A Visual Workflow to Explore the Web of Data for Scholars

Anastasia Dimou[1]
anastasia.dimou@ugent.be

Laurens De Vocht[1]
laurens.devocht@ugent.be

Mathias Van Compernolle[2]
mathias.vancompernolle@ugent.be

Erik Mannens[1]
erik.mannens@ugent.be

Peter Mechant[2]
peter.mechant@ugent.be

Rik Van de Walle[1]
rik.vandewalle@ugent.be

[1]Ghent University - iMinds - Multimedia Lab
Ghent, Belgium

[2]Ghent University - iMinds - MICT
Ghent, Belgium

ABSTRACT

As the Web evolves in an integrated and interlinked knowledge space thanks to the growing amount of published Linked Open Data, the need to find solutions that enable the scholars to discover, explore and analyse the underlying research data emerges. Scholars, typically non-expert technology users, lack of in-depth understanding of the underlying semantic technology which limits their ability to interpret and query the data. We present a visual workflow to connect scholars and scientific resources on the Web of Data. We allow scholars to move from exploratory analysis in academic social networks to exposing relations between these resources. We allow them to reveal experts in a particular field and discover relations in and beyond their research communities. This paper aims to evaluate the potential of such a visual workflow to be used by non-expert users to interact with the semantically enriched data and familiarize with the underlying dataset.

1. INTRODUCTION

The usefulness of Linked Data to most scholars is evident, but a lack of in-depth understanding of the underlying semantic technology limits their ability to interpret and query the Web of Data. A key solution in overcoming this, is to visualize Linked Data in a coherent and legible manner, allowing scholars to implicitly compose queries, identify links between resources and intuitively discover new pieces of information [4]. Our work has been motivated by the presence of ongoing efforts for resource reuse and exchange indicating that the idea of Linked Data plays a substantial role in the context of digital libraries and archives, too [8]. This is because following the Linked Data design principles, outlined as a set of 'best practices' for publishing data on the Web [1], allows treating conceptual entities as dereferencable resources and therefore to not only retrieve metadata for a specific resource, but also to retrieve all referred resources from multiple scholars.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579039.

The resolution of vague or complex information problems, such as when multiple publishers provide resources, requires exploratory behaviors. We examine how *exploratory search* and *exploratory data analysis* applied to Linked Open Data (LOD) can fruitfully solve such problems. Exploratory Data Analysis (EDA) [13] is an approach to data analysis that allows the data itself to reveal its underlying model and their relationships without requiring any formal statistical modeling and inference (non hypothesis-driven). Graphical EDA employs a variety of techniques to present the underlying data, maximizes the insight into a dataset and uncovers the underlying data patterns, allowing the users to discover the dataset.

Exploratory search, on the other hand, can describe either the problem context that motivates the search or the process by which the search is conducted [11]. This means that the users start 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. During exploratory searches and analysis, it is likely that the problem context becomes better understood, allowing the searchers to make more informed decisions about interaction or information use [15].

Interactive visualization of information is a technique to facilitate exploring and analyzing data repositories. The advantage of visual representations is that the users see multiple aspects of the data while controlling its views. We consider that their combination can play a determinant role in the exploration of Linked Data by (non-expert) users, revealing new potential as it is subversive to the current user behavior on the Web but immanent to humans cognitive behavior and the Linked Data representation. We applied each approach separately to an application that provides visualizations on top of selected datasets from the Web of Data. We then evaluated the users' interaction with these visual representations as a mean to discover the underlying datasets. The potential of scholars' interaction with the Web of Data using visualizations is high, as a combination of graphical approaches for exploratory search and analysis could have a decisive determinant role in the exploration of the Web of Data by non-expert users.

This paper is structured in three parts: Firstly, related work in the area of visualizations of LOD and research metadata is presented. Then, a detailed description of our workflow follows. Last, the evaluation of the visualizations' workflow is described and the results are discussed.

2. RELATED WORK

To the best of our knowledge, no previous evaluation of exploratory search and data analysis combined to facilitate exploring Web of Data for scholars exists. In this section, we summarize earlier work related to (research) Linked Data visualizations and to evaluation of information visualizations.

2.1 Visualizing Linked Data

Providing software that offers a unified experience can simplify the use of Linked Data by people who are not Semantic Web developers. A survey by Dadzie et al. summarizes efforts about visualizing datasets structured as Linked Data [4]. Graves et al. developed Visualbox to demonstrate that it is possible to facilitate the process of creating web-based visualizations on top of Linked Data [7]. They concluded that their editor was still too general for users working with visualizations, but the users valued that once a query was ready, the construction of a visualization was trivial. In our workflow we make abstraction of the query creation process and use of pre-defined query templates to facilitate the creation of the visualizations. Last, in VisLink coordinated views link existing visualizations [3], while we use a coordinated view to align the narrowing and broadening views in the workflow.

2.2 Visualizing Research Networks

In the past there were attempts to visualize research networks but most of them did not rely on Linked Open Data. Recent works based on research Linked Data consider visualizations as a supportive mean to the presented information. None of them puts the focus on the visualizations and, therefore they do not take into consideration a certain approach for the data exploration. Recently, the Semantic Web Journal published its own Drupal-based journal management system [9] with the focus on providing a novel user interface. Among others, they provide graph-based research networks that visualize the emerging research networks as researchers author papers together or they review the different submissions. ArnetMiner [12], on the other hand, distinguishes between the networks (star graph of co-authors) and the communities of researchers (simple graphs) but does not provide any weights on the graph's visualization nor any interaction. Finally, TalkExplorer [14] takes into consideration bookmarks and tags for the visualizations of the research groups and puts the focus on providing recommendations rather than exploring the underlying dataset. Overall, none of the aforementioned applications provide an interactive visualization interface which is appropriate for exploratory search and analysis of the underlying Linked Data, as our visualization workflow does.

2.3 Evaluating Information Visualizations

The goal of information visualization evaluation is to measure the effectiveness on supporting people in their information tasks and how people conduct their information related tasks so that visualizations can be better designed [2]. Reasons for choosing an evaluation method are often related to typical human computer interaction empirical research practices and its constraints (time, finding the right participants, previous knowledge of the participants, learning tempo, etc.). As in all studies, depending on the preferred results, information visualization evaluations happen via quantitative or qualitative methods. Carpendale [2] describes possible methods such as laboratory or field experiments and studies, sample surveys, judgements studies, formal theory and computer simulation. Quantitative methodologies are used to find relations between variables and to make generalizations to a broader population. Qualitative methods often almost rely on interviews or observations. Methodologies supporting these are: experimenter observations, think-aloud protocol, opinion collecting, sets of heuristics (usability, collaboration, etc.), in situ observational studies, contextual interviews and participatory observations.

3. INTEGRATED VISUAL WORKFLOW

Our workflow enables users to discover, search and analyse research metadata. The workflow of the visualization is streamlined through a coordinated view of two different views centralizing the link focused on a specific resource that binds them. Figure 1 shows how users start with an overview of the groups in a dataset (1a) through which the users "dive" in a more narrow perspective (1b) by selecting a group to find out details and see the internal relations of the subdivisions (1b). A coordinated view (1c) of selected resources leads them through a broadened view (1d) by exploring relations of these resources.

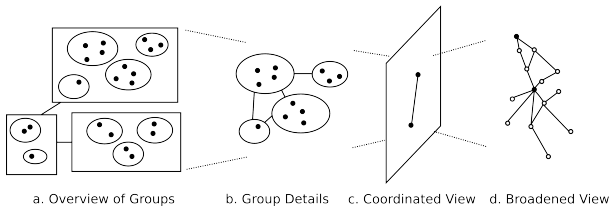

Figure 1: A coordinated view combines the workflow.

In our use case, the narrowing view is achieved based on exploratory analysis techniques applied on the research groups' relations, represented as an aggregation of individual researchers, considering their publications and projects in total rather than on individual researchers' achievements and relations. The broadening view is achieved using exploratory search techniques over individual researchers, their publications and their relationships. The combination of the two allows users to gradually discover the dataset and enables them to discover, search and analyse LOD on the Web.

Narrowing View. In the narrowing view, the visualisations are more focused on the collaboration networks and communities of practice but also on research networks. We provide visualizations based on the LOD provided by the "Research Information Linked Open Data" (RILOD)[1] dataset. RILOD is the result of the integration of heterogeneous sources related to research in Flanders, ending up in a rich and diverse dataset. The LOD contains resources of researchers, publications and projects which are associated with the corresponding research groups and institutes and classified under the IWETO Discipline classification[2].

The narrowing view relies on grouping and aggregating resources based on their types and properties. For example, visualisations that provide a broad view of a research

[1] http://ewilod.be/ewilod/html/sparql-test.html

[2] http://ewilod.be/ewilod/lod/0.1/ontology/taxonomies/iwetoDisciplineCodes

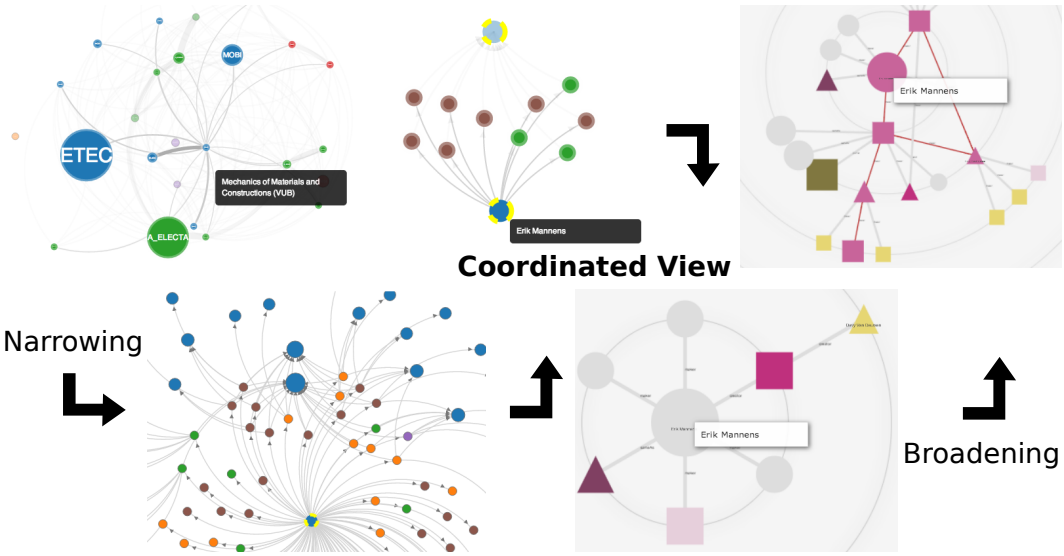

Figure 2: Scholars narrow down from disciplines to research groups and further to the individual researchers in this group. To find out relations between researchers they select two researchers and can then use the coordinated view to shift to the broadening view and expand to resources beyond their research community.

discipline are achieved by aggregating the researchers under their research groups and providing their collaboration links considering their co-publications. The research groups are demonstrated as graph nodes that diversify in size depending on the total number of projects they have and in color depending on the institute they belong to, while the strength of the links depends on their researchers' co-publications. Since there is no explicit assumption regarding the dataset, as EDA ordains, the dataset itself reveals its underlying model and the relationships between its resources.

Coordinated View. As the scholars, supported by the visualisations, narrow down to more detailed resources, namely a certain resource or the links between two resources, they reach the resources that cannot be further decomposed and thus act as the coordinated view between the two views. In our use case such a resource can be an individual researcher or the links between two researchers of the research group whose extensive collaboration network is demonstrated. Starting from this coordinated view, the scholars, being aware of the underlying dataset, can define a goal and explore further towards a "controlled" perspective, discover the emerging relations and rearrange the central focus.

Broadening View. The goal of this view is that scholars can find novel relations between existing and known to them resources such as authors, publications or conferences. Therefore, an optimized pathfinding algorithm [5] is used to this end. We make use of the "Digital Bibliography and Library Project" (DBLP), an on-line reference for computer science bibliography[3] [10]. Users interact with a visualization of resources from DBLP combined with the contributions of researchers on Social Media, such as conferences, publications and proceedings. It exposes affinities between researchers and those resources. Each shown affinity between researchers and resources captures the amount of shared interests and other commonalities [6].

4. EVALUATION

We evaluated the workflow at *iMinds The Conference 2013* in Brussels[4], a yearly gathering for people active in several aspects of research and digital innovation. We implemented a demo of the workflow for the evaluation, integrating our visualisations for scholars, as shown in Figure 2: The *LOD Visualization Suite*[5] (LOD/VizSuite) which implements the narrowing view and *ResXplorer*[6] which implements the broadening view.

Test users evaluated the tool in two ways: a user test and a questionnaire. These methods give us insight in how the users perceive the tools and show us quickly potential bottlenecks [7]. Sixteen (16) test users were selected to participate to the observation and they received no information about the demo that implements the workflow in advance. They were asked to execute certain assignments and to fill in a questionnaire afterwards. During the user test, users were asked to think aloud and their actions were recorded while an evaluator observed the comments and took notes. Each test took about 30 to 45 minutes. The questionnaire took 5 to 10 minutes and it was completed also by another twenty (20) additional test users who only attended the demo presentation. There was a good match between the test users and the conference participants; we selected diverse profiles of test users (both researchers and innovation policy-makers).

The survey included questions regarding the usefulness, learnability, complexity and explorability as perceived by the respondents using a five-point Likert scale and questions regarding the workflows components (narrowing and broadening view) where the users score statements. The evalua-

[3] http://dblp.rkbexplorer.com

[4] http://conference2013.iminds.be

[5] http://ewi.mmlab.be/academic

[6] http://www.ResXplorer.org

tion targets the assessment of the system's functionality on extent and accessibility, the assessment of users' experience of the interaction, and the identification of specific problems with the system. Due to the integrated and nature of our interactive visual workflow, we do not compare our proposed solution against a purely keyword based search engine such as Google Scholar[7], DBLP (L3S) Faceted Search[8], etc. We therefore did not incorporate any of traditional information retrieval metrics, beside precision, that could be expected in a more comparative evaluation.

4.1 Assignments

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.
3. Users tried to find the relations to another researcher by searching and exploring related resources until the users were satisfied with the results.

4.2 Data Gathering

Data was gathered using a multi-method approach. We used the think-aloud protocol during the experiments to collect feedback from participants. We recorded the screen actions of participants using QTrace[9]. Afterwards we asked the participants to fill out a questionnaire about their background and their perception of the usefulness of the workflow's components and the extent the existence of such tools would affect their behaviour on the Web in this domain.

4.3 Results

To determine the impact and quality of the workflow considering their use for scholars, we analyzed how the users explored and perceived the visualizations. We measured the *positive predictive value* and *true positive rate* in the narrowing view case and the *precision* and *productivity* in the broadening view case.

4.3.1 Survey Results

Complexity. Considering the perceived complexity, shown in Figure 3, almost all respondents agree or strongly agree with the statement that after a learning period, users should be able to get benefit out the visualisations (median = 4, agree). They perceive that they have found relevant insights about the researchers they were looking for (medians=4). Finally, the majority of the test users agreed that they can learn quickly to interpret the visualizaton (median = 4, agree) where many of the respondents agree with this for the narrowing view (median = 4, agree).

Explorability. We see in Figure 4 that test users agree on the fact that they would use the workflow to explore opportunities for collaborations (median = 4, agree). The middle box plots indicate that many of the respondents agree with

[7] http://scholar.google.com
[8] http://dblp.l3s.de/
[9] http://www.qasymphony.com/qtrace.html

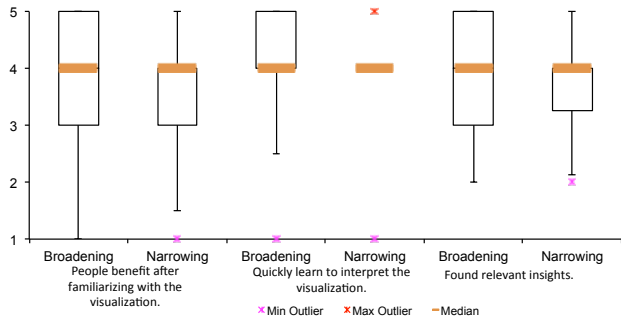

Figure 3: Complexity

the fact that the visualizations facilitate the exploration of the published Linked Open Data (median = 4, agree). Finally, the test users strongly agree that broadening the view supports gaining insights into the published data (median = 5, strongly agree), while test users agree for the narrowing view case as well but they are not the same confident (median = 4, agree).

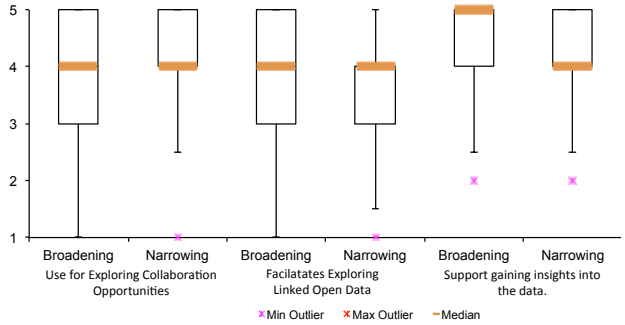

Figure 4: Explorability

4.3.2 Narrowing View evaluation

In the first assignments the test users were asked to interpret the displayed visualizations. In assignment 1, a network of research units which are active in a certain discipline was displayed as a weighted graph. In assignment 2, the collaborations of a research unit with others were displayed in the form of a weighted star graph. The users were expected to interpret the visualized result considering the differences in the displayed graph's vertices and edges sizes. We measured the *Positive Predictive Value (PPV)*, how often users perceived (TP + FP) a distinctively evident node or edge compared to the others displayed (TP). Additionally, we evaluated the visualizations' *True Positive Rate (TPR)*, namely how good the displayed visualization is at prominently representing (TP) the prevalent vertices and edges (P).

$$PPV = \frac{TP}{TP + FP} \quad (1)$$

$$TPR = \frac{TP}{P} \quad (2)$$

All test users perceived spontaneously the differentiation of the vertices size and only one did not understand the differentiation of the edges size in the first assignment. On the

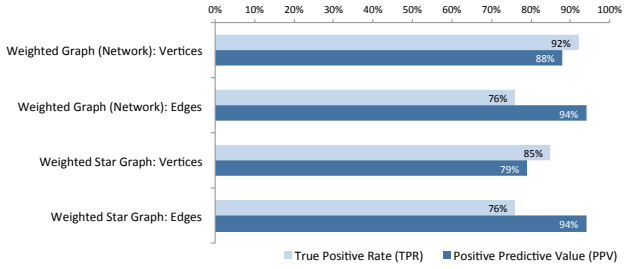

Figure 5: TPR and PPV results

other hand, 75% of the test users ($^{12}/_{16}$) understood spontaneously the evidently prevalent vertices and 56% of them ($^{9}/_{16}$) the evidently prevalent edges. This coincides with the visualizations' true positive rate. As shown in Figure 5, the visualizations proved to better display the prevalent nodes (the true positive rate is equal to 92% for the first assignment and 85% for the second assignment), compared to the prevalent edges (the true positive rate is equal to 0.76 in both assignments), and better interpretation of the scaling is achieved in the cases of more dense graphs as in the first assignment (average density 0.30), compared to the second assignment (average density 0.14).

Despite the aforementioned, it is remarkable that the test users identify and better interpret the evidently prevalent edges compared to the evidently prevalent vertices. In more details, the positive predictive value in the first assignment was 88% for the vertices and 94% for the edges. In the second assignment the positive predictive value was exactly the same for the vertices and almost the same (85%) for the edges. Thereafter, it is evident that, by the time the test users observe the differentiation of the edges' size, they better interpret it. In contrast, the test users tend to misunderstand more the diversification of the vertices' nodes. The positive predictive value is equal to 88% in the case of the first assignment and equal to 79% in the case of the second assignment. Again, the test users performed better in more dense graphs, such as in the first assignment, than in less dense graphs, as in the second assignment.

4.3.3 *Broadening View evaluation*

We asked users to find a relevant person to contact or a conference to go to. The users had to mark all visualized nodes if they were relevant to them. For visualizing resources, users could choose between three actions: searching, adding top related resources and expanding neighbours of visualized resources. In the last case they could choose between direct or indirect neighbours of the centrally focused node in the visualization.

Precision as measure for *effictivity* (E) defines how often a visualized result (R) related to a resource was marked relevant by the user (M).

$$\textbf{precision} = E = \frac{\mid M \cap R \mid}{\mid R \mid} \tag{3}$$

Additionally we checked *Productivity (P)* for each type of action A, we verified that after each action that delivered new resources to the result set resulted in an increase of quality of the result set. The quality of a result set is the number of marked relevant resources compared to the total number of visualized resources.

$$\textbf{productivity} = P_A = \sum_{k \in A} \frac{E_k - E_{k-1}}{\mid A \mid} \tag{4}$$

The *E/P Ratio* indicates the impact of the newly visualized nodes on the existing visualized resources. We get this ratio by dividing productivity by precision.

Adding a top related resource was not done often by the users and added only a couple of resources to the result set. It is however the most effective action as the users marked $^{13}/_{26}$ (50%) of the visualized resources relevant. We also see that the productivity in Figure 6 is 12%, this means that on average over all test users. Adding top related resources resulted in a result set that contained 12% more relevant nodes as before adding top related nodes. The E/P ratio is 24% ($^{0.12}/_{0.5}$).

Searching for a resource was still more productive and has an E/P ratio of 81% ($^{0.25}/_{0.31}$). This is remarkable as the precision of searching is much lower than adding a top related resource. This means that the impact of each added resource when searching is much bigger, because the quality of the result set was not relatively high at the moment users decided searching: on average less than 31% of the resources. This would result in an increase in productivity if of the newly added resources at least 31% was marked relevant according to users.

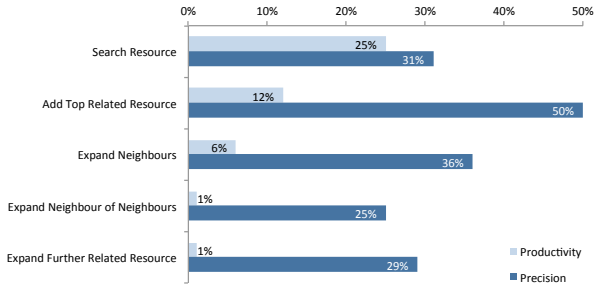

Figure 6: Productivity and precision results

We see that the precision of expanding resources $^{53}/_{166}$ (32%), is about the same as searching for a resource $^{54}/_{174}$ (31%). As the user actions resulted in about as many new resources in the case of searching and expanding, this is a very reliable comparison. Expanding the direct neighbours is the most productive 6%. Expanding further related neighbours does not really positively impact the result set, productivity is only 1%. It retains the quality of the result set. Negative productivities were also noted for some users as their action leaded to a decrease in quality of the visualized result set. Users mostly expanded direct neighbours, which leaded to 94 new resources compared to 72 expansions of indirect neighbours.

4.3.4 *Think-aloud analysis*

The *think-aloud* analysis gave us information regarding how the participants perceive the visualizations (information visualized by vertices and edges), during the execution of the assignments. Considering the direct feedback of the users, we conclude that they are able to reason via the tools: e.g. by bringing the size of the nodes into discussion in LOD/VizSuite or by appointing missing groups.

The test users declared that they are not used to explore the information in this way but most of them found it interesting. The major drawback observed was related to the incremental appearance of visualized information in both tools. When there is a high amount of vertices and edges displayed (e.g. when higher level research groups are chosen with the LOD/VizSuite or after expanding actions with the ResXplorer). The test users stated that getting an overview gets more difficult (if there are lots of links) and that information is displayed in segments in LOD/VizSuite; keeping track of the former steps is not possible as with ResXplorer.

5. CONCLUSIONS AND FUTURE WORK

In this paper we present preliminary results of a novel workflow for exploring resources in the Web of Data and the evaluation of a demo implementation that we developed. To the best of our knowledge, we are one of the first to present and evaluate a solution for exploratory search and analysis. After analyzing the observed users' behavior and their answers in a questionnaire, we highlight some results:

An interactive visual workflow is feasible in the Web of Data for scholars. We tested how well it affects the user's behavior while exploring the information 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 workflow streamlined across the different views we provide 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. Therefore, the workflow can prove to be rather useful on research policy makers who want to figure out the impact of individuals on the overall performance and efficient collaboration of research groups but also on individuals level.

In the future we expect to provide an integrated solution that aligns with the users' streamline workflow as they explore the information that will be more thoroughly evaluated. The distance between the nodes will be taken into consideration to give meaning to the position of the nodes in the visualization. Additionally, we plan to support multigraphs introducing multiple relations between displayed resources and subgraphs and direct expansion of the aggregated resources (e.g. researchers of a research group or publications contributing to the edge's weight). Finally, we would like to evaluate its applicability to other domains such as cases of exploring other type of actors and their actions' output.

6. ACKNOWLEDGEMENT

The research activities described in this paper were funded by Ghent University, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

7. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[2] S. Carpendale. Information visualization. chapter Evaluating Information Visualizations, pages 19–45. Springer-Verlag, Berlin, Heidelberg, 2008.

[3] C. Collins and S. Carpendale. Vislink: Revealing relationships amongst visualizations. *IEEE Trans. on Visualization and Computer Graphics (Proc. of the IEEE Conf. on Information Visualization (InfoVis))*, 13(6), 2007.

[4] A.-S. Dadzie and M. Rowe. Approaches to visualising Linked Data: A survey. *Semantic Web*, 2(2):89–124, 2011.

[5] L. De Vocht, S. Coppens, R. Verborgh, M. Vander Sande, E. Mannens, and R. Van de Walle. Discovering meaningful connections between resources in the Web of Data. In *Proceedings of the 6th Workshop on Linked Data on the Web*. CEUR-WS, 2013.

[6] L. De Vocht, E. Mannens, R. Van de Walle, S. Softic, and M. Ebner. A search interface for researchers to explore affinities in a Linked Data knowledge base. In *Proceedings of the 12th International Semantic Web Conference Posters and Demonstrations Track*, pages 21–24. CEUR-WS, 2013.

[7] A. Graves. Creation of visualizations based on Linked Data. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, page 41. ACM, 2013.

[8] B. Haslhofer and B. Schandl. The oai2lod server: Exposing oai-pmh metadata as Linked Data. 2008.

[9] Y. Hu, K. Janowicz, G. McKenzie, K. Sengupta, and P. Hitzler. A linked-data-driven and semantically-enabled journal portal for scientometrics. In *The Semantic Web - ISWC 2013*, volume 8219 of *Lecture Notes in Computer Science*, pages 114–129. Springer Berlin Heidelberg, 2013.

[10] M. Ley. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *String Processing and Information Retrieval*, pages 1–10. Springer, 2002.

[11] G. Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[12] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. KDD '08, pages 990–998. ACM, 2008.

[13] J. Tukey. *Exploratory Data Analysis*. Addison-Wesley series in behavioral sciences. Addison-Wesley Publishing Company, 1977.

[14] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 351–362, 2013.

[15] R. W. White and R. A. Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.