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

Ingo Frommholz
frommholz@ipsi.fraunhofer.de

Fraunhofer Institute
Integrated Publication and Information Systems (IPSI)
Darmstadt, Germany
http://ipsi.fraunhofer.de/
Motivation and Background:
Digital Libraries in Cultural Heritage

Valuable **historic document collections** exist, but scattered in national archives

- Sources mostly not available online
- Difficult-to-use database and referencing systems
- Lack of content-based indexing and access

Valuable **expert domain knowledge** exists, but mostly inaccessible to externals

- Tacit knowledge, insufficiently documented
- Professional communities lack technology support for collaborative knowledge working
The COLLATE Project (IST-1999-20882)

- Constructing a “Collaborative Information Space”
  - Integrate historical collections and expert knowledge in a distributed multimedia repository
    - European historical film documentation (20s and 30s)
    - Historical film censorship records, press material, photos, film posters, digital film/video fragments
    - Open metadata standards

- Ensure accessibility
  - Work environment for content indexing and annotation
  - Assist collaborative knowledge working processes
  - Content and context-based retrieval

- Evaluate acceptability
  - Empirical studies of real-life user behaviour
Censorship / Registration Cards

BESONDERES STADTAMT II/3
im selbständigen Wirkungsbereiche.

Karten Nr. 6601. Zur behördlichen Vorführung eingereicht von Ufa
000144

Titel des Bildes: "Die Drei von der Tankstelle"

Erzeuger des Bildes: Universum-Film AG, Berlin

Länge des Bildes: 2823 m

Inhalt des Bildes: Ein heiteres Singspiel in 10 Akten. Hauptdarsteller:
Lillian Harvey, Willy Fritsch


Wien, am 30. Jan. 1930

Mag. Prot.
Rückkehr zum europäischen Film

Ein paar wirtschaftliche Entwicklung der letzten Tage mögen darüber deutlich machen, daß die Filmwirtschaft des internationalen Films vollkommen vorbei ist. Was möge das böse nicht doch wahr werden? Es wird angeblich nur noch Filme geben, die in Deutschland einmal so gefilmt wurden, und die wir in Amerika als unseres gehen lassen, damit wir die Kinozuschauer nicht wieder belehren müssen. Aber es scheint doch, als ob auch die Auslandspartei, eine österreichische oder eine amerikanische, auf ähnliche Weise wie die Franzosen und Belgier, sich selbst für das Publikum der Weltkunde, ihm zu bieten, was es verdient, eine Filmwelt zu schaffen, die den Anstoß gibt, den sie verdient, selbst die europäische Marktfähigkeit wieder abzunehmen. Es wird von der neuen und feurigen Nachkriegszeit, die zur weiteren Entwicklung des Films gebracht, die deutsche Kultur auch zu neuen Schöpfungen gebracht haben."
Posters
In der Kammersitzung vom 8. Juni 1931, an der teilgenommen haben:

als Vorsitzender: Reg. Rat Zimmermann
als Beisitzer: Kahlenberg (Filmindustrie)
Wagner (Kunst und Literatur)
Czech (Volkswohlfahrt)
Züh-Köln

ist über den Bildstreifen:

"Im Westen nichts Neues"

folgende Entscheidung verkündet worden:

Der Bildstreifen wird zur Vorführung im Deutschen Reich zugelassen, darf jedoch nur vor den gesetzgebenden Körperschaften des Reiches und der Länder sowie in ge-
Collaboration in COLLATE

- **Film-centered collaboration**
  - people working on documents related to some specific film

- **Document-centered collaboration**
  - Annotations and documents are discussed
  - current document context is used to determine potential collaborators

- **Discourse-centered collaboration**
  - register interest in a particular discourse/discussion
We previously presented a task-oriented architecture, but:

- Film scientist’s collaborative work is not divisible into static tasks

Instead:

- Discourse about specific topics
- Discourse can be resumed if new contributions are added

Shift from task-oriented towards discourse-oriented view
COLLATE Architecture
Document-centered Discussions

- Source Analysis in the Humanities is an interpretative process
- Points of view expressed in annotations
- Experts’ tacit knowledge becomes explicit
- **Establishing a discourse context**
  - Modelling discourse as interrelated nested annotations
  - Annotation thread reflects **scientific collaborative discourse**
  - Typed links (DSR) between
    - Document and annotation
    - Annotations
Discourse Structure Relations

elaboration

comparison

cause

analogy
difference

background information

interpretation

argumentation

support argument
counterargument

interpersonal
Information Objects

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

Interpretation

Counterargument

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.

Cataloguing

Keyword

<controlled_keyword> obscene actions </controlled_keyword>

Document Interpretation

<assessors_chairman> Oberregierungsrat Dr. Becker 
Beisitzer: Justizrat Dr. Rosenthal... </assessors_chairman>

<filmtitle> Kuhle Wampe </filmtitle>

© Ingo Frommholz

ECDL 2003 Trondheim, Norway
Many results of the film scientists’ collaborative work are manifested in the discussion about documents.

Expert users are taking these results as a basis for their further work, which means they have very specific information needs.

Documents have to be accessed not only by their content but also based on their associated discussions.

Annotation Thread conveys valuable, but subjective information.

Deep analysis of annotations and their interrelations.

Retrieval functions stimulate scientific collaboration.
Access: Document Retrieval

- For a query $q$, a ranking of documents is returned. Therefore, a retrieval weight $r(d, q)$ has to be calculated for each document $d$.

- Documents are ranked according to descending retrieval weights.

- The retrieval is based on the document’s metadata (cataloguing, keywords) and on the annotation thread.
In COLLATE, we deal with the **discourse context**.

A document is seen in the light of its interpretations, i.e. *what did the experts discuss about the document?*

➤ Analyzing content and structure

Example: Consider the request for “*all censorship decisions made for political reasons*”.
Example: Considering only the Metadata

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.
Example: Considering a Direct Interpretation

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.

...<controlled_keyword>
obscene actions
</controlled_keyword>
...

...<assessors_chairman>
Oberregierungsrat Dr. Becker
Beisitzer: Justizrat Dr. Rosenthal...
</assessors_chairman>
...

0.32
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.
Calculation of the Retrieval Weight $r(d, q)$

$r_{\text{meta}}(d, q)$ is the retrieval weight of the document $d$ w.r.t. the query, considering the metadata (cataloguing, keywords) alone.

$r_{\text{dis}}(d, q)$ is the retrieval weight considering the annotation thread (i.e. the scientific discourse about the document) only.

With this:

$$r(d, q) = \alpha \cdot r_{\text{dis}}(d, q) + (1 - \alpha) \cdot r_{\text{meta}}(d, q)$$
COLLATE User Interface
Initial Results

- Two prototypes have been implemented
- Gathered valuable empirical input from three European film archives
- Annotation tool supporting DSR introduced to our users

Next prototype:
- Context-based retrieval and collaboration methods are featured in the next prototype

Evaluations of context-based retrieval and collaboration approach
Conclusions

- COLLATE is a system to maintain and work with historical cultural material
- We presented the COLLATE system and its ingest, data management, access and collaboration methods
- Collaboration of experts
  - Manifests itself in scientific discussions
  - The value-added information gained there can be used to support context-based retrieval
- Appropriate means for supporting proactive collaboration behaviour have been developed
Thanks for your attention! 😊

More info:
http://www.collate.de/
frommholz@ipsi.fraunhofer.de