Margret Schild

Text-Based Film Retrieval 2006

A New Concept to Index, Manage and Present Films, Their Content and Context

This paper will emphasize the perspective of a librarian and information specialist, being myself responsible for the museum libraries of the Film Museum and Theatre Museum at Düsseldorf. I deal with books, journals and other printed material within the library. Another department is responsible for other types of objects – photos, posters, handwritten material, films on reels, video tapes, etc. My task is to acquire printed material, to archive and preserve it, to record it and put it at the disposal of library users. Users are members of staff of the museum, free lancers and volunteers working on projects, and the interested public. The approach of Text-based Film Retrieval (TFR) 2006 was developed by the Institute of Terminology and Applied Knowledge Research (itaw),[1] located at Berlin and affiliated to the Humboldt University. The tool combines the experience of daily practical work with methods of knowledge structuring and presentation in order to organize information around film in an efficient way and to preserve it in the long term.

Information Flood / Chaos

Every time when a film is made, when visual media are produced, the development, the actual shooting and post-production are accompanied by rich textual material: scripts, results of information retrieval during the planning stages, collections of scientific, popular and technical information concerning the subject. When the film is distributed, more material is produced: advertising brochures, film stills, features for broadcasting on TV, trailers, websites on the Internet etc. Since film is not only seen as entertainment but as a cultural phenomenon and subject of special interest, the producers, the distributors, institutions dedicated to the cultural heritage and others try to make this textual material, acquired with great expenditure of human and financial resources, available for potential users (e.g. the audience of a film, film historians, teachers, students etc.). The materials enable users to get to know the process of

1 Further information concerning the institute and the project: www.itaw.hu-berlin.de/imb-prod.htm

film making, distribution and reception. Users and information specialists interested in the subject of film or media know that a lot of material of this kind exists, but is not easy to know how much there actually is and where it is available in the long term.

Currently, accompanying materials are normally issued by the industry, administrations, or public broadcasters. They address the general public as well as special target groups or mediators like teachers or instructors.

Problems

In the past, the material normally was split up into two parts: the film, archived on reels, video tapes, or as data files on the one hand; printed materials on the other. Fairly recently, a third group of materials emerged: digital information – e.g. information on CD or DVD, or websites on the Internet. In Germany, a lot of film-related collections exist: collections and archives for printed material, collections and archives for films, film libraries and information centres. The German National Library and its regional counterparts have the task to collect all printed and digital publications issued by publishers; however, neither brochures, leaflets and other film-related ephemera nor the films themselves fall under this rule. Actually, there is no institution in Germany that has the official task to collect and archive films and film-related materials.

Another problem lies in the difficulty to link between films and the textual or digital material about them, especially when the material refers to a particular image or scene. Especially when the film itself is stored by analogue means on celluloid or video tape, it is difficult to refer to particular places, or vice versa. With the development of DVDs and film presentations on multimedia-equipped computers, this has become much easier.

From the perspective of libraries and archives, the long-term preservation of information comes into focus. Because of technical changes and developments, some facilities have already disappeared (or will have in the near future), e.g. video recorders. Within the field of digital information, archives and libraries have to develop strategies in order to preserve the information in the long term, taking into consideration changing data formats, operating systems, and software, which are often not compatible to earlier versions. And we simply have no experience concerning the archival possibilities of CDs or DVDs. Within the archival community, we only have experience with microfilm and normal film.

The Concept of Text-Based Film Retrieval

TFR 2006 offers a way to present and edit moving images together with the textual material on the same level and on one shared platform – the computer screen. The accompanying material has not to be printed and dispatched. The complete material is available on the Internet without geographical or time limits. If the user wants to store the textual material, he is able to store it on his computer like any other information found on the Internet.

The greatest advantage of this concept is the possibility to link sequences of a film to relevant text parts or key words. It is also possible to add context information, for example by linking to the terms in a dictionary or other resources. The sequence of a film can be seen as a basic structure. By linking between film sequences and texts, the user can easily switch between the two, or navigate through the context of the film.

In TFR 2006, film images and textual materials are displayed simultaneously. Complex structures and highly organized information are presented through multimedia presentations. Different retrieval facilities with regard to the needs of the respective subject can be supplied. The accessibility and the view to the information can be defined free and open for everyone as well as restricted to special users or user groups through accounts and rights management.

The Technical Basis

Starting point was the development of multimedia implementations at the Humboldt University of Berlin. This means the integration of different types of information to achieve visualized content units. One media type is, of course, film.

The time code of a film is used to link film sequences with the relevant textual information. The time code allows the annotation and enrichment of film sequences by textual means. This additional information also enables the user to be directed quickly and precisely to a particular film sequence by free text search, field oriented search or index search. In order to realise this solution, the itaw MediaBase (iMB), an XML platform, was implemented and the Text-based Film Retrieval application was developed.

Within the application, the film files and the text files are stored strictly separate. They are only brought together for the presentation on the computer screen. The film format does not matter - every format can be integrated in the multimedia application: film on DVD (MPEG2), film via streaming (MPEG4 or ASF), or film as download.

The iMB offers tools to record, edit or import information as well as tools for the visualisation and for interactive facilities. The format of the textual files is XML. This format enables the user to copy and paste any part of the text in order to transfer it into his own database without any formatting problems, independent of the word processing software. The transfer of information can be reduced to the two commands "copy" and "paste" without the danger of loosing information.

Which Kinds of Film Can be Presented?

The concept is already used for different film genres. In the case of documentaries, additional information can comprise a lot of different aspects, for example the text of the commentary, original statements of the presented persons, additional information for a better understanding of the content and the presentation of associated subjects. The context is presented together with the content of the film. The necessary context information is normally acquired during the formation process of the film. This means: additional benefit can be achieved with little additional work (costs), using the existing knowledge (i.e. the context information). More images, short or long citations – from secondary literature –, historical or subject related additions can be integrated. The Internet and the internal linking options increase the multidimensional approach to the information.

In the case of movies, the text of the dialogue (or the complete script) and any other material, for example film-historical analyses, the literary model, or other film sequences for comparison can be brought together on one computer screen. The context information is the second level of information, displayed and presented parallel to the film itself (first level of information).

All parts of the presented film sequences can be linked to different levels of background information (context). The film presentation can be stopped at any point. All relevant text parts are presented now, even if there is more than one level of background or additional information available. It is also possible to use this the other way round: If a citation or the title of a chapter is marked, the linked film sequence will be found and displayed immediately. TFR 2006 can be used to provide access to scholarly and technical information about the medium film as well as to media-pedagogical solutions at schools, universities or museums.

Advantages

The information provider defines the structure and depth of indexing on the basis of his needs and the habits and needs of potential target groups. Because of the separate storage of the different kinds of data (text, image, film, sound), it is possible to follow the technical development for each type. Existing electronic information can be imported into the data base and controlled, using the rules defined within the document type definition (DTD) or the XML scheme for the data import. Forms can be used to record new data without knowledge about the hidden XML structure as well as guaranteeing consistency and quality for the data input. Different ways of information retrieval can be offered: free text search, field oriented search, the use of subject headings, the implementation of a thesaurus or a classification.

Some Examples

TFR 2006 was used in a project in cooperation with the public broadcaster Rundfunk Berlin-Brandenburg (RBB). A prototype was developed for the TV documentary *Deutsche und Polen* (Germans and Poles), a four-part history programme (see Figure 1 and dvd.deutsche-und-polen.de). The table of content on the left side allows direct access to a chapter or a special segment of a chapter by a mouse-click on the appropriate title. Playback of the film jumps to the new position as soon as another item is chosen. It is possible to search for terms in the field search. The result list of a search is presented on the left side; again, choosing one of the hits, the relevant film image is displayed on the right side screen. If the user prefers, he can read the text of the found segment (button "Text") rather than watch the part of the film. It is possible to switch between the search results, the relevant film segment and the text pertaining to that segment. The subject index on the left side enables the user to change directly to a special theme. These key words are linked to textual material, to images that are not part of the film, and to bibliographic references, as available on the Internet (context). After broadcasting, the film was commercially available as DVD-ROM. The user can insert the DVD-ROM into his computer drive and use it together with the information presented on the Internet.[2]

The Museum for Film and Television in Berlin has implemented TFR 2006 in its recently enlarged permanent exhibition – the history of broadcasting and television in Germany, especially in the program gallery. The visitor is invited to choose a film and get information about the film, the involved per-

2 In other cases, the film can be made available via streaming at different rates.

sons or institutions. The database with cinematographic information was a file maker data base that was exported within an XML structure. The information had to be combined with the process of choosing and the screening of the film on TV from the video tape.

Further examples:

- The movie *Bis zum Horizont und weiter* (1998, Peter Kahane) was published on CD-ROM together with different versions of the script (final printed version, realized version) as well as information about the making of and involved persons.
- A choice of documentary films was presented during the 50th International Festival of Documentary Film in Leipzig.

TFR 2006 can be implemented throughout the whole process of film making, film distribution and reception. It allows all information input to be documented and re-used for other purposes – resulting in an increasing information data basis, collecting all available and used material. It allows to combine different platforms and to combine finished parts of a project with not finished or additional material. The use of XML for data management and structuring enables the user to define the structures with regard to the content, the needed depth of indexing and to manage all types of information in an adequate way, following the development of the technological basis.

Figure 1. *Deutsche & Polen*