

# **New Document Concept and Metadata Classification for Broadcast Archives**

## **1 Introduction**

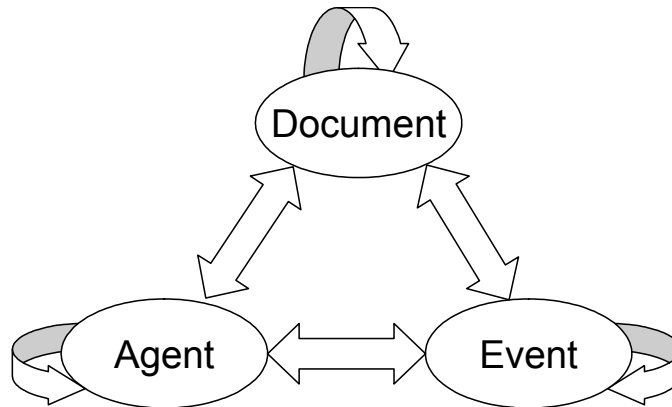
It is evident that document is the central concept in any archive, and the basic task is to make these documents searchable and retrievable. Textual documents can be full text indexed and very efficient search engines exist as Google, however non-textual documents as images, sounds or other audiovisual documents require associated textual information, metadata to make search possible.

Several document definitions and metadata exchange schemes can be found for all types of documents. However, when we started to design the information system for the Hungarian National Audiovisual Archive that was established to be the digital legal deposit of the Hungarian related broadcast audiovisual programs, we have to face some inevitable problems. The main questions were: how to define the bibliographic unit, the document in broadcast environment and which standardized metadata scheme can serve the most efficiently archival processes?

This paper describes the selection process that lead to the multilevel, multidimensional document model and classified metadata scheme that proved to be adequate both for archival and retrieval purposes. The concepts introduced can be useful for all kinds of document archives to extend their capabilities.

## **2 Basic data model**

The audiovisual archive's basic data model is very simple. The three main entities are the Documents, Agents and Events. There are relations defined between Documents and Agents (e.g. Writer, Director, Actor), between Agents and Events (got a prize), Documents and Events (about the 2<sup>nd</sup> World War). And there are also internal relations among Documents, Events and Agents.



**Fig. 1.** The basic data model

Using this simple model and applying the principles of the Dublin Core Metadata Initiative a data structure of more than a hundred tables was designed. That is very useful for precise description and refined search, however metadata association is not an easy task. To exploit the strength of the model, efficient support should be provided for the archivists. In the case of Agents and Events the techniques of authority lists could be adapted, however the document concept and the document-document relationship has to be revised for broadcast audiovisual content.

### 3 What is a 'document'?

Studying reference works and searching the web, several definitions can be found for the concept *document*. The number of definitions show itself, that the problem isn't solved at all.

#### 3.1 Problems with document definition

Let us see some typical approaches for document definition:

- "A document is a writing that contains information" [1]
- "A physical entity of any substance on which is recorded all or part of a work or multiple works" [2]
- "Recorded information or object which can be treated as a unit" [3]
- "An information resource is defined to be anything that has identity" [4]

Any definition referring to writing is inadequate for audiovisual material. Physical objects also can hardly be associated to documents in a digi-

tal environment, because storage systems with hundreds of terabytes capacity don't make the physical distinction of logical units possible. The last two examples use the tricky solution of the problem: the obscure concept of 'document' is defined by the similarly obscure 'concept' of 'unit' or resource respectively.

Conventional archives like libraries have already solved some definition problems with series and periodicals or compound volumes, but they have the book, the physical object as a starting point. In the case broadcast audiovisual programs only the practically infinite media stream can be considered as physical reality, all program items can be identified technically as a time interval in the stream by a more or less arbitrary human decision. Broadcast items are typically multiple level compounded documents, and can be elements of several series, so the solution used by librarian practice can be used only partially. Preliminary program guides can help to select basic units of the stream, but these guides are usually not exact by time and not detailed enough. Human processing time for the audiovisual units is critical, because the stream flows, so the birth of new documents is continuous. However filling the transmission time with all original items is almost impossible for the broadcasting companies, so replay, reuse of items is a general practice, so the number of documents to be processed is much less than it can be expected by technical calculations.

To summarize the broadcast document model requirements, it has to be able to:

- Define exact technical parameters, because recorders and encoders do need exact commands, algorithms.
- Enable multilevel hierarchy or relations to make any level of itemization possible.
- Support processing by the exploitation of document-document relations, identity, similarities and inheritance.

### **3.2 Identifying documents in broadcast environments**

For a real archive a practical, method oriented definition is have to be given, omitting undefined or obscure concepts.

The document's birth is its transmission, broadcasting to the public. The document's source (by the definition of the DCMI [4]), so the recorded sub-stream itself technically can be identified by the frequency, the transmission network (including geographical location) and the time interval. The first two parameters can be considered as static (at least rarely changing), so the key issue is the determination of the time interval. The raw re-

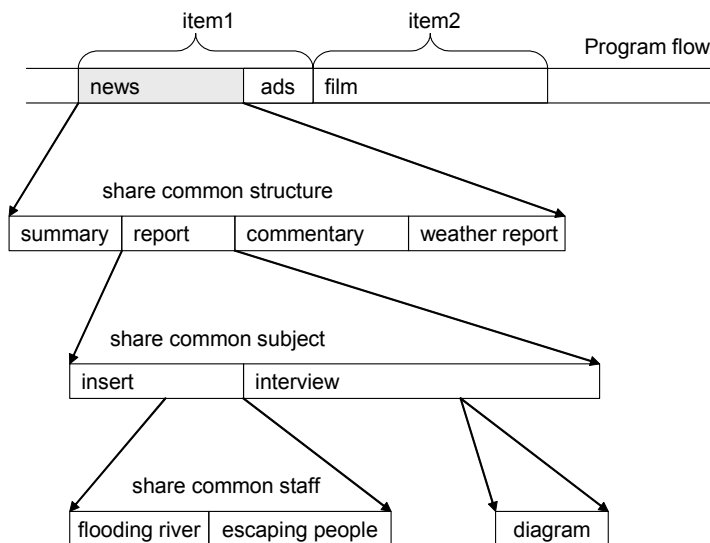
cordings are available for about ten days (then the cyclic buffer overwrites unsaved data), so the time for archiving limited.

As the first step, to take the authors intention into account, the program guide's items are used as the rough selection points. It has to be mentioned, that the program guides don't include any reference to the advertisements, program recommendations that wedged between program items, so the starting points of the interval van differ significantly. By the definition used in our archive, the first level document starts at the beginning frame of the program item and lasts until the first frame of the following, preliminary announced program item. The first level of the document delimitation often includes several foreign parts and most often defines a multilevel compounded document. The basic metadata are associated to the first level documents based on preliminary information, so at the end of the first step practically an interactive program guide is obtained.

After the first step we arrive to a decision point. Because the capacity of the storage system, and therefore the processing time is limited, we have to decide that a certain first level document has to be included in the long term archive or not. If the preservation is refused, the processing is finished, but if it is enabled (considered as Hungarian related, and not yet processed), the first level document is stored in the long term archive (to be retained for eternity). Obviously valuable data can be lost, because seemingly neutral program items can contain Hungarian relations (contributors, locations, events), or the already processed document occurs in a definitely different context, but we have to live with this limitations.

The second step is the further partitioning of the selected first level documents. We have now time, the selected first level documents are retained for ever. Generally the selected document has several compounds. According to our model these compounds can be selected as the parts of the *whole* document, and treated as a new, independently described document, however inherently related to the mother document. (The relation type determines the metadata inheritance or import rules.)

After the second step, since the document is archived, third, fourth and more steps can be done for more and more refine the resolution according to the desired level as far as the frame level.



**Fig. 2.** Document identification process

The document hierarchy is illustrated on the figure. This simple example shows that a single news magazine can generate dozens of documents with different staff, different subject, different structure, however sharing some common properties. Extrapolating this tendency to a 24-hour program flow, we can calculate with hundreds of separate documents on a single day, on a single television or radio channel, each has its own metadata set.

#### 4 What does metadata describe?

Metadata provide information about information (content, document) by definition. Several standardized metadata schemes coexist such as MARC in libraries or Dublin Core in the world of electronic documents. Metadata are classified by Gilliland-Swetland [5]: administrative, descriptive, technical, preservation and use categories were defined. However the mentioned examples stand mainly for the descriptive metadata class. In an operational environment other metadata categories are to be used intensively to control workflow, or standard descriptive metadata are associated with a new meaning.

It was mentioned earlier that in audiovisual broadcast archives the time factor is of the key success factors due to the enormous quantity of documents. A lot of human effort can be saved by the utilization of the relations among documents. Even in the librarian world the role of document-

document relations is increasing. The number of pre-defined relation types is growing [6], or additional document levels are defined to join documents that cannot be hold together using the one level model [7].

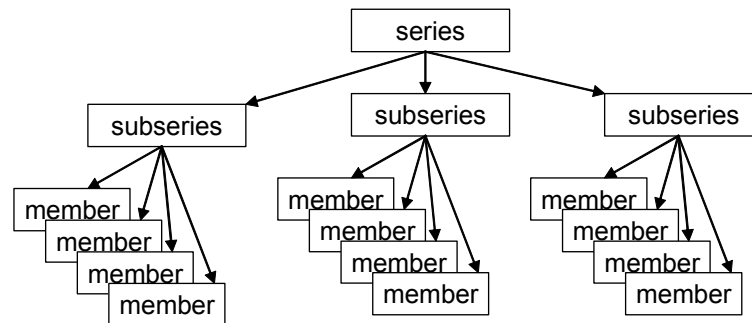
In the following chapters we give a few examples on the utilization of the *Relation* metadata field for enhancing effectiveness of processing but maintaining the original meaning as well.

Relation traditionally means a semantic connection between documents, however it can serve as a very efficient tool also to help (or even make unnecessary) filling metadata values, and on the contrary, the similarity in metadata values can suggest relation.

#### 4.1 Document relationship - Virtual documents

Virtual documents are documents that have no essence in the digital repository, virtual documents are pure metadata sets that are not associated directly to a mediastream.

Using virtual documents can play extremely useful role in describing periodicals and series. If a program has a regular staff, and it runs daily, the staff has to be filled at the virtual document level and different series members inherit these metadata through the *IsMemberOf* relation type. Only the specific differences (guests, subject) have to be recorded for the actual item. For example the metadata of Star Track can be normalized like a well structured relational database using three hierarchy levels minimizing human processing time, and minimizing the number of potential errors.



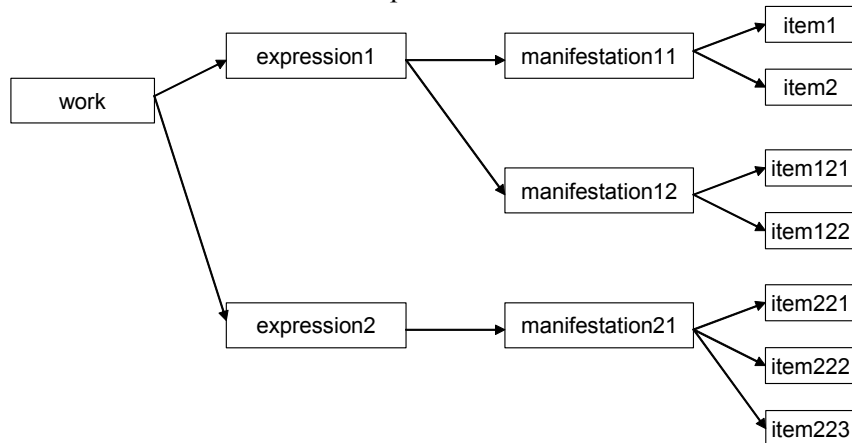
**Fig. 3.** Inheritance from series

The same theory, but a different situation can be found in the four-level document model of the International Federation of Library associations [7]. It separates levels from the result of theoretical authoring (Work) through the materialization process (Expression, Manifestation) to the

physical object (Item). In broadcast archives the replays can be considered as items of the same manifestation, so all metadata of the manifestation level except timing can be inherited to items through the *IsExemplifiedBy* relation. The several (shorter or longer) versions of a commercials are examples of the expression to manifestation relationship, like different editions of a novel. The associated relation is *IsEmbodiedIn* and almost all metadata can be inherited except duration and time. The application of the work to expression relationship often goes beyond the limits of the audiovisual archive. The *IsRealizedThrough* relation can link documents that have the same origin. Several treatments (drama, ballet, orchestral music) of the story of Romeo and Juliet can share the same content description using the work concept. The Work, Expression and Manifestation are always a virtual documents, the only physical entity is Item.

**Fig. 4.** The Work-Expression-Manifestation-Item levels

The utilization of logical document relations and inheritance is very useful to shorten processing time (and save storage capacity), but there is no common rule for the extent of inheritance. From case to case the fields to be inherited and the fields with prohibited inheritance have to be carefully



selected.

## 4.2 Document types - Document templates

Virtual documents in the previous chapter were applicable if there were a logical, intentional relation between documents. However a there are looser connections that can be exhausted. These connections are related to the Type field in some way.

The Type metadata is used to determine the properties of the content. In the Dublin Core Type Vocabulary a general, but rather coarse typology is given restricted only for the basic types. In the case of audiovisual content Type is often associated more or less with Genre. In the practice of broadcast companies the theoretical genre and the intended audience is often confused. The European Broadcasting Union's P/META project [8] gave a clear vision and a four dimension encoding scheme to solve these problems. The broadcaster's intention, the formal structure, the content and the relevant groups were separated. If we study the programs classified by EBU according their formal structure, several similarities can be noticed in structure and main metadata groups and fields.

Determining and considering the formal structural type of the document, common templates can be used. These templates utilize only a small part of the whole model, so filling fields is much easier and faster. At the application level document templates can be derived from real documents, retaining only the filled fields of the model.

### 4.3 Looking for Similarities – Embedded Search

The methods described in the previous chapters are usable only if the archivist has the knowledge of essential or formal relations. Certainly this knowledge develops after a while, but the large number of documents makes the decisions hard. The collection management system of our archive helps this process by the means of informatics.

Let's suppose that the archivist responsible for a program item doesn't recognize any essential or formal relations, considers the document as a standalone item and starts to fill the fields of the entire model. After completing the formal section of metadata set a background process starts and looks for similar documents. If a certain correspondence found a document template, or another relationship is offered to the user. The keyword is the 'certain correspondence'. The adequate measure of similarity should be derived from practice, however the categories can be presented according to the model.

**Table 1.** Cases of similarity

Correspondence	Technique/relation
Same format fields are filled	Same Type
Same values in the same fields	Serial member or Same Item
Similar values in content fields	Expressions or Manifestations of the same Work

## Conclusion

One of the main problems of the operation of audiovisual archives rises from the large number of individually retrievable documents. This large number arises from the nature of broadcasting, the use of complex, multiple level compounded program items. However analyzing the structure and the elements of the program flow several identical or similar documents or document variations can be found.

Based on the document-document relation models borrowed from the librarian practice and adapted for broadcast environment, an efficient tool was conceptualized and experimentally implemented to help archivist to recognize and utilize relations, and spare processing time, reduce database complexity and minimize storage capacity without any restrictions to the usability of the archive.

## References

1. Wikipedia, the free encyclopedia, [en.wikipedia.org](http://en.wikipedia.org)
2. U.S. National Archives and Records Administration, [www.archives.gov](http://www.archives.gov)
3. National Archives of Australia, [www.naa.gov.au](http://www.naa.gov.au)
4. Dublin Core Metadata Initiative, [www.dublincore.org](http://www.dublincore.org)
5. Anne J.: Setting the Stage Introduction to Metadata
6. The European Library Project, <http://www.europeanlibrary.org/>
7. Functional Requirements for Bibliographic Records, IFLA, 1998
8. The EBU Metadata Exchange Scheme, EBU Tech 3295