Metadata-based access to cultural heritage collections: the RHCe use case

Metadata-based Access to Cultural Heritage Collections: the RHCe Use Case

Kees van der Sluijs[1] and Geert-Jan Houben[1,2]

[1] Technische Universiteit Eindhoven, Computer Science, PO Box 513, 5600 MB Eindhoven, the Netherlands
{k.a.m.sluijs, g.j.houben}@tue.nl
[2] Vrije Universiteit Brussel, Computer Science, Pleinlaan 2, 1050 Brussels, Belgium

Abstract. More and more cultural heritage organizations see a great opportunity by opening up their collections via the Web to expand their user-base. In this paper we look at our current work in a specific use case, a cultural heritage organization called RHCe that wanted to open up its photo and video archives to the public. We demonstrate in this paper how we can utilize metadata to offer a homogeneous multi-faceted view over their heterogeneous archives. We also discuss what to do if metadata is not available for resources and how we can use a simple mechanism like tagging to still get high quality annotations. We do this by relating the user tags to concepts in an ontology and we discuss some mechanism to do this (semi-) automatically. We also show how these techniques can be used to build a user model and how we can identify the most probable annotations that can be used by domain experts to improve their annotation-time efficiency.

Keywords: cultural heritage, data access, personalization, metadata, tags, ontologies, semantics.

1 Introduction

Collections of cultural heritage content have long been accessible only from within the institutions hosting the collections. With the emergence of the Web, many of these institutions have started attempts to make the content available from the World Wide Web, and experimenting with this new role of the content and their own new role in offering access to this content.. Some examples of such projects are FinnONTO[1] (and FinnONTO 2.0[2]), CHIP[3], CATCH[4], SmartMuseum[5]. All these efforts share the desire to open up the collections of cultural heritage content to the wider public and they investigate how to do that such that the individual users can get effective access.

[1] http://www.seco.tkk.fi/projects/finnonto/
[2] http://www.seco.tkk.fi/projects/sw20/
[3] http://www.chip-project.org/
[4] http://www.nwo.nl/nwohome.nsf/pages/NWOP_66EUM7_Eng
[5] http://smartmuseum.eu/

A key element in this endeavour of opening up the cultural heritage collections is the availability of metadata. The metadata describes the content and allows the tools for data access to know which content is there and can be supplied as part of an answer to a user's request for information. Both in searching the content as in browsing the content, metadata describing the content is a necessity. Typical for the scenarios in the institutions and for the early experiments is that not a lot of high quality metadata is available and first has to be created: often, by hand by the professionals from the institutions or by automatically extracting it from the content. This lack of metadata is a problem that has triggered several attempted solutions.

Often the metadata is organized with the aid of concept structures or ontologies that structure the metadata, for example with classes and relationships. In many domains, consolidated concept structures or ontologies have been obtained and can then be used for the organization of the access to the data based on the metadata. However, these structures have often been obtained through a consolidation process involving professionals in the domain, and that makes them not always directly suitable for average end-users as road map for their access to the content: the concepts sometimes do not come with an intuitive meaning, nor do the relationships between concepts.

Both for the purpose of easy understanding of the structure and for creation of metadata by end-users, user-annotation or tagging in Web 2.0-speak has therefore come into the picture. Whereas tagging in the sense of associating free keywords to the content is easy to do, systems that have chosen for good reasons to be based on more carefully crafted concept structures or ontologies cannot use those tags without difficulty. That is why in such cases it is interesting to see how the end-user tags can be related to the concept structures.

This had led to two interesting questions. First, it is relevant to investigate how metadata can be exploited for browsing and searching. Second, it is relevant to see how good metadata can be obtained, which includes the question how tags can be related to concepts. In one of our use cases, in RHCe, we exactly had these questions. In this workshop paper we report on the current standings of this research and lay out the path to the future.

In section 2 we introduce RHCe and explain the goal it had and the associated problems it was facing. In section 3 we show how a navigation and search structure was made over RHCe's heterogeneous datasets using metadata. Then in section 4, we discuss which collaborative-based approaches were used to offer similar navigation structures over the datasets for which no metadata or full-text is available. We then end the paper with some observations on our system and a discussion of the currently planned work.

2 RHCe

The Regional Historic Centre Eindhoven (RHCe) governs all historical information related to the cities in the region around Eindhoven in the Netherlands. The information is gathered from local government agencies or private persons and groups. This includes not only enormous collections of birth, marriage and death

certificates, but also posters, drawings, pictures, videos and city council minutes. Most of the fragile material is stored in vaults and is thus physically inaccessible to the public. A first step in opening up the collections has been the digitization of many of the collections, and as in many similar cases this enormous effort has been done in a more or less literal transformation of the physical structures into their digital multimedia representations.

One of the main goals of our collaboration with RHCe was to experiment with technology that could help to further expose these collections to the general public. However, especially for the videos and pictures very little metadata is available which makes indexing this data for navigation or searching very hard. The original metadata was mainly targeted at the professionals working at the offices of the centre, and was therefore not suitable for the larger public. A more specific goal of RHCe is therefore to have high-quality metadata of all their collections for easy retrieval (both online and offline, and both for the general public and for the officials of the local government).

RHCe employs a number of domain experts (cultural heritage experts) whose full-time job is to provide high quality metadata over multimedia documents based on a carefully constructed topic ontology by RHCe's domain metadata specialists. However, in spite of all their efforts by far most of their collections have no metadata at all. Worse yet, new material arrives more frequently and in larger quantities than the domain experts can hope to annotate in any near future: it is easy to see that their capacity will not be sufficient to supply all the desired metadata.

We therefore designed a prototype application called CHI. The goal of this prototype was twofold. First, it has to disclose the data to the end-users for browsing and searching. Second, it has to support the users in collaboratively providing the metadata and then to support the application and consolidation of that metadata. These aspects will be explained in the following sections of the paper.

3 Browsing and searching the RHCe collection

As a first step to open up the digital collections to the larger public, the challenge was addressed to demonstrate how metadata could be exploited in browsing and searching. To this end CHI was built, a prototype Web application framework with the purpose to offer the digital multimedia versions of the collection content to the public in a meaningful way. For this first version, the photo and video collection was considered since it was judged that this would offer the best short-term gain in terms of public access and interest and in terms of insights in the role of metadata in the process.

The framework offers a faceted browser view over the data (inspired by work like [1], [2], [3]). This means that the data can be browsed or searched via a number of different dimensions. The photo and video collections that were considered first in these experiments carry three such dimensions: time, location, and keywords. All three of those dimensions are used to describe the subject of the photos and video scenes. In CHI these dimensions are described by detailed domain-specific ontologies. These ontologies are under the control of RHCe's professionals and they

also ensure that the metadata of the content aligns with these ontologies. Due to this alignment between metadata and ontologies, CHI can offer the end-users the navigation along the collections in a homogeneous way.

For every dimension CHI has a separate visualization in the user interface. For time we use the Simile Timeline[6], for location we use Google Maps[7] and for the keywords we built a graph representation which represents the relatedness of terms. In this way, we can cluster data elements that share a characteristic in one dimension in these interfaces. The user can either navigate directly through the datasets via one of the views, or can use a search interface to search and present the clustered search results in one of these interfaces (besides the regular search result list). This representation is created in such a way that for a given picture or video, other pictures and videos that share a characteristic can be found.

Figure 1 for instance is a screenshot of the Google Maps visualization of our application. In the screenshot the search results for photos and videos related to the Eindhoven city-centre are represented. Clicking on one of the locations one can see all clustered elements that belong to the specific location; in case of the screenshot all pictures related to the Eindhoven Market can be seen in the popup window.

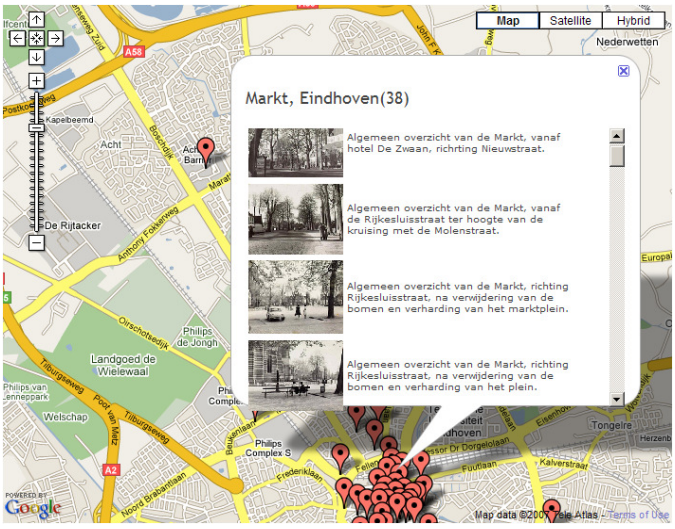

Figure 1: Screen shot of the Google Maps visualization with clustered results

In order connect the objects in the RHCe dataset to Google maps we had to align the RHCe location ontology to location information that could be used by Google Maps.

The location of an object in the RHCe metadata ontology consists of a location name. This name is not always available for an object. This has several reasons, e.g. it

[6] http://simile.mit.edu/timeline/
[7] http://maps.google.com/

might not be known at which specific location a photo is taken and it also might be impossible to find out, and sometimes there are several conflicting assertions over where a picture is taken. The locations also differ in granularity. Photos of specific buildings can be quite specifically pinpointed to a location, while the location of an aerial photo of the entire city is much broader. Another challenge for the alignment is that location names change over time, e.g. during redivision or complete renovation of districts.

These problems are overcome by maintaining a location ontology. Location metadata refers to concepts in this ontology. Every concept has a label and if a location name changes the new name can be added as a label to the concept as the new name (including a time indication to indicate when the name has changed). The location hierarchy also includes building names, to simplify the task of the annotators. In this way the domain experts can for instance annotate a photo with "city hall" instead of having to look up the exact address. They do have to use context of the hierarchy to indicate exactly which city hall they refer to (e.g. the one in Eindhoven or one in the neighboring municipalities).

As the domain experts prefer to annotate objects with conceptual names like "city hall" instead to be bothered with coordinates during the annotation process we add these coordinates in the location ontology (instead of the annotation). We only have to associate the locations in that ontology to the coordinates once. Explicitly providing coordinates for all locations in the Eindhoven area is still ongoing work, but with our approach with every new coordinate-pair added to a location concept in the ontology we have effectively obtained the exact location information of a large set of objects. The advantage of having coordinates is that they can be directly translated in Google Maps locations. In the cases where we don't have coordinates we try to use the location names, but this sometimes leads to faulty locations on the map.

The hierarchical structure of the location also allows us to differ in the granularity of our clustering. We can for instance cluster on the most detailed level (e.g. only cluster all photo's of the city hall), but also on street level, district level or even city level.

With the time dimension (see Figure 2 for a screenshot) we do something similar as with locations. RHCe has a custom time description in their metadata ontology. For reasoning purposes we aligned that with the OWL time ontology[8]. Via the time ontology we are able to populate the XML input for the timeline. The ontology can for instance be used for querying, (e.g. for queries with restrictions like "before" or for the use of intervals). It can also, like the Google Maps, be used to differ in granularity (e.g. day-based, year-based or era-based levels). When clustering the results in the timeline we can also use the keyword ontology, so for instance for the period of the second world war (for which a lot of material exists and is annotated in the collection), we are able to not only cluster al results from that era together, but add further clustering via the keywords (e.g. clustering all objects that depict bunkers from that time).

[8] http://www.w3.org/TR/owl-time/

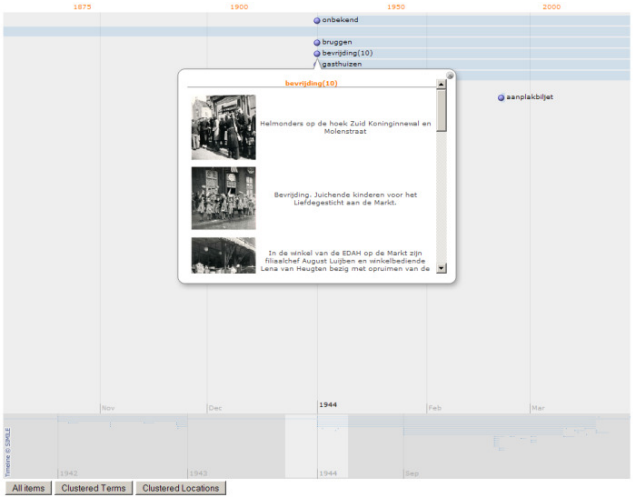

Figure 2: Screenshot of the Simile Timeline visualization

4 Creating metadata through tagging in the RHCe collection

Ontology- or concept-based approaches ask for quality metadata. Besides opening up their datasets to the public, one of RHCe's goals is actually to obtain high-quality metadata of all of its archived data not only to improve on searching and browsing the datasets but also to adhere to quality standards as specified by the (local and national) government. However, RHCe's few domain specialists have only limited time and the collections are huge. For them, the biggest benefit from CHI is to exploit the access by the users for getting metadata from them. However, for many obvious reasons which we will not specify here in detail, users do not want to fill in large forms to provide well-structured data about the photos and videos for example: many will find this too time-consuming or too complicated (e.g. a typical part of RHCe's user group consists of elderly people with little computer (and typing) experience, but with great knowledge and interest in the domain). Therefore, simplicity is a key feature for CHI and we use several simple mechanisms to keep the system as easy accessible as possible while still obtaining this information (and the construction of the prototypes is actually part of the effort to experiment with this demand).

At the core of this approach is a tagging mechanism ([4],[5],[6],[7]) by which users can enter keywords or small sentence fragments, called tags, to describe a scene on a photo or video. An inherent property of tagging is that it is schema-less. This means that the user does not need any prior knowledge of some domain for annotating resources. This is what makes tagging inherently simple, and what convinces RHCe that this will be an effective tool in the circumstances they are in with their photo collection.

Using a tagging mechanism introduces also some problems, however, and some that we also experience here. One problem is that the semantics of tags are not always clear, i.e. what is precisely the meaning and intention of a tag? There are several causes for doubts, for instance spelling mistakes, disambiguation concerns (e.g. the Dutch word "bank" can mean "bench" or a "financial institute"), words that have more then one common spellings (e.g. "chic" versus "sjiek", both meaning "classy" in Dutch) or morphology.

Another problem is that tags are often not very well structured. It is not clear which tags are related to which other tags, or what property of a resource is actually described. For example, the tags differ in how specific they are. A picture that depicts the building called "Catharina church" could for instance be tagged with "Building", "Church" or "Catharina church". However, if you would know that "Catharina church" is a type of "Church" and "Church" is a kind of "Building", this information could be used during a search for buildings and then you could also find resources only labeled with "Catharina church".

Of course a tag-only-based approach could be used, meaning that we build an ontological structure based on co-occurrence relationships between tags which is used in various approaches (e.g. consider [8],[9]). However, this approach has some disadvantages. One problem is that relationships between tags that co-occur are not clear. If two terms often co-occur, does that mean that they are synonyms, that one is more specific than the other, or is there some other relationship? Another problem, as explained in [4], is that the groups of terms that all people agree on are usually very general and will lead to a shallow ontology, except for some specific 'hot' topics that tend to be over-specific. The largest problem is however the lack of quality control. As explained, RHCe tries to adhere to metadata quality standards and they have their own carefully constructed ontology. The ontology is extendable if necessary, but this should be under total control of RHCe. If the users experience the freedom of providing tags but in fact (perhaps with a little bit of help) annotate the photos or video scenes with concepts from the ontology, then this would significantly increase the quality and effect of the metadata. Therefore, the aim of the support is that the resources should be somehow annotated with concepts from their ontology. And if we have this annotation, then this should lead to a high-quality well-balanced browsing experience. Therefore, we chose to look at relating tags to ontological concepts in the controlled ontology.

In order to relate tags to ontological concepts we use several techniques that correspond to techniques in ontology building (e.g. [9]) and ontology matching (e.g. [10],[11]). The situation at hand differs in solely relating tags to concepts which gives some specific problems, and we also make use of the fact that we have a relatively controllable (in size) user group for personalization. In the next sections we briefly discuss the techniques we use in this specific setting.

4.1 Lexical matching

Our first step to relate tags to concepts from the ontologies that we use is based on lexical matching, i.e. we are going to compare tags and concepts on the basis of the lexical representations associated with them.

Tags basically are strings, so that does not pose a problem, however with ontologies things are different. We use RDF and OWL to express the ontologies, which is a natural choice as these language are especially designed to express ontologies and because of their widespread use many additional sources are available (both in terms of tools and "helper" ontologies and tools). In those languages a concept is denoted by a URI. The textual representation of a concept can be modeled in different ways. A common way to represent textual representations of concepts is by using the rdfs:label. However, many other candidate properties exist, like the skos:preflabel and skos:altlabel to discern between preferred and alternative labels, but also custom properties are used. Sometimes the label schema has a more complex structure, like the example in Figure 3. On the other hand, many ontologies actually do not use labels at all, but use the fragment identifier of an URI as the only labeling.

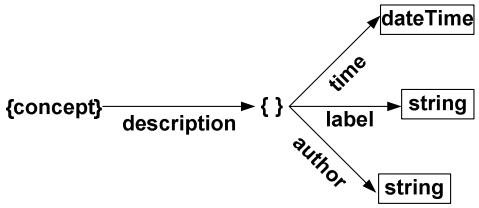

Figure 3: Complex label structure

In CHI, to identify the labels of concepts in ontologies we use a configuration where the label of a concept can be specified in four ways. Default configuration behavior is to look for all the well-known label properties, like rdfs:label and the skos:preflabel and skos:altlabel properties. For complex structures the configuration can be specified using a SPARQL query. For URI decoding the configuration has to specify which delimiter schema to use (e.g. mixed case nouns, underscores). The RHCe metadata ontology uses complex labeling structures.

After identifying the labels for the concepts in an ontology, we calculate the string similarity between the tags and the string representations of the concepts. We do this to accommodate for small spelling variations, morphology, etc. Many methods and libraries to calculate this similarity exist and we chose to use the simmetrics library[9] for this. After calculating the similarity values for tags and concept labels we select those above a configurable threshold. The result of this process is a set of concepts and certainties (similarity value) for every tag.

4.2 Exploiting the Ontological Structure

When we have a relationship between tags and concepts in the ontology we can also exploit the structure of the ontology. We can exploit this structure at several points in the application. For instance at query time, e.g. with a search on the concept "Church" we can traverse the skos:broader relation and find the *narrower* terms, e.g. "Catharina Church", and then also all images and video that have a tag that matches (one of the) label(s) of the concept for "Catharina Church". In CHI we however also

[9] http://sourceforge.net/projects/simmetrics/

exploit the ontology relationships at an earlier stage, as we show the user suggestions for a newly inputted tag consisting of concept labels from the ontology. In this way we let the user verify the matches and select the most appropriate suggestion. By not only showing syntactic matches but also semantically related matches we give the user a richer choice of labels and thus we get more precise feedback which concept the user actually meant. For this we configure CHI to know which properties to traverse (e.g. the skos:broader property). If the user selects a label we store this action as a relation between the original user tag and the concept that the label belongs to. To give an idea of the quality of the suggestions consider the suggestions for the input tag "bevrijd" (liberated) in Table 1.

Suggestion	Certainty	Suggestion	Certainty
Bevrijding (liberation)	0.96	vrijheid (freedom)	0.90
intocht (parade)	0.94	dekollonisatie (decollonization)	0.88
vrijlatingen (setting free)	0.94	onderdrukking (suppression)	0.85
emancipatie (emancipation)	0.90	oorlogen (war)	0.85
Onafhankelijkheid (independence)	0.90	verkieziningen (elections)	0.76

Table 1: Suggestions for the input tag "bevrijd" (liberated)

4.3 Collaborative filtering and Personalization

The techniques we discussed up until now mainly utilize semantics of concepts. Another promising route is the use of the contribution of the users. We exploit this to improve the matching process, but also for user management and personalization.

First we use the user verification as a feedback mechanism. As the user gets concept suggestions we record their choice. They can not only indicate if they think a certain suggestion is good, but can also give negative feedback for bad suggestions. By accumulating this data we adjust the certainties of user suggestions. Suggestions for tags that many users agree on are considered to be better matches than suggestions that are often disapproved.

Next, we can use the user feedback to build up user models that can be used for personalization and verification. By noting which resources are tagged and used by a user and which terms the user uses in his tags we can say something about the user interest.

The most important for RHCe however is quality control using collaborative techniques. By a small addition in the user interface users have the possibility to rate current tags (and concepts) for a given resource. By measuring which users usually agree with the RHCe domain experts we can calculate which users might be the most valuable for RHCe and might give the opinion of these people more weight (i.e. make them more important in the system). This can be applied recursively by looking at users that have a high degree of agreement with the important users in the system, which might increase their importance as well. We are currently investigating several ways to further exploit the information we have.

5 Conclusion

Most important for us is to continue and finalize the implementation of the different techniques that we described in this paper so that we obtain a rich toolset to time-efficiently assist a domain expert to provide high quality metadata for large uncharted datasets. One problem we are for instance also trying to tackle is how to add totally uncharted resources in the dataset. What we do not want is to consolidate a set of objects that every user has seen and tagged, while a large set of data is never seen. The challenge is find a way to integrate objects without metadata into user query results without giving the user the feeling that he gets wrong results.

We are also collaborating with a selection of users to evaluate our techniques and features. In this way we work toward evaluation of our system by a representative set of users.

6 References

1. Mäkelä, E., Hyvönen, E., Saarela, S.: Ontogator - A Semantic View-Based Search Engine Service for Web Applications. In: Proceedings of the International Semantic Web Conference, pp. 847--860, Springer, Heidelberg (2006)
2. Yee, K.P., Swearingen, K., Li, K., Hearst, M.: Faceted metadata for image search and browsing, In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 401--408, ACM, New York (2003)
3. Michal Tvarožek, Mária Bieliková, Personalized Faceted Navigation for Multimedia Collections, In Proceedings of the Second International Workshop on Semantic Media Adaptation and Personalization, pp. 104--109, IEEE Press, New York (2007)
4. Golder, S., Huberman, B.A.: Usage Patterns of Collaborative Tagging Systems. In: Journal of Information Science, vol. 32, no. 2, pp. 198--208, Sage Publications, Inc. Thousand Oaks, CA (2006)
5. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging, In: Proceedings of the 16th international conference on World Wide Web, pp. 211--220, ACM, New York (2007)
6. Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics, In: Proceedings of the International Semantic Web Conference, pp. 522--536, Springer, Heidelberg (2005)
7. Marlov, C., Naaman, M., Boyd, D., Davis, M.: HT06, tagging aper, taxonomy, Flickr, academic article, to read. In: Proceedings of the seventeenth conference on Hypertext and hypermedia, pp. 31--40, ACM, New York (2006)
8. Choi, S.O., Lui, A.K.: Web Information Retrieval in Collaborative Tagging Systems, In: Proceedings of the International Conference on Web Intelligence. pp. 352--355, IEEE Press, New York (2006)
9. Specia, L., Motta, E.: Integrating Folksomonies with the Semantic Web. In: The Semantic Web: Research and Applications, pp. 624--639, Springer, Heidelberg (2007)
10. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. In: The VLDB Journal, vol. 10, no. 4, pp. 334--350, Springer-Verlag, New York (2001)
11. Aleksovski, Z., ten Kate, W., van Harmelen, F.: Ontology matching using comprehensive ontology as background knowledge, In: Proceedings of the International Workshop on Ontology Matching at ISWC 2006, pp. 13--24, CEUR (2006)