

Assisting User Browsing over Linked Data: Requirements Elicitation with a User Study

Dhavalkumar Thakker, Vania Dimitrova, Lydia Lau, Fan Yang-Turner,
and Dimoklis Despotakis

University of Leeds, Leeds LS2 9JT, UK
{D.Thakker, V.G.Dimitrova, L.M.S.Lau,
F.Yang-Turner, scdd}@leeds.ac.uk

Abstract. There are growing arguments that linked data technologies can be utilised to enable user-oriented exploratory search systems for the future Internet. Recently, search over linked data has been studied in different domains and contexts. However, there is still limited insight into how conventional semantic browsers over linked data can be extended to empower exploratory search, which is open-ended, multi-faceted and iterative in nature. Empirical user studies in representative domains can identify problems and elicit requirements for innovative functionality to assist user exploration. This paper presents such an approach – a user study with a uni-focal semantic data browser over several datasets linked via domain ontologies is used to inform what intelligent features are needed in order to assist exploratory search through linked data. We report main problems experienced by users while conducting exploratory search tasks, based on which requirements for algorithmic support to address the observed issues are elicited. A semantic signposting approach for extending a semantic data browser is proposed as a way to address the derived requirements.

Keywords: Linked Data, Exploratory Search, Requirements Elicitation.

1 Introduction

Linked Data technologies have received wider acceptance, both in industry and academia. One of the major factors for this success has been the availability of large amount of semantic data in various formats and domains. In parallel with engineering solutions for seamless generation of semantic data, efforts have been made to facilitate user interaction with such data. There are growing arguments that Linked Data technologies can be utilised to enable user-oriented exploratory search systems for the future Internet [1]. In contrast to regular search, exploratory search is open-ended, multi-faceted, and iterative in nature, and is commonly used in scientific discovery, learning, and sense making [2].

There are a wide range of tools available for offering exploratory search using semantic web technologies (state of the art in [3] and [4]). However, exploratory search over linked data is still insufficiently studied. As pointed in a recent keynote focusing on interaction with Linked Data[5], although the technological platforms for

exploring linked data are growing, enabling citizen users to explore inter-connectable links associated with structured data is still a key challenge. This calls for an urgent attention by researchers and technology developers to identify major issues with user exploration of linked data, derive requirements for new methods, and engineer solutions to implement these methods utilising semantic technologies and tools. Experimental studies with existing systems in domains well-presented in linked data can be used to elicit requirements for engineering new methods for user exploration.

The work presented in this paper follows the above arguments, and specifically focuses on *providing intelligent functionality embedded in a data browser to assist users in their exploratory search tasks over linked data*. This is part of an ongoing research examining intelligent interfaces for interactive sensemaking over linked data, conducted in the framework of the EU project Dicode (<http://www.dicode-project.eu>). We have built a fairly traditional semantic data browser – Pinta - which provides a base line for identifying key issues users face with conventional uni-focal exploratory search interfaces over linked data. An instantiation of Pinta in the Music domain is used in an experimental study with users to elicit requirements for intelligent assistance based on observations of challenges users face while interacting with MusicPinta; and suggesting a way to address them by adding signposting features.

Section 2 will present the base line system. A user study with MusicPinta is presented in Section 3. Section 4 reports observations of main interaction issues faced by users, based on which requirements for adding intelligent functionality are elicited. Following the requirements, a signposting approach for adding intelligent assistance is proposed. The paper concludes by pointing at future work.

2 Baseline System for Browsing through Semantic Data

In this section, we present a traditional semantic data browser called Pinta, which provides a uni-focal interface for browsing through several linked semantic datasets.

2.1 Pinta: A Generic Uni-focal Semantic Browser Shell

The main goal of Pinta is to enable users to easily tap into resources built from the Web and, in particular, exploring the use of the Linked Data paradigm. Figure 1 depicts three-layer architecture for Pinta which comprises: (i) Data Layer, including knowledge sources and content, (ii) Processing Layer, including modules for semantic augmentation and query, and (iii) Presentation Layer for content browsing. The **Data Layer** contains domain specific *ontological knowledge sources* and *content* assembled from the Web (Linked Data and other domain specific sources). The knowledge sources consist of graphs of ontological concepts relevant to the domain of interest. They provide the foundation for semantic augmentation of the content in the Processing Layer, and the structure for semantic trajectories for browsing in the Presentation Layer. The **Processing Layer** has two main services: (i) semantic augmentation of the assembled content designed using GATE and (ii) semantic queries to retrieve content for the Presentation Layer. The **Presentation Layer** provides a front-end for the output of semantic queries from the Processing Layer.

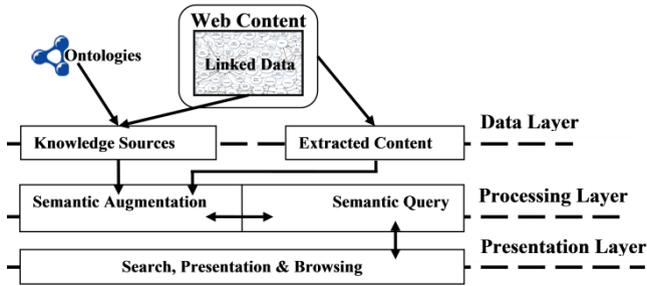


Fig. 1. Architecture of the generic uni-focal semantic data browser Pinta

The interface layout includes three main facets and a description (at the top) extracted from the knowledge datasets for the *focus entity* (being currently explored): (i) Facet 1 includes facts about the focus entity; (ii) Facet 2 includes terms related to the focus entity; and (iii) Facet 3 shows related content.

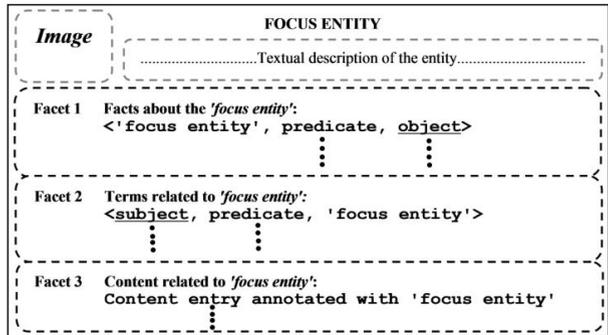


Fig. 2. A faceted-layout template for presenting a focus (currently explored) entity in Pinta

2.2 MusicPinta: An Instantiation of Pinta in the Music Domain

The Music domain has been selected for an instantiation of Pinta and has been used as a testbed to observe exploratory search and derive requirements for intelligent support. The Web of data is rich in music-related datasets and content. As of 2011, there were at least 13 datasets identified, with a diverse range of concepts and ambiguous entities covering instruments, performances/events, artists, and music genres. The data sets used for MusicPinta comprise the following resources.

DBpedia: for musical instruments and artists. This dataset is extracted from dbpedia SPARQL endpoint using CONSTRUCT queries. **DBTune**(<http://dbtune.org/>): for music-related structured data made available by the DBTune.org in linked data fashion. Among the datasets on DBTune.org we utilise: (i) **Jamendo** - a large repository of Creative Commons licensed music; (ii) **Megatune** - an independent music label; and (iii) **MusicBrainz** - a community-maintained open source encyclopaedia of music information. **Amazon reviews** for musical instruments shown in Pinta. All datasets, except the reviews, were available as RDF datasets and the **Music ontology** (<http://musicontology.com/>) was used as schema to interlink them. The Amazon

reviews were converted in RDF using Pinta's semantic augmentation of textual content in the Processing Layer. MusicPinta is available at: <http://imash.leeds.ac.uk/services/pinta/musicpinta/>.

3 User Study and Interaction with MusicPinta

To observe user exploratory search behaviour and elicit requirements for adding intelligent functionality in uni-focal semantic browsers over linked data, we conducted an exploratory study with MusicPinta.

The study involved 12 participants recruited on voluntary basis. All participants had IT background, good experience in web search. Each participant attended an individual session, conducted and observed by an experimenter for an hour: (i) using a Pre-study questionnaire [5 min] for collecting information about the user and test his/her domain awareness; (ii) introducing MusicPinta [10 min]; (iii) conducting Task 1 [15 min] aiming at identifying distinctive characteristics of the musical instrument "bouzouki"; (iv) conducting Task 2 [15 min] for identifying usage and features of the musical instrument "electric guitar"; (v) a Post-study questionnaire [10 min] for testing again the participant's domain awareness and gathering usability feedback; and, (vi) briefly interviewing [5 min] for eliciting the overall impression of using MusicPinta for exploratory search. After each task, the users were asked to fill-out a short questionnaire to assess cognitive load using the NASA-TLX questionnaire[6].

Table 1. User tasks in the experimental study

Task 1: Characteristics of a musical instrument [bouzouki]	Task 2: Usage and features of a musical instrument [electrical guitar]
<p>The music shop is extending its collection of instruments with international musical instruments. You work in an advertising agency which has been asked to prepare an advertisement script for some of the new instruments that will appear in the shop. A key part of the preparation of the advertisement script is the research of the product.</p> <p>You have been asked to conduct a research of one of the new instruments, called bouzouki, using the information available in MusicPinta. You have to identify:</p> <ul style="list-style-type: none"> • the <u>main characteristics</u> of bouzouki; • up to <u>five similar instruments</u> to bouzouki; • features that make bouzouki <u>distinctive</u> from the similar ones you have chosen. <p>Go to 'Semantic Search' in MusicPinta and type bouzouki. Browse the content and follow links. Complete the provided form.</p>	<p>The music shop wants to increase the sales of its traditional musical instruments, such as electrical guitars. It intends to do this by adding links to creative commons album recordings with electric guitars, together with some interesting information about these albums to inspire customers to play/buy electric guitars or other musical instruments.</p> <p>Furthermore, when displaying its electric guitar items, the shop wants to highlight key features people look for when purchasing electric guitars.</p> <p>You are asked to conduct the research to address the above requirements by using information provided in MusicPinta. You have to review the information about electric guitar and identify:</p> <ul style="list-style-type: none"> • three <u>interesting album recordings</u> that include electric guitars and specify what is interesting; • <u>key features</u> that people look for when purchasing an electric guitar. <p>Go to 'Semantic Search' in MusicPinta and type electric guitar. Browse the content and follow links. Complete the provided form.</p>

The study required participants to complete two tasks related to exploring musical instruments and was positioned within an advertising scenario for a fictitious UK music shop (see Table 1). In both tasks, the participants were given an entry point for browsing and asked to fill in their answers in a provided template. The tasks exhibit the characteristics of exploratory search tasks summarised in [7]: the main goal is learning and/or investigation of a musical instrument; there is a low level of specificity about the information needed and how to find it; search is open ended, requires finding several items and involves a degree of uncertainty; tasks are ‘not too easy’ and include multiple facets.

4 Requirements for Assisting User Browsing over Linked Data

Two musical instrument experts (one for Bouzouki, one for Electric guitar) have marked the outcome of participants for the two tasks. The marking is to measure how successful the participants have been in completing the tasks using MusicPinta. Participant achieved 70% average score for the task 1 and 48% for the task 2. The detailed analysis on the task performance and learning outcome in the user study is covered in our recent publication[8]. In this paper, we have only focused on the observations related to the task outcome and browsing behaviour that allow us to elicit requirements for supporting exploratory environments.

Observation 1: Abstraction Conundrum. While browsing specific instruments (e.g. Bouzouki), performances and performers, two participants clicked on abstract concepts, such as **instrument**, **performance** and **performer**, from the Music Ontology. In both cases, the participants were looking for concrete information (e.g. participant-12 clicked on **instrument** in task 1 when seeking for more detail about a musical instrument, while participant-05 clicked on **performer** and **performance** in task 2 when seeking more detail about an album). The aggregated datasets in MusicPinta have large number of instances for the abstract concepts (which is typical of linked datasets), which led to confusion as the result was a long list of performers, performance and instruments, and the participants quickly pressed back button.

Requirement 1: Offering Semantic Links at an Appropriate Level of Abstraction. The above observation motivates consideration on identifying what can be algorithmically offered as the right level of abstraction on various browsing junctures. This is important when the abstract concepts have large amount of concrete instantiations.

Observation 2: Exploring Entities/Content with Insufficient Information. Another interesting case is the high number of ‘empty clicks’ - the user clicks on a link and is taken to a page with no information, sees that this link is not helpful and quickly returns to the previous page. In task 1, such clicks concerned similar instruments, e.g. there was no information about **bajitar**, **xalam**, **rebab**. In task 2 such clicks concerned **performances** (music albums) and happened quite often. ‘Empty clicks’ leading to pages with no information was seen as one of the main reasons for user’s frustration. At the same time, may be due to their experience of links that lead

to dead ends, some links were perceived as empty without exploring them further and the users missed to click on important for the tasks information. With linked datasets, it is typical to find entities that do not have much explanation or links to other entities.

Similar issues were observed with content (Amazon Reviews in our user study). Users clicked to view some of the Amazon reviews to find out more information about an instrument and its review. However some of the reviews were deemed to have insufficient information to be useful. This observation is in line with relevant research conducted which concludes that not all reviews are equally helpful (for example, [9]).

Requirement 2: Reduce Entity Link Options. Avoid showing entity links that do not lead to any new information. Reduce number of entity links shown to the user based on their browsing value; allowing reduction of clutter and confusion. The challenge here is to define what ‘browsing value’ is and how to calculate it for an entity with respect to other entities from the same entity page.

Requirement 3: Reduce Content Link Options. Avoid showing content links that do not lead to any new information. Reduce number of content links shown to the user based on their helpfulness/usefulness.

Observation 3. Varied Selection Strategies while Facing Too Many Choices. Both tasks (deliberately) put the users in situations where they had too many choices. This means that the users had a large number of links to review while on a focus entity page. For example, the *bouzouki* page included 12 different links in the facts facet and 51 links in the terms. This is a typical situation with the datasets from linked data. For example, for the DBpedia dataset, which has 3.5M entities and 627M triples, on average, a user might have to review 192 links while exploring a focus entity.

We observed users following different strategies when presented with too many choices in the browsing interface: (i) clicking on the nearest classification link from the ‘facts facet’ (e.g. *plucked string instruments* or *string instruments*) to see general characteristics in the case of *bouzouki* as part of task 1. However, users rarely clicked on links from the facts facet as part of the task 2, as the task did not require this; (ii) clicking on instruments mentioned in the ‘related terms facet’– (e.g. *lute* and *mandolin* mainly in task2; (iii) clicking on something (e.g. ‘an instrument’) that ‘sounds familiar’ (e.g. *sitar*, *banjo*, *pipa* in task 1); (iv) click on something (e.g. ‘an instrument or an album’) that sounds interesting or unusual (e.g. *oud*, *xalam* in task 1 and noticing a women artist or something interesting in the album name in task 2); (v) clicking on something that looks important (e.g. an artist has several albums in task 2); and, (vi) clicking randomly (after exhausting other strategies).

This observation is in line with the latest research in search engines and HCI; increasing numbers of options can make designers and users feel less confident when deciding and less happy with the results[10]. To support varied level of selection strategies, following requirement is derived.

Requirement 4. Take into Account Context to Cater for Interests and Importance. People when faced with many choices do select what they find useful/familiar/interesting/unusual/important. Hence, there is a merit in making it easier for users to decide/spot easily these values. The challenge here is how to measure and decide these values from the available options for a specific user or holistically.

5 Semantic Signposting to Assist Exploratory Search

The identified requirements from the study indicated the need for further algorithmic support to realise the exploratory search potential of semantic data browsers. One possible approach to address these requirements is **semantic signposting**.

In uni-focal exploration, a user focuses on one entity at a time represented on a page. This entity page contains links to various descriptions, image and links to other entities. Such entity page can be treated as a juncture in journey where the explorer has to make few choices. Some of the requirements elicited can be addressed by providing *signposts* guiding the explorer in making a choice about paths she can take.

Only showing 'important' links which are a subset of all possible links for the user to review as part of next path he/she can take. Let us call '*candidate entities*' *all the* links possible to navigate from a focus entity page. Importance of each candidate entity can be computed based on density parameters such as – number of further entities available from a candidate entity (i.e. number of directly connected entities to the candidate entity), number of potentially reachable entities from this entity (i.e. number of entities connected to candidate entity via directly connected entities) and type and weight of the connection (e.g. semantic relationship between candidate entities and it's directly connected entities). The judgement of creating subset of links can be implemented using density metrics for the semantic graphs[11], where density function shall allow comparing how dense/informative each of the path is originating from a candidate entity. The subset of links to be shown to the user in this case will be based on the density value of each link (i.e., candidate entity).

For more creative tasks (such as task 2 in our study) which require browsing through a large amount of content, the study appeared to provide indication that it will *not be very beneficial to limit the user entity choices*, as this can affect the free content exploration. Instead, signposting can include some *indicators about the 'importance' or 'value' of a content item*, e.g. if there is any description (or any multimedia content), its source of the content (e.g. DBpedia), if further semantic links are available in the content (e.g. albums with several musical instruments) to facilitate user choices. There can be some ordering based on the value. Again the judgement of importance can be implemented using density metrics for the semantic graphs.

Adaptive Signposts. One of the other parameters to consider while judging importance of links is consideration of user's prior knowledge, e.g. does user already know about a particular entity or class of entity? Such consideration in creation of signposts (i.e. reducing number of links shown to the user) can allow users to decide what is useful/familiar/interesting/unusual/important(R4). A possible way to 'sense' previous knowledge is to analyse the user clicks on the low classification level links – clicking on an instrument can indicate some familiarity with its most specific classification category (e.g. in the study, users familiar with Russian musical instruments clicked on Balalaika and users familiar with Chinese musical instruments clicked on Pipa). The necessary techniques to address such requirements can benefit from the research in the user modelling, adaptation and personalisation. Such solution can allow creating signposts that include familiar and new knowledge together. Putting familiar and new items together in such a way can deepen the learning by association[12].

6 Conclusions

We have presented a study with a traditional uni-focal semantic data browser to observe browsing behaviour of users while interacting with several linked semantic datasets aiming at deriving requirements to inject intelligent features. We have found several intricate challenges that are applicable to typical interaction over linked semantic datasets. For example, disparity of the options available while browsing from an entity. In some cases large number of links available from an entity, hence posing too many options for the user to choose from and in other cases no links or information available making users frustrated. We have also observed and reported varied levels of selection strategies when a user is faced with too many options.

Acknowledgements. The research reported in this paper is supported by the EU 7th Framework Programme under grant agreement ICT 257184 (DICODE project).

References

1. Waitelonis, J., Knuth, M., Wolf, L., Hercher, J., Sack, H.: The Path is the Destination—Enabling a New Search Paradigm with Linked Data. In: *Linked Data in the Future Internet at the Future Internet Assembly* (2010)
2. Marchionini, G.: Exploratory search: from finding to understanding. *Communications of the ACM* 49, 41–46 (2006)
3. Ferré, S., Hermann, A.: Semantic Search: Reconciling Expressive Querying and Exploratory Search. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) *ISWC 2011, Part I. LNCS*, vol. 7031, pp. 177–192. Springer, Heidelberg (2011)
4. Popov, I.O., Schraefel, M.C., Hall, W., Shadbolt, N.: Connecting the Dots: A Multi-pivot Approach to Data Exploration. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) *ISWC 2011, Part I. LNCS*, vol. 7031, pp. 553–568. Springer, Heidelberg (2011)
5. schraefel, m.: What does It Look Like, Really? Imagining how Citizens might Effectively, Usefully and Easily Find, Explore, Query and Re-present Open/Linked Data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) *ISWC 2010, Part II. LNCS*, vol. 6497, pp. 356–369. Springer, Heidelberg (2010)
6. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *I. J. Human Mental Workload* 11, 139–183 (1988)
7. Wildemuth, B.M., Freund, L.: Assigning search tasks designed to elicit exploratory search behaviors. In: *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval, HCIR 2012*, pp. 1–10 (2012)
8. Dimitrova, V., Lau, L., Thakker, D., Yang-turner, F., Despotakis, D.: Exploring Exploratory Search: A User Study with Linked Semantic Data. In: *ACM Workshop on Intelligent Exploration of Semantic Data (IESD 2013)*, pp. 9–16 (2013)
9. Mudambi, S.M., Schuff, D.: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly* 34, 185–200 (2010)
10. Oulasvirta, A., Hukkinen, J., Schwartz, B.: When More Is Less: The Paradox of Choice in Search Engine Use. *J. Evaluation*, 1–7 (2009)
11. Alani, H., Brewster, C.: Ontology Ranking based on the Analysis of Concept Structures. In: *Proceedings of the 3rd International Conference on Knowledge Capture* (2005)
12. Roschelle, J.: Learning in Interactive Environments: Prior Knowledge and New Experience. *Knowledge Creation Diffusion Utilization, American Association of Museums*, 37–51 (1995)