Multimedia Markup Tools for OpenKnowledge

David Dupplaw, Madalina Croitoru, Antonis Loizou, Srinandan Dasmahapatra,
Paul H. Lewis, Mischa M. Tuffield, Liang Xiao

School of Electronics and Computer Science, University of Southampton,
Southampton, UK, SO17 1BJ

Abstract. OpenKnowledge is a peer-to-peer system for sharing knowledge and is driven by interaction models that give the necessary context for mapping of ontological knowledge fragments necessary for the interaction to take place. The OpenKnowledge system is agnostic to any specific data formats that are used in the interactions, relying on ontology mapping techniques for shimming the messages. The potentially large search space for matching ontologies is reduced by the shared context of the interaction. In this paper we investigate what this means for multimedia data on the OpenKnowledge network by discussing how an existing application that provides multimedia annotation (the Semantic Logger) can be migrated into the OpenKnowledge domain.

1 Introduction

In this paper we discuss what multimedia metadata issues occur in open systems, by investigating how multimedia data is being used in the OpenKnowledge system. OpenKnowledge provides an interaction-based open network that provides ontology mapping and sharing. It utilises the shared context of the specific interaction to reduce the search space of the mapping calculation thereby making the mapping tractable. We show how this functionality impacts on multimedia ontologies by introducing an existing multimedia annotation tool onto the network.

Our purpose is to be able to exchange multimedia data in a flexible and reusable manner aimed at addressing the "semantic gap" - the interpretation differential between the data that can be extracted automatically from a media item and the meaning that humans might attribute to it. We show that the flexible approach provided by OpenKnowledge will have clear benefits for building multimedia applications by easing multimedia data exchange. We show a simple application built using OpenKnowledge that provides, in a semi-automatic way, low to high level mappings for multimedia data. These mappings can be obtained using interaction models that could then be reused for different sets of mappings. We thus show, on one hand the OpenKnowledge project's ability to handle multimedia data and, on the other, specific advantages that arise from addressing multimedia in an OpenKnowledge scenario.

2 Motivation

An important challenge arises when the interoperation between multimedia systems requires the understanding of data that is not easily annotated. Text-based semantic extraction methods are well-used and, on the whole, accurate. However, their multimedia counterparts, used to extract semantics from image, audio or video data, are not as mature and less accurate because the underlying representation of objects in multimedia is further from semantic representations than text. Automatic annotation attempts to bridge this gap between low level descriptors and the symbolic labels by learning mappings between combinations of low-level media descriptors representing data objects and textual labels indicating the identity of those objects.

One of the early attempts at automatic image annotation was the work of Mori et al [10] which attempted to apply a co-occurrence model to keywords and low-level features of rectangular image regions. Current techniques for auto-annotation generally fall into two categories; those that first segment images into regions, or 'blobs' (e.g. [3, 9]) and those that take a more scene-orientated approach using global information (e.g. [6, 7]).

Creating semantic-web metadata representations for multimedia is a vibrant research area, and many of the projects base their work on MPEG-7[1], an XML standard developed by the Moving Pictures Expert Group (MPEG) for metadata representation. Like the major metadata protagonists, MPEG-7 is independent of the multimedia itself. However, it has flaws in its representational schema that mean, for example, there are many ways to represent the same data. One of the latest schemas for describing metadata is COMM (Common Ontology for Multimedia)[1] that builds upon MPEG-7 in a way that is familiar yet overcomes many of its shortcomings. COMM is "semantic-web compatible" by being delivered in OWL. Other multimedia description schemes are in development and the W3C Multimedia Semantics Incubator[2] have put together a list of multimedia vocabularies that could be used to annotate data. Most of these have some form of relationship between them such that some form of mapping could be afforded.

The aspiration of OpenKnowledge is to allow knowledge to be shared freely and reliably, regardless of the source or the consumer. Reliability here is interpreted as a semantic issue, as the Internet is in the fortunate position that transport-level reliability has already been established. Reliably sharing semantic data requires consistent interpretation by either a shared conceptualisation where there is consensus, or the mapping of semantic terms between locally defined models. However, the supplier and consumer of such knowledge need a context in which mapping can be performed, and models of interaction can provide this context. It is for this reason that OpenKnowledge has its core mechanism in the sharing of *interaction models*, and it is expected that communities can form in which the sharing of knowledge mappings can be exploited.

[1] MPEG-7 *http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm*

[2] MMSEM *http://www.w3c.org/2005/Incubator/mmsem*

OpenKnowledge uses the interaction modelling language LCC [11] (Lightweight Coordination Calculus) that provides hooks for forcing user interaction or visualisation of data. Placeholders in the interaction definition represent data items that are assumed to be annotated by the components in the network that produced the data. In most interactions this can be handled by ontology mapping based on textual representations of concepts. Providing similar functionality to multimedia data items poses extra challenges but the benefits of an "open" approach to multimedia data manipulation is clear.

The OpenKnowledge system is based upon a subscription paradigm, where peers may join interactions by subscribing to roles in those interactions. Interactions can be sought by searching the peer network by keywords. Only when all necessary roles in a specific interaction model have subscriptions will the interaction bootstrapping begin. This involves negotiating with each peer to decide on the optimal set of peers to play in the interaction. Once all peers have agreed to play the interaction the interaction will start. A coordinator for the interaction is selected from the peer network who will control the interaction's execution state. Peers playing roles will be called to execute specific functions that their role requires; these are encoded as constraints on message operations in the model definitions.

In the following section we will describe an existing exemplar multimedia annotation application, before describing how OpenKnowledge can allow annotations to be shared.

3 Approach

3.1 Generating Contextual Annotations

The Semantic Logger[3] (SL) is the outcome of previous work looking into automatic, auto-biographical metadata acquisition, published in [13]. The semantic logger builds on the ideas brought forward in the original *Scientific American* Semantic Web article[2], with a particular focus on the notion of assembling, and integrating web accessible resources. It has been presented as a means to populate the Semantic Web with personal metadata and has now been ported to the OpenKnowledge system. The integration of the two systems is seen as illustrative of how legacy systems can be utilised from within the OpenKnowledge framework.

The intuition behind the SL system is that, since analysing multimedia objects to obtain high-level descriptors is a hard task, it may be easier to first associate a media object with more readily extracted semantics about the same abstract event and then use the assembled *context* to produce annotations.

The Semantic Squirrel Special Interest Group (SSSIG)[4] is a group of researchers who aim to automate the process of logging any available raw data (or *'nuts'*) that describe aspects of one's personal experience. This raw data is

[3] Semantic Logger, http://akt.ecs.soton.ac.uk:8080/

[4] http://www.semantic-squirrel.org/

captured or summarised into RDF representations. The intent is to utilise these RDF fragments to construct the context of a particular event, at a particular time. By virtue of the fact that each event logged by the system is time-stamped and related to the user's FOAF[5] URI, we are able to choose variable levels of granularity to describe any given context. The squirrels are mediated at the heart of the system by the AKT Project's[6] SPARQL-compliant RDF triplestore 3store[8]. Figure 1 shows some possible raw data sources, although the system is not limited to any specific set.

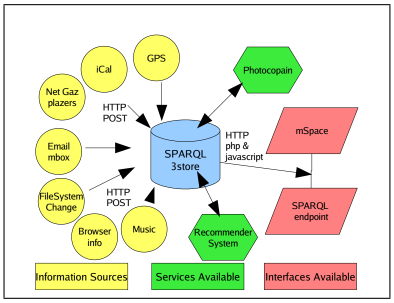

Fig. 1. Overview of the Semantic Logger architecture

The following squirrels have been implemented that propogate their information to the SL in RDF: calendar entries, geographical data, email, weather information as well as other information related to a user's context including music listening trends, web-browsing trends, file system changes and personal news feeds.

3.2 Generating Annotation for Images

The PhotoCopain project[12] is an exemplar of how the contextual information stored in the Semantic Logger can be used to annotate images. Once an image is uploaded to the system, the knowledge base is queried for objects created at the same place and time as the image. Calendar entries that span the time the picture was taken are processed to indicate the location and subject of the image. If GPS entries or network addresses are available for the same time, these are resolved against gazetteers to provide geographical annotations.

Part of the work undertaken in the PhotoCopain project was to use content-based techniques to supplement the context-based information captured by the Semantic Logger in order to annotate a user's photos. A set of low-level and mid-level feature extraction modules were built to extract the feature vectors that were used in reasoning tools and classifiers to provide high-level semantic concepts for annotating the images in RDF.

[5] Friend of a Friend (FOAF) *http://www.foaf-project.org*

[6] Advanced Knowledge Technologies, http://www.aktors.org/

The available image-analysis modules that the semantic logger has available are:

- **Scene Type Detector**: Detects whether the scene contains natural features or man-made features by using an edge-direction coherence vector. This assumes that man-made objects generally have strong straight-lines in them.
- **Face Detector**: Detects faces in the image using colour coherence providing their location, size, and count.
- **EXIF Extractor**: Extracts the EXIF data from an image providing values for reasoning, such as the time and date of capture, details about the camera and capture conditions such as whether the flash went off and the type of lens the photo was taken with.
- **Focus Detector**: Extracts focus information from the image for backing up the hyperfocal distance calculations from the EXIF data to help infer a particular scene type.

The initial evaluation has focused on the performance of individual feature classifiers, based on KNN classifier scheme, with $K = 3$. The classifiers have been evaluated for indoor/outdoor images and for artificial/natural environments. In both cases, a training set of 150 images, with 75 instances representing each classification, and a test set of 30 images were used. For the classification of images as indoor or outdoor we found that a combination of the EXIF data and the CIELab colour map data without any dimension reduction performed the best, yielding only 4 errors in the 30 tests. Our classification of natural and artificial environments yielded its best results with a combination of the edge direction coherence vector, the CIELab colour map and the nearest neighbour clustering algorithm, giving 3 errors in the 30 test cases. Further reasoning takes place over the results from these modules to decide if the photo is a portrait of a person (by aggregation of the camera details, focus detector and the face detector results).

A difficulty arises in aggregating the context if modules produce annotations that have not been derived from an ontology for which there is a global consensus. In an open system this is not practical. In the following section we introduce how the OpenKnowledge deals with the problem of ontology sharing.

3.3 Sharing Annotations

The OpenKnowledge system defines two levels of ontological mapping. As the OpenKnowledge system uses a subscription paradigm, the first mapping can occur during subscription to a role in an interaction model. The ontology matcher (currently based on SMatch[5]) checks how the peer's functionality can be mapped to the model's role. This mapping is based on the static information present in the interaction model's definition and so can be done offline. The 'Global Good Enough Answer' provides information about how well the peer is skilled to proceed in a specific interaction model, and the 'Local Good Enough Answer' provides information on how skilled a peer is for playing a specific role in an interaction.

$$\begin{array}{l}\mathbf{a}(\text{faceDetector}, FD) :: \\ \quad msg(I) \ \Leftarrow \ \mathbf{a}(\text{faceProvider}, FP) \leftarrow \mathbf{isImage}(I) \ then \\ \quad count(N) \ \Rightarrow \ \mathbf{a}(\text{faceProvider}, FP) \leftarrow \mathbf{detect}(I, N) \\ \quad \ or \ noFaces() \ \Rightarrow \ \mathbf{a}(\text{faceProvider}, FP)\end{array} \tag{1}$$

Model 1 shows the role in the model for the face detection squirrel (for a full definition of LCC syntax and semantics see [11]). The role expects a message containing an image from a peer in the $faceProvider$ role and will return the number of faces in the message $count(N)$ or it will return the message $noFaces()$ if the $\mathbf{detect}(I, N)$ constraint fails.

During subscription the mapping will map the $\mathbf{isImage}(I)$ and the $\mathbf{detect}(I, N)$ constraints with the functions that are available on the peer subscribing to the role. This includes the re-ordering of the arguments and, if necessary, mapping of the arguments types from the type defined in the interaction model (if one is defined) to a type the peer will understand. These mappings are created semi-automatically.

The second level of mapping is still an ongoing research topic in the OpenKnowledge project and occurs during run-time when instances of the variables I and N are instantiated in the model. If no default type overlay is provided by the interaction model author, the mapping algorithm attempts to map the relevant parts of the peers conceptual formalisations to each other. As N represents the name of a person, mapping the semantics of N to the semantics understood by the peer processing the constraint may be a passive step requiring only concept mapping (e.g. $PersonName$ to $FullName$). However, as I represents an image, information about that image must be made available to the peer in the correct format such that the peer can process the image. This involves a two level mapping: the passive step as above, plus, if required, an active mapping step where the image is downloaded and converted into another format. This requires that the content of I is in some ontological format that can be mapped using the concept mapping techniques.

As a simple example of this, we might take a peer that is gathering a user's email and populating a semantic logger with this data. The peer creates RDF representations for the email based on iCal RDF[7]. When the user adds a photo to his peer an interaction model (Model 2) is played out that calls PhotoCopain to detect faces, and extract EXIF.

[7] iCal RDF: *http://www.w3.org/2002/12/cal/ical*

$$
\begin{array}{l}
\mathbf{a}(\text{sl\_photo\_provider}, ID1) :: \\
\quad \mathit{image}(I) \ \Rightarrow \ \mathbf{a}(\text{photocopain\_module}, PCM) \leftarrow \mathbf{getImage}(I) \ \mathit{then} \\
\quad \mathit{imageMetadata}(M) \ \Leftarrow \ \mathbf{a}(\text{photocopain\_module}, PCM) \ \mathit{then} \\
\quad \mathit{storeMetadata}(M) \ \Rightarrow \ \mathbf{a}(\text{semantic\_logger}, SL) \\
\\
\mathbf{a}(\text{photocopain\_module}, PCM) :: \\
\quad \mathit{image}(I) \ \Leftarrow \ \mathbf{a}(\text{photo\_provider}, PP) \ \mathit{then} \\
\quad \mathit{imageMetadata}(M) \ \Rightarrow \ \mathbf{a}(\text{photo\_provider}, PP) \leftarrow \mathbf{getMetadata}(I, M)
\end{array}
\tag{2}
$$

The data returned to the peer in M, on the second line of the model, will be represented using different ontologies to the peer's and so disparate fragments would be created in the logger. This means no aggregative context could be formed for the user's data. It is at the reception of the metadata that the mapping will take place to transform M into a format understood by the peer. Once this mapping has been achieved (possibly manually by the user), the mapping can be stored and shared over the P2P network, such that the mapping algorithms of other peers may take advantage of the successful mapping.

The variable I in Model 2 is used to transfer metadata about the image, because sending the actual raw data may be costly (the image may be very large) and possibly unnecessary (if only the width and height of the image is required, for example). OpenKnowledge does not enforce any specific format on this metadata and any appropriate RDF representation could be used; for example, the RDF representation of MPEG-7 or the COMM[1]. One advantage to using multimedia-specific ontologies is that they allow content-based extraction results to be transferred as metadata, which might allow further processing without costly network transfer of data. As this is ongoing work, we hope to evaluate how well the mappings provide interoperability in the near future.

4 Discussion and Future Work

The aspiration of OpenKnowledge is to allow knowledge to be shared freely and reliably, regardless of the source or the consumer.

We have presented a multimedia annotation demonstrator for gathering of semantic data from a user's computer. The demonstrator shows how contextual information can be used to support content-based annotation of images.

We discussed how OpenKnowledge can be used as a base on which image analysis modules can build and share annotations. OpenKnowledge has the ability to provide ontological mapping between two previously unseen conceptualisations, allowing producers of metadata to annotate their objects in the way which best suits them. The mapping then allows the re-user of the metadata to understand the data and, by sharing the resulting mapping, allows other re-users to benefit from the community's knowledge of ontological mappings. We have discussed an extension to this ontological mapping that we hope to integrate into OpenKnowledge to allow the download and conversion of large media items to be deferred until the raw data is required.

We have shown that OpenKnowledge provides a good basis on which to build open, scalable networks of services that include multimedia annotation services.

References

1. Richard Arndt, Raphal Troncy, Steffen Staab, Lynda Hardman, and Miroslav Vacura, *COMM: Designing a well-founded multimedia ontology for the web*, 6th International Semantic Web Conference (ISWC'2007), November 2007.
2. T. Berners-Lee, J. Hendler, and O. Lassila, *The Semantic Web*, Scientific American **284** (2001), no. 5.
3. P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth, *Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary*, ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV (London, UK), Springer-Verlag, 2002, pp. 97–112.
4. Peter G. B. Enser, Yiannis Kompatsiaris, Noel E. O'Connor, Alan F. Smeaton, and Arnold W. M. Smeulders (eds.), *Image and video retrieval: Third international conference, CIVR 2004, Dublin, Ireland, July 21-23, 2004. proceedings*, Lecture Notes in Computer Science, vol. 3115, Springer, 2004.
5. F. Giunchiglia, M. Yatskevich, and P. Shvaiko, *Semantic matching: Algorithms and implementation*, Journal of Data Semantic, vol. VIII, Springer, 2006.
6. Jonathon S. Hare and Paul H. Lewis, *Salient regions for query by image content.*, in Enser et al. [4], pp. 317–325.
7. Jonathon S. Hare, Paul H. Lewis, Peter G. B. Enser, and Christine J. Sandom, *Mind the gap*, Multimedia Content Analysis, Management, and Retrieval 2006 (San Jose, California, USA) (Edward Y. Chang, Alan Hanjalic, and Nicu Sebe, eds.), vol. 6073, SPIE, January 2006, pp. 607309–1–607309–12.
8. Stephen Harris, *SPARQL query processing with conventional relational database systems.*, WISE Workshops, 2005, pp. 235–244.
9. Jiwoon Jeon and R. Manmatha, *Using maximum entropy for automatic image annotation.*, in Enser et al. [4], pp. 24–32.
10. Y. Mori, H. Takahashi, and R. Oka, *Image-to-word transformation based on dividing and vector quantizing images with words*, Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99), 1999.
11. David Robertson, *Multi-agent coordination as distributed logic programming.*, ICLP, 2004, pp. 416–430.
12. Mischa M. Tuffield, Stephen Harris, David P. Dupplaw, Ajay Chakravarthy, Christopher Brewster, Nicholas Gibbins, Kieron O'Hara, Fabio Ciravegna, Derek Sleeman, Nigel R. Shadbolt, , and Yorick Wilks, *Image annotation with PhotoCopain*, Proceedings of the First International Workshop on Semantic Web Annotations for Multimedia (SWAMM) (Edinburgh), May 2006, held as part of 15th World Wide Web Conference (22-26 May, 2006).
13. Mischa M Tuffield, Antonis Loizou, David Dupplaw, Sri Dasmahapatra, Paul H Lewis, David E Millard, and Nigel R Shadbolt, *Semantic logger: Supporting service building from personal context*, Proceedings of Capture, Archival and Retrieval of Personal Experiences (CARPE) Workshop at ACM MM, ACM MultiMedia, October 2006.