Using Linked Data to Diversify Search Results a Case Study in Cultural Heritage

Chris Dijkshoorn[1], Lora Aroyo[1], Guus Schreiber[1],
Jan Wielemaker[1], and Lizzy Jongma[2]

[1] Computer Science, The Network Institute, VU University Amsterdam
{c.r.dijkshoorn,lora.aroyo,guus.schreiber,jan.wielemaker}@vu.nl
[2] Rijksmuseum Amsterdam, The Netherlands
l.jongma@rijksmuseum.nl

Abstract. In this study we consider wether, and to what extent, additional semantics in the form of Linked Data can help diversifying search results. We undertake this study in the domain of cultural heritage. The data consists of collection data of the Rijksmuseum Amsterdam together with a number of relevant external vocabularies, which are all published as Linked Data. We apply an existing graph search algorithm to this data, using entries from the museum query log as test set. The results show that in this domain an increase in diversity can be achieved through adding external vocabularies. We also analyse why some vocabularies have a significant effect, while others influence the results only marginally.

Keywords: Linked Data, Diversity, Semantic Search, Cultural Heritage.

1 Introduction

One of the promises of Linked Data is that it can be used to achieve richer, more diverse search results. This paper reports on a case study on diversifying search results that we performed in the domain of cultural heritage. Cultural heritage is a knowledge-rich domain, in which many external vocabularies are available, often as Linked Data. Also, we expect that, given the nature of the domain, a significant group of users would be interested in retrieving a more diverse palette of search results than the standard ones. For example, searching for "Rembrandt" should provide you with much more insights and related artworks than just his master pieces.

As data for this study we used the collection data of the Rijksmuseum Amsterdam. We enriched the collection metadata with a number of external vocabularies that have been published as Linked Data, such as the Art & Architecture Thesaurus, WordNet and Iconclass. We employ an existing graph search algorithm to find search results [13]. This algorithm finds paths in the graph from the search term to target objects, in this case artworks. The algorithm also clusters the results by classifying paths and grouping the results with similar path classes. Two example clusters for the search term "rembrandt" would be works with as creator "rembrandt' and works in a location labelled with "rembrandt"

K. Janowicz et al. (Eds.): EKAW 2014, LNAI 8876, pp. 109–120, 2014.

(e.g., the Rembrandt House). In this study we use the number of resulting clusters and the path length as indicators of diversity. As sample queries we collected the terms in the museum's query log for the duration of one month. We see this study as a step towards showing how Linked Data could be used to create a richer search experience. Showing this might help to make institutions aware of the potential added value of investing in Linked Data.

This paper is structured as follows. In the next section we discuss related work. Section 3 describes the collection data and Linked Data used in the study. In Section 4 we discuss the experimental setup, including the test set and the graph search algorithm. Results are discussed in Section 5. In Section 6 we reflect on the results and consider future directions.

2 Related Work

A lot of work has been done on integrating cultural heritage collections and linking them to external sources. Hyvonen et al. created a portal to integrated collections of Finnish museums, using semantic web techniques [5]. Europeana is an initiative which supports the integration of European cultural heritage collections [6]. It comprises over 26 million metadata records, originating from more than 2,000 cultural heritage institutions. The aggregation comes at a cost, the collections have to adhere to the Europeana Data Model, which can result in a loss of semantic relations.

De Boer et. al. describe a methodology of publishing collections as Linked Data while preserving the rich semantics [2]. Similar methodology is applied by Szekely et al. in [9]. They stress the importance of high data quality for museums. By integrating collections and linking them to external vocabularies the amount of available data increases, giving rise to the need for structured means to access the information [3].

Researchers at Europeana clustered artworks at different granularities to create an overall picture and provide users with related objects [11]. The clustering approach is useful for identifying duplicate records, although at lower granularities users had difficulties explaining why artworks were clustered together. Regularities in the Linked Data cloud can be used to cluster, while still being able to explain how the objects are related. Hollink et al. use predefined patterns to improve image search [4] and similar paths are successfully used in a content based-recommender system [12].

There is a growing interest in the diversification of search results in the field of information science. Introducing more diverse results is used to address the ambiguity of keyword queries. Agrawal et al. assign topics to the user intent and documents and optimise the chance that the user is satisfied by the results [1].

An increasing number of systems use Linked Data to support users. Mismuseos[1] lets you search in integrated Spanish art collections and refine results using filters and facets. Constitute includes RDF representations of over 700

[1] `http://www.mismuseos.net/`

constitutions and lets users search and compare them[2]. Seevl uses semantic web techniques to provide search and discovery services over musical entities [7]. The BBC is developing a system to open up their radio archives, automatically annotating audio fragments and using crowdsourcing mechanisms to refine the data [8].

3 Data

In this section we describe the characteristics of the collection data of the Rijksmuseum and its links to external vocabularies. An overview of the datasets and how these are connected is given in Figure 1.

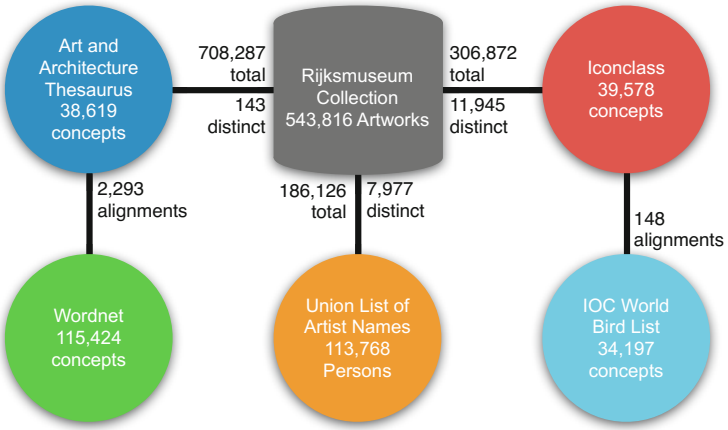

Fig. 1. Overview of Rijksmuseum collection data and links to external vocabularies

The **Rijksmuseum collection** consists of around 1,000,000 artworks. Around 25% of the collection is available in a digital form: some 180,000 prints and 70,000 other works have been digitised. Catalogers have added annotations to the records, when possible originating from structured vocabularies.

The Rijksmuseum provides through a non-public API access to 550,000 collection objects. This server provides data in the format of the RDF-based Europeana Data Model (EDM) [6]. Figure 2 shows an example of an object represented in EDM. The format makes a distinction between the unique "work" and (possibly multiple) digital representations and metadata descriptions of the work. EDM thus caters for alternative representations of the same object in different (sub)collections.

The **Iconclass vocabulary**[3] is designed to be used to annotate subjects, themes and motifs in Western art. Iconclass is available as linked data since

[2] `https://www.constituteproject.org/`
[3] `http://www.iconclass.org/help/lod`

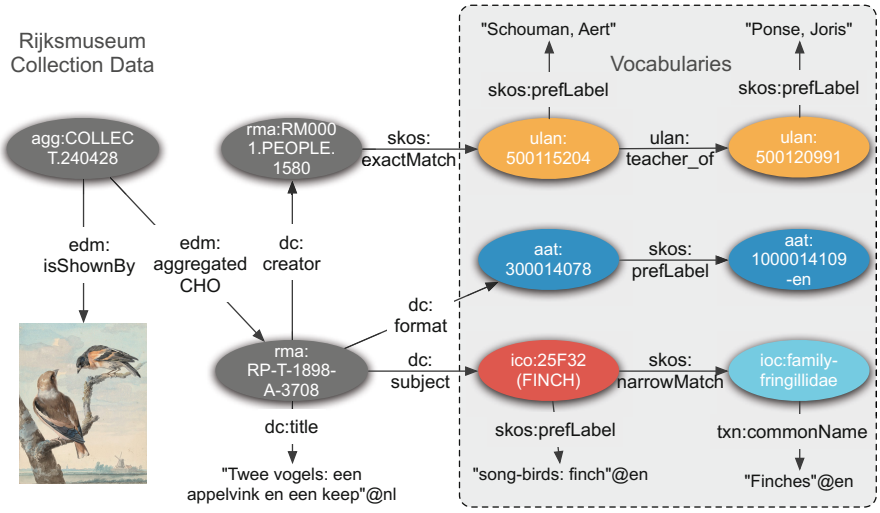

Fig. 2. Rijksmuseum artwork "Two birds". The top-left resource (`agg:COLLECT.240428`) represents the actual "work". The work has a digital representation (bottom-left) and a metadata description (`rma:RP-T-1898-A-3708`)[5]. Four pieces of metadata are shown: the title is represented as a literal, the subject is an Iconclass category, the format points to an AAT concept, and the creator of the work is represented with a resource from an in-house vocabulary of artists. The Iconclass subject category has an alignment with a bird class in the IOC vocabulary. The creator resource is aligned with a corresponding resource in the ULAN vocabulary. The link between the two ULAN concepts is one example of the type of extra information accessible through this alignment.

2012, containing almost 40,000 resources. Iconclass categories are defined using a code grammar. For example, the top-level category *Nature* has code `2`; the category *song birds* has code `25F32`. The category hierarchy is modelled using `skos:broader` and `skos:narrower` predicates. In this study we use some 300,000 links from collection objects to Iconclass categories.

The Getty research institute compiles, maintains and distributes vocabularies that focus on visual arts and architecture, in particular (i) the **Union List of Artist Names** (ULAN), (ii) the **Art & Architecture Thesaurus** (AAT), and (iii) the **Thesaurus of Geographic Names** (TGN). For this study we chose to link the Rijksmuseum collection to AAT and ULAN. AAT has 77,470 resources describing techniques, materials and styles which artworks can have in common. In this experiment we use the recently released Linked Data version of AAT[4]. ULAN includes biographic information about 113,768 artists and is in addition a valuable source of relations between persons, such as "collaborated with" and "teacher of". We use the XML version of ULAN converted to RDF and create links with the collection based on string matching.

[4] `http://www.getty.edu/research/tools/vocabularies/lod/`

WordNet is a source of lexical information about the English language. It provides short descriptions of words, groups words with the same meaning into synsets and defines the semantic relations between those sets. We use the WordNet 2.0 version published by W3C [6], comprising over 79,000 nouns, 13,000 verbs and 3,000 adverbs. We reuse the 2,293 alignments made between AAT and WordNet by Tordai et al. [10].

The International Ornithologists Union maintains a comprehensive list of bird names. We convert the XML version of this **IOC World Bird**[7] to RDF, adding labels from the multilingual version. This results in a taxonomy of 34,197 concepts describing the orders, families, genera, species and subspecies of birds and the corresponding structure. We manually align the bird concepts of Iconclass to matching concepts in the IOC vocabulary.

4 Methods

4.1 Experimental Setup

Firstly, we investigated how many query terms have matches in the dataset. For this purpose we collected query terms on the Rijksmuseum website during one month (see Section 4.2 below). The terms are then matched with the literal index of the triple store containing the collection data and the five external vocabularies. As the frequency of the query term might be a factor influencing the number of matches, we split the list of query terms into three sublists, containing respectively the high, medium and low-frequency query terms. The query terms are split in such a way that the three sums of the number of times, that the queries in a sublist are used, are equal for each split.

Secondly, we explore to what extent the external semantics improve semantic search results. To this end we use an existing semantic search algorithm (see Section 4.3 below for details) to perform a search on all query terms. We do this five times, each time with a different dataset configuration: (i) only collection data, (ii) AAT and WodNet added, (iii) Iconclass and IOC added, (iv) ULAN added, and (v) all vocabularies added. The reason for combining AAT with WordNet and Iconclass with IOC stems from the dependencies between these vocabularies, as shown in Figure 1.

The graph search algorithm delivers the results in clusters of semantically similar results. Per obtained cluster we analyse the path length in the graph as well as the number of clusters. This gives us per query information about the average path length, average number of clusters and average number of results. The results of this analysis are again split in three parts according to the query frequencies (high, medium, low).

The code developed for these experiments as well as the resulting data are available online[8].

[6] `http://www.w3.org/TR/wordnet-rdf/`
[7] `http://www.worldbirdnames.org/ioc-lists/`
[8] `https://github.com/rasvaan/cluster_search_experimental_data`

4.2 Query Logs

We used the query logs of January 2014 of the Rijksmuseum. From these logs we extracted all distinct query terms used plus their frequency. This provided us with 48,733 unique query terms. We filtered out 4,074 terms because they were either object IDs[9] or were in some other way erroneous. The resulting set of 44,659 query terms was used in the experiments. The split into frequency groups of query terms resulted in 2,393 terms in the high split (high frequency), 16,963 query terms in the medium split (medium frequency), and 25,303 terms in the low split (low frequency).

It should be noted that these queries were made against the collection data without the external vocabularies. This causes a bias because the collection data contain mainly Dutch terms and therefore users who have used the search interface before are likely to refrain from using English search terms, knowing that these are of limited value.

4.3 Graph Search

We developed a semantic search system that enables the exploration of collections by using a keyword query to produce clusters of semantically related objects[10]. We use the graph search algorithm as described in [13]. The algorithm matches the query term with literals in the triplestore, using stemming. When the match exceeds a given threshold it is added to a list. The literals in this list are used as a starting point to traverse the graph structured data. This traversal continues registering the times a specified target class is found, all the while recoding the steps it makes. The starting literal and successive properties and resources used as steps, form the path in the graph which serves as the basis for clustering. For clustering the properties in the path are abstracted to their root properties when possible. In addition resources are abstracted to their class, unless they are a concept. This allows to merge clusters based on similar semantics.

5 Results

We have collected data of four types:

- The number of query terms in the test set that have, in principle, matches in the dataset.
- The number of search results for each of the query terms in the dataset with and without the external Linked Data.
- The number of clusters of search results for each of the query terms in the dataset with and without the external Linked Data.
- The distribution of path lengths of search results for each of the query terms in the dataset with and without subsets of the external Linked Data.

[9] Some problems with the existing query interface can be circumvented by entering directly an object ID of an artwork, e.g. `SK-A-4979`. For the purposes of this study we left out the query terms resulting from this practice.

[10] `http://sealinc.ops.few.vu.nl/clustersearch/`

Matches between Query Terms and Dataset. In Figure 3 can be seen that 94% of the query terms in the high frequency split match literals in the collection data. ULAN and AAT match over 50% and continue to match many query terms in the lower splits. WordNet, Iconclass and IOC have less matches, with the IOC percentage on all splits below 12%. There is a decrease between the query frequency splits, were in the low split all the matches in external vocabularies are less than 23%.

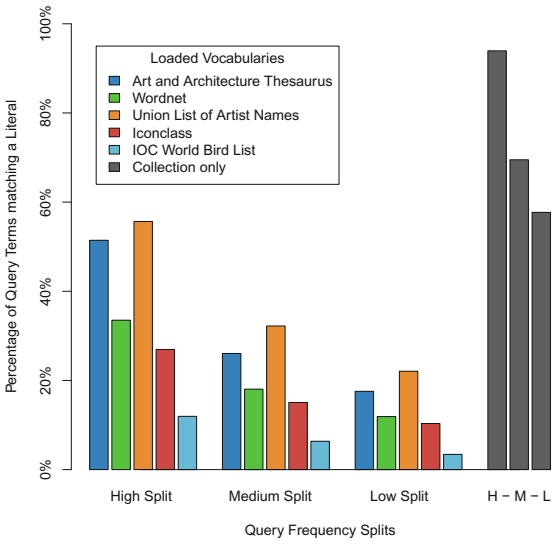

Fig. 3. Bar chart showing the percentage of query terms that match with a label in the vocabularies

To illustrate, the query term "rembrandt" matches in AAT, ULAN, and the collection data. Where in the collection and ULAN labels of "Rembrandt van Rijn" are matched, AAT matches "Rembrandt frames". The query term "watercolor" has no match in the collection data, but does match in AAT, ULAN, and WordNet. In AAT it matches materials and a technique, in ULAN descriptions of painters and in WordNet the type of paint in addition to the watercolour painting as an object.

The numbers above give an indication of the potential in the data to be used for search. It depends on actual links between resources in the dataset on whether these can actually be used during search.

Search Results Per Query. Figure 4 shows the overall increase of search results when the external vocabularies are loaded. The increase is marked but moderate. The increase is highest in the third quartile of the high split; the quartile raises

from 214.0 to 268.8. The mean increases from 81.5 to 104.5 search results. To give an example, when the external vocabularies are loaded the query term "rembrandt" has 674 instead of 636 results. Instead of no results, "watercolor" increases to 9 results. It should be pointed out that the number of clusters (not shown here, see below) influences the maximum number of results, as the algorithm imposes a maximum of 100 search results per cluster.

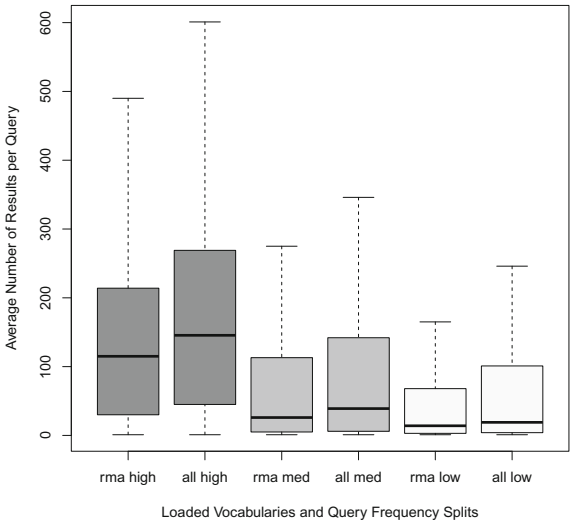

Fig. 4. Overall number of search of results per query term. The boxes marked as "rma" represent the baseline (collection data only); the boxes marked with "all" represent the search results with all vocabularies loaded.

Clusters Per Query. Figure 5 shows how the number of clusters of search results increases when the external vocabularies are loaded. The median increases with one for the medium and high splits. There is also a marked increase in the range: some queries apparently lead to a large number of clusters. Thus, the external vocabularies not only lead to more results, but also more diversified results.

The number of clusters for the query term "rembrandt" increases from 12 to 15, adding for example a cluster of paintings of "Pieter Lastman" who was a teacher of Rembrandt and paintings of "Salomon Koninck" who was, according to ULAN, an ardent follower of Rembrandt. One cluster is found for "watercolor", containing water colour paintings by "Pieter Withoos", based on the descriptive note "He specialized in watercolors of insects and flowers.". An example of a query term leading to a large number of clusters is "rubens": 8 clusters are created

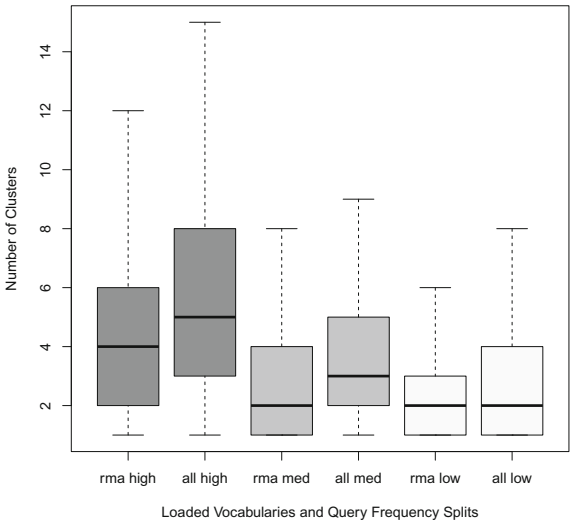

Fig. 5. Box plot of the number of clusters per query. The boxes marked as "rma" represent the baseline (collection data only); the boxes marked with "all" represent the search results with all vocabularies loaded.

with the collection data loaded, 15 with the external vocabularies loaded, adding among others clusters about students and assistants of Peter Paul Rubens.

Path Length Per Query. Finally, we look at the path length of search results. A longer path length suggests a diversification of results. For path length we have looked at the contribution that the different vocabularies give to the path length. This can provide us an indication which vocabularies are most useful.

Figure 6 shows how the path length of the search results changes when particular vocabularies are loaded. The first column shows the baseline, where the path length is either 1 or 2. We see that adding AAT plus WordNet or Iconclass plus IOC has hardly any effect on the path length. The Union List of Artist Names (ULAN) has a significant effect on the path length. ULAN leads to 22% of the paths being longer than 2, up to paths of length 15.

ULAN is actually responsible for almost the complete path length diversity (see last column). We can see an example of this phenomenon when we look again at the keyword query "rubens". The following path generated a cluster with artworks of a student of Peter Paul Rubens:

Rubens, Peter Paul → *teacher of* → *Dyck, Anthony van*
Dyck, Anthony van → *creator* → <*several artworks*>

Why does only ULAN contribute significantly to the diversity of path lengths? If we look at Figure 1 we see that ULAN has the highest number of links from the

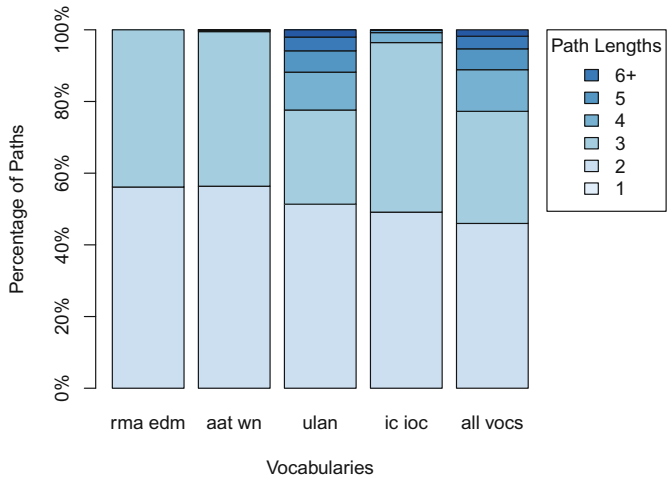

Fig. 6. Path lengths of search results, shown as percentage of all search results. The first column shows the baseline (only collection data); columns 2-4 depicts the three (groups of) vocabularies separately; column 5 shows the situation when all external vocabularies are loaded.

collection data to distinct resources in ULAN. Also, the structure of ULAN (with many crosslinks such as "teacher" and "collaborator") makes it suitable for search diversification. In contrast, the links to AAT involve only a limited number of AAT resources, and are possibly not of much interest to users (typically things like "canvas", "oil paint" and "print").

Iconclass does have, like ULAN, many links from the collection data to distinct Iconclass resources and also interesting internal links. The likely reason why this does not lead to more search results is probably that Iconclass does not have Dutch labels. The test set of query terms came from the current search facility which works only with Dutch-language metadata. Assuming this has led to a limited usage of other languages for search, Iconclass resources were of little use for this test set. So, this part of the results is likely to be biased by the test set.

6 Discussion

This case study suggests that Linked Data in the form of vocabularies can indeed be used to diversify search results to meet the needs of a significant group of users, e.g. general audience and humanities scholars, that would be interested in retrieving a more diverse palette of search results beyond just the standard and popular ones. Moreover, diversifying search results helps also the collection owners in promoting specific parts of their collection.

The results show that for this application domain we can achieve (i) an increase in number of results, and (ii) indirectly through the number of clusters and the path length, an increase in the semantic variety of search results. However, not all vocabularies appear to be equally useful. Thus, it is important for collection owners to know the influence of each vocabulary on the accessibility of their collections, and further integrate this in the strategies for collection annotation. Based on this study we hypothesise that the usefulness of vocabularies for diversification of search results depends on the following two factors:

1. the number of links between distinct vocabulary resources and the metadata of target search objects;
2. the richness of the internal links between vocabulary objects;

Results clearly show that vocabularies, such as ULAN and Iconclass, which provide rich semantics for additional context (e.g. relation between people and their roles) have significant influence on the diversity of the results. In previous studies [12] and related work [11], [4] we also show that these vocabularies are a valuable source of context and relevance for end users.

This study has a number of limitations. Firstly, our test set is a set of query terms (44.5K) that came from logs of the existing search log of the institution involved. People, who use a search interface multiple times, are likely to limit their search to terms that work well with this interface. Therefore, the fact that the Iconclass vocabulary did not contribute a lot to diversity, may be a result of this bias. Secondly, there is not yet a set of standard semantic-search algorithms. It could well be the case that other algorithms lead to different results with the same data and test set. Also, clusters and path length are indirect indicators of diversity. More studies are needed to show how valuable these indicators are and how they compare to diversity measures as introduced in [1]. Thirdly, the data set we used is limited in nature. It would be good to perform studies like these also in large, more heterogeneous datasets.

With this study we show the promise and added-value of Linked Data for diversifying search results. We plan on extending this work by adding additional external vocabularies and investigating ways of increasing the density of the links between the collection and vocabularies.

Acknowledgements. This publication was supported by the Dutch national program COMMIT/. We are grateful to all our project colleagues for the discussions on this subject.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM 2009, pp. 5–14. ACM, New York (2009)

2. de Boer, V., Wielemaker, J., van Gent, J., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 733–747. Springer, Heidelberg (2012)
3. Grimnes, G.A., Edwards, P., Preece, A.D.: Instance based clustering of semantic web resources. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 303–317. Springer, Heidelberg (2008)
4. Hollink, L., Schreiber, G., Wielinga, B.: Patterns of semantic relations to improve image content search. Web Semantics Science Services and Agents on the World Wide Web 5(3), 195–203 (2007)
5. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: Finnish museums on the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web 3(2-3), 224–241 (2005)
6. Isaac, A., Haslhofer, B.: Europeana linked open data – data.europeana.eu. Semantic Web Journal 4(3), 291–297 (2013)
7. Passant, A.: seevl: mining music connections to bring context, search and discovery to the music you like. Semantic Web Challenge 2011 (2011)
8. Raimond, Y., Ferne, T.: The bbc world service archive prototype. Semantic Web Challenge 2013 (2013)
9. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E.E., Allen, R., Goodlander, G.: Connecting the smithsonian american art museum to the linked data cloud. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 593–607. Springer, Heidelberg (2013)
10. Tordai, A., van Ossenbruggen, J., Schreiber, G., Wielinga, B.: Aligning large SKOS-like vocabularies: Two case studies. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part I. LNCS, vol. 6088, pp. 198–212. Springer, Heidelberg (2010)
11. Wang, S., Isaac, A., Charles, V., Koopman, R., Agoropoulou, A., van der Werf, T.: Hierarchical structuring of cultural heritage objects within large aggregations. CoRR (2013)
12. Wang, Y., Stash, N., Aroyo, L., Gorgels, P., Rutledge, L., Schreiber, G.: Recommendations based on semantically enriched museum collections. Web Semantics: Science, Services and Agents on the World Wide Web 6(4), 283–290 (2008)
13. Wielemaker, J., Hildebrand, M., van Ossenbruggen, J., Schreiber, G.: Thesaurus-Based Search in Large Heterogeneous Collections. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 695–708. Springer, Heidelberg (2008)