COLLATE – A Web-Based Collaboratory for Content-Based Access to and Work with Digitized Cultural Material

Jürgen Keiper [*], *Holger Brocks* [#], *Andrea Dirsch-Weigand* [#], *Adelheit Stein* [#] *and Ulrich Thiel* [#]

[*] Deutsches Filminstitut – DIF, Frankfurt am Main, Germany
E-mail: keiper@deutsches-filminstitut.de

[#] GMD-IPSI, German National Research Center for Information Technology, Integrated Publication and Information Systems Institute, Darmstadt, Germany
E-mail: {brocks, dirsch, stein, thiel}@darmstadt.gmd.de

ABSTRACT

Many important historic and cultural sources are fragile and scattered in various archives, so that their full knowledge and usage are severely impeded. Project COLLATE develops a distributed Web-based repository with dedicated knowledge management facilities to support users in their work with digitized cultural material. As example domain COLLATE uses historical documents referring to films of the 20ies and 30ies. The repository set up by three major European film archives comprises a large corpus of digitized film censorship documents, related articles, photos, posters and film fragments. In-depth analyses of such documents give, for example, evidence about different film versions and cuts, which can be used for reconstruction of lost or damaged films or identification of persons (actors) and film fragments of unknown origin. As a virtual knowledge and working environment for distributed user groups, COLLATE 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 derived user knowledge into the system. This growing body of metadata is exploited by the system using intelligent document processing and advanced XML-based content management and retrieval functionalities. This paper describes the conceptual approach of COLLATE, focusing on the tasks and requirements of the archives and users of the system. The functional system components are outlined and assessed, describing how various user types and related complex tasks can be supported by comfortable task-based user interfaces in a collaborative working environment.

KEYWORDS: collaboratory, digital archive, historical film documentation, content-based indexing, annotation and information access

INTRODUCTION

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 that 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.

Our answer to the problems of the laborious and expendable document access and deficient community organization in the domain of cultural heritage is the European-funded project "COLLATE – Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material" (IST-1999-20882). Technologically, the World Wide Web can serve both as gateway for document-centered digital library applications and as standard communication platform for such professional communities. Therefore, the COLLATE project set out to design and implement a WWW-based collaboratory for archives, researchers and end-users working with digitized historic-cultural material.

In the remainder of this paper we introduce a new and innovative approach to offering access to a comprehensive online digital collection in the domain of historical film documentation. First, we describe the conceptual approach of COLLATE, focusing on the content domain and the tasks and requirements of the archives and users of the system. In the second half of the paper we outline and assess the functional system components, describing how various user types and related complex tasks can be supported by comfortable task-based user interfaces in a collaborative working environment.

WHY A CULTURAL COLLABORATORY?

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). Whereas various collaboratories have been developed since the early 90ies mainly in natural and computer sciences, there exist so far – aside from some pre-studies and systems with very limited functionality – only few comparable efforts in arts and humanities.

As example domain COLLATE uses historic film documentation, employing digitized multi-format documents on several thousands European early 20th century films. Three major European film archives from Germany, Austria and Czechia provide source material. The developed tools and interfaces, however, are designed to be generic, i.e. to be easily adaptable to other content domains, types of applications and user types.

The historic film domain demonstrates the necessity of an international collaboratory in an impressive way. Not only are films sometimes co-produced by several countries, but also they are usually distributed and shown in many countries – in different language versions and sometimes with cuts, e.g. due to censorship restrictions. Copies of the film complemented by secondary material like censorship documents, film reviews, press articles, photos and advertising material are widely distributed in archives. Nonetheless, they represent altogether the cultural heritage concerning the medium film.

Today, such material is stored in national archives and provides only an incoherent understanding of film and cultural history. The complex cultural phenomenon film is disintegrated into a scattered puzzle – inaccessible and unknown. Therefore, archive work aims to reconstruct this "*unity*" and to define an integrated whole of this cultural heritage. This means, for instance, to put together film fragments from various copies in order to obtain a historically correct reconstruction (cf. Schaudig 1988), to use secondary material for intellectual reconstruction of the contents of lost films or cut film scenes, to add information from various sources to a coherent filmography and to compile all textual and visual documents concerning film in order to understand the meaning of this cultural representation.

This unity is the prerequisite for the archivists' work and scientific research. A shared knowledge and information space replaces individual, distributed work and puts together all information into one database. From now on the distribution and reception, which had only poorly been interlinked across nations, is being represented in a system that is accessible worldwide via the Internet.

A collaboratory also provides tools and techniques for deep content indexing of the documents by the film and archive experts. Thus, rich knowledge about the contents can be stored, which allows content-based access and support of complex tasks like, for instance, the preparation of a historic museum's exhibition. Using COLLATE, people from different archives can view photos, discuss a selection of them and enlarge this selection by a detailed new search, e.g., for specific motifs of photos combined with certain visual and esthetic features. Or, a journalist who intends to describe the relationship between violence and media is now enabled to search with exact questions in historical censorship documents.

A great deal of the archivists' work like film reconstructions or text editions rely on collaborations, which so far have mostly been realized by personal contacts, or stepping on information by chance. An online collaboratory allows us to establish this information flow and to access distributed sources with refined search options – and therefore to broaden the archive's work in an invaluable way.

As a virtual knowledge and working environment for distributed user groups, the COLLATE system supports individual work and collaboration of domain experts who are analyzing, evaluating, indexing and annotating the material. It continuously integrates the hereby-derived user knowledge into its digital data and metadata repositories, and on this basis can offer improved content-based retrieval functionalities within the information system. Users are thus enabled to create and share valuable knowledge about the cultural, political and social contexts, which in turn allows other end-users to better retrieve and interpret the historic material.

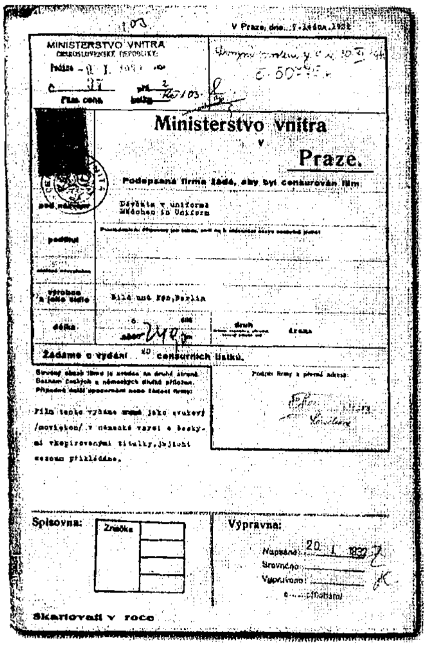
COLLATE PROJECT APPROACH

Developing a "collaboratory in use", we pursue two complementary overall goals:

- Ensuring wide *accessibility* of cultural heritage: Implementation of a content-centric, user-driven information system and working environment on top of a distributed multimedia repository, employing comfortable Web-based tools and interfaces for collaborative work with and content-based access to the digital repository.
- Establishing evidence for the *acceptability* of a collaboratory in the historic domain: Documentation of experiences of the professionals' real-life work with the system, and empirical evaluation of the actual usage of the collaboratory by different user groups, e.g., for "preservation case studies" or other complex scholarly work.

In this approach, technology development and empirical evaluation of the developed system in a real-life environment (the COLLATE archives as pilot users) are closely intertwined. Outputs from both areas of project work strongly influence each other to allow an iterative, dynamic system development. Evaluation steps are explicitly built in, and the users themselves are actively involved throughout the various development cycles – be they administrators, archivists, film scholars or other interested end-users.

The major part of the digital repository set up in COLLATE consists of rare historic *film censorship documents* from the 20ies and 30ies related to several thousands of films (see below some sample documents provided by the Czech National Film Archive in Prague, NFA, and a "forbidden" film advertisement photograph provided by the German Film Institute, DIF).



103.

V Praze, dne 7. ledna 1932

MINISTERSTVO VNITRA
ČESKOSLOVENSKÉ REPUBLIKY
Podáno -8. I. 1932
č. 37 příl. 2
Film. cenz. balík Ko/103.

5-5077

Ministerstvo vnitra
v
Praze.

Podepsaná firma žádá, aby byl censurován film

[redacted]

pod názvem: Děvčata v uniformě
Mädchen in Uniform

podtitul

výrobce a jeho sídlo: Bild und Ton, Berlin

délka: 2440 m

druh: drama

Žádáme o vydání 20 censurních lístků.

Stručný obsah filmu je uveden na druhé straně.
Seznam českých a německých titulků přiložen.
Případné další upozornění nebo žádost firmy:

Film tento vydáme jako zvukový /mluvící/ v německé verzi s českými vkopírovanými titulky, jejichž seznam přikládáme.

Podpis firmy a přesná adresa:

Spisovna: Značka

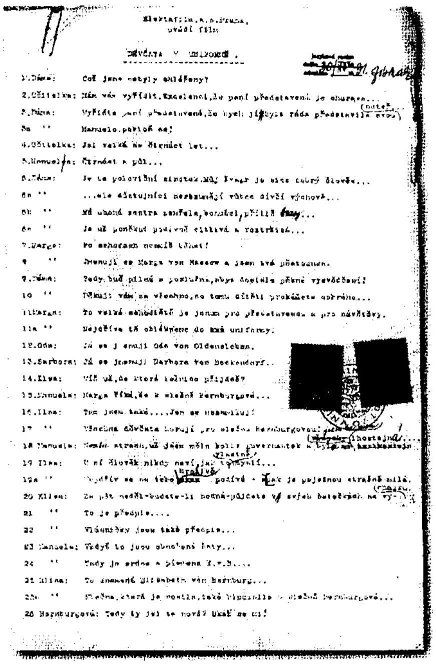
Výpravna:
Napsáno 20. I. 1932
Srovnáno
Vypraveno
s ... přílohami

Skartovati v roce

Figure 1: Application for censorship/permission (cover page)

Most of these documents are to date unavailable in electronic form. They are scattered as paper prints or copies in various archives and are mostly even not yet analyzed and catalogued. Part of this material has been digitized in previous national projects (see, for example, the various projects of the German Film Institute – DIF, in Frankfurt *http://www.filminstitut.de/zensur.htm*).

However, the relevant sources and archive material have so far never been indexed by content or subject matter. They have only been formally catalogued by the respective archives, and even by not all of them. Additional collections provided by the other archives are being digitized, catalogued, indexed and comprehensively annotated during the COLLATE project's runtime.



Elektafilm, a.s. Praha,
uvádí film

DĚVČATA V UNIFORMĚ.

20/IV 31.

```
1.Dáma:       Což jsme nebyly ohlášeny?
2.Učitelka:   Mám vám vyřídit, Excelenci, že paní představená je churava...
3.Dáma:       Vyřiďte paní představené, že bych jí byla ráda představila svou (neteř)
3a  "         Manuelo, pokloň se!
4.Učitelka:   Jak veliká na čtrnáct let...
5.Manuela:    Čtrnáct a půl...
6.Dáma:       Je to polovičí sirotek. Můj švagr je sice dobrý člověk...
6a  "         ...ale důstojníci nerozumějí vůbec dívčí výchově...
6b  "         Má ubohá sestra zemřela, bohužel, příliš brzy...
6c  "         Je už poněkud podivně citlivá a roztržitá...
7.Marga:      Po schodech nesmíš utíkat!
8   "         Jmenuji se Marga von Massow a jsem tvá představená.
9.Dáma:       Tedy buď pilná a poslušná, abys docílila pěkné vysvědčení!
10  "         Děkuji vám za všechno, co tomu dítěti prokážete dobrého...
11.Marga:     To velké schodiště je jenom pro představenou a pro návštěvy.
11a "         Nejdříve tě obléknene do naší uniformy.
12.Oda:       Já se jmenuji Oda von Oldensleben.
13.Barbora:   Já se jmenuji Barbora von Beckendorf...
14.Ilse:      Víš už, do které ložnice přijdeš?
15.Manuela:   Marga říkala, že k slečně Bernburgové...
16.Ilse:      Tam jsem také....Jen se nezamiluj!
17  "         Všechna děvčata horují pro slečnu Bernburgovou!
18 Manuela:   Hloupé stvoření, už jsem měla tolik guvernantek a byly mi lhostejné...
                                                  vlastně
19 Ilse:      U ní člověk nikdy neví, jak to myslí...
                                 hrozivě
19a "         Nejdřív se na tebe        podívá - a pak je najednou strašně milá.
20 Elien:     Za pět neděl - budete-li hodná - půjdete v svých botičkách na vy...
21  "         To je předpis....
22  "         Vlásničky jsou také předpis...
23 Manuela:   Vždyť to jsou obnošené šaty...
24  "         Tady je srdce a písmena E.v.B....
25 Ilse:      To znamená Elisabeth von Bernburg...
25a "         Slečna, která je nosila, také blouznila o slečně Bernburgové...
26 Bernburgová: Tedy ty jsi ta nová? Ukaž se mi!
```

[redacted]

Figure 2: Dialogue list attached to application

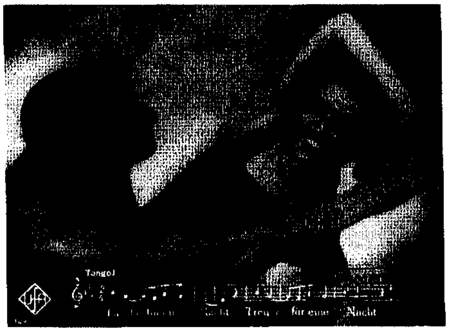


Figure 3: Forbidden photograph (DIF)

The importance of censorship for film production and film distribution lies mainly in the fact that it is often impossible to identify the one unique film. Often, there are a lot of different film versions with cuts, changed endings and new inter-titles (e.g., for silent movies), depending on the place and date of release. Exactly these differences are documented in censorship documents and allow statements about the original film. With respect to political and historical aspects, censorship documents can give valuable information about the development of mass media and the democratic public sphere (cf. Barbian 1993).

For a subset of significant films the data repository additionally offers *enriched documentation* in the form of, e.g., digitized newspaper articles, photos, stills, posters and film fragments. In-depth analyses and comparison of such documents give, for example, evidence about different film versions and cuts, which can be used for reconstruction of lost or damaged films or identification of persons (actors) and film fragments of unknown origin.

COLLATE's idea is to reproduce original text documents as images that can be *marked up* and *annotated*, and to present at the same time their critical revision in the form of previous annotations in a common user interface, i.e., presenting work in progress. All interim findings are conserved and made available as a basis for ongoing analyses. The users have the opportunity to participate in the full process of annotating, revising their annotations and indexing the contents of the digitized documents.

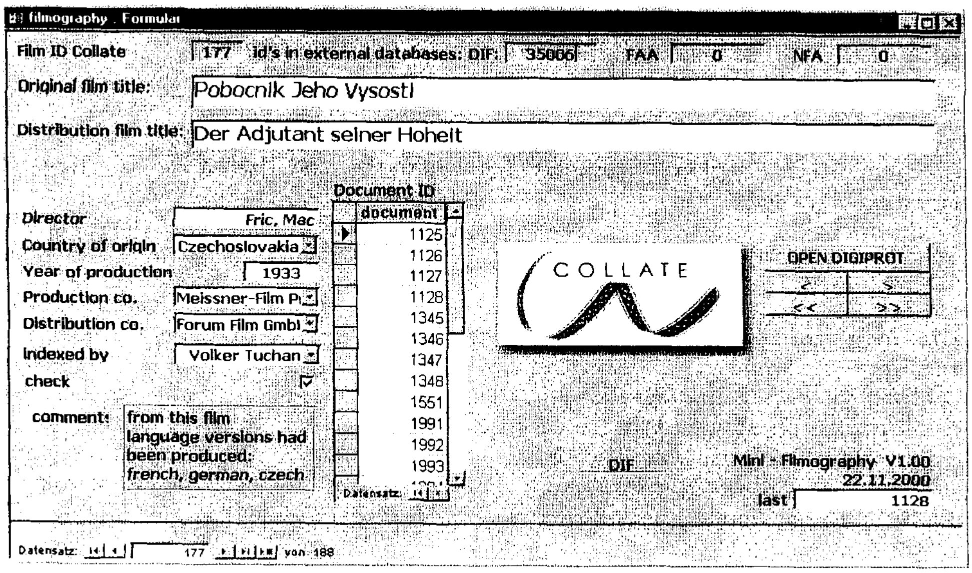
Digital signatures mark the intellectual property rights for entries of individual users, i.e. their annotations of the documents and – recursively – the discursive annotation of annotations. Thus, collaborative in-depth evaluation of documents is supported by the system.

Current technology offers the means to easily reproduce a document as an electronic facsimile in an image format which preserves much more visual information details than a transcribed text, e.g., paleographic characteristics, non-transcribable symbols, margin notes, illustrations, etc. In an XML standard browser with windows/frame technique and with the help of Java and CGI-scripting we provide users the facility to mark single passages in the image (text document, graphic or photo), annotate these marked sections using special forms and menus, and to re-edit the marked document passage or the additional annotation nearly in real time under the same Web interface. In this way, a kind of dynamic document edition can be performed. Using the XML standard format guarantees easy publishing and public access via WWW and a sustainable character of the generic interface.

The digitized documents are indexed at various levels:

First, a *digitization protocol* is recorded in an easy-to-use database (*DIGIPROT*), representing basic technical data of the document (e.g., physical characteristics and scan parameters).

The next step is to assign the document to its designated film title and *filmographic data*, because the film title printed on a document often differs from the original title stored in a filmographic database. *Figure 4* shows the current input form for entering basic data into the "Mini-Filmography" of COLLATE (of course, this can be combined with data from larger external filmographic databases). Note that the same tool is used for entries in *DIGIPROT* – on the right hand side the user may switch between the two input forms.



filmography : Formular

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, Mac
Country of origin: Czechoslovakia
Year of production: 1933
Production co.: Meissner-Film P
Distribution co.: Forum Film Gmbl
Indexed by: Volker Tuchan
check ☑

comment: from this film language versions had been produced: french, german, czech

Document ID

document
1125
1126
1127
1128
1345
1346
1347
1348
1551
1991
1992
1993

Mini - Filmography V1.00
22.11.2000
last 1128

Datensatz: 177 von 188

Figure 4: Input form for the "Mini-filmography" of COLLATE

Subsequently, *formal cataloguing* and manual *content-based indexing* (assigning to the document or parts of the document controlled keywords or free annotations) ensure effective structural retrieval of the documents based on the metadata. This method can then feasibly be combined with automatic indexing and retrieval mechanisms such as probabilistic *full-text retrieval* and *content-based image retrieval* based on automatic document analyses (see, e.g., Stein et al. 1998, Thiel et al. 1999).

Finally, the *annotation work* focuses not only on the document contents – topics or subject matters – but also on (subjective) assessments and comments of the film experts. The last work embodies the heart of the collaboratory: access and retrieval of distributed digital documents, free annotations, and references to other editors by interlinked documents.

Combining results from the manual and automatic indexing procedures, elaborate content-based retrieval mechanisms can be applied. This helps users find what they are actually looking for, to combine evidence from various sources and to interrelate so far unrelated sources and knowledge. Thus, not only the size and richness but also the quality, affordability and acceptability of the information repository can constantly be improved.

THE COLLATE SYSTEM

COLLATE aims to support the proceeding digitization of cultural heritage corpora by innovative models in the following areas: (1) it employs practicable methods of content and knowledge processing for traditionally isolated document collections; (2) it proposes a new concept of content-based organization of impaired and precarious historical material; and (3) it supports the to date only informal cooperative community in arts and humanities by offering a comfortable online working environment to transfer tacit knowledge of professionals into explicitly represented knowledge.

To become a suitable platform for scholarly work in historic/cultural domains, the *support of collaborative work* must go beyond contemporary groupware products, featuring significant innovative functions such as:

- integration of advanced document processing and groupware functions to allow for collaborative inspection and interpretation of source material;
- support of specific tasks in scholarly work such as protection of intellectual property where individuals or groups contribute unpublished parts of their work or assets;
- organization of discussions and typical procedures of scholarly work, such as preparing a source edition or assembling and creating material for an exhibition or publication.

Content-based access to cultural heritage collections must also employ innovative features. To offer an information system with advanced access functions, we definitely need to overcome the current practice of merely providing digital reproductions of and simple online access to historic sources. Instead, results from current and previous scholarly work – such as evaluating and indexing these sources – must be incorporated into the information system, e.g., in the form of metadata and annotations, which in turn allows improved content-based data access.

COLLATE as an interactive knowledge environment enables access to a *distributed data repository.* Its users are directly and indirectly involved in system development because they can actively participate in enriching the document repository through successive annotations and indexation. In-depth analyses are very often done in a team effort, therefore COLLATE implements features supporting such collaborations, e.g., annotation of annotations, collaborative evaluation and comparison of documents. As a result, a large amount of value-added information is provided in addition to the digitized documents. This dynamic accumulation of additional data through annotations in return requires the data structures to be scaleable and extensible.

In order to capture these dynamics we chose *XML* as the de-facto standard for the encoding of generic document and metadata representation schemata. Through the use of XML we are able to guarantee 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 addition, XML is the basis for the integration of knowledge processing methodology and retrieval functionality in the system. Therefore, COLLATE is capable of capturing the dynamics of collaboration without neglecting the necessary flexibility of scaleable and extensible representation schemata, which can be transferred to other content domains as well.

As COLLATE focuses on acceptability

as well as on accessibility, it is essential to facilitate the complex workflows in its domain of film documentation. For this reason we have developed comprehensive models for task-based (semi-automatic generation of) user interfaces, content-based document analysis and annotation, and advanced information retrieval mechanisms.

FUNCTIONAL SYSTEM STRUCTURE

The COLLATE collaboratory is a multi-functional software package integrating a large variety of functionalities, which is realized by cooperating software modules. It comprises several databases and different document representation schemata. XML is used as the uniform internal representation language for the documents in the repository and the associated metadata as well as for the implementation of the communication protocol among its system modules.

There are three main document pre-processing modules:

Digital Watermarking Engine – Through the use of digital watermarking COLLATE ensures all intellectual property rights in a public working environment by incorporating copyright watermarks, integrity watermarks, as well as digital signatures for the users' annotation and edition work. Copyright watermarks are used to tag an image with information about its owner whereas integrity watermarks are employed to guarantee the authenticity and originality of the digitized historical material delivered via the Internet. An example of use would be the production of a CD-ROM containing data from the archives' annotation and edition work. If images of this CD-ROM are found somewhere else, the copyright watermark uniquely identifies their owners – in this case the archives – therefore enabling them to undertake appropriate measures to protect their rights (cf. Dittmann et al. 1999).

Intelligent Document Processing and Classification module – By using machine learning techniques knowledge about digitized documents can be semi-automatically acquired and organized. This is achieved by automatic segmentation, layout analysis and classification of the scanned material. Each page of a historical document is scanned into a binary image. The digital representation is then segmented into rectangular blocks, which contain textual/graphic information or just images. These blocks are labeled according to their type of content, so that subsequent processing stages can benefit from this information. In the layout analysis phase structures among these blocks are detected. This layout structure is used to identify the logical components of the digitized document. According to its logical structure a document is classified into appropriate semantic categories, its identified logical components can be marked for the annotation work or can be further processed by OCR software (for detailed descriptions of the learning mechanisms see Esposito et al. 1994 and Semeraro et al. 2001). Application in COLLATE is particularly challenging, since the repository offers a large variety of manuscripts with non-standardized formats and nowadays unusual typesets and fonts.

Image and Video Analysis module – This module supports conceptual indexing and classification of photos and small film fragments by suggesting options to the human indexers who may modify and enrich them using their background knowledge. Given a new image, the module first automatically extracts pictorial features such as color values, edges, textures from the picture processed as a pixel array. In a second step, it employs a set of rules to relate various thematic concepts, e.g., objects, persons, events and motifs, to corresponding patterns of pictorial features that were extracted from the image. This mapping reduces high-level user concepts, e.g., "person in front of building" to combinations of simpler concepts, e.g., "Object: Building", which can in turn be derived from statistical features of the image (for detailed descriptions of the automatic image classification mechanisms see Thiel et al. 1999, Hollfelder et al. 2000).

The rule set used in this step is derived in a training phase, where the system is fed with characteristic images from a domain, together with corresponding lists of index terms provided by human indexers. Thus, the system is enabled to generate rules which allow to infer appropriate conceptual interpretations of the picture contents. As the training set needs to contain only a small fraction of the potentially huge set of pictures at hand, the efforts spent in preparing the examples and processing them will pay off, because it is then possible to automatically categorize and index pictorial material according to the COLLATE metadata schema for images. Video fragments are handled by extracting appropriate key-frames, which, in turn, can be treated like images.

The COLLATE system can be structured into several *functional layers* (see *Figure 5*; for a detailed presentation of the system architecture see Brocks et al. 2001):

Operational Layer – The Operational Layer can be described as a digital data repository. It comprises a variety of data, ranging from scanned-in text documents to multimedia data and the accumulated annotations related to onc or more of these original data. On the physical level the data can be distributed over several databases, each located in a different computer in the network.

Domain Metadata Layer – In order to organize the stored data in a way that supports the complex knowledge-intensive tasks users perform on the repository contents, suitable tools for metadata management are provided. In this sense the Domain Metadata Layer operates as mediating middleware for accessing the underlying database systems. The knowledge structures, which are represented by specific XML schemata, constitute the Domain Model. They comply with metadata standards such as the TEI (Text Encoding Initiative, *http://www.uic.edu:80/orgs/tei*) and CES (Corpus Encoding Standard, *http://www.cs.vassar.edu/~ide/CES*), but we definitely need extensions to cope with the rich structure of our domain. RDF (Resource Description Framework) can be employed to link the digital documents to their associated metadata. Appropriate indexing aids like controlled vocabularies are then used to ensure the consistent processing of terminological knowledge.

Task Layer – The COLLATE system allows a wide variety of user types to access, work with and evaluate the digitized archive material. It is designed to support complex working tasks in historic film documentation. Therefore, a generic task model has been developed for complex work scenarios like comparative text analyses, source edition, identification of lost or cut film scenes, preparation of a virtual exhibition, etc. Concrete subtasks comprise, e.g., formal cataloguing, keyword indexing, annotating documents or other annotations, interlinking documents or document parts, entering transcriptions and translations, etc.

As the COLLATE focus is on collaboration, groupware functionality is also included, thus allowing for collaborative inspection and interpretation of source material. Specific concurrency control schemata and authentification mechanisms have been developed. When, for example, a user is annotating a document the annotation is stored together with a digital signature, which ensures the authenticity of the contribution.

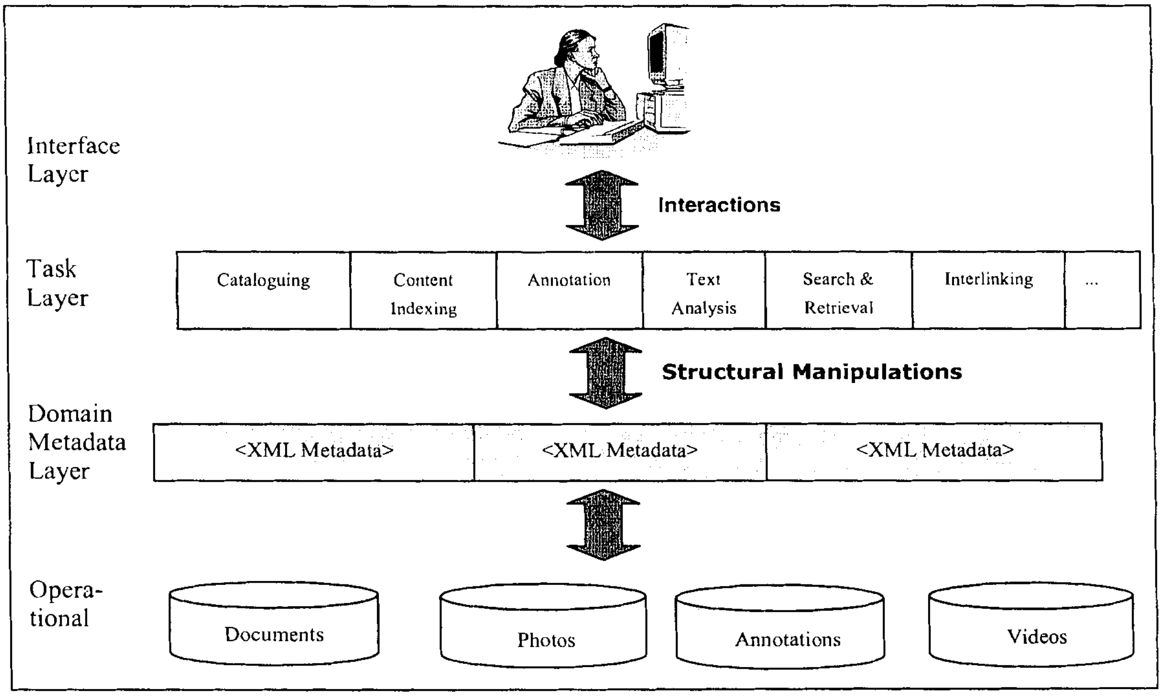


Figure 5: COLLATE System Layers

During the project phase access to the COLLATE system is restricted to specific user groups, i.e. members of the participating archives and other individual collaborators from related institutions. The rationale behind this lies in the intellectual property rights for the digitized documents and in the fact that a sufficiently large body of evaluated and indexed material must be set up to allow for a comfortable retrieval before the system is opened to the public by the end of the project.

Some tasks are mainly performed by specialized user groups, e.g., formal cataloguing by archivists from the digitization departments or content analyses of the censorship documents by film scientists with a special research interest. However, regarding the content-based work there is no generally prescribed division of labor between user groups. In most cases users can decide on their own which tasks and steps they want to perform in which order, according to their knowledge and research interest. Of course, there is a system of hierarchical user permissions: certain users are assigned administrator rights whereas others may have limited access rights. When the COLLATE system will finally be opened to the public, it has to be decided which user groups are permitted to enter further annotations and comments to the system, and how these inputs are to be treated by the content manager.

Interface Layer – In order to support the users in accomplishing their tasks COLLATE provides appropriate interfaces for convenient work with the digital documents. These interfaces are semi-automatically derived from the underlying task model. If tasks have to be revised due to practical experiences, changes in the task model automatically result in appropriate modifications of the corresponding interfaces. Manually recoding of the interface is thus being reduced to a minimum.

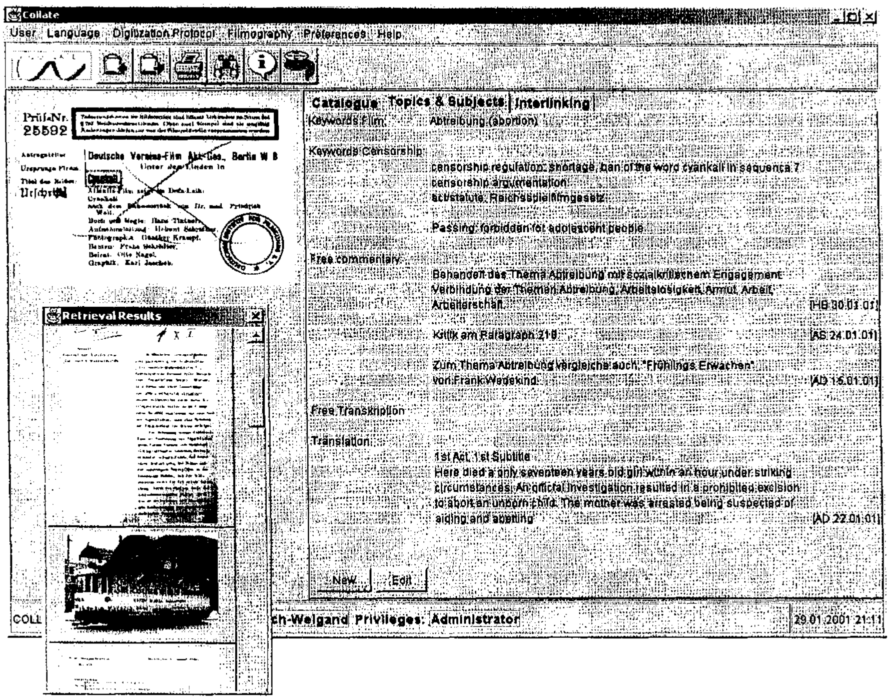
Certain specialized interface components for annotation, mark-up, editing, search and retrieval have been developed to facilitate user interaction. The specification of the interface structure also utilizes XML to allow for a generic mapping to concrete instantiations (e.g., Java Swing).

As indicated above, communication between these layers is realized through XML-based communication protocols. The implementation employs SOAP (Simple Object Access Protocol, *http://www.w3.org/TR/SOAP*) as a basis for the communication infrastructure for the COLLATE system components to ensure maximum flexibility within a shared environment.

USING COLLATE – AN EXAMPLE

The archives participating in the project supply both the document sources and value-added contents to the digital repository while working as pilot users of COLLATE. They also – especially during the first project year – have extensively been involved in the development of the domain-specific classification and indexing schemes, e.g., the document classification and controlled vocabularies for the censorship domain (keywords and indexing aids). The thus developed indexing schemes are put to the test and are being continuously refined during the project's runtime, evaluating the real-life experiences of users during their interaction with the COLLATE system.

Figure 6 shows a screenshot from the first design study of the COLLATE user interface. It depicts some basic characteristics of the content indexing situation and gives a vivid impression of how a working session with the COLLATE system might pass. Subsequent interface versions will be more complex, incorporating additional functionalities such as more comfortable retrieval facilities, structured visualization of additional information (e.g., filmographic data) and document types (e.g., pictorial material) and the respective indexing aids.


Collate
User Language Digitization Protocol Filmography Preferences Help
Retrieval Results
New
Edit
COLL

Prüf-Nr. 25592

Deutsche Vereins-Film Akt-Ges., Berlin W 8
Unter den Linden 1a

Urschrift

Cyankali
nach dem Bühnenwerk von Dr. med. Friedrich Wolf.
Buch und Regie: Hans Tintner.
Photographie: Günther Krampf.
Beirat: Otto Nagel.

Catalogue Topics & Subjects | Interlinking |

Keywords Film: Abtreibung (abortion)

Keywords Censorship:
censorship regulation: shortage, ban of the word cyankali in sequence 7
censorship argumentation
act/statute: Reichsspielfilmgesetz

Passing: forbidden for adolescent people

Free commentary:
Behandelt das Thema Abtreibung mit sozialkritischem Engagement
Verbindung der Themen Abtreibung, Arbeitslosigkeit, Armut, Arbeit, Arbeiterschaft. [HB 30.01.01]

Kritik am Paragraph 218 [AS 24.01.01]

Zum Thema Abtreibung vergleiche auch "Frühlings Erwachen" von Frank Wedekind. [AD 15.01.01]

Free Transkription

Translation:
1st Act, 1st Subtitle
Here died a only seventeen years old girl within an hour under striking circumstances. An official investigation resulted in a prohibited excision to abort an unborn child. The mother was arrested being suspected of aiding and abetting [AD 22.01.01]

ch-Weigand Privileges: Administrator 29.01.2001 21:11

Figure 6: Example Screen of Indexing interface

Following our software engineering philosophy of an evolutionary, cyclic, iterative and user-driven approach, the system and its interfaces evolve over the time. Archivists and system developers are continuously discussing, modeling and re-engineering the functionalities of the system, as well as the look and feel of the graphical user interface, in order to improve its use and usability.

What is more, one of the designated COLLATE project partners – specialized in human-computer interaction evaluation studies – is responsible for empirical system evaluations. Performing field studies and lab experiments for observing users during their usage of the COLLATE system, they will evaluate the respective prototype versions and – on this basis – suggest improvements of the supported functionalities and interaction options. Hence, several refined versions and upgrades of the user interface are subsequently being produced.

Let us assume a sample case, where an archivist or other system user plans to analyze and index a digitized document. He/she starts the working session with the intention to select a specific document, entering a *query*, e.g., searching for all documents in the collection dealing with a specific film title. In our example the search term "Cyancali" was entered, and the system displays all of the available text documents and pictures.

The COLLATE interface provides query forms for *simple searches* (e.g., using attributes like film title, document type and some tree text search field) as well as query forms for *advanced searches* (e.g., for attributes like the subject matter of the document or film, censorship arguments and motifs, the signing authority of a decision, contemporary dates and other significant content matters).

The query results in our example appear in a special window (*Retrieved Results*), which may show the list of document titles retrieved, or thumbnails of the images as shown in the snapshot. Clicking on an item, the user gets displayed the selected document in the upper left part of the annotation window (workspace). Automatically, the user's individual name (digital signature) and the current date are registered by the system, indicated in the status line of the annotation window at the bottom.

The main *menu bar* offers options for registering as indexer, read-only-user, system administrator or content manager (*User*), to choose a certain language version of the interface (*Language*), to consult the digitization data of documents in the collection (*Digitization Protocol*) and to look up certain detailed information about the censored film or photograph in an external filmographic database (*Filmography*). In the *icon bar* below, the user finds pictograms for scrolling, printing, searching, help and access to COLLATE's own filmographic database. Here, the icons invoke document-specific and situation-specific functions or information related to the current dialogue state.

If a document has previously been indexed, the entries show in the main window in the *indexing and annotation working area*, otherwise an empty form is being displayed. The archivists can decide which kind of indexing they are going to perform: (1) formal indexing (*Catalogue*), e.g., entering bibliographic data; (2) content indexing (*Topics & Subjects*) or (3) interlinking the current document or parts of it with related documents. We call this whole process of analyzing, interpreting and recording a specific document "*annotation*".

In our sample case the archivist chooses *Topics & Subjects*, which allows her/him to represent the interpretative content analysis of the document in the form of keywords or free comments. Prerequisite are indexing aids like controlled vocabularies such as structured keyword lists for both film contents (e.g., topics of scenes) and censorship-specific topics (e.g., reasons for banning film scenes, references to authorities, regulations, etc.). Appropriate rules for the indexing must be provided, and other indexing and interpreting aids may be added as necessary, e.g., multilingual indexing terms or phrases and translation aids.

The archivists can annotate the document as a whole or may mark a single *passage* or *word* they want to annotate. The system will automatically assign their following input to this highlighted document detail (on the screenshot, the film title "Cyancali" has been marked by a red rectangle).

The archivist may be the first user to annotate the selected document, or otherwise will find previous annotations of colleagues displayed in the annotating area. These entries are certified by the respective author's name, date and signature as indicated in the right column of the *Topics & Subjects* table. In this case, the current user can append a new comment or keyword input with name and date. New comments may refer to the same selected document part as previous comments, or they may directly refer to another comment, e.g., adding details, giving alternative interpretations or counter arguments. The latter accounts for discursive annotation of annotations through different authors.

By this procedure, it is possible to gather in this collaborative way by and by the knowledge of several experts about a given document or semantically clustered documents in the collection. Step by step, the whole collection of documents is being catalogued and indexed – not in a mechanical way but in a content-related approach. The result is a network of semantically clustered texts that can be browsed and searched by both formal and content-related categories.

CONCLUSIONS

In this paper we have introduced a new type of collaboratory which widens the range of applications for Web-based collaborative platforms by including historic research as an important area from the humanities.

Whereas collaboratories in natural sciences and engineering are usually a means for accessing data, models and software from other sites, in our application the focus is on annotating, comparing, arranging and retrieving historic documents. To that end, the system provides adequate functions for storing, exchanging and annotating scanned-in documents based on contemporary technologies such as XML, OCR, automatic picture analysis, document segmentation and digital watermarking.

The COLLATE system is being tested by a group of film institutes and archives from Germany, Austria and the Czech Republic, members and collaborators of which work on a collection of historic documents referring to films produced in the 1920ies and '30ies.

Performing empirical lab and field experiments for analyzing the experiences made by the film scholars and archivists during their use of the system, we are able to continuously adopt suggestions for extensions and refinements coming from the real-life experience of the experts in their daily work with historic and cultural material.

ACKNOWLEDGEMENTS

This work has partially been funded by the European Commission in the 5th Framework Programme, IST Key Action "Multimedia Contents and Tools", Action Line "Digital Preservation of Cultural Heritage", COLLATE Project No IST-1999-20882.

REFERENCES

1. Barbian, J.-P. Filme mit Lücken: Die Lichtspielzensur in der Weimarer Republik: von der sozialethischen Schutzmaßnahme zum politischen Instrument. In Jung, U., Ed. *Der deutsche Film: Aspekte seiner Geschichte von den Anfängen bis zur Gegenwart. (Filmgeschichte international; Vol. 1)*, Trier 1993, 51-78

2. Brocks, H., Thiel, U., Stein, A., Dirsch-Weigand, A. Customizable Retrieval Functions Based on User Tasks in the Cultural Heritage Domain. In *Proceedings of the ECDL 2001 – 5th European Conference on Research and*

Advanced Technology for Digital Libraries (Darmstadt, Germany), 4-8 September 2001

3. Dittmann, J., Steinmetz, A., Steinmetz, R. Content-based digital signature for motion pictures authentication and content-fragile watermarking. In *Proceedings of IEEE Multimedia Systems, Multimedia Computing and Systems,* June 1999, 574-579

4. Esposito, F., Malerba, D., Semeraro, G. Multistrategy learning for document recognition. In *Applied Artificial Intelligence: An International Journal*, 8 (1), 1994, 33-84

5. Hollfelder, S., Everts, A., Thiel, U. Designing for Semantic Access: A Video Browsing System. In *Multimedia Tools and Applications,* 11 (3), 2000, 281-293

6. Kouzes, R.T., Myers, J.D., Wulf, W.A. Collaboratories: Doing science on the Internet. In *IEEE Computer,* 29 (8), August 1996

7. Schaudig, M. Mit der Zensurkarte auf 'Rattenfang'. Die Ratten (1921): Aspekte der Überlieferung, Edition und Rekonstruktion eines Stummfilm-Fragments. In Ledig, E., Ed. *Der Stummfilm: Konstruktion und Rekonstruktion (Diskurs Film. Münchner Beiträge zur Filmphilologie; Vol 2),* München 1988, 163-207

8. Semeraro, G., Ferilli, S., Fanizzi, N., Esposito, F. Document Classification and Interpretation through the Inference of Logic-based Models. In *Proceedings of the ECDL 2001 – 5th European Conference on Research and Advanced Technology for Digital Libraries* (Darmstadt, Germany), 4-8 September 2001

9. Stein, A., Gulla, J.A., Müller, A., Thiel, U. Abductive dialogue planning for concept-based multimedia information retrieval. In Fankhauser, P. and Ockenfeld, M., Eds. *Integrated Publication and Information Systems. 10 Years of Research and Development,* Sankt Augustin, GMD – Forschungszentrum Informationstechnik 1998, 129-148

10. Thiel, U. Everts, A. Lutes, B., Stein, A. Can rule-based indexing support concept-based multimedia retrieval in digital libraries? Some experimental results. In Draper, S.W. et al., Eds. In *MIRA 99: Evaluating Interactive Information Retrieval.* Berlin, Springer Verlag (eWiC, electronic Workshops in Computing series), 1999, see *http://www.ewic.org.uk/ewic/workshop/view.cfm/MIRA-99*

11. Twidale, M.B. and Nichols D.M. *A Survey of Applications of CSCW for Digital Libraries. ARIADNE Project on Digital Libraries.* Technical Report CSEG/4/98, Computing Department, Lancaster University, 1998, see http://tina.lancs.ac.uk/computing/research/cseg/projects/ariadne/docs/survey.html

12. Wulf, W. A. *The National Collaboratory - A White Paper.* In "Towards a National Collaboratory", Unpublished report of a workshop held at Rockefeller University, March 1989 (co-chaired by Joshua Lederberg and Keith Uncapher)

ABOUT THE AUTHORS

Jürgen Keiper is the scientific coordinator of the content-providers and film archives of the COLLATE project. He helds a masters degree from the University of Frankfurt, and his special background is in Theater, Media and Film Sciences. Since several years he has worked for the DIF (German Film Institute) in various national research projects. His research interests include: film theory & criticism and social history of film.
E-mail:
keiper@deutsches-filminstitut.de

Holger Brocks works as a computer scientist and full-time researcher at GMD-IPSI in the COLLATE project. He is responsible of the design and implementation of the COLLATE system, supervising student assistants and practicals in their programming work. His research interests are: dynamic generation of task-based user interfaces and intelligent information retrieval dialogues.
E-mail: *brocks@darmstadt.gmd.de*

Dr. Andrea Dirsch-Weigand works as a part-time researcher at GMD-IPSI. She holds a PhD in History, and she is about to finish her second studies of Information Science at the Darmstadt Polytechnic, Dept. of Information Science. Her research interests include: cultural heritage, information and knowledge management, and interactive information systems.
E-mail: *dirsch@darmstadt.gmd.de*

Dr. Adelheit Stein is the head coordinator of the COLLATE project. She has been a senior researcher at GMD-IPSI since several years. Her background is in Sociology and Philosophy, with a special focus on cognition and social interaction. At GMD she was involved in several European IT projects and university teaching. Her current research interests include: collaborative human-computer interaction, dialogue planning, and intelligent, adaptive user interfaces.
E-mail: *stein@darmstadt.gmd.de*

Dr. Ulrich Thiel is responsible for the coordination of the technical tasks of the COLLATE project. He holds both a diploma in Computer Science and a PhD in Information Science. Since several years he has been a senior researcher at GMD-IPSI. He was manager of many EU-funded projects His research interests include: intelligent information retrieval, dialogue planning, conversational, adaptive information systems.
E-mail: *thiel@darmstadt.gmd.de*