Int J Digit Libr (2004) 4: 8–12 / Digital Object Identifier (DOI) 10.1007/s00799-003-0069-1

COLLATE – A collaboratory supporting research on historic European films

Ulrich Thiel[1], **Holger Brocks**[1], **Ingo Frommholz**[1], **Andrea Dirsch-Weigand**[1], **Jürgen Keiper**[2], **Adelheit Stein**[1], **Erich J. Neuhold**[1]

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

Published online: 23 July 2004 –

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 Material[1] – 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

of task-guided collaboration includes the recognition of structures as well as relations between different types of annotations. By taking the users' roles, tasks, and goals into account we aim to provide comprehensive support for the various levels of indexing.

The COLLATE system supports asynchronous collaboration in indexing for nontechnical users. In our understanding, the domain objects (scanned documents, metadata) represent the main focus of collaborative work, i.e., collaboration is performed through annotating the digitized artifacts or their associated metadata objects.

The interrelations between the various domain objects can either be unspecified or modeled in a more explicit way by defining specific types of admissible relations. In addition, certain communicative acts on the metalevel (e.g., request for clarification) are part of the COLLATE collaboration model.

In addition, 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 that 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 the value-added information in the 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 will be part of a thematic thread, e.g., some newsgrouplike discussion about a certain topic. Digital signatures are employed to ensure authenticity of the individual contributions as well as the chronological order of the annotations.

But to properly represent scientific discourses, especially in the arts and humanities, annotations have to consist of more than unstructured and uncontrolled text that comments on another domain object (binary image file, XML metadata object). For this reason we have devised a comprehensive 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.

Our document-centered discourse model is loosely based on the theory of discourse structure relations. Even though they have been developed for monologs in the linguistic context of text coherence, we think that discourse structure relations can also be adapted to describe admissible relations between various data and metadata objects in the COLLATE context, especially annotations or comments. In particular, we employ a specific subset of relations ranging from factual to more interpersonal levels, i.e., focusing on certain qualities of the participants of a discourse. Figure 1 illustrates the discourse structure relations as used in the COLLATE system (originating from text linguistics, cf. [9, 10]).

A detailed discussion of the discourse structure relations can be found in [4]. The seamless transition from factual to interpersonal discourse structure relations depicted in Fig. 1 also corresponds to the illocutionary aspects of an annotation, i.e., the specific communicative intention its author had in mind at the time of creation (e.g., from stating factual information to active participation in a discussion thread).

Figure 2 displays a fictious example discourse about the partial ban of the movie *Kuhle Wampe* by Berthold Brecht.

Even though discourse structure relations proved adequate for modeling the interrelations between annotations, it turned out that there were some relevant pragmatic aspects of collaborative indexing work that had not yet been covered. In the next section we describe how discourse structure relations can be complemented

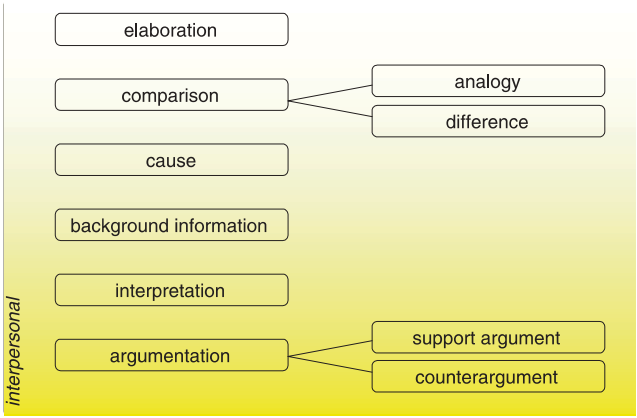

Fig. 1. Discourse structure relations

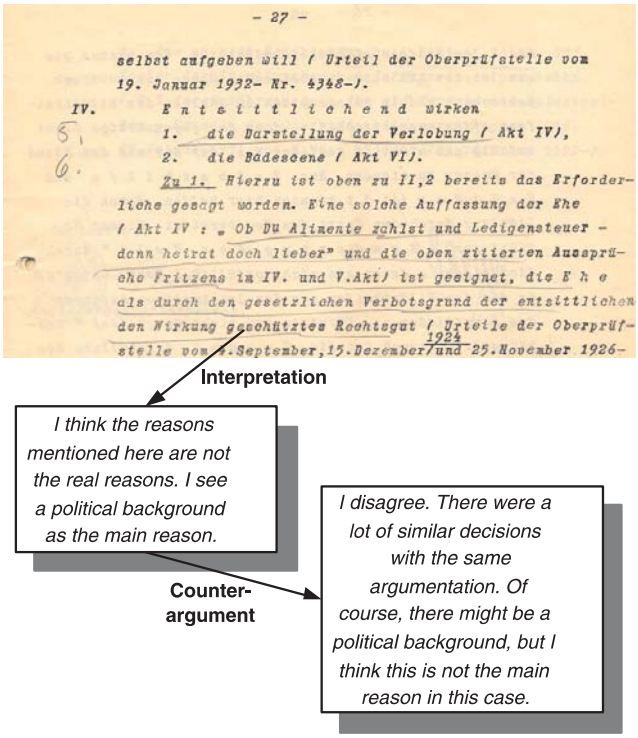

Fig. 2. Example discourse

by communicative acts [12–14] to introduce metacommunication, i.e., explicit communication about domain objects, in a seamless way.

Conceptually, discourse structure relations and communicative acts can be considered as complementary: communicative acts focus on illocutionary aspects of a specific dialog situation (cf., e.g., [1, 14]), whereas discourse structure relations describe characteristic relationships between assertive acts, e.g., annotations or comments.

But on 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 can be considered as specific instances of comments, i.e., they are treated as assertive communicative acts.

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, that are complemented by a certain set of admissible acts for metacommunication (at the dialog level) referring to the various types of COLLATE domain objects (e.g., annotations, cataloging information).

At its current stage of development, the COLLATE prototype system already supports various tasks like cataloging, indexing (structured, free), and nested annotations using discourse structure relations (Fig. 3). Users can select a whole document, a document page, or an area of a document for cataloging, indexing, and annotation. Furthermore, the prototype offers simple search options based on filmographic information as well as on cataloging, keywords, and annotations.

5 Context-based retrieval of documents

Appropriate search and retrieval functionality represents a fundamental requirement for enabling users to access a cultural digital library in a reasonable way. To allow for advanced content- and context-based search, the documents in the digital collection must be indexed by content and subject matter.

In context-based retrieval, not only the document itself, but also the actual context of this document is con-

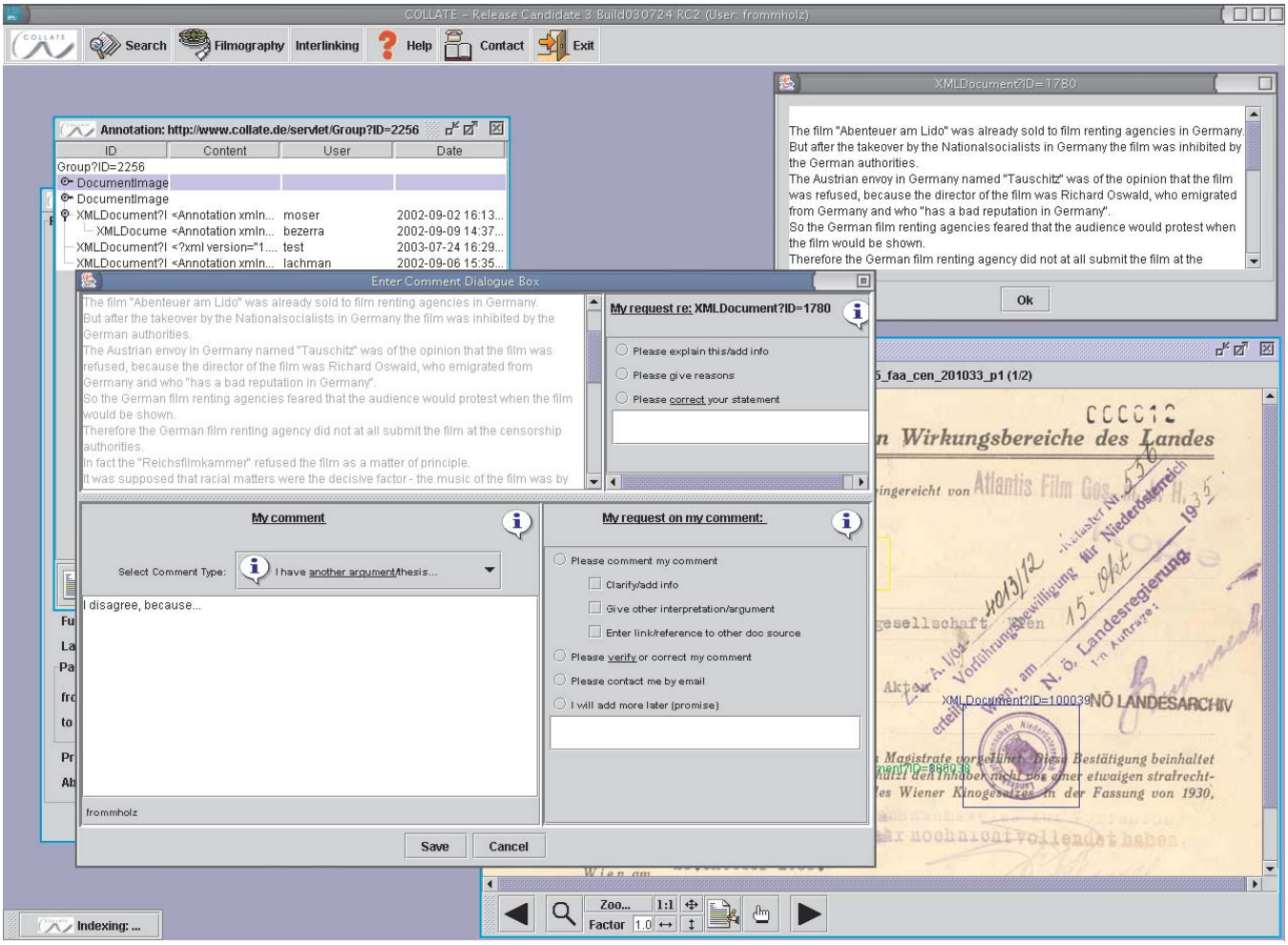

Fig. 3. The COLLATE system

sidered. In the case of COLLATE, this means that we are dealing with the discourse context. Various domain objects (digitized documents, metadata represented in XML, annotations) are given unique URIs and interrelated by RDF statements. These statements are typed according to the discourse structure relation they represent. With this information, we then know the specific type of an annotation with respect to its context, e.g., an elaboration or an example.

Taking a second look at Fig. 3 one can see that only the inclusion of the associated annotations would yield it as relevant for a query like "political background". But it also becomes evident that the annotations within a certain discourse context cannot be treated in isolation, e.g., the second counterargument weakens the statement it comments upon in this context.

The introduction of discourse structure relations allows for novel retrieval options with respect to the discourse context. They can be used to create a ranking of relevant documents according to user queries. Depending on the specific type of its connecting relation, an annotation can possibly raise or lower the overall relevance weight of its discourse context. An algorithm for this kind of context-based retrieval is presented in [5].

6 Conclusion

The COLLATE system represents a new type of collaboratory supporting content- and concept-based work with digital document sources. Innovative task-based interfaces support professional domain experts in their individual and collaborative scholarly work, i.e., analyzing, interpreting, indexing, and annotating sources. The metadata provided in this way are managed by an advanced XML-based content manager and an intelligent content- and context-based retrieval system.

References

1. Allen J, Ferguson G, Stent A (2001) An architecture for more realistic conversational systems. In: Proceedings of Intelligent User Interfaces (IUI '01). ACM Press, New York
2. Bentley R, Horstmann T, Sikkel K, Trevor J (1995) Supporting collaborative information sharing with the World-Wide Web: the BSCW shared workspace system. In: Proceedings of the 4th international WWW conference, Boston, December 1995, pp 63–74
3. Brocks H, Thiel U, Stein A, Dirsch-Weigand A (2001) Customizable retrieval functions based on user tasks in the cultural heritage domain. In: Constantopoulos P, Sølvberg I (eds) Proceedings of the European conference on research and advanced technology for digital libraries (ECDL 2001), Darmstadt, Germany, September 2001. Lecture notes in computer science, vol 2163. Springer, Berlin Heidelberg New York, pp 37–48
4. Brocks H, Stein A, Thiel U, Frommholz I, Dirsch-Weigand A (2002) How to incorporate collaborative discourse in cultural digital libraries. In: Proceedings of the ECAI 2002 workshop on semantic authoring, annotation and knowledge markup (SAAKM02), Lyon, France, July 2002
5. Frommholz I, Brocks H, Thiel U, Neuhold E, Iannone L, Semeraro G, Berardi M, Ceci M (2003) Document-centered collaboration for scholars in the humanities – the COLLATE system. In: Proceedings of the European conference on research and advanced technology for digital libraries (ECDL 2003), Trondheim, Norway, August 2003. Lecture notes in computer science, vol 2769. Springer, Berlin Heidelberg New York, pp 434–445
6. Kahan J, Koivunen MR, Prud'Hommeaux E, Swick RR (2001) Annotea: An open rdf infrastructure for shared web annotations. In: Proceedings of the WWW10 international conference, Hong Kong, May 2001, pp 623–632
7. Keiper J, Brocks H, Dirsch-Weigand A, Stein A, Thiel U (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), Milan, Italy, September 2001, Politecnico di Milano, Milano, pp 405–511
8. Kouzes RT, Myers JD, Wulf WA (1996) Doing science on the Internet. IEEE Comput 29(8):40–46
9. Maier E, Hovy EH (1993) Organising discourse structure relations using metafunctions. In: Horacek H, Zock M (eds) New concepts in natural language processing, Pinter, London, pp 69–86
10. Mann WC, Thompson SA (1987) Rhetorical structure theory: a theory of text organization. In: Polanyi L (ed) The structure of discourse, Ablex, Norwood, NJ, pp 85–96
11. Nichols DM, Pemberton D, Dalhoumi S, Larouk O, Belisle C, Twindale MB DEBORA: Developing an interface to support collaboration in a digital library. In: Borbinha JL, Baker T (eds) Proceedings of the 4th European conference on research and advanced technology for digital libraries (ECDL 2000), Lisbon, Portugal, September 2000. Lecture notes in computer science, vol 1923. Springer, Berlin Heidelberg New York, pp 239–248
12. Searle JR (1979) A taxonomy of illocutionary acts. In: Searle JR (ed) Expression and meaning: studies in the theory of speech acts, Cambridge University Press, pp 1–29
13. Sitter S, Stein A (1992) Modeling the illocutionary aspects of information-seeking dialogues. Inf Process Manage 28(2): 165–180
14. Stein A, Gulla JA, Thiel U (1999) User-tailored planning of mixed initiative information-seeking dialogues. User Model User-Adapted Interact 9(1–2):133–166