COLLATE – A collaboratory supporting research on historic European films

Ulrich Thiel¹, Holger Brocks¹, Ingo Frommholz¹, Andrea Dirsch-Weigand¹, Jürgen Keiper², Adelheit Stein¹, Erich J. Neuhold¹

¹Fraunhofer IPSI, Darmstadt, Germany
e-mail: {thiel, brocks, frommholz, dirsch, stein, neuhold}@ipsi.fraunhofer.de
²Deutsches Filminstitut – DIF, Frankfurt am Main, Germany
e-mail: keiper@deutsches-filminstitut.de

Abstract. In the COLLATE project, we aim to design and implement a Web-based collaboratory for archives, scientists, and end users working with digitized cultural material. Our example domain is the historic film documentation comprising digitized material about European films of the early 20th century. Designed as a content- and context-based knowledge working environment for distributed user groups, the COLLATE system supports both individual work and collaboration of domain experts who are analyzing, evaluating, indexing, and annotating material in the data repository. The system provides appropriate task-based interfaces for indexing and annotating. As a multifunctional means of in-depth analysis, annotations can be made individually but also collaboratively, for example in the form of annotation of annotations. Combining results from manual and automatic indexing procedures, elaborate content- and context-based information retrieval mechanisms can be applied.

Keywords: Cultural heritage – Collaboratories – Digital libraries – Annotations – Context-based retrieval

1 Introduction

Various Web-based collaboratories [8] with advanced digital library functions have been employed since the early 1990s, mainly in the natural sciences, but in the arts and humanities we mostly find only systems with limited functionality. One reason might be that work processes in the social sciences are different from the procedures established in the natural sciences and engineering and require appropriate system designs. However, the process of compiling arguments, counterarguments, examples, definitions, and references to historical source material – which is the prevailing method in the humanities – may derive particular benefit from an electronic environment that improves the capacity and reach of the individual knowledge worker (cf., e.g., [6, 11]).

Another aspect of collaboratories is their ability to enable virtual teams to work together almost as if they were at the same location. Although many (informal) contacts between cultural archives constitute specific professional communities, they so far lack effective technology support for collaborative knowledge working. The World Wide Web can serve both as communication platform for such communities and as gateway for document-centered work in such digital libraries and archives (e.g., [2]).

The EU-funded project COLLATE – Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material¹ – started in the fall of 2000 (IST-1999-20882) and will run for 3 years [3, 7]. An international team of content providers, film domain experts, and technology providers works together to develop a new type of collaboratory in the domain of cultural heritage. The implemented system offers access to a digital repository of historic text archive material documenting film censorship practices for several thousand European films from the 1920s and 1930s. For a subset of significant films it provides enriched context documentation including selected press articles, film advertising material, digitized photos, and some film fragments. Major film archives from Germany, Austria, and the Czech Republic provide the sources and work as pilot users with the COLLATE system.

2 COLLATE’s goals and approach

Designed as a content- and context-based knowledge working environment for distributed user groups, the COLLATE system supports both individual work and collaboration of domain experts who are analyzing, evaluating, indexing, and annotating the material in the data repository. The example application focuses on historic film documentation, but the developed tools are designed to be generic and as such adaptable to other content do-

¹ http://www.collate.de/
mains and application types. This is achieved by model-based modules; exchanging, for instance, the annotation types DTD allows us to support different collaboration structures. Other domain ontologies and document description types can be plugged in as well.

The system provides appropriate task-based interfaces for indexing/annotation and collaborative activities such as preparing a joint multimedia publication or assembling and creating material for a (virtual) exhibition, contributing unpublished parts of the film scientists’ work in the form of extended annotations and commentaries. Appropriate knowledge management tools, e.g., indexing aids and domain-specific controlled vocabularies, have been developed jointly by system developers and film domain experts, thus exploiting the benefits of a participatory design. Using the tools for manual cataloging and indexing, users create a growing body of metadata with a special focus on subject indexing and content-based annotations of documents. Automatic indexing of textual and pictorial parts of a document can be invoked to receive suggestions for index terms from the system. In addition, users can rely on the support from automatic layout analysis for scanned documents, which allows for the annotation of individual segments. Annotations are a central concept in COLLATE. As a multifunctional means of in-depth analysis, annotations can be made individually but also collaboratively, for example in the form of annotation of annotations, collaborative evaluation, and comparison of documents.

The system exploits user-generated metadata and annotations by advanced XML-based content management and retrieval methods. The final version of the online collaboratory will integrate cutting-edge document preprocessing and management facilities, e.g., XML-based document handling and semiautomatic segmentation, categorization, and indexing of digitized text documents and pictorial material. Combining results from the manual and automatic indexing procedures, elaborate content- and context-based information retrieval mechanisms can be applied.

3 COLLATE system features

Collaboration support in the COLLATE working environment makes use of some contemporary groupware products and additional system functions based on an explicit model of collaborative indexing and annotation. Through interrelated free-text annotations users can enter into a (direct or indirect) discourse on the interpretation of documents and document passages, e.g., adding information, interpretations, arguments, etc. Possible relations between annotations can be predefined or inferred by the system in order to represent the discourse structure. Additionally, explicit communication about the interpretation of contents and the interrelation of annotations are supported by a built-in discussion forum and in the final system version by an intelligent dialog/collaboration manager.

The COLLATE collaboratory is a multifunctional software package integrating a large variety of functionalities that are realized by cooperating software modules. It comprises several databases and document representation schemata. XML is used as the uniform internal representation language for documents in the repository and the associated metadata as well as for the implementation of the communication protocol among its system modules. An XML-based content manager is responsible for the integration of knowledge processing methodology and retrieval functionality in the system.

The main modules of the COLLATE system architecture are (see [5] for further details):
- Three document preprocessing modules for digital watermarking of documents (copyright and integrity watermarks), intelligent, automatic document structure analysis and classification, and automatic, concept-based picture indexing and retrieval.
- A distributed multimedia data repository comprising digitized text material, pictorial material such as photos and posters, and digital video fragments.
- Tools for the representation and management of the metadata, the XML-based content manager incorporating an ontology manager and a retrieval engine, which are implemented as SOAP-based Web services.
- A collaborative task manager for complex individual and collaborative tasks such as indexing, annotation, comparison, interlinking, and information retrieval, including tools for online communication and collaborative discourse between the domain experts and other system users.
- The user interface of COLLATE comprises several workspaces for various tasks performed by distributed user groups and user types allowing for different access rights and offered interface functions. The final system version will be generated semiautomatically by exploiting knowledge from the underlying task model and the user-specific dialog history.

4 Collaboration in the COLLATE environment

Digital libraries offer new opportunities for collaboration and communication that were unfeasible in traditional libraries [13]. Our goal is to develop a cultural collaboratory, supporting interpretative work on mostly textual material. Our starting point for collaborative work comprises already existing data in the form of binary image representations of the digitized source documents. Therefore, we do not focus on cooperative data acquisition but rather on collaborative content-based indexing.

In COLLATE, we go beyond the mere replication of traditional domain-specific workflows by providing a comprehensive model of the various COLLATE domain objects and their potential interrelations. Our notion