Interactive Information Access on the Web of Data

Lynda Hardman, Jacco van Ossenbruggen, Raphaël Troncy,
Alia Amin and Michiel Hildebrand

Interactive Information Access, CWI Amsterdam, The Netherlands
firstname.lastname@cwi.nl

Abstract. The Web of data enables fragments of information to be identified, described and connected together in a rich information environment. Users requiring information are faced with the problem of finding out what information is available, and obtaining sufficient fragments to successfully carry out their task. Systems supporting these tasks can use the fragments, descriptions of them and relationships among them, to improve both the selection and presentation of the information. Questions to be answered are: which information needs can be better supported, and how can the Web of data help.

While the construction of the "linked data cloud" is necessary to even start thinking about providing this type of support for users, our claim is that we first need to establish the user's information needs before establishing the potential roles the linked data can play in information selection and presentation. In this paper, we discuss potential uses of linked data to support users' information needs, give examples of using linked data to support user information seeking tasks and highlight future research directions.

1 Introduction

The web of data is the growing collection of data sets which are made available and linked together on the Web, providing a rich collection of vocabularies and data sources that users can interact with in an attempt to satisfy their information needs.

User have information needs, such as fact-finding or information gathering, which can be satisfied using various information sources and search mechanisms. We use the phrase *interactive information access* to convey the notion of the processes a user goes through to query information sources, investigate the returned results and potentially adjust the query, until either the information need is satisfied, or the user gives up. Key research challenges include:

- *i)* designing and evaluating effective user interfaces for tasks that go beyond the simple fact-finding and question answering tasks that dominate the current state of the art and
- *ii)* taking advantage of the inherent structures, patterns and semantics of the data while still supporting heterogeneous data collections.

In this paper we first discuss the information needs we found when talking to expert users of rich information sources. We then look at a number of prototypes we have created and reflect on the role linked data played in supporting users. We draw conclusions about where linked data can play a role in supporting interactive information access.

2 Identifying professional users' information needs

Before investigating how to incorporate linked data into user applications, it is crucial to know for which tasks linked data may be beneficial. While much is known from the literature about different types of information needs, one goal was to identify those that can usefully be supported by linked data resources. We carried out a qualitative study with information specialists to improve our understanding of the types of information needs that these users have [1]. Within the context of a nationally funded project we had access to cultural heritage professionals who make use of many different information sources for their daily work. 17 cultural heritage professionals were interviewed about the information sources they used, and the tasks they carried out. A total of 110 information seeking tasks were identified and classified into different categories, such as fact-finding and information gathering. Information gathering encompasses a number of sub-tasks, such as comparison, relationship search and topic search, where users carry out several searches to fulfill a higher level goal, such as writing a report, preparing an exhibition, or collecting information to make a decision.

Among the insights gathered are that fact-finding tasks, such as *"To which tribe/culture does this object belong?"*, are only a small percentage (10%) of the information needs that experts have. The majority of their needs (63%) can be classified as "information gathering". One type of information gathering activity is *comparison search,* where the differences or similarities between objects or sets of objects are compared, e.g., *"What objects from the Middle-East do other museums in the Netherlands have? Is there any tribe or region not represented in our collection or in the collection of other museums? If there is, we need to find out exactly what kind of object we should get."* Another example is *topic search*, where multiple related searches are carried out, e.g. when looking for information about a specific Jewish ceremonial coat *" Where and when was this coat made? Was there any restoration done to the coat? What is the purpose of the coat? What does it symbolize? Is there any meaning behind the embroidery? Where was it used? Who used it? Was it ever used in an important historical event?".*

Within many of the information gathering tasks identified, it is apparent that the pieces of information being gathered are linked to each other in some way, e.g., which tribes and regions are associated with the Middle-East. In other cases, such as looking for information related to the ceremonial coat, it is less clear beforehand what the connections are, but as different individual searches are carried out, other related information can be searched for.

It is these types of tasks which make use of explicit underlying relations among pieces of information from multiple sources that we hope to be able to provide support for using linked data resources. For example, for topic search, whenever experts search for information centered around a particular topic, they need information related to a single term as well as suggestions for related terms, e.g. nearby geographical locations or related cultures. Current challenges are how to deal with the potentially large number of results and how to differentiate the interesting from the large number of trivial relations, since these notions are subjective and context dependent.

3 Potential Support using Linked Data

Within existing national and international projects, we have constructed prototypes in the cultural heritage and news domains. While assembling the vocabularies and visual content has been a vital part of enabling the research, the working relations with the domain experts have been indispensable. These have enabled us to understand the problems of information seeking and how these can be supported through flexible access to rich linked data repositories.

We present here three different systems that were created to investigate linked data support in different domains and for different users. For each subsection we briefly describe the system, explain the anticipated user's information need, the role of the linked data within the support provided and our conclusions.

3.1 Exploring cultural heritage repositories

In parallel with the user study described in the previous section, we have contributed to the creation of an exploratory environment for cultural heritage assets. Many of the goals behind its creation were related to the underlying knowledge and reasoning infrastructure [4, 6]. We also, however, used the environment to explore how the relations in linked data can be used to improve the presentation of results at the user interface. No specific user profile was identified, but rather an exploratory approach was taken as to how the underlying linked data could improve search results and their presentation. In particular, more results can be retrieved for a single term query because of the underlying relations linking terms through thesauri incorporated within the infrastructure. To compensate for the potentially larger number of results presented, we looked for ways to group these to allow the user to understand the breadth of available results, e.g. that *"picasso"* is a type of marble from which an artwork can be made, as well as the name of an artist. In addition, artworks may depict the artist, or may be painted by him. Both such distinctions were used in grouping sets of results together, giving the user a broad overview while giving access to many results in limited screen space. A problem with increasing the numbers of terms found is that some relationships derived from the data and underlying thesauri, such as *"artworks in an art style related to the art style used by Picasso"* may not be meaningful to the user. Currently an empirical approach is used to determine a balance between including potentially interesting results while keeping less intuitive relations to a minimum.

3.2 Contextualized exploration of multimedia news content

Building on the infrastructure created for the cultural heritage domain in the E-Culture project, we explored the creation of a news-based application. Materials from video (INA) and news archives (AFP) were integrated with diverse information sources, such as `DBpedia`, `geonames` and `imdb`, together with a basic classification of the main subject of the news expressed in `newscodes`. The environment enables contextualized exploration of multimedia news content [5]. The users who inspired the direction of the integration are journalists at AFP, who are often faced with queries that return thousands of closely related results, but no satisfactory means for exploring these and finding the information sought. This environment allows journalists to find connections among people, places and events, e.g., that Ryan Babel, the Dutch football player, is related to Amsterdam because he is born in the city and has played for the AJAX club, although he now plays for Liverpool.

A further refinement to the interface was made allowing not only groupings of results based on the linked data properties but also on real-time image analysis of the result set, allowing, e.g. photos of Zinedine Zidane to be grouped into those with different visual characteristics, e.g., predominantly green for on the football field, and predominantly grey when in a suit while receiving an award.

3.3 Developing support for identifying annotation terms

In more recent work we have aimed at providing support for a specific task [2]. Cultural heritage professionals enter information about prints into a catalogue, for which they have 15 minutes to investigate the print, study the background information and enter information about the main theme of the print. They use a number of different thesauri for selecting terms to describe the object. As the thesauri developed within individual museum are limited in coverage and only provide an "art historic" perspective, external thesauri have to be integrated into the system. In close collaboration with experts, we developed a prototype system that allows these professional users to search for appropriate terms in multiple heterogeneous thesauri. As part of their task, if they discover that a required term is missing, then a new term has to be added. They thus need to search for existing terms as well as make exhaustive searches that non-existing terms do indeed need to be added. One role of linked data here was to connect terms occurring in different thesauri, saving time through using a single search. In addition, the interface presenting potential terms used different organization mechanisms, depending on the type of the annotation. For example, for people the terms were ordered alphabetically, locations were shown within a geographical hierarchy and the terms used to describe what was depicted on the image were ordered by the thesaurus the term comes from. Also, confirming that a term was absent from different thesauri required a decision based on a number of individual searches establishing the absence of the term, or spelling variations of it.

4 Conclusions

We have identified information tasks that can be supported using linked data by making non-obvious connections among related pieces of information explicit, such as exploratory tasks or topic search. When providing such "query expansion", the number of results potentially increases, and the relations available can be used to group results to provide overviews of the result diversity. How exactly results should be grouped depends on the domain, the user and their task.

While the above studies illustrate individual cases where linked data plays a role in supporting users accessing information, one of our main observations during the creation of these systems is that the data modeling, the quality of the data set and the underlying search and inference model play a crucial role and have a direct influence on the end-user interfaces and their evaluation [3]. This makes any study of a specific part of the interaction chain difficult. It is thus of vital importance to select sufficiently focused tasks, for which realistic support can be provided on useful data sets.

Two studies for which we are currently preparing will investigate the design of support for a comparison search task, using feedback from cultural heritage professionals who encounter such tasks frequently, and topic search support, for cultural heritage professionals who receive queries from historical researchers who have some notion of what they need to find, but are unable to translate this into terms directly accessible in the museum catalogues. We see these as specific instances of the types of higher level support that could be provided in other domains with similar information seeking needs where linked data resources are available.

Acknowledgements

We thank the participants from the Dutch cultural heritage institutions who made these studies possible. This research was supported by the MultimediaN E-Culture project, funded through the BSIK programme of the Dutch Government and by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content, K-Space.

References

1. A. Amin, J. van Ossenbruggen, L. Hardman, and A. van Nispen. Understanding Cultural Heritage Experts Information Seeking Needs. In *8th ACM/IEEE Joint Conference on Digital Libraries (JCDL'08)*, Pittsburgh, PA, USA, 2008.
2. Michiel Hildebrand, Jacco van Ossenbruggen, Lynda Hardman, and Geertje Jacobs. Supporting subject matter annotation using heterogeneous thesauri, a user study in web data reuse. Technical Report INS-E0902, CWI, February 2009.
3. J. van Ossenbruggen, A. Amin, and M. Hildebrand. Why Evaluating Semantic Web Applications Is Difficult. In *Semantic Web User Interface (CHI'08 Workshop)*, pages 1–4, Florence, Italy, 2008.
4. 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. *J. Web Sem.*, 6(4):243–249, 2008.
5. R. Troncy. Bringing the IPTC News Architecture into the Semantic Web. In *7th International Semantic Web Conference (ISWC'08)*, pages 483–498, Karlsruhe, Germany, 2008.
6. J. Wielemaker, M. Hildebrand, J. van Ossenbruggen, and G. Schreiber. Thesaurus-based search in large heterogeneous collections. In *7th International Semantic Web Conference (ISWC'08)*, pages 695–708, Karlsruhe, Germany, 2008.