Requirements for practical multimedia annotation[*]

Joost Geurts, Jacco van Ossenbruggen, Lynda Hardman[**]

CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
Firstname.Lastname@cwi.nl

Abstract. Applications that use annotated multimedia assets need to be able to process all the annotations about a specific media asset. At first sight, this seems almost trivial, but annotations are needed for different levels of description, these need to be related to each other in the appropriate way and, in particular on the Semantic Web, annotations may not all be stored in the same place.
We distinguish between technical descriptions of a media asset from content-level descriptions. At both levels, the annotations needed in a single application may come from different vocabularies. In addition, the instantiated values for a term used from an ontology also need to be specified. We present a number of existing vocabularies related to multimedia, discuss the above problems then discuss requirements for and the desirability of a lightweight multimedia ontology.

1 Introduction

We motivate the need for processing complex semantic annotations of media assets with our work on the generation of multimedia presentations [11]. This assumes that a large number of semantically annotated media assets is available, a subset of these can be selected from a repository and can be assembled into a multimedia presentation tailored to a particular viewer.

Many fundamental problems related to multimedia metadata exist (see, for example, [12,6,8] for an overview). These problems fall into three categories. First, the relationships need to be made explicit between real world objects and digital media assets representing them, and between metadata expressing concepts in the domain and metadata about the media assets. Second, multimedia applications require a distinction between content-level annotations and (technical) descriptions of the media asset. Third, multimedia applications often require metadata from a variety of different vocabularies. Combining these into a single, coherent and flexible annotation scheme is a non-trivial and often overlooked problem.

In this paper we first present a number of vocabularies that describe different aspects of multimedia, ranging from descriptions of media types from a software engineering perspective to graphic design style characteristics. We devote a section to each of the three problems related to rich annotations of media assets. We then list a set of requirements for a lightweight multimedia vocabulary. The following section discuss whether or not all the requirements can be met simultaneously.

[*] This research was supported by the Dutch NWO/NASH and BSIK/MultiMediaN projects.

[**] Lynda Hardman is also affiliated with the Technical University of Eindhoven.

2 Multimedia Vocabularies

A number of vocabularies that deal at some level with multimedia content currently exist. The list of vocabularies presented here is not intended to be complete, but rather to illustrate how a domain shapes a vocabulary. These have all been relevant to our work on multimedia presentation generation and our Cuypers engine [10].

MPEG-7 [7] aims to standardize a core set of quantitative measures of audio-visual features and structures of descriptors and their relationships. MPEG-7 attempts to keep the application domain as broad as possible. This results in an elaborate standard in which a number of fields ranging from low level encoding scheme descriptors to high level content descriptors are merged. Because there is no explicit assumed use case MPEG-7 supports multiple ways of structuring annotations. In addition, domain dependent descriptors can be added if the default descriptors do not provide enough detail.

Dublin Core Element Set[1] is a standard for cross-domain information resource description. There are no fundamental restrictions to the types of resources to which Dublin Core metadata can be assigned. Dublin Core is a commonly used annotation scheme across different domains. It is small set of relations, identified by domain experts in the field of digital libraries.

VRA[2] is used to describe visual resources in the cultural heritage domain. Mostly these are visual representation of physical objects. To avoid confusion between annotation records about the object and its visual resources the difference between physical object and digital representation has been made explicit. Furthermore VRA defines a vocabulary to annotate material in which it makes suggestions to use terms from other vocabularies, such as AAT.

Media Streams [3] provides a detailed vocabulary to describe video content, for search and automatic editing. Media Streams uses a modular approach to support stratified annotations. Moreover the used vocabulary is general but precise in order to support a large domain and reduce implicit contextual knowledge.

Art and Architecture Thesaurus (AAT) is a thesaurus developed by Getty, partly to index their image collection. It focuses on terms from the art and architecture domains, and related vocabularies from Getty that are often used include the Union List of Artist Names (ULAN) and the Thesaurus of Geographic Names (TGN)[3].

MIME [5] was initially designed to allow mail tools to identify the media type of email content, but is now also used by web browsers and many other multimedia applications.

CSS[4] is a style sheet language that allows authors and users to attach style (e.g., fonts, spacing, and aural cues) to structured documents (e.g., HTML documents and XML applications). Although not specifically meant for multimedia we used CSS (and XSL Formatting Objects) as inspiration for our hypermedia formatting objects [11].

[1] `http://www.dublincore.org/documents/dces/`

[2] `http://www.vraweb.org/vracore3.htm`

[3] `http://www.getty.edu/research/conducting_research/vocabularies/`

[4] `http://www.w3.org/Style/CSS/`

Composite Capabilities/Preference Profiles (CC/PP [5]) describes the structure and vocabulary of profiles that can be used to describe device capabilities and user preferences. It is mainly used for content adaptation in mobile environments.

PREMO [4] defines a vocabulary which can be used to describe multimedia systems. It focuses on synchronization and dependencies from a system components perspective. PREMO is referenced in SRM-IMMPS [2], a reference architecture for multimedia generation system, as potential candidate to provide specifications how a presentation specification is rendered to be perceived by a user.

Modality Theory [1] aims at providing a taxonomy of output modalities which are used for creating multi-modal output in human computer interfaces. Modality theory is used to combine media assets in an effective way for the user. Modality theory defines the properties which define whether modalities can be combined.

Web Content Accessibility Guidelines [6] explain how to make Web content accessible to people with disabilities. One of the guidelines state to explicitly provide content for a different medium (the ALT attribute in HTML image tag) in case the original content cannot be served.

3 Relationships between topic, media assets and annotations

As in all metadata applications, there is the association between the metadata and the digital artifact being annotated. In multimedia, the digital artifact itself is often also “about” another (often non-digital) artifact. This results in a complex, at least four-way relationship between (1) the concept (physical artifact), (2) its annotation, (3) the digital media item and (4), its annotation. Figure 1 gives an example of such a relationship. In the upper left, we see the primary topic, a painting from the collection of the Rijksmuseum Amsterdam. Note that this represents the real, physical object, as it is hanging on the wall in the museum. In the lower left, there are a series of digital media items that are directly related to the painting: digital images of the painting in different sizes and resolutions, X-ray images of the same painting, etc. The right half represents the metadata[7], with, on top, the annotations describing the painting, and at the bottom the annotations describing the media items that depict the painting.

All multimedia applications need to deal with the bottom side of this figure to be able to show the media items (bottom left) and how to process the different media types (bottom right). Most of them also need access to information about the topic of discussion (the top right), for example to make decisions about the relevancy of the media items. Unfortunately, the distinction between the four quadrants is not always made, and, if it is, the relationship between the quadrants has not been standardized or is even not made explicit at all.

[5] http://www.w3.org/Mobile/CCPP/

[6] http://www.w3.org/TR/WAI-WEBCONTENT/

[7] Note that in some sense, the media items can also be thought as being metadata about the painting. In this paper, we stick with the more conventional definition of metadata.

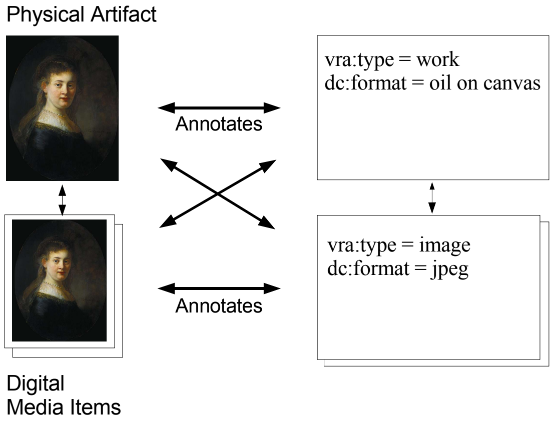

Fig. 1. Media Item and Annotation Record

4 Content-level versus media asset descriptions

As in all metadata applications, the metadata depends on its intended use. For multimedia, however, this is even more relevant since the content also needs to be described. In text-based applications, it is often sufficient to annotate only the generic properties of the whole asset (such as `dc:title` and `dc:creator`) and perform keyword search based on full-text information retrieval approaches. For such applications, simple and generic annotation schemes such as Dublin Core are often sufficient. In addition, it is often sufficient to employ only the annotation properties that describe the entire asset.

For non-textual assets, however, full text search is only an option if there are sufficient associated textual assets (cf. Google's image search based on surrounding text in HTML or video retrieval based on subtitles, closed captions or audio transcripts). In many other cases, however, content descriptions are inevitable. Unfortunately, content descriptions are much more problematic than media asset annotations. First, there is the issue of granularity: content descriptions are often not about the entire document, but only about a specific part (e.g. a specific shot in a film or a specific region in a picture) of it. Second, there is the issue of complexity and size of the annotations: stating the name of the author, the MIME type of the asset or its file size, is in general much simpler that describing its content using a formal description scheme. In this section, we describe both problems in more detail.

Granularity of annotation The granularity issue requires that the annotator is able to specify to which part of the asset an annotation is associated. For example, to express that the first 20 seconds of a video is shot in black and white, one need a way to address only the first 20 seconds. There are basically three levels in which this problem can be solved.

First, one can assume that it is the responsibility of the media assets data format to provide sufficient hooks to identify parts that need to be annotated. This is essentially the approach taken in the original HTML specification, where one could only address a portion of the document if that portion was already explicitly marked up with an identifier by an `<a name=` construct. While this may seem a simple solution from the point of view of the annotator, the major drawback is that it only works in cases in which every potential target of annotation, or if the annotator has (write) access to the underlying material and can add tags when required.

A second, more scalable approach is to move the problem from the media assets to the annotations themselves. In this case, the annotation schema does not only need to provide sufficient vocabulary to describe the content, but also the vocabulary to describe to which part of the asset these annotations apply. This is the approach taken by, for example, MPEG-7 and Media Streams.

The third approach is to solve the problem during the construction of the links that bind the annotations to the target asset. This moves the required vocabulary from the annotation schema to the language that describes the links. This is the approach taken for hyper links on the Web and semantic relations on the Semantic Web. Both assume that the target of the link is addressable using a single URI. Fragments of a resource are typically addressed by a so called fragment identifier (e.g. the part of a URI that comes after the '#'). Note that the meaning of the fragment identifier is dependent on the MIME type of the target resource, and many types still lack such a definition.

Note that independent of which of these three approaches is chosen, finding a flexible, effective and robust way to identify parts of multimedia assets is a non-trivial problem. For example, the need to annotate on multiple layers of granularity in the case of video annotation has been described by Aguierre Smith and Davenport in [9].

Issues related to content descriptions The problems related to multimedia content descriptions are well known and described in the literature (see, for example, [12]). Subscribers to the old cliché, "a picture tells more than a thousand words" might find it not surprising that describing the content of multimedia assets using a controlled vocabulary often leads to bulky and complex descriptions. In addition, they tend to be not very objective: content annotations vary significantly depending on the author of the annotations, and the intended use of the annotations.

The complexity and bulkiness of content annotations is also reflected by the vocabularies that are available for content descriptions. Commonly used examples include the thesauri Getty uses to annotate their image collection, which are themselves as complex and bulky as the annotations that use them. In addition, these vocabularies tend to be domain specific and their widespread use is often hindered by copyright issues.

Note that while the typical use case is that the terms in a vocabulary are used to annotate the media assets, a specific example of the reverse situation is found in many applications that focus on semantic descriptions, and use media assets to illustrate the concepts in their semantic network. A text-based example is the use of text labels on the Semantic Web. To allow Semantic Web applications to display a human-readable name for RDF-defined concepts (instead of their URI), RDFS has a predefined property `rdfs:label` that links such a label with a resource. An non-textual example is the use of the `foaf:depicts` in the FOAF project. While it may be initially intended to annotate images with information about the persons that they depict, its use is often the other way round: to find a suitable picture of a person that can be used to illustrate an automatically generated page in a FOAF network browser.

Media specific asset descriptions Even when we ignore the issues related to granularity issues and content annotation and stick to generic descriptions that apply to the entire asset, there are still a number of multimedia-specific issues that need to be addressed. First, in order to process an media asset, most multimedia applications need to have access to metadata that characterizes the system-oriented properties of the media asset. First of all, they need to be able to identify the type of media they are dealing with. An simple and commonly used vocabulary to identify media type is the MIME [5] media type hierarchy. Other attributes may depend on the media type, e.g. the aspect ratio may be a relevant property for visual media, but not for audio. Also the semantics of values of common attributes may depend on the media type. For a piece of text, the value of the `dc:creator` field will typically be interpreted as the author of the text, but for a video fragment the interpretation is not *a priori* clear. Also note that many media types already have a set of established metadata properties. For example, the properties of ID3 tags used by most MP3 application and JFIF tags added to JPEG images by digital photo cameras. Other metadata formats need to take into account and support such existing practices.

5 Annotation Schemes for Content-level metadata

The structure of an annotation is an often overlooked problem. In the digital library community, it is often assumed to be a flat list of attribute/value pairs. In the Semantic Web community, annotations are often assumed to be an instance of an ontology, of which the syntax structure is irrelevant. In the multimedia community, annotations are grouped in a hierarchy that mirrors to the structure of the media asset.

Making the structure of the annotations explicit by means of a machine processable schema could have advantages. It allows, for example, validation, moreover, a generic annotation tool could use the annotation structure as a template for driving an annotation form in its GUI [8]. Current Semantic Web tools provide little support for structured annotations. Most tools presents a list of RDF triples, of which the order is random or alphabetical at best. An annotation schema can be used to enforce syntactic relations between properties. For example, that begin-time and end-time should be grouped, or begin-time should be entered first. Rules defining required and optional properties or dependencies between properties can also be part of the schema, for example to define that, when there is a dc:publisher there should also be a dc:creator, but dc:publisher is optional. The schema could also enforce semantic constraints, for example by requiring values to come from a particular ontology. For example, that vra:material should be filled in with a value defined by a certain branch of the AAT thesaurus[8].

[8] Note that even for Semantic Web applications,it is not clear that RDF is the best way to encode annotation schemata. The syntactic examples given above suggest the use of XML Schema, while the semantic constraints could be better defined by RDFS or OWL.

6 Multimedia Annotation Requirements

Based on lessons learned from using the vocabularies described in section 2 and the problems discussed in section 3, 4 and 5, we can now list the requirements for a practical multimedia annotations environment.

Lightweight and extensible. We believe that the size and complexity of approaches such as MPEG-7 is a major issue that blocks a more widespread use. Even commercially developed vocabulary such as AAT and IconClass require extensive training before they can be used effectively. For many applications, a lean and mean multimedia core vocabulary with a small number of often used properties would be more appropriate. This core should however be sufficiently flexible to serve as a basis for more detailed, modular extensions. Such a modular approach not only allows for a small core, but also for relatively small extension modules.

Reuse existing vocabularies. For most frequently used properties a vocabulary already exists. The core should reuse existing vocabularies, rather than defining yet another one. By selecting core properties, arbitrary decisions will be unavoidable (for example, one might decide to include dc:title in the core at the cost of excluding vra:title). Mapping schemes should be provided to map commonly used vocabularies to one and other.

Relate Concept to Media Asset. The relationship between a concept and an asset should be made explicit and defined in an interoperable way. The nature of this relationship might dependent on the media type, and should be bidirectional: media-centric applications should be able to find the associated annotations, but knowledge-centric applications (e.g. FOAF) should be able to find the media assets from the annotations. Especially, vocabulary describing the often used "depicts" relation should be standardized, if only to provide a multimedia counterpart to the rdfs:label relation.

Structured annotations. The approach should allow arbitrary grouping of annotations, so that the structure of the media asset can also be used to organize the metadata. An explicit definition of the annotation structure in a schema would enable tools to validate annotations or could even drive the GUI of a generic annotation tool.

Unrestricted Fair Use. At least the core of a multimedia vocabulary should be free of fair use restrictions to allow for widespread use, including non-commercial applications.

Functional specification of media assets. The functional role of a media item should be made explicit. In case the context (e.g. device, user) requires an alternative asset this should not disrupt the rhetoric of the presentation.

7 Discussion

How realistic is it to expect that these requirements can be met in the near future? On the down side, the political situation does not look as bright. To what extent standards such as MPEG-7 will ever be widely adopted is uncertain, but its very existence seems to make many organizations reluctant to invest in a new, Semantic Web based, approach to multimedia annotation. At the moment, the field is still dominated by expensive and inflexible proprietary solutions that hardly contribute to the requirements of interoperability and general deployment of multimedia annotations.

On the positive side, the technical infrastructure of the Semantic Web seems to be a unique opportunity to move multimedia annotation from the realm of a selected professional specialists into the realm of widespread use by non-professional end-users. In addition, the Semantic Web has the potential to provide the required interoperable annotation languages that are needed to allow cross-platform annotations and the ubiquitous use of URIs might help to solve the current problem of defining explicit links between artefacts, media assets and their annotations.

Meeting the requirements sketched in this paper would require a joint effort of the communities involved such as the Semantic Web, multimedia and digital library communities. A possible first step is to use existing vocabularies to develop a minimal but common agreed upon annotation schem for multimedia assets.

References

1. N. O. Bernsen. Defining a Taxonomy of Output Modalities from an HCI Perspective. *Computer Standards and Interfaces*, 18:537–553, 1997.
2. M. Bordegoni, G. Faconti, M. Maybury, T. Rist, S. Ruggieri, P. Trahanias, and M. Wilson. A Standard Reference Model for Intelligent Multimedia Presentation Systems. *Computer Standards & Interfaces*, 18(6-7):477–496, December 1997.
3. M. Davis. *Readings in Human-Computer Interaction: Toward the Year 2000*, chapter Media Streams: An Iconic Visual Language for Video Representation., pages 854–866. Morgan Kaufmann Publishers, Inc., 1995.
4. D. J. Duke, I. Herman, T. Rist, and M. Wilson. Relating the primitive hierarchy of the PREMO standard to the standard reference model for intelligent multimedia presentation systems. *Computer Standards & Interfaces*, 18(6-7):525–535, December 1997.
5. N. Freed and N. Borenstein. Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types. RFC 2046, November 1996. Obsoletes: 1521, 1522, 1590 Category: Standards Track.
6. W. I. Grosky. Managing multimedia information in database systems. *Communications of the ACM*, 40(12):72–80, December 1997.
7. ISO/IEC. Overview of the MPEG-7 Standard (version 8). ISO/IEC JTC1/SC29/WG11/N4980, Klagenfurt, July 2002.
8. A. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga. Ontology-based Photo Annotation. *IEEE Intelligent Systems*, 16(3):66–74, May-June 2001.
9. T. G. A. Smith and G. Davenport. The Stratification System. A Design Environment for Random Access Video. In *ACM workshop on Networking and Operating System Support for Digital Audio and Video.*, San Diego, California, 1992.
10. J. van Ossenbruggen, J. Geurts, F. Cornelissen, L. Rutledge, and L. Hardman. Towards Second and Third Generation Web-Based Multimedia. In *The Tenth International World Wide Web Conference*, pages 479–488, Hong Kong, May 1-5, 2001. IW3C2, ACM Press.
11. J. van Ossenbruggen, J. Geurts, L. Hardman, and L. Rutledge. Towards a Formatting Vocabulary for Time-based Hypermedia. In *The Twelfth International World Wide Web Conference*, pages 384–393, Budapest, Hungary, May 20-24, 2003. IW3C2, ACM Press.
12. J. van Ossenbruggen, F. Nack, and L. Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part I). *IEEE Multimedia*, 11(4):38–48, October – December 2004. based on http://ftp.cwi.nl/CWIreports/INS//INS-E0308.pdf.