CWI

Searching in semantically rich linked data: a case study in cultural heritage

M. Hildebrand, J.R. van Ossenbruggen, L. Hardman,
J. Wielemaker, G. Schreiber

INS-1001

Centrum Wiskunde & Informatica

Centrum Wiskunde & Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Science Park 123, 1098 XG Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3681

Searching in semantically rich linked data: a case study in cultural heritage

Michiel Hildebrand [a,b,*] Jacco van Ossenbruggen [a,b] Lynda Hardman [a,1] Jan Wielemaker [b] Guus Schreiber [b]

[a] *Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands*
[b] *VU University Amsterdam, de Boelelaan 1081a, 1081 AH Amsterdam, The Netherlands*

Abstract

Traditionally the relations between concepts from a controlled vocabulary, such as the hierarchical and associative relations in a thesaurus, have been used to support users in their search process. In the context of the Semantic Web, multiple interlinked vocabularies are becoming available, providing a large number of different relations between concepts. However, for a specific search task, only a small fraction of these will be meaningful to the user, and currently we have little understanding of which methods can be used to determine this.

In this paper, we describe a case study in the cultural heritage domain that investigates support for the specific task of finding artworks in a data set of multiple linked art collections and vocabularies. In a first experiment a number of use cases from domain experts are collected and the paths in the data graph by which artworks can be found are analysed. A number of different types of paths are identified and their usefulness is qualitatively evaluated. In a second experiment we explore how the different path types can be used in a semantic search algorithm to support the intended search behavior indicated by the experts. We conclude that effective end-user support requires a highly interactive application in which the user can explore multiple search strategies. Based on our findings we discuss the implications on the design of such an interactive search application.

Key words: Semantic search; linked data; cultural heritage; user study.

1. Introduction

The term "semantic search" has been used to refer to a wide variety of search strategies. Some of these are based on logical inference, others on smart use of statistics, and yet others on natural language processing. Within the Semantic Web community, the focus of semantic search has been on improving the search process by explicit use of knowledge encoded in RDF/OWL. In previous work [5], we surveyed several RDF and OWL-based search systems. The survey showed that different systems use different types of relations from the RDF data for different tasks. We have, however, little experimental evidence to support claims about what types of relations in real world RDF data are relevant to which search tasks and how these relations are best deployed in a semantic search engine to support the user with a particular task. In this chapter we present a case study that investigates which relations between queries and search results are present in the data, and to what extent these relations are

* Corresponding author. Phone: +31 (0)20 5987740; Fax: +31 (0)20 59287728
Email address: michielh@few.vu.nl (Michiel Hildebrand).
[1] Lynda Hardman is also affiliated with University of Amsterdam.

relevant for users with a specific search task.

Our case study makes use of the cultural heritage domain, where searching for artworks can be a time consuming task, even for domain experts. To satisfy their non-trivial information needs, they often need to formulate multiple queries and manually combine and integrate the various search results into a single coherent set of answers [1]. We investigate how linked data can be used to support the user with the task of finding artworks. In particular, we focus on semantically-rich and heterogeneous linked data, where artworks from multiple collections have been annotated with terms from multiple structured and interlinked vocabularies.

To better understand how the relations in the data can be used to link queries to artworks we analyse three concrete use cases that are collected during interviews with domain experts. Our first finding is that the queries from these use cases can be successfully matched to literals in our data set, and that many of these literals are indeed directly or indirectly related to artworks. Our second finding is that, because of the heterogeneity of the data, there is a large number of different types of related terms that are potentially useful.

To deal with the heterogeneity of the data, we classify the relations into six path types. In a second round of interviews with the domain experts, we solicit their feedback on the relevance of these path types. Our key finding here is that while experts find the information resulting from all path types potentially relevant for their search process, if and how they would like to practically use it depends on many factors. This suggests that effective semantic search for experts in this domain can only be realised in a highly interactive search application.

To support different types of search strategies in an interactive search process we explore the applicability and configuration of the six path types in a graph search algorithm. For this purpose we collected the 25 queries most frequently submitted to a semantic search engine for cultural heritage. Using these queries we investigate the effect of different configurations of the graph search algorithm on the results. Based on our findings we discuss the implications for the design of an interactive semantic search application in the cultural heritage domain.

The chapter is organised as follows. In the next section we explain our study setup. In section 3 we describe the linked data set used in the study. In section 4 we explain the expert use cases and the selection of the test queries. Section 5 investigates how the queries in the use cases could be matched to literals in the data set and how these literals can be related to artworks. The large number of paths are abstracted to six path types. A qualitative evaluation of the relevance of these path types for the expert use cases is presented in section 6. How to implement the path types in a semantic search algorithm is explored in section 7. We discuss the implications of our findings on the design of interactive search applications in section 8, here we also include references to related work. Finally, section 9 presents the conclusions.

2. Study setup

For this study we collected two sets of test queries used to find artworks in a large collection. For the first set, we collected the top 25 most popular queries from the logs of the online Europeana "ThoughtLab" search engine. [2] The queries cover a variety of different categories. In addition to this set of queries, we collected in-depth information about the use of text-based queries in expert use cases. We interviewed three domain experts from the Rijksmuseum Amsterdam about information needs they recently encountered in their own work. The experts were chosen to cover different areas of expertise, including a librarian assisting in external requests over the whole collection, a specialist in Japanese prints and an expert in middle age prints. Details of the collected test queries, the domain experts and their use cases are discussed in Section 4.

We use the collections of annotated artworks and controlled vocabularies used in the Europeana "ThoughtLab" as the data set for our experiments. Details about this data set are given in the next section.

We focus our study on the investigation of two research questions: (i) Which relations in linked data are useful in professional artwork search, and (ii) how can these relations be used in a semantic search application? To answer the first research question we analyse the different path types in the linked data and qualitatively evaluate these types in a user study with domain experts. Based on the findings of the first experiment, we perform a second experiment where we explore how the path types indicated useful by the domain experts can be exploited in a semantic search application.

[2] `europeana.eu/portal/thought-lab.html`

First experiment — For the selected queries, we generate the paths of relations available in the data. We use a graph search algorithm to compute all paths up to path length 6. By analysing the paths we collect first insights on how the relations in the data can be used to relate artworks to the queries. In follow up interviews with the domain experts we collect qualitative feedback for different path types.

Second experiment — For the 25 queries from the search logs we analyse the results that can be found with the different path types. We explore how the algorithm should be tuned to approach the desired behaviour indicated by the domain experts.

3. Data set

We use an existing data set: the data from the Europeana "ThoughtLab". Figure 1 provides an overview of the sources in the data. There are three types of data sources: *collections* describing works of art (the circles with a coloured fill the figure), *vocabularies* used to annotate the artworks (the remaining circles) and *alignments* among the vocabularies (the arrows between the vocabularies). Note that there are many different vocabularies, and only three different collections.

Collections — The artwork collections used are the internal collection database of the Rijksmuseum Amsterdam, the RKDimages database of the Netherlands Institute for Art History (RKD) and the Atlas database of the Musée du Louvre. For all three sources, only the artworks for which images are available on the Web were converted to RDF. The results are RDF descriptions of about 170,000 artworks.

The three institutions have described their artworks using different and rich metadata schemata. These original schemata have been translated directly to RDF, resulting in 469 different metadata properties on subjects of type `vra:Work`. All these properties have then been mapped to VRA, a specialisation of Dublin Core for visual resources. The properties have a wide variety of values. Some are short RDF literals, such as titles, measurements and dates. Longer RDF literals include descriptions, literature references and editorial notes. Some values point to terms defined in one of the controlled vocabularies, while others are structured (blank node) values. The latter are used to capture relations with arity > 2, e.g. that an artwork was part of a collection, but only during a specific period. In these structured values, `rdf:value` is used to indicate the "main" value of the property, in the example above the name of the collection. In an informal context we can often ignore this use of blank nodes and simply consider the main value as a direct property of the artwork.

In some cases there is no clear choice between modelling a metadata property as a literal or as an object property. For example, some institutes use literal values for `dc:creator`, often with conventions on how to spell an artist's name, while others fill the same field with a pointer to an entry in a predefined list of artists. In fact, any literal property value can be replaced by a term which has the same literal as an `rdfs:label`, and in the data set both modelling conventions are used. Some fields, however, have a such a wide range of values that it becomes virtually impossible to predefine in a vocabulary. Titles, descriptions and editorial notes, for example, are typically free text fields in most collections, and represented as literal RDF properties in the data set.

Vocabularies — All three institutions use and maintain their own in-house vocabularies, that typically describe people, locations, events and concepts. The Rijksmuseum and RKD also use concepts from the IconClass [3] classification system, currently maintained by RKD. In addition, the data contains several external vocabularies. From Getty it includes the United List of Artist Names [4] (ULAN), the Thesaurus of Geographic Names [5] (TGN) and the Art and Architecture Thesaurus [6] (AAT). Finally, three interlinked lexical sources, the W3C's RDF version of Princeton's WordNet [7], the Dutch lexical semantic database Cornetto [8] and the French Wolf version of WordNet [9] are included. All vocabularies are either directly modelled in SKOS or mapped to SKOS using `rdfs:subClassOf` and `rdfs:subPropertyOf`.

Cross-vocabulary relations — The in-house vocabularies have been (partially) aligned with those from Getty, for example, vocabularies for persons are aligned with ULAN, those for locations

[3] `www.iconclass.nl`
[4] `www.getty.edu/research/conducting_research/vocabularies/ulan`
[5] `www.getty.edu/research/conducting_research/vocabularies/tgn`
[6] `www.getty.edu/research/conducting_research/vocabularies/aat`
[7] `www.w3.org/2006/03/wn/wn20`
[8] `www2.let.vu.nl/oz/cltl/cornetto`
[9] `alpage.inria.fr/~sagot/wolf-en.html`

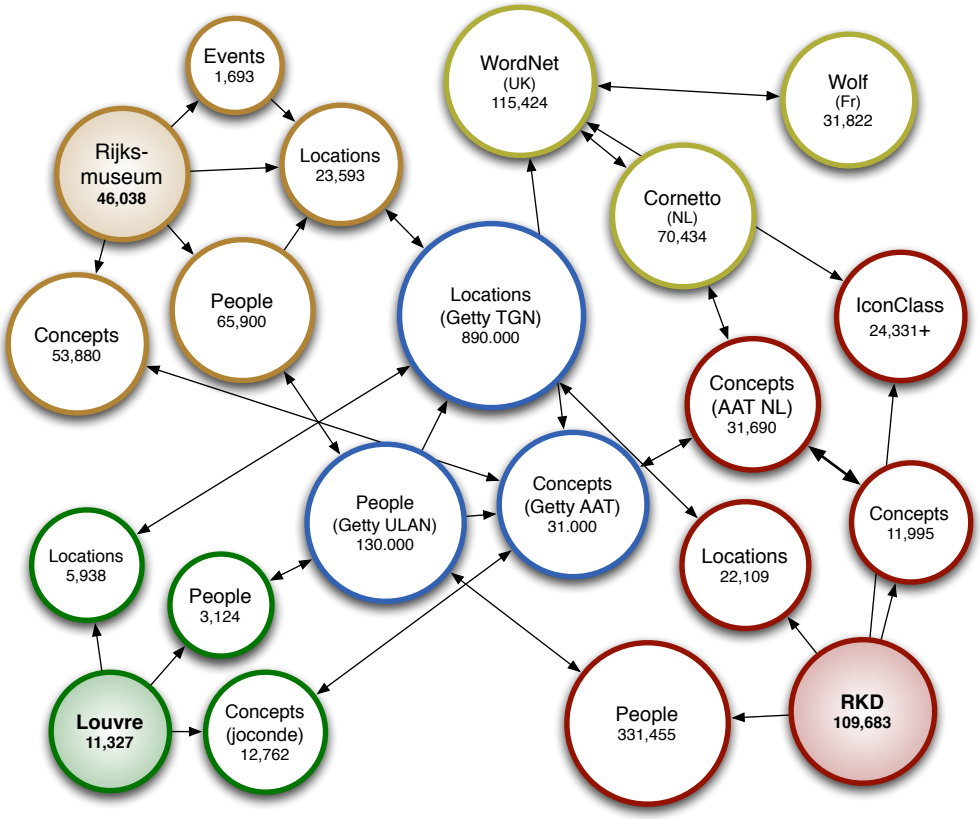

Fig. 1. Datacloud of the Europeana "ThoughtLab" data set

with TGN and other concepts with AAT. IconClass and AAT are also aligned with WordNet. For most alignments `skos:exactMatch` is used, some others use `owl:sameAs`. In the data set alignment relations may also occur within a single vocabulary, typically to state that two terms that were once thought as being distinct, are now to be considered equivalent. These alignments were already present in the original vocabulary data provided to the Europeana project. The majority of the alignments, however, relates terms from different vocabularies and are the result of automatic or manual vocabulary alignment efforts.

In addition to alignment relations, there are also vocabulary-specific relations that link terms from different vocabularies. For example, the locations from TGN are used as values for the birth places of persons in ULAN.

For more details on the conversion of the original data to RDF we refer to [15,18] and for details on the vocabularies alignment to [16].

4. Collecting the query test set

Category	Query
`Concept:`	book, war
`Location:`	portugal, spain, rome, italy, greece, paris, poland, romania
`Museum:`	prado, louvre
`Painting:`	mona lisa
`Style:`	renaissance
`Person:`	klimt, van gogh, vermeer, rubens, goya, shakespeare, munch, da vinci, monet, renoir, hitler

Table 1
Top 25 queries from the January 2009 logs of the Europeana online "ThoughtLab" demonstrator and the inferred categories.

The 25 queries in Table 1 are taken from the January 2009 logs of the Europeana online "ThoughtLab" demonstrator. While we do not have exact demographic data, we assume most visitors are lay persons and not art experts. In total, the log of that

month contained almost 13,500 session cookies, and 7,330 unique queries. After removal of the queries listed as examples on the website, and those used by the project members for demonstrations, the remaining top 25 queries were selected. When we interpret these queries we infer that two refer to general concepts. All other queries refer to names: eight to location names, two to museum names, one to the name of a painting, 11 to person names and one to the name of a style/period.

These types of queries are comparable to those found in a more extensive study by Trant on the search log data from the Guggenheim Collection online [17]. We find the same focus on named entities as Trant, and, in particular, artists names are searched on the most. Only we find fewer references to styles and periods, and more to locations.

4.1. *Expert use cases*

The first set of interviews with the domain experts from the Rijksmuseum Amsterdam were centred around a search session within area of the expertise of the participant. We asked the participants to reproduce an information need that they had recently encountered, and perform the actions to satisfy this need with their own tools. We asked them to think aloud and explain their actions. For each interview, we recorded the search terms the experts entered and their motivation to choose these. The interviews and the search queries were performed in Dutch. In consultation with the participants we translated the topics and queries into English.

4.1.1. *Use case: peddler*

The first participant (P1) is a librarian of the Rijksmuseum's reading room. One of her tasks is to assist cultural heritage researchers in finding museum objects. In the interview she explains how she assisted a researcher with a study into the different ways "peddlers" (a kind of travelling merchant, in Dutch: "marskramers") have been depicted on historical prints and paintings. The initial task of the librarian is to collect evidence that the collection contains a variety of artworks depicting peddlers.

The librarian starts her session by searching for museum objects in the museum's collection management system, using the query `peddler`. She first searchers in the title and second in the description field. These searches return only a few objects. To find more objects she tries a new query `street vendor`, which is a different type of travelling merchant. This search term also returns a few objects. She also tries `street salesman`. To get other ideas she turns to the library system to find books about the topic. Using again the query `peddler` she finds a book about the topic. She selects the book from the library and finds several prints depicting peddlers. As these prints are made by `Pieter Breughel`, she returns to the collection management system to find objects made by this artist. Although there are many artworks that do not depict peddlers, one of the artworks that was displayed in the book is found. She concludes that there is sufficient material about the topic. She reports back to the researcher, expecting to return later for a more thorough investigation.

Because `peddler` was the query initially entered by the user, and a typical example of a vocabulary concept, we will use this query in the remainder of the article as the main query associated with this use case.

4.1.2. *Use case: Fuji*

The second participant (P2) is a cataloguer of the Rijksmuseum print room, specialised in Japanese prints. At the time of the experiment she was not performing her own research, so we together discussed a possible query topic: artworks by Ando Hiroshige that depict mountains. She initially explained that this topic was too broad to be realistic. After the search session, however, she explained that the session was typical for her search behavior.

Within the Rijksmuseum's internal database there are 163 prints from Ando Hiroshige. In the search session, the participant searches within this set. She starts by entering the query `mountain` in the title field. As there are only 3 artworks, she adds the description field. Ten more artworks are found. She states: "The chance is high that there are landscape paintings that contain mountains, but this has the disadvantage you might get artworks without mountains". She adds the search term `landscape` as a disjunctive query in the description field. Four additional artworks are found, from which one indeed contains a mountain. She recognises that this is the Japanese mountain `Fuji`, which she tries as her next search term. There are 11 artworks that depict this mountain. She also tries the related term `valley` in the description, which does not have any results. She explains that she could continue with this process for a while.

When asked for other methods to search, she explains that if you know that an artist has created artworks about a topic in a specific period, you can look for other artworks within this period. Or if you know when a volcano erupted, you can search for artworks that are created shortly after that date within the same region.

Most queries used in this use case are typical thesaurus concepts, which are represented in the data in a way similar to that of the "peddler" concept of the previous use case. The query `Fuji`, however, refers to a geographical name, and has different characteristics. Because we know from the Europeana logs that location names are frequently queried for, we will focus on this query as the main query associated with this use case.

4.1.3. *Use case: Gregory*

The third participant (P3) is also a cataloguer of the print room. She is specialised in prints from the Middle Ages. In addition to her work as a cataloguer, she investigates a specific technique used for illustrations in Middle-Age books. The Rijksmuseum print collection also contains prints that were originally parts of books, and she tries to discover if any of these are made using this specific technique.

As the technique of her interest is very rare and not named or described yet, her main strategy is to query on topics she knows that are used in the books. In previous research, she discovered that the `Gregorian mass` is one of these topics. She uses this as a search term to search for prints in the Rijksmuseum collection. Several prints contain the search term in the title, but upon further study none of these are made using the technique. To find more prints she also tries `gregory` to find artworks depicting the pope "Gregory the Great" who was involved in this mass. This returns fewer than 20 artworks, which can be studied one by one. She also tries the search term `mass`. For this, many more artworks are found and she wants to further constrain this set by place and time. In previous research she discovered that the print technique is used in Germany between 1400 and 1500. She demonstrates how a different database supported her by allowing constraints between 1400 and 1500. She also mentions that she would like to search on all locations within Germany.

Again, most queries correspond to concepts. Only `gregory` refers to a person's name, the other main type of query we found in the Europeana logs. We will therefore focus on this query when further discussing this use case.

5. Analysis of relations found

To find the relations between queries and potentially related artworks in the RDF data set, we apply a graph search algorithm [19]. To match queries to RDF literals, we use the algorithm's default string matching technique based on Porter stemming [12] and tokenization. The directional graph search traverses the graph from objects to subjects, but not the other way around: only symmetric properties and properties with an explicitly defined inverse are traversed in both directions. We set the maximum path length to 6, counted by the number of properties. In our experience path lengths above 6 become unrealistic to compute in reasonable time. For more details about the algorithm used, we refer to [19].

5.1. *Expert use cases: path analysis*

We analyse the paths from query to artwork that can be found for the three queries described in the expert use cases. We define a path as a series of triples, where the object of the one is the subject of the following. The last object is the literal that matched with the query and the first subject is the artwork found. At path length 1, artworks are, thus, related by an RDF property with a literal value that matches the query. At path length 2, the artworks are related to a vocabulary term that is in turn related to a literal matching the query. As we are interested in the different ways to find artworks and not how to find vocabulary resources, we consider all properties to find artworks and only one path for each unique vocabulary term.

Table 2 shows for each query and path length the total number of artworks and the total number of different paths needed to find these artworks (displayed as #path #artworks). At path length 1, for all queries only a small number of artworks are found and almost as many paths are required to find these. In other words, almost all artworks are found via a different literal. At path length 2, relatively few paths are required to find the artworks. In this case the different values by which the artworks are found are, thus, represented by a single vocabulary term.

At path length 6, more than 1,000 artworks are found for all queries. For the queries `peddler` and `gregory` over 1,000 results are already found for path lengths 4 and 5. For the query `Fuji`, however,

path length:	1		2		3		4		5		6	
`peddler`	46	47	2	104	1	128	5	2,106	137	14,189	260	33,909
`fuji`	15	15	2	3	0	0	2	48	1	1	15	1,514
`gregorius`	56	68	5	5	4	9	1	40	33	235	119	1,146
`gregory`	8	9	11	57	2	3	19	158	64	3,714	154	8,992

Table 2
For each query from the expert use cases the number of different paths and the artworks found these paths (#paths #artworks).

only a small number of artworks is found at path lengths 3,4 and 5. For comparison we also generated the paths for the 25 queries from the search logs (not shown in Table 2). From the 25 queries, 16 are already related to more than 1,000 artworks at path length 4. The queries with generic concepts, e.g. "war" and "book", well known persons, e.g. "van gogh", and locations, e.g. "rome" and "paris", have over 10,000 results at path length 4. The small number of results for the query `Fuji` at lengths 3 and 4 should thus be seen as an exception, caused by a limited amount of information available on the topic.

For the first expert use case (`peddler`) we describe in detail the different paths found at different path lengths and discuss our findings from the analysis of these paths. For the other two use cases we highlight the similarities and differences compared to the first use cases.

5.1.1. *Use case: peddler*

At path length 1 artworks are related by an RDF property with a literal value that matches the query. The query `peddler` matches with literals used as titles, descriptions and editorial notes of 47 artworks. The Rijksmuseum provides 15 of the artworks, while the other 32 are from RKD. With the exception of the title "The Peddler", which occurs twice, all other literals are unique. An example path of length 1 is:

rma:SK-C-1346	dc:title	"The peddlers"

At path length 2, the concept peddler in the RKD subject thesaurus is found via a label matching the query. It is used to describe the depicted subject of 104 artworks from the RKD collection. Note that almost all artworks found at length 1 had a different path, while the 104 artworks at length 2 are found by only two vocabulary terms (see the second column in Table 2). All artworks found by this path are from the RKD collection, as only these are described with the concept peddler from the RKD in-house thesaurus. An example path is:

rkd:68359	dc:subject	rkd:peddler
rkd:peddler	skos:prefLabel	"peddler"

One artwork is also found by a different path at length 2, because its title is modelled as a compound object with the title as the value of the `rdf:value` property.

At path length 3 there is only one path, resulting in 128 related artworks. These are described with the concept salesman, a more general concept of the RKD concept peddler. An example path is:

rkd:9429	dc:subject	rkd:salesman
rkd:salesman	skos:narrower	rkd:peddler
rkd:peddler	skos:prefLabel	"peddler"

At path length 4 there are five different paths, resulting in 2,106 artworks. Four of these paths contain vocabulary terms related to the concept salesman in the RKD thesaurus. One of these contains the more generic concept professions. The other two are more specific terms of salesman: market_salesman and fish_salesman. These concepts are thus siblings of peddler. The activity of trade is found by a `skos:related` property. An example path is:

rkd:60688	dc:subject	rkd:trade
rkd:trade	skos:related	rkd:salesman
rkd:salesman	skos:narrower	rkd:peddler
rkd:peddler	skos:prefLabel	"peddler"

The largest number of artefacts (1,772 out of 2,106) are found via the concept basket from the RKD thesaurus. This concept is found through an equivalent concept in Cornetto WordNet. The concept is a more generic term of a type of basket used by peddlers, in Dutch named a "mars".

rkd:57252	dc:subject	rkd:basket
rkd:basket	skos:exactMatch	wn:basket
wn:basket	wn:hypernymOf	wn:mars
wn:mars	wn:gloss	"peddler basket"

At path length 5 the number of paths drastically increases and results in over 14,189 related artworks.

All 137 paths use relations from the RKD subject thesaurus, of these, 117 lead to sibling concepts of salesman. In fact, we have now reached all professions in the RKD thesaurus. Eight terms are related to the activity of trade: six are `skos:related`, such as scale and market stall, one is the more specific concept money trade and another is more generic concept. Seven paths contain vocabulary terms related to the concept basket, including specific types of baskets, such as fruit_basket. The final five paths are concepts found via `skos:related` and `skos:broader` properties from salesman and `market_salesman`. As we start to drift off topic we do not discuss the more than 33,000 results at path length 6.

We conclude the analysis of this use case with three findings. First, some artworks can only be found via literal properties, because they have not been explicitly annotated with a vocabulary term that matches the query `peddler`. Searching with the vocabulary concept peddler as the value of a `dc:subject` property only finds artworks from the RKD collection, and none from the Rijksmuseum. All results from the Rijksmuseum collection are found via literal properties, such as `dc:title` and `dc:description`. This explains why expert P1, who is familiar with the collection, searched on these literal properties during the first interview.

Second, a large number of additional artworks were found at lengths 3 and above. The majority of these different paths involves some combination of thesaurus and alignment relations. The more than 14,000 artworks found at path length 5 is overwhelming. From our analysis it is unclear *a priori* which paths include results relevant to the search task.

Our third finding is that some of the paths contain concepts that are, to a large extent, similar to the alternative queries used by our expert user, but not exactly the same, e.g. salesman. In addition, the paths contain many other vocabulary terms for which it is also not clear *a priori* if they are relevant to the search task.

5.1.2. *Use case: Fuji*

For the query `fuji` similar types of relations are used at path length 1 as in the first use case: titles, descriptions and editorial notes. In this case only 15 artworks are found, but again all results are found via different literals. Interpreting the literals showed that most were about Mount Fuji, and one was about the Fuji Photo Film Company.

At path length 2 there are two paths, resulting in only three artworks. Here we also find the two different interpretations of the query, represented by two vocabulary terms: Mount Fuji and Fuji Photo Film Company. At length 3 there are no artworks found.

At length 4 there are two paths, resulting in 48 artworks from the Louvre collection. Both paths include the term Fuji-san from the Joconde thesaurus. One path leads to two artworks depicting mountains via the concept mountain, which is related by two `skos:broader` relations to Fuji-san. Another path contains the sibling concept of Fuji, the vulcano Vesuvio. At path length 5 another sibling is found, in this case an additional step is required as the path goes via the term the Alps.

At path length 6, 15 paths are found, resulting in 1,514 artworks. The majority of the paths (13), contain geographical concepts from the JOCONDE thesaurus. 453 artworks from the RKD collection are found. For these artworks the concept mountain from the RKD thesaurus is used as a value for the `dc:subject` property. This concept is found through an alignment with WORDNET, where it is related to Fuji by three `wn:hypernymOf` relations. Via a similar path, artworks from the RKD collection are found that depict cherry trees. In WORDNET, fuji is also defined as a specific type of cherry, and through `wn:hypernymOf` relations the generic concept of cherry tree is found.

For the query `fuji`, the number of artworks found is considerably less than for the query `peddler`. Most other findings are, however, similar. We do not find all artworks depicting Mount Fuji by looking only at artworks with concept Mount Fuji as a `dc:subject`. At longer path lengths we find related vocabulary terms, including some related to the queries from the expert use case, e.g. mountain. Only on a few of these artworks mount Fuji is depicted. An additional finding is that the query has multiple interpretations. At path length 1 these interpretations are implicit in the individual literals by which the artworks are found. At path length 2 the interpretations are explicitly represented by different vocabulary terms.

5.1.3. *Use case: Gregory*

We analyse the paths for the query in Latin (`gregorius`) and in English (`gregory`), as there are no alignments between the concepts in the different languages.

At path length 1 there are 68 artworks found for the query `gregorius`. The artworks are from RKD and the Rijksmuseum. Some of these are related to the "mass of gregorius", others to "pope gregorius" and again others to topics unknown to us. For the query `Gregory` the matching artworks are about many different topics. For example, several artworks are found because the background literature used to describe the object is written by "J. Gregory".

At path length 2 the query `Gregory` leads to nine paths with a vocabulary term from IconClass, of which seven are events that include "Gregory the Great". For the query `gregorius`, the matching vocabulary terms are other persons. Two persons are found because the query matches with their biography. Reading this biography we discover that one is a cousin of Gregory the Great.

At path lengths greater than 3, the vocabulary terms found for the query `gregorius` are linked to interpretations of the query that are not related to Gregory the Great. A large variety of relations are used in these paths, such as the collection-specific properties `granted_privilege` and `assigned_to` relations between persons `assisted_by`, `teacher_of` and `sibling_of`. Other persons are found because they are born in a place with a matching name.

For the query `Gregory` similar types of paths are found. We also find at path length 5 relations from WordNet. For example, 58 artworks are found because they depict a pope. WordNet contains a `wn:hypernymOf` relation between Gregory the Great and the concept pope. For a different vocabulary term, a `wn:hypernymOf` relation leads to the concept saint, resulting in an overwhelming 2,849 artworks.

Again we conclude that relevant artworks are found by matches on literal properties as well as via vocabulary terms. At longer path lengths there are many different paths and artworks found. Multiple interpretations of the query are also found. An additional finding is that different interpretations of the query all lead to more related results, whereas only the paths from one or a few interpretations are useful. Another finding is that vocabulary terms are found via labels as well as descriptions, e.g. a biography, where the relation to the query is implicitly captured in the text.

5.2. *Abstraction of the paths*

From the use cases above we conclude that a large number of artworks can be found via many different paths. The large number of resulting artworks makes evaluation based on a domain expert scoring the relevance of each artwork unrealistic. Even the number of different paths is too large to be scored individually, also because the semantics of many of the longer paths and the subtle differences between them are hard to express in natural language or by some other understandable means. We therefore look for frequently recurring types in the paths we found, and see if we can use these path types to classify all relations into a small number of abstractions.

5.2.1. *Metadata properties*

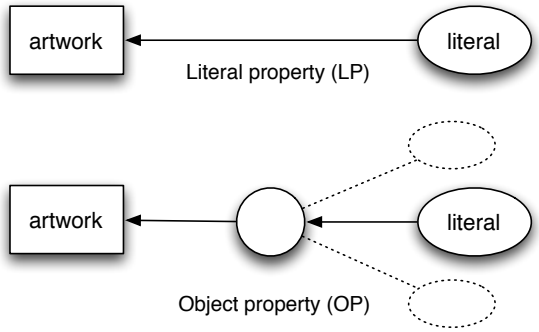

Fig. 2. **Metadata paths**: Two types of paths between a literal and an artwork. The arrows are shown in the search direction, from object to subject.

The paths found in the three use cases show a clear distinction between those that directly use the metadata properties on the artworks, and longer paths that include relations between terms. With metadata properties we refer both to the paths of length 1, where the query is matched to the value of a literal property of an artwork and of length 2, where the query matches a label of a vocabulary term. These two types of paths are represented in Figure 2. The path type labelled *literal property* represents all paths with a direct relation between the matching RDF literal and an artwork via an RDF property. In the three use cases artworks were found using RDF properties for titles, descriptions and notes. Note the arrows are shown in the search direction, thus from object to subject.

The path type labelled *object property* represents all paths where the matching literal and artwork are connected through a single resource. In the three use cases, artworks were found by vocabulary terms, such as persons, locations, events, domain specific concepts and collection names. These terms were related to the artworks using high level properties,

such as `dc:subject`, as well as collection specific properties, such as `granted_privilege`. Some artworks were found by an object with a matching literal from the `rdf:value` property.

5.2.2. *Relations between vocabulary terms*

The relations between the terms used in the paths longer than 2 follow from different schemata in the data set. The vocabularies are either directly modelled in SKOS or they have there own schema, which is mapped to SKOS. At an abstract level the relations between vocabulary terms can, thus, be defined in SKOS using the hierarchical relations, `skos:broader`, `skos:narrower`), the associative relation `skos:related` and alignments, such as `skos:exactMatch`. Our data set only contains equivalence alignments. We describe how these relations create different path types to find artworks.

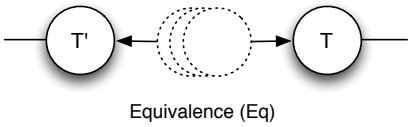

Fig. 3. **Equivalence**: Vocabulary terms defined as equivalent.

The path type in Figure 3 labelled *Equivalence* represents a path that aligns two equivalent terms. This path can be used both directions, as the equivalence relation is symmetric. Equivalence between two vocabulary terms can be defined directly, for example, in the first use case the concept basket from the in-house thesaurus of RKD was found via an equivalence alignment with WordNet. An equivalence alignment can also cover multiple terms.

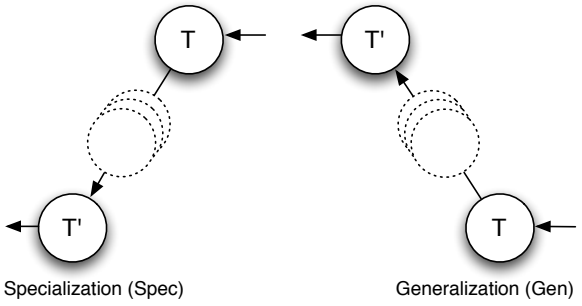

Fig. 4. **Hierarchical**: Two types of hierarchical paths between vocabulary terms: specialisation and generalisation.

Figure 4 shows two types of paths to connect vocabulary terms by hierarchical relations. The path type labelled *specialisation* defines the relation by which a more specific (or narrower) term is found. The path type labelled *generalisation* defines the opposite relation by which a more generic (or broader) term is found. For example, in the "peddler" use case the concept salesman was found as generalisation of the concept peddler. In WordNet the concept mountain was found as a generalisation of Fuji via three relations. By combining the a generalisation and a specialisation a sibling term can be found. For example, in the "peddler" use case the concept fish_salesman was found as a sibling of the concept peddler.

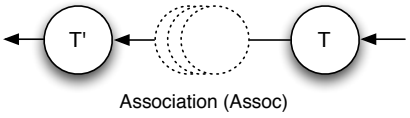

Fig. 5. **Association**: Vocabulary terms associated by one or more relations.

The path type in Figure 5 labelled *Association* represents a path between two vocabulary terms connected by an associative relation. By this path type an associated term can be found. For example, in the first use case the concept the activity trade was found via an association to salesman. A single path can contain multiple association relations.

The four types of paths between vocabulary terms provide a means to extend the object property path, replacing the single node with one or more related terms. Inclusion of the different path types has different effects on the relation between the artwork and the query. In the following two sections we investigate the role of the different path types in the search process, by (i) a qualitative evaluation with the domain experts and (ii) a qualitative analysis of the results found.

6. Qualitative evaluation of path types

Having identified the six path types through analysis of the data when then evaluated to what extent the domain experts deem the information found by these path types relevant for their search activities. We would like to evaluate the potential relevance of the information itself, so we need to avoid that the experts' feedback is influenced by RDF modelling flaws in the RDF data set, bugs in the search engine or confusing elements in the user interface of the Europeana "ThoughtLab". To achieve this, we manu-

ally select web pages on the websites of the organisations that originally provided the data before it was converted to RDF, where each web page shows some of the information directly resulting from applying one or more path types to the original query. In this way, we collect six sets of web pages for each query, where each set covers a specific type of information, and each set consists of examples from the different websites of the original content providers.

Each interview took place in the museum, using a museum computer to show the participants the various webpages. Each interview lasted one hour, was voice recorded and notes were taken by the conductor of the experiment and a second observer. After reading a short description of the goals of the study and the outline of the experiment, the participants were shown the six sets of pages. After seeing each set, they were asked to comment freely on what they had seen and were also asked five or six directed questions.

6.1. *Searching free text fields*

To get feedback on the role of the different metadata properties in the search process, we showed the experts a set of pages from the collection websites. We used pages that resulted from clicking on an individual search result associated to one of the queries used in the first interview. We selected results that were found by matches on different properties, e.g. a page showing a painting with "peddler" in the title, another page with a print with "peddler" in the description and a final one with "peddler" in the depicted subject field. We asked the experts in which fields they would search.

All experts commented that searches on a controlled field, in general, yield incomplete results. [P1]: *"When I only need one or two examples, the thesaurus-based search is best, because it will yield exactly the examples I need . . . but if I really need all depictions of a specific topic, I will also search on free text fields such as title and description, because cataloguers never add* all *relevant terms to the subject field."* [P3]: *"It depends if you know how thorough a collection has been annotated. I know that the subject field has not been used for all objects of this collection."*

Experts strongly prefer to search in fields using controlled vocabularies. In practice, they need to search on the literal properties of free text fields, because they know the annotations with terms from controlled vocabularies are often incomplete.

6.2. *Search using controlled vocabularies*

A key difference between searching object property fields and searching in literal property fields is that the vocabularies can be used to explicitly disambiguate the query if it has multiple interpretations (i.e. homonymy). We used web pages of the thesauri providers' websites (i.e. from the Getty AAT, ULAN and TGN website, Princeton's Wordnet site, the Joconde site of the French ministry of culture, the RKD site for Iconclass). Each web page showed the search results for the query in the thesaurus, showing all different interpretations of the query.

All experts appreciated being able to see the different interpretations of the query explicitly . [P1]: *"Yes, if you know there are multiple meanings, you know in advance you can expect lots of noise in the search results"*. All experts also like the feature of thesaurus-based systems to search on only one specific meaning of a term. [P2]: *"If I search only on a thesaurus-controlled field, I would trust the list of search results more, and would not click every result to check why it is a match"*. Again, P3 uses this as a strategy to deal with errors in the data: [P3]: *"I would trust the results more than those of a free text field . . . but I would also search on the other interpretations, just in case the cataloguers have made a mistake"*.

We conclude that when a query has multiple interpretations, experts would not only like to be able to disambiguate and search using the intended interpretation, but also to be made aware of other possible interpretations.

6.3. *Search using equivalences*

The key feature of the equivalence path type is that it links terms from different thesauri that have the same or similar meaning. We again selected web pages from the thesauri web sites. We showed the participants that there are often multiple thesauri containing terms relevant to their query, and showed the experts that different thesauri encode different types of information related to their query. Again, we asked them how useful these different information elements would be in their search process.

All experts found the thesauri that provided name variants (e.g. Fuji versus Fujisan) and spelling variants (e.g. Fujisan versus Fuji-San) extremely

useful, especially for person and location names. P2, looking at the name variants for Fuji listed in TGN: [P2]: *"Yes, this is very useful indeed. I would search on all variants listed to see if the results would yield additional results."* Also the multilingual aspects were considered useful: [P1]: *"Having a domain-specific thesaurus in another language is very useful, as normal dictionaries do not always cover the jargon I am looking for. In addition, we are an internationally oriented museum, and the search interface software on our website is multilingual. But the content is often still in one language, which is confusing."* [P3]: *"For the names of saints, even if you are using the Latin name, there are always subtle differences in different languages. So this is very useful."*

We conclude that the experts consider equivalent relations across thesauri more generally useful: their applicability is not limited to specific cases. They seem most useful when the links to other thesauri bring in extra name or spelling variants, or translations to other languages.

6.4. *Search using specialisation and generalisation*

We selected web pages from the thesauri websites, with the term hierarchy fully expanded wherever possible, showing all broader terms to the top of hierarchy, and narrower terms wherever applicable.

When confronted with the full hierarchy experts responded positive [P3]: *"Being able to move down the hierarchy is useful for query refinement when you have too many results, or for broadening when you have too few"*. Most focussed on the narrower relations: [P1]: *"In general, the more specific you can be, the better."* All three also indicated a possible use for the more generic terms, but only in cases where few results are found. P2 explained she might use more generic terms, but only in combination with other terms or restrictions. After seeing that "Fuji" was a narrower term of "mountain" in one thesaurus and of "Japan" in another: [P2]: *"I would do a new search by combining both "mountain" and "Japan""*.

The experts did not express a clear preference for the hierarchy from one thesaurus above the other. One, however, expressed the need for semantic integration of the different hierarchies. [P1]: *"In the ideal case, you should be able to use all different thesauri in a way that is fully integrated . . . but I do not know if that is possible . . . it should be done right though, otherwise I would not trust it."*

Experts had mixed opinions about the relevance of "sibling" terms, e.g. terms with the same parent. [P3]: *" Maybe I would use these if the original query yields insufficient results . . . but even then, only if the terms are semantically close to the original query."* [P1]: *"No. I would never use siblings in this case, as it would not give new results. Fuji is by far the most important feature in this region. If a print has been annotated with something else from this region, it would not depict Fuji . . . otherwise Fuji would have been added as an annotation as well."*

We conclude that experts see potential in using hierarchical relations for search, but only if the other terms are semantically close, and even if this is the case, they would only use them in specific cases.

6.5. *Search using associations*

As this path type is also thesaurus related, we again showed similar thesaurus web pages, this time drawing the participant's attention to the section on the page dedicated to horizontal relations. These relations might differ widely, ranging from general `skos:related` in the RKD thesaurus to specific `ulan:brotherOf` in ULAN.

Experts were all positive about this type of relations. After seeing Fuji being related to "volcano" in WordNet: [P2]: *"People always refer to Fuji as a mountain, never as a volcano . . . but now I see this, I would also search for volcanoes in Japan . . . I would not have thought of this myself."* One expert even uses these relations as part of a strategy to deal with errors in the data. [P3]: *"For artists, knowing family or apprenticeship relations is very important. If I know an artist has a brother, for example, I would always search on the brother too, because works are sometimes attributed to the brother by mistake as the names are very similar. Wrong attributions to students or teachers of an artist are also common."*

Again we conclude that associative relations can potentially yield relevant search results, but *how* to use them varies from case to case.

We conclude that experts prefer searching on a field that clearly indicates the relation to the artwork and for which the annotation terms are taken from a controlled vocabulary: this gives high precision, with results that can be quickly assessed on relevance. They mainly use literal properties when striving for completeness: searching in free text fields will yield additional results, but with lower preci-

sion and the results typically require time consuming inspection to assess their relevance. Equivalence paths seem useful, but mostly for name, spelling and language variants. In addition, experts consider the information provided by the hierarchical and association path types potentially useful in an interactive search application. When and how they like to use this information depends, however, on the context of the search task.

7. Exploring path type configurations

The analysis of the query types showed that a large number of paths relate artworks to queries. In the previous section, we described how experts value the information related to different types of paths, where we manually collected the related information. The question remains how these path types should be applied in a search application. From the analysis in Section 5 we conclude that using all path types for all queries results in too low precision. This is confirmed by the experts who indicated that they would use specific paths only in specific contexts.

To better understand how the different types of paths can be used to effectively support artwork search, more analysis is required. For this purpose we use the 25 queries that were most frequently submitted to the Europeana "ThoughtLab" search engine. For each query we compute the number of artworks found for the different path types. In addition, we compute the number of vocabulary terms found. You find an overview of this recall data in Table 3 (at the end of the chapter).

In the following sections we discuss the different columns of the table. We provide a qualitative analysis for a number of topics, and based on observations, discuss the precision of the retrieved artworks. To cover the full spectrum of relevant topics further research is required.

7.1. *Alternative matches for free text fields*

The experts prefer to search in fields with terms from controlled vocabularies. For the 25 search log queries we investigate if the vocabulary terms and the relations between them provide an alternative to finding artworks by free text fields. For this purpose we first compute, for each query, the artworks found via a literal property (LP). Accordingly we compute how many of these artworks are also found by a matching vocabulary term, either used directly as an object property (OP) or indirectly by a series of relations up to path length 5 (P5).

The column in Table 3 labelled LP shows for each query the total number of artworks found via a matching literal property. The $\cap OP$ column shows how many of these artworks are found via a vocabulary term with a label matching the query. For the `concept` queries none of these artworks are found by an object property. For most of the other types of queries the object property provides an alternative. For example, the painting with the matching title "Van Gogh's bedroom in Arles" is also found via the vocabulary term Vincent van Gogh, which is used as the value of the `dc:creator` property.

On average 42% of the artworks found via a free text field can also be found via an object property. Of the 58% of artworks without an object property many are found via a match on an editorial note. For example, several artworks have a literal property that describes the background literature used for cataloguing, e.g. "the tulip book". These editorial notes match with the queries, e.g. `book`, but the artworks found are in most cases not relevant. For some queries there are, however, also relevant artworks that can only be found via a literal property. For example, in the peddler use case the relevant artworks from the Rijksmuseum collection could only be found via a matching title or description. In general, the artworks found via literal properties cannot be excluded from the search results, but some properties should be excluded to increase precision.

By using longer paths in the graph even more alternative paths become available to find artworks. The column in Table 3 labelled $\cap P5$ shows the number of artworks that can also be found by paths up to length 5. At path length 5, 79% of the artworks found via a free text field can be found via a vocabulary term. For example, a painting created by Giovanni Antonio Boltraffio was found for the query `da vinci` because it matched with the artwork's text of a description property. The same artwork is also found via the vocabulary term for Boltraffio, which is the value of the `dc:creator` property, and this person is related to `Leonardo Da Vinci` by the associative relation `ulan:student_of`. We discuss in the following paragraphs how to exploit these types of paths.

7.2. *Using vocabulary terms for query disambiguation*

The experts considered the vocabulary terms useful to disambiguate the query. We investigate which vocabulary terms are good candidates for this type of query disambiguation. As a baseline we collect all vocabulary terms with a literal matching the query. Next, we compute the different literal properties by which these vocabulary terms are found and the paths that relate them to artworks.

The *All* column in Table 3 shows the total number of vocabulary terms with a literal matching the query. For the `concept` and some `location` queries many different terms are found. Also for some `person` queries more than 100 different matching vocabulary terms are found. On further analysis we observe that a large part of these vocabulary terms are found via descriptions, such as a biography. The precision of the results found via these terms will be low. For example, the vocabulary term for the Dutch politician Hendrikus Colijn matches the query `war`, as this query occurs in his biography "...He served as minister of **war**...". However, most of the artworks where Hendrikus Colijn is the `dc:subject` do not depict war.

To increase precision we can consider using only the terms for which the query matches with a "label property", `rdfs:label` or one of its sub-properties. The *Label* column in Table 3 shows the number of vocabulary terms with a matching label. Only 31% of terms matching the query are via a label property. For some queries the number of different vocabulary terms is still large. However, for disambiguation not all the terms are required. The next column, labelled *in OP*, shows that only 14% of the terms with a matching label are directly related to an artwork. When we are only interested in artworks found via object properties, these terms are sufficient for the query disambiguation.

Where longer paths are used to find artworks, more vocabulary terms are related to the artworks and thus more interpretations of the query become available. The column in Table 3 labelled *in P5*, shows the number of vocabulary terms matching the query and related to artworks at path length 5. For the 25 queries 51% of the vocabulary terms with a matching label are related to an artwork at path length 5. In other words when longer paths are considered more interpretations of the query lead to artworks.

7.3. *Including equivalence*

The external vocabularies provide information that the experts deem useful in the search process. We investigate the effect of automatically including information from external sources in the search process. As a baseline we collected all artworks found via vocabulary terms with a matching label. The column labelled *OP* in Table 3 shows the results. The next column, labelled *+Eq*, shows the number of artworks found when equivalence relations are also included.

The effect of the equivalence relations is in general small. Only for the *concept* queries significantly more artworks are found. This increase is caused by the external sources that provide English labels that can be matched with the query, whereas the in-house vocabularies only provide Dutch or French labels. Including the equivalence alignment relations, thus, provides support for multilingual search.

7.4. *Integrating specialisation and generalisation*

The experts indicated that hierarchically related terms could be useful query suggestions. We investigate how specialisation, generalisation and siblings can be integrated into the search process. As a baseline we use the artworks found via object properties and equivalence alignments (shown in the column labelled *+Eq* in Table 3). The equivalence alignments are included to also make the hierarchical relations from the external sources available to find artworks. Compared to this baseline we investigate the additional artworks found by including specialisations, generalisations and siblings.

The column labelled *+Spec* in Table 3 shows for each query the total number of artworks found via one or more `skos:narrower` relations. For the `concept` and `location` queries we see several large increases compared to the number of artworks found via the equivalence path type alone. Automatically including these specialisations can, however, result in many irrelevant artworks for a number of queries. For example, the query `rome` matches with a vocabulary terms for the city in Italy, but also with several mythological events from IconClass. Including specialisations for all these terms will reduce precision, as typically only one or a few interpretations of the query are intended. In these case it is better to apply specialisation after the query is disambiguated.

The `location` and `concept` queries can also be generalised. However, these generalisations lead to overly generic concepts, such as the continent Europe. We do not further investigate generalisations of the queries here, but only note that for more specific queries, such as `peddler` in the first expert use case, generalisations were more useful. Instead we take a closer look at sibling terms, the combination of generalisation and specialisation. As shown in the *+Sib* column in Table 3, the inclusion of sibling terms has dramatically increased the number of results. For example, for the `location` queries artworks related to all other countries in Europe are found. Again for the specific concept `peddler`, located deep in the hierarchical structure, we observe that the sibling terms are closer related to the query. The use of sibling terms should, thus, be in control of the user.

7.5. *Integrating associations*

The experts indicated that associations could be useful query suggestions. We investigate the additional artworks found via association relations and the vocabulary terms by which these artworks are found. The column labelled *+Assoc* in Table 3 shows for each query the total number of artworks found via a path with one or two association relations. For all types of queries we see an increase in the number of artworks compared to the artworks found via object properties alone. For all queries, on average 11 times as many artworks are found. In particular, large increases are shown for `person` queries. These are predominantly found via the associative relations in ULAN. Large numbers of artworks are also found for the queries `rome`, `italy` and `paris`. The locations matching these queries are related to persons, for example by properties such as `birthPlace`, and these persons are themselves associated to other persons.

The column labelled *Term* in Table 3 shows the number of associated vocabulary terms per query. A large number of vocabulary terms are associated with a number of queries, with a maximum of 2,768 for the query `paris`. To get more insight into the specific types of associations we compute the different types of properties by which these terms are found, the column labelled *Rel.* In particular, for the queries with a large number of associated vocabulary terms we observe that these are found by a relatively small number of relation types. We also compute the different types of associated terms e.g. person, location, event, collection and concept, the column labelled *Type*. We observe that most queries are associated to more than one type of term, but on average the queries have 2 different types of associated terms. We conclude that the types of the terms and the different relations, provide some categorisation of the large number of associations.

8. Implications for design

Based on the findings from the experiments we discuss the requirements to effectively support domain experts in artwork search. We observed that the search process typically consists of multiple iterations:

- The user starts with a basic keyword search to get an idea of the artworks that are available in the collection,
- if insufficient or irrelevant results are found the user reformulates the query,
- if the result set is too large the user adds additional filters.

In this section we describe a number of implications on the search functionality to support basic search, query reformulation and faceted result filtering. For each of these we describe how the search algorithm should be configured and discuss the implications on the presentation of the search results and navigation paths. Where applicable we discuss related work.

8.1. *Basic search functionality*

If the goal is to support the user in finding artworks "directly" related to the query, only literal and object property paths should be searched, in combination with the equivalence relations to cater for name, spelling and language variants. To increase precision of the obtained results, specific properties can be excluded. First, the free text fields found via editorial notes and sub properties of `rdfs:comment` could be excluded, as these "meta" properties are unlikely to contain results that are relevant for most users. Second, only vocabulary terms with a matching literal value of a sub-property of `rdfs:label` could be included, as the vocabulary terms matching on other literal properties make the relation between query and result indirect. Additionally, assessment of the results is, in both cases, more difficult. The

user should, however, have the opportunity to disable these restrictions when high recall is important.

There are tasks where the result set should also contain artworks related to specialisations of the query. For example, for a query on works made in Germany, it makes sense to also include works made in a city within Germany, as we observed in the "Gregory" use case. The `skos:narrower` relation, however, is used for different types of specialisations in our data set and yields low precision for many queries. For example, the concept war specialises in WordNet to battle; battle specialises to soldier, but also to horse. It is unlikely that returning depictions of horses on a query for depictions of war is the intended behavior for most search tasks. The user should thus be able to control the in- or exclusion of specialisations.

In our data set, we observed that there are different interpretations for most queries and typically only one of these is intended by the user. The user experiment (Section 6) and the result analysis (Section 7), showed that automatically including the indirect relations for the other interpretations may dramatically reduce precision. We thus advise not to include specialisations before the query has been disambiguated. For the other types of relations we should be even more cautious. Hollink et al. showed that there are only a few combinations of hierarchical relations from WordNet that actually yield good precision and recall [7]. We thus advise to omit relations other than specialisation in the basic search functionality. We discuss below how to use them for query reformulation.

In the presentation of the search results, the relation between the results and the query should be communicated to the user, as the domain experts assess the results found via controlled vocabulary terms differently from results found via a plain text field. In [14], for example, the results are clustered based on the relation between results and query. As mentioned by Hearst, such clustering has the advantage that irrelevant groups of results can be quickly eliminated [3].

8.2. *Interactive query reformulation*

The large number and the diversity of the relations make it difficult to effectively include them automatically in basic search functionality. Koenemann and Belkin also concluded that interactive query expansion improves effectiveness and user satisfaction over automatic expansion [9].

As most search sessions require several queries before the desired results are obtained, effective support for interactive query reformulation would thus be a useful feature of a semantic search application. The relations between vocabulary terms are likely candidates for such functionality. The experts indicated that they want to explore multiple search strategies. We distinguish three such strategies based on expert users feedback in Section 6: disambiguation of the query with vocabulary terms, specialisation or generalisation of the query and recommendation of associated vocabulary terms.

Query disambiguation To disambiguate multiple interpretations of a keyword query, the vocabulary terms should be provided as suggestions. The suggestions should include at least all the vocabulary terms used in the basic search functionality, as for these it is known they are directly related to artworks. For further exploration other related vocabulary terms could be suggested to the user. Presenting them separately makes the user aware of the difference.

A selected vocabulary term provides the query for the basic search functionality. The URI of this term can be directly used to filter the results. This, however, will not find results by free text fields. It will require further research to discover if and how the labels of the vocabulary terms can also be used for disambiguation of the free text fields.

In the presentation of the suggested navigation paths, ranking could help the user choose the appropriate vocabulary terms. Meij et al. demonstrated the use of DBPedia to discover the concepts contained in text-based queries [11]. They show that the corresponding concepts can be effectively re-ranked by learning the most effective features. The literature also provides suggestions for grouping similar results. For example, the terms can be grouped by different types [2]. This requires vocabulary terms to have more specific types than `skos:Concept` alone. In previous work [6] we concluded that in term search additional information is often required to disambiguate terms that have similar labels, for example, by showing the profession and birth date of people.

Query specialisation or generalisation The hierarchically related vocabulary terms could be presented to the user as specialisation or generalisation suggestions, including at least the narrower terms and

a broader term. The equivalence alignments could also be included, as different thesauri provide their own hierarchical structures.

The hierarchical relations of similar types of thesauri may need to be integrated into a single structure. For geographical thesauri this is often straightforward. In previous work [6] we demonstrated that the integration of TGN and the in-house location thesaurus of the Rijksmuseum created a useful extension for both sources. TGN providing the top level of the hierarchy, with the Rijksmuseum thesaurus contain specific details, such as street names [6]. The hierarchical structures of different types of thesauri are, however, often better presented as alternatives, providing different perspectives on the topic (e.g. art specific in AAT, religious, biblical and mythological in IconClass and lexical in WordNet).

For the interface to support the navigation, we advise a design that provides interactive expansion, as this gives the user control over the path length and the direction, preventing the explosion of related terms. In addition, Joho et al. also showed that the presentation of a hierarchical structure can significantly reduce the time users need for query refinement compared to suggestions presented in a list [8].

Recommending associated terms After disambiguation of the query, vocabulary terms that are associated to the query, or otherwise related, could be made available as query suggestions. The suggestion algorithm could even include combinations of all path types. For example, in the first use case the concept of trade was associated to the concept salesman, which was a generalisation of the query `peddler`.

In the presentation of suggested navigation paths it is important that the relation to the query is communicated. The experts indicated that the type of relation helps to determine if a suggestion should be explored. Magennis and Rijsbergen showed that it is often difficult for end users to determine which suggestions are more useful [10] and Ruthven concluded that the identification of relationships among related information can help the user make such a decision [13].

As the number of associated vocabulary terms become large, additional organisation needs to be provided. In previous work [4] we demonstrated the use of sub-property relations to hierarchically organise the properties in the interface. To be helpful to the user, however, the sub-property hierarchy needs to be well designed.

8.3. *Result filtering*

In addition to reformulation of the query, the user also needs to be able to filter the result set on other dimensions. For example, P3 wanted to search for artworks related to the query `gregory`, but only when they were made in Germany at a particular time. In addition, the user should be able to combine query reformulation with result filtering. For example, P2 wanted to generalise the `Fuji` in combination with a constraint on the location, e.g. querying for volcanos (a generalisation of Fuji) but constrained to results made or depicting scenes in Japan.

A popular method to interactively add these types of constraints is faceted browsing [20]. In previous work [4] we showed that this functionality can be effectively applied to RDF data. The precise integration of facet browsing with basic search functionality and query reformulation requires further research.

The main implications for design are that basic search functionality should only include literal and object properties, combined with equivalence. Hierarchical and associative relations are best used after query disambiguation. In some cases, specialisation of the query can be directly included after disambiguation, whereas the inclusion of generalisation and associations always needs to be under control of the user. The interactive search functionality for query reformulation needs to be combined with methods for result filtering.

9. Conclusion

We conclude that there is no one-size-fits-all solution for semantic search. Instead effective end-user support requires the user to explore different search strategies, such as direct search on the artworks metadata, query disambiguation, query specialisation and generalisation, suggestions of associated terms and result filtering. The search functionality to support these strategies require different configurations of the path types. A graph search algorithm should be able to support these configurations. We analysed the potential paths and their configurations for a specific cultural heritage data set. In addition, the different types of search results

and the large number of candidates for query reformulation require different types of organisation and presentation methods.

We consider this study a first exploration to better understand how to search in semantically-rich and heterogeneous linked data. On the one hand, the qualitative analysis confirms the results already known in Information Retrieval, such as the need for interactive solutions to word sense disambiguation, query expansion and result filtering. On the other hand, the study explored new aspects introduced by linked data. First, the presence of multiple (partially) aligned vocabularies introduces both new opportunities as well as new problems. Second, the annotations from controlled vocabularies and the relations between the terms from these vocabularies provide semantically-rich background knowledge. The explicit types of the terms and relations within this background knowledge can be exploited in the search functionality and result presentation.

We are currently working on implementations of specific types of search functionality. In future work we plan to perform quantitative evaluations of these individual solutions by conducting user experiments with a larger number of participants performing a specific search task.

Acknowledgements

We would like to thank Geertje Jacobs, the three participants of the experiment and all other people at Rijksmuseum for their feedback, time and enthusiasm. This research was supported by the MultimediaN project funded through the BSIK programme of the Dutch Government and the EuropeanaConnect project funded through the eContent*plus* programme of the European Commission.

References

[1] A. Amin, L. Hardman, J. van Ossenbruggen, A. van Nispen, Understanding cultural heritage experts' information seeking tasks, in: JCDL '08: Proceedings of the Joint Conference of Digital Library, ACM Press, New York, NY, USA, 2008.

[2] A. Amin, M. Hildebrand, J. van Ossenbruggen, V. Evers, L. Hardman, Organizing suggestions in autocompletion interfaces, in: 31st European Conference on Information Retrieval, Toulouse, France, 2009, to be published, based on techreport: http://ftp.cwi.nl/CWIreports/INS/INS-E0901.pdf.

[3] M. A. Hearst, Clustering versus faceted categories for information exploration, Commun. ACM 49 (4) (2006) 59–61.

[4] M. Hildebrand, J. van Ossenbruggen, L. Hardman, /facet: A Browser for Heterogeneous Semantic Web Repositories, in: The Semantic Web - ISWC 2006, 2006.
URL http://dx.doi.org/10.1007/11926078_20

[5] M. Hildebrand, J. van Ossenbruggen, L. Hardman, An analysis of search-based user interaction on the Semantic Web, Tech. Rep. INS-E0706, CWI (July 2007).
URL http://www.cwi.nl/ftp/CWIreports/INS/INS-E0706.pdf

[6] M. Hildebrand, J. R. van Ossenbruggen, L. Hardman, G. Jacobs, Supporting Subject Matter Annotation Using Heterogeneous Thesauri, A User Study In Web Data Reuse, International Journal of Human-Computer Studies 67 (10) (2009) 888 – 903.
URL http://dx.doi.org/10.1016/j.ijhcs.2009.07.008

[7] L. Hollink, G. Schreiber, B. Wielinga, Patterns of semantic relations to improve image content search, Journal of Web Semantics 5 (3) (2007) 195–203.

[8] H. Joho, C. Coverson, M. Sanderson, M. Beaulieu, Hierarchical presentation of expansion terms, in: SAC '02: Proceedings of the 2002 ACM symposium on Applied computing, ACM, New York, NY, USA, 2002.

[9] J. Koenemann, N. J. Belkin, A case for interaction: a study of interactive information retrieval behavior and effectiveness, in: CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, New York, NY, USA, 1996.

[10] M. Magennis, C. J. van Rijsbergen, The potential and actual effectiveness of interactive query expansion, in: SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, 1997.

[11] E. Meij, M. Bron, L. Hollink, B. Huurnink, M. de Rijke, Learning semantic query suggestions, in: 8th International Semantic Web Conference (ISWC 2009), 2009.

[12] M. F. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.

[13] I. Ruthven, Re-examining the potential effectiveness of interactive query expansion, in: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, New York, NY, USA, 2003.

[14] G. Schreiber, A. Amin, L. Aroyo, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, B. Omelayenko, J. van Ossenbruggen, A. Tordai, J. Wielemaker, B. J. Wielinga, Semantic annotation and search of cultural-heritage collections: The multimedian e-culture demonstrator, J. Web Sem. 6 (4) (2008) 243–249.
URL http://dx.doi.org/10.1016/j.websem.2008.08.001

[15] A. Tordai, B. Omelayenko, G. Schreiber, Thesaurus and metadata alignment for a semantic e-culture application, in: K-CAP '07: Proceedings of the 4th international conference on Knowledge capture, ACM, New York, NY, USA, 2007.

[16] A. Tordai, J. R. van Ossenbruggen, G. Schreiber, Combining Vocabulary Alignment Techniques, in: Proceedings of The Fifth International Conference on Knowledge Capture, IAAA, 2009.

[17] J. Trant, Understanding searches of a contemporary art museum catalogue: A preliminary study, http://tinyurl.com/yl5lttk (2006).

[18] M. van Assem, M. R. Menken, G. Schreiber, J. Wielemaker, B. Wielinga, A Method for Converting Thesauri to RDF/OWL, in: Proceedings of the Third International Semantic Web Conference (ISWC'04), No. 3298 in Lecture Notes in Computer Science, Springer, Hiroshima, Japan, 2004.
URL `http://www.cs.vu.nl/~mark/papers/Assem04a.pdf`

[19] J. Wielemaker, M. Hildebrand, J. van Ossenbruggen, G. Schreiber, Thesaurus-based search in large heterogeneous collections, in: A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, K. Thirunarayan (eds.), International Semantic Web Conference, vol. 5318 of Lecture Notes in Computer Science, Springer, Berlin Heidelberg, 2008.
URL
`http://dx.doi.org/10.1007/978-3-540-88564-1_44`

[20] K.-P. Yee, K. Swearingen, K. Li, M. Hearst, Faceted Metadata for Image Search and Browsing, in: CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press, Ft. Lauderdale, Florida, USA, 2003.

	#artworks			#terms				#artworks		#artworks			#terms		
	LP	*∩OP*	*∩P5*	*All*	*Label*	*in OP*	*in P5*	*OP*	*+Eq*	*+Spec*	*+Sib*	*+Assoc*	*Terms*	*Rel*	*Type*
book	114	0	106	2,247	598	135	306	1,810	4,194	6,336	73,771	7,014	95	32	2
war	17	0	3	2,080	291	21	139	885	1,123	4,414	73,660	6,504	51	15	2
portugal	57	18	33	155	56	8	33	56	59	73	20,519	335	28	12	3
spain	5	0	3	572	20	0	8	0	71	171	21,930	4,762	51	22	3
rome	1,012	439	859	1,454	695	62	401	795	822	4,129	4,716	28,195	1,752	61	4
italy	326	25	262	1,142	390	55	265	902	1,059	2,280	19,322	26,784	617	51	4
greece	2	0	1	263	11	1	8	10	20	79	18,790	26	5	6	2
paris	646	241	403	1,283	561	100	275	2,089	2,207	2,238	5,448	28,557	2,768	60	4
poland	4	0	0	182	12	2	5	2	7	10	19,026	410	16	13	3
romania	0	0	0	60	7	0	6	0	0	1	19,024	0	0	0	0
prado	243	180	183	48	46	1	5	254	254	254	262	4,064	8	8	2
louvre	435	117	315	104	85	34	48	254	254	254	427	3,390	10	8	2
mona lisa	2	0	1	2	1	0	1	0	0	0	0	14	2	4	2
renaissance	316	23	291	278	77	3	31	143	143	549	8,364	2,304	27	20	3
klimt	18	0	0	10	9	0	8	0	0	0	345	728	13	8	2
van gogh	331	286	329	63	48	4	12	374	374	374	715	2,384	26	14	2
vermeer	28	12	17	108	100	12	21	172	172	172	172	4,923	28	18	3
rubens	572	420	489	203	160	8	68	2,046	2,046	2,046	2,046	9,339	105	31	3
goya	40	7	8	45	41	5	12	139	139	139	484	455	13	11	2
shakespeare	12	1	2	109	13	0	4	0	0	0	16	0	2	2	1
munch	2	0	0	33	32	0	9	0	0	0	345	7,211	12	8	2
da vinci	17	9	10	35	22	9	14	18	18	18	25	956	22	15	2
monet	5	3	4	33	29	3	10	6	6	6	351	1,049	13	7	2
renoir	2	1	1	13	9	3	6	12	12	12	12	2,614	15	9	2
hitler	5	3	3	44	10	1	3	12	12	12	12	16	2	2	2
	4,211	1,785	3,323	10,566	3,323	467	1,698	9,979	12,992	23,567	289,782	142,034	5,681	437	59
		42%	**79%**		31%	**14%**	51%		1.3x	1.8x	**22x**	11x			

Table 3
Results from the analysis of the 25 search log queries. The first three columns show the number of artworks found via literal properties and the subset that are found via alternative paths. The next four columns show the total number of vocabulary terms matching the query and the subset that are found via a label. The columns labelled *in OP* and *in P5* show the subset directly or indirectly related to artworks. The columns labelled *OP* and *+Eq* show the number of artworks found via an object property, with or without equivalence relations included. The columns labelled *+Spec*, *+Sib* and *+Ass* show the number of artworks found via paths including equivalence and specialisation, siblings or 2 association relations. The final three columns show the vocabulary terms found by 2 association relations, the number of different relations by which they are found and their different types.

Centrum Wiskunde & Informatica (CWI) is the national research institute for mathematics and computer science in the Netherlands. The institute's strategy is to concentrate research on four broad, societally relevant themes: earth and life sciences, the data explosion, societal logistics and software as service.

Centrum Wiskunde & Informatica (CWI) is het nationale onderzoeksinstituut op het gebied van wiskunde en informatica. De strategie van het instituut concentreert zich op vier maatschappelijk relevante onderzoeksthema's: aard- en levenswetenschappen, de data-explosie, maatschappelijke logistiek en software als service.

Bezoekadres:
Science Park 123
Amsterdam

Postadres:
Postbus 94079, 1090 GB Amsterdam
Telefoon 020 592 93 33
Fax 020 592 41 99
info@cwi.nl
www.cwi.nl

CWI

Centrum Wiskunde & Informatica