations and search on keyword, genre, person names and companies. The prototype should also allow users to search for documents and films, which are similar to a document or film they are working on.

Mandatory: Among these design suggestions, it is most important to make sure that

- Technical problems with the search functions in Filmographic Search are solved
- Full text search should as a minimum search in both annotations and in the cataloguing abstract.
- Search in indexing is also absolutely necessary either in the full text search or separate.
- Finally, it is mandatory that the user can specify if the chosen search criteria is the exact title of a film.

Redesign: All the mandatory suggestions for improvement of the search functionality have been fully implemented.

Results from the empirical evaluation of Indexing

Collaborative management of knowledge is a key problem in COLLATE and is defined by the project as collaboration on indexing and annotation of collections of documents. Collaboration on indexing is supported by a separate indexing interface, The problem of censorship indexing is complex. This is partly due to the ambiguity or interpretive flexibility of censorship domain semantics. The complexity might be reduced through the availability of an overall stable structure, reflecting the semantics of the work with censorship materials. The interface does not adequately make visible a distinction between the semantics of indexing on the one hand (options for subject analysis and keyword assignment) and the syntax of indexing on the other hand (options for supporting a particular indexing process).

Our focus on indexing work revealed the fact, also pertinent to annotation work, that the documents the archivists work with cannot always be considered or treated in isolation. It was seen, for example, that even prior to the point of cataloguing and indexing a document, certain pre-selection criteria would be applied. These criteria were, moreover, not merely due to the cognitive dispositions of the actors involved but, rather, integral to the nature of the work. This points to the fact that it was not the individual document as such which is important but more the censorship case of which it could be considered a part. The censorship case is what determines when and how a document should be catalogued and indexed. Once a document is identified at the core of a censorship case the information it contains will bear on all subsequent documents and information pertaining to the same case. In this light it was

concluded that COLLATE, which is necessarily biased towards individual documents, does not provide much in the way of support concerning subject analysis. This is not ideal but it was impossible to know prior to observing the archivists actually working with censorship documents that this would be the case and it does, at any rate, appear that the archivists are able, through secondary sources and general experience, to circumvent the obstacles that this presents.

Keywords were a much-discussed topic during the course of the Vienna workshop and a good deal of the input from the archives was couched in quite negative terms. To a large extent this disaffection appeared to derive from the fact that the archivists did not feel that they had exerted a sufficient influence on the design and organization of the keyword list. The structure of the keyword list seemed to be the issue that vexed them most, for while there appeared to be little wrong with the keywords that had been selected the archivists felt that the sub-categories were not always helpful. This observation should, though, be seen against the background that the archivists had requested more specific sub-categories than were present in the original prototype. The fact that they were not altogether satisfied with those they obtained is not necessarily due to a design failure, however, since the 'interpretive flexibility' of terms relevant to the censorship domain sometimes resist easy classification. A further difficulty appeared to rest in the anticipated target audience, for while some felt the structure of the keyword list was understandable from the point of view of an expert there remained a concern that end-users would find the keywords more difficult to work with.

Main design suggestions to improve Indexing

The analysis of the evaluation data identified further need for support of the following indexing activities in the COLLATE prototype:

Choice of a starting document for indexing. Subject analysis. Planning entries. Choice of keywords from lists. Entering keywords for a document. Evaluating the indexing of a censorship document. Search for information resources.

Mandatory: It is absolutely necessary:

- To be able to enter more than one keyword in each indexing category.
- The keyword categories are problematic when the archivists want to use a keyword from one category for another purpose. The example given is that war from the reality topic can sometimes be used to describe a film scene.
- Sometimes keywords are used across the categories, and it is necessary that the user is able to use free keywords instead of controlled keywords.

Redesign: The necessary improvements of the indexing facility has been undertaken.

Results from evaluation of the Annotation

A collaboratory system should by definition contain some support for cooperative activities. For the time being cooperation is ensured via a shared access to a common work space, a structured discourse module for handling annotation content ("My comment") and a module for structuring communication based on a conversation for action approach ("My request re:" and "My request on my comments"). Different styles of annotation suggest different forms of collaboration. This needs to be taken in account when considering what type of support the COLLATE collaboratory should provide for annotation work, as we need to know what form of collaboration provides most in the way of added value for the archives.

Coordination of annotation work

If we look at the users reaction to this type of structured collaboration (structured comments and discussion on annotation) the users bring forth that the communicative acts used in the current version reflects coordination work carried out in the COLLATE project with respect to establishing a common ground for collaboration amongst them, for example to establish a common understanding of the different practices applied in the different archives. However, there are no explicit reference in making requests and promises. In the current version the interpretation of the state of affairs may be rather difficult. It may

be difficult to engage in decision-making activities, with respect to judging a certain request, since it seems that the system does not, in an adequate way, provide the users facilities for engaging in the collaborative decisions needed to coordinate their work. For example in terms of exploring contextual information related to requests (or promises). This means, that it may be difficult for users to establish a coherent analysis of a given situation and identify the relevant information entities that might smooth the progress of the coordination activities.

Categorizing annotations with discourse elements

With respect to the list of discourse typologies that is meant for categorizing annotations, the users seem to like the idea of having the possibility to give comments on annotations in a structured way. It is positive that the users can give annotations to a parts or selections in a document and that they can ask for more information using the "My request on my comment" facility. In addition, the quotation indicates that there is a positive attitude towards having categories that are similar to those implemented in the "My comment" feature, but it seems that the current list of categories does not fully encompass the way user would categorize their comments on comments. Especially there is a concern that it is not possible to combine the categories - that there could be a need for being able to combine different types of discourse elements to give a proper background of a certain comment.

Constraints from discourse elements

Another concern is that the predetermined list of discourse elements could impact current work practices in an unwanted way by imposing a fixed list of discourse elements that do not meet the requirements imposed on the archivists, by the state of affairs in the work space and the wider work environment. There is a concern that forcing archivists to follow a predestined type of discourse logic could impinge on individual's way of working. In other words, there might be a semantic incongruity between the list of discourse typologies that is meant for categorizing annotations and the semantic categories in which the archivists' express their domain knowledge during their daily work. Similarly, the domain terminology used in the list and the domain terminology applied by archivists also differ. Since there exists both a semantic categorization and a terminology inconsistency, the archivists will probably to a large extent only use the "Unspecified Relation" menu choice.

Annotation task strategies and discourse elements

The importance of the censorship case and the interests and knowledge of potential users were also found to be important factors in the case of annotation. It was witnessed here, for example, that very different strategies with regard to the aims and ends of annotation could be applied on the same type of documents. It was also apparent that the strategies observed could be used with a particular view of collaboration in mind. One strategy focused on rigorously providing all the available context information about a document. Establishing this foundation of relevant, more factual context information is, in this view, the prerequisite for collaboration, since it provides a body of knowledge that can be accessed and interrogated by autonomous experts. The other strategy focused on a thematic type of annotation and this strategy takes collaboration to be an emergent process in which the annotation is directly part of the collaboration. Both strategies have their virtues although the fact that they anticipate very different ways of using the collaboratory suggests that they are not particularly compatible with one another. The lack of compatibility might be overcome if the structured discourse model for handling annotation content was used by the archivists. Our findings here, albeit tentative, are that the archivists find the discourse model too structured at present. This was particularly opposite to the design of the annotation function, which allows the archivists to choose only one of the possible annotation categories. The widespread feeling among the archivists was that annotations tend to cover a number of annotation categories. There was also some skepticism concerning the designation of an annotation category to mean this or that when such decisions are prone to be subjective and might even be misleading for others with a different perspective.

Alerts and online communication

Nonetheless, the principle of providing the archivists with some form of independent communication channel where they could exchange information unsuitable or unusable for end-users is important. It was notable, for example, that we received very positive feedback regarding the experience of using Sharepoint. More to the point, the archivists seemed to regard that the structuring of their

communications allowed by Sharepoint was beneficial in terms of providing a framework for their discussions with each other. This was reflected in the fact that the archivists, though wary of classifying their annotations, were more open to communicative options allowed to them in the prototype. The only problem with this was that the comments on comments etc. did not allow for any direct notification to the interested parties. So in order to be aware that another archivist had made a comment to an annotation, the author of the annotation would actually have to access his or her own comment. Quite clearly this is not something that the archivists would, in the normal course of events, choose to do, since, having written the annotation, there is very little reason why they would return to read it again. For the archivists, experienced working with Sharepoint, it seemed much more logical to have some communication forum where relevant information is immediately available. In the absence of this then there preference would, if they wished to obtain a response, to email directly with the person concerned. Given this, the option of requesting a response with respect to an annotation or a comment on an annotation, while an important function for collaborative work, is redundant in the context of the COLLATE prototype. The question, then, is whether efforts in the next phase of development should focus on integrating this collaborative functionality more fully into the COLLATE interface or whether it would be better to establish some independent collaborative layer using a program such as Sharepoint or Messenger.

Main design suggestions to improve Annotation Functions

Although annotation tasks appear in the archivists' daily work, it should be noted, however, that the annotation task has never been conducted as a collaborative task or with censorship material among different experts from different archives, and never by means of any support tool. Therefore, the design of the annotation facilities aiming at the creation of new collaborative task activities, is challenged by the fact that the empirical work analysis of users' perception of how such a task should be performed suffered from the limitations of having no such experience. In this situation, the Annotation functions of the prototype used by several international participants during the evaluation process actually worked as an instrument for crystallizing the collaborative activities that the users' realize are possible and which they want to embark on. Concurrently with this recognition they used the opportunity to discuss a number of emerging visions and requirements to the Collaboratory they were using.

These ideas and requirements addressed further support facilities in: Writing an annotation. Targeting annotations. Categorizing comments. Linking and marking documents. Initiating collaboration. Enter comment. My Comment" list of discourse elements. "My request re" and "My request on my comment". Alert and online communication.

Mandatory: Users should be notified automatically about changes to the database and give dynamic feed back. The system should offer a possibility to see what changes has been made to the database by whom, and when, with some kind of alert.

Redesign: Most important is the implementation of a new online communication and notification facility in "My request...", for which the archives argued very enthusiastically. This new facility haven't been well tested in detail but the archivists have given positive feedback to this implementation. They found it to be a really important feature, since it provides some of the functions otherwise offered by Sharepoint. Just a small subset was defined, i.e. those allowing explicit requests that are directly related to individual annotations and comments or documents. This meta-communication in COLLATE is treated differently from other annotations/comments and will not be displayed to others than the contributing partners.

Results from evaluation of support of work tasks

The four different work tasks, cataloguing, searching, indexing and annotation, chosen as the frame for this evaluation were selected because of their relevance to the work-domain independent of technology. This approach, besides serving as a means for organizing the data obtained during the Vienna workshop, also has the virtue of being consistent with the work domain analysis conducted at the outset of COLLATE (see *Deliverable 9.1*). The Vienna workshop achieved more than confirming our original findings, however, since the opportunity to observe the archivists actually working with the prototype COLLATE system emphasized the relative importance of the different work tasks in relation to the overall goals and priorities of the archives as work domain. One of the most positive findings in our evaluation,

therefore, is that the COLLATE prototype was structured in a way consistent with the tasks that archivists need to perform. All of the tasks that we identified were supported to a greater or lesser extent by the COLLATE prototype and, as importantly, the system also allowed them to be performed sequentially in a way that was consistent with the archivists' interests. This is an important achievement. Our findings also show that further work needs to be undertaken, though, since while all the critical work tasks are supported, the extent to which they are supported is in certain cases still insufficient.

Analytical evaluation of COLLATE Version 2 and 3

The analytical evaluation addressed how well aspects of the final version of the prototype met the requirements from the Vienna workshop as described above. All the user requirements were met that were deemed mandatory for having a full scale prototype working. One of the most positive findings in our evaluation, therefore, is that the COLLATE prototype was structured in a way consistent with the tasks that archivists need to perform. All of the tasks that we identified were supported to a greater or lesser extent by the COLLATE prototype in a way that was consistent with the archivists' interests. The diverse visions and requirements from the end users to the Annotation facilities (some of which were more ambitious than the a priori defined boundaries of the COLLATE project) were very encouraging for an even more tight and dynamic collaboration among film experts. They new ideas emerged late in the project and they had to be postponed due to the closure of the project.

Evaluation of the collaboratory's impact on the quality of work

The overall objective of the development of the COLLATE tool was to ensure:

- Enhancement in the *quality* and effectiveness of work by providing access to materials and knowledge that are otherwise not accessible
- Increasing international *collaboration* among experts and users of the materials stored in the collaboratory
- *Knowledge acquisition* through professionals' analysis, evaluation, indexing, and annotation of film materials
- Collaborative *accessibility* of a content-centric, user-driven information system and working environment
- *Acceptability* of a collaboratory in the film domain by the professionals' through real-life experiences of collaborative work with the tools of the collaboratory

The evaluation of the final COLLATE prototype furnished evidence that all these goals have been met to a greater or larger extent. This is an important achievement in itself, also providing the necessary conditions for further work to be undertaken to meet all the new ambitions (beyond the scope of this project) that were developed among the film domain experts during this project.

5 European Added Value

The goals and use intentions of project COLLATE precisely correspond with the EU-specific objectives of the relevant IST Programme's action line. Hence, we aimed to develop "systems for creating, processing, managing, networking, accessing and exploiting digital cultural content" to be applied and used – first of all – by European shareholders, and to focus on "the sustainable development of valuable digital repositories in Europe's libraries, museums and archives". These objectives concern both the *contents* (the final COLLATE system offers content-based access to a large repository of material on historic European film productions from various European countries) and the developed *technologies* (advanced software supporting analysis, management and usage of the digital repository for scientific analyses and access by the wide public, jointly developed by European partners from Germany, Italy and Denmark).

COLLATE addresses issues of the preservation of European heritage on a multi-national level, an inherently and pre-eminently European concern. The cultural treasures preserved are products of the film industries of several countries, and their documentation. Community added value has been achieved by facilitating scientific cooperation across European borders through employing the collaboratory in which scientists have and will contribute to cultural preservation by analyzing, documenting and making accessible to a wide European public an important cultural medium of this century.

The chosen domain itself, i.e. film documentation from the 20ies and 30ies, has provided European added value by allowing more insight into one of the darker chapters of European history and its impact on cultural activity, i.e. the censorship of films rampant before and during World War II. The historic film domain per se demonstrates the necessity of an international collaboratory like COLLATE in an impressive way, since moving pictures have never been restricted to national borders. Films were not only often co-produced by several countries, they also have always been distributed worldwide. Political scientists, historians and film scholars agree that, especially in the case of cultural artifacts such as films, it is only through study of their international reception that a scientifically complete documentation and critical analysis is possible. The film archives used existing cooperation with other similar institutions and memberships in international film organizations to make available further relevant material throughout Europe – and they will continue these activities after the end of the project.

Some of the achieved results in COLLATE were only made possible by cross-national collaboration on the European level. The countries of the film archives/institutes in Austria, the Czech Republic and Germany were historically involved in many film co-productions and were strongly affected by censorship in the 20ies and 30ies. Before showing their films each production company in Germany, Austria and Czechoslovakia (or the importers of foreign productions) had to submit their films for approval at the (national/local) censorship authorities, which decided about permission or ban as well as about age restrictions or cuts. These censorship processes were documented in the digitized collections used within COLLATE, and access to the integrated collection allow cross-national comparisons of film censorship practices in this period in Europe.

The project has provided synergistic effects in all relevant areas of Community interest: preservation of cultural heritage, facilitation of cross-border scientific cooperation and historical study in a political context. Moreover, if we consider future potential applications, such as to share knowledge at different levels in various economic and professional sectors, we can say more generally that follow-up systems of COLLATE will allow also European citizens to share important, cross-nationally and scientifically authorized areas of knowledge and information at a European level.

6 Outlook

The implemented COLLATE system is one of the first working Web-based collaboratories in the Humanities and Cultural Heritage. It offers a comfortable work environment for the de-centralized collaboration of scholars from various EU-countries in the field of film history. Interconnecting cultural archives via advanced technology supports their institutional effectiveness and visibility and allows broader dissemination of work results across Europe. Since the respective content-providing community (concerned with film archiving and documentation) has never before been involved in such a collaboration-intensive, multinational project, the outcomes of COLLATE will be substantially guiding for similar or comparable application domains.

After the end of the COLLATE project, the participating film archives/institutes (DIF, NFA and FAA) will concentrate on the dissemination of their research results to the interested public and the film industry throughout Europe. In particular, they plan to win other contributors for the COLLATE collaboratory through (1) contacts with large international scholarly societies, e.g., FIAF, Fédération Internationale des Archives de Film, or ACE, Association des Cinémathèques Européennes and existing connections to various national state archives; (2) cooperation with individual film institutes and universities (e.g., by conducting seminars) in order to attract students of film science and history to participate in the future COLLATE collaboratory.

The technology developers, on the other hand, also intend to reuse the implemented results as far as possible both in their future research and for extended and new applications of the software system. Although the COLLATE system accounts for user requirements in the film documentation domain, the developed technology is largely generic and easily adaptable to other application domains that are similarly information-intensive and profit from collaborative knowledge work. Therefore, some of the technology partners aim to adapt the COLLATE system as a whole and/or specific modules to new content domains, i.e. assisting the new user groups in adapting the software to their specific user and usage requirements. For application domains and user groups whose work processes are comparable to those of the current COLLATE users, e.g. other film documentation and research centers or other memory institutions and associations, pilot installations and respective consultation services can be offered shortly. In this case some accompanying evaluation of the work processes may be feasible in order to guide future research. Less similar domains and user requirements would require a more extensive tailoring of the system, which is possible as well due to its modular configuration (e.g., the document pre-processing modules could be isolated, the currently employed search engine might be replaced by another one, etc.).

All COLLATE partners are most interested in continuing their individual research and developments performed during the project, looking for follow-up projects and cooperations whether within the relevant research areas or addressing the relevant industries, i.e. IT developers and application partners/users. Currently starting European projects and NoEs in the cultural heritage domain will benefit from the results, components and experiences of the COLLATE project. In the integration project BRICKS, for instance, the annotation management of COLLATE will be reused, whereas our experiences with a challenging real-life application contributes to the empirical knowledge assimilated in the DELOS NoE. Intensifying the dissemination and exploitation activities we also hope to win over future potential partners that are willing to put forward some of the open research issues or to employ the COLLATE collaboratory in different real-life application contexts.

7 Conclusions

In this report we have summarized the project objectives, the methodologies applied and the major project results achieved in COLLATE. Rather than describing all project activities and achievements in much detail, the report gives a relatively concise overview and points out the main characteristics that distinguish our developments from other, comparable projects or implemented systems. For extended

descriptions of specific work packages and resulting developments we refer to the 18 comprehensive project deliverables (see *Section 9.1*) and – more importantly – to the large number of publications produced and published in the course of the COLLATE project (see *Section 9.2*).

Main achievements of the COLLATE project were the development, implementation and real-life evaluation of a “collaboratory in use” for the preservation of and access to historic cultural heritage material. To this end we followed two complementary strategic overall goals:

- Ensure content-based accessibility of the digital document collection:
A Web-based system supporting document-centered collaborative work of distributed user groups was developed. It offers content-based access to a newly set up, comprehensive digital library on European historic film documentation, which has been analyzed, indexed and annotated by a multi-national team of film experts during the project. The final version of the COLLATE collaboratory additionally includes innovative/advanced document processing and management facilities, e.g., XML-based document handling, tools for automatic document structure recognition, digital watermarking and semi-automatic indexing of pictorial material.
- Establish evidence for the acceptability of the collaboratory approach in the current domain:
We evaluated in detail experiences and requirements of professional real-life users who worked as pilot users with the COLLATE system in the last two project years, i.e. a team of film experts who indexed and annotated the document collection and collaborated in performing case studies in historic film research.

As a prerequisite for content-based *accessibility*, the COLLATE work environment provides domain experts with facsimiles/images of digitized text documents that can be marked up to assign (controlled or free) index terms and free-text annotations. The users are supported in the full process of analyzing, cataloguing and indexing the document contents, entering free-text annotations to documents or marked-up document passages and annotate other annotations by other users, thus building up a complex discursive annotation thread. All user-generated metadata are stored and made available as a basis for ongoing scholarly document evaluation, and at the same time allow employment of highly precise search and retrieval functionalities.

Combining results from the manual indexing by the users and automatic indexing procedures, elaborate content-based retrieval mechanisms can be applied, combining evidence from various so far unrelated sources and the constantly evolving knowledge in the system. To capture these dynamics we chose XML as the de-facto standard for the encoding of generic document and metadata representation schemata. The use of XML guarantees the generality of our approach since these schemata can be enriched and tailored to additional sources and knowledge incorporated into our system without any need for re-modeling the whole system.

In order to establish evidence of the *acceptability* of the collaboratory it was essential to support the complex workflows in the domain of film documentation. For this reason comprehensive models for task-based user interfaces supporting content-based document analysis and annotation were employed and tested under real-life work conditions by real users and domain experts. Members of the three COLLATE film archives/institutes (collection administrators, archivists, film scholars, etc.) worked as pilot users; they productively used the COLLATE system for their cataloguing, indexing and annotation work throughout the last two project years, starting with the first working prototype. In this approach, technology development and empirical evaluation of user experiences with the system were closely intertwined to allow an iterative, dynamic system development. Evaluation steps were explicitly built in and the users were actively involved in all development cycles of the COLLATE system.

By this means the COLLATE system incorporates much of the user requirements in the present domain of film history research. Most of the developed technologies, on the other hand, were designed to be generic and easily adaptable to different content and application domains that also require information-intensive, document-centered knowledge working. Employing the COLLATE collaboratory for new user groups – either in similar domains or in quite different content and applications domains (adapting the technologies to new requirements) – is a major concern of all partners with respect to the future use and exploitation of results of the COLLATE project.

8 References

- Brocks, Holger; Stein, Adelheit; Frommholz, Ingo; Thiel, Ulrich & Dirsch-Weigand, Andrea (2002).
How to Incorporate Collaborative Discourse in Cultural Digital Libraries. In: *Proceedings of the ECAI 2002 Workshop "Semantic Authoring, Annotation & Knowledge Markup" (SAAKM '02) at the 15th European Conference on Artificial Intelligence (ECAI '02), 21-26 July 2002, Lyon, France.*
- Brocks, Holger; Thiel, Ulrich & Stein, Adelheit (2003).
Agent-Based User Interface Customization in a System-Mediated Collaboration Environment. In: Harris, D. et al. (Eds.): *Human-Centred Computing*. Mahwah, New Jersey, London: Lawrence Erlbaum, pp. 664-669.
- Buschmann, F.; Meunier, R.; Rohnert, H.; Sommerlad, P. & Stal, M. (1996).
Pattern-oriented Software Architecture. John Wiley and Sons 1996.
- CCSDS, 2002
Consultive Committee for space Data Systems, January 2002
Reference Model for an Open Archival Information System (OAIS),
<http://www.classic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf>.
- Craver, S., Memon, N., Yeo, B.L. (1998).
Resolving Rightful Ownerships with Invisible Watermarking Techniques: Limitations, Attacks, and Implications. In: *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 4, 1998, pp. 573-586.
- Eakins, J.P.; and Graham M. E. (1999)
Content-Based Image Retrieval. A report to the JISC Technology Applications Programme
<http://www.unn.ac.uk/iidr/research/cbir/report.html>
- Fridrich, J. (1999).
Methods for Tamper Detection of Digital Images. In: Dittmann, J., Nahrstedt, K., Wohlmacher, P. (Eds.), *Multimedia and Security, Workshop at ACM Multimedia '99, Orlando, Florida, USA, Oct 30 – Nov 5 1999*, pp. 29-34.
- Hollfelder, S; Everts, A.; and Thiel, U. (2000):
Designing for Semantic Access: A Video Browsing System In: *Multimedia Tools and Applications*, Volume 11, Issue 3, August 2000. Pages. 281-293
- Kalker, T. (1998).
A Security Risk for Public Available Watermark Detectors. In: *Benelux Information Theory Symposium*, Veldhoven, The Netherlands, May 1998.
- Keiper, Jürgen; Bezerra, Laura; Stein, Adelheit; Brocks, Holger & Thiel, Ulrich (2003).
Collaborative Film Documentation and Research: The COLLATE Collaboratory. Paper presented at *1st International Workshop on Web-Based Collaboratories (WbC-2003) – From Centres Without Walls to Virtual Communities of Practice*, held in conjunction with the *IADIS WWW/Internet 2003 Conference*, 5-8 November, 2003, Carvoeiro, Portugal. Updated version to be published 2004 in: *Journal of Digital Management*, special issue on Web-Based Collaboratories.
- Keiper, Jürgen; Brocks, Holger; Dirsch-Weigand, Andrea; Stein, Adelheit & Thiel, Ulrich (2001).
COLLATE – A Web-Based Collaboratory for Content-Based Access to and Work with Digitized Cultural Material. In: Bearman, D. & Garzotti, F. (Eds.), *Proceedings of the International Cultural Heritage Informatics Meeting (ICHIM '01), Milano, Italy, 3-7 September 2001*. Milano: Politecnico di Milano, 2001, pp. 495-511.
- Kohl, U., Lotspiech, J., Kaplan, M.A. (1997).
Safeguarding Digital Library Contents and Users. In: *D-Lib Magazine*, September 1997, ISSN 1082-9873.
- Kouzes, R.T.; Myers, J.D. & Wulf, W.A. (1996). Collaboratories: Doing science on the Internet. In: *IEEE Computer*, Vol. 29, No. 8. See also <http://www.wvu.edu/~research/DOE/IEEEcollaboratory.htm>.

- Maes, P. (1994). Agents that Reduce Work and Information Overload. In: *Communications of the ACM*, 37(7), 1994.
- Maier, Elisabeth & Hovy, E.H. (1993).
Organising Discourse Structure Relations Using Metafunctions. In: Horacek, H. & Zock, M. (Eds.), *New Concepts in Natural Language Processing*. London, Pinter, 1993, pp. 69-86.
- Mann, William C. & Thompson, Sandra A. (1987).
Rhetorical Structure Theory: A Theory of Text Organization. In: Polanyi, L. (Ed.), *Discourse Structure*. Norwood/NJ: Ablex, 1987, pp. 85-96.
- Searle, John R. (1979).
A Taxonomy of Illocutionary Acts. In: Searle, J.R. *Expression and Meaning. Studies in the Theory of Speech Acts*. Cambridge/MA: Cambridge University Press, 1979, pp. 1-29.
- Sitter, Stefan & Stein, Adelheit (1992/1996).
Modeling the Illocutionary Aspects of Information-Seeking Dialogues. *Information Processing & Management*, 1992, 28(2):165–180. Largely revised version: Modeling Information-Seeking Dialogues: The Conversational Roles (COR) Model. *Review of Information Science* [on-line journal], 1996, 1(1).
- Stabenau, M., Dittmann, J. (1998).
Digitale Wasserzeichen für MPEG Video. *GMD Report 34*, GMD-Forschungszentrum Informationstechnik GmbH, September, 1998.
- Stein, Adelheit; Gulla, Jon Atle & Thiel, Ulrich (1999).
User-Tailored Planning of Mixed Initiative Information-Seeking Dialogues. *User Modeling and User-Adapted Interaction*, 1999, 9(1-2): 133-166. Also available as reprint in: S. Haller, A. Kobsa & S. MsRoy (Eds.), *Computational Models of Mixed-Initiative Interaction*. Dordrecht, Boston, London: Kluwer, 1999, pp. 317-350.
- Stein, Adelheit & Maier, Elisabeth (1995).
Structuring Collaborative Information-Seeking Dialogues. *Knowledge-Based Systems*, 1995, 8(2-3): 82-93.
- Thiel, U.; Everts, A.; and Hollfelder, S.(1999a);
Beyond Similarity Searching: Concept-Based Video Retrieval and Browsing In: Proc. of the 10. DELOS Workshop, Santorini, Greece, June, 24 - 25, 1999
- Thiel, U.; Everts, A.; Lutes, B.; and Stein, A. (1999b).
Can rule-based indexing support concept-based multimedia retrieval in digital libraries? Some experimental results. In: Draper, S.W. et al. (Eds.), *MIRA 99: Evaluating Interactive Information Retrieval*. Berlin: Springer Verlag (eWiC, electronic Workshops in Computing series), see <http://www.ewic.org.uk/ewic/workshop/view.cfm/MIRA-99>
- Tomsich, P., Katzenbeisser, S. (2000).
Towards a Secure and De-centralized Digital Watermarking Infrastructure for the Protection of Intellectual Property. In: *Electronic Commerce and Web Technologies, First International Conference, Proceedings*. Springer Lecture Notes in Computer Science vol. 1875, 2000, pp. 38-47.
- O'Rourke, J. (1994).
Computational Geometry in C. Cambridge University Press, 1994.
- Wolf, G., Pfitzmann, A. (1999).
Empowering Users to Set Their Protection Goals. In: *Multilateral Security in Communications*. Addison-Wesley, 1999.
- Wulf, W. A. (1989). The National Collaboratory - A White Paper (cited in Kouzes/Myers/Wulf 1996). In: *Towards a National Collaboratory*. Unpublished report of a workshop held at Rockefeller University, March 1989 (co-chaired by Joshua Lederberg and Keith Uncapher).
- Zielhofer (2000).
"KOM-D-113: Entwurf und Optimierung von Bildwasserzeichen für den Einsatz im Internet-Szenario. *DiplomaThesis*, Aug. 2000.

9 Annex

9.1 COLLATE Deliverables

Del. No.	Deliverable name	Lead Partner	Del. Type ²	Delivery Date
D0	Project Web site and project description	IPSI	O	22.12.00
D1	Analysis of user requirements and project specification	IPSI	R	13.03.01
D2.1	Definition and set up of the digital MM test collection	DIF	R	13.03.01
D2.2	IPR rights: Digital signatures and watermarking interfaces	IPSI	R	18.09.02
D3	Task/Domain model and set up of the metadata base	IPSI	R	13.06.01
D4.1-1	Part 1: Document processing	Uniba	P	19.03.02
D4.1-2	Part 2: XML-based content management (versions V1 and V2)	Sword	P	19.03.02 01.10.02
D4.2	Application of annotation and retrieval methods	IPSI	P	29.04.02
D5	Integration of knowledge processing tools	IPSI	R	19.09.02
D6	Second generation image and video analysis tools	IPSI	P	24.10.02
D7.1	Design and implementation of the indexing/annotation interface Addendum: Collaboration support in P2	IPSI	P	19.09.01 19.09.02
D7.2	Integration of system components	Sword	P	30.10.03
D8	Experiences from the preservation case studies	DIF	R	12.11.03
D9.1	Study of user needs and behavior	Risoe	R	14.06.01
D9.2	Evaluation of the COLLATE prototype	Risoe	R	12.11.03
D10.1	Dissemination and use plan (versions V1 – V3)	Sword	R	13.03.01 15.10.02 12.11.03
D10.2	Dissemination activities of COLLATE	Uniba	R	12.11.03
D10.3	Technology implementation plan (TIP) – DRAFT	Sword	R	27.11.03
D11.1	Final project report - DRAFT	IPSI	R	27.11.03

² Type of deliverable: **R** = Report; **P** = Prototype; **D** = Demonstrator; **O** = Other.

9.2 COLLATE Publications and Conference Presentations

Fraunhofer IPSI

Publications

- Brocks, Holger; Dirsch-Weigand, Andrea; Keiper, Jürgen; Stein, Adelheit & Thiel, Ulrich (2001). **COLLATE – Historische Filmforschung in einem verteilten Annotationssystem im WWW.** In: Schmidt, R. (Ed.), *Information Research & Content Management - Orientierung, Ordnung und Organisation im Wissensmarkt. Proceedings der 23. DGI-Online-Tagung 2001.* Frankfurt am Main: DGI, 2001, pp. 183-196.
- Brocks, Holger; Stein, Adelheit; Frommholz, Ingo; Thiel, Ulrich & Dirsch-Weigand, Andrea (2002). **How to Incorporate Collaborative Discourse in Cultural Digital Libraries.** In: *Proceedings of the ECAI 2002 Workshop "Semantic Authoring, Annotation & Knowledge Markup" (SAAKM '02) at the 15th European Conference on Artificial Intelligence (ECAI '02), 21-26 July 2002, Lyon, France.*
- Brocks, Holger; Thiel, Ulrich & Stein, Adelheit (2003). **Agent-Based User Interface Customization in a System-Mediated Collaboration Environment.** In: Harris, D. et al. (Eds.): *Human-Centred Computing.* Mahwah, New Jersey, London: Lawrence Erlbaum, pp. 664-669.
- Brocks, Holger; Thiel, Ulrich; Stein, Adelheit & Dirsch-Weigand, Andrea (2001). **Customizable Retrieval Functions Based on User Tasks in the Cultural Heritage Domain.** In: *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001).* Berlin: Springer, 2001, pp. 37-48.
- Croce-Ferri, Lucilla (2003). **IPRs Protection for Digitized Historical Documents.** Paper presentation at: *Electronic Imaging & the Visual Arts EVA 2003 Florence; 24-28 March 2003,* p. 190-194.
- Frank, Markus; Spielmann, Simon; Croce Ferri, Lucilla & Dittmann, Jana (2003). **Schutz digitaler Dokumente mittels digitaler Wasserzeichen.** In: *Proceedings of the 25. DGI-Online-Tagung 2003. 3.-5.Juni 2003; Frankfurt am Main,* S. 100-112.
- Frommholz, Ingo; Brocks, Holger; Thiel, Ulrich & Stein, Adelheit (2002). **Kontextbasiertes Retrieval unter Verwendung verknüpfter Annotationen.** In: Schubert, S., Reusch, B. & Jesse, N. (Eds.), *Informatik bewegt. Proceedings der 32. Jahrestagung der Gesellschaft für Informatik, Dortmund, Germany, 30 Sept – 3 Oct 2002,* pp. 161-165.
- Frommholz, Ingo; Brocks, Holger; Thiel, Ulrich; Neuhold, Erich; Iannone, Luigi; Semeraro, Giovanni; Berardi Margherita & Ceci, Michelangelo (2003). **Document-Centered Collaboration for Scholars in the Humanities - The COLLATE System.** In: Koch, T. & Sølvberg, I.T. (Eds.), *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science 2769,* Berlin: Springer, 2003, pp. 434-445.
- Keiper, Jürgen; Bezerra, Laura; Stein, Adelheit; Brocks, Holger & Thiel, Ulrich (2003). **Collaborative Film Documentation and Research: The COLLATE Collaboratory.** Paper presented at *1st International Workshop on Web-Based Collaboratories (WbC-2003) – From Centres Without Walls to Virtual Communities of Practice,* held in conjunction with the *IADIS WWW/Internet 2003 Conference,* 5-8 November, 2003, Carvoeiro, Portugal. Updated version to be published 2004 in: *Journal of Digital Management,* special issue on Web-Based Collaboratories.
- Keiper, Jürgen; Brocks, Holger; Dirsch-Weigand, Andrea; Stein, Adelheit & Thiel, Ulrich (2001). **COLLATE – A Web-Based Collaboratory for Content-Based Access to and Work with Digitized Cultural Material.** In: Bearman, D. & Garzotti, F. (Eds.), *Proceedings of the International Cultural Heritage Informatics Meeting (ICHIM '01), Milano, Italy, 3-7 September 2001.* Milano: Politecnico di Milano, 2001, pp. 495-511.
- Stein, Adelheit; Thiel, Ulrich & Keiper, Jürgen (2002). **Going Beyond Traditional Digital Libraries for Cultural Heritage: The COLLATE Collaboratory.** *Cultivate Interactive, Issue 6, 2002,* (www.cultivate-int.org/issue6/collate/)

- Thiel, Ulrich; Brocks, Holger; Dirsch-Weigand, Andrea; Keiper, Jürgen & Stein, Adelheit (2002). **A Collaborative Archive Supporting Research on European Historic Film – The COLLATE Project.** In: *Proceedings of the DLM-Forum 2002. @ccess and preservation of electronic information: best practices and solutions, Barcelona, Spain, 6-8 May 2002.* Luxembourg: Office for Official Publications of the European Communities, 2002, pp. 228-235.
- Thiel, Ulrich; Brocks, Holger; Frommholz, Ingo; Dirsch-Weigand, Andrea; Keiper, Jürgen; Stein, Adelheit & Neuhold, Erich.J. (2003). **COLLATE – A Collaboratory Supporting Research on Historic European Films.** In: *Journal of Digital Libraries Special Issue "Digital Libraries as Experienced by the Editors of the Journal"* (in print).

Conference presentations (in addition to conference presentations of papers cited above)

- Brocks, Holger (August 2002). **COLLATE presentation and demonstration of the COLLATE system** at the exhibition of the *World Library and Information Congress: 68th IFLA General Conference and Council (IFLA 2002), Glasgow.*
- Dirsch-Weigand, Andrea (August 2003). **COLLATE presentation and demonstration of the COLLATE system** at the exhibition of the *World Library and Information Congress: 69th IFLA General Conference and Council (IFLA 2003), Berlin.*
- Dirsch-Weigand, Andrea (October 2002). **Entwicklung und Integration von Ontologien zur Ausarbeitung von historischen Dokumenten. Wissensrepräsentation für ein digitales Archiv der Filmwissenschaften.** Presentation invited for the *ISI 2002. Internationales Symposium für Informationswissenschaften, Regensburg, Germany, 7-10 October 2002.*
- Stein, Adelheit (Sept 2001). **COLLATE – A Web-Based Collaboratory for Content-Based Access to and Work with Digitized Cultural Material.** Paper presentation at the *International Cultural Heritage Informatics Meeting (ICHIM '01), Milano, Italy, September 3-7, 2001.*
- Stein, Adelheit; Keiper, Jürgen & Thiel, Ulrich (Sept 2002). **Web-based Collaboration Support in a Digital Library on European Historic Films - the COLLATE Project.** Paper presentation at the *7th Conference on Digital Resources in the Humanities (DRH2002), Edinburgh, Scotland, 8-11 September 2002.*

University of Bari

Publications

In the following the set of publications produced by Uniba is listed and ordered by argument.

The transformation of scanned paper documents to a form suitable for an Internet browser is a complex process which is supported by the system WISDOM++. WISDOM++ is a document processing system that operates in five steps: document analysis, document classification, document understanding, text recognition with an OCR, and text transformation into HTML/XML format. WISDOM++ makes extensive use of machine learning techniques in order to achieve a high degree of adaptivity which is necessary to face the complexity of the tasks.

- Altamura, Oronzo; Esposito, Floriana & Malerba, Donato (2001). **Transforming Paper Documents into XML Format with WISDOM++.** *International Journal of Document Analysis and Recognition*, 2001, 4(1): 2-17.
- Ferilli, Stefano (2001). **Management of Cultural Heritage Material: The COLLATE project.** In: L. Bordonì, G. Semeraro (Eds.), *Proceedings of the Workshop on Artificial Intelligence for Cultural Heritage and Digital*

*Libraries in the 7th Congress of the Italian Association for Artificial Intelligence (AI*IA-2001)*, Bari, 25 September 2001, pp. 29-33.

- Malerba, Donato; Ceci, Michelangelo & Berardi, Margherita (2003)
XML and Knowledge Technologies for Semantic-Based Indexing of Paper Documents. In: V. Marik, W. Retschintzegger & O. Stepankova (Eds.), *Database and Expert Systems Applications*, (DEXA 2003), 256-265, LNCS 2736, Springer, Berlin.
- Esposito, Floriana (Nov 2001).
Knowledge Acquisition and Discovery in Document Processing: From Paper to Digital Archives. Paper presentation at the *AI*IA workshop on Knowledge Management: Roles and Perspectives of Artificial Intelligence*, Milano, November 2001.

The problem of association rules discovery from document images has been faced. Document images are initially processed to extract both their layout structures and their logical structures. To take into account the inherent spatial nature of the layout structure, a spatial data mining algorithm is applied, which returns spatial association rules.

- Michelangelo Ceci, Margherita Berardi, Donato Malerba (2003) **Mining association rules in document images.** Workshop on Multimedia Discovery and Mining, September 22, 2003. Dubrovnik, Croatia.
- Berardi, Margherita; Ceci, Michelangelo & Malerba, Donato (2003)
Mining spatial association rules from document layout structures. In: *Proc. of the 3rd Workshop on Document Layout Interpretation and its Application (DLIA 2003)*, 9-13, Deutsches Forschungszentrum für Künstliche Intelligenz, GmbH, Germany.

WISDOM++ transforms paper documents into XML format, for further improvements of Wisdom++ are mainly devoted to supporting the retrieval of Web documents. This activity is supported by a new method for the classification of a HTML/XML document into a hierarchy of categories. The hierarchy of categories is involved in all phases of automated document classification, namely feature extraction, learning, and classification of a new document.

- Malerba Donato; Esposito, Floriana & Ceci Michelangelo (2002).
Mining HTML pages to support Document Sharing in a Cooperative System. In R. Unland, A. Chaudri, D. Chabane & W. Lindner (Eds.) *XML-Based Data Management and Multimedia Engineering – EDBT 2002*, Lecture Notes in Computer Science, 2490, Berlin: Springer, 2002, pp. 190-201
- Ceci, Michelangelo; Malerba, Donato; Lapi, Michele & Esposito, Floriana (2003).
Automated Classification of Web Documents into a Hierarchy of Categories. In: Ö M.A. Klopotek, S.T. Wierzchon, K. Trojanowski (Eds.), *Intelligent Information Processing and Web Mining*, Series: Advances in Soft Computing. Berlin: Springer, 2003, pp. 59-68.

Layout analysis is the process of extracting a hierarchical structure describing the layout of a page. In the system WISDOM++, the layout analysis is performed in two steps: firstly, the global analysis determines possible areas containing paragraphs, sections, columns, figures and tables, and secondly, the local analysis groups together blocks that possibly fall within the same area. We investigated the possibility of supporting the user during the correction of the results of the global analysis. This is done by means of the application of ILP techniques.

- Malerba, Donato; Esposito, Floriana & Altamura, Oronzo (2001).
Learning Rules for Layout Analysis Correction. In: *Workshop on Document Layout Interpretation and its Applications (DLIA 2001)*, Seattle, 9 September 2001.
- Altamura, Oronzo; Esposito, Floriana & Malerba, Donato (2001).
Learning to Correct the Layout Extracted from Document Images. In: Chella, A. & Malerba, D. (Eds.), *Proceedings of the Workshop on Artificial Intelligence, Vision and Pattern Recognition in the 7th Congress of the Italian Association for Artificial Intelligence (AI*IA '01)*, Bari, 24 September 2001, pp. 63-73.
- Malerba, Donato; Esposito, Floriana & Altamura, Oronzo (2002).
Adaptive Layout Analysis of Document Images. In: Hacid, M.-S., Ras, Z.W., Zighed, D.A. & Kodratoff Y. (Eds.), *Foundations of Intelligent Systems*. Lecture Notes in Artificial Intelligence 2366. Berlin: Springer, 2002, pp. 526-534.

- Berardi, Margherita; Ceci, Michelangelo; Esposito, Floriana & Malerba, Donato (2003)
Learning Logic Programs for Layout Analysis Correction. In: *Proc. of the 20th International Conference on Machine Learning (ICML 2003)*, 27-34, AAAI Press, Menlo Park, California, (USA).
- Malerba, Donato; Esposito, Floriana; Altamura, Oronzo; Ceci, Michelangelo & Berardi, Margherita (2003)
Correcting the Document Layout: A Machine Learning Approach. In: *Proc. of the 7th International Conference on Document Analysis and Recognition (ICDAR 2003)*, 97-102, IEEE Computer Society Press, Los Alamitos, California, USA.

The applicability and effectiveness of incremental symbolic learning techniques to the task of document image understanding (i.e., the induction of classification rules for the different document types, and of interpretation rules for the significant components therein, based only on the layout appearance of the documents themselves) was deeply investigated, along with the use of multiple reasoning strategies to face the peculiar complexity of historical material, obtaining confirmation of their profitable application.

- Esposito, Floriana; Ferilli, Stefano; Fanizzi, Nicola; Basile Teresa & Di Mauro Nicola (2003).
Incremental Multistrategy Learning for Document Processing. *Applied Artificial Intelligence Journal*, 17 (8/9): 859-883, Taylor & Francis, London, 2003.
- Esposito, Floriana; Ferilli, Stefano; Fanizzi, Nicola; Basile, Teresa & Di Mauro, Nicola (2002).
Cooperation of Multiple Strategies for Automated Learning in Complex Environments. In: M.-S. Hacid, Z.W. Ras, D.A. Zighed and Y. Kodratoff (Eds.), *Foundations of Intelligent Systems. Lecture Notes in Artificial Intelligence* 2366. Berlin: Springer, 2002, pp. 574-582.
- Ferilli, Stefano; Fanizzi, Nicola & Semeraro, Giovanni (2001).
Learning Logic Models for Automated Text Categorization. In: Esposito, F. (Ed.), *AI*IA 2001: Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence* 2175, Springer: Berlin, 2001, pp. 81-86.
- Ferilli, Stefano; Fanizzi, Nicola; Basile, Teresa & Di Mauro, Nicola (2002).
Learning Family Relationships Exploiting Multistrategy. In: *Proceedings of the AI*IA Workshop su Apprendimento Automatico: Metodi e Applicazioni - Eight Convegno of the Italian Association for Artificial Intelligence (AI*IA-2002)*, Siena, Italy, 11 September 2002, pp. 71-81.
- Ferilli, Stefano; Esposito, Floriana; Di Mauro, Nicola & Basile, Teresa M.A. (2003).
Automatic Induction of Rules for Classification and Interpretation of Cultural Heritage Material. In T. Koch and I.T. Sølvberg (Eds.), *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science* 2769, 152-163, Springer:Berlin, 2003.
- Ferilli, Stefano; Basile, Teresa M.A.; Di Mauro, Nicola & Esposito, Floriana (2003).
Incremental Induction of Rules for Document Image Understanding. In A. Cappelli and F. Turini (Eds.), *AI*IA 2003: Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence* 2829, 176-188, Springer:Berlin, 2003.
- Ferilli, Stefano; Iannone, Luigi; Palmisano, Ignazio; Semeraro, Giovanni; Basile, Teresa M.A. & Di Mauro, Nicola
Automatic Annotation of Historical Paper Documents *Proceedings of the Workshop on Artificial Intelligence for Cultural Heritage in the 8th Conference of the Italian Association for Artificial Intelligence (AI*IA-2003)*, Pisa, 23 September 2003, 99-103.
- Semeraro, Giovanni; Esposito, Floriana; Ferilli, Stefano; Fanizzi, Nicola; Basile, Teresa M.A.; Di Mauro, Nicola (2002).
Multistrategy Learning of Rules for Automated Classification of Cultural Heritage Material. In: E. Lim, S. Foo, C. Khoo, H. Chen, E. Fox, S. Urs & C. Thanos (Eds.), *Digital Libraries: People, Knowledge, and Technology. Lecture Notes in Computer Science* 2555, Berlin: Springer, 2002, pp. 182-193.
- Semeraro, Giovanni; Ferilli, Stefano; Fanizzi, Nicola & Floriana Esposito (2001).
Document Classification and Interpretation through the Inference of Logic-Based Models. In: *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001)*. Berlin: Springer, 2001, pp. 59-70.
- Malerba, Donato; Esposito, Floriana; Lisi, Francesca A. & Altamura, Oronzo (2001).
Automated Discovery of Dependencies Between Logical Components in Document Image

Understanding. In: *Proceedings of the Sixth International Conference on Document Analysis and Recognition, Seattle, 10-13 September 2001*, pp. 174-178.

The need of speeding up the learning systems performance on particularly complex documents, in order to make it acceptable where standard techniques failed, a new procedure for computing the matching of first-order clauses under the theta-subsumption generalization model was developed.

- Ferilli, Stefano; Di Mauro, Nicola; Basile, Teresa M.A. & Esposito, Floriana (2003).
q-subsumption and Resolution: A New Algorithm. *14th International Symposium on Methodologies for Intelligent Systems*, Lecture Notes in Artificial Intelligence 2871. Berlin: Springer, 2003, pp. 384-391.
- Di Mauro, Nicola; Basile, Teresa M.A.; Ferilli, Stefano; Esposito, Floriana & Fanizzi, Nicola (2003):
An Exhaustive Matching Procedure for the Improvement of Learning Efficiency In T. Horváth and A. Yamamoto (Eds.), *Inductive Logic Programming – 13th International Conference (ILP-2003) Proceedings*, Lecture Notes in Artificial Intelligence 2835. Berlin: Springer, 2003, pp. 112-129.
- Ferilli, Stefano; Fanizzi, Nicola; Basile, Teresa & Di Mauro, Nicola (2002).
Efficient Theta-subsumption under Object Identity. In: *Proceedings of the AI*IA Workshop su Apprendimento Automatico: Metodi e Applicazioni - Eight Convegno of the Italian Association for Artificial Intelligence (AI*IA-2002), Siena, Italy, 11 September 2002*, 59-69
- Ferilli, Stefano; Di Mauro, Nicola; Basile, Teresa M.A. & Esposito, Floriana (2003).
A Complete Subsumption Algorithm. In A. Cappelli and F. Turini (Eds.), *AI*IA 2003: Advances in Artificial Intelligence*, Lecture Notes in Artificial Intelligence 2829, 1-13, Springer:Berlin, 2003.

In sum, the research carried out at Uniba succeeded in developing strategies and techniques to improve both the effectiveness and the efficiency of automatic document processing by means of machine learning techniques. This is particularly true if the peculiar difficulties associated with the management of historical documents are taken into account. The results are highly valuable not only in the specific context of Cultural Heritage, but most of them represented significant advances in the Artificial Intelligence research in general.

COLLATE Archives

The intended purpose of the film studies is to show to a wide range of professional users the different possibilities that are inherent in such a collaboratory. For this reason we placed great importance in a comprehensive and meaningful Web presentation of the case studies' results. On the site <http://www.deutsches-filminstitut.de/collate/index.html> the archives DIF, FAA and NFA present the results of their wide experience of film studies using COLLATE. The efforts of the case studies in COLLATE – 35 essays and 300 links to historical documents – have thus been made available to the public. This outcome arises not only from a close collaboration between the DIF, NFA and FAA but also from lively interchanges with other colleagues in several countries.

DIF – Deutsches Filminstitut

Publications

See also the six joint papers with IPSI that are listed above.

- Bezerra, Laura & Keiper, Jürgen (2003).
Indexing and Annotation: Collaborative in-Depth Analysis of Film-related Material. In: Loiperdinger, M. (Ed.), *Celluloid Goes Digital, Historical-Critical Editions of Films on DVD. Proceedings of the First International Trier Conference on Film and New Media*. Trier: WVT Wissenschaftlicher Verlag Trier, 2003, pp. 67-73 plus figures on CD-ROM.
- Keiper, Jürgen (2001).
COLLATE. In: Klemp, K. & Ribbe, A. (Eds.), *Die Frage nach der Frage. Zukunft durch*

Wissenschaft, Zukunft der Wissenschaft. Themenfelder und Probleme zukünftiger Forschungen. Vorträge der Veranstaltung vom 1. - 3. März 2001 in der Deutschen Bibliothek. Frankfurt am Main: Waldemar Kramer, 2001, pp. 120-126.

- Kopf, Christine (2001).
Der Schein der Neutralität - Institutionelle Filmzensur in der Weimarer Republik.
<http://www.deutsches-filminstitut.de/news/dt2n13.htm>

Conference presentations (in addition to conference presentations of papers cited above)

- Bezerra, Laura & Keiper, Jürgen (November 2002).
COLLATE - Historische Filmforschung mit einem verteilten Annotationssystem im WWW.
Invited presentation at "AUDIOVISUELLE WISSENSMEDIEN ONLINE" in the section *Metadata. National Conference, Institut für den wissenschaftlichen Film – IWF Göttingen, Germany, 3-4 December 2002.*
- Bezerra, Laura & Keiper, Jürgen (May 2003).
Using the Web to Work Together: COLLATE – Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material. Invited talk at the *International Conference on Cinema Archives: The Memory of Cinema. An Exchange between Archivists, Librarians and Conservators (Section: Conservation and treatment of films and "non-film" material in cinema libraries and museums).* Associazione Nazionale Archivistica Italiana (ANAI), Italy, 28-31 May 2003.

FAA – Filmarchiv Austria

Publications

- Ballhausen, Thomas (2003).
Verboten! Eine kurze, aber wahre Geschichte der Filmzensur in Österreich bis 1938. In: *peng. zeitschrift für film kunst kultur* #5, pp. 16-23.
- Ballhausen, Thomas (2002).
Notizen zur Geschichte der Wiener Filmzensur 1918 – 1938. In: *Biblos. Beiträge zu Buch, Bibliothek und Schrift* 51, 2. edited by the Österreichische Nationalbibliothek. Phoibos: Vienna, 2002, pp. 203-214.
- Caneppele, Paolo (2001).
Beschnittene Schaulust. Entstehung und Entwicklung der Filmzensur in Österreich. Ein Abriß. (1900-1938). *Medien und Zeit*, 2001, 16 (2): pp. 22-34.

In the following books about censorship history the COLLATE project is described shortly:

- Materialien zur österreichischen Filmgeschichte 3 - Entscheidungen der Tiroler Filmzensur - 1919-1921 mit einem Index der in Tirol verbotenen Filme 1916-1922; ISBN 3-901932-11-9
- Materialien zur österreichischen Filmgeschichte 4 - Entscheidungen der Tiroler Filmzensur – 1922-1938; ISBN 3-901932-12-7
- Materialien zur österreichischen Filmgeschichte 8 - Entscheidungen der Wiener Filmzensur – 1922-1925; ISBN 3-901932-14-3
- Materialien zur österreichischen Filmgeschichte 9 - Entscheidungen der Wiener Filmzensur – 1926-1928; ISBN 3-901932-20-8

Conference presentations

- Moser, Karin & Streit, Elisabeth (10 March 2001).
Presentation of COLLATE at the conference of "SYNEMA" at the University of Vienna, Altes AKH, 9-11 March 2001 in Vienna.

NFA – Národní Filmový Archiv

Publications

- Lachman, Tomáš (2002).
COLLATE. Nová dimenze filmové historického výzkumu. (COLLATE. A new Dimension of the Filmhistoric Research.) In: *Illuminace*, 2002, 14 (2): pp: 141-150.
- Lachman, Tomáš & Uzlová, Eva (2002).
COLLATE - vytváření nového informačního zdroje v oblasti kinematografie. *Ikaros* [online]. 2002, No. 12, <http://www.ikaros.cz/Clanek.asp?ID=200212005>, ISSN 1212-5075.

Conference presentations

- Lachman, Tomáš (September 2003).
Surrogates of missing films. Presentation of COLLATE & the results of WP 8.3 "Production of surrogates for lost films" on ARCHIMEDIA 2003, Prague, 25. September 2003.
- Uzlová, Eva & Lachman, Tomáš (November 2002).
Presentation of COLLATE at seminary "*Informační zdroje v humanitních a společenských vědách*", Praha. (Information sources on social science, Prague, 18. November 2002).

Risø National Laboratory

Publications

- Cleal, Bryan; Andersen, Hans K.H. & Albrechtsen, Hanne (2004).
Collaboration, Communication and Categorical Complexity: a Case Study in Collaboratory Evaluation. In: *Journal of Digital Management*, special issue on Web-Based Collaboratories (in press).
- Andersen, Hans K.H.; Albrechtsen, Hanne & Cleal, Bryan (2003).
Structuring Collaborative Research: Experiences from an Evaluation Study of a Collaboratory. In: *Proceedings. 2. Danish human-computer interaction research symposium*, Roskilde (DK), 27 Nov 2003. Hertzum, M.; Heilesen, S. (eds.), (Roskilde Universitetscenter, Datalogisk afdeling, Roskilde, 2003) (Datalogiske Skrifter, 98) p. 13-16
- Albrechtsen, Hanne & Pejtersen, Anneliese M. (2004).
Cognitive Work Analysis and Work Centered Design of Classification Schemes. In: B. Hjørland (Ed.), *International Journal of Knowledge Organisation*, special issue on Domain Analysis and Knowledge Organization (in press)
- Albrechtsen, Hanne; Andersen, Hans K.H.; Cleal, Bryan & Pejtersen, Annelise M. (2003).
Categorical Complexity in Knowledge Integration: Empirical Evaluation of a Cross-Cultural Film Research Collaboratory. Accepted for the *International Conference on Knowledge Organization and the Global Information Society (ISKO8)*, to be held 13-16 July, London, UK.
- Albrechtsen, Hanne; Pejtersen, Annelise M. & Cleal, Bryan (2002).
Empirical Work Analysis of Collaborative Film Indexing. In: Bruce, H.; Fidel, R.; Ingwersen, P. & Vakkari, P. (Eds.), *Emerging Frameworks and Methods. Proceedings of the fourth International Conference on Conceptions of Library and Information Science (CoLIS4)*. Greenwood Village: Libraries Unlimited, pp. 85-108.
- Hertzum, Morten; Pejtersen, Annelise M.; Cleal, Bryan & Albrechtsen, Hanne (2002).
An Analysis of Collaboration in Three Film Archives. A Case for Collaboratories. In: Bruce, H.; Fidel, R.; Ingwersen, P. & Vakkari, P. (Eds.), *Emerging Frameworks and Methods. Proceedings of the fourth International Conference on Conceptions of Library and Information Science (CoLIS4)*. Greenwood Village: Libraries Unlimited, pp. 69-84.
- Pejtersen, Annelise M. & Albrechtsen, Hanne (2002).
Models for Collaborative Integration of Knowledge. In: Lopez-Huertas, M. (Ed.), *Challenges*

in Knowledge Representation and Organization for the 21st Century: Integration of Knowledge across Boundaries. Würzburg: Ergon Verlag, 2002, pp. 412-421.

Conference presentations (in addition to conference presentations of papers cited above)

- Cleal, Bryan (8 November 2003).
Presentation of the **empirical evaluation of the COLLATE prototypes** at the *first international research workshop on web-based collaboratories (Wbc1)*, held at the IADIS www/Internet Conference, Algarve, Portugal.
- Albrechtsen, Hanne (8 November 2003).
Presentation of **indexing and annotation in COLLATE** at the *first international research workshop on web-based collaboratories (Wbc1)*, held at the IADIS www/Internet Conference, Algarve, Portugal.
- Albrechtsen, Hanne (30 October 2003).
Presentation of the **COLLATE project** at *Department of Information Studies, Royal School of Library and Information Science, Copenhagen, Denmark*.
- Albrechtsen, Hanne (7 October 2003).
Presentation of the **empirical evaluation of the COLLATE prototypes** at the Research Symposium part of the International Workshop on Innovations in Digital Asset Management - Concepts, Tools, Solutions (6-8 October 2003), held at IPSI, Darmstadt, Germany, organised by the COLLATE consortium.
- Pejtersen, Annelise M. (22 June 2003).
Tutorial on Cognitive Systems Engineering and Evaluation of Collaboratories, held in connection with *HCI International 2003, 22-27 June 2003, Crete, Greece*.
- Albrechtsen, Hanne (November 2002).
Classification schemes and common workspaces in film research collaboratories - COLLATE. Presentation at *research workshop on Human-Computer Interaction held at Risoe National Laboratory, 19 November 2002*.
- Albrechtsen, Hanne (July 2002).
Conceptual perspectives in knowledge organization. The case of collaborative ordering of cultural heritage. Presentation invited for the *Workshop on Philosophical, Historical, Rhetorical, and other Conceptual Approaches to Library and Information Studies, University of Washington, Seattle, USA, 26 July 2002*.
- Albrechtsen, Hanne (February 2003).
Work-based Classification Schemes - Designing Common Workspaces for Collaborative Film Indexing. Presentation invited for the *International CHMI Workshop on Human-Machine Interaction, Risø National Laboratory, Roskilde, Denmark, 12 February 2003*.
- Pejtersen, Annelise M. (November 2002).
Methods for empirical evaluation of film research collaboratories - COLLATE. Presentation at *research workshop on Human-Computer Interaction held at Risoe National Laboratory, 19 November 2002*.
- Pejtersen, Annelise M. (July 2002).
Applying the framework for cognitive work analysis for empirical evaluation of research collaboratories. Presentation invited for the *fourth International Conference on Conceptions of Library and Information Science (CoLIS4), Seattle, USA, 21-25 July 2002*.

Sword Information and Communication Technology S.r.l.

Sword's contribution to dissemination activities mainly consists of a variety of presentations and meetings. For publications see also the joint paper with IPSI and Uniba (Frommholz et al. 2003) listed above.

Conference presentations

- IFTM, *International Forum on Technology Management*, 12-16 Nov., 2001 in Bangalore, India
- ECKM, *2nd European Conference on Knowledge Management*, Bled School of Management, Slovenia, Nov. 2001
- KM Europe 2001, *The 2nd Knowledge Management Europe Exhibition & Conference*, NCC, Den Haag, Netherlands, Nov. 2001
- Valente, Antonio & Golia, Pierpaolo: XML Content Management – a middleware for comfortable user access for distributed digital repositories. Presentation at International Workshop on “*Innovations in Digital Asset Management*”, Fraunhofer IPSI, Darmstadt, Germany, 6-8 October 2003

As a commercial technology developer in COLLATE, the major interest of Sword is to promote and reuse the technologies, such as software, interfaces, etc., from the COLLATE system for future projects and products. For this reason, several meetings and prototype presentations for pre-commercial business activities were organized.

10 List of Figures

Figure 1	Hierarchy of COLLATE goals	6
Figure 2	User requirements and tasks	8
Figure 3	COLLATE system architecture.....	10
Figure 4	Interface / input form of DIGIPROT.....	15
Figure 5	Interface / input form of the Mini-Filmography.....	16
Figure 6	Organization of digitization data	17
Figure 7	COLLATE's document description puzzle	19
Figure 8	Procedures and tools for document description in the COLLATE system	20
Figure 9	Conceptual integration of COLLATE indexing vocabularies	22
Figure 10	Examples of documents to be processed.....	27
Figure 11	Automated document understanding: result of matching process against learned rules	28
Figure 12	WISDOM++ architecture	30
Figure 13	Eakin's three level model of image retrieval systems	32
Figure 14	COLLATE image & video indexing approach.....	34
Figure 15	Screenshot of the picture classification tool with topic keywords	37
Figure 16	COLLATE screenshot of picture retrieval result (automatically indexed image).....	39
Figure 17	Model of the interconnections of a generic digital library	41
Figure 18	Intellectual property rights graph.....	43
Figure 19	Relationship servers – clients.....	44
Figure 20	Embedding scheme	46
Figure 21	COLLATE task model	50
Figure 22	Discourse Structure Relations (DSR).....	51
Figure 23	The COLLATE annotation interfaces	54
Figure 24	The MACIS framework.....	55
Figure 25	Cataloguing, indexing and annotation interfaces in COLLATE.....	57
Figure 26	Annotation of a German letter: Local ban of "Battleship Potemkin".....	58
Figure 27	Work phases in the COLLATE project	59
Figure 28	Source Edition on the Web: censorship and genre (case study KING KONG).....	62