LINKED OPEN DATA AND RESEARCH KNOWLEDGE MANAGEMENT: AN EXPLORATORY SEARCH AND VISUALIZATION FRAMEWORK

P. Mechant[1], M. Van Compernolle[1], A. Dimou[2], L. De Vocht[2] & L. De Marez[1]

[1]MICT-iMinds-UGent (BELGIUM)
[2] MMLab-iMinds-UGent (BELGIUM)

LINKED OPEN DATA AND RESEARCH KNOWLEDGE MANAGEMENT: AN EXPLORATORY SEARCH AND VISUALIZATION FRAMEWORK

P. Mechant[1], M. Van Compernolle[1], A. Dimou[2], L. De Vocht[2] & L. De Marez[1]

[1]MICT-iMinds-UGent (BELGIUM)
[2] MMLab-iMinds-UGent (BELGIUM)

Abstract

This paper describes and evaluates an exploratory search and visualization framework that enables non-expert users to interact and intuitively explore academic and bibliographic information. The framework, called the 'Linked Open Data Visualizations Suite' (LOD/VizSuite), uses the Linked Open Data (LOD) of RILOD, a database with academic bibliographic data from heterogeneous sources on research in Flanders (northern part of Belgium), and supports exploratory search and data analysis by means of interactive graph-based visualizations.

First a literature review on (Linked) Open Data and exploratory search is presented. Second, we describe LOD/VizSuite and its goals. Third, we investigate the impact and quality of the visualization with test users (n=36) using a multi-method approach.

Results show that a flow of graph-based visualizations is feasible with Linked Open Data and that users perceived that they found relevant insights about the researchers or disciplines they were exploring or looking for. We conclude that search and visualization frameworks on Linked Open Data and LOD/VizSuite in specific, can be supportive for scholars in academia as well as for policymakers in the research field.

Keywords: Linked Open Data (LOD), visualisation, research projects, research knowledge management, exploratory search.

1 INTRODUCTION

With more than one third of the world's population being online [1], the Internet has increasingly become part of modern living. For example, social media sites enable Internet users to interact with unprecedented levels of richness, affording them to easily consume, share, communicate, facilitate and create online [2,3]. Society has evolved from an analog one to one that is converting into a digital landscape making the amount of data available online today overwhelming. The sheer size of the data often inhibits researchers, policymakers or ordinary people to gain insight in these data in an efficient and effective way creating the need for solutions that enable users to discover, explore and analyze data.

In the first part of the paper the stage is set for describing the LOD/VizSuite by theoretically unpacking concepts such as 'Open Data' (data that can be freely used, reused and redistributed by anyone – subject only, to the requirement to attribute), 'Linked Data' (structured data so that it can be interlinked), and 'Big Data' (a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools). We also elaborate on the differences between exploratory search and exploratory data analysis, an approach that enables the data itself to reveal its underlying model and relationships. After reviewing the literature and describing the state of the art, the second part of the paper describes LOD/VizSuite and its goals. A screenshots of the user interface as well as a hyperlink to a demonstrator application are provided. The third and final part of the paper determines the impact and quality of the visualization tool and analyses how the test users (n=36) explored and perceived the visualizations, using a multi-method approach including observations, experiments and questionnaires. We then formulate a conclusion on how the LOD/VizSuite can be supportive for scholars in academia as well as for policymakers in the research field.

2 LITERATURE REVIEW

In order to set the stage for the description of the LOD/VizSuite we start by theoretically unpacking and describing various concepts including 'Open Data', 'Linked Data' and 'Big Data', as well as by elaborating on the differences between exploratory search and exploratory data analysis.

2.1 Open Data

Open data is data that can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and share-alike [4]. Open data is actually the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. The goals of the open data movement are similar to those of other 'Open' movements such as Open Source, Open Content, and Open Access. The phrase 'Open Data' itself is recent, gaining popularity with the rise of the Internet and World Wide Web and, especially, with the launch of open-data government initiatives such as Data.gov.

Open Data entails that the data must be available as a whole and in a convenient and modifiable form, thus data must be provided under terms that permit reuse and redistribution and everyone must be able to use, reuse and redistribute. Open Data supports interoperability and is considered an important impetus for innovation, often referred to as *'the new oil for the digital age'* [5].

However, as Rufus Pollock, co-founder of the Open Knowledge Foundation, stated: "Openness for data and content is not an end in itself, it's a means to an end" [6]. In short, in order for data to become useful Open Data needs to be used. Individuals and institutions should analyze it and should act on that analysis, companies and communities should build apps and services with it… To enable this the LOD/VizSuite, which is at the centre of this paper, affords users the means to discover, explore and analyze (Big/Linked) Open Data sets and gain insight in these Open Data sets in an efficient and effective way.

2.2 Linked Data

Liked Data is structured data so that it can be interlinked. Heath et al. [7] define Linked Data as "a set of best practices for publishing and connecting structured data on the Web". Linked Data describes data on the web connected through hyperlinks. Here comes the 'semantic web' or the 'web of data' in the picture as Linked Data is about making links so that a person or machine can explore the web of data: "With linked data, when you have some of it, you can find other, related, data" [8]. Linked (Open) Data brings consistency in information and displays the contents of web documents more meaningful. Links are shaped as so-called 'RDF triples' "(…) in which the triple's subject is a URI in the namespace of one server, and the triple's object is a URI in the namespace of the other. The triple's predicate URI determines the type of the link." [9].

In a nutshell, 'Linked Data' refers to a method of publishing structured data so that data can be linked and useful. Linked Data builds on standard web technologies such as HTTP and URIs, but rather than using these technologies to provide web pages to human readers, these technologies are used to provide information that can be automatically read by computers.

2.3 Big Data

In information technology, big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools. So you have a lot of challenges from a technical viewpoint. Big data is also a shorthand label that typically means applying the tools of artificial intelligence, like machine learning, to vast amounts of data.

In this context, the 6 V's are often mentioned as Big Data affords (1) radical new possibilities because of the vast amounts of data (volume), (2) from different sources (variety), (3) which are usually analysed in real time (velocity). The multitude of data gives rise to a potentially larger margin of error requiring (4) other forms of data authentication (veracity) and (5) other approaches to defining the relevant value of the data (value) as well as (6) other ways to make data understandable (visualization).

In reports published by McKinsey Global Institute [10] and Pew Research Centre [11] experts in the field acknowledge that analysing Big Data will help people be more agile and adaptive, and that Big Data has swept into every industry, business function and research discipline. However, these reports also point to several issues that need to be addressed to capture the full potential and impact of Big

Data such as organizational change, the development of new technologies for storage, computing and analysis, or policies with regards to accessing, searching, aggregating, and cross-referencing large data sets. Mantelero [12] points out that these huge amounts of data represent a strategic and economically relevant asset resulting in a centralized power held only by a few subjects, creating information asymmetries in which public and private entities (who collect, manage and create data), intermediaries (e.g. ISPs) and entities who have the skills to manage an overdose of information, hold the informational power.

2.4 Exploratory search and exploratory data analysis

Exploratory search refers to information exploration by people who are (1) unfamiliar with the domain of their goal, (2) unsure about the ways to achieve their goals, or (3) unsure about their goals in the first place. Exploratory search can describe either the problem context that motivates the search or the process by which the search is conducted [13]. This means that users starts from a vague but still goal-oriented defined information need and are able to refine their need upon the availability of new information to address it, with a mix of keyword look-up, expanding or rearranging the search context, filtering and analysis.

Exploratory data analysis refers to an analysis approach that summarizes the main characteristics of the data, often with visual methods in order to see what the data can tell us beyond formal modelling or hypothesis testing task. It is an approach to data analysis that allows the data itself to reveal its underlying model and their relationships without requiring any formal statistical modelling and inference.

During exploratory search and analysis, it is likely that the problem context will become better understood by the searcher, allowing him/her to make more informed decisions about interaction or information use [14]. With the LOD/VizSuite we bring together these two information exploration approaches and we apply them to graph visualizations for the exploration of (Big/Open) Linked Data.

3 LOD/VIZSUITE

3.1 Description

The LOD/VizSuite is a tool supporting exploratory data analysis of Flemish Universities and their research metadata, exposing research and collaboration networks or communities of practice in a certain discipline. The LOD/VizSuite aims to be data and schema agnostic so it can be easily transferable to visualize different datasets. The LOD/VizSuite provides visualizations based on the Linked Open Data provided by the ‘Research Information Linked Open Data’ (RILOD) dataset [15]. It offers an overview and details of the groups in the dataset (discipline level in this case), through which the users dives in a more narrow perspective (research groups) until they reach a specific resource (researchers), see Figure 1. The LOD/VizSuite focuses on the exploratory analysis of the research groups' relations based on publications and projects rather than on the individual researchers' relations.

3.2 Goal

The goal of the LOD/VizSuite is to create a framework for exploratory search and exploratory data analysis, on top of the DataTank, a (Linked) Open Data management system [16]. Its functionality is based on SPARQL queries which are published as SPARQL templates at a DataTank instance. LOD/VizSuite exposes research and collaboration networks, communities of practice in a certain discipline and timelines to monitor a discipline's evolution over time.

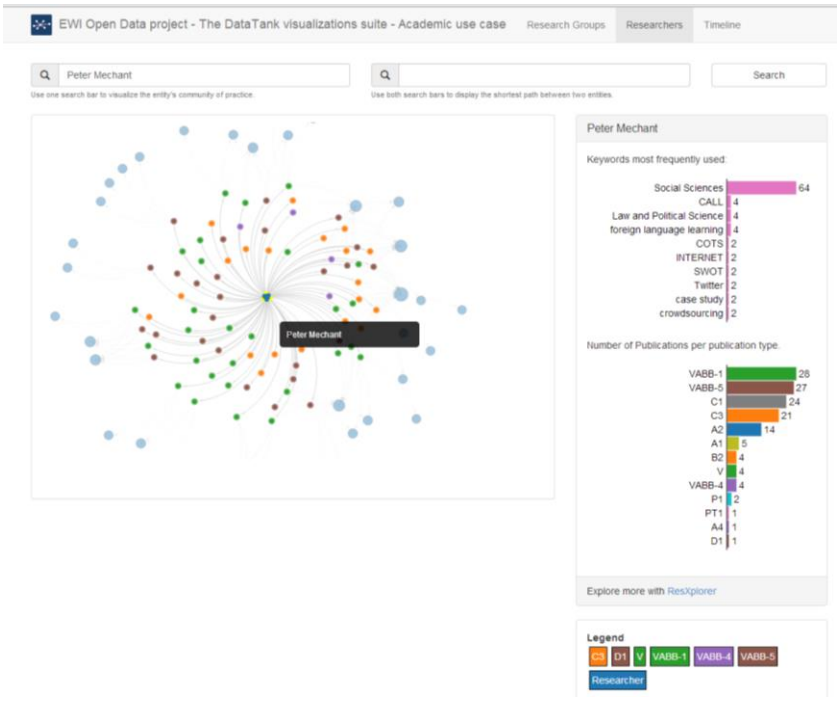

Figure 1: Screenshot of the LOD/VizSuite (http://ewi.mmlab.be/academic/)

4 RESEARCH

4.1 Methodology

We evaluated the LOD/VizSuite tool at iMinds The Conference 2013 in Brussels, a yearly gathering for research active in several aspects of digital innovation, where almost half of the attendees have a background as a researcher. Test users evaluated the tool in two ways: a user test and a questionnaire. They received no information about the tool in advance. They were asked to execute certain assignments and to fill in a questionnaire afterwards.

4.1.1 User tests and assignments

During the user test, users were asked to think aloud. An evaluator observed the comments and took notes. The users' actions were recorded using QTrace [17]. Each test took about 30 to 45 minutes. We gave the user to complete three assignments: (1) The test users were asked to search for their preferred discipline, and try to understand and analyse the collaborations between the displayed research units, (2) The test users were asked to navigate from the discipline's visualization to or search for their preferred research group and try to understand and analyse the collaborations of the displayed research groups, and (3) Users tried to find the relations to another researcher by searching and exploring related resources until the users were satisfied with the results.

4.1.2 Online survey

The survey included questions on: (1) the perceived goals of the tool; (2) a series of statements on a five-point Likert scale measuring the usefulness learnability, complexity explorability and transparency as perceived by the respondents; and (3) open questions on data quality and suggestions.

4.1.3 Sample

Sixteen test users participated in the test during iMinds The Conference 2013 and filled out the questionnaire. Also, twenty additional test users were recruited and filled out the same questionnaire. There was a good match between the test users and the conference participants; we selected diverse roles of test users (both researchers and innovation policy-makers).

4.2 Results

Test users were asked to interpret the displayed visualizations the received by using the LOD/VizSuite for executing three assignments: (1) search for preferred discipline and look at the collaborations between the displayed research units, (2) navigate from the discipline's visualization to or search for their preferred research group and look at the collaborations of the displayed research groups, and (3) find relations to another researcher by searching and exploring related.

To understand how users perceive the LOD/VizSuite, we asked them in an online survey to score possible purposes. The respondents indeed perceived the LOD/VizSuite as adequate tools to explore, discover and search, agreeing that the LOD/VizSuite is useful for broadening views (65% - data consumer find novel relations between existing and known resources) and for the narrowing views (82% - data consumer familiarize with a certain dataset, as they are not aware of its context).

With regards to perceived complexity, almost all respondents agree or strongly agree with the statement that after a learning period, users should be able to get benefits out the visualisations (median = 4). They stated that they have found relevant insights about the researchers they were looking for (median = 4) and the majority of the test users agreed that they can learn quickly to interpret the visualizations using the LOD/VizSuite (median = 4).

With regards to explorability, the test users agreed on the fact that they would use the tool to explore opportunities for collaborations (median = 4) with many agreeing with the fact that the visualizations facilitate the exploration of the published Linked Open Data (median = 4).

Most test users agreed (24/36) that the visualizations helped in revealing the potential of (Linked) Open Data. The same number of test users believes that these visualizations can become an encouraging factor for data publishers to provide more Open Data and of better quality. Nevertheless, less than half of the test users (16/36) believed that the visualizations are adequate to gain insights into innovation and economic growth, but it should be remarked that this question had the most undecided test users (11/36). Finally, 22/36 of the test users agreed that the government becomes more transparent if the public is able to explore the published Open Data with such visualizations.

5 DISCUSSION AND CONCLUSION

In this paper we presented preliminary results of novel work. To the best of our knowledge, we are one of the first to present and evaluate a solution for combined exploratory search and analysis applied to Linked Open Data.

Our results indicate that the new behavioural flow of graph-based visualizations is feasible with Linked Data. We tested how well it affects the user behaviour while exploring the information and vice versa and found that our visualizations proved to be better optimized for the interpretation of prevalent resources, especially in more dense graphs. Furthermore, we observed that searching for resources increases the visualized set of resources with the most new relevant resources, while it is on average as effective as expanding resources. The results of our questionnaire indicate how end-users perceive the visual work streamlined across the different views we provided them. Overall, user interfaces based on graph visualizations allow the scholars to have a unique, multifaceted experience when combined with techniques for information exploration and enhanced with optimized search in Linked Data. Such visualisations enable scholars to view and navigate through combined aspects of research data and come up spontaneously with observations whose potential reasoning can be investigated by narrowing down their view.

This paper contributes to the field of knowledge management, open data and data visualization by empirically assessing the effectiveness and efficiency of LOD visualization technology, by theoretically unpacking and discussing key concepts, and by describing a case study analysis of LOD/VizSuite.

ACKNOWLEDGEMENTS

The research for this paper was funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) and was conducted in the context of an iMinds research project (http://www.iminds.be).

REFERENCES

[1] ITU (2012): *Measuring the information society*, International Telecommunication Union Report. Geneva, Switzerland: United Nations. Available at http://www.itu.int/ITU-D/ict/publications/idi/material/2012/MIS2012_without_Annex_4.pdf.

[2] Beer, D. (2009). Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society, 11*(6), 985-1002.

[3] Slot, M., & Frissen, V. (2007). *Users in the 'Golden' age of the information society*. Paper presented at the Cost 298 Conference - The good, the bad and the unexpected, Moskou. Available at http://www.userproducer.nl/Documents/P036-final.pdf.

[4] OKFN. (2013). Open Definition. Available at http://opendefinition.org.

[5] Kroes, N. (2012). *From Crisis of Trust to Open Governing*. Available at http://europa.eu/rapid/press-release_SPEECH-12-149_en.htm.

[6] Pollock, R. (2011). Open Data: a means to an end, not an end in itself. Available at http://blog.okfn.org/2011/09/15/open-data-a-means-to-an-end-not-an-end-in-itself.

[7] Heath, T., Hepp, M., & Bizer, C. (2009). *Linked Data—The story so far*. Special Issue on Linked Data, International Journal on Semantic Web and Information Systems (IJSWIS).Available at http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

[8] Berners-Lee, T. (2006). *Linked Data*. Available at http://www.w3.org/DesignIssues/LinkedData.html.

[9] Bizer, C. (2009). The Emerging Web of Linked Data. *Intelligent Systems, IEEE, 24*(5), 87-92.

[10] McKinsey Global Institute (2011). Big Data: the next frontier for innovation, competition and productivity. Available at http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx.

[11] Pew Research Center. (2012). Big Data: Experts say new forms of information analysis will help people be more nimble and adaptive, but worry over humans' capacity to understand and use these new tools well. Available at http://www.pewinternet.org/files/old-media/Files/Reports/2012/PIP_Future_of_Internet_2012_Big_Data.pdf.

[12] Mantelero, A. (2012). Masters of Big Data: concentration of power over digital information. Available at http://dx.doi.org/10.2139/ssrn.2048236.

[13] Fu, W.-T. (2009). *A Semantic Imitation Model of Social Tagging.* Paper presented at the Proceedings of the IEEE conference on Social Computing.

[14] Marchionini, G. (2006). Exploratory search: From finding to understanding. *Commun. ACM, 49*(4):41-46.

[15] Research Information Linked Open Data (RILOD). Available at http://ewiscloud3.vlaanderen.be/ewilod/html/sparql-test.html.

[16] The DataTank. Available at http://thedatatank.com.

[17] QTrace. Available at http://www.qasymphony.com/qtrace.html.