Multimedia Annotations and the Semantic Web

Jacco van Ossenbruggen[1], Giorgos Stamou[2] and Jeff Z. Pan[3]

[1] Centrum voor Wiskunde en Informatica, Kruislaan 413, NL-1098 SJ Amsterdam, The Netherlands
[2] Department of Electrical and Computer Engineering, National Technical University of Athens, Zographou 15780, Greece
[3] School of Computer Science, The University of Manchester, Manchester, M13 9PL, UK

Abstract. Multimedia and the Semantic Web: in theory, it is a perfect match. The Semantic Web, on the one hand, provides a stack of languages and technologies for annotating Web resources, enabling machine processing of metadata describing semantics of web content. Multimedia applications, on the other hand, require metadata descriptions of their media items to facilitate search and retrieval, intelligent processing and effective presentation of multimedia information. This need for multimedia metadata was recognized by the media industry long ago. Semantic Web technologies, however, still play a very minor role within multimedia applications and most approaches employ non-RDF based techniques. This paper describes a number of current approaches to multimedia metadata and provides an inventory of the open issues to achieve a practical integration of multimedia metadata into the Semantic Web.

1 Introduction

Those who deal with multimedia, from professional archivists to amateur photographers, are faced with daunting problems when it comes to storing, annotating and retrieving media items from multimedia repositories, whether via the Web or otherwise. Although the standardization activities of ISO (and other) communities (MPEG-7, MPEG-21, Dublin Core etc) [5, 10–12, 17, 22] have provided standards for describing content, these standards have not been widely used, mainly for the following reasons. Firstly, it is difficult, time-consuming, and thus very expensive to manually annotate multimedia content. Secondly, many organizations feel that the complexity of many standards makes multimedia annotation unnecessarily difficult. Thirdly, there is little incentive for organizations to provide, for example, MPEG-7 metadata because there are insufficient applications that would benefit from its use.

We believe that these problems could be solved by merging and aligning existing practices in multimedia industry with the current technological development of the Semantic Web. First, such integration would give metadata providers immediate payoff because they could directly benefit from the Semantic Web software that is (publicly) available. Second, it would enable the deployment

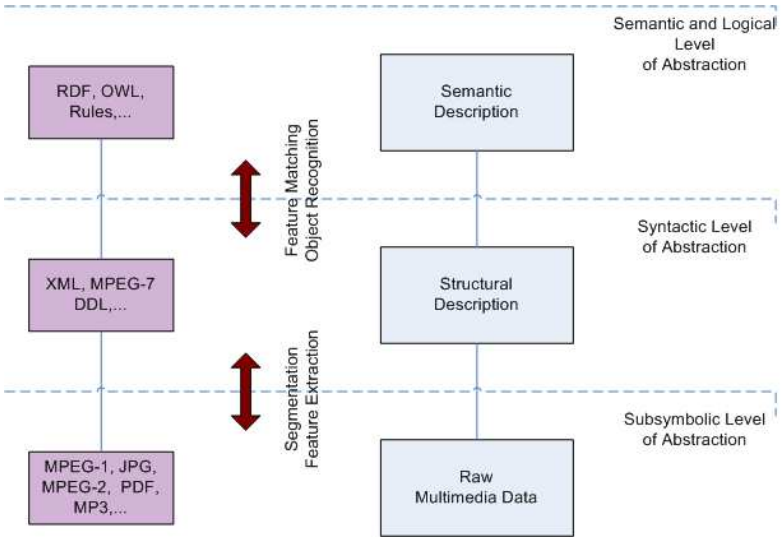

Fig. 1. Abstraction levels of multimedia annotation

of more intelligent applications to reason over multimedia metadata in a way that is currently not possible because current multimedia metadata standards are usually (XML) syntax-oriented, and thus lack a formal semantics. Third, the "open world" approach of the Semantic Web would simplify the integration of multiple vocabularies from different communities. Finally, it could provide small, simple but extensible vocabularies. These vocabularies should be suitable for private use (e.g. simple annotation of online photo albums à la Flickr) but at the same time be sufficiently flexible to be extended for more complex and professional annotation tasks.

2 Multimedia semantic annotation

The information conveyed by a multimedia document can be formalized, represented, analyzed and processed in three different levels of abstraction: the subsymbolic, the symbolic and the logical (see Figure 1).

The subsymbolic level of abstraction covers the raw multimedia information represented in well known formats for video, image, audio, text, metadata, etc. Note that these are typically binary formats, typically optimized for compression and streaming delivery. They are not necessarily well-suited for further processing that uses, for example, the internal structure or specific features of the media stream. To address this issue, one can introduce a level of abstraction, the middle layer in Figure 1, which provides this information. This is the approach of MPEG-7, which allows one to use the output of feature detectors, (multi-cue)

segmentation algorithms, etc. to provide a structural layer on top of the binary media stream. Note that information on this level is typically serialized in XML.

The problem with this XML-based, structural layer is that the semantics of the information encoded in the XML is specified only in the specification of, for example, the MPEG-7 standard and needs to be hard wired into the code by the programmer of the MPEG-7 application software. It also makes it hard to re-use this data in environments that are not based on MPEG-7, or to integrate non-MPEG metadata in an MPEG-7 application.

To address this, one could simply replace the middle layer by another one that is open and has formal, machine processable semantics. This, however, would not take advantage of existing XML-based metadata, and, more importantly, ignore the advantages of an XML-based structural layer (more on that later). So, rather than replacing the middle layer, a solution is to add a third layer that provides the semantics for the middle layer.

These semantics are mappings between the structured information sources and a formal knowledge representation of the domain, for example in OWL. In this layer, the implicit knowledge of the multimedia document description can be made explicit and reasoned upon, for example to derive new knowledge not explicitly present in the middle layer.

Several standards have been proposed and used in the literature for the representation of multimedia document descriptions (Dublin Core, MPEG-7, MPEG-21 etc), mainly operating in the middle layer of Figure 1. The stack of RDF-based languages and technology provided by the W3C community are well-suited to the formal, semantic descriptions of the terms used in the middle layer. However, since they often lack the structural advantages of the XML-based approach, a combination of the above standards seems to be the most promising way for multimedia document description in the near future [8, 9, 13, 14, 20, 23, 24].

3 Open issues

To realize such an integrated scenario, several open issues need to be addressed.

Interoperability and tool support The main problem we see is that the Semantic Web technologies do not interoperate with existing approaches in the multimedia production chain. In the longer term, integration of Semantic Web technologies in the major multimedia tool is essential. In the shorter term, we need to show how RDF-based software can take advantage of popular existing, non-RDF metadata such as ID3 tags in MP3 music files[4], EXIF metadata added to JPEG images by digital cameras[5], informal tagging of images (with terms from so called 'folksonomies') etc.

The problem mentioned above of aligning Semantic Web-based approaches with MPEG-7 is also a major issue, for a thorough comparison of and a list

[4] As is done by the content handlers of Kowari, `http://kowari.org/`
[5] As is done by JpegRDF, `http://sourceforge.net/projects/jpegrdf`

of the open issues in integrating the MPEG-7 and Semantic Web approaches, see [15, 26].

Linking media data with metadata Since metadata is just data *about* other data, the link between the metadata and the target media item is of crucial importance. On the Semantic Web, the link between the two is simple: all you need is the URI of the media item which you use as the value of the `rdf:about` attribute somewhere in your metadata:

```
<rdf:Description rdf:about='http://www.example.com/mymovie.mpg'
  dc:title='My First DVD Soccer Movie'/>
```

This approach, however, makes a number of assumptions that do not always hold in the multimedia domain.

First of all, it assumes that the thing being annotated can be addressed by a commonly agreed upon URI scheme. While this may be a safe assumption for HTML and XML-based resources, this is not the case in multimedia. The example above works because the dc:title is an annotation that applies to the entire resource. Annotations that apply only to a part of the resource are much harder. Imagine you would like to provide annotations for the 7th scene on that DVD, or for a specific sequence of frames, a specific region in a frame, a specific object (the ball, a specific player), a part of the sound track (the audience singing), etc, etc. Standardizing URIs for such targets is not trivial. For example, when sticking with the current URL schemes, it requires standardizing a powerful fragment identifier[6] syntax for all common multimedia MIME types used.

Second, the example above assumes that the link can be embedded within the metadata. A disadvantage of this approach is that it is geared to 1-to-1 relations and it becomes harder to model n-to-m relations [16]. On a more practical level, it becomes harder to associate existing annotations with other media items, since this requires modification of the original RDF. This might be unwanted from a maintenance perspective or downright impossible if the person creating the new link has no write access on the metadata. From the (pre-Web) hypertext literature (see [25] for an overview) we know that links between two pieces of information can be embedded at the source (as is the case in, for example, HTML), in the target, or in an independent location (often called a link base). All three solutions have different characteristics when it comes to flexibility and complexity, so the key issue is to know what solution to use in what context.

Third, the example assumes that the URI unambiguously identifies the target. However, in many multimedia resources the URI of the digital artifact is used also for the physical object it represents. For example the URI of an image of a painting is also used to for the painting itself. Vocabularies such as the VRA Core 3.0 [27] make the distinction explicit by distinguish metadata records describing the "work" from records that are about the "image". How to link "work" records to associated "image" records remains, however, unspecified.

[6] Informally, fragment identifiers are what comes after the '#' in a URI. So in `http://example.com/index.html#section1`, `section1` is the fragment identifier. Note that the syntax and semantics of fragment identifiers depend on the MIME type of the resource

Vocabularies for multimedia annotations While it is true that the Semantic Web allows everyone to create his or her own vocabulary, sharing and reusing information benefits from having only a few widely used vocabularies for a specific purpose, and having these vocabularies widely available in a Semantic Web compatible format. Good examples of such vocabularies are, however, still hard to find[7].

Part of the problem is that many vocabularies for multimedia predate the Semantic Web. Another explanation is that describing the content of audiovisual material in general requires a large vocabulary basically covering the entire (visual) world around us. Developing such vocabularies is a long and costly process, and organizations that have invested large sums of money in creating such vocabularies are often not willing to make the results publicly available on a royalty free basis. Well known examples (see [7] for a more extensive overview) include Getty's Art and Architecture Thesaurus [6], that needs to be licensed and is not available in RDF. Another example is Mark Davis's MediaStreams iconic ontology [4] developed in the early nineties, also predating the Semantic Web. In addition, many national audiovisual archives (e.g. INA in France and Beeld en Geluid in The Netherlands) have developed in house vocabularies to describe and index the large quantities of audiovisual material they need to archive, and these vocabularies have also been developed long before the Semantic Web took shape.

Uncertainty in multimedia annotations Several issues of multimedia information systems are often subject to uncertainty and imprecision. The representation of multimedia annotations, the automatic extraction of this annotation, the retrieval of multimedia documents etc are processes that involve uncertainty and inconsistency in several levels. For example, the extraction (automatic or manual) of the key entities that semantically describe the multimedia document is always a matter of degree. Moreover, the visual characteristics of an object (i.e. its color) possess usually imprecise information with its accuracy being a matter of the measurement process, though most of times is not really important (usually in the retrieval process only linguistic terms like "red" are needed). For the above reasons, theories and methods covering the framework of uncertainty (fuzzy logic, probabilistic reasoning, evidence theory, neural networks etc) are very important and sometimes crucial in multimedia information systems. Although several papers have been published in this area [1,21], the issue remains open for further research.

Datatypes The output of many multimedia feature detectors is described by complex datatypes. MPEG-7, for example, uses XML Schema to describe multimedia objects, such as video, audio and images, as instances of XML schema datatypes [2].

[7] During the first Workshop on Multimedia and the Semantic Web at ESWC 2005 on Crete, the participants agreed to collect publicly available multimedia ontologies on a central website, `http://www.acemedia.org/aceMedia/reference/resource/`.

In the Semantic Web standards, such as RDF and OWL, datatypes are defined in a more formal way [3]. More specifically, a datatype (such as boolean) is characterized by a lexical space (such as {T,F,1,0}), a value space (such as $\{true, false\}$) and a lexical-to-value mapping (such as {T$\mapsto$ $true$, F$\mapsto$ $false$, 1$\mapsto$ $true$, 0$\mapsto$ $false$}). Although RDF and OWL only allow some built-in XML Schema simple types, OWL-Eu [19] has been designed to support user-defined XML Schema simple types based on restriction and union. Furthermore, OWL-E [18] (the n-ary extension of OWL-Eu) supports user-defined datatype predicates.

An obvious issue here is that MPEG-7 also requires the structuring support of XML Schema complex types, which are not compatible with the above RDF/OWL datatype model. The main issue is, however, whether it is proper to introduce the structuring support into datatypes, or simply use concept languages provided by OWL to represent the structure of multimedia objects. Another issue is that even XML Schema complex types are not enough for MPEG-7, which extends XML Schema datatypes with array and matrix datatypes (among others), with both fixed size and parameterized size.

4 Conclusions

In this paper we analyzed the open problems for enabling multimedia metadata on the Semantic Web. Obviously, solving these problems will require effort from both the multimedia and the Semantic Web communities. We feel, however, that the Semantic Web community has a special obligation to prove that their models and techniques can add sufficient added value to convince the multimedia content owners to go beyond the current, XML-based, approaches. From this perspective, this paper is "a call to arms" to the Semantic Web community to address the following issues.

First, we should show how our RDF-based environments can interoperate with current, non-RDF metadata practices in the multimedia field, for example by developing RDF tools that can handle embedded metadata in MP3 and JPEG files, or build upon XML-based approaches such as MPEG-7.

Second, we should collect and publish example multimedia vocabularies using Semantic Web languages, and show that annotating multimedia data with these vocabularies is not only practical but also provides more useful functionality than that provided by current multimedia metadata tools.

Third, we should be able to flexibly attach metadata to media resources. To be able to link metadata to the appropriate part of the target media item, fragment identifier schemes need to be standardized and widely implemented for a wide variety of commonly used media types. We also need a standard way of describing the link between a piece of metadata and its target media item independently of the media item and the meta data. In addition, we need a common way to discriminate between, and relate the metadata about, a physical object and the metadata about a (digital) representation of that object.

Fourth, we need to extend our languages and tools to be able to formally express the uncertainty and imprecision inherent to many statements about multimedia data.

Last but not least, we need to extend the Semantic Web datatype formalism to deal with the often complex media types that are required to express statements about specific multimedia features. We need to harmonize the structure information, specified by XML Schema complex types, about multimedia objects in the symbolic level and the semantic descriptions, represented by concept and datatype constraints, in the logical level.

Even when all of the above issues have been solved, multimedia annotation will remain a difficult, time consuming and expensive process. The question is whether we can develop the required standards in a way that reduces, and not adds, to the complexity of the task, and develop the tools and applications with an added value that makes multimedia annotation payoff in practice.

5 Acknowledgments

We wish to thank our colleagues Joost Geurts, Frank Nack and Lynda Hardman for their contributions to this work. Part of this work was funded by the European Knowledge Web and Dutch MultimediaN and NWO NASH projects.

References

1. Special issue on management of uncertainty and imprecision in multimedia information systems. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems,*, Volume 11, February 2003.
2. P. V. Biron and A. Malhotra. Extensible Markup Language (XML) Schema Part 2: Datatypes – W3C Recommendation 02 May 2001. Technical report, World Wide Web Consortium, 2001. http://www.w3.org/TR/xmlschema-2/.
3. J. J. Carroll and J. Z. Pan. XML Schema Datatypes in RDF and OWL. Technical report, W3C Semantic Web Best Practices and Development Group, Nov 2004. Editors' Draft, http://www.w3.org/2001/sw/BestPractices/XSCH/xsch-sw/.
4. 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.
5. Dublin Core Community. Dublin Core Element Set, Version 1.1, 2003. ISO Standard 15836-2003 (February 2003), http://www.niso.org/international/SC4/n515.pdf; NISO Standard Z39.85-2001 (September 2001), http://www.niso.org/standards/resources/Z39-85.pdf; CEN Workshop Agreement CWA 13874 (March 2000), http://www.cenorm.be/isss/cwa_download_area/cwa13874.pdf.
6. Getty Research Institute. Art & Architecture Thesaurus (Online). http://www.getty.edu/research/tools/vocabulary/aat/, 2000. Version 2.0.
7. J. Geurts, J. van Ossenbruggen, and L. Hardman. Requirements for practical multimedia annotation. In *Workshop on Multimedia and the Semantic Web*, pages 4–11, May 2005.

8. J. Hunter. Adding Multimedia to the Semantic Web — Building an MPEG-7 Ontology. In *International Semantic Web Working Symposium (SWWS)*, Stanford University, California, USA, July 30 - August 1, 2001.
9. J. Hunter, J. Drennan, and S. Little. Realizing the hydrogen economy through semantic web technologies. *IEEE Intelligent Systems Journal*, January 2004.
10. ISO/IEC. Overview of the MPEG-7 Standard (version 6.0). ISO/IEC JTC1/SC29/WG11/N4980, Pattaya, December 2001.
11. ISO/IEC. Text of ISO/IEC 15938-5/FDIS Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Schemes. ISO/IEC JTC 1/SC 29/WG 11/N4242, Singapore, September 2001.
12. ISO/IEC. MPEG-21 Overview v.5. ISO/IEC JTC1/SC29/WG11/N5231, Shanghai, October 2002.
13. A. Jaimes and J. R. Smith. Semi-automatic, data-driven construction of multimedia ontologies. *Proc. IEEE Intl. Conf. on Multimedia and Expo (ICME)*, March 2003.
14. S. Little and J. Hunter. Rules-b-example - a novel approach to semantic indexing and querying of images. *In 3rd International Semantic Web Conference (ISWC2004)*, November 2004.
15. F. Nack, J. van Ossenbruggen, and L. Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part II). *IEEE Multimedia*, 12(1):54–63, January – March 2005. based on http://ftp.cwi.nl/CWIreports/INS//INS-E0309.pdf.
16. Natasha Noy and Alan Rector. Defining N-ary Relations on the Semantic Web: Use With Individuals. Work in progress. W3C Working Drafts are available at http://www.w3.org/TR, 21 July 2004.
17. NewsML. The NewsML Home Page. http://www.newsml.org/, 2000.
18. J. Z. Pan. *Description Logics: Reasoning Support for the Semantic Web*. PhD thesis, School of Computer Science, The University of Manchester, 2004.
19. J. Z. Pan and I. Horrocks. OWL-Eu: Adding Customised Datatypes into OWL. In *Proc. of Second European Semantic Web Conference (ESWC 2005)*, 2005.
20. G. Stamou and S. Kollias (eds). *Multimedia Content and the Semantic Web: Methods, Standards and Tools*. John Wiley & Sons Ltd, 2005.
21. G. Stoilos, G. Stamou, V. Tzouvaras, J. Pan, and I. Horrocks. A fuzzy description logic for multimedia knowledge representation. *Multimedia and the Semantic Web Workshop, European Semantic Web Conference,*, pages pages 12–19, May 29-June 1 2005.
22. The TV-Anytime Forum. The TV-Anytime Forum Home Page. http://www.tv-anytime.org/.
23. R. Troncy. Integrating Structure and Semantics into Audio-visual Documents. In *Second International Semantic Web Conference (ISWC2003)*, pages 566 – 581, Sanibel Island, Florida, USA, October 20-23, 2003. Springer-Verlag Heidelberg.
24. C. Tsinaraki, P. Polydoros, and S. Christodoulakis. Interoperability support for ontology-based video retrieval applications. *In Proceedings of Third International Conference on Image and Video Retrieval (CIVR),*, pages 582–591, July 21-23 2004.
25. J. van Ossenbruggen, L. Hardman, and L. Rutledge. Hypermedia and the Semantic Web: A Research Agenda. *Journal of Digital Information*, 3(1), August 2002.
26. 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.
27. Visual Resources Association. Visual Resources Association Website.