European CHronicles On-line

Final Report

IST-1999-11994

Information Society Technologies

Authors

Franciska de Jong	University of Twente
Diego Giuliani	IRST-ITC
Martin Keiser	Eurospider
Stephan Kopf	University of Mannheim
Lori Lamel	LIMSI-CNRS
Jonan Oomen	NAA
Felix Rauch	Memoriav
Marco Rendina	Istituto Luce
Stephan Schneider	Tecmath
Howard Wactlar	Carniege Mellon University
Pasquale Savino	ISTI-CNR

Editor

Pasquale Savino

ECHO – European Chronicles on Line (IST-1999-11994), final report, March 2003

The ECHO project has been funded by the IST programme of the European Commission under the V Framework; the project began in February 2000 and will be completed in March 2003. Its objective has been to develop a digital library service for historical films owned by large national audio-visual archives. Actually being able to see and hear an account of a historical event, filmed in the original context, is very different from reading about it. The ECHO services allow users to search and access these documentary film collections and exploit the content for their own particular requirements, whether commercial, educational, leisure, or whatever.

The project has involved a number of European institutions holding or managing unique collections of documentary films, dating from the beginning of the century until the seventies: Istituto Luce, Italy; Institut Nationale Audiovisuel, France; Netherlands Audiovisual Archive; and Memoriav, Switzerland. Academic partners are ISTI-CNR (coordinator), CNRS-LIMSI, ITC-irst, University of Twente, University of Mannheim and Carnegie Mellon; industrial partners are Techmath and Eurospider Information Technology. These collections are of great value as they document different aspects of life in Europe (social, cultural, political, economic) during this period of time. The set of services implemented by ECHO thus provides users with access to significant portions of their cultural heritage, which would otherwise be almost inaccessible. In addition, users can compare the way in which an event or phenomenon is documented in their own country with how it is reported in others, or can investigate how different countries have documented a particular historical period of their life, etc.. This means that the ECHO services have to operate over linguistic, cultural and national boundaries, while respecting the requirements of international standards.

System Functionality

The services to be provided by the system were defined on the basis of the results of a user needs analysis performed by the project. A main requirement was to support interoperability over distributed, heterogeneous digital collections and services. Achieving interoperability in the digital library setting

is facilitated by conformance to an open architecture as well as agreement on items such as formats, data types and metadata conventions. These issues have already been addressed with varying degrees of success by digital libraries handling textual collections; the challenge in ECHO was to solve the numerous technical problems that up until now have obstacled the inclusion of audio-visual material in a searchable digital environment. The aim has been to make the film collections available and searcheable to as broad a range of users as possible. To achieve this goal, the following components were developed and included in the system.

Audio/Video Metadata Model

When the project began, there were no well-defined metadata models for an adequate description of film data. A major effort of the project has been to define a suitable metadata model to represent the audiovisual contents of the archive. The model that has been implemented is an adaptation of the IFLA model, a general conceptual framework used to describe heterogeneous digital media resources. The model is composed of four levels describing different aspects of intellectual endeavour: work, expression, manifestation and item. The entities of the model are hierarchically ordered from the top level (work) to the bottom (item).

Intelligent Access

The ECHO system assists the user during the indexing and retrieval of A/V documentaries. Semiautomatic indexing is supported: the system automatically extracts a number of metadata information such as the scenes composing the video, key-frames that describe each scene, image features describing each key frame, spoken dialog (automatically transformed into text through a speech recognition process), faces, specific objects. Later on, the user can complete the indexing by specifying metadata that cannot be automatically extracted. Search and retrieval via desktop computer and wide area networks is performed by expressing queries on the audio transcript, on the metadata, or by image similarity retrieval. Retrieved films or their abstracts are then presented to the user. By the collaborative interaction of image,

speech and natural language understanding technology, the system compensates for problems of interpretation and search that arise when handling the error-full and ambiguous data sets.

Multilingual User Interface

The ECHO film archives are made up of language dependent (speech, text) and independent (video) media. Thus although users querying over collections in different languages may not understand the spoken dialog, they can still identify useful documents (or parts of documents) via the images. This has facilitated the implementation of a relatively simple multilingual search interface that can still provide useful functionality. The approach adopted has been to implement on-line cross-language search tools based on the use of standard metadata formats and mechanisms that provide a mapping between controlled vocabularies agreed between the content providers. Access is provided by local site interfaces in the local languages but a common user interface in English is also maintained on the project Website for external access.

Creating Visual Summarization

The project has developed techniques to produce visual summaries. The aim is to capture the content and structure of the underlying documentary film in a brief visual abstracting process. The summary consists of a sequence of moving images, much shorter than the original film, but preserving the essence of the original message. It should provide a good overview of the entire film documentary. The creation of visual summaries is based on the use of a number of video features, such as the different scenes recognized, faces, text, objects, and action scenes detected. After this initial analysis, the more relevant clips are determined and assembled in order to maintain the flow of the story. The abstract is usually set to 8\% of the length of the original video, but other values can be specified, depending on user and application needs. The video abstracting process is performed off-line, after video archiving, since it requires approximately 10 times the video duration.

Security

In order to make a digital library of films possible, the copyright owners must be guaranteed that their property will be protected and that its use will be measured in order to provide them with appropriate compensation. ECHO thus includes mechanisms which support access control, authentication, security and privacy

.

Analyzing user's needs

The list of new features included in ECHO has been defined by taking into account the functionality provided by existing commercial systems, including Informedia and Media Archive. In addition, the experience of ECHO content providers was taken into consideration. However, given the unique characteristics of ECHO's audio-visual material and the specific characteristics of potential users of ECHO, we concluded that it was important to collect user requirements for the ECHO system.

In order to collect the user requirements we have prepared an extensive questionnaire that has been submitted to approximately 15 users per participating country (giving a total of 58 users). The primary purpose of the questionnaire was to capture major functional aspects of ECHO, resulting in the user requirements list. Moreover, the questionnaire was constructed extensively enough to give as good an idea as possible on what the interviewees think about using a digital video archive like ECHO.

Three different user types were identified per target market or user community. Each identified user type was expected to have different interests in the ECHO system and to use ECHO in different ways. Furthermore, the user types identified had different levels and special focuses of experience in searching traditional film archives catalogues.

The selection of interviewees was very thorough; it represented the four national archives and the customers of these national archives. Experts on indexing video, on information retrieval, on film making and film history, and on scientific historical research were chosen because they were expected to provide interesting comments on the functionality of ECHO and because they were identified as potential future users of ECHO.

The interviewees themselves indicated their expected future use of ECHO in the questionnaire as follows

- research on historical film material
- browsing historical film material
- reuse of historical film material in new AV products
- data entry in ECHO

- reuse of historical film material for teaching

ECHO can be used for different purposes. The project team distinguished the following types of usage:

- Data entry for archiving
- Retrieval for reference searching, without interest in viewing the content
- Retrieval and viewing the content for the general public
- Professional retrieval and viewing the content for scientific study (history or film science)
- Professional retrieval, viewing and (non-)profit reuse of the content. (including general educational usage

Each type of usage refers to a specific goal for which someone might want to use ECHO.

To actually collect the user requirements, the questionnaire was submitted to users and specific support was provided during the activity.

Given that users could be split in two broad categories, the first including users who were professionally concerned with the database creation process and the second with those concerned with database consultation, two questionnaire were prepared.

The questionnaire was analysed using a numerical approach. The method of analysis was carried out in different steps.

1 Quantitative aspects: thorough analysis of the spreading of the percentages over the answers. The contact persons had the strong impression that the choice in answering *Yes, mandatory* or *Yes, desirable* depended more on the character of the interviewee than on the actual necessity of the functionality concerned. It was therefore decided to analyse these answers together.

2 Qualitative aspects: in the case of a rather equal spreading of the percentages over the answers or when a relatively large number of interviewees answered *No* or *Yes*, optional to a functionality that is

generally regarded as essential to a digital video-archive, the following was taken into account:

The macro-level of the user requirements could be obtained, while:

- the answers to the open questions that were added, to give as good an idea as possible on what the interviewees think about using a digital video archive like ECHO, were not always suitable for transformation into directly related user requirements
- those questions in the questionnaire which asked for a detail of the same major functional aspects of ECHO were analysed together
- the questions concerning the content (category 6) were left out of the analysis for this user requirements report. This category is analysed in Workpackage 2: Content Selection.
- only those questions referring to automatically created metadata were analysed. All other questions that might have an impact on the metadata model of ECHO, concerning manual formal and subject indexing, were not analysed for the general user requirements list of ECHO. The results concerning the manual formal and subject indexing will be discussed and decided upon in the context of Workpackage 3 on Metadata.

The ECHO audio/video collection

The four participating National Archives house enormous collections of audiovisual material starting from the first experiments in the end of the nineteenth century till the present. The origin of the material is diverse and includes cinema newsreels, television programmes, commercials, amateur film etc. etc. These collections keep growing rapidly as audiovisual production increases.

The characteristics of the collections owned by four ECHO archives are briefly listed in the following:

- **Istituto Luce**.

The archive of the Istituto Luce conserves the vastest historical audio-visual memory of Italy, from the first film clips at the time of the Lumière Brothers to the documentaries of the 80's. History, entertainment, customs: a cross-section of their most recent past documented by various cinematographers and news agencies. The most significant part of this patrimony is available on-line. Seventy years of history have been catalogued and computerised and are today available to whoever feels the need to examine their past. The entire audio-visual patrimony of the Istituto Luce contains 3,500 hours of film to which must be added a million photographic negatives. These cover the period between 1916-1960. Within the cinematographic sector, about 3,000 hours have been digitised and computerised. The Istituto Luce archives films that are mainly documentaries and newsreels, mainly in black and white, and from 1932 with sound. The productions with sound were made between 1925 and 1977 for the big screen and not for television.

- **INA**

Created in 1975, INA is the legal guardian of the French audio-visual heritage of 3 millions of radio and television programmes of the public and private broadcasters in France.

The collections go back to 1949 for television programs, 1940 for the news (35mm cinema newsreels) and 1933 for radio archives.

The material include discs and tapes for radio assets and films (mainly 16mm), tapes and cassettes for television recordings (from 2 inches to ¾ inch,

at various standards analogous or digital).Those archival materials are kept, restored, distributed for professional uses : re-broadcasting and new broadcasting productions.
Since 1995, the public service mission is extended, making INA, the radio and television Legal Deposit. This mission consists in managing the collections allowing scientific use for researchers and graduate students.

- **Memoriav**

For different reasons (high costs, federal tradition) it was not possible to create a national audiovisual archive in Switzerland. Therefore subsequent considerations led to the foundation of the Association for the preservation of the audiovisual heritage of Switzerland, known as Memoriav, at the end of 1995. The essential concept of Memoriav is creating and sponsoring a network of all resources in Switzerland which are already engaged in different fields in the task of rescuing and conserving audiovisual cultural heritage.
The most ambitious project of Memoriav is the digitisation of reports on Swiss politics in films and TV programs.
One important part of this project is the realisation of the accessibility to the Swiss cinema newsreels (Schweizer Filmwochenschau) and to the TV-News (1954-1989), which have been re-archived by a team of television documentalists.

- **NAA**

The Dutch Audiovisual Archive was founded in 1996 when the four major keepers of Dutch audiovisual Cultural Heritage merged. The NAA houses the material of the public broadcasters, documentary films, the National Music Depot, The Broadcast Museum, The Smalfilmmuseum, Commercials, a research department and other audiovisual material dating from 1898 to the present day.
In total the collection of the NAA consists of more than 600.000 hours of radio, television, film and music. The usergroup of the NAA is diverse and ranges from producers, publishers and scholars to families that visit the museum.
The NAA searched for original documents such as Dutch Newsreels from the period 1920 to 1980 and for films and documentaries made for cinema as well as television broadcasts by Dutch directors. It's also interesting to show documentaries and items produced for current affairs programmes in the 60's

– 90's, which refer back to certain developments in European society and to European main events in the twentieth century. In order to have a reality look at the European daily life the NAA, in collaboration with the Filmarchive Smalfilmmuseum, will select a number of amateur films
NAA is accessible on the internet: http://www.naa.nl (unfortunately only in Dutch) and http://www.smalfilmmuseum.nl .

Taking their catalogue as starting point, a set of themes that could be relevant for the project has been selected. The final theme framework consists of 5 themes and 28 subthemes. The ECHO digital collection is composed of 200 hours of documentaries belonging to all these themes and subthemes, taken from the four archives.

THEME 1 **Post War**	The second World War had a devastating effect on every aspect of the society and international affairs. It also revived the wish for a economic community. Apart from these matters directly linked to WW II, the ECHO collection will also include developments that indirectly originated from the war, like the consumer society and the emigration movements.
European Communities	Films documenting the founding of the various organisations prior to the EC. Successively: Marshall Plan, European Coal and Steel Community, European Economic Community, European Community.
Continuing Life in the City	After the war the cities and the infrastructure were rebuilt and the economic, environmental, cultural and social life were slowly normalised. See also 1.6 Society.
Emigration Movements	Films concerning various forms of emigration: Post War exodus, political (human rights), economical (from Third World to Europe), humanitarian (War, Third World famine), etc. Immigration also falls under this heading.

Rebuilding the Military Forces	Films on NATO, dissuasion and nuclear weapons, etc.
Cold War and International Relationships	Euro-missile crisis (1977-1984), Afghanistan (1973-1990), Cold War 1945-1963 (Berlin Crisis 1958-1962, Cuban Missile Crisis 1962), Nicaragua (1978-1990), Nuclear non-proliferation policy (1945-1991), etc.
(Changes in) Society	This subtheme concerns matters of labour, feminism, demography, poverty, consumer society, and technical inventions.

THEME 2 The World Wars	The World War section ECHO collection will include material from the 1920's to 1945.
Aftermath World War 1	The first World War had serious consequences on the developments of the already deteriorating international relations and, in particular, some events determined the somehow predicted economic and political crisis at the end of the 1930's.
1920-1945 Major Events	The fundamental events in those twenty years that influenced the political developments in the second world conflict. Some important topics: Mussolini at power, Stalin at power, Hitler at power, Civil war in Spain, Monaco Conference .
1920-1945 Propaganda	Provides material of how not only dictatorships but also democratic countries used audiovisual material to support their political beliefs among the people
1920-1945 International Relationships	The international relations between the most important European countries: official visits, diplomatic meetings, etc., selected to underline the new developments in the international alliances and the changes from the old system.

1920-1945 Socio-economic Factors	A documentation on the economic transformations which caused Europe's economic quick recovery through rearmament after the crisis at the end of the first World War. Some important topics: The economic crisis after WWI, The 1929 crisis in the USA and Europe, Rearmament.
1939-1945 The Development of the Second World War	Some selected events, such as: Occupation of Poland, Drôle de Guerre, Occupation of Paris, War in the USSR, Stalingrad, Fascist defeat, Normandy, Occupation of Berlin

THEME 3 Sports in the 20th Century	The importance of sports in European societies has developed during the 20th century. The goal is not only to show nice motion pictures of different kinds of sport but also to give an impression of the social and national meanings of sports in Europe.
Sociological Developments	Films showing the (changing) importance of sports for a society or for a social group like youth, women, working class etc.
National Sports	Films on sport with a large national identification, sometimes even for people who aren't interested in sports at all; often practised by a national team. E.g. ski-races in Switzerland.
Mass-Events	Films on sports events with a large audience like Olympic Games, World Championships or famous bike races.
European Contests	Films on European Championships in different sports.
Sponsoring	Films showing the growing importance of money in sports and the discussions about it.
Vandalism	Films on the ever growing violence around sports

events.

THEME 4 Daily Life	Daily life is a broad theme, including matters like: work, education, food, drink and health matters.
Work and Leisure	Films dealing with the production of goods and services and the time available for ease and relaxation following these activities. Labour movements is also placed under this heading. (this subtheme includes religious feasts, folklore and holidays)
The European Family	Films on parents and their children in the 20th century. This is also a very broad subtheme, ranging from housing, marriage and funerals to more demographic matters.
Education	Films concerned with teaching, pedagogy, upbringing of children and government legislation will be assembled here.
Food and Drink	Films on the production, distribution and possible lack of food in Europe. European legal matters and legislations will also be discussed here.
Sickness and Health	Films dealing with various health matters, ranging from hospitals, the work of relief organisations to more scientific subjects.

THEME 5 (Youth) Culture in Europe	The theme Culture not only includes different art forms, like literature, film, painting. It also includes the system of socially acquired values, beliefs, and rules of conduct in any given society. This system is created by lifestyle, television and forms of personal expression.
Fashion, Clothing, Lifestyle	Films concerning the manner of living that reflects the people's values and attitudes, expressed in their

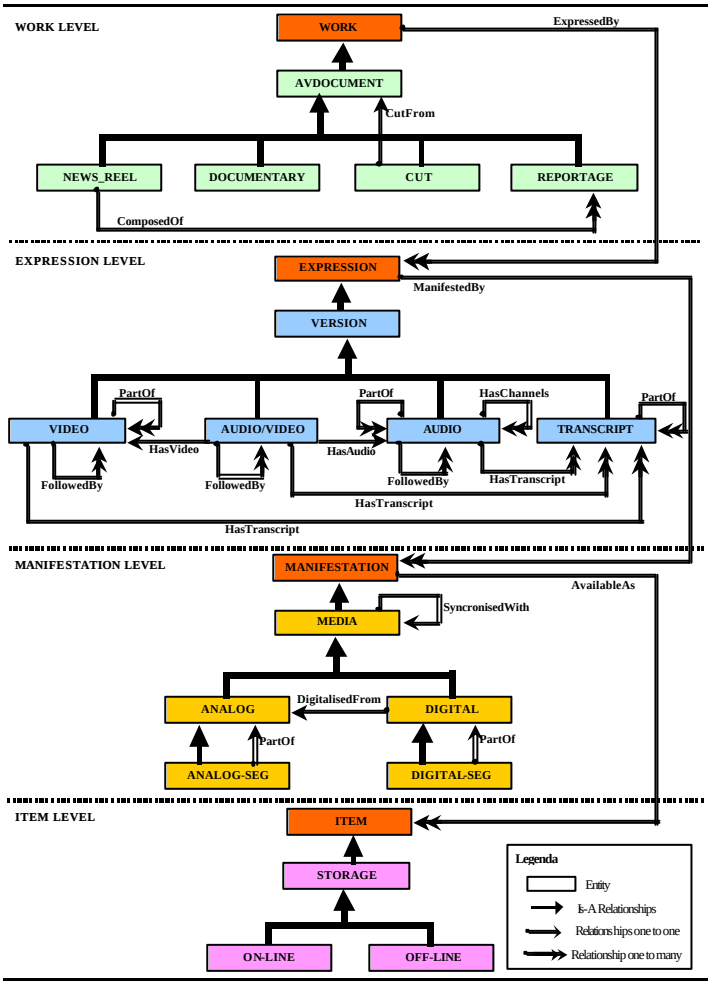
	appearance.
Student Revolts	Revolts started by students against the ruling classes and their school board.
Sexual Revolution	The alteration of the right or good conduct concerning matters of sex and relationships.
Television and Radio	How the technique behind film and radio was invented and how it was introduced to the public. Films about television personalities, legendary broadcasts and the social impact of these modes of mass communication .
The Arts	This very broad subtheme is filled with various products of human creativity ranging from literature and cinema to painting, theatre and music.

A metadata model for A/V documentaries

The ECHO audio/video metadata model extends the IFLA-FRBR model in order to support the description of audio visual documents and to cope with the specific requirements of audio/video digital libraries.
IFLA-FRBR suggests to describe resources using four different entities: (1) *Work* entity to describe the abstract idea of a resource, (2) *Expression* entity to describe different versions of the same work, (3) *Manifestation* entity to describe different physical supports where an expression can be embodied, (4) *Item* entity to describe different copies of the same manifestation.

The ECHO metadata model includes new sub entities for each of the previous IFLA-FRBR entity to offer more specialised mechanism to describe audio visual documents. The overall schema of the metadata model is sketched in the Figure.
The Work entity was extended by defining the sub entity *AVDocument* that contain attributes specialised for describing abstract ideas of audio visual documents as, for example, Director, Event, Date, Person, Location, Description. Some of these attributes can be defined in such a way that cross-language search can be performed effectively. For instance, a controlled vocabulary can be used where terms are chosen in such a way that a translation exists in all supported languages.
A specific version of a work can have several forms. For instance, it can be a silent movie, an audio stream, a black and white movie, the Italian translated version, etc. To this aim, the *Expression* entity was extended by defining the *Version* entity, which contains specialised attributes like VersionTitle, Duration, etc. The *Version* entity was also extended by other entities specialised to describe specific versions: the *Video* entity, corresponding to a silent video expression, *Audio*/*Video* entity, corresponding to an audio/video expression, *Audio* entity, corresponding to an audio only expression, and *Transcript* entity, corresponding to the word spoken in an audio stream.

It is also important to describe separately relevant portions of the entire expression as, for instance, sequences and shots. Scenes and shots of a video are represented by the *Video* entity itself (since portions are just videos). *PartOf* relationships relate composite videos and video components. The



The metadata model schema

same technique is used to represents portions of Audio/Video, Audio and Transcript portions. Other appropriate relationships model the possible links among these expressions such as for example *FollowedBy, HasAudio*, *HasVideo, HasChannel* and *HasTranscript*. Using these entities and relationships, it is possible to separately describe audio streams and corresponding video streams. It is possible, for instance, to separately keep track of audio streams produced in different languages, associated with the same video stream, and their corresponding transcripts. It is also possible to describe the fact that the same sequence (for instance a video sequence) was used in different documents.

Expressions are associated with features that are extracted from them in order to support similarity retrieval. Typically, these features are extracted automatically by means of image, audio, and text processing. Examples of features for visual documents are colour histograms, shapes, motion vectors, etc.

Audio video documents can be stored on analog and digital supports. Examples of analog supports are VHS and Betacam, while examples of digital supports are DVD and MPEG files. This was modelled by extending the *Manifestation* entity with the *Media* entity that was further refined by means of the *Digital* and *Analog* sub entities.

In order to relate expressions, corresponding to relevant portions of entire documents to their physical position in the corresponding manifestations, *Digital* and *Analog* sub entities were further refined by defining the sub entities *Analog-Sequence* and *Digital-Sequence*. A relation *SinchronisedWith*, included in the *Media* entity, permits to relate two manifestations that may be played synchronously as, for instance, a movie and its soundtrack, in the case they are stored separately.

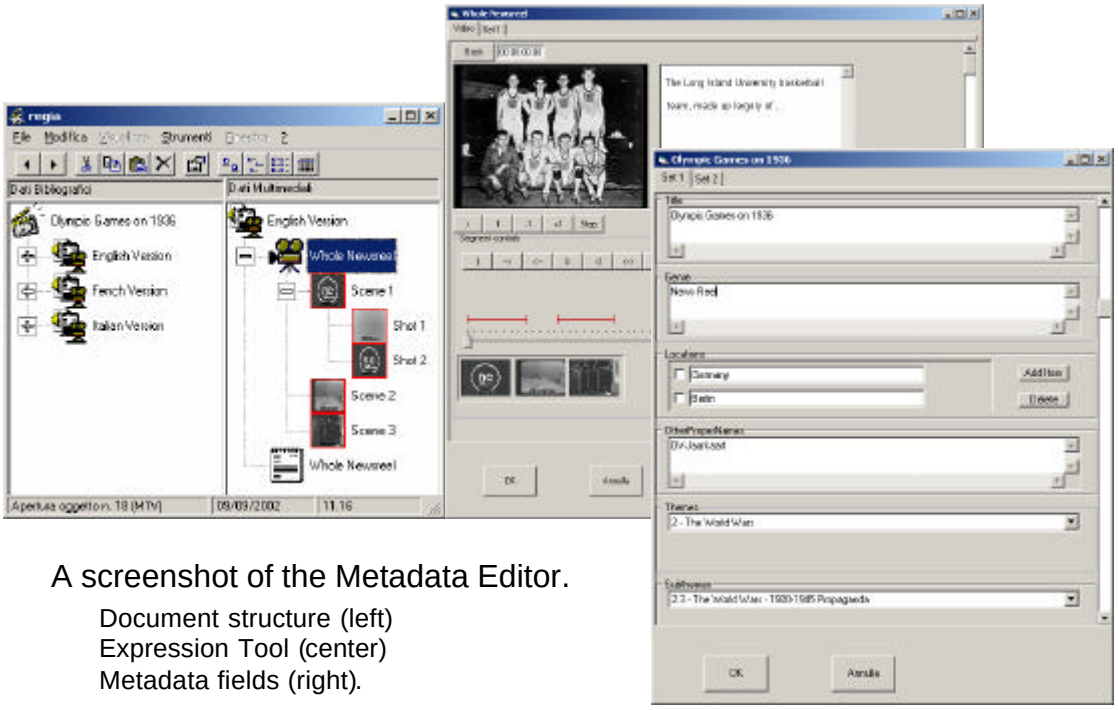
A library may have several copies of the same manifestation. For instance the same MPEG file may be stored on several video servers, or the library may have available several copies of the same VHS tape. To represent this situation, the *Item* entity was refined by the *Storage* entity, containing attributes for the right access management and ownership, which was further refined by *On-Line* and *Off-Line* entities. The first represents and describes resources that can be accessed directly by a computer, the second, resources

that are stored, for instance, in some shelf of the library. In the first case, it is possible to describe that on-line resources may have different access speed, and the systems where they are stored have different workload. In the second case, it is possible to describe the quality of preservation of different copies.
In designing this metadata model great care was taken to the provision for interoperability. We are currently experimenting the mapping of this model into other metadata models to evaluate its interoperability level. Another aim in designing the model was to render it extensible. This is a key features of any metadata model since it permits a successive expansion to cover new emerging needs.
The current version of the model does not cover yet the typical requirements for an efficient “right approach”. Future work will inquire at what degree of detail of rights description a content information system should supply and how: by providing a right issue specialization, or by activating a systematic flux with a right management system.

The Metadata Editor

The Metadata Editor is client tool used by the cataloguer of the archive for editing the metadata automatically inserted by the indexer. The typical cataloguer workflow is the following:

1. A new audiovisual document is digitalized or transformed from one digital format into another;
2. The document is archived by the system in the Essence Manager;
3. The document is processed for automatic indexing (extraction of scene cuts, speech recognition, etc.);
4. When the automatic indexing has been completed, the user is informed by



A screenshot of the Metadata Editor.

Document structure (left)
Expression Tool (center)
Metadata fields (right).

the system and the manual indexing can start;

5. The user typically edits the textual description for typos or factual content, reviews or sets values of the metadata fields, adjusts the bounds of the document segments, removes unwonted segments and merges multiple documents. This phase is usually performed starting from the top level of the model (the AVDocument), and continuing by modifying/editing the lower-level objects connected to the AVDocument (i.e., Version, Media and Storage objects).

Since the metadata model is relatively complex, the design of the metadata editor is of primary importance.

The interface of the editor is designed in such a way that it is possible to browse the tree structure of an audiovideo document. The Figure shows a screenshot of the interface: the window on left side displays a document like a folder navigation tool. On the top level of the tree, there is an icon representing an AVDocument object (the work of the "Olympic Games on 1936" in our example). Connected to the work object the editor presents all the entity instance of the document.

The navigation tool on the left side of the window shows only the main expressions belonging to the documents. The editor allows to browse a single Version one at a time by using a second frame on the right side of the window. In this way it is possible to see the possible Video, Audio and Transcript Versions (at least one of them must exist) of the document and, for each Version, to browse the video segmentations in scenes, shots, etc.

By clicking on the icon corresponding to a metadata object, it is possible to modify, in a separated window, the metadata fields of the object. A particular attention has been paid to the expression window design, i.e, the Expression Tool. The Figure gives an example of the Expression tool interface. Besides the textual fields, the Expression Tool allows the access to the metadata relative to the video segmentation, and allows one to modify them. More precisely, the user can view the video, and read the transcript. The window shows also an overview of the video segmentation, by means of a timeline tools (see the bottom of the Expression Tool window), which represent the

video and a the list of the keyframes of the video segments (scene, shots, etc.). The timeline is subdivided in partition that represent the media this segmentation. By selecting a segment (or the corresponding keyframe), the Expression Tool shows the video corresponding to the segment.

The ECHO digital library system

All the functionalities of the ECHO have been integrated into a digital library system. This system bases upon the media archive® Enterprise Asset Management System. This is a fully distributed archiving system running under WindowsNT and 2000 and relying upon CORBA as middleware.
The base system provides a separate management of audiovisual assets (i.e. audio, video files) and the metadata related to them using SQL databases. Basic functionalities are the ingest and the indexing of a video clips and a full-text search upon the archive's metadata. The system provides client applications e.g. to import video clips and to annotate them. A web based interface provides search and browse functions to access the content.

Modules that have been developed in the ECHO project have been integrated seamlessly into the existing framework. The hierarchical ECHO metadata model has been integrated without any loss of the full-text search performance. Special import tools have been designed to import video assets and metadata into the system.
The object and the face detection modules are integrated as extension of the standard video analysis. The metadata of these are imported into the ECHO metadata model. The video abstracting module works as an extension to the video conversion service – it converts not the format but the length of the video clip. The audio transcripting in Dutch; French German or Italian works remotely using distant speech recognition engines.

Two extra search engine have been integrated
1. the cross-language retrieval provided by EIT and
2. the Image Retrieval Engine provided by CMU and IEI.

The web service has been extended to show all the extended information and to provide convenient navigation facilities. The web interface itself is multilingual it can be configured for 4 languages.

As a result of the project there are two prototypes available now:

1. a commercial prototype with all standard functionality of the ECHO project, such as metadata management, video browsing etc. and
2. an experimental prototype demonstrating extended features such as face detection and image search.

Speech recognition

Large vocabulary continuous speech recognition is a key technology that can be used to enable content-based information access in audio and video archives. State-of-the-art speech recognition technology was used to produce automatic transcriptions of the historical documents for the French, Italian and Dutch languages, provided by the Echo partners.

The archives consist primarily of radio and television broadcasts and are challenging to transcribe as they contain signal segments of various acoustic and linguistic nature. Two main problems needed to be addressed to deal with the continuous flow of inhomogenous data. These concern the varied acoustic nature of the signal (signal quality, environmental and transmission noise, music) and different linguistic styles (prepared and spontaneous speech on a wide range of topics, spoken by a large variety of speakers).

The archive documents are different from contemporary broadcast news, which raises several issues for processing them: they are often noisier, the speaking style has evolved, there are epoch-specific proper names that need to be added to the lexicon, and the range of topics is very broad compared to contemporary news.

Training the statistical models of the transcription system requires a large amount of acoustic and text data, representative of the documents to be processed and in electronic format. Finding a sufficient amount of data for this adaptation, especially electronic texts about historical periods, is one of the biggest challenges.

The LIMSI broadcast news transcription system was used to transcribe the French part of the corpus, spanning the latter half of the 20th century. The word error rate with the system designed to transcribe recent data was about 42%. By adapting the models to the acoustic conditions of the Echo data and using additional texts from a French video archive web site (documentary

summaries covering the period from 1945 to 1979) reduces the word error rate by about 20% relative.

The ITC-irst broadcast news transcription system was initially applied to transcribe the Italian part of the corpus, consisting of films produced between '30s and '60s of the 20th century. Due to an evident acoustic and linguistic mismatch between the broadcast news and the new application domain, the word error rate was about 46%. By exploiting a limited amount of manually annotated training data from the ECHO domain, adaptation of all components of the transcription system was performed, namely the audio partitioner, the acoustic model, and the language model. The improved system achieved a word error rate of about 30%.

Twente's broadcast news recognition system was applied to he Dutch part of the ECHO corpus, which contains films from the period 1930-1980. Due to the extremely low audio quality, in addition to the linguistic mismatch, initial recognition figures were poor, especially for the older material. Efforts to obtain text data for training language models for outmoded speech by scanning available paper transcripts of some of the material did not succeed due to the low quality of the paper texts (carbon copies), and not enough data was available to apply clustering techniques to train domain specific acoustic models. The focus was therefore put on improvement of the contemporary language models. This resulted in word error rates between between 30% and 40% for more recent ECHO material.

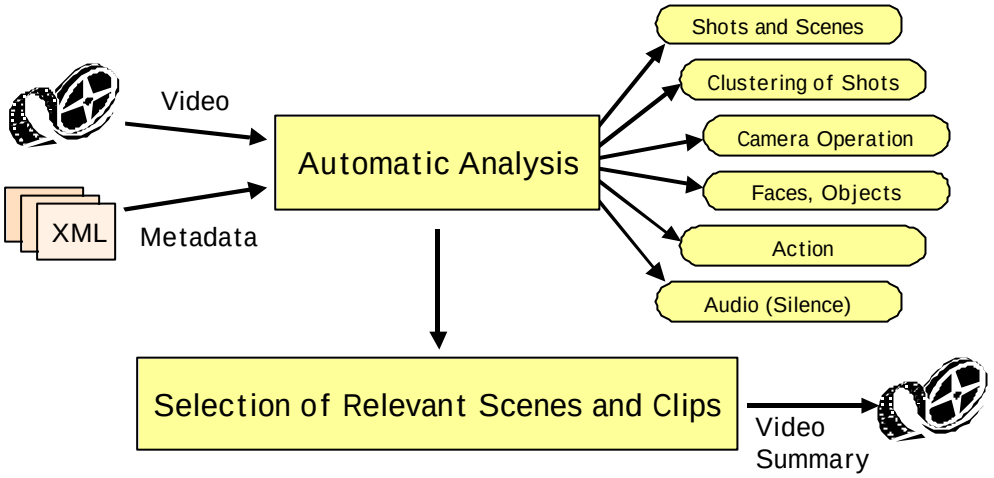
The automatic transcriptions have word error rates in the range of 30 to 40% relative to detailed manual transcriptions in the three languages. However, many of the errors are on function words or are due to lack of agreement (number and gender) which are not important for information retrieval purposes. As demonstrated by document retrieval experiments, the transcriptions are of sufficient accuracy to locate relevant audio segments.

Video summarization

The abstract of an article is a short document summary often used to preselect material relevant to the user. In the age of multimedia it would be desirable to use as video abstracts short clips that contain the essence of a longer video, without any break in the presentation medium. A video abstract or summary is a sequence of moving images that have been extracted from a longer video. Much shorter than the original, the abstract nonetheless preserves its message.

The video summarization application is a component of the ECHO system developed at the University of Mannheim. A summary is created in two steps. The first encompasses the automatic analysis of the video and the calculation of relevant features. If the data of a specific feature are available in the ECHO database, the application uses the information available. The following image depicts the summarization process.

Video summarization



Automatic analysis includes the decomposition of the input video into semantic units (shots and scenes). Semantically rich pieces of information such as frontal faces, moving objects or action scenes are calculated as well.

In the second step, the most important clips are selected. These are assembled into their final form, then the corresponding audio parts are selected. This summary is stored as an MPEG video. Users of the ECHO system may specify the length of the abstract

Indexing of video key frames

The ECHO prototype incorporates the image similarity search engine developed by CMU for the Informedia project. The image similarity search operates only on key frames that have been extracted during the video ingestion.

Image retrieval is supported by two distinct modules: an Image Indexer and an Image Retrieval Server (Image Search Engine). The Image Indexer processes the key frames identified in video streams and extracts salient features (e.g. color histograms, textures, etc.). Extracted features, along with the key frame identifiers, are given to the Image Retrieval Server that organizes them appropriately in its access structures. Query execution is performed as follows:

- An image is given to the Image Indexer, which extracts salient features.
- Those features are given to the Image Retrieval Server that, using the access structures returns a list of key frames Ids, which are similar to the query image.

Image feature ranges from lower level colour and texture to higher level segmented objects. A feature extraction technique can be built on one or a combination of such image features. The following discusses three image feature extraction techniques utilizing colour and combination of colour and texture information. In general, the format of an image feature vector is an array of floating point values. The number of such values depends on the corresponding feature extraction technique. According to the experience from our empirical experiment and others' research, there is no single dominant feature extraction technique to cover all kinds of images. That is, colour feature is sufficient to support image similarity search for some images while texture only feature or the combination of colour and texture is sufficient for some others.

The similarity search process is based on the ability of measuring the similarity between two items. This is obtained by defining a *similarity function*. Of course, there is not a general and universal definition of this function, since

it depends on the specific application. For instance, the similarity function used to compare images will be different than that used to compare stock quotes. Similarity functions are typically defined by application experts depending on the aspect they are interested in.
The similarity function is typically considered as a black box provided by the application expert. General techniques developed for similarity search should not be tied to a particular similarity function. However, it might be required that similarity functions satisfy some specific properties.
Typical queries that can be expressed in order to retrieve multimedia objects are *range queries* and *k-NN queries*. They can be defined as follows:

Definition 5.1: *The range query q(r) requires the retrieval of all objects o Î M ½ sim(q,o) £ r.*

Definition 5.2: *The k-NN query q(k) requires the retrieval of a set L of k objects (L = (o_1, … , o_k)) such that L Í M and " o_i, o_j | (o_i Î M Ù Ï L) Ù o_j Î L then sim(o_i, q) sim(o_j, q).*

In ECHO we use an approach based on so-called *metric trees* which directly consider only pair-wise distances between objects to recursively partition the search space without considering positions of objects in a multidimensional space - dimensionality of space need not even be known.
Specifically, the following access structures have been used in ECHO

- **M-tree**. An M-tree can be viewed as a hierarchy of metric (ball) regions. A region is defined by a database object O_i and radius $r(O_i)$, which represents the maximum distance between O_i and any other object, including its region (if any), in the region of O_i. An M-tree is a multi-way-branching tree, thus each node can contain several object entries which are all members of a region centred around a *parent object*, O_p, stored in a higher-level node. Notice that the region of objects from the root is assumed to be the entire universe, because these objects do not have any actual parent object. Each entry is represented by the object's features and, in the case of non-leave entries, by their region radii which restrict minimum regions in which all descendant objects and/or regions can be found. For efficiency reasons,

child to parent object distances, computed during the tree construction phase, also form a part of the objects' entries.

- **D-Index**. Another interesting hashing approach for similarity search in metric spaces is *the D-Index*. It is a multilevel hash structure that takes advantage of the idea of the *excluded middle partitioning*. In each level of the hash structure a certain number of buckets are defined. Each level is associated with a hash function that, given an object, assigns it to either a specific bucket or no buckets in the corresponding level. Objects that are associated with no buckets in a level become candidate to be stored in the next level. It might happen that some objects are excluded to be stored in all levels. These remaining objects are stored in a separate *excluded bucket*. In this access structure, each object is stored in one bucket of the multilevel structure or in the excluded bucket. Queries can be processed, by accessing at most one bucket per level plus the excluded bucket. The upper bound on the number of disk access is limited by the number of levels, and the number of required distance computations can be significantly reduced by using some pre-computed distances. Distance computation between query objects and some accessed objects sometime is not performed since, by using pre-computed distances, it is possible to infer that they do not belong to the result set.

Furthermore, we have also integrated a search method that supports "approximate similarity search". Indeed, though the problem of similarity search seems to be very well defined and several storage structures have been proposed the current technology cannot certainly be considered stable. In fact, in some situations, performances of existing access methods are worse than sequential scan of the whole data set.

Given this inefficiency problem, the notion of *approximate similarity search* has emerged as important research issue. The basic idea behind the approximate similarity search is that queries are processed faster at the price of some imprecision in search results. In general, these approaches are motivated by the following two observations. First, a good data partitioning of many metric data sets is simply not possible, so resulting data regions have typically a high overlap, and many regions must be accessed to answer a

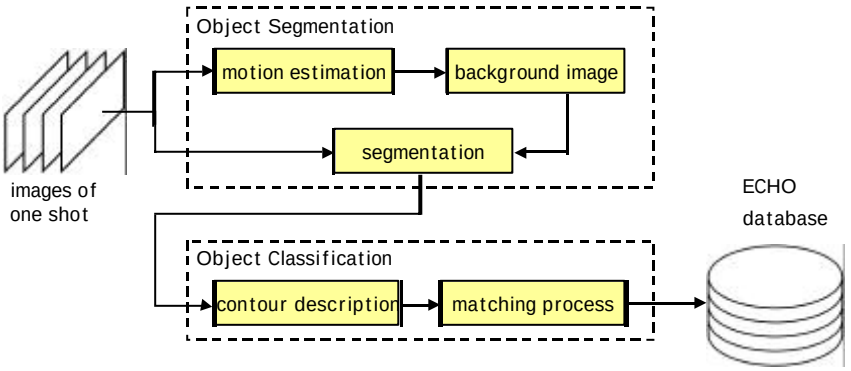
query. Second, similarity based search processes are intrinsically iterative. In many cases, users redefine queries depending on the results of previous retrievals. In this case, an efficient execution of elementary queries is of particular importance and users easily accept some imprecision, especially in the initial and intermediate search results, if much faster responses can be achieved.

Moving-Object recognition

Moving objects are segmented on the basis of motion cues. Each shot is analysed and the global camera motion (pan, tilt, zoom) calculated.

For each shot, a background panorama image is constructed. The foreground objects in this background image are removed by means of temporal filtering. The object is segmented by comparing each frame of the video to the background image (segmentation).

For each detected object, an efficient shape-based representation is calculated (contour description). This description is then matched to pre-calculated object descriptions. The background image and the recognized objects (cars) are depicted in the following image.

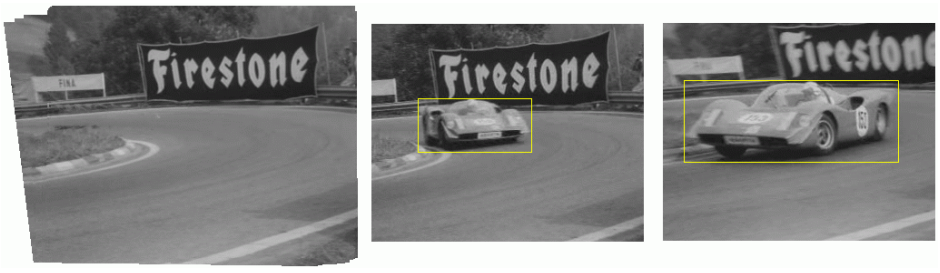


Moving object recognition

Moving objects are segmented on the basis of motion cues. Each shot is analysed and the global camera motion (pan, tilt, zoom) calculated.

For each shot, a background panorama image is constructed. The foreground objects in this background image are removed by means of temporal filtering. The object is segmented by comparing each frame of the video to the background image (segmentation).

For each detected object, an efficient shape-based representation is calculated (contour description). This description is then matched to pre-calculated object descriptions. The background image and the recognized objects (cars) are depicted in the following image.



Face detection and recognition

The main challenge in face detection is the amount of variation in visual appearance. Visual appearance also depends on the surrounding environment. Light sources will vary in their intensity, color, and location with respect to the face. Nearby objects may cast shadows on the face or reflect additional light on the face. The appearance of the face also depends on its pose; that is, its position and orientation with respect to the camera. For example, a side view of a human face will look much different than a frontal view. A face detector must accommodate all this variation and still distinguish the face from any other pattern that may occur in the visual world.

To cope with all this variation, a two-part strategy for fact detection is used. The first part is to use a view-based approach with multiple detectors to deal with variation in pose. Each of these detectors specializes to a specific face orientation. The second is to use statistical modelling within each of these detectors to account for the remaining variation.

View-based detectors

Separate detectors are developed independently. Each detector specializes to a specific orientation of the face. For example, one detector is specialized to the right profile views of faces and one is specialized to the frontal views. Applying these view-based detectors in parallel gives individual results, which are then combined. If there are multiple detections at the same or adjacent locations, the strongest detection will be chosen.
Our empirical experiments determined two orientations to model for faces. The two view-based detectors are frontal and right profile face. To detect left-profile faces, the right detector is applied to a mirror-reversed input image. Each of these detectors is not only specialized in orientation, but is trained to find the face only at a specified size within a rectangular image window. Therefore, to be able to detect the object at any position within an image, the detectors are re-applied for all possible positions of this rectangular window. Then to be able to detect the face at any size, the input image is iteratively

resized and the detectors are re-applied in the same fashion to each resized image.

Statistical form of detector

Within each view-based detector, statistical modelling is used to account for the remaining forms of variation. Each of these detectors share the same underlying statistical form. They differ only in that their models use statistical gathered from different sets of images.

Two statistical distributions are modelled within each view-based detector. The statistics of the given face is modelled as *P(image | face)* while the statistics of the rest of the visual world, called "nonface" class, is modelled as *P(image | nonface)*. Then the detection decision is computed using the likelihood ratio test:

$$\frac{P(image \mid face)}{P(image \mid nonface)} > \frac{P(nonface)}{P(face)}$$

If the likelihood ratio (the left side) is greater than the right side, the face is decided to be present. This is equivalent to Bayes decision rule or MAP decision rule and will be optimal if the representations for *P(image | face)* and *P(image | nonface)* are accurate. For the details of the representation of statistics, please refer to Henry Schneiderman and Takeo Kanade's work[1].

[1] Schneiderman, H. and Kanade, T. "A Statistical Method for 3D Object Detection Applied to Faces and Cars" IEEE Computer Vision and Pattern Recognition (CVPR), Hilton Head, SC, June 13-15, 2000, pp.746-751

Cross-language information retrieval on audio transcripts

The final ECHO prototype incorporates a text search, which supports search not only on metadata but basically on transcripts. The searchable item will be the entire transcript of a newsreel, the transcript of a single scene, respectively of a shot.

The transcripts will be derived out of the speech part of the newsreel and aligned to the scene-structure of the video. The transcripts will be in Italian, French or Dutch. The transcripts for Swiss newsreels in German language will be derived from written transcripts. As these transcripts are subjected to an OCR process, the quality of text and alignment to scenes and shots will be quite different to the other parts.

The Cross-Language Information Retrieval (CLIR) is Eurospider's latest cutting-edge technology and a additional modules of the standard RotondoSpider system. A module to retrieve relevant information simultaneously in several predefined languages – by typing in a query term only once in one language.

The method used for multilingual information retrieval allows the user to formulate the query in his/her preferred language, in order to retrieve relevant documents in any of the languages contained in a multilingual document collection. This is an extension to the classical Cross-Language Information Retrieval (CLIR) problem, where the user can retrieve documents in a language different from the one used for query formulation, but only one language at a time.

Our approach to multilingual information retrieval is corpus-based, using so-called document alignments. The alignment process associates documents that cover similar stories. This leads to an unidirectional mapping between texts in different collections.

The individual collections in different languages and the mapping given by the alignments together form a *multilingual comparable corpus*. A related resource would be a *parallel corpus*. In such a parallel corpus, the paired documents are not only similar, but high-quality manual translations. One of the benefits of the presented method is the excellent availability of comparable corpora as opposed to rare and expensive parallel corpora.

Using a derived comparable corpus, the system accomplishes multilingual IR by doing pseudo (local) feedback[2] on sets of aligned documents. We show how this process can easily be combined with a dictionary based approach to the translation problem.
Related work on alignment has been going on in the field of computational linguistics for a number of years. However, most of this work relies on parallel corpora and aligns at sentence or even word level. Usually, the amount of data processed is also smaller by several magnitudes.
Other corpus based approaches to cross-language IR have been proposed in the past, including the use of similarity thesauri. Our approach is based on document/document similarities, whereas the similarity thesaurus approach and LSI approach are based on term/term similarities, with the similarity thesaurus working in a dual space, and LSI using a low-dimensional vector space obtained by singular value decomposition.

In the ECHO prototype the a full text search will be performed on newsreels *transcripts*. The most adequate document entity in this collection will be the transcript of a single shot. The transcripts text is aligned to single shots by automatic shot detection. The audio is aligned to the detected shots and saved as text in the ECHO data repository, after automatic audio-text conversion. In other word the searchable item and also the aligned document for our multilingual information retrieval method will be the textual representation of a *single newsreel shot*.
As to the selection of the newsreel according to specific *themes* for ECHO, we will find certain alignment on shot level, although the chosen newsreel do not cover exactly the same theme.

The alignments are used for multilingual information retrieval by using a simple strategy: In case the two collections to be searched were parallel, i.e., real translations of each other, it would be possible just to search the

[2] Xu, J. and Croft, B. W.: Query Expansion Using Local and Global Document Analysis. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (1996) pages 4-11.

collection that corresponds to the language of the query, and return a result list produced by replacing every found document by its counterpart in the other collection. This is of course not a very interesting case, as the collection to be searched is seldom available in translated form. We therefore replace the requirement for a parallel corpus with one for a comparable corpus.

First, the user's query is run against the source collection, thus obtaining a ranked list. Instead of replacing the found documents by their translations, the document-level mapping produced by the alignment process is used to replace them with their most similar counterparts from the other collection, if available. This produces a new result list containing documents in the target language. Because a lot of documents are not part of an alignment pair, however, they would never be retrieved using this strategy.

This problem is approached by using a pseudo *relevance feedback* (in this case *local feedback*) loop *after* the replacement step (but *before* a search in the target language takes place) . A certain number of the highest ranked documents are *assumed* relevant and terms are extracted from these documents that are thought to represent them well. These terms form a query used for a new search. Because the documents are already in the target language, so is the query produced.

This simple strategy works surprisingly well for certain queries. It fails however if the initial query doesn't retrieve any relevant documents. This can be amended by combining the strategy with a dictionary translation of the query. The same wordlist used for producing the alignments can be used for a *crude* word-by-word translation of the query. Such translations normally have problems, even if the dictionary is of high quality. Translations are often very ambiguous, and including many extraneous wrong translations hurts retrieval performance a lot. Another problem is missing entries from the dictionary, because the query term is inflected or because even the base form itself is absent.

Therefore, simple word-by-word replacement with all possible translations usually doesn't perform well and is less appropriate than for computing alignments. Additional efforts are needed, like using part-of-speech, sense disambiguation, lexically correct word normalization etc. It is however easily possible to *combine* such a simple dictionary translation with the relevance

feedback process described above. Instead of deleting the original query before feedback, it can be replaced with such a "pseudo-translation". Because the terms get re-weighted in the feedback process, even the problem of assigning weights to the two sources of information (dictionary vs. alignments) is automatically taken care of. The alignments as additional source of information help lessen the mentioned translation problems.

In case the collection to be searched is not aligned, two independent aligned collections can be used to produce the query. The aligned collections are then only used for the transfer of the query into the target language, whereas the search takes place on the third collection.

System evaluation

The purpose of a software product evaluation is to provide any interested party with quantitative results concerning a software product that are comprehensible, acceptable and trustworthy. Existing standards define the features for ensuring software quality. For ECHO, the reference what ISO/IEC 9126, and this defines six features to be taken into account: functionality, reliability, usability, efficiency, maintainability and portability.

Some features concern the software development, others deal with the actual usage of the system. Below, we will go deeper into how the usability of the system was defined.

Measuring usability

We need 'usability' in order to achieve system acceptability. Acceptability is most usually thought of as consisting of a number of different factors, in particular the usefulness and ease of use of the system. A consequence of this is satisfactory use of the system and thus acceptance. However, it is important to note that satisfaction is not always a consequence of usability, but ought to be considered a factor in its own right.

These different elements can be defined as follows:

- **Usability:** The extent to which a product can achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use;
- **Effectiveness:** The accuracy and completeness with which users archive specified goals;
- **Efficiency:** The resources expended in relation to the accuracy and completeness with which users achieve goals;
- **Satisfaction:** the comfort and acceptability of use.

In the ECHO project, measuring points regarding these elements are embedded in different phases of the project. Three prototypes were delivered as part of the development approach. Each prototype extends the functionality

of the previous one(s). From an end-user point of view, the most interesting and advanced possibilities of the system were integrated exclusively in the third prototype.

Both the second and third prototype where subject to extensive user testing in two application domains: that of education and entertainment. The tests were conducted on an extensive corpus of audiovisual material, taken from the collections of INA, Sound and Vision, Memoriav and IL. The efforts provided useful feedback for adjustments and further developments of the system at one hand, and more general on the use of advanced searching possibilities on a wider scale.

Setting up the test

From an application point of view, the ECHO project developed a digital library service for historical films belonging to large, national, audiovisual archives. The system will allows different user groups to see an event that is recorded and documented in the country of origin and compare it with the way the same event has been documented in other countries.

Apart from the differentiating between different users that use the system, the different purposes for which the system is used also had to be taken into account. The user requirement report that was compiled in the early stages of the project distinguished between the following types of usage for which someone might want to use ECHO:

- Data entry for archiving;
- Retrieval for reference searching, without interest in viewing the content;
- Retrieval and viewing the content by the general public;
- Professional retrieval and viewing the content for scientific study -history or film science-;
- Professional; retrieval, viewing and -non--profit reuse of the content;
- Educational usage.

With this in mind, a test methodology was set up. In test methodology, there is a clear division between the methods that can be deployed and on which features of usability they are able to collect data from. There's a widely accepted and practised set of usability methods available, each with their specific advantages and disadvantages. After studying the possible testing methodologies, it was decided to use a mixture of conducing interviews, questionnaires and observing the users.

The next stage was to translate these ideas into a more practical programme of work. User-testing follows a more or less standard sequence, leading eventually to a final usability report. These are, consecutively: select users, design test, set up test, run test, analyse results and write the report.

After first three preparatory stages the actual test was conducted following a strict scenario. After a short introduction on the prototype for the whole group of users, users were asked to provide some statistic details, captured by means of a pre test questionnaire. The subject sits at a desk with a computer. Additional seating was provided in the room for the observer(s). The observers are concerned with observing interactions between the subject and the browser. They are responsible for capturing data throughout the test such as task time, error rate, verbal responses and heuristics. They record their findings on specially designed report forms.

The subject is asked to carry out a number of set tasks given to them by the test monitor. They are encouraged to talk through what they are doing during the test to help the observers gather information.

For these tests a number of tasks have been designed to bring into use functions associated with using the application. The aim is to probe the user interface to uncover potential usability problems. The set of tasks has been ordered in a way to expose the subject to the functions and features of the application and assess their responses.

- General task to guide through the system -simple and advanced-;
- Task on national collection -transcript-;
- Task on national collection -boolean search-;

- Task on Theme 4 -also advanced only-;
- Additional time to 'play' with the system.

Following the test the subject attends a post-test interview to provide a qualitative assessment on the use of the application. Questionnaires are very often used in usability testing. The questionnaire that needed to be to be filled out is a standard questionnaire for usability assessment in accordance with the ISONORM 9241/10 (developed by Prümper & Anft). Additionally, the subjects were asked some additional questions, in order to complete the measurement of utility and satisfaction of the system. This is a useful source of information and many of the subjects provided information that was not apparent through the task led testing or observation.

The interviews provide a lot of information on qualitative assessment. Finding out what was appealing and easy to use as not normally evident from task testing. Subjects commented more often on what they found difficult during tasks, rather than what was useful or enjoyable.

Analysing the results

All the information gathered was analysed and a usability report was compiled.

The objective –quantitative- data gathered from the questionnaires was entered into spreadsheets for tabulating and summarising the data. In the end report, these results are displayed as a whole and broken down for more detail into specific profile user groups. The more subjective data gathered from post-test interviews have been assessed in reference to the list of heuristics appearing earlier in this report. Satisfaction has been measured through the post-test interview and questionnaire.

The results and findings are arranged into their respective sections of Utility for performance measures, Usability for heuristic (human interaction) results and satisfaction for subjective reporting. A number of charts are provided with task achievement data and results from the subjective rating scale in the post-test interview.

Tests were carried out at all archive sites, and the results are most promising for the solution designed and build by the ECHO consortium. Users appreciate the possibility to be able to browse through different collections. For most of them, it was the first time to work with a digital asset management system. So, even the standard search option and display of results made a big impression. Add to this the breakthrough features, such as the ability to browse through the transcripts, launch content based image queries and the possibilities provided by multilingual access and it will be clear the users were impressed by the systems abilities. Users in both the entertainment and education domain saw ways how the use of ECHO could be an valuable asset to their work: respectively to reuse audiovisual sources inaccessible before and be able to access well annotated historical cultural heritage in a flexible way.

ECHO Contact points

Echo web site **http://pc-erato2.iei.pi.cnr.it/echo/**

Project coordinator: **Pasquale Savino**
ISTI-CNR
Via Moruzzi, 1
56124 Pisa (I)
e-mail: pasquale.savino@isti.cnr.it

Project partners

Consiglio Nazionale delle Ricerche	*Coordinator*	*I*
ITC-Irst	*Associate Partner*	*I*
INA	*Partner*	*F*
CNRS-IDF Sud – LIMSI	*Associate Partner*	*F*
Tecmath AG	*Partner*	*D*
University of Mannheim	*Associate Partner*	*D*
NAA	*Partner*	*NL*
CTIT/University of Twente	*Associate Partner*	*NL*
Istituto Luce	*Partner*	*I*
C.M.U.	*Partner*	*USA*
Memoriav	*Partner*	*CH*
EIT	*Partner*	*CH*

Information Society Technologies