INVENiT: Exploring cultural heritage collections while adding annotations

Chris Dijkshoorn[1], Jacco van Ossenbruggen[2], Lora Aroyo[1], Guus Schreiber[1]

[1] Computer Science, The Network Institute,VU University Amsterdam
{c.r.dijkshoorn, lora.aroyo, guus.schreiber}@vu.nl
[2] Centrum Wiskunde en Informatica, Amsterdam, the Netherlands
jacco.van.ossenbruggen@cwi.nl

Abstract. The growing number of cultural heritage collections published as Linked Data has given rise to a vast source of collection objects to explore. To provide an experience which goes beyond traditional search, the links from objects to terms from structured vocabularies can be used to create new paths to explore. We present INVENiT, a semantic search system which leverages these paths for result diversification and clustering. Users can freely explore the collection, but are also able to contribute their knowledge by annotating collection objects. The added information is directly incorporated in the search results. The demo can be found at `http://sealinc.ops.few.vu.nl/invenit/`.

1 Introduction

The increasing number of cultural heritage collections published as Linked Data promises to be an incredible source of rich content for end users to explore [3,2]. Explorability of the collections heavily depends on the quality of the metadata describing the objects [3]. The ability to explore collections is increased when a dense network of links between objects is created. These relations can be realised by linking objects to other collection objects and entities from structured vocabularies.

For example, the Rijksmuseum Amsterdam publishes its collection online and for this purposes employs catalogers to register, annotate and digitise collection objects. They use a limited set of structured vocabularies to annotate the subject matter, the material, techniques and artists. However, there is a multitude of LOD vocabularies that can be used in addition to support the desired exploration of the collection.

Many catalogers have a background in art-history allowing them to only provide basic information about different subject matter domains. To fill the missing domain expertise, and provide annotations in all the domains represented in the Rijksmuseum collection, we involve in the curation process people from outside the museum that have expert knowledge in each of those domains. In this paper we discuss a use case demonstrator, which allows external experts to annotate parts of images with terms from structured vocabularies. The contribution of this work is three-fold. First, we align the new vocabularies to the existing annotation vocabulary structure of the Rijksmuseum, following standardised data models, e.g. Europeana Data Model. Second, we explore linked data patterns to optimise the use of these aligned vocabularies in the presentation and exploration of search results. Finally, we integrate the annotation results of the external annotators in a common semantic search system `http://sealinc.ops.few.vu.nl/invenit/`.

2 Data

The Rijksmuseum collection comprises around 1,000,000 artworks, of which 159,661 have a digital representation. The RDF data is modelled according to the Europeana Data Model [2]. Objects are linked to multiple vocabularies: the Iconclass vocabulary[3] for describing subject matter, the Art and Architecture Thesaurus (AAT) for materials and techniques and the Union List of Artist Names[4] (ULAN) for artists.

The Rijksmuseum collection contains links to 11,945 of the 39,578 concepts in Iconclass. These concepts are hierarchically structured, with more specific resources further down the hierarchy. While there are many links from collection object to AAT these concern a limited number of materials and techniques. In contrast many distinct links are made to ULAN, the Rijksmuseum has a diverse collection with works made by many different artists. In addition ULAN defines interesting relations between the concepts, for example *teacher_of* and *uncle_of*.

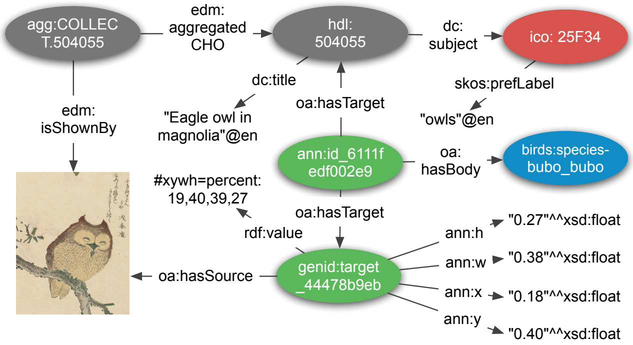


Fig. 1: The print “Eagle owl in magnolia” modelled according to the Europeana Data Model, with an annotation added specifying the species of depicted bird. The annotation resource has two targets, the cultural heritage object and a generated target resource, linking the digital representation with the area specified by the annotator.

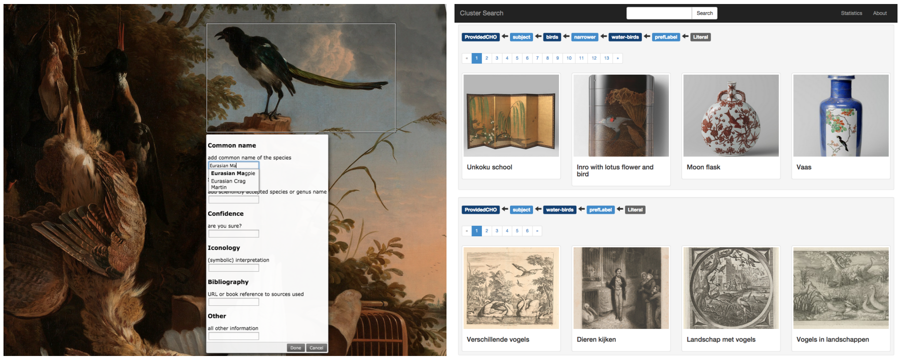
For the current demo we take a subset of 1,598 object from the Rijksmuseum collection: artworks with depictions of birds. The catalogers might not

[3] `http://www.iconclass.org/`
[4] `http://www.getty.edu/research/tools/vocabularies/`

have enough knowledge to classify which species of bird is depicted while there are many bird enthusiasts who do. To test the use of additional structured vocabularies we made a conversion of the IOC world birdlist[5], including 31,644 species and sub species. Figure 1 shows an example of an artwork with an added annotation. The INVENiT demonstrator has been also instantiated with other Rijksmuseum sub-collections, e.g. prints related to biblical topics and books `http://invenit.wmprojects.nl/`.

3 System

INVENiT is based on the Cliopatria semantic web server [4], extended with an annotation module and a cluster search module, of which the corresponding interfaces are depicted in Figure 2. The **annotation module** provides functionality to add annotations to images. The annotation fields can be tailored to the use case and autocompletion is based on a specified vocabulary. Relevant objects in the image can be identified by drawing bounding boxes and all of the provided information is stored in a triple store.


Common name
add common name of the species
Eurasian Ma
Eurasian Magpie
Eurasian Crag
Martin
add scientificly accepted species or genus name

Confidence
are you sure?

Iconology
(symbolic) interpretation

Bibliography
URL or book reference to sources used

Other
all other information

(a) Annotation interface showing a bounding box and autocompletion.

(b) Search interface showing clustered search results.

Fig. 2: Annotation and search interface of the INVENiT demo system.

The **cluster search module** utilises a graph search algorithm, which matches keyword queries with literals, uses the graph structure to find connected artworks and clusters similar objects together [4]. Literals in the database are assigned a matching score according to a ranking function. Literals with a score above a specified threshold are used as starting point for backward graph traversal. The graph is traversed in a backward fashion until a specified class of resource is reached, in this case *edm:ProvidedCho*. Objects with similar paths are clustered together.

[5] `http://github.com/rasvaan/naturalis`

Users can use this search functionality to explore the collection and find artworks to annotate. We adapted the algorithm to interpret the added annotation as subject matter metadata, which allows the user to directly inspect the result of their efforts in the search results. The demo in its initial state (without user contributed content), supports exploration based on the metadata provided by the Rijksmuseum Amsterdam.

The presented clusters are generated based on paths in the graph. These paths can be based on a direct link between a literal and artworks, but also longer paths are used. When possible properties are abstracted to their (SKOS) root properties. The recourses used in paths are abstracted to their class. Below three examples of paths can be found:

1) Literal → *title* → *Artworks*
2) Literal → *subject* → *Owls* → *broader* → *Birds* → *subject* → *Artworks*
3) Literal → *prefLabel* → *Artist* → *teacherOf* → *Artist* → *creator* → *Artworks*

These examples illustrate the characteristics of the dataset and vocabularies. The first example includes results based on metadata in the collection. The second example generalises the results based on links in Iconclass. The third example uses links within the ULAN vocabulary to cluster results.

4 Discussion and Future Work

The INVENiT demo uses the semantics in structured vocabularies to diversify and cluster results. Users can make new connections by annotating collection objects with terms from structured vocabularies. We believe that providing users with the possibility to find the objects they like to annotate and directly inspect the results of their efforts will have a positive effect on their motivation.

Currently all annotations are accepted and incorporated in the system. This is not something a museum would allow, since unknowledgeable or malicious users might add incorrect information. We therefore plan to incorporate trust assessment in current work, providing an indication whether an annotation is trustworthy or not, for example based on the assessed expertise of an annotator [1].

There are still open issues to address regarding the clustering of search results based on paths in the graph. The relevance of the results in a cluster are influenced by the path used to retrieve them. At this moment the clusters are ranked according to the number of results within the cluster. We plan on improving this, since the meaningfulness of a path depends on the perception of the user.

The clusters of results are named by the paths used to create them. These paths are hard to interpret for a user. Take for example *Literal* → *title* → *Artworks*. This could be translated into the name "works titled". Especially longer paths are difficult to concisely describe. Automatically generating more user-friendly names is a problem we want to address in future work.

Acknowledgements. This publication is supported by the Dutch national program COMMIT/. We like to thank the members of the SEALINCMedia worktable and in particular the Rijksmuseum for their support.

References

1. Ceolin, D., Nottamkandath, A., Fokkink, W.: Efficient semi-automated assessment of annotations trustworthiness. Journal of Trust Management 1(1), 3 (2014)
2. Isaac, A., Haslhofer, B.: Europeana linked open data – data.europeana.eu. Semantic Web Journal (2013)
3. Szekely, P., Knoblock, C., Yang, F., Zhu, X., Fink, E., Allen, R., Goodlander, G.: Connecting the smithsonian american art museum to the linked data cloud. In: The Semantic Web: Semantics and Big Data (2013)
4. Wielemaker, J., Hildebrand, M., van Ossenbruggen, J., Schreiber, G.: Thesaurus-Based Search in Large Heterogeneous Collections. In: ISWC2008 (2008)