Downloaded from UvA-DARE, the institutional repository of the University of Amsterdam (UvA)
http://dare.uva.nl/document/168227

File ID	168227
Filename	Summary

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type	Dissertation
Title	End-user support for access to heterogeneous linked data
Author	M. Hildebrand
Faculty	Faculty of Science
Year	2010
Pages	vi, 172

FULL BIBLIOGRAPHIC DETAILS:
http://dare.uva.nl/record/334809

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.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)

Summary

On the Web, organisations are opening up their data and services for re-use. This openness allows information from different sources to be combined, enabling access in ways unforeseen by the original data providers. The Web applications that combine different data sources and services, also known as mash-ups, have become a common solution to support specific end-user tasks. Creating these web applications, however, requires a developer to provide precise instructions on how to use and integrate the data from different sources.

Semantic Web technologies promise to simplify reuse of data from different sources. The Semantic Web representation languages provide a standardised way to model and share knowledge on the Web. In addition, these languages allow aspects of the semantics available in the data to be made explicit in a machine-accessible way, enabling intelligent technologies to automatically infer how data should be used and integrated.

A growing number of machine-accessible data sets is now being published and linked together on the Web. The result is a large graph of linked data, describing a variety of different types of objects. In a survey study, we analyse applications that provide access to this heterogeneous linked data. From this study we conclude that Semantic Web applications support a wide variety of different types of tasks and provide access to different types of data sets. It, however, remains unclear how well these semantic technologies improve support for end-users, as commonly agreed upon evaluation methods are lacking. Given a specific domain and search task, it is thus difficult to determine how the semantics in the data can be used to benefit the end user.

Within this thesis, cultural heritage is chosen as an application domain to study end-user support for access to heterogeneous linked data. The semantically-rich Web of culture data created in the MultimediaN E-Culture project is used as a data set. The practical results of this research are made available as applications and web services in ClioPatria, the open source framework of the project.

This thesis takes a first step in formulating, for a specific domain and a number

of tasks, the requirements to support end-user access to semantically-rich and heterogeneous linked data. Different aspects of the search problem are explored in three case studies: (i) artwork annotation to study how users can effectively find terms in multiple vocabularies, (ii) faceted browsing on multiple collections of annotated artworks to study the formulation of structured queries, and (iii) semantic artwork search to study how the different relations in linked vocabularies can be used to find objects semantically related to a query.

Annotation In professional annotation the task of the user is to find vocabulary terms to describe an artwork. In existing collection management systems the annotation fields each provide access to the terms of a single vocabulary. The scope of the internal vocabulary used for this purpose are not always sufficient. In particular, when describing the content depicted on artworks, the cataloguers need to spend valuable time on extending the vocabularies with missing terms. External sources on the Web can be used to extend the scope of annotation terms, but this requires end-user support to find terms in multiple vocabularies. How can text-based search be used to find terms in multiple vocabularies with different schemata and characteristics?

In a study with professional cataloguers at the Print room of the Rijksmuseum Amsterdam, we investigate how multiple vocabularies can be used in an annotation tool. The initial requirements on the data, algorithms and interface design are formulated based on an analysis of their current practices. In a process of iterative prototyping, the requirements are refined and different solutions are explored. The solutions in the final prototype are qualitatively evaluated with feedback from the cataloguers. We found that end-users can be effectively supported with existing technologies, such as client side interface components for autocompletion, and server side algorithms for term search and result organisation and presentation. However, the user interface and algorithms need to be carefully configured for different annotation fields and vocabularies. We identify the required parameterization of the algorithms and user interface, and demonstrate successful configuration for a specific task and domain.

Faceted browsing Faceted browsing is a popular interface paradigm to support the formulation of structured queries to explore (artwork) collections. How can the faceted interface paradigm be used to support the formulation of structured queries for linked data? Traditional faceted browsers require a homogeneous collection with a single schema, whereas heterogeneous Semantic Web repositories may contain different types of objects each with their own facets. In addition, structured queries on linked data involve indirect constraints that cannot be formulated in traditional faceted browsers.

Based on a use case, we formulate the requirements for faceted browsing on heterogeneous Semantic Web repositories. Solutions for the required search func-

tionality and presentation methods are explored by the implementation of a prototype system. We found that the required functionality for any small to medium sized RDFS repository can be supported with a completely data-driven solution. The semantic relations in the data can improve end-user support by organising the large number of facets. The formulation of indirect constraints is supported in the prototype by allowing users to browse different types of objects in the same interface, and using the constraints on one type as constraints of a semantically related type. However, we also found that to support a specific task the existing properties in the data might not match with the user's conceptualisation of the domain. In this case, appropriate facets oriented to the end-user should be identified and manually configured.

Semantic search In many professional search tasks users want to find artworks that are somehow related to a topic. To satisfy their non-trivial information needs, domain experts often need to formulate multiple queries and manually combine and integrate the various search results into a single coherent set of answers. Our hypothesis is that the artwork annotations with terms from multiple structured and interlinked vocabularies, and the relations among these terms, can help end-users in the search process. To provide such semantic search we need to understand which and how the different types of relations in semantically-rich linked data can support end-users.

Support for semantic search is investigated in two experiments. First, the usefulness of different paths of relations 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. Based on the findings of these experiments the implications for the design of an interactive semantic search application in the cultural heritage domain are discussed. We observed that the process of searching for artworks contains different phases. Different types of relations in the data are useful in these phases and different types of end-user interaction are required to support the user. For example, the initial search results need to include artworks related by literal and object properties as well as equivalence alignments. Interaction should be provided for query disambiguation by selecting vocabulary terms and query reformulation using different types of hierarchical and associative relations between vocabulary terms. To support these strategies multiple interactive interface components are required using different configurations of a graph search algorithm.

From the case studies we derived several architectural requirements on the search functionality and result presentation methods. To support annotation and semantic search, configurable term and graph search are required, as well as configuration

of the result organisation and presentation methods. Interactive solutions are required to support the user in trying different queries and explore different search strategies. In the presentation of navigation paths and search results appropriate abstractions in the data are required to support the user with the large number of different types of termen and relations.

The functionality for text-based search is implemented with generic algorithms on the RDF data model, which can be configured in a number of dimensions. The algorithms are made available in ClioPatria as parameterized web services. The web services are extended with configurable algorithms for organisation and presentation of the search results. The required user interaction is supported by reusing and extending Web interface components. To support the configuration of these components and the web services used by them, we propose a method that captures the functionality in a model. Accordingly the configuration becomes a mapping task that can be performed by a domain expert, reducing the need for a programmer.

Within the explored solutions four types of semantic relations were used to improve end-user support. First, the thesaurus-specific relations in the original data, as defined in SKOS, provide a useful abstraction for vocabularies on which specific functionality and presentation methods can be defined. Second, the equivalence alignments between terms from different vocabularies allow the inclusion of data from external sources, increasing recall in text-based search. They also enable the removal of duplicate search results. Third, lightweight schema mappings enable integrated access to heterogeneous data, while preserving the richness of the individual collections and vocabularies. They also enable the use of different abstractions of the result presentation. Finally, ontological descriptions of properties enable integrated search functionality over heterogeneous data.

In this thesis we showed that a Web of culture data can be used to improve the support for domain experts with a number of tasks. To apply our results to other domains and user populations, the appropriate abstractions in this domain and the specific user needs have to be identified, and the configurations of the search functionality and presentation methods need to be investigated. Based on the conclusions from this thesis we expect that future comparison of different semantic search systems is best studied for a specific type of functionality and considering a specific stage of the search process.