A Common Multimedia Annotation Framework for Cross Linking Cultural Heritage Digital Collections

Hennie Brugman[1], Véronique Malaisé[2], Laura Hollink[2]

[1]Max Planck Institute for Psycholinguistics
P.O. Box 310
6500 AH Nijmegen, Netherlands

[2]Free University
De Boelelaan 1105
1081 HV Amsterdam, Netherlands

E-mail: hennie.brugman@mpi.nl, vmalaise@few.vu.nl, laurah@few.vu.nl

Abstract

In the context of the CATCH research program that is currently carried out at a number of large Dutch cultural heritage institutions our ambition is to combine and exchange heterogeneous multimedia annotations between projects and institutions. As first step we designed an Annotation Meta Model: a simple but powerful RDF/OWL model mainly addressing the anchoring of annotations to segments of the many different media types used in the collections of the archives, museums and libraries involved. The model includes support for the annotation of annotations themselves, and of segments of annotation values, to be able to layer annotations and in this way enable projects to process each other's annotation data as the primary data for further annotation. On basis of AMM we designed an application programming interface for accessing annotation repositories and implemented it both as a software library and as a web service. Finally, we report on our experiences with the application of model, API and repository when developing web applications for collection managers in cultural heritage institutions.

1. Introduction

The Dutch national research program CATCH (Continuous Access to Cultural Heritage) [1] currently consists of ten research projects, each working on one of the three following research themes: semantic interoperability, knowledge enrichment, and personalization. Each project is hosted by a large Dutch cultural heritage institution to help solving some of their existing problems related to the construction and exploitation of digital collections, and uses the actual Institution's collections as use case for performing their research. Among these institutions are for example the Rijksmuseum in Amsterdam, the Dutch National Archive, the Dutch National Library and the archive of the Dutch public broadcasting corporations (Netherlands Institute for Sound and Vision). CATCH aims at developing methods and tools to help collection managers make their collections more accessible. It is an explicit objective to make these methods and tools applicable across the heterogeneous collections of the participating cultural heritage institutions.

Annotation in some form plays a role in all ten CATCH projects. These annotations range from automatic semantic annotation of text, manual or automatic transcription of scanned handwritten manuscripts or speech, to manual annotation of images. These annotations associate a given descriptive (semantic or primitive) value with segments of documents (usually multimedia resources) from one of the ten museum, library and archive collections involved.

Collection integration could benefit a lot from a (virtual) repository of these heterogeneous annotations, generated externally from the related institution's official collection and metadata: this repository could be used for searching cross-institution's resources and segments of them, but also for adding different layers of semantics to these. This annotation repository would be even more powerful if we would allow annotations to refer to a (set of) controlled vocabulary that is shared by the same community as the annotation repository. This repository of vocabularies should contain both the controlled vocabularies and the mappings between elements of the different vocabularies. Within the CATCH community software developers actually collaborate on such a Repository of Vocabularies and Mappings.

For the realization of this shared annotation repository and the modeling of its content we need a formalism that meets the following requirements (for a discussion of general requirements for multimedia annotation, see Geurts et al. (2005)). The formalism:

1. Enables to describe and implement the anchoring of annotations to segments of resources of all the types used within the CATCH project (plain and semi-structured text, like XML/HTML data, images including

[1] CATCH - http://www.nwo.nl/catch

scanned handwriting, video, audio, music). Of course, anchoring to complete digital resources should also be covered.

2. Enables to associate a semantic annotation (annotation based on an ontology or a controlled vocabulary element) as well as annotations of primitive data types (such as date, string, etc) with the anchor.
3. Enables to create annotations of annotations: for example, a video segment can be annotated, making explicit that this segments contains a Gesture of a certain type. The Gesture annotation, in turn, should be able to carry further annotations, for instance specifying a particular Hand shape.
4. Enables to create annotations of annotations' segments. For example, segments of a scanned handwriting can be annotated with a text transcription, individual words of which should further be accessible for annotations, for instance with geographical concepts from an ontology.
5. Enables to include project and media specific extension, for particular types of annotation (these different annotation sub-models should be compatible with and specialize the general one).
6. It should be possible to pre-define annotation schemes or templates for specific annotation projects.
7. Queries on a set of annotations should be possible both at a general level and at more specific levels (through the use of inferencing).
8. The annotation model should be expressive and simple at the same time. It should be easy to learn and apply to a specific annotation related task or project.
9. The model should refer to or specialize existing annotation models, schemes and standards such as the Dublin Core[2] or the TEI[3]. It should function as 'glue', rather than try to cover every case by its own.

Most of these requirements have at least been already partly underlined by (Geurts et el 2005) and others, and are often met by existing annotation formalisms. However, the combination of all of them is challenging. For some of the requirements we were not able to find examples of models covering them: especially the ambition to be able to layer annotations by supporting annotation of annotations and annotation of annotation segments seems unaddressed.

The next section (section 2) discusses existing annotation models and their drawbacks in terms of the above listed requirements. In section 3, we present the annotation model that we have developed to meet these requirements, and several of its specializations. We proceed in section 4 with a discussion of the implementation of the model as a software library and a web repository, as well as the exploitation of these in the context of the annotation infrastructure that is currently being built within CATCH. We conclude in section 5 with a first evaluation of our model and its implementation, and a presentation of our plans for the future.

[2] Dublin Core - http://dublincore.org/documents/dces/
[3] Text Encoding Initiative - http://www.tei-c.org/

2. Related work

A number of different communities have been developing annotation formalisms for multimedia resources for a long time. Some of these formalisms made it into standards carried by organizations like ISO or the W3C.
For an overview of formalisms playing a role in the Semantic Web community, see for example (Troncy, 2007). Focus here is on annotation of either complete digital resources (usually not segments of them) or web pages. The overview in (Troncy, 2007) explicitly lists the supported media types per annotation vocabulary discussed. None of the vocabularies supports all of the media types that we need to deal with, because none of them offers the possibility to explicitly and easily define coordinates of a given annotation in the context of original resources of these media types. In general, the semantic web community's main interest is in the semantics of annotation values, not in the anchoring of annotations to multimedia resources.
The linguistic annotation community mainly focuses on annotation of texts and time series data (audio and video recordings). For these media types considerable effort was spent on developing models, principles and standards see for example (Bird & Liberman, 2001), standoff annotation, (Romary & Ide, 2007), (Brugman, 2003). For linguistic annotation, anchoring to text or to a media time axis is much more important and is therefore usually carefully modeled. On the other hand, most of the time less attention is paid to modeling of (semantic) restrictions on annotation values.
Finally, the media industry has spent a lot of effort on standardization of annotations, with MPEG-7 (ISO/IEC, 2002) as the best example. Although MPEG-7 supports anchoring to a wide number of media types in complex ways, not all of our required media types are supported.

Although most of the existing models allow annotation by means of simple data types or with elements from controlled vocabularies or ontologies, as far as we know none provides a straightforward way of defining annotation values and anchors in an annotation in such a way that they can be further annotated. Moreover, many models have been created with specific use cases in mind (e.g. VRA[4] [5] for visual objects description, MPEG-7 for audio, visual and audiovisual resources and CIDOC-CRM (Crofts et al, 2007) for an event-centered description of museum collections) that makes them

[4] http://www.vraweb.org/projects/vracore4/index.html
[5] for VRA in RDF-OWL see: http://www.w3.org/ /2001/sw/BestPractices/MM/vra-conversion.html

either too specialized or too complex for our use cases. It is not realistic to expect the projects and institutions in a project like CATCH to adopt a complex standard like MPEG-7 to share their relatively simple annotations. Therefore, we designed the Annotation Meta Model, which we present in more details in the following section.

3. AMM

Different communities often use the terms *annotation* and *metadata* with slightly different and overlapping meanings. This gives rise to many misunderstandings and debates. For the context of this paper we use *annotation* in the widest possible sense: any association of a value with a resource or part of a resource. This definition includes what is often called resource metadata, as well as "Web 2.0" tagging and descriptions of resource bundles.

We defined the Annotation Meta Model (AMM) in OWL[6] (using Protégé[7]). This choice has not so much to do with OWL's reasoning possibilities, but is motivated by the observation that RDF, RDFS and OWL seem especially suited as modeling languages for our problem. They provide solutions for some of our requirements out of the box: class and property inheritance, expressive and explicit constraints (like domains and ranges for properties) or seamless integration of semantic annotation values. Furthermore, structurally complex annotation models as we find in for example linguistic annotation typically use graph structures (as RDF does) instead of hierarchical ones (as in XML). Many tools and models in linguistic annotation are inspired by Annotation Graphs (Bird &Liberman, 2001).

We defined the Annotation Meta Model as a very simple model that can be extended to cover specific needs of communities or projects. This core model consists of only 11 classes and 23 properties, several of which are each other's inverse. It focuses on generic modeling of *anchoring to annotatable objects*, the objects that can be annotated, and offers some specializations to deal with anchoring to a number of different media types: images, text and temporal multimedia documents, like TV programs. Project specific annotation schemes can be defined by means of sub classes and sub properties of the AMM core classes and properties.

3.1 Core AMM

The central AMM class *AnnotatableObject* represents an object that can be annotated with either features (semantic annotations) or data properties (primitive datatype annotations) (see figure 1). Preferably, these are existing properties from vocabulary standards like for example Dublin Core or VRA.

AnnotatableObjects can be *anchored to* other AnnotatableObjects. E.g. an annotated image region is

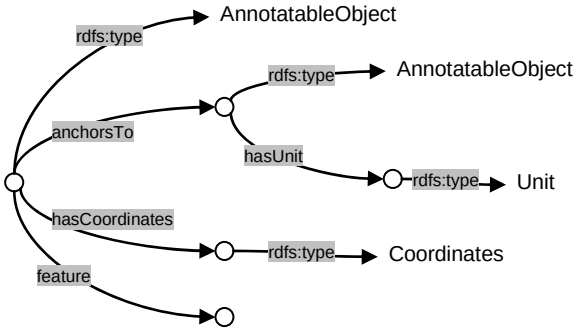

Figure 1: graph representation of the core AMM classes and properties

anchored to a complete (annotated) image. Also, the features of an AnnotatableObject can be AnnotatableObjects again. E.g. an image can be annotated with a *TextObject*, where (segments in) the text can be further annotated.

Finally, an AnnotatableObject can have *Coordinates*, indicating the boundaries of the object with respect to the object it is anchored to. These boundaries are specified in terms of *Units* associated with the anchor (e.g. pixels in case of some *ImageObject*, characters for a *TextObject*, etc).

For three basic resource types the Annotation Meta Model includes subclasses of AnnotatableObject: *ImageObject*, *TextObject* and *TimeSeriesObject*. Each of these subclasses is associated with its own subclass of Coordinates: spatial regions for image data, begin and end character offset for text data and begin and end time for time series data.

With a limited set of classes and properties, all of our requirements seem to be met. To test this we applied the model to a wide range of annotation cases.

3.2 AMM models for annotation use cases

To test and fine tune AMM we defined special models for a wide range of annotation use cases by sub classing the core AMM classes and properties. These cases are: automatically generated semantic annotations of texts, text transcription of scanned handwriting images, annotated image regions, complex linguistic annotation of co-occurring speech and gesture, and syntactic annotation of text. They are mainly contributed by member projects of the CATCH community, some cases come from the domain of linguistic annotation. To test the representative power of our model, for each of these special cases we manually created instance data.

As first illustration we describe the case of scanned handwriting annotation (see figure 2), inspired by the CATCH project called SCRATCH. The scanned handwriting document with title "handwriting.jpg" is

[6] http://www.w3.org/2004/OWL/
[7] Protégé ontology editor - http://protege.stanford.edu/

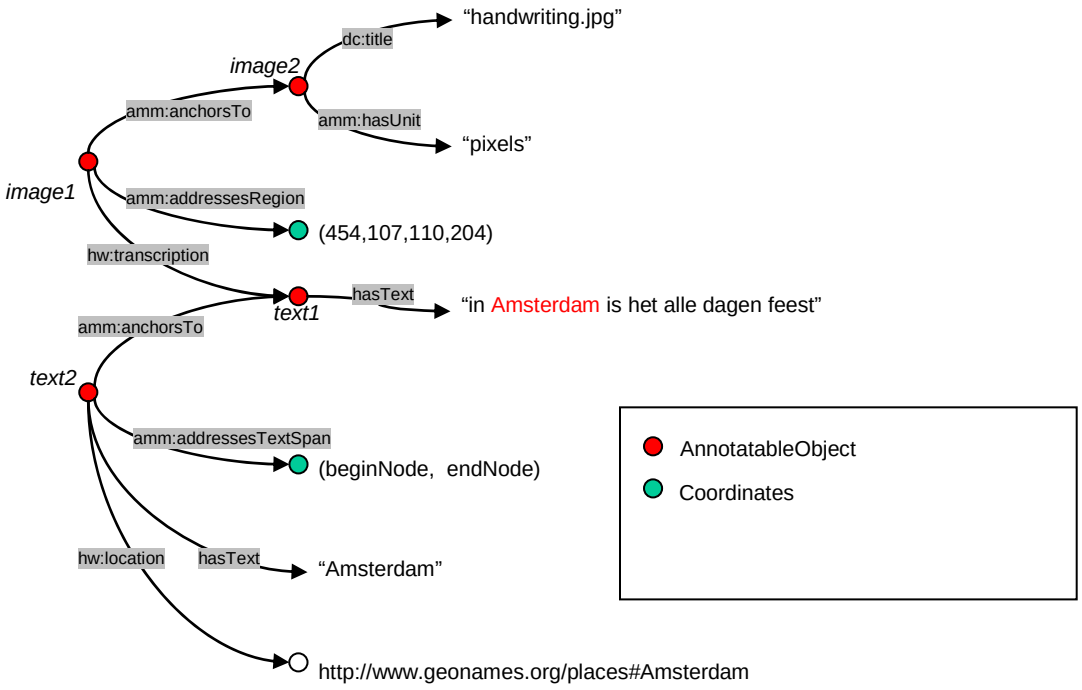

Figure 2: graph representation of annotation of scanned handwriting images

represented by an AnnotatableObject *image2*. AnnotatableObject *image1* represents a region of this image, for example a bounding box enclosing one line of written text. *image1* has a feature 'transcription' that has the type *amm:TextObject* as its range of possible values. Segments of the text contained in the TextObject can be further annotated. The property 'location' associates a geographical concept with the text segment between beginNode and endNode. Note that to support this special annotation case we only had to define the sub properties 'location' and 'transcription'. ('amm:addressesRegion' and 'amm:addressesTextSpan' are sub properties of amm:hasCoordinates in the core model).

The second illustration deals with a digitized video recording of speech accompanied by gestures. We represent complex annotations of a gesture overlapping a speech utterance in time (figure 3). Additional annotations for part of speech and hand shape are attached to an utterance transcription and a gesture annotation respectively. SpeechUtterances and Gestures are represented as special cases of TimeSeriesObjects. SpeechUtterances have a 'transcription' property with a TextObject as value. As in the previous example, segments of these transcription values are annotated further, this time with a part-of-speech.
Gestures are TimeSeriesObjects with additional properties "description", "hand shape" and "orientation". This illustrates how a project specific *template* for gesture annotation can be defined: as a subclass of AMM TimeSeriesObject with a number of AMM sub properties of amm:feature.

4. Implementation and infrastructure

The AMM model, being designed as a simple OWL model, allows us to share software (tools and tool components), infrastructure (web services, web repositories) and annotation data between different projects, starting with the different CATCH projects. Moreover, since the model supports annotation of (parts of) annotations it allows CATCH projects to cooperate in a special way: one project can take the annotations of another project, treat it as their primary data and add new layers of annotation. For example, transcription text of a project working on automatic speech recognition of radio archives can automatically be semantically annotated with the help of an annotation web service developed by another project.

The AMM OWL definition, project specific annotation models and annotation instances are currently all stored in one central store. This store is implemented using a web based Sesame RDF repository. To prevent that users of the repository have to formulate complex RDF queries we developed a web service encapsulating this repository. The interface of this web service contains operations to retrieve annotations and annotation information, and to manage (create, delete, modify) annotations and annotation models. A Java software library is available to use the web service transparently (without having to make explicit web service calls) from Java programs or Java based web applications.

Programmers in CATCH currently collaborate on a web based annotation environment that supports manual and semi-automatic annotation of different types of digital media from heterogeneous cultural heritage collections.

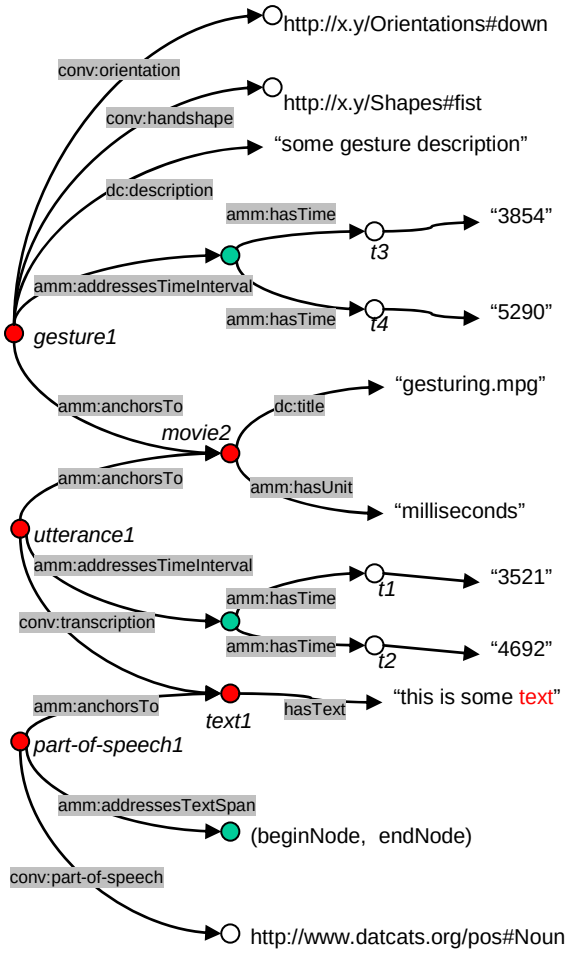

Figure 3: graph representation of overlapping gesture and speech utterance annotations for a video recording

This environment explicitly supports creating, searching, browsing and visualizing of layered annotations. Also, the CATCH project two of the authors are participating in, CHOICE[8], implemented an annotation web service that takes a textual resource as input and returns a list of automatically generated semantic annotations anchored to that resource (Gazendam, 2006). All projects can store their annotations in one central AMM repository using the Java web service library mentioned before. In this way, a semantically rich resource is created, that enables access to (fragments of) digital objects across the boundaries of cultural heritage institutions, collections and media types.

5. Conclusions and future work

Our practical experiences with testing and applying the Annotation Meta Model described in this paper so far indicate that it seems to meet all of our initial requirements. All special annotation cases that we applied it to fitted in without having to leave out information or come up with forced reinterpretations of existing annotations. The cases that we tested are quite diverse and extreme. Therefore, we are confident that the AMM can be applied very well to combine annotation data of quite different nature. Also, several cases of adding new layers of annotations to previously created annotations were handled successfully.

Accessing annotations that were stored in a web based Sesame repository using our AMM API works efficiently. We tested with approx 50.000 heterogeneous AnnotatableObjects in the store. Nevertheless, this is not a large number, and we should therefore carefully test how well it works with substantially larger numbers of annotations.

API and repository are successfully applied in the context of building a *Documentalist Support System* for the Netherlands Institute for Sound and Vision that is based on automatic annotation a text documents. In this DSS AMM is used to represent all annotations involved: newly created catalog metadata for radio/TV programs, automatic semantic annotations of fragments of texts describing these radio/TV programs and the associations of the programs with the descriptive texts.

In the near future we first of all hope to improve the AMM itself, for example by testing it on even more extreme use cases (e.g. interlinear text, a complexly structured annotation practice used field linguistics). It may also be of interest to extend the model to cover relations between anchors, like segment overlap, nesting, sequence or distance relations, in both generic and media specific ways.

We will go on defining more project or media specific sub models for AMM, and we might even consider building an interactive tool to help users do this themselves without having specific competences in RDF/OWL.

With respect to annotation infrastructure, we will convert and import existing annotation data, and/or include existing annotation repositories by wrapping them with a web service that implements our AMM API's interface. We will probably have to face issues of scalability and system's distribution.

Finally, it may be interesting to investigate how AMM is perceived by potential users. Is it as easy to understand, adopt and implement as we expect?

6. Acknowledgements

We thank the Netherlands Institute for Sound and Vision for hosting our projects and providing valuable input and feedback. We would like to thank NWO for making this research possible.

[8] http://www.nwo.nl/catch/choice

7. References

Bird, S., Liberman, S. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33(1,2):23-60.

Brugman, H. (2003). Annotated Recordings and Texts in the DoBeS project. In *proceedings of EMELD 200*3 http://emeld.net/workshop/2003/proceeding03.html

Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M., (editors) (2007), Definition of the CIDOC Conceptual Reference Model**,** August 2007.

Gazendam, L., Malaisé, V., Schreiber, G., Brugman, H. (2006). Deriving Semantic Annotations of an audiovisual program from contextual texts. In *Proceedings of First International workshop on Semantic Web Annotations for Multimedia (SWAMM 2006).* 23 May 2006, Edinburgh, Scotland.

Geurts, J., van Ossenbruggen, J., Hardman, L. (2005). Requirements for practical multimedia annotation. In: *Workshop on Multimedia and the Semantic Web.* http://www.acemedia.org/ESWC2005_MSW/ (pages 4-11), May 2005, Heraklion, Crete.

Hunter, J. (2001). Adding Multimedia to the Semantic Web – Building an MPEG-7 Ontology, in *Proceedings of the 1st International Semantic Web Working Symposium (SWWS 2001)*, 2001.

ISO/IEC (2002). Overview of the MPEG-7 Standard (version 8). ISO/IEC JTC1/SC29/WG11/N4980, Klagenfurt, July 2002.

Romary, L., Ide, N. (2007). International Standard for a Linguistic Annotation Framework. *Natural Language Engineering* 10, 3-4 (09/2004) 211-225.

Schreiber, A., Dubbeldam, B., Wielemaker, J., Wielinga, B. (2001). Ontology-based Photo Annotation. *IEEE Intelligent Systems*, 16(3):66-74, May-June 2001.

Troncy, R., van Ossenbruggen, J., Pan, J., Stamou, G. (2007). Image Annotation on the Semantic Web. W3C Incubator Group Report 14 August 2007.