Chapter 1

Introduction

1.1 The Semantic Web as a Web of data

The openness of the World Wide Web is changing the expectations on information access. Internet users expect information to be available at anytime, with the potential to contribute and link everything to everything else. As a consequence, organisations are starting to open up their previously isolated data silos and services, enabling users to combine data from multiple sources and apply the services of one organisation to the data of another.

These expectations on information access are underpinned by several technological developments. The move towards Web 2.0 has popularised sharing of content, using lightweight data structures such as RSS and JSON, sharing services over public APIs, and sharing interface components, such as JavaScript widgets. These front-end technologies have enabled the construction of mash-ups, web applications that combine data and services from different providers, such that information can be accessed in unforeseen ways.

At the back-end, Semantic Web technologies promise to simplify reuse of data from multiple sources. In a database or XML document, the intended meaning of the data, the semantics, is captured implicitly in the database or XML schema. A mash-up has to be aware of these schemata, which, typically, requires a developer to provide precise instructions on how to use and integrate the data. Semantic Web representation languages, such as RDF, OWL and SKOS, allow aspects of these semantics to be made explicit in a machine-accessible way, enabling intelligent technologies to infer how data should be used and integrated.

In the development of the Semantic Web, we can distinguish the "semantics" from the "web" aspect. In the early years, the focus was on the formal "semantics" suited for knowledge representation on the Web. As a result, the World Wide Web Consortium (W3C) standardised RDF (W3C 1999), and the more expressive OWL (W3C 2004), as languages for representing information on the Web. A set of statements, or triples, defined in these languages, which express different levels of

interoperable semantics, is called an RDF graph. The community provides off-the-shelf "triple stores" to store such RDF graphs and provide basic reasoning facilities over them.

In more recent years, the "web" aspect has gained momentum. Inspired by the Linking Open Data Project[1], a growing number of data collections is being published online and linked together as graph structures distributed over the Web. The result is heterogeneous linked data that to date, for example, contains the structured data from Wikipedia, geographical locations from Geonames, books from Amazon, publications from DBLP, subject headings from the Library of Congress and music related information from MusicBrainz. Furthermore, for some content providers, such as the BBC, publishing linked data has become part of their daily routine (Kobilarov et al. 2009). Although the Web of Data contains some formal ontologies, the bulk of data is formed by domain specific relations and controlled vocabularies. The Simple Knowledge Organisation System (SKOS) (W3C 2005) can be used to publish these vocabularies in an interoperable manner.

The amount of linked data has now reached a state that we can start asking how end users will access this information. Web 2.0 front-end technologies have already resulted in intuitive interfaces for end users to access information from a single data source or, when a mash-up has been created, to a combination of data sources and services. A general problem is how these front-end technologies can be applied to heterogeneous linked data, or the other way around, how the interoperability at the back-end can be exploited by interfaces at the front-end. The central problem is: *how can we support end users with access to heterogeneous linked data.*

An important consideration in information access is the trade-off between task-specific or generic support and homogeneous or heterogeneous data. Task-specific support, typically, requires some control over the data, which limits information access to data with a specific schema. On the other hand, a generic approach, such as a direct visualisation of the data structure, can give access to any collection of heterogeneous data, but provides limited support to the user. In this thesis, we investigate which models and tools can provide task-specific support to access heterogeneous linked data. In particular, we investigate how the semantics in linked data can be used for this purpose.

1.2 Project context: A Web of culture data

Within this thesis we use cultural heritage as an application domain. This domain is well suited for our research. First, the data in this domain is heterogeneous, as it describes many different types of objects and the descriptions vary per institution. Second, the domain contains semantically-rich background knowledge, as it has a

[1] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

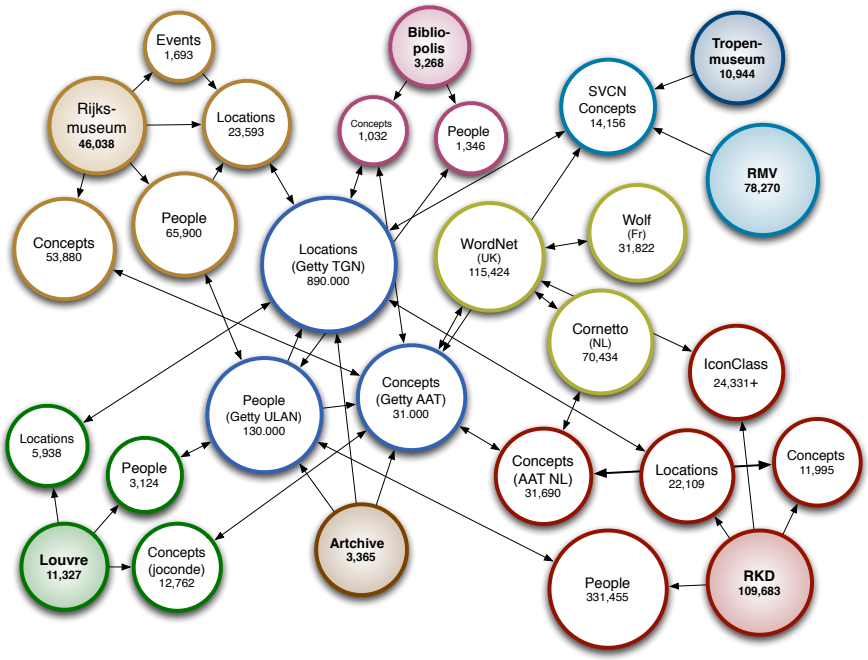

Figure 1.1: Datacloud of the linked cultural heritage data created in the MultimediaN E-Culture project.

long tradition of maintaining rich vocabularies and carefully describing objects with terms from these structured vocabularies. Third, there are users, from cultural heritage experts to interested museum visitors, with information needs that require data from different sources.

This research is performed in the context of the MultimediaN E-Culture project (Schreiber et al. 2008). This project has converted collections, their corresponding thesauri and additional background knowledge to RDF, OWL and SKOS, and captured the semantic relationships among them. Figure 1.1 shows the data cloud from the resulting web of culture data. Typical for this culture web is the large number of controlled vocabularies. The circles in the figure with a coloured fill represent the *collections* of works described with annotations from many different vocabularies. In addition, the project has included different external vocabularies and aligned them with other sources to create a highly interconnected web of data.

The project also created a flexible software architecture, where a triple store is tightly integrated with reasoning and Web server facilities (Wielemaker 2009a). This architecture is used as the platform for experimentation.

1.3 Research questions

The general problem investigated in this thesis is:

> *How can we support end-users with access to heterogeneous and semantically-rich linked data?*

An important characteristic of *linked data* is its representation in a graph structure, where the nodes and edges (or resources and relations) are explicitly typed. In *heterogeneous* linked data this graph structure contains different types of resources and relations. On the Web, new data — containing different types of resources and relations — can be added at any time. We, therefore, assume that there is limited control over the the available types in the data. We focus our research on access to objects that are described with terms from multiple, structured and interlinked vocabularies. We call these heterogeneous and interlinked vocabularies *semantically rich* background knowledge. Our hypothesis is that the structures in this background knowledge can be used to support end-user access to heterogeneous data.

We consider information *access* as a process where the user formulates an information need, in the form of a query, and a computer system responds by selecting and presenting a number of objects (TRECVID organizers 2007). This requires an interface to submit a query and an algorithm to select objects. We refer to this combination as the search functionality. It remains unclear how search functionality can exploit semantically-rich graph structures. The first research question, therefore, is:

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

Information access also requires an algorithm to organise the selected results and an interface to present these to the user. We refer to this combination as the presentation method. The second research question focusses on the presentation of results found in linked data:

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

The effectiveness of the applications to access heterogeneous linked data is affected by the quality and the modelling of the data, the behaviour of the algorithms and the design of the user interface. These three aspects influence each other

and are difficult to isolate in a realistic experimental setting. The methods to evaluate access to semantically-rich and heterogeneous linked data are, therefore, an intrinsic part of this research.

1.4 Approach

The two research questions mentioned above are first addressed in a literature study. In a survey of search and browsing applications for Semantic Web data, different types of search functionality and presentation methods are analysed. While the applications provide a large variety of different solutions, semantic search and browsing is still an open area. Given a specific domain and search task it is difficult to determine how the semantics in the data should be used to benefit the end user.

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 linked data. As already motivated in the project context (Section 1.2), cultural heritage is chosen as an application domain. To minimise the effect of the data quality and the modelling decisions we focus on support for expert users, who are already familiar with interaction in a knowledge-rich domain.

The tasks are chosen to cover different dimensions of the search problem encountered in the survey. In the search functionality (RQ 1), we distinguish controlled input for the formulation of structured queries and uncontrolled text-based search to formulate more open-ended requests. In the presentation (RQ 2), we distinguish the search results and the navigation paths used for further interaction. From a data perspective we distinguish artwork collections (foreground) and vocabularies (background). We investigate different combinations of the dimensions in three case studies: annotation, faceted browsing and semantic search.

Annotation: text-based search for vocabularies The first case study investigates the research questions in the context of artwork annotation. In professional annotation the task of the user is to find vocabulary concepts to describe an artwork. The study addresses the first research question by exploring the application of text-based search functionality to multiple vocabularies. It addresses the second research question by exploring the visualisation and organisation of the search results (vocabulary concepts).

The starting point is an existing collection management system where each annotation field provides access to only a single vocabulary. In a user study with professional cataloguers we explore how experts can be supported in effectively finding concepts from multiple vocabularies. The initial requirements are formulated based on an analysis of the current situation. 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 end-users.

Faceted browsing: structured query formulation for collections The second case study investigates the research questions in the context of faceted browsing, a popular interface paradigm to explore (artwork) collections. It addresses the first research question by exploring an extension of traditional faceted browsing to support the formulation of structured queries for linked data. It addresses the second research question by exploring methods to organise the large number of navigation paths.

The starting point is traditional faceted browsing for a homogeneous collection with a single schema. Based on a use case, the requirements for faceted browsing on heterogeneous Semantic Web repositories are formulated. Solutions for the required search functionality and presentation methods are explored by the implementation of a prototype system.

Semantic search: text-based search for collections The third case study investigates the research questions in the context of artwork search. In many professional search tasks users want to find artworks that are somehow related to a topic. The study addresses the first research question by investigating how different types of relations in linked data can be used in text-based search functionality. We speculate upon the second research question by deriving implications for the presentation of search results and navigation paths in an interactive search application.

The starting point is the search engines currently used by domain experts, where results are found by a syntactic relation to the query. In a user study with cultural heritage experts we explore how to support the user in finding artworks that are semantically related to the query. The initial requirements for semantic search are formulated based on interviews with a number of domain experts, collecting realistic use cases and feedback on the use of different types of relations. The performance of an initial graph search algorithm is qualitatively evaluated by the analysis of the results found for a number of search log queries.

In addition to these three case studies, the architectural support for the proposed solutions are an intrinsic part of this thesis. The algorithms to support the required server-side search functionality and presentation methods are implemented as web services. To support re-use of our solutions for the development of web applications on other data sets or other domains a solution is explored to support configuration of the individual components.

The studies in this thesis are all carried out in a single domain, namely cultural heritage. Some of our findings, however, are also relevant to other domains with collections of annotated objects and controlled vocabularies. We make the implemented solutions available as open source software and provide support for other researchers to apply them to their own data.

The explored solutions focus on support for expert users and do not directly translate to applications for novices. At this explorative stage of the research field, lacking methods for meaningful quantitative analysis, we make the assumption that domain experts provide the most valuable qualitative feedback.

1.5 Contributions

The work presented in this thesis contributes to the development of interactive applications for end user access to heterogeneous linked data. The theoretical contribution of this thesis is an analysis of the problems related to selecting and presenting results from heterogeneous linked data. We provide:

- An analysis of the search functionality (RQ 1) and result presentation techniques (RQ 2) in state of the art Semantic Web applications. In addition we report on the methods used to evaluate these technologies.

- A statement of requirements on the search functionality (RQ 1) and the result presentation techniques (RQ 2) to apply keyword search and faceted browsing to heterogeneous linked data. In three case studies different aspects of the search problem are explored: finding vocabulary concepts to describe artworks, faceted browsing to formulate structured queries and finding artworks that are semantically related to a query.

The practical results are made available within the open source ClioPatria architecture.[2] This research has contributed to ClioPatria in two significant ways:

- Co-development of the ClioPatria architecture. In particular, the design and implementation of server-side algorithms and their APIs, for parameterized access over HTTP, and the implementation of cross-browser client-side interface widgets. The paper describing the ClioPatria architecture, written together with Jan Wielemaker, received an honourable mention at the In-Use track of the International Semantic Web Conference 2008.

- Implementation of configurable applications within ClioPatria for vocabulary-based annotation, faceted browsing and semantic search. Early versions of the applications were part of the E-Culture demonstrator that was awarded first prize at the Semantic Web challenge 2006[3]. A later version is used as a demonstrator for the Europeana thought lab[4], and as demonstrators in the news (Troncy 2008) and music (Raimond and Sandler 2008) domains and to browse events (Shaw et al. 2009).

[2] http://e-culture.multimedian.nl/software/ClioPatria.shtml
[3] http://challenge.semanticweb.org/
[4] http://www.europeana.eu/portal/thought-lab.html

1.6 Structure of the thesis

The high level structure of this thesis consists of a discussion of the related work (Chapter 2), three case studies that each address the two research questions (Chapters 3, 4 and 5) and a discussion of the architectural support that is required for the proposed solutions (Chapters 6 and 7). Finally, Chapter 8 provides the conclusions.

In Chapter 2 we present a survey of semantic search and browsing applications and analyse how results are found and how results are presented.

In Chapter 3 we study term search in multiple semantically rich vocabularies to support professional cataloguers with subject matter annotation.

In Chapter 4 we explore faceted browsing as a generic solution for structured query formulation on heterogeneous Semantic Web repositories.

In Chapter 5 we study the use of semantically-rich vocabularies to support text-based search in artwork collections.

In Chapter 6 we describe the ClioPatria architecture, including the server-side algorithms for term and graph search and the APIs that make them accessible over HTTP.

In Chapter 7 we describe the models of client-side interface widgets for term search and faceted browsing and a method to configure these widgets with domain specific information.

In Chapter 8 we provide our conclusions and discuss our work in a broader context.

1.7 Publications

The research was proposed at the Doctoral Consortium of the International Semantic Web Conference 2008, where it was awarded Best Paper Doctoral Consortium:

- Michiel Hildebrand. Interactive Exploration of Heterogeneous Cultural Heritage Collections. In *Proceedings of the 7th International Semantic Web Conference*, pages 914–919, Karslruhe, Germany, 2008.

Publications on which the chapters of this thesis are based:

- Chapter 2 contains material that will appear as: Michiel Hildebrand, Jacco van Ossenbruggen and Lynda Hardman. The role of explicit semantics in search and browsing. Chapter in *Multimedia Semantics: Metadata, Analysis and Interaction*, Raphal Troncy, Benoit Huet and Simon Schenk. Wiley, 2010.

- Chapter 3 was published as: Michiel Hildebrand, Jacco van Ossenbruggen, Lynda Hardman and Geertje Jacobs. Supporting subject matter annotation

using heterogeneous thesauri, a user study in Web data reuse. *International Journal of Human Computer Studies*, pages 887–902, Volume 67, Issue 10, October 2009.

- Chapter 4 was published as: Michiel Hildebrand, Jacco van Ossenbruggen and Lynda Hardman. /facet: A Browser for heterogeneous Semantic Web repositories. In *Proceedings of the 5th International Semantic Web Conference*, pages 272–285, Athens, USA, 2006.

- Chapter 5 is submitted as a journal article: Michiel Hildebrand, Jacco van Ossenbruggen, Lynda Hardman, Jan Wielemaker and Guus Schreiber. Searching in semantically-rich linked data: a case study in cultural heritage.

- The material in Chapter 6 is the result of joint work with Jan Wielemaker, Jacco van Ossenbruggen and Guus Schreiber and published as: Jan Wielemaker, Michiel Hildebrand, Jacco van Ossenbruggen and Guus Schreiber. Thesaurus-based search in large heterogeneous collections[5]. In *Proceedings of the 7th International Semantic Web Conference*, pages 695–708, Karslruhe, Germany, 2008. Honourable mention at the In-Use track.

- Chapter 7 was published as: Michiel Hildebrand and Jacco van Ossenbruggen. Configuring Semantic Web interfaces by data mapping. In *Workshop for Visual Interfaces to the Social and the Semantic Web*, Sanibel Island, USA, 2009.

Other publications resulting from the research described in this thesis:

- Jan Wielemaker, Michiel Hildebrand and Jacco van Ossenbruggen. Using Prolog as the fundament for applications on the Semantic Web. In *Proceedings of the ICLP'07 Workshop on Applications of Logic Programming to the Web, Semantic Web and Semantic Web Services*, Porto, Portugal, 2007.

- Jacco van Ossenbruggen, Alia Amin and Michiel Hildebrand. Why evaluating Semantic Web applications is difficult. *Semantic Web User Interaction Workshop*, Florence, Italy, 2008.

- Guus Schreiber, Alia Amin, Lora Aroyo, Mark van Assem, Viktor de Boer, Lynda Hardman, Michiel Hildebrand, Borys Omelayenko, Jacco van Ossenbruggen, Anna Tordai, Jan Wielemaker and Bob J. Wielinga. Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. *Journal of Web Semantics* 6 (4), pages 243–249, 2008. Winner of the Semantic Web Challenge of 2006.

[5] This publication also included in the PhD thesis of Jan Wielemaker, Logic programming for knowledge-intensive interactive applications (Wielemaker 2009a).

- Alia Amin, Michiel Hildebrand, Jacco van Ossenbruggen, Vanessa Evers and Lynda Hardman. Organizing suggestions in autocompletion interfaces. In *31st European Conference on Information Retrieval*, Toulouse, France, 2009.
- Lynda Hardman, Jacco van Ossenbruggen, Raphal Troncy, Alia Amin and Michiel Hildebrand. Interactive information access on the Web of Data. In *Proceedings of the WebSci'09: Society On-Line*, Athens, Greece, 2009.
- Alia Amin, Michiel Hildebrand, Jacco van Ossenbruggen, Lynda Hardman. Designing a thesaurus-based comparison search interface for linked cultural heritage data. To appear in *Proceedings of the International Conference on Intelligent User Interfaces 2010*, Shanghai, China, 2010.