# Regia: a Metadata Editor for Audiovisual Documents

Claudio Gennaro (claudio.gennaro@isti.cnr.it)
*ISTI-CNR Pisa, Italy*

**Abstract.** Although the Metadata Editor is an important part of any digital library, it becomes fundamental in the presence of audiovisual content. This is because the metadata produced by automated support tools (such as speech recognizers and shot detection procedures) is error-prone and often needs correction. In addition, scenes are manually annotated.

This paper describes *Regia*, a prototype application for manual editing metadata for audiovisual documents developed in the ECHO project. Regia allows the user to manually edit textual metadata and to hierarchically organize the segmentation of the audiovisual content. An important feature of this metadata editor is that it is not hard–wired with a particular metadata attributes set; for this purpose the XML schema of the metadata model is used by the editor as configuration file.

**Keywords:** Metadata, Metadata Editor, IFLA/FRBR, Audiovisual Content, Digital Video Library

## 1. Introduction

With the reduction in price of digital devices for multimedia production, audiovisual material is progressively becoming ubiquitous. DVD, Digital TV, and the Internet are some examples of sources of audiovisual contents in digital form. Moreover, nowadays everyone can buy a digital video camera and can become a producer of material, which can be easily distributed and published. Furthermore, this process will be even more simplified with the advent of digital camcorders able to produce video already compressed in MPEG-2 or MPEG-4 format.

On the other hand, all this information will practically remain wasted without available means to effectively discover it. Making a parallelism with the textual content, it is the same as we had all the pages of the Internet available as single images, and we were unable to apply some effective OCR tool for extracting the textual information. However, the situation for the audiovisual content is even worse, since OCR tools are quite effective as well as cheap; instead, information from audiovisual content requires much more effort for being extracted. In effect, the audiovisual information can be found in the speech, in the audio (e.g., music, explosions, etc.), in the keyframe (which is an image), or in the moving objects in the video. Furthermore, a sequence or a scene can be associated with a textual annotation (just like for the text we have

titles for chapters and sections) and can contain faces of well-known persons, objects, etc.

Often, the content providers send, along with the audiovisual stream, precious information such as text for subtitling; however they do not make it available for search but just for presentation. In contrast, television broadcasters, such as CNN or BCC, dedicate many resources to add descriptions to their audiovisual contents with the aim of making them more fruitful.

In general, the process of indexing and cataloguing of audiovisual content needs the human intervention. This is because the metadata produced by automated support tools (such as speech recognizers and shot detection procedures) is error-prone and often needs correction. In addition, scenes are usually manually annotated. For these reasons, a manual metadata editor tool is a fundamental component of a Video Digital Library.

A comprehensive metadata editor for audiovisual content should meet the following requirements:

1. *User-friendliness.* This feature is in general a requisite of any metadata editor. However, since audiovisual metadata are usually more complex than textual metadata, user-friendliness is practically obligatory.

2. Metadata *Customizability.* It enables the editor to be customized in order to be adapted to different metadata domains.

A satisfactory support of retrieval and management of documents requires a correspondingly adequate description of document contents. This is enabled by adopting of a suitable metadata model to represent the audiovisual documents, which should have the following two important features:

1. Metadata *Multilevel Abstraction of Bibliographic Data.* It allows to take into account different levels of specialization of the same abstract document. For instance the abstract idea of audiovisual object "The Lord's of the Rings" can be realized through different versions: "Original Version", "Extended Edition", etc. Moreover, the "Original Version" can be embodied in different formats, such as DVD, VHS, DivX, etc.

2. *Hierarchical Segmentation.* The audiovisual content can be recursively segmented in sub–elements. For instance, a segment of the video (e.g., a sequence) can be decomposed in more sub-segments (e.g., scenes), etc.

The objective of this paper is to present *Regia*, the metadata editor of the video digital library developed in the ECHO project (European CHronicoles On-line). The ECHO system assists the user during the indexing and retrieval of audiovisual documentaries. 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 are performed by expressing queries on the audio transcript, on metadata or by image similarity retrieval. The rest of the paper is organized as follows. In Section 2, we discuss related work. Section 3 presents the application scenario of this work. In Section 4 we discuss the metadata model adopted in Regia. Section 5 presents our metadata editor. Finally, Section 6 concludes the paper.

## 2. Related Work

The problem of editing metadata for audiovisual content is the topic of work of several companies and research groups. The main objective of a metadata editor is to enable the cataloguer to easily associate metadata to the audiovisual documents, in order to help the user in finding the contents. Sometimes these metadata are referred as *annotations*.

Here we give an overview of the most relevant existing work in this area of research.

*FilmEd* [10] is a prototype system developed at the Distributed Systems Technology Centre, of the University of Queensland, which enables the real-time collaborative indexing, browsing, description, annotation and discussion of high quality digital films or video contents. Beyond other important features of the FilmEd prototype, this system allows the user to annotate video content using the Annotea system of W3C [4]. Annotea uses an RDF based annotation schema for describing annotations as metadata and XPointer for locating the annotations in the annotated document. The model implemented in FilmEd, and hence also the annotation tool, supports hierarchical video segmentation. The annotations (which can be notes, explanations or other types of external subjective remarks) can be associated with segments, keyframes, or still regions within frames. Although, FilmEd is one of the most advanced system of annotation for audiovisual content its metadata model does not give any support for the multilevel abstraction of bibliographic data.

The *VideoAnnEx* annotation tool, developed by IBM [2], is a metadata editor for MPEG-1/2 video that is able to generate metadata in MPEG-7 format. Its interface is very essential but effective, and allows to annotate segments and regions of still images of the video. The annotations inserted can be taken from a lexicon (which is extendible and stored in a separated MPEG-7 document) or can be free text. However, the lexicon VideoAnnEx does not allow any further customization. Moreover, it does not support hierarchical segmentation and multilevel abstraction.

*VIDETO* of ZGDV [6] enables users to quickly generate video descriptions, which can be adapted to the specific domain of the content provider by using customization templates. For the metadata storage, it supports different data output formats and in particular the MPEG-7 format. The customizability is guarantied by domain specific templates in XSLT format. Unfortunately, it does not support the multilevel abstraction of bibliographic data.

Richoh's *MovieTool* [3] is a metadata editor for creating video content descriptions conforming to MPEG-7 syntax interactively, which supports hierarchical organization of the segments. The MovieTool is able to generate descriptions based on the structure of a video by fully exploiting the complexity of the MPEG-7 schema. Although the user interface of MovieTool allows to edit any element of the MPEG-7 description schema, it is very complicated to use. For this reason, it requires an users with a good knowledge of the MPEG-7 schema definition.

## 3. The Application Scenario of Regia

Regia has been developed during the ECHO project [1] funded by the IST programme of the European Commission under the V Framework [1]. The objective of the ECHO project has been to develop and experiment a video digital library system for historical documentaries owned by four different European audiovisual archives. The ECHO services allow users to search and to access all these documentary film collections and to exploit the content for their own particular requirements, whether commercial, educational, leisure, or whatever. This means that the ECHO services have to operate over linguistic, cultural and national boundaries, while respecting the requirements of international standards. The video digital library system has been designed with particular emphasis to the description of the characteristics

[1] the project began in February 2000 and was completed in March 2003

of the ECHO metadata model and the support for editing metadata in ECHO.

The ECHO system assists the digital library manager during the indexing and retrieval of audiovisual documentaries. Semiautomatic indexing is supported: the system automatically extracts several items of metadata information such as the scenes composing the video, keyframes that describe each scene, image features describing each keyframe, spoken dialog (automatically transformed into text through a speech recognition process). 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. Furthermore, retrieval over multiple languages is supported by using a cross-language controlled vocabulary. Retrieved films or their abstracts are then presented to the user. The project has also developed techniques to produce visual summaries. 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 detected action scenes. After this initial analysis, the more relevant clips are determined and assembled in order to maintain the flow of the story.

### 3.1. Lifecycle of the documents

Since the metadata model of ECHO 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, who 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 unwanted segments and merges multiple documents. As we will see, this phase is usually

performed starting from the top level of the model structure, and continuing by modifying/editing the lower–level objects.

## 4. The Metadata Model

When the project began, there were no well-defined metadata models for an adequate description of film data. An important 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 in the ECHO project [7] is an adaptation of the IFLA model, a general conceptual framework used to describe heterogeneous digital media resources [9]. In this section we present the ECHO metadata model.

### 4.1. Global Metadata and Internal Metadata

The ECHO metadata model includes a number entities related to different aspects of the document to be described. To simplify the presentation of the model, we decided to group its entities in two sets: *Global Metadata* and *Internal Metadata*.

The former set includes all the entities which concern pure cataloguing aspect of the document, without going into the peculiarity of the type of multimedia object described. From the perspective of the global metadata the document catalogued is a black box and it is seen as one monolithic object.

The latter set of metadata entities concerns the nature of the catalogued object and it is related to the internal organization of the documents. In general, any kind of metadata which can help the user to identify and retrieve the object (or part of it) can be used. For instance, if the object is a computer video game, there could be entities for describing the saved games, the character of the players, etc. In particular in our model we mainly focused on the time partitioning aspect of the multimedia object, i.e., how the audiovisual objects are divided in scene, shots, etc.

### 4.2. The metadata Model

The ECHO metadata model [7] has been defined as an extension to the IFLA/FRBR model. This model is composed of four levels describing different aspects of intellectual or artistic endeavor: *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

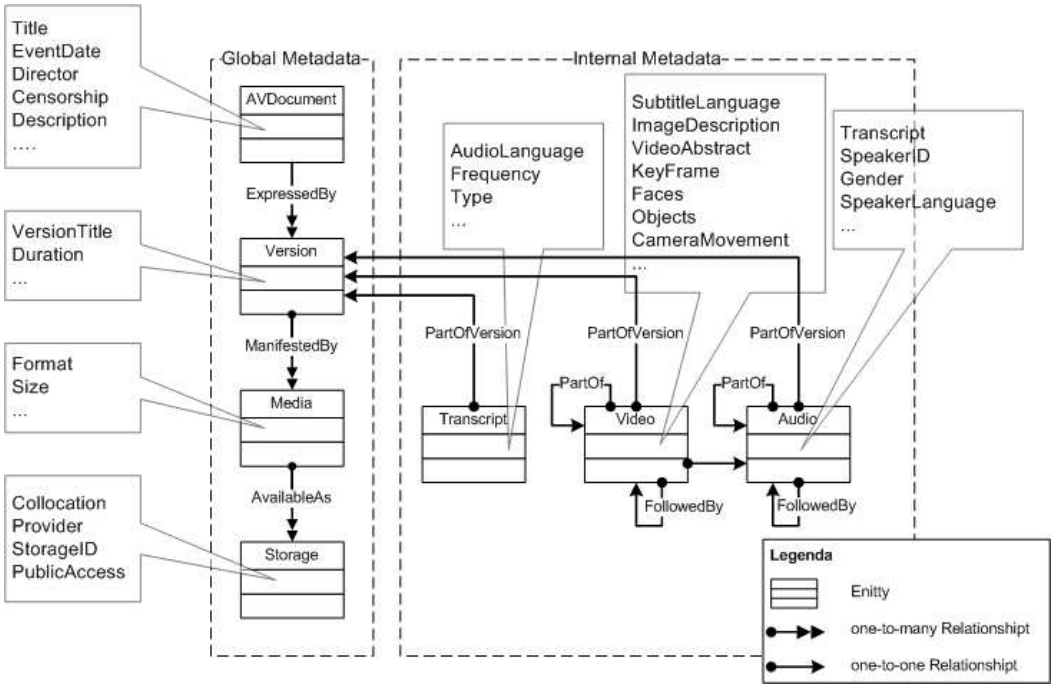


*Figure 1.* Schematic representation of the ECHO Metadata Model.

top level (work) to the bottom one (item). Figure 1 shows a schematic representation of the ECHO Metadata Model.

The entities of the global metadata set are *AVDocument*, *Version*, *Media*, and *Storage*, which belong to the four levels Work, Expression, Manifestation, and Item, respectively.

According to the IFLA/FRBR methodology the AVDocument entity is the most abstract one; it provides the general intellectual or artistic view of the document. For 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 instances, of the abstract entity Version, could represent different language versions of the same film, e.g., versions in Italian or in German.

However, the Version entity does not represent any specific implementation of the film. This aspect can be expressed 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., files, DVD, etc).

In Figure 1 some metadata fields belonging to entities of the model are also shown.

The three entities of the internal metadata set (*Video*, *Audio*, and *Transcript*) maintain the `PartOfVersion` reference to the Version they belong to. These entities are identified in the IFLA/FRBR as *Expressions* (since formally they actually belong to the Expression level). This means that for each instance of the Version entity we can associate a Video, an Audio, and a Transcript Object, which correspond to the entire document from the temporal point of view. The most important aspect to remark is that each of these entities has a one-to-one `PartOf` relationship to itself. This relationship permits to associate a document sub–segment to the parent segment which contains it. For instance, a Video object, which represents a specific scene Film document, is part of the Video object which represents the whole document. This allows to recursively partition the Video in sequences, scene, shots, etc. The relationship `FollowedBy` can be used to express the temporal succession of the document parts and the type of transition between a part and the following one (cut, fading, etc.).

### 4.3. Implementation

The ECHO model has been realized using the XML format. When Regia works in online mode, it communicates with the Echo system in order to obtain the XML metadata of the audiovisual documents from the database. In particular, once the user has found (by means of the video retrieval tool) a relevant document the URI of the document is obtained and passed to the metadata editor. In order to retrieve the document, this URI is sent to the ECHO database, which returns the XML document metadata. In particular each instance of the model is implemented by a set of XML files, each one corresponding to an instance of an entity of the model.

For simplicity, we refer to an instance of the model as "ECHO document", which is composed of an instance of the AVDocument entity, and one or more instances of the other entities (Version, Video, etc). Figure 2 depicts an example of an XML ECHO document. The example shows how the relationships between the objects are realized by means of specific XML elements. For instance, the element `ExpressedBy` of the AVDocument object contains the list of the URIs of its Version objects. Note that, Regia is able to work also in off-line mode, in this case the ECHO document can be stored in the filesystem, and the URIs simply become local path names.

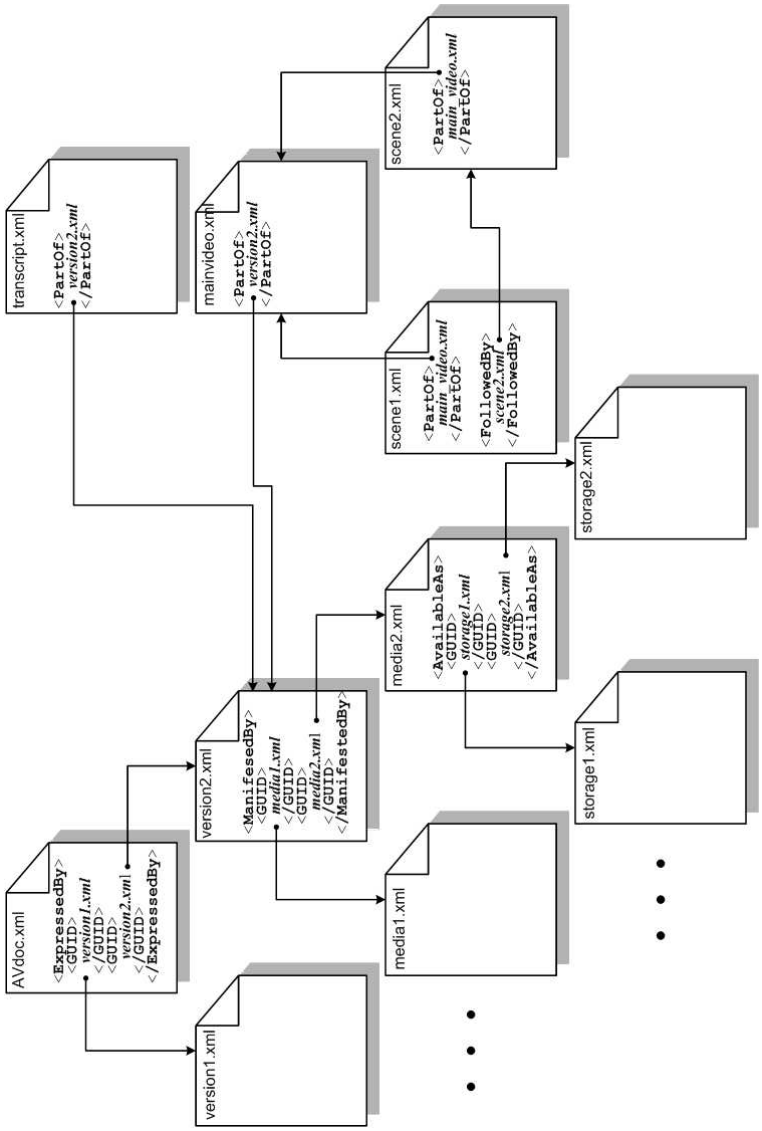


*Figure 2.* Example of an XML ECHO document.

The ECHO metadata model has been defined using the W3C XML Schema Definition (XSD). However, as we will see in the next section, the metadata fields defined in the schema are adaptable and can be customized to the needs of the users.

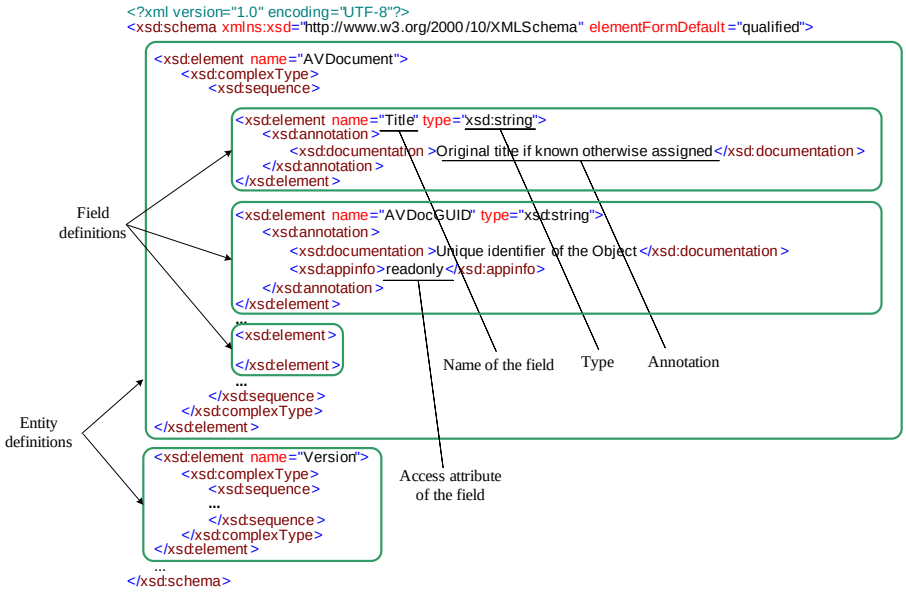


*Figure 3.* Illustration of the schema definition of the Metadata Model.

## 5. The Metadata Editor

### 5.1. Customizability - the XML Schema as Configuration File

An important feature of Regia is that it is not hard–wired with a particular metadata attributes set, for this purpose the metadata XML schema is used by the editor as configuration file for the metadata model. 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, such as: strings, boolean, dates, integers, etc.

In Figure 3 an illustration of a piece of ECHO schema definition is given. The XML element `sequence` (see the XSD documentation for more details [5]) includes the definitions of the metadata fields of the AVDocument entity (which are instantiated as XML elements). Each of them is defined through the XML element `element`. Besides the name of the field and its type, `element` can provide other useful information

```
<AVDocument xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance "
            xsi:schemaLocation="http:///EchoServer.it/Echo.xsd>

    <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>
```

*Figure 4.* Illustration of the AVDocument instance.

```
<xsd:simpleType name="CutType">
    <xsd:restriction base="xsd:string">
        <xsd:enumeration  value="HardCut "/>
        <xsd:enumeration  value="FadeIn"/>
        <xsd:enumeration  value="FadeOut "/>
        <xsd:enumeration  value="FadeInFadeOut "/>
    </xsd:restriction>
</xsd:simpleType>
```

*Figure 5.* Illustration of a Closed List Type definition.

to the editor. This information is contained in the predefined XSD element `annotation`, which includes the subelements `documentation` and `appinfo`. `documentation` is in general useful for the reader in order to better understand the meaning of the field and it is also visualized by the editor as a *tooltip text* when the mouse pointer is moved over the TextBox control associate with the field (see Figure 7). `appinfo` can be used, in general, to provide information for tools, stylesheets and other applications. We use `appinfo` to associate the metadata field with an *access attribute*, which specifies, for instance, whether the field is read-only or not. Regia reads this access attribute and in case of a read-only field it disables the editing. This is useful, for instance, for disabling the editing of some metadata fields that contain GUIDs (Global Unique IDentifier) to object of the database.

The instance of an AVDocument object based on the presented schema could be as in Figure 4.

By means of the XML schema it is also possible to define elements, which contain a closed list of strings. This feature is used to help the cataloguer during the metadata ingestion phase, but it is also useful during the retrieval phase, since these types of elements cannot contain misspelled words. In Figure 5 an example of a XML schema closed list type definition is given.

The closed lists are automatically recognized by the editor and represented in the editing interface as ComboBoxes controls.

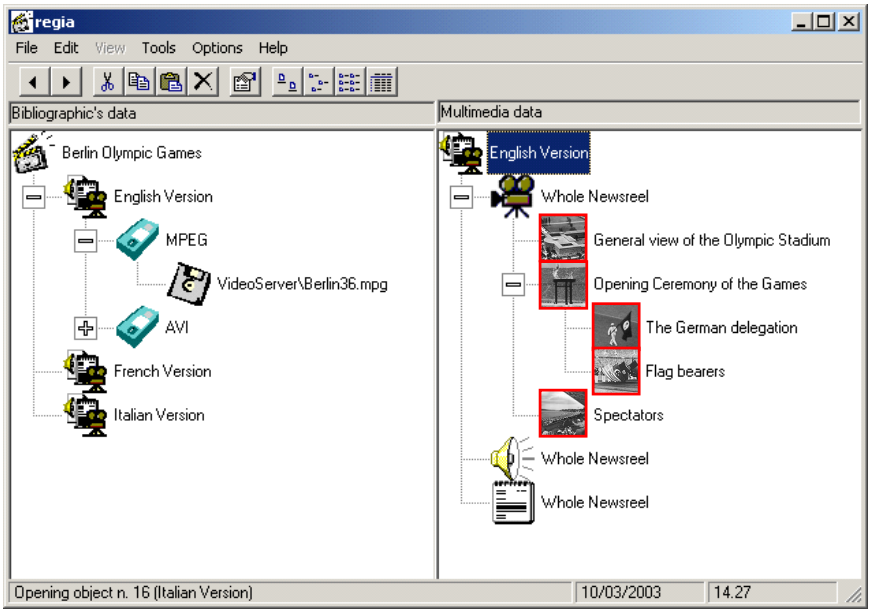


*Figure 6.* A screenshot of the main window of Regia.

## 5.2. The User Interface

The interface of the editor is designed in such a way that it is possible to browse the tree structure of an audiovisual document. Figure 6 shows the screenshot of the main window of editor: it displays a document like a folder navigation tool. The editor main interface reflects the separation between Global Metadata and Internal Metadata. On the left side of the window, we find the entities of the former set (referred as Bibliographic's data) and on the right side, the entities of the latter set (referred as Multimedia data).

On the top level of the tree, there is an icon representing the AV-Document object (the work of the "Olympic Games on 1936" in our example) with three children icons representing its Versions. By selecting one of them (the English Version in figure), it is possible to see the Media instances of the Version and, hence, the corresponding Storage objects.

By selecting the icon corresponding to a Version, the editor allows to browse the corresponding expression objects. This is achieved by using a second frame on the right side of the editor interface. In this way it is possible to see the existing Video, Audio and Transcript expressions of the document and, for each of them, to browse their segmentation.

By clicking on the icon corresponding to an object, it is possible to modify, in a separated window, its metadata fields. Figure 7 shows

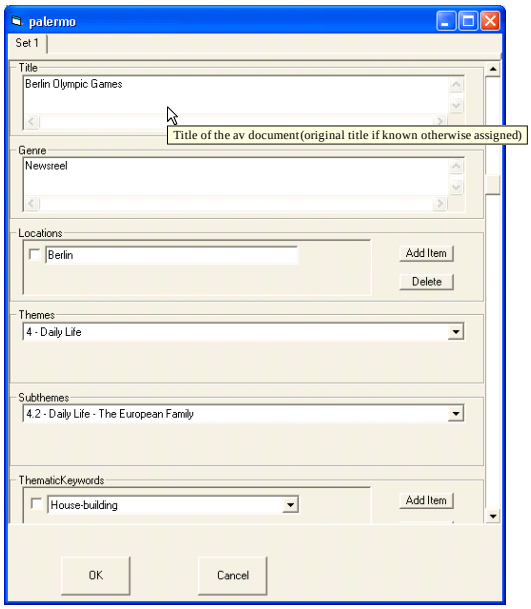


*Figure 7.* A screenshot of the Textual Edit window of AVDocument.

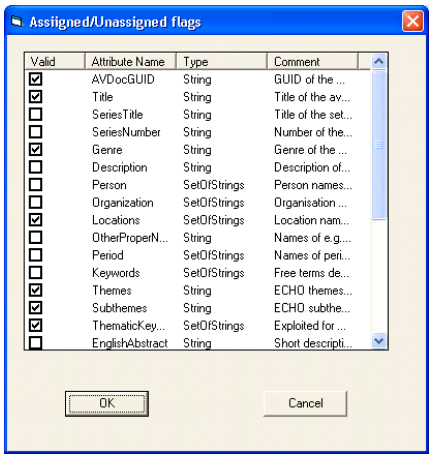


*Figure 8.* A screenshot of the window that allows to change the status of the textual metadata fields.

this edit window in case of AVDocument. Note that, the type of input control chosen for each field is decided at run time on the basis of the XML schema.

Sometimes, only a subset of the available textual fields contains meaningful values. Moreover, the value contained in the corresponding XML element may be not able to indicate a null or not assigned field. For instance a boolean element must be `true` or `false`, no other value is

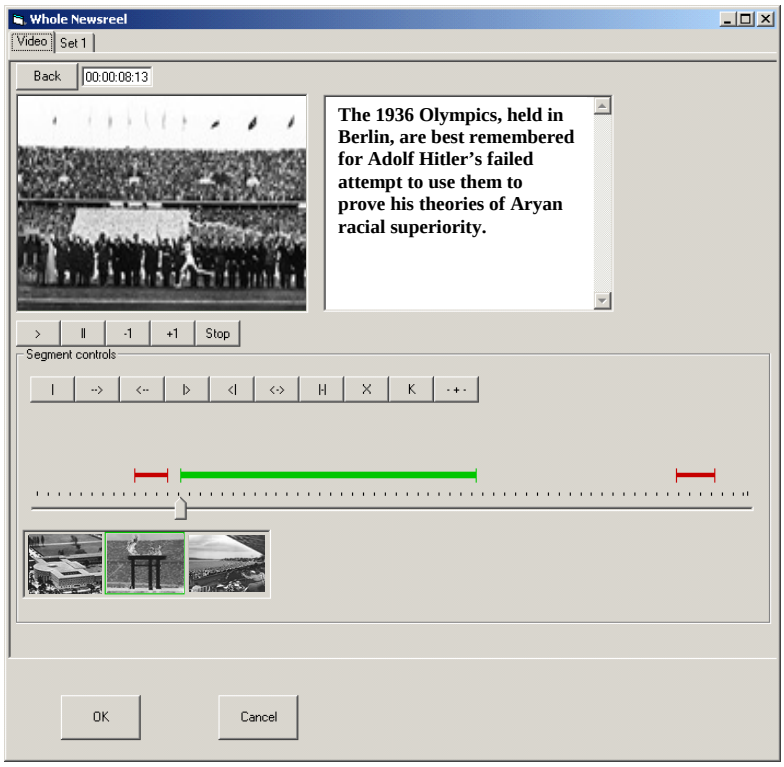


*Figure 9.* A screenshot of the Expression Tool of Regia.

accepted by the XML parser. The solution adopted is to allow the XML element of an "unassigned field" to not appear in the document, which, for convenience, is not made accessible for editing on the interface. In order to change the status of a field from *unassigned* to *assigned* and viceversa, we use the separate dialog window of Figure 8. If the checkbox is checked then the corresponding field is *Assigned* (and will appear in the edit window) if checkbox is unchecked then the corresponding field is *Unassigned* (and will not appear in the edit window).

A particular attention has been paid to the expression window design, i.e., to the Expression Tool (Figure 9). 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 segments, by means of a timeline control (see the bottom of the Expression Tool window). This timeline is subdivided in partitions that represent the media segmentation. By selecting a segment, the Expression Tool shows the "sub–Expression" corresponding to the segment of the media (for in-

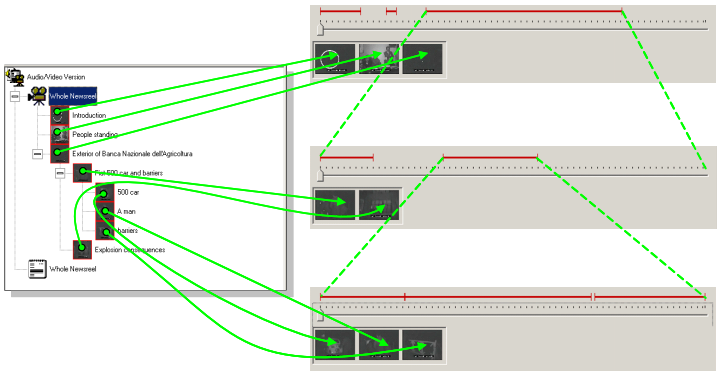


*Figure 10.* Navigation of the hierarchical structure of the audiovisual segments.

stance, a scene or a shot). For convenience, the same segments are also represented by their keyframes, allowing the user to see their video content. By double–clicking on a keyframe associated with a segment (or a segment on the timeline) it is possible to follow the corresponding sub-Expression. In this manner, Regia is able to provide two type of navigation through the hierarchical structure of the audiovisual segments (See Figure 10): a tree view (main window of Regia) and an hyperlink–like access (Expression Tool).

### 5.3. Cross–language support

The ECHO film archive is made up of language-dependent (speech, text) and language-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 cross–language search interface that can still provide a useful functionality. The approach adopted has been to implement online 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 Web site for external access.

In particular, all the AVDocuments of the collection are classified on the basis of three categories: *Themes*, *Subthemes*, and *Thematic Keywords*. For each category a closed list of possible values has been specified. Moreover, these categories are organized in a tree structure: each Theme includes a disjoint set of Subthemes and each Subtheme includes a disjoint set of Thematic Keywords. Each ECHO document

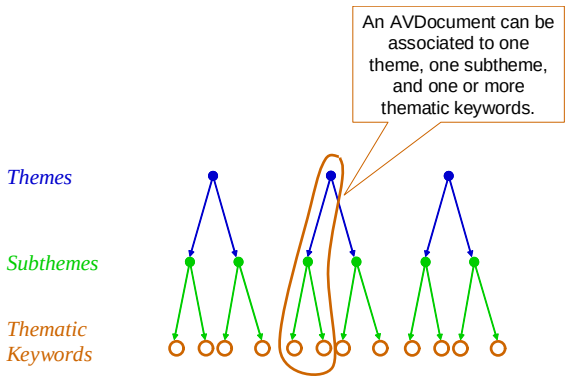


*Figure 11.* Closed list tree of the cross–language of ECHO.

is then associated to exactly one Theme, one Subtheme and one or more Thematic Keywords as shown in Figure 11. In order to enable the cross–language browsing/editing, the text associated to these categories is translated in five languages (English, Italian, Dutch, French, and German). A configuration file in XML format maintains these cross–language closed lists. For a specific AVDocument, Regia uses the English entry for actually storing the value of these three categories, but uses the current user language for presentation. In other words, the English language is used as pivot language.

## 6. Conclusion

Given the complex structure of metadata for audiovisual content, it is essential to integrate the automatic and manual generation by way of a 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 manually generating highly structured metadata. 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 this paper offer a valid solution.

Regia has been proved to match our four assumptions: it is user friendly and allows to be customized to support metadata of different domains. Moreover, both Regia and its metadata model support Multilevel Abstraction of Bibliographic Data and Hierarchical Segmentation.

In order to demonstrate the validity of the approach of the metadata editor developed in ECHO, we have migrated the entire archive of ECHO to the MILOS system [8]. MILOS (Multimedia dIgital Library Object Server) is a general purpose software component tailored to support design and effective implementation of digital library applications. Since Regia was designed to natively manage the documents of ECHO directly in XML format, we had only to replace its interface of communication (originally written for Corba) with a new one based on SOAP able to call the methods of middleware of MILOS. The prototype implementation of Regia developed in Visual Basic can be downloaded from the Web at `http://www.nmis.isti.cnr.it/gennaro/regia.htm`.

Future directions of our research include the support of MPEG-7 standard, many improvements of the user interface and the realization of a new version in Java.

## 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: European CHronicles On-line. http://pc-erato2.iei.pi.cnr.it/echo/.
2. IBM MPEG-7 Annotation Tool. http://www.alphaworks.ibm.com/tech/videoannex.
3. Ricoh MovieTool. http://www.ricoh.co.jp/src/multimedia/MovieTool/.
4. W3C Annotea Project. http://www.w3.org/2001/Annotea/.
5. W3C xml schema. http://www.w3.org/TR/xmlschema-0/.
6. Zentrum fuer Graphische Datenverarbeitung e.v. (ZGDV). VIDETO - Video Description Tool. http://www.rostock.zgdv.de/ZGDV/ Abteilungen/zr2/Produkte/videto/ndex_html_en.
7. G. Amato, D. Castelli, and S. Pisani. A metadata model for historical documentary films. In José Luis Borbinha and Thomas Baker, editors, *Proc. of the 4th European Conference ECDL*, pages 328–331. Springer, 2000.
8. G. Amato, C. Gennaro, P. Savino, and F. Rabitti. Milos: a multimedia content management system for digital library applications. In *Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2004)*, volume 3232 of *Lecture Notes in Computer Science*, pages 14–25. Springer, September 2004.
9. K.G. Saur Mnchen. Functional requirements for bibliographic records - final report, 1998. http://www.ifla.org/VII/s13/frbr/frbr.htm.
10. Ronald Schroeter, Jane Hunter, and Douglas Kosovic. Filmed - collaborative video indexing, annotation and discussion tools over broadband networks.

In *10th International Multimedia Modelling Conference*, pages 346–353. IEEE Computer Society, January 2004.