

## Users, producers & other tags

### *Trends and developments in metadata creation*

*by Annemieke de Jong*

*Netherlands Institute for Sound and Vision-Media Management Commission FIAT/IFTA*

#### **Introduction**

Evolving media technology is the cause of more and more metadata on our collections being created outside of the context of the archive, outside of the reach of the professional cataloguer. This is done in various ways and in various places. Today I want to try and give an overview of the developments and see how they relate to our work as audiovisual metadata professionals.

Not so very long ago – and in many cases still- archivists and documentalists were in full control of their catalogues. All metadata was being manually produced and they had the first and the last say about what was coming in and what was going out. It was the professional who decided what was to be made accessible in the first place and how this should be done. Once in the catalogue an item, a description was there to stay. In the exact same format as it had been put in initially.

This situation is changing fast. More and more metadata on the materials that we maintain, is being generated away from us, outside of the archival realm, outside of our reach and control. This concerns metadata that is derived automatically, but also metadata that is put in by humans. The creation can proceed inadvertently, temporal, as a consequence of the processing of files. It's also done in a deliberate, organized way, meant to keep permanently. In some cases the metadata is attached to the actual video or audio file. In other cases it's stored in a database somewhere. Metadata creation outside of the archive may concern all categories, be it technical or administrative and formal. It may also describe the content of the programs, and contain semantic information therefore.

#### **Metadata creation by 'producers' and systems**

One of the main reasons this is happening is that we are part of a networked environment. We are interoperable, our systems and databases have connections to other systems and databases inside and outside. This environment may be very complex like the one around the archive that I work for, with many internal and external information flows, interfaces and protocols. It can also be a much simpler one. The point is that the archival processes and the archival systems do not any longer stand-alone but are embedded in an infrastructure where metadata on audiovisual programmes is permanently created, modified and passed on. Inadvertently and automatically, but just as well deliberate and manually.

We - as the archive, as the professionals – are gradually getting used to the fact that this 'external' metadata is flowing into our domain almost permanently nowadays. That it's non-expert cataloguers, like technicians, rights owners, logistic people, programme makers and journalists that are creating them. And that it's scheduling systems, editing software and transmission devices doing the same.

Metadata come into being right at the source through to the whole production process to eventually end up in a catalogue or a catalogue-like environment, where it can be retrieved for re-use or consultation. What we as cataloguers do, when it eventually arrives at our desks (that is when we are given the chance) is embedding it into our own world as good as we can. We correct it, we contextualize it and we classify it, by having it linked to our thesauri and controlled vocabularies, for better and more sustainable search results. In this way we add both quality and the long-term perspective, as genuine archivists would.

### **Metadata creation by algorithms**

Another emerging new way of metadata creation springs from software development. Software to be applied during production or encoding, which basic technique is segmentation of the content into shots or clips, that are then analyzed according to certain parameters. Image, audio and natural language processing result in low level and high level features. Application of these technologies delivers metadata, to be stored in a database. It can also provide a series of keyframes, a storyboard, graphics or timelines as an on-the-fly generated search result. Automated indexing deals with shot changes, color distribution, texture and camera movements. But it can also inform us about the actual content, the higher semantics. What's the programme about, what's actually to be seen and heard? This kind of metadata is derived from algorithms that detect persons, settings and objects and from processing text and language elements.

Again, we're dealing with a form of 'description' that is done externally, outside of the traditional professional domain, by computers. Now these processes may be automatic, but they're also deliberate, the algorithms are being especially designed for the purpose. Cataloguing experts have an important job here, helping researchers in defining - from our experience, from our user requirements- the algorithms and parameters that are at the core of this technology. On a practical level, once implemented in the workflow, automatic indexing turns out to be still (and will probably be for a long time) quick and dirty. The results therefore do not only have to be corrected in the professional domain, it's there that they will have to be interpreted and placed into a meaningful context.

### **Metadata creation by users**

Next to these two developments, metadata from the production environment and automated indexing, there is another metadata movement. It is done manually, it is done deliberately and it involves higher semantics. It's the metadata created by users, directly and indirectly. I distinguish three forms.

The first is done by outside experts on a certain field This practice is a result of the increasing role of audiovisual materials in both the educational- and the cultural heritage domain. For general consultancy and to support re-use in a production environment, the materials may be catalogued by professionals in the archive, but for specific use in school or in museums the same materials is made accessible from the perspective of the intended users in these domains. For the educational environment, didactic experts assign domain-specific keywords that relate to the various curricula. For media historical use, context information may be added on the reception of a certain program, the viewers rate, the societal effects and so on.

Crowdsourcing is the second way in which users can be involved in cataloguing. Wikipedia defines crowdsourcing as "the act of taking a job traditionally performed by an employee, and outsourcing it to an undefined group of people in the form of an open call". The public may be invited to analyze large amounts of data. Wikipedia is itself of course one of the most famous examples of crowdsourcing. In the domain of digital documentary information more and more use is for instance made of crowd sourcing for helping to catalogue photo collections. Via the web people are being asked to provide as many details and context information on the published photographs as possible.

With the third type of user generated metadata, social tagging, I do not just mean the massively growing social software movements like Flickr, Technorati, Delicious and YouTube, where people tag their own data but I mainly refer to the online tagging of 'traditional' cultural collections. Examples: Steve Museum in the United States being one of the first initiatives. Here paintings are being tagged within a controlled environment.

But we can also see it emerging in the audiovisual domain: BBC Radio's project tagging audio clips.

Expert cataloguing, crowdsourcing and social tagging all employ the knowledge of the user, directly or indirectly. This save costs, as archive professionals don't have to do their own research. It 's also a way of getting to know the user, getting to know the context the archive operates in. This may help a collection owner to better target his goods, by employing profiling and recommendation techniques that make use of the implicit information behind all this user input. User participation is extremely interactive, so it increases the engagement with archival collections, which again strengthens the cultural and societal role of the archive.

From a managerial point of view the advantages are clear. To professional cataloguers however, user generated metadata represent a new challenge. This is in particular true for social tagging. In a way, al lot of the other metadata creation that I have mentioned, for instance in production and by outside sources, all have some parallel in the pre-digital world. Social tagging hasn't. It's a completely new phenomenon. Main dilemma for the professionals lies in the control. Should they correct, contextualize and classify this user input, as they do with other outside metadata? Or should tagging be left free and used as autonomous entries to the content?

### **Folksonomy vs Taxonomy**

This question touches upon an interesting debate that is going on these days between representatives of the professional cataloguers (the followers of Aristotle and Dewey as I've named them, as the founders of the classification tradition) and the fans of free tagging. For the first group -mainly book libraries and document archive people- letting the tagging be done free is not advised. Main reason is that traditional cataloguing description has the authors' intent as the leading principle, whereas free tagging allows multiple equivalent viewpoints. They fear that free tagging thus will eventually make 'meaning' relative.

Followers of the free tagging movement agree, but they do not think this is a bad thing. Sure, every user will tag the same object slightly, or even completely different. The result may seem like a mess but it isn't, because automatic clustering and filtering techniques will create meaning and context in the end. The implicit information, the relations behind all the single tags, users, objects and collections can be statistically processed and thus create all the necessary context.

### **Audiovisual cataloguing & social tagging**

I would like to elaborate on the implications of this debate for us. Were do we stand as cataloguers of audiovisual content? Actually, I think we have a special position here. The concept of free tagging might not be so very distant from our practice, as they may be from the book and document world. In fact, it could fit in very well with our way of working. This is has everything to do with the nature of audiovisual materials. But it's also strongly connected to our descend as production archives.

The description we make of an audiovisual programme has to take a two demands into account: Because of the temporal nature of av-content, the catalogue description has to function as a substitute of the program itself. A text description of the shots and scenes, preferably time aligned, is the only way to quickly grasp the contents of a program without having to view it. Then there is the re-use purpose, meaning that the description should facilitate easy re-use of parts. Both these demands lead to a cataloguing approach where an audiovisual product is seen an aggregation of separate parts or elements, a collection of clips. Our descriptions therefore, are clip based, item focused and it's this approach where our practice clearly diverges from the way books and archival documents are usually being made accessible: which is not on chapter level, not on paragraph level.

Then there is the semantic richness of our content. This implies that one and the same catalogue description has to deal with several different levels of meaning, and consequently, include different viewpoints.

-Information content: The who, what, when, how and why

-Audiovisual content: what's to be seen and heard?

-Stock shots: shots that can re-used in a different context

Summarizing: The re-use objective makes us catalogue our materials in a clip based, item focused way. Because of the rich semantics the clips are made accessible from different perspectives, from multiple viewpoints. These are the basics of audiovisual cataloguing. And in a way, they remind us strongly of the basics of social tagging!

### **Audiovisual cataloguing & social tagging**

When assigning keywords we have to do justice to these different semantics. This is not an easy job. The information content doesn't give us any problems here, it's facts: the who, when, why and where. But the other levels, the audiovisual content and the stock shots have always rendered many dilemmas and complications. Audiovisual content as such, simply holds too much details, too much ambiguity, too much possible viewpoints.

From my own experience (but I know this goes for many other archives) I remember that in the pre-digital era we had our own practical solution to deal with these semantic complications. We simply didn't use any rules for assigning keywords. So, next to a simple subject catalogue – we owned a huge collection of single keywords in alphabetical order. This enormous pile gradually built up, unstructured, overlapping, unbalanced, with no relations or internal references. The keywords in it, mainly dealt with the 2<sup>nd</sup> two layers, the audiovisual content and the stock shots, and they were our way to accommodate all possible sorts of re-use, every semantic perspective. The many details and visual characteristics of an image would be tagged in a completely de-contextualized way. Consequently, our pile of keywords grew with every broadcast programme that entered the archive. It became a real messy lot, that lacked all rationality. Extremely unprofessional, we would say now, Yet, this enormous card catalogue, that eventually contained millions of single keywords, was extremely useful, because - a jumble as it was- it contained every imaginable viewpoint on the content and subsequently, many different ways to access and exploit our shots and sequences.

Ofcourse at one point, we were forced to list all of our keywords and structure them in more professional way. Our pile was gradually cleaned up and thinned out, to be arranged in an electronic thesaurus and several controlled vocabularies. What we gained was a well-designed, interoperable metadata structure where some ten thousand keywords, each had their own place in a network of fixed relations. But what we lost were the multiple access points, the wonderful subjectivity, the inspiring randomness and the serendipity of our old messy pile, that had so well served many of our users.

I'm not saying we should go back to this old rather messy way of working. I simply mean that, in employing the social tagging movement around us, in combining it with our own professional practices, we –as audiovisual collection owners- may be getting the best of both worlds.

As everyone well knows creating metadata, tagging on clip level in order to build up a rich collection of re-usable content is highly labor intensive. In our case, at Beeld en Geluid, as in many other cases, the increasing flood of incoming materials has been preventing us from doing this for quite a number of years now. There is simply no time for it. Maybe at one point, algorithms may help out here, but at the moment they're by far not fit to this on a high semantic level. So here social tagging may come in. Free tagging by the general public could be of enormous help in making our collections accessible, on clip level and from multiple viewpoints.

We – that is to say our classifications, our thesauri, our vocabularies – should of course still control the first semantic plane, the information content, the who, what, where and when, in respect to the makers' intent, and thus create at least one professional context. But for the pure audiovisual content we might very well employ the enthusiasm of the user and his communities, and create platforms where they can freely tag our clips. In this way, much more, much more detailed and much more diverse access points to our collections may be created, than we could ever establish ourselves.

### **Conclusions**

It's clear that we will have to get used to regular feeds from outside of the professional context into our systems and catalogues, from the production environment, from software for automated indexing and from the user side, directly or indirectly.

In employing this metadata, in allowing it into our world we have to start regarding metadata creation as a common activity, which can be performed by professionals and non-professionals, experts and non-experts alike.

In general the use of professional cataloguing standards will decrease because of this. There will be more recall, but less precision. There will be satisfying search results instead of maximizing results, something we have always aimed at as professionals.

We will need to start regarding the catalogue as a collection of metadata description that will never be finished. The catalogue will become a process, an open source.

Because of all these developments- the 'semantic gap' will be bridged, the cataloguing and searching process will approach each other, because users will tag with the same words they would search with.

Professional cataloguers may feel then that these developments will cause them to lose a bit of their of their exclusiveness, that they will have to share their job, their expertise with non-experts, with systems and software even. This is probably true, but they're bound to have to develop a few very important new focuses too, like: Maintaining the metadata quality in the production environment (doing quality checks, advocate the use of controlled metadata); Help defining the algorithms from a user requirements point of view, and still, bring in the higher semantics; Controlling the information content and the context of digital audiovisual materials.

A final note. Digital archives have a great lot to do. The pressure of permanent digital production, the great piles of digitized legacy materials we are starting to build up: they all need to be exploited and made available to as much users as possible, on as much semantic levels as possible. Therefore - in my view- we should be happy with all the help that we can get, be it users, producers or systems and fully embrace the reinforcements they offer us.

*14<sup>th</sup> of October 2007,  
Annemieke de Jong  
Netherlands Institute for Sound and Vision  
Media Management Commission FIAT/IFTA  
FIAT/IFTA Conference Lisbon*