Document-Centered Collaboration for Scholars in the Humanities – The COLLATE System

Ingo Frommholz[1], Holger Brocks[1], Ulrich Thiel[1], Erich Neuhold[1], Luigi Iannone[2], Giovanni Semeraro[2], Margherita Berardi[2], and Michelangelo Ceci[2]

[1] Fraunhofer IPSI, Darmstadt, Germany
{frommholz, brocks, thiel, neuhold}@ipsi.fraunhofer.de
[2] Dipartimento di Informatica, University of Bari, Italy
{iannone, semeraro, berardi, ceci}@di.uniba.it

Abstract. In contrast to electronic document collections we find in contemporary digital libraries, systems applied in the cultural domain have to satisfy specific requirements with respect to data ingest, management, and access. Such systems should also be able to support the collaborative work of domain experts and furthermore offer mechanisms to exploit the value-added information resulting from a collaborative process like scientific discussions. In this paper, we present the solutions to these requirements developed and realized in the COLLATE system, where advanced methods for document classification, content management, and a new kind of context-based retrieval using scientific discourses are applied.

1 Introduction

Much scientific work with historical document collections is characterised by additional requirements compared to those usually found in contemporary digital libraries. For instance, the original source material might be lost and no longer be available, hence research has to rely on references found in secondary documents describing the original artifacts. Since working with cultural content is highly interpretative and incremental, the examination of scientific discussions about the material might grant more insight than the documents themselves. Therefore, a digital library dealing with historical material should offer support for storage, identification and access to the cultural documents, as well as providing the means to assist collaborative knowledge working processes where additional knowledge is gained based on the discussions about the material at hand. In this context, annotations entered by domain experts serve as building blocks for establishing scientific discourses. In addition to metadata generated by traditional formal indexing (e.g., cataloguing, controlled keywords) the value-added information represented in those discourses can be exploited in order to provide advanced content- and context-based access to the underlying digital repository.

The COLLATE system, which will be presented in this paper, employs advanced techniques to provide adequate access to historic film-related documents and their associated metadata, as well as to support collaboration between its

T. Koch and I.T. Sølvberg (Eds.): ECDL 2003, LNCS 2769, pp. 434–445, 2003.

professional users working with the material. Based on the reference model for an OAIS (Open Archival Information System, [8]) we will first outline the motivation behind the COLLATE system architecture, which is a revised version of an older one presented in [6], and then demonstrate the necessity to extend OAIS in order to support collaborative work processes. Then, we will take a closer look at some of the major system components, in particular those responsible for automatic document classification, XML-based content management and advanced, discourse-related retrieval functions.

2 The COLLATE System

The COLLATE[1] system focuses on historic film documentation, dealing with digitized versions of documents about European films of the 20ties and 30ties of the last century. Such documents can be censorship documents, newspaper articles, posters, advertisement material, registration cards, and photos. The system has to support several tasks required for managing its cultural content like, e.g., metadata creation, in-depth analysis and interpretation, and collaborative discussions about the documents and their related metadata. Advanced embedded search and retrieval functionality, both concept- and context-based, represents a fundamental requirement to maintain a continuous flow of information between the various actors involved.

2.1 Adding Collaboration to an Open Archival Information System

In [6], we presented an overall architecture of the COLLATE system which is based on a task-oriented layer concept. However, it turned out that the work of film scientists is not dividable into static tasks, but into phases manifesting themselves as discourses about specific topics. In contrast to tasks, such phases may be resumed if new contributions are added. In this case, a user can refer to an earlier point in the discourse. Our analysis of user behavior resulted in a revised COLLATE system architecture, as shown in Figure 1, which reflects our shift from a task-oriented towards a discourse-oriented view. The architecture is based on the reference model for an Open Archival Information System (OAIS). According to the definition in [8], "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 designated 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[2], which was founded in the area of e-print 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 [8]. An OAIS is surrounded

[1] Collaboratory for annotation, indexing and retrieval of digitized historical archive material IST-1999-20882, `http://www.collate.de/`.

[2] `http://www.openarchives.org/`

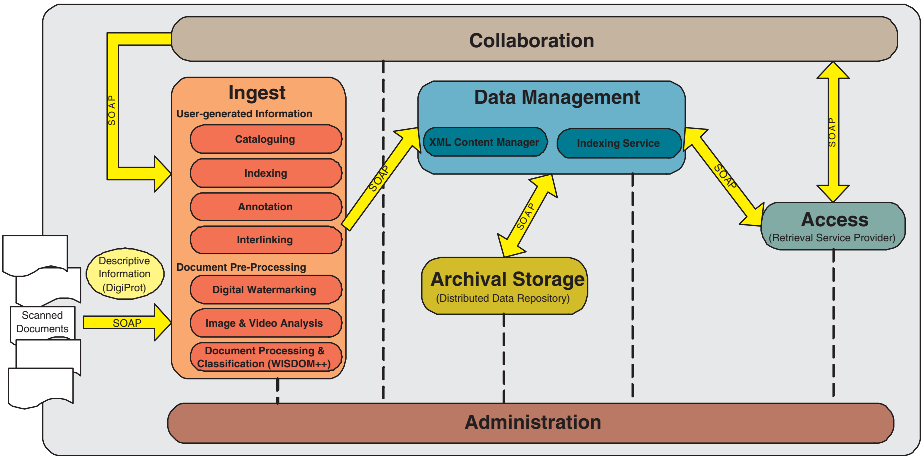


Fig. 1. The COLLATE system architecture

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* use the services provided by an OAIS in order to find the information they are interested in.

The *Ingest* component provides functionality for the submission of information or objects to be stored in the system. Producers can insert scanned documents (together with some descriptive information) and user-generated information. The pre-processing contains the creation of digital watermarks, an image & video analysis, and certain document processing and classification techniques for the automatic generation of metadata. User-generated information is created by producers in a collaborative process. Documents and information prepared by Ingest services are sent to the *Data Management* component. In COLLATE, Data Management consists of two modules: The *XML Content Manager* realizes the storage of and access to digitized documents and user-generated information, which is serialized in XML. The *Indexing Service* updates the retrieval index on the arrival of new data objects. All Data Management modules are closely connected to the *Archival Storage* component, which is responsible for the maintenance of the distributed data repository by providing some low-level database access. 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 described in Section 2.2. For the communication between all system components, the Simple Object Access Protocol (SOAP)[3] is used. Therefore, the services of all components are implemented as SOAP-based web services, making them scalable and their usage platform-independent, hence enabling interoperability. In this way, COLLATE services can easily be made available to other applications.

2.2 Enabling Collaboration

Scientists in the Humanities maintain highly effective mechanisms for collaboration which have not been supported by systems so far. Ensuring 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 stored, user-generated metadata (cataloguing, indexing, annotating) is created collaboratively. If, for instance, a user retrieves a specific document and the metadata already associated with it, she might be willing to contribute additional knowledge, e.g., comment upon an annotation by another user or complete missing cataloguing information. This kind of rather passive collaboration alone would be insufficient to justify complex collaboration services. In COLLATE we focus on active, system-internal support for collaboration, in particular proactive notifications about, e.g., newly submitted documents, and requests for comments broadcast to relevant domain experts. It should also be possible to bring together experts working in similar contexts, but who did not know of each other until now. In COLLATE's *collaboration* component, we therefore apply an agent-based approach, the MACIS (Multiple Agents for Collaborative Information Systems) framework, which has been developed to implement collaborative, distributed information environments (see also[5]).

3 System Components

3.1 Ingest

The Ingest component is responsible for document and metadata submission. In COLLATE, we distinguish between two kind of metadata: user-generated information and metadata which are automatically generated during document pre-processing.

The cultural material in COLLATE, which consists of scanned versions of the original resource, cannot be used for access and retrieval as it is. Therefore, methods have to be applied to extract as much information from both textual and pictorial material and make them as machine-accessible as possible. This requires us to go beyond mere OCR techniques for textual documents, and to apply methods for image analysis in order to automatically index pictorial documents. *WISDOM++*[4] is a document analysis system that can transform textual

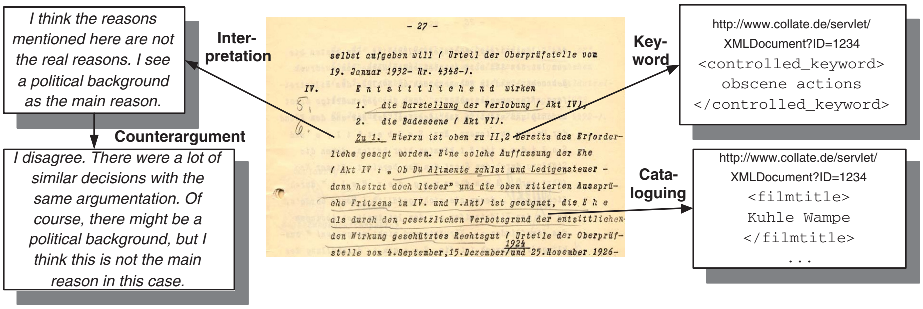
[3] `http://www.w3.org/TR/2002/CR-soap12-part1-20021219/`

[4] `http://www.di.uniba.it/~malerba/wisdom++/`

paper documents into XML format [2]. This is a complex process involving several steps performed by WISDOM++. First, the image is *segmented* into basic layout components (non-overlapping rectangular blocks enclosing content portions). These layout components are *classified* according to the type of their content which can be, e.g., text, graphics, etc. Second, a perceptual organization phase called *layout analysis* is performed to detect structures among blocks. The result is a tree-like structure which is a more abstract representation of the document layout. This representation associates the content of a document with a hierarchy of layout components, such as blocks, lines, and paragraphs. Third, the *document image classification* step aims to identify the membership class (or type) of a document (e.g. censorship decision, newspaper article, etc.), and it is performed using some first-order rules which can be automatically learned from a set of training examples [13]. *Document image understanding* (or *interpretation*) [18] creates a mapping of the layout structure into the *logical structure*, which associates the content with a hierarchy of logical components, such as title/authors of a scientific article, or the name of the censor in a censorship document, and so on. In many documents the logical and the layout structures are closely related. For instance, the title of an article is usually located at the top of the first page of a document and it is written with the largest character set used in the document. Once logical and layout structure have been mapped, OCR can be applied only to those textual components of interest for the application domain, and its content can be stored for future retrieval purposes. Document image understanding also uses first-order rules [13]. The result of the document analysis is an XML document that makes the document image retrievable. In this way, the system can automatically determine not only the type of document, but is also able to identify interesting parts of a document and extract the information given in this part plus its meaning. As an example, we can automatically identify a document as being a censorship document coming from a specific authority and can additionally identify, e.g., the name of the censor (which is usually in a certain part of censorship documents by this institute).

For pictorial material, we can automatically extract metadata by performing an image & video analysis. The result of such an analysis is the extraction of basic image features like, e.g., edge analysis values, grayscale, and entropy, which support the classification of the pictorial material and the extraction of index terms describing the picture. See [11] for further details. Documents can optionally be supplied with digital watermarks. Watermarking is beyond the focus of this paper; refer to [7] for a description.

User-generated metadata ranges from formal indexing to content-based information gathered in collaborative processes (such as, e.g., in discussions about certain interpretations of documents or annotations). Well established in library science, formal indexing in COLLATE corresponds to collecting bibliographic metadata and the assignment of keywords to a document or certain passages of it. While formal indexing represents a fundamental prerequisite for enabling access to the documents, our focus is set on collaborative, content-based indexing and the resulting discourses established in scientific discussions about certain topics. Source analysis in the Humanities is an interpretative process, which reflects the current subjective point of view of the scientist. If in a col-



I think the reasons mentioned here are not the real reasons. I see a political background as the main reason.

I disagree. There were a lot of similar decisions with the same argumentation. Of course, there might be a political background, but I think this is not the main reason in this case.

- 27 -

selbst aufgeben will (Urteil der Oberprüfstelle vom 19. Januar 1932- Nr. 4348-).

IV. Entsittlichend wirken

5, 1. die Darstellung der Verlobung (Akt IV),

6. 2. die Badescene (Akt VI).

Zu 1. Hierzu ist oben zu II,2 bereits das Erforderliche gesagt worden. Eine solche Auffassung der Ehe (Akt IV : „ Ob Du Alimente zahlst und Ledigensteuer - dann heirat doch lieber" und die oben zitierten Aussprüche Fritzens im IV. und V.Akt) ist geeignet, die Ehe als durch den gesetzlichen Verbotsgrund der entsittlichenden Wirkung geschütztes Rechtsgut (Urteile der Oberprüfstelle vom 4.September,15.Dezember 1924 und 25.November 1926-

http://www.collate.de/servlet/XMLDocument?ID=1234

```
<controlled_keyword>
obscene actions
</controlled_keyword>
```

http://www.collate.de/servlet/XMLDocument?ID=1234

```
<filmtitle>
Kuhle Wampe
</filmtitle>
...
```

Fig. 2. Annotation thread, keywords, cataloguing

laborative digital library like COLLATE this point of view can be expressed and manifested (in an annotation), the expert tacit knowledge becomes explicit and can be accessed by other users. This in turn means that certain subjects in the experts' statements themselves become the focus of interest and are controversially discussed, i.e. they are commented upon. Since we interpret discussion threads associated to a document as coherent linguistic entities, we have empirically devised a set of admissible discourse structure relations which classify the interrelations between the annotations. Ranging from factual (e.g., providing additional information) to more interpersonal levels (i.e. focusing on certain expertise of the participants of a discussion), we use these relations to structure the resulting discourses (see [4] for a detailed definition of discourse structure relations). Using the Resource Description Framework[5] (RDF) to represent the interrelations[6], we obtain a directed acyclic graph with the digitized document as the root node. The set of nodes in the graph is the set of the document and the annotations occurring in the discussion; the typed links between the annotations, or the document and its direct annotations, respectively, form the vertices of the graph. We call this graph the *annotation thread*. An annotation thread forms a *collaborative discourse* about a specific topic (which is, in our case, the digitized document). With the typed links and the interpretation of annotations as node, an annotation thread can be seen as a hypertext (according to the definition of hypertexts in [1]). Figure 2 shows an example of an annotation thread (on the left side) with the digitized document as root node, and additional cataloguing and indexing information (on the right side). We can see an interpretation of the document, which is attacked by another scientist, using an annotation together with a "counterargument" relation type.

[5] http://www.w3c.org/RDF/

[6] The corresponding RDF Schema can be found at http://www.collate.de/RDF/collate.rdfs.

3.2 Data Management and Archival Storage

Data submitted to and set up by the Ingest component is forwarded to the Data Management component, which is coupled to the Archival Storage. Archival Storage provides functionality for the distributed data repository, which consists of relational database management systems. Archival storage offers low level access and storage capabilities, mainly based on SQL. Data Management modules make use of these capabilities. Data Management consists of two components, which are the XML Content Manager and the Indexing Service.

XML Content Manager. The COLLATE system has been devised as a distributed architecture. In fact, the idea of a collaboratory entails the new concept of entities - software components and users - that need to work together, but both in different locations (distribution along space) and asynchronously (distribution in time). This vision justifies our approach to developing a platform that is capable of tackling the distribution issue along these two dimensions while providing complete transparency w.r.t. Data Management and Archival Storage to end users. Hence the task of content management is delegated to a dedicated component named *XML Content Manager* (XMLCM)[7]. Content management components have to deal with all kinds of processes, e.g., the insertion of a scanned document in the repository, and the insertion/access of some metadata on a specific resource. XMLCM comprises three layers: the integration layer, core components, and the persistence layer. The *integration layer* is the handle which external applications can rely on to use XMLCM services. It has been developed within the Web Services paradigm, using SOAP technology. Thus, the integration layer allows the communication with services of other COLLATE components, like those from Ingest and Access. *Core components* are those components that have to manage COLLATE resources represented as XML documents inside XMLCM. As sketched in Figure 3 they provide: access to XML documents at different levels of granularity (DocumentManager, ElementManager), the possibility of managing XML Schemas or DTDs for the documents in the Repository (MetadataManager), full support for accessing the repository (QueryManager), the possibility of managing the underlying RDBMSes, thanks to the BridgeXMLSQL component (in this way allowing the storage/retrieval of non XML resources such as scanned documents), and a complete layer for managing RDF Descriptions (models as well as single RDF statements) used for connecting COLLATE resources. *XML Persistence layer* is the set of components that cope with the problems of effective storage/retrieval of XML resources, using the low-level access and storage functionality provided by the Archival Storage.

Indexing Service. The *Indexing Service* is responsible for the maintenance of the index used for retrieval. Every time new data arrive from the Ingest, XMLCM stores the data in the repository and contacts the indexing service,

[7] XMLCM has been developed by one of our project partners, SWORD ICT (`http://www.sword.it/english/index.htm`).

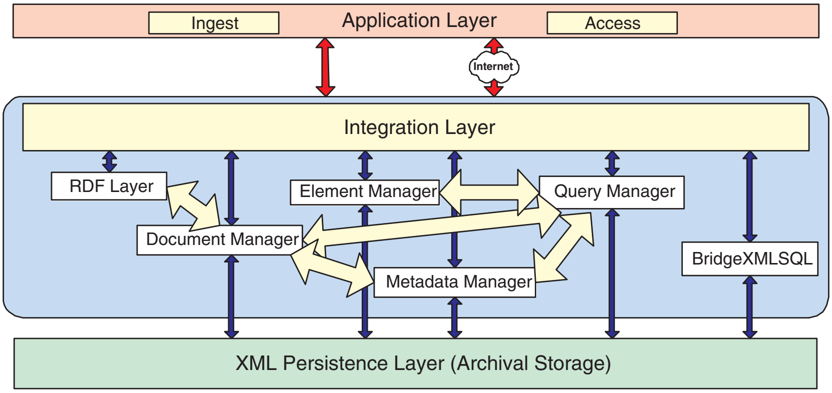


Fig. 3. XML Content Manager Architecture

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.

3.3 Access

To provide access to the cultural material, we have devised a set of advanced customizable retrieval functions for COLLATE [6]. Given a query q, we calculate for each document d the *retrieval status value* of d w.r.t. q, denoted by $r(d, q) \in [0, 1]$. Having such retrieval weights for all documents, we rank the documents based on descending retrieval status values. Using standard IR techniques, COLLATE provides functions for calculating $r_{meta}(d, q)$, which is the retrieval weight based on cataloguing and keyword metadata or data automatically derived from document pre-processing. Hence $r_{meta}(d, q)$ is computed on the information we can gain from the document itself. In contrast to that, we focus on a more advanced retrieval method called *context-based retrieval* which regards the annotation thread as extension of the document it belongs to, conveying additional information which could not be derived from the document itself and has an interpretative, thus subjective nature.

Context-based Retrieval. The context we are talking about in context-based retrieval is the *discourse context*; not only a statement in the discourse is considered, but also its position in the discourse and is type, given by a discourse structure relation. Thus, we do not only consider *what* is said (in an annotation), but also *where, when* and to *what purpose* it was said (the position of the annotation in the annotation thread plus the link type connecting it with its source). Furthermore, a document is judged in the light of the discussion about it. To demonstrate the value of context-based retrieval, we provide a simple example. Suppose a user is looking for censorship decisions mainly taken for political reasons. Returning to the document d and metadata depicted in Figure 2, we do not find any evidence on political reasons in the cataloguing information or keywords. Based on this only, $r(d, q)$ is very low for this document, say 0.01. Nevertheless, another film scientist has put her interpretative analysis of the document into an annotation, stating that she thinks the censorship decision has a political background, even though it is not mentioned explicitly in the censorship document. Therefore, the document (together with this annotation) can be interesting for a user seeking political censorship. The retrieval engine would take this fact into account by raising the retrieval weight for the document d to the value of, say, 0.4. Going further, a second film scientist has attacked the statement of the first one by annotating the first annotation and using the "counterargument" relation type [4]. This means the statement of the first scientist is controversial and far away from being safely regarded as a fact. To reflect this situation, the retrieval engine now lowers the previously raised retrieval weight for document d to, e.g., 0.25. If the discussion were to go on, all contributions would have an impact on the overall weight of document d, depending on their position in the discourse, their content and the type of the incoming link. The kind of retrieval weight for d which is based on the discourse on d is referred to as $r_{dis}(d, q)$, which is computed in a recursive way. For each annotation A in the annotation thread, we need to calculate $r_{dis}(A, q)$, the retrieval status value of this annotation w.r.t. the query, taking the annotation subthread with A as the root element into account. The direct relation between a source annotation A and the destination A' with a link of type X (with X being a discourse structure relation) is defined as the triple $rel(X, A, A')$. Then, $r_{rel}(A, A', rel(X, A, A'), q)$ is the retrieval weight of A w.r.t. q, having a directly connected annotation A' linked with type X. To compute $r_{dis}(A, q)$, we look at each direct successor of A (the set $succ(A)$) in the annotation thread, with

$$r_{dis}(A, q) = \frac{1}{|succ(A)|} \sum_{A' \in succ(A)} r_{rel}(A, A', rel(X, A, A'), q) \tag{1}$$

It is $r_{rel}(A, A', rel(X, A, A'), q) = f(r_{ann}(A, q), r_{dis}(A', q), X) \in [0, 1]$ with $r_{ann}(A, q)$ as the retrieval status value of annotation A without taking any context into account (calculated, e.g., by applying full-text retrieval methods), so (1) is a recursive function. Furthermore, it is $r_{dis}(A, q) = r_{ann}(A, q)$, if A is a leaf in the annotation thread, so the recursion terminates. We are currently evaluating several strategies for the calculation of $r_{rel}(A, A', rel(X, A, A'), q)$. For X being a counterargument, it should be $r_{rel}(A, A', rel(X, A, A'), q) < r_{ann}(A, q)$ (A' being direct successor of A in the annotation thread), since counterarguments weaken the argument made in A. On the other hand, if X is a supportive argument, then $r_{rel}(A, A', rel(X, A, A'), q) > r_{ann}(A, q)$. Supportive arguments therefore strengthen a statement made in A. Finally, for a document d, let

$$r_{dis}(d, q) = \max_{A \in succ(d)} r_{dis}(A, q) \tag{2}$$

(2) is justified by our view of direct annotations to a document being interpretations; we take the weight of the best-matching interpretation since we need a

measure which is independent of the number of interpretations. To achieve one single ranking, it should be possible to use both $r_{meta}(d,q)$ and $r_{dis}(d,q)$ in a balanced way. This leads to $r(d,q) = \alpha \cdot r_{dis}(d,q) + (1-\alpha) \cdot r_{meta}(d,q)$. Depending on the user's preferences, $\alpha \in [0,1]$ can be high (emphasizing the scientific discourse) or low (emphasizing the "hard facts" only).

Retrieval Engine and Result-Set Enrichment. We apply *HySpirit*[8][10], which is an implementation of probabilistic Datalog, providing the required support for retrieval based on metadata, full texts, and even hypertexts. HySpirit can access Datalog clauses stored in an RDBMS. After submitting a query, the retrieval engine calculates, depending on the kind of retrieval to be performed, a document ranking. This ranking is enriched with appropriate metadata obtained from the Data Management, and then set up in order to present it to the user.

4 Related Work

Some efforts have been made before to create collaborative information spaces. *BSCW* [3] provides web-based access on shared objects for collaborative environments. A commercial groupware product providing means for information sharing, but with limited collaboration support is *Lotus Notes*[9]. Collaboratories (e.g., [12]) more thoroughly support studies of the source material, which is stored as a distributed set of digitized source documents. *DEBORA* [14] enables digital access to books of the Renaissance, also offering annotation functionality, but without an explicit discourse model. *Annotea*[10] is a web-based annotation system using RDF. Similar to COLLATE, annotations are seen as statements about web pages, but do not establish a scientific discourse. Hypertext Information Retrieval (HIR) has been a research topic for many years. Besides direct search, hypertext information systems also offer the means to navigate and browse through the information space, resulting in the definition of combined models covering text retrieval and benefits gained from hypertext structures [1]. *Google* regards the whole World Wide Web as a hypertext and makes use of its link structure by applying the PageRank algorithm [15], but, as a Web search engine, does not take any link types into account. Frei and Stieger present an algorithm using typed links based on spreading activation [9], which is similar to the one presented in this paper, but cannot be applied to special hypertexts modeling a discourse like we have in COLLATE. Besides these few examples, there exist a number of other HIR systems. Link types similar to those defined in the COLLATE project, but not focused on scientific discourses, were introduced in 1983 by Randall Trigg [17] as well as, to state another example, in the authoring environment *SEPIA* [16].

[8] `http://www.hyspirit.com/`

[9] `http://www.lotus.com/home.nsf/welcome/notes`

[10] `http://www.w3.org/2001/Annotea/`

5 Conclusions

In this paper, we presented the COLLATE system and its advanced ingest, data management, access and collaboration methods which make it a system capable of dealing with the requirements arising from maintaining and working with historical cultural material. WISDOM++, a tool for advanced automatic document processing and classification, is applied. XMLCM is used to manage the digitized and user-generated content stored in the distributed data repository. In order to make use of the results coming from collaborative discussions, we have modeled scientific discourses as hypertext and, with discourse structure relations, introduced appropriate link types. Context-based retrieval directly uses the value-added information contained in scientific discourses, taking its subjective nature into account.

Lessons learned. The solutions described in this paper are rooted in user requirements we identified in [6]. Taking these as a starting point, we had many discussions with our targeted users, namely three European film archives which are part of the COLLATE consortium. Considering the valuable empirical input we got from them, many of the methods described in [6] have been implemented and are in use; other concepts had to be modified (as can be seen in particular by our shift from a task-oriented to a discourse-oriented view as described in Section 2.1). We provided our users with an external web-based discussion tool; evaluating their discussions formed the basis for our model of scientific discourses using discourse structure relations, and for the context-based retrieval. The need to offer a proactive collaboration component was derived from empirical data as well. A COLLATE prototype has been implemented and is used at the three film archives' sites. Discourse structure relations and context-based retrieval have recently been introduced to our users. We will collect feedback from the users in order to evaluate the acceptance of our approaches as well as to measure the efficiency and effectiveness of our context-based retrieval approach.

References

1. Maristella Agosti. An overview of hypertext. In Maristella Agosti and Alan F. Smeaton, editors, *Information Retrieval and Hypertext*, chapter 2. Kluwer Academic Publishers, Boston et al., 1996.
2. O. Altamura, F. Esposito, and D. Malerba. Transforming paper documents into XML format with WISDOM++. *International Journal for Document Analysis and Recognition*, 4:2–17, 2001.
3. R. Bentley, T. Horstmann, K. Sikkel, and J Trevor. Supporting collaborative information sharing with the World-Wide Web: The BSCW shared workspace system. In *4th International WWW Conference*, Boston, 1995.
4. H. Brocks, A. Stein, U. Thiel, I. Frommholz, and A Dirsch-Weigand. How to incorporate collaborative discourse in cultural digital libraries. In *Proceedings of the ECAI 2002 Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM02)*, Lyon, France, July 2002.

5. H. Brocks, U. Thiel, and A. Stein. Agent-based user interface customization in a system-mediated collaboration environment. In *Proceedings of the 10th International Conference on Human-Computer Interaction*, 2003.
6. H. Brocks, U. Thiel, A. Stein, and A. Dirsch-Weigand. Customizable retrieval functions based on user tasks in the cultural heritage domain. In Panos Constantopoulos and Ingeborg Sølvberg, editors, *Proceedings of the ECDL 2001, Darmstadt, Germany, Sept. 2001*, Lecture Notes in Computer Science, pages 37–48, Berlin et al., 2001. Springer.
7. J. Dittmann. Content-fragile watermarking for image authentication. In *Proceedings of SPIE: Security and Watermarking of Multimedia Contents III*, volume 4314, San Jose, California, USA, 2001.
8. Consultive Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS), January 2002. `http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf`.
9. H. P. Frei and D. Stieger. The use of semantic links in hypertext information retrieval. *Information Processing and Management: an International Journal*, 31(1):1–13, 1995.
10. N. Fuhr and T. Rölleke. Hyspirit — a Probabilistic Inference Engine for Hypermedia Retrieval in Large Databases. In H.-J. Schek, F. Saltor, I. Ramos, and G. Alonso, editors, *Proceedings of the 6th International Conference on Extending Database Technology (EDBT), Valencia, Spain*, Lecture Notes in Computer Science, pages 24–38, Berlin et al., 1998. Springer.
11. Silvia Hollfelder, Andre Everts, and Ulrich Thiel. Designing for semantic access: A video browsing system. *Multimedia Tools and Applications*, 11(3):281–293, August 2000.
12. R.T. Kouzes, J.D. Myers, and W.A. Wulf. Doing science on the internet. *IEEE Computer*, 29(8), 1996.
13. D. Malerba, F. Esposito, and F.A. Lisi. Learning recursive theories with ATRE. In *Proc. of the 13th European Conf. on Artificial Intelligence*, pages 435–439. John Wiley & Sons, 1998.
14. D.M. Nichols, D. Pemberton, S. Dalhoumi, O. Larouk, C. Belisle, and Twindale M.B. DEBORA: Developing an Interface to Support Collaboration in a Digital Library. In J.L. Borbinha and T. Baker, editors, *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2000)*, Lecture Notes in Computer Science, pages 239–248, Berlin et al., 2000. Springer.
15. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
16. M. Thüring, J.M. Haake, and J. Hannemann. What's ELIZA doing in the Chinese Room? Incoherent hyperdocuments - and how to avoid them. In *Hypertext '91 Proceedings*, pages 161–177, New York, 1991. ACM Press.
17. Randall Trigg. *A Network-Based Approach to Text Handling for the Online Scientific Community.* PhD Thesis, Department of Computer Science, University of Maryland, November 1983.
18. S. Tsujimoto and H. Asada. Understanding multi-articled documents. In *Proceedings of the 10th International Conference on Pattern Recognition*, pages 551–556, 1990.