Indexing and Retrieving documentary films: managing metadata in the ECHO System

Giuseppe Amato
ISTI-CNR
Via G.Moruzzi 1, 56124, Pisa, Italy
+39 050 3152906
G.Amato@iei.pi.cnr.it

Claudio Gennaro
ISTI-CNR
Via G.Moruzzi 1, 56124, Pisa, Italy
+39 050 3152888
C.Gennaro@iei.pi.cnr.it

Pasquale SAvino
ISTI-CNR
Via G.Moruzzi 1, 56124, Pisa, Italy
+39 050 3152888
P.Savino@iei.pi.cnr.it

ABSTRACT

Wide access to large information collections is of great potential importance in many aspects - economic, environmental, health, cultural, social, etc. - of everyday life. Historical video documentaries hold by national audiovisual archives, constitute one of the most precious - from a historical and cultural viewpoint - and less accessible cultural information.

This paper presents the approach adopted in the ECHO (European CHronicles On line) system, to provide an effective support for indexing and retrieval of historical documentary films A combination of automatic and manual indexing is adopted. The paper describes in detail the Metadata Editor that simplifies the task of manual indexing; metadata are represented in an XML form, thus enabling the extension of the model according to specific needs..

Categories and Subject Descriptors

H.3.1 [**Information Storage and retrieval**]: Content analysis and Indexing. H.3.7 [**Information Storage and retrieval**]: Digital libraries. H.5.1 [**Information Interfaces and presentation**]: Multimedia information systems.

General Terms

Management, Documentation, Performance, Design, Human Factors

Keywords

Audio/Video, Digital Library, Video Documentary, Information Retrieval, Metadata, Metadata Editor, XML schema

1. INTRODUCTION

Wide access to large information collections is of great potential importance in many aspects - economic, environmental, health, cultural, social, etc. - of everyday life. However, limitations in information and communication technologies have, so far, prevented the average person from taking much advantage of existing resources.

The ECHO project [1,2] aims at developing a Digital Library (DL) service for historical films belonging to large audiovisual 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 a user to search and access these documentary film collections. Users can be able, for example, to see an event which is documented in the country of origin and how the same event has been documented in other countries, or to investigate how different countries have documented a particular historical period of their life, etc.

This paper describes the ECHO system, looking at its functionality first and then considering the system architecture and the components used for A/V indexing and retrieval. All these functionality have been defined according to the requirements collected from a user needs analysis performed in the project. Particular emphasis is given to the combination of metadata automatically extracted from videos (e.g. audio transcripts, image features extracted from key frames, features extracted from faces and recognized objects) and metadata manually associated by the user through the use of a Metadata Editor. The editor supports a metadata model that allows one to describe all characteristics of a video object: its structural composition (e.g. scenes, shots, etc.), the information associated to its content (e.g. the audio transcripts, the description of the scenes, etc.), as well as its bibliographic data (e.g. the creation date, the author, etc.).

2. The ECHO system

The ECHO system assists the user during the indexing and retrieval of A/V documentaries. The indexing is semi-automatic. Using a high-quality speech recogniser, the sound track of each video source is converted to a textual transcript, with varying word error rates. The transcript is then stored in a full-text information retrieval system. Multiple speech recognition modules, for different European languages are included. Likewise, video and image analysis techniques are used for extracting visual features and segmenting video sequences by automatically locating boundaries of shots, scenes, and conversations. Metadata are then manually associated with film documentaries in order to complete their classification.

Search and retrieval via desktop computer and wide area networks is performed by expressing queries on the audio transcript, on metadata or by image similarity retrieval. Retrieved documentries or their abstract, are then presented to the user. By the collaborative interaction of image, speech and manually associated metadata, the system compensates for problems of interpretation and search in the error-full and ambiguous data sets. Exploration of the ECHO repository is based on these same techniques, allowing for spoken or typed natural language access to the information space.

2.1 Overall system architecture

The architecture of the ECHO system (Figure 1) is composed of three main components: *client interface*, *automatic processor* and *middleware*. The client interface is the component directly employed by the users to interact with the system. The automatic processor component analyses multimedia documents, to automatically extract metadata. The middleware component manages accesses to data stored in the video database and metadata database, on behalf of the other two components. In the following, we will give a more detailed description of each of these components.

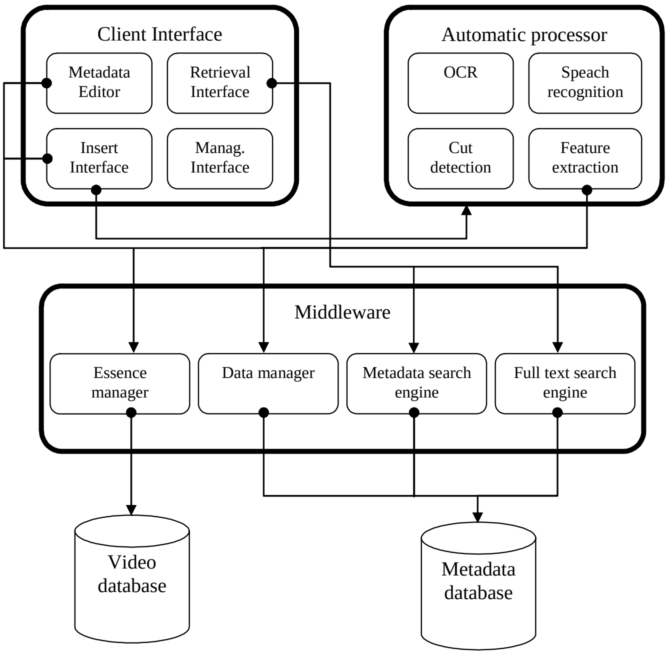


Figure 1: ECHO system architecture.

2.1.1 *Client interface*

The Client interface is composed of four main modules, related to the corresponding activities that can be carried out by users of the system. The *metadata editor* allows users to manually edit and review metadata associated with multimedia documents. The user can either edit automatically generated metadata, as for instance scene boundaries, or he can add additional metadata manually. The *insert interface* is used when new documents are inserted. This module interacts with the metadata editor and the automatic processor components that automatically analyse the documents being inserted. The *retrieval interface* is used to search the system for interesting documents. Various possibilities are offered by this interface. Users can retrieve documents by performing full text retrieval, on the transcript or descriptions associated with documents, or selecting specific fields of the metadata structure. Finally, the *management interface* can be used to configure and fine-tune the system.

2.1.2 *Automatic processor*

The Automatic processor is composed of four main modules, each one dedicated to a different automatic processing technique. The *OCR* module recognises textual video captions. The *speech recognition* module is able to generate a transcript in correspondence of an audio or audio/video document. The generated transcript is indexed and the corresponding document can be retrieved by performing full text retrieval. The *cut detection* module analyses a video document and automatically identifies scene changes. In this way, metadata can be associated with specific portions of the document, instead of the whole document. Finally, the *feature extraction* module analyses multimedia document in order to extract physical properties, which can be used to perform similarity retrieval. Typical features extracted are colour distribution, texture, shapes, and motion vectors [3].

2.1.3 *The middleware*

This component manages the accesses to the underlying databases: the video database, that physically stores video documents managed by the system, and the metadata database, where all metadata associated with the documents are stored. The middleware component is constituted of four modules. The *essence manager,* that handles the access to the video server, when other system modules, e.g. the metadata editor or the automatic processor, need to access the video server. The *data manager,* that handles the access to the metadata database. Main operations supported are *read, insert, delete,* and *modify* metadata fields. The *metadata search engine* supports document search based on the full metadata content. The *full text search engine* can be used to search for documents by using textual parts of the metadata. For instance, the descriptions and the transcripts associated with documents are used to perform this type of search. The result of the two search engines can be combined to allow users to perform an integrated complex search.

2.2 Editing Metadata

A fundamental aspect of the ECHO project is the metadata model [4] used for representing the audiovisual contents of the archive. The proposed model is based on the IFLA model, a general conceptual framework used to describe heterogeneous digital media resources [5]. This metadata model is composed of four levels describing different aspects of intellectual or artistic endeavour: *work, expression, manifestation,* and *item*. The entities of the model are organized in a structure that reflects the hierarchical order of the entities from the top level (*work*) to the bottom (*item*). Figure 2 shows a schematic representation of the ECHO Metadata Model [4]. As it is possible to see, the metadata which belong to different classes comprised in the model, are logically divided in two sets *Bibliographic Metadata* and *Time/Space related Metadata*. This classification is also reflected in the Metadata Editor interface.

The AVDocument entity is the most abstract one; it provides the general intellectual or artistic view of the document. For

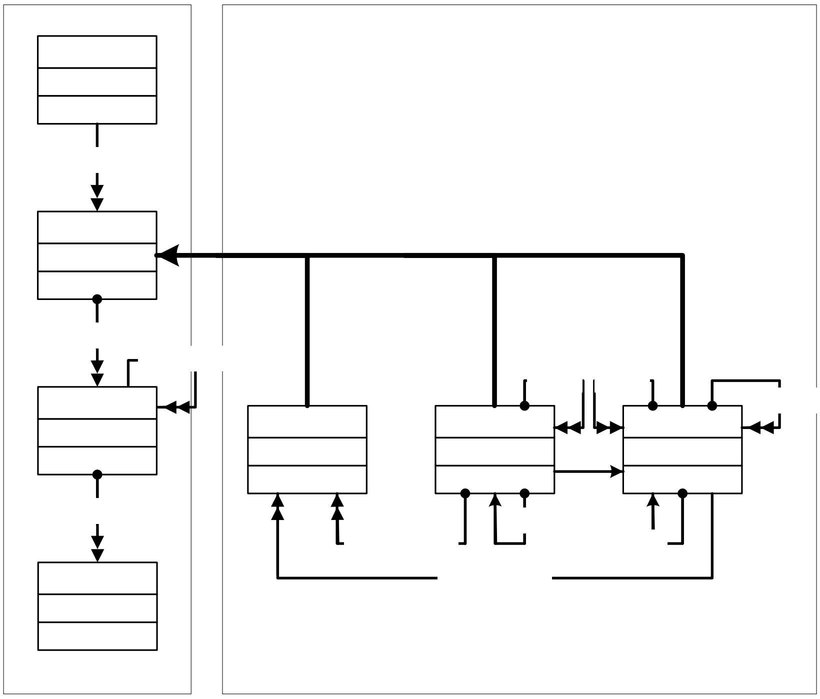


Figure 2: schematic representation of the Echo Metadata Model.

instance, let us suppose we want to describe a document about the Berlin Olympic Games in 1936. An AVDocument object will represent the abstract idea of the documentary film on the Games. A number of objects, of the abstract entity Version, could represent different language version of the same film, e.g., versions in Italian or in German. The Italian version could have three different objects that are specializations of the entity Version: a Video object, a Transcript object and an Audio object. Moreover, Version objects can be also part of other Version objects. For example, the Video object representing the Italian version of the whole documentary film can have other Video objects representing specific scenes of the film.

However, the Version entity does not represent any specific implementation of the film. This aspect can be represented by means of the manifestation level. For instance, a Media object could represent a digital realization of the document in MPEG format. More than one manifestation of the same Version, e.g. MPEG, AVI, etc., may exist.

Nevertheless, the Media object does not refer to any physical implementation. For instance, the MPEG version of the Italian version of the Games can be available on different physical supports, each one represented by a different Storage object (e.g., videoserver, DVD, etc).

Since the metadata model is relatively complex, the design of the metadata editor is of primary importance. The editor is intended to be used by the cataloguers of the archive, that insert new audiovisual documents and that specify the metadata of the documents. 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 videoserver;
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 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).

The interface of the editor is designed in such a way that it is possible to browse the tree structure of an audiovideo document. Figure 3 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 the three main Versions that belong to the AVDocument. Moreover, selecting an icon representing a Version (the Italian Version in Figure), it is possible to see the Media instances of the Version and, hence, the corresponding Storage objects.

The navigation tool on the left side of the window shows only the main expressions belonging to the documents (i.e., the

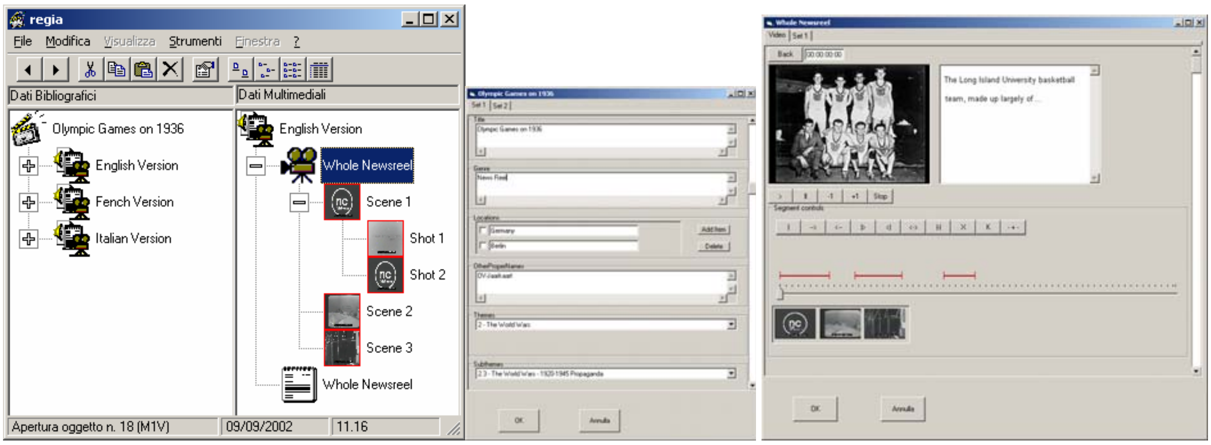


Figure 3: A screenshot of the Metadata Editor. Document structure (left), AVDocument metadata (center), and Expression editing tool (right).

expression which correspond to the whole audiovideo document). 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. Figure 3 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, hear the audio and read the transcript. The window shows also an overview of the video segmentation, by means of three slide tools (see the bottom of the Expression Tool window), which represent the video, the audio and the transcript (if any) of the whole expression. These slides are subdivided in partition that represent the media segmentation. By selecting a segment, the Expression Tool shows the Version corresponding to the subpart of the media (for instance, a scene or a shot).

2.2.1 The editor architecture

As shown in section 2.1, the metadata editor communicates with the middleware of the ECHO system in order to obtain the metadata of the A/V documents from the database. In particular, once the user has found a relevant document, by means of the video retrieval tool (see section 2.2.2), the URI (Uniform Resource Identifier) of the document is obtained and passed to the metadata editor. This URI is sent to the datamanager which returns the document metadata. Metadata are represented in an XML format. The metadata editor is not hard wired with a particular set of metadata attributes; indeed, the metadata schema is defined in the W3C XML Schema Definition (XSD) and it is used by the editor as a configuration file for the metadata model (see Figure 4). The advantage of this choice is that it is possible to add/remove fields in the schema of the metadata of the audiovisual document. This is achieved by giving the editor the ability of recognizing a subset of the types available for the XSD schemas; which are: **xsd:string**, **xsd:bool**, **xsd:date**, **xsd:float**, **xsd:integer** and **SetOfString** (the tag "**xsd:**" means that the type is built in to XML schema). The special type **SetOfString** has been introduced in order to allow the schema designer to have multiple fields.

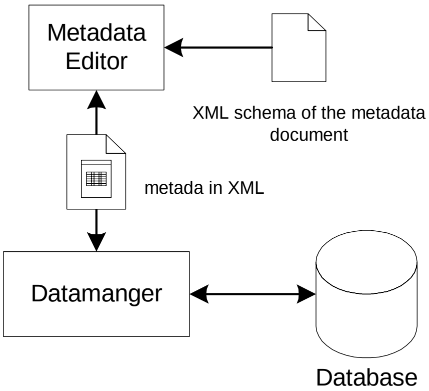


Figure 4: The architecture of the metadata editor

In Figure 5 an example a subset of the AVDocument entity schema is given:

Inside the tag **<sequence>** (see the XSD documentation for more details [6]) a metadata field for the AVDocument entity can be defined by means of the tag **<xsd:element>**. The tag **<xsd:annotation>** is useful for the reader in order to better understand the meaning of the field and it is used also by the editor as a ToolTip when the mouse pointer is moved over the form's input control related to the field.

As an example, suppose that a new field, indicating the creation date of the A/V Document, (called *ProductionDate*) is needed. It is sufficient to put a new element, as follows:

```
<xsd:element name="ProductionDate"
type="xsd:date">
   <xsd:annotation>
      <xsd:documentation>
         Date of the production
      </xsd:documentation>
   </xsd:annotation>
</xsd:element>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2000/10/XMLSchema" elementFormDefault="qualified">
   <xsd:element name="AVDocument">
      <xsd:annotation>
         <xsd:documentation>work level entity</xsd:documentation>
      </xsd:annotation>
         <xsd:complexType>
            <xsd:sequence>
               <xsd:element name="Title" type="xsd:string">
                  <xsd:annotation>
                     <xsd:documentation>Original title if known otherwise
                     assigned</xsd:documentation>
                  </xsd:annotation>
               </xsd:element>
               <xsd:element name="Genre" type="xsd:string">
                  <xsd:annotation>
                     <xsd:documentation>Genre of the Document</xsd:documentation>
                  </xsd:annotation>
               </xsd:element>
               ...
               other fields...
               ...
            </xsd:sequence>
      </xsd:complexType>
   </xsd:element>
   <xsd:complexType name="SetOfStrings">
      <xsd:sequence>
         <xsd:element name="string_item" minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
   </xsd:complexType>
   <xsd:simpleType name="string_item">
      <xsd:restriction base="xsd:string"/>
   </xsd:simpleType>
</xsd:schema>
```

Figure 5: Subset of the AVDocument entity schema

Eventually, the instance of a A/V Document based on the presented schema could be the following:

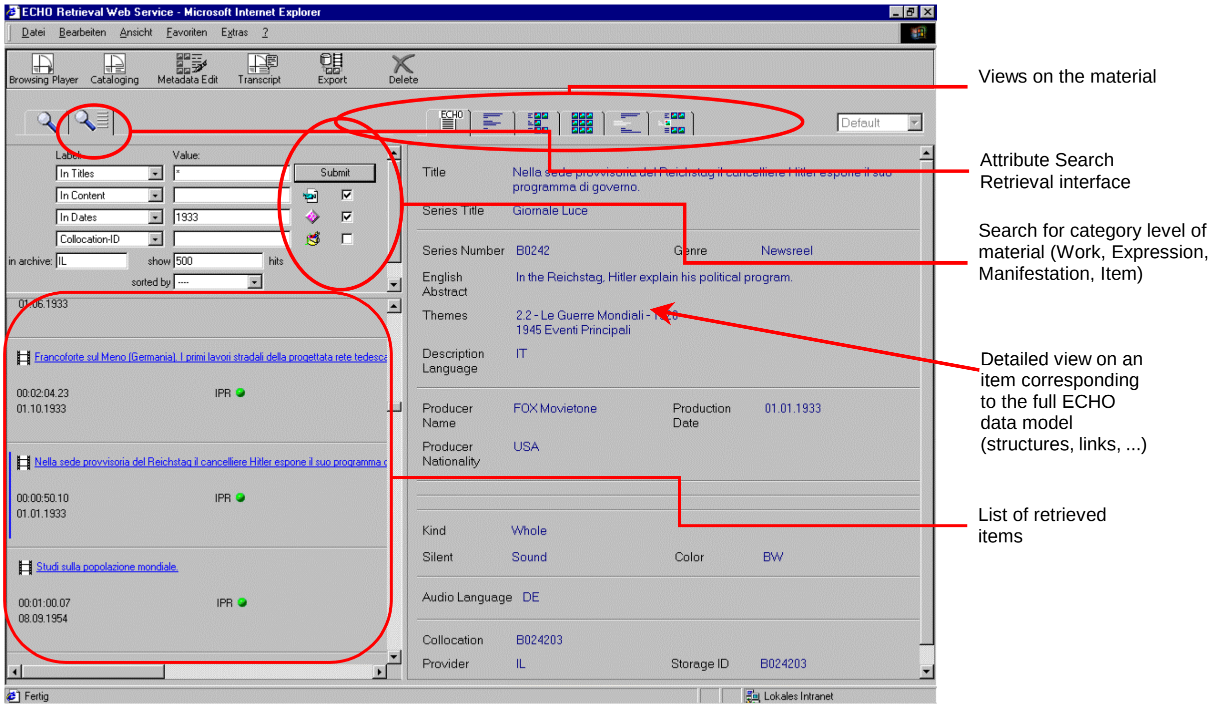
```
<?xml version="1.0" encoding="UTF-8"?>
<AVDocument ... >
   <Title> Olympic Games on 1936</Title>
   <Genre>Documentary</Genre>
   <Description>Documentary on 193 Berlin
       Olympic Games</Description>
   <Person_names>
      <string_item>Jesse Owens</string_item>
      <string_item>Hendrika
         Mastenbroek</string_item>
   </Person_names>
   ...
</AVDocument>
```

2.2.2 *The Video Retrieval Tool*

The video retrieval tool allows the user of the archive to search for the A/V documents in the database. The interface of the system offers three types of searches: a monolingual free text search, a cross language search and a boolean structured search. Figure 6 shows a simplified version of the retrieval tool interface. On the left side of the windows, three search forms are displayed and on the right side, the documents retrieved are listed.

The monolingual free text search looks for documents whose metadata contain the words specified in the form according to the type of search specified (i.e, “any word” or “all words”).

The cross language search allows one to look for documents over three specific metadata fields: Keywords, People and Location. These cross language fields can contain only a closed list of words, which are stored in a database table that contains the translations in all four languages.



Label:	Value:	
In Titles	*	
In Content		☑
In Dates	1933	☑
Collocation-ID		□

in archive: IL show 500 hits

sorted by ----

01.06.1933

Francoforte sul Meno (Germania). I primi lavori stradali della progettata rete tedesca

00:02:04.23 IPR ●
01.10.1933

Nella sede provvisoria del Reichstag il cancelliere Hitler espone il suo programma c

00:00:50.10 IPR ●
01.01.1933

Studi sulla popolazione mondiale.

00:01:00.07 IPR ●
08.09.1954

Title	Nella sede provvisoria del Reichstag il cancelliere Hitler espone il suo programma di governo.		
Series Title	Giornale Luce		
Series Number	B0242	Genre	Newsreel
English Abstract	In the Reichstag, Hitler explain his political program.		
Themes	2.2 - Le Guerre Mondiali - 1920 1945 Eventi Principali		
Description Language	IT		
Producer Name	FOX Movietone	Production Date	01.01.1933
Producer Nationality	USA		
Kind	Whole		
Silent	Sound	Color	BW
Audio Language	DE		
Collocation	B024203		
Provider	IL	Storage ID	B024203

Views on the material

Attribute Search Retrieval interface

Search for category level of material (Work, Expression, Manifestation, Item)

Detailed view on an item corresponding to the full ECHO data model (structures, links, ...)

List of retrieved items

Figure 6: Video Retrieval Tool

The structured search looks for documents by matching the metadata fields given in the form. This type of search is accomplished by translating the query in SQL and by submitting it to the database.

When the search is executed, the system shows some keyframes of the video part of the retrieved documents. By selecting a document the user can see the multimedia content and the metadata content by using an interface similar to the interface of the metadata editor.

3. Conclusions

We have presented the ECHO system that can be used for managing archives of documentary films. One important aspect of the system is the way in which metadata are managed. Given the complex structure of the metadata used, automatic and manual generation was integrated by way of the metadata editor. It is well known that fully automatic generated metadata contain noise that may affect the effectiveness of the retrieval. On the other hand, it is very costly generating highly structured metadata manually. In fact manually generated metadata have generally flat structure as, for instance, the Dublin Core. We believe that the integration of automatic generation and manual revision offer the possibility of using the power of complex metadata reducing the cost of generating them and guaranteeing a good accuracy of their content. To this aim the metadata editor presented in previous sections offer a valid solution.

Acknowledgement

This work has been funded by the 5th F.P. IST Research Program, Project No. 11994, ECHO (European CHronicles On line). We would like to thank the members of the ECHO Consortium for fruitful discussions which inspired the ideas presented in the paper.

References

[1] ECHO Web site: http://pc-erato2.iei.pi.cnr.it/echo

[2] ECHO User Requirement Report, ECHO Project Deliverable D1.2.1, June 2000, http://pc-erato2.iei.pi.cnr.it/echo/workpackages/wp1.html

[3] B. Furht, S.W. Smoliar, H. Zhang, “Video and Image Processing in Multimedia Systems”, Kluwer Academic Publishers, 1996. ISBN 0-7923-9604-9

[4] G. Amato, D. Castelli, S. Pisani, “A Metadata Model for Historical Documentary Films”, Proc. of the 4th European Conference ECDL 2000, Lisbon, Sept. 2000

[5] K.G. Saur München, “Functional Requirements for Bibliographic Records”, Final Report, 1998, http://www.ifla.org/VII/s13/frbr/frbr.pdf

[6] David C. Fallside, “XML Schema Part 0: Primer W3C Proposed Recommendation”, 30 March 2001, http://www.w3.org/TR/xmlschema-0/