

Downloaded from UvA-DARE, the institutional repository of the University of Amsterdam (UvA)
<http://hdl.handle.net/11245/2.74702>

File ID	uvapub:74702
Filename	Chapter 8: Conclusions
Version	unknown

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type	PhD thesis
Title	End-user support for access to heterogeneous linked data
Author(s)	M. Hildebrand
Faculty	FNWI: Informatics Institute (II)
Year	2010

FULL BIBLIOGRAPHIC DETAILS:

<http://hdl.handle.net/11245/1.318913>

Copyright

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content licence (like Creative Commons).

Chapter 8

Conclusions

The goal of this thesis is to explore end-user access to semantically rich and heterogeneous linked data. To achieve this goal we investigated:

RQ 1. *How semantically-rich graph structures can be used in search functionality to support the user in finding objects in heterogeneous linked data?*

and,

RQ 2. *How semantically-rich graph structures can be used in the presentation of the results found in heterogeneous linked data?*

The research questions were first addressed in a literature study (Chapter 2). In a survey of search and browsing applications for Semantic Web data different types of search functionality and presentation methods are analysed. The survey provides an overview of the semantic relations, algorithms and interface designs used in existing applications. Without a common evaluation method¹ it, however, remains unclear how well these technologies improve support for end-users. Given a specific domain and search task it is, therefore, difficult to determine which semantic relations, algorithms and interface designs should be used to provide effective access.

This thesis studied the search functionality and result presentation in three case studies using the linked cultural heritage data: annotation, faceted browsing and semantic search. Chapter 3 studied the required support for artwork annotation in a user study with professional cataloguers from the Rijksmuseum Amsterdam. The final prototype was successfully used for professional annotation in an experimental setting and was judged positively by the cataloguers. The Rijksmuseum Amsterdam is currently investigating integration of such a system in their workflow.

¹At the time of writing the 3rd Semantic Search workshop has announced the organisation of a small-scale evaluation campaign for the first time. <http://km.aifb.uni-karlsruhe.de/ws/semsearch10/>

In Chapter 4, a generic solution for faceted browsing of heterogeneous linked data was explored by the implementation of a prototype. The prototype provides a completely data-driven solution that can be applied to any small to medium sized RDFS repository. In addition to the cultural heritage domain it has been used to give access to news images and articles (Troncy 2008), music songs and artists (Raimond and Sandler 2008) and historical events (Shaw et al. 2009).

In Chapter 5, support for semantic search was investigated in two experiments. First, the usefulness of different paths of relations in linked data is investigated in a user study with a small number of domain experts. A number of path types are identified and their usefulness is qualitatively evaluated. In the second experiment, the implementation of the path types in a semantic search application is investigated with the most frequently used queries from a search log. Implications for the design of an interactive search application for cultural heritage are derived.

In Chapter 6, the architectural support for the required search algorithms is investigated and their implementation as web services is discussed. The web service of the MultimediaN E-Culture project is used by the CHIP project (Aroyo et al. 2007) to provide text-based search functionality within their recommendation system. In Chapter 7 the required interface designs, in the form of several JavaScript Widgets, are discussed. To tailor the behavior of these widgets to a specific domain and task a method was presented to support configuration by data mapping.

Below, we revisit the two research questions posed in Chapter 1. We then reflect upon our work and discuss future research.

8.1 The research questions revisited

We discuss the high level conclusions for the two research questions based on the findings of the literature study, the three cases studies and our experiences in providing architectural support.

8.1.1 Search functionality for semantically-rich linked data

Semantically rich graph structures can be used in search functionality to provide effective end-user support when the search algorithm is configured for the specific data and task. In addition, interactive support should be provided for different phases of the search process. We discuss the need for configuration of term search and graph search algorithms, and the support for different types of end-user interaction. Finally, we discuss how the semantic relations provided by RDF(S), OWL and SKOS are used in the explored solutions.

Configuring search algorithms The literature study indicates that generic text-based search functionality can be defined directly on top of the RDF data model (Chapter 2). In the user studies on annotation (Chapter 3) and semantic search (Chapter 5) we showed that this does not, however, provide sufficient support for end-user applications. To provide effective task-specific support search algorithms need to be configured in several dimensions.

- *Effective support for annotation requires different configurations of a term search algorithm for different search fields.* The interface described in the case study on annotation (Chapter 3) required search fields to describe the who, what, where and when depicted on artworks. The fields required different configurations of the search algorithm, for example, different filters had to be provided to constrain the candidate results.
- *Effective support for semantic search requires configurations of a graph search algorithm at different phases of the search process.* In the case study on semantic search (Chapter 5) we derived that the process of searching for artworks contains different phases. Different types of relations in the data are useful in these phases. For example, the initial search results need to include artworks related by literal and object properties as well as equivalence alignments. For query reformulation, however, different types of hierarchical and associative relations between vocabulary terms need to be included.

To support the required search behavior for annotation (Chapter 3) and semantic search (Chapter 5) we implemented configurable algorithms for term and graph search. The algorithms are made available in ClioPatria as parameterized web services (Chapter 6).

Supporting end-user interaction A search process often consists of multiple cycles in which the user tries different queries and explores different search strategies. Interactive solutions can effectively support the user in this process.

- *Faceted browsing provides a generic solution to interactively navigate linked data repositories.* In the literature it is shown that faceted browsing can effectively support the user with the exploration of annotated objects (Chapter 2). In the case study on faceted browsing (Chapter 4) we showed that this interface paradigm can be applied to linked data. In addition, the traditional functionality, which only allows direct constraints on the results, can be extended to support navigation of the graph structure to create indirect constraints.
- *Autocompletion provides effective end-user interaction when searching for specific vocabulary terms.* In the case study on annotation (Chapter 3) autocompletion was successfully used by professional cataloguers to find annotation terms from multiple vocabularies. Autocompletion enabled users

to quickly try multiple queries, as it allowed them to easily switch between scanning the list of results, reducing or extending the query or creating a new query.

- *Effective support for access to semantically-rich linked data requires different types of end-user interaction.* In the case study on annotation (Chapter 3) text-based search alone showed to be not always sufficient to find vocabulary terms. In some cases the cataloguers wanted to navigate the hierarchical structures in which the search results were contained. In the case study on semantic search (Chapter 5), we observed that domain experts require support for multiple search strategies. To support these strategies we identified different types of end-user interaction, for example, query disambiguation by selecting vocabulary terms and query reformulation by navigating the hierarchical structure.

Many reusable interface components that support user interaction are already available on the Web. To apply these components, such as JavaScript widgets, to heterogeneous linked data they have to be configured for the specific domain and task. Typically this requires a developer for the programming and a domain expert that knows how the domain specific data should be used in the component. In Chapter 7 we proposed a method to capture the functionality of interface components in a model. Once such a model is implemented there is no more need for a programmer and the component can be configured by mapping domain specific data to this model.

Using semantics for search functionality We explored the use of three² types of semantic relations available in the data: (i) thesaurus specific relations in the original vocabularies, (ii) alignments between concepts from different vocabularies, (iii) lightweight schema mappings and (iv) ontological descriptions of properties. We discuss how these semantic relations were used in the three case studies.

- *SKOS provides a useful abstraction for vocabularies on which specific functionality can be defined.* As all vocabularies in the data set were modelled or mapped to SKOS, we could use the SKOS relations to provide search functionality for all vocabularies. For example, in the annotation prototype (Chapter 3) the `skos:prefLabel` and `skos:altLabel` provided a generic means to define the values for string matching with the query and define the preferred label in case multiple matches are found. The `skos:broader` relation defines the hierarchical structure of a thesaurus, providing a generic means to support hierarchical navigation of different thesauri. This was used in the prototypes for annotation (Chapter 3), faceted browsing (Chapter 4) and we

²[http://en.wikipedia.org/wiki/The_Spanish_Inquisition_\(Monty_Python\)](http://en.wikipedia.org/wiki/The_Spanish_Inquisition_(Monty_Python))

proposed it as a method for query reformulation in an interactive semantic search application (Chapter 5).

- *Alignment relations allow the inclusion of data from external sources, increasing recall in text-based search.* In our data, equivalent terms from different vocabularies are aligned using the `skos:exactMatch` and `owl:sameAs` relations. In the search functionality these equivalence relations allow information from external sources to be included. For example, in the case study on annotation (Chapter 3), the alignment with external vocabularies increased the available information by which the terms from in-house thesauri could be found. They provided useful spelling variations, synonyms, nicknames and multiple languages. In the user study on semantic search (Chapter 5), the experts also indicated a need for the integration of information from external vocabularies. Here they provided additional literals as well as useful relations between terms for query reformulation.

In our work we only considered equivalence alignments. Other alignment relations, such as `skos:closeMatch` and `skos:broaderMatch`, also allow the inclusion of external information, but functionality should consider their associative and hierarchical nature.

- *Schema mappings enable integrated access to heterogeneous data while preserving the richness of the individual collections and vocabularies.* Instead of a single unifying data model, the different schemata of the collections and vocabularies are aligned by lightweight mappings with `rdfs:subPropertyOf` and `rdfs:subClassOf` relations. In the search functionality these mapping relations enable integrated access. For example, in the faceted browsing prototype (Chapter 4) the mappings from the collection specific properties to a common super property enabled integrated access. For example, a facet corresponding to the `dc:creator` property provides integrated access to the large number of specific creator properties from the different collections. At the same time the rich information was maintained by allowing the user to still select a facet corresponding to a collection specific property.
- *Meta properties of the relations in the data enable integrated search functionality over heterogeneous linked data.* The schemata in our data set contains “meta” properties, such as `owl:inverseOf` and `owl:symmetricProperty`. These are useful for data integration purposes. For example, the hierarchical relations are defined by some organisations using `skos:broader`, while others use `skos:narrower`. SKOS defines that these two are each other inverse, using the “meta” property `owl:inverseOf`. Applications only need to support the semantics of such “meta” properties to provide integrated search functionality.

8.1.2 Presenting the results found in heterogeneous linked data

Semantically rich graph structures can be used in the result presentation to provide effective end-user support when the presentation information and organisation method is configured for the specific data and task. In addition, appropriate abstractions in the data are required to make the large diversity in heterogeneous data manageable. We discuss the need for configuration of result organisation and presentation methods, and the need for common abstractions in the data. Finally, we discuss how the semantic relations can be used to support such configuration and abstraction.

Configuring presentation algorithms Linked data contains various characteristics and many specific RDF properties that can be used for the organisation and presentation of the search results. To provide task-specific support different methods are required for different types of terms.

- *Unambiguous presentation of vocabulary terms requires different types of additional information for different types of terms.* In the case study on annotation (Chapter 3) it became clear that a label alone was not always sufficient to disambiguate different vocabulary terms and additional information had to be presented. The required information varied for different types of terms. For example, profession, nationality and birth/death dates for people, and the country and place type for geographical locations.
- *Search results from different vocabularies require different organisation strategies.* In the case study on annotation (Chapter 3) it became clear that professional cataloguers prefer different sorting and grouping strategies for different annotation fields. For example, they preferred to sort people alphabetically, as this is a natural and transparent ordering for names. For ICONCLASS they preferred presentation of the concepts towards the top of the hierarchy first, as they are used to navigating from these to more specific concepts. Searching in multiple thesauri, e.g in WORDNET and ICONCLASS, the presentation of the results in separate groups for each vocabulary helps to distinguish between different types of term.

To support configurable result organisation and presentation the web service for term search had to be extended (Chapter 6). In addition, the organisation and presentation dimensions had to be included in the RDF model describing the auto-completion widget. This provides a solution to configure the selection, organisation and visualisation of the widget in a single configuration file (Chapter 7).

Providing appropriate abstractions Heterogeneous data contains a large number of different types of resources and relations. To avoid overwhelming the

user, this diversity should not always be presented. For a specific domain, appropriate abstractions, which match with the conceptual model of the user, should be identified. In our studies on the cultural heritage domain artefacts, persons, locations, events and domain specific concepts were reoccurring abstractions. These abstractions should be used in the presentation of the navigation paths and search results to provide a manageable and coherent view.

- *For a specific task a small number of abstract facets should be presented to the user.* For non-trivial data sets a generic solution for faceted browsing will result in a very large number of facets. The `rdfs:subPropertyOf` relations in the data provide a means to organise the facets and provide the user with a smaller number of abstract facets (Chapter 4). To support a specific task the existing properties in the data might, however, not match with the user's conceptualisation of the domain. In this case, appropriate end-user facets should be identified and manually configured using mappings to the underlying data.
- *Effective presentation of the search results requires the identification of common abstractions in the domain that cover the richness of the data.* In the user study on semantic search (Chapter 5), the domain experts indicated that their assessment of the search results depends on the relation to the query. Results found by a literal property need manual assessment, whereas the results found by a term from controlled vocabulary are trusted by the relation to this term. The individual collections contain a large number of different types of relations by which artworks can be related to the query, e.g. many specific types of creator relations. In the presentation of the initial search results this level of detail should be hidden from the user. Instead, a number of common high-level properties in the domain should be used, e.g. creator and subject, to provide a simple unified view on the results.

In the analysis of the graph search algorithm for the search log queries (Chapter 5), a large number of related vocabulary terms were found as potential candidates for query reformulation. The association relations by which these terms are found and the classes they belong to provide a means to organise the suggestions for query reformulation. Using the right abstractions for these properties and classes a large number of suggestions for query reformulation can be organised while providing an intuitive navigation structure.

Using semantics for organisation and presentation In the result presentation we explored the use of the same four types of semantic relations available in the data: (i) thesaurus specific relations in the original vocabularies, (ii) equivalence alignments between concepts from different vocabularies (iii) lightweight schema mappings and (iv) ontological descriptions of properties.

- *SKOS provides a useful abstraction for vocabularies on which specific presentation methods can be defined, but the use of domain specific vocabularies is still required.* In the case study on annotation (Chapter 3), the presentation properties of the results were partially available as the values of SKOS properties, e.g. the preferred and alternative labels, a description and the hierarchical structure. To this extent a generic presentation solution could be provided. For the persons and locations, however, additional properties were required, such as the profession, nationality and birth/death dates for people. As different vocabularies provided their own types of properties for these values, mappings were required between these properties from different vocabularies.
- *Equivalence alignments should be used to remove duplicate search results.* In the case study on annotation (Chapter 3), similar concepts from different vocabularies need to be presented as a single search result to prevent duplicate results in the interface. The study showed that for this search task a conservative alignment method is most suited. Incorrect alignments are harmful, because they remove possible candidates from the search results whereas a few duplicates in the interface are acceptable.
- *Hierarchical relations between different schema provide a useful dimension to provide different abstractions of the result presentation.* In the case study on faceted browsing (Chapter 4), it was shown that the `rdfs:subPropertyOf` relations can be used as a dimension where along different abstractions of the available facets can be presented. This enables the user to interactively choose the facets at the appropriate level of abstraction for the specific tasks. For example, the facets capturing the common abstractions in the domain provide a high level integrated view on all collections, while the existing properties from the individual collections provide a detailed view but only applies to a single collection.

In (Chapter 6) we showed how the paths in the graph, indicating the relation to the query, can be used to group similar types of results together. Using the collection and vocabulary specific properties and classes many queries result in many different types of paths. Abstracting the path using both the `rdfs:subPropertyOf` and the `rdfs:subClassOf` relations enables groups of results at different levels of abstraction.

- *Meta properties of the relations in the data enable a unified view over heterogeneous linked data.* The “meta” properties of the relations were used in a similar fashion as in the search functionality, but in this case to generate the presentation structures. For example, in the presentation of a single SKOS concept, the use of `owl:inverseOf` guarantees that the view on the concept

properties is independent of modelling decisions with respect to `skos:broader` versus `skos:narrower`.

8.2 Discussion and future research

We identify three limitations in research described in this thesis: (I) the studies are performed only in a single domain, (II) we focussed only on support for domain experts (III) the studies only provide qualitative evaluations. We discuss each limitation in turn. In addition, we discuss the required support for representing the semantics used in this research.

Application to other domains A limitation of the research described in this thesis is that all studies are performed in the cultural heritage domain. We can not make general claims about the applicability of our solutions to other domains. We expect, however, that our solutions can be used to support end-users in other domains where objects are described with terms from structured vocabularies. Evidence for this is provided by the application of the software described in this thesis in the news domain (Troncy 2008), the music domain (Raimond and Sandler 2008) and for historical events (Shaw et al. 2009). At the time of writing, the software is primarily used to showcase linked data in these other domains. To provide effective end-user support for annotation, faceted browsing and semantic search in these domains, further research is required. We expect that it will be sufficient to model the common abstractions in the domain and configure the search algorithm and presentation methods. Finding the appropriate abstractions and configurations for a particular domain will require experimentation with the end-users in this domain.

For some of the collections shared as open linked data on the Web, the objects are not described with terms from vocabularies. In general, our solutions can still be applied to access such collections. For example, the objects can be found by their direct metadata with text-based search or by faceted browsing. The specific search functionality and result presentation methods that are enabled by the semantic relations are, however, not available. By enriching the metadata of these collections additional functionality can be enabled. For example, a literal value describing the creation site of an object can be replaced with a reference to a term from a geographical vocabulary. This makes the semantics from the external source available when accessing the collection, e.g. enabling access through the geographical containment relations.

Support for domains with stronger ontological relations and/or rules was not within the scope of this thesis. Exploiting the added value of formal reasoning in end-user tasks such as annotation, search and faceted browsing requires further research. On the other hand, we expect that in domains with strong ontological

commitments there will also be different end-user tasks than the search oriented tasks considered in this thesis, requiring altogether different support.

Broader user population A limitation of the studies performed in this research is that they focused on support for expert users. We cannot make general claims about the applicability of our solutions to other types of users. We expect, however, that some of our solutions can be applied to support other user populations, but that the configuration of the search functionality and the result presentation requires adaptation to their specific needs.

In the case study on annotation (Chapter 3), we investigated support for domain experts when creating high quality descriptions. To support annotation by the general public, future research is needed to better understand issues around quality control, the type of annotations the general public would want to make and the effort they would put into selecting vocabulary terms. We found some indication that users that are less experienced in annotation cannot be expected to carefully select the most appropriate term when multiple options are available. Two of the study participants, who only occasionally provide artwork annotations, often selected the first likely candidate without considering other options. The end-user support should take this type of behavior into consideration.

In the case studies on search (Chapter 5) and browsing (Chapter 4) we explored data-driven approaches. This assumes that users are familiar with the representation of domain-specific knowledge, as this is directly exposed in the user interface. We cannot expect all types of users to be familiar with highly specific representations. We believe that the solutions we explored can deal with this aspect. We identified the need for the use of common abstractions in the user interface. We focused on this need to handle the heterogeneity of the data. This, however, also provides a solution to adapt the information that is presented to the user at an appropriate level. We experienced that these abstractions can themselves be modelled in linked data and thus automatically appear in the data driven solutions. In addition, they can be used in the configuration of the interface components, as explained in Chapter 7.

Towards quantitative evaluation In our studies we limited the evaluations to qualitative analysis. Without commonly agreed upon evaluation methods or benchmarks we can not make any claims about the comparison to other semantic search systems. Working towards commonly agreed upon evaluation methods is, however, not straightforward. The effectiveness of semantic search systems for a specific task is affected by the quality and the modelling of the data, the behavior of the algorithms and the design of the user interface. In addition, semantic systems provide support in different stages of the search process: query formulation, search algorithm and result presentation.

Based on the conclusions of this thesis we expect that comparative studies are

best performed using a specific type of functionality and considering a specific stage of the search process. We hope that the community works in this direction, for example in initiatives, such as the evaluation track of the Semantic Search 2010³ workshop.

RDFS plus? The semantics used in our data set correspond to a subset of what Allemang and Hendler (Allemang and Hendler 2008) defined as “RDFS *plus*”: a subset of RDFS/OWL that provides sufficient utility for data integration and is computationally practical to apply. Our experiences confirm that in a domain with collections of objects annotated with terms from structured vocabularies, RDFS *plus* combined with SKOS provides sufficient utility for integrated access in the tasks we studied. Furthermore, computation with these semantic relations could be performed on the fly the performance required for interactive applications. For our purposes the following was sufficient:

- **rdf:type** to provide typing information. For vocabulary terms it is useful to have more information than **skos:Concept** alone, such as **Person**. In the search functionality the types of objects and vocabulary terms help to restrict and in the presentation to explain the results.
- **rdfs:subClassOf** and **rdfs:subPropertyOf** to define lightweight mappings between different schemata. In the search functionality this enables integrated access. In the presentation it provides a unified view on the results and enables different levels of abstraction.
- **rdf:value** to represent default values in N-ary relations. In the search functionality it indicates that the value can be considered as a direct object property. In the presentation it indicates what should be shown to the user.
- **skos:broader** and **skos:narrower** to define hierarchical relations between vocabulary terms. When more detailed hierarchical relations are already available they should be mapped to the relations in SKOS. In search functionality this enables specialisation and generalisation. In the presentation it provides a means to organise the results in an hierarchical structure.
- **skos:related** to define associations between vocabulary terms. When more detailed association relations are already available they should be mapped to the relation in SKOS. In text-based search functionality this enables suggestions for query reformulation and in the formulation of structured queries to define indirect constraints.
- **skos:exactMatch** and **owl:sameAs** to define equivalence between resources. In search functionality this enables the inclusion of information from external sources. In the result presentation it allows the removal of duplicate terms.

³<http://km.aifb.uni-karlsruhe.de/ws/semsearch10/>

- `owl:inverseOf` and `owl:symmetricProperty` to define characteristics of relations. In search functionality and result presentation this enables the unification of modelling decisions.

Although we used transitive reasoning in the back-end, for example, in the facet browser (Chapter 4) to find all objects transitively related to a selected value, we did not consider it as a characteristic of the data. In our experiences the use of transitivity varied per search functionality and task, instead for the type of data. Furthermore, in the result presentation transitivity was used in the opposite direction to find all ancestors for a given set of resources.

8.3 Looking ahead

Linked data on the Web has become reality. In 2009 both the US⁴ and UK⁵ governments have started to share information on the Web and parts of this are being published according to the principles of linked open data. Publishing this data is, however, only the start. Designing the applications so end-users can benefit from all this data is still a grand challenge. The work described in this research has explored several aspects of this challenge. Above all, it has made clear that there is no one-size-fits-all solution to access linked data. Instead, specific tasks require different types of search functionality and result presentation methods and these need to be carefully configured to provide effective end-user support.

Although much work lies ahead, we believe that with the right web services in place, and configurable interface components at our disposal end-user access to heterogeneous linked data has become within reach. Where on Web 2.0 a programmer was required to create a Mash-up, integrated access on Web 3.0 will follow from the relations in the data itself.

⁴<http://www.data.gov/>

⁵<http://data.gov.uk/>