CRIS 2014

Visualizing the information of a Linked Open Data enabled Research Information System

Anastasia Dimou[a], Laurens De Vocht[a], Geert Van Grootel[b], Leen Van Campe[b], Jeroen Latour[c], Erik Mannens[a], Rik Van de Walle[a]

[a]Ghent University – iMinds – Multimedia Lab, Gaston Crommenlaan 8, bus 201, B-9050 Ledeberg-Ghent, Belgium
[b]Flemish Government, Department of Economy, Science and Innovation, Koning Albert II-laan 35, bus 10, 1030 Brussels
[c]IBM Netherlands BV, Johan Huizingalaan 765, 1066 VH Amsterdam, Netherlands

Abstract

The Open Access movement and the research management can take a new turn if the research information is published as Linked Open Data. With Linked Open Data, the management of the research information within institutions and across institutions can be facilitated, the quality of the available data can be improved and their availability to the public is assured. However, it can be difficult for non-expert users to take advantage of the interlinked information offered by Linked Open Data as they lack of in-depth knowledge. In this paper, we present a use case of publishing research metadata as Linked Open Data and creating interactive visualizations to support users in analyzing the Flemish research landscape.
© 2014 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of euroCRIS.

Keywords: Research Information System, Linked Open Data, Linked Open Data visualizations

1. Introduction

Publishing research information as Linked Open Data promotes the Open Data Management in research administration and complies with the Open Access trends. Thanks to the semantic annotations, research information metadata are interlinked in such way that complex queries can be fairly easily answered for the first time. Thereafter, research management can take a new turn as the research information management within institutions and across them can be better coordinated, the quality of the available data can be improved and their availability to the public is facilitated and assured. However, forming such queries and traversing the research information relying on the published links remains a complicated task for non-Semantic Web experts lacking of in-depth knowledge. In this paper, we present a use case of publishing research data focusing on their consumption through visualizations, which aim to reveal the potential of (Linked) Open Data for research information management. Publishing research

metadata as Linked Open Data and visualizing them makes it possible to reveal links and relationships that are not obvious without any deeper analysis of the available data.

This paper is structured as follows: Firstly, we present other tools that use visualizations to display their data. In section 3 we explain how the data is published as Linked Open Data. In section 4 we explain the architecture of the visualizations and in section 5 we demonstrate different visualizations. Finally, in section 6 we present some results of a preliminary evaluation of the visualizations.

2. Related Work

Similar visualizations on top of research data are explored by other applications, too. However, in most of the case, the visualizations are used to demonstrate the provided information in a complementary fashion to the information they present.

Recently, the Semantic Web Journal published its own Drupal-based journal management system [4] and provides graph-based research networks that visualize the research networks that emerge as researchers author papers together or review the different submissions based on its own dataset. ArnetMiner [5] distinguishes the networks (unlinked co-authors) and the communities of researchers (linked co-authors) into two different types of visualizations. On the other hand, ResXplorer [6] differs significantly as it provides the optimal path between two entities (researchers, publications, papers, or conferences) and allows the user to view the emerging graph from different perspectives bringing in the center of the focus any of the existing or not entities.

Finally, the Science Atlas[1]. is the only corresponding visualizations system that provides information about the disciplines the researchers are active in. In the same context, beyond the research networks of the co-authors, the LOD/VizSuite instance for RILOD can present alternative paths between two researchers. It uses the Everything Is Connected Engine EICE [3], the same engine that ResXplorer uses, whose function relies on an optimized path-finding algorithm [7]. In this way, alternative paths, and thus alternative opportunities to establish collaborations, are represented.

3. Publishing Research Information as Linked Open Data

The Flemish public administration understands the crucial importance of having a well-planned research and innovation policy to support and stimulate the region's economic growth. Hence the *Brussels declaration on Open Access*[2] was signed as well as a long-term plan from the *Department of Economy, Science and Innovation*[3] aiming to improve the availability and quality of the information about the Flemish research.

A high-performance environment for publishing the research metadata information following the principles of Linked Open Data was implemented in Flanders as a result of the RILOD pilot project by the Flemish government. Heterogeneous sources from different research institutions and from the Flemish Academic Bibliographic Database (VABB)[4] were integrated and semantically annotated resulting in the Research Information Linked Open Data (RILOD) dataset[5] [1], [2].

In total, the triplestore contains information about 22.006 researchers, 2.022 research groups, 24.635 projects and 608.729 publications, resulting in more than 400.000.000 triples. Beyond the published research metadata, the abstract and texts of Open Access publications were crawled and published publicly accompanied by metadata derived through text analysis.

---

[1] http://scienceatlas.ijs.si/

[2] http://goo.gl/zKZbBM

[3] http://www.ewi-vlaanderen.be/

[4] http://www.ecoom.be/en/research/vabb-shw

[5] http://ewisclod3.vlaanderen.be/ewilod/html/sparql-test.html

3.1. Annotating the Research Information

The data are semantically enriched using different vocabularies, with Common European Research Information Format (CERIF)[6] and the custom-defined EWI Linked Open Data Ontology (EWILOD)[7] being the dominant ones. Beyond those vocabularies, the IWETO[8] Discipline and subject domains taxonomy were described as an ontology and were used to further semantically annotate the data.

3.2. Publishing Research Information as Linked Data

The data were initially on three databases. The first database contained data about the projects and the research groups that participated to these projects, including the promoters of the research groups who lead the project on behalf of their research group. The second one aggregates data of different universities regarding the publications they had, the authors of the publications and their affiliation. The last one contained data as a result of the text mining applied to the open accessed publications.

In order to achieve this enrichment, a D2R server[9] was used and the generated triples are loaded in a Jena TDB instance, a native high performance triple store accompanied by a FUSEKI component that exposes the triples as a SPARQL end-point accessible over HTTP. The Open Access publications were crawled and the abstract and the text were enriched and published accompanied with more information derived from the applied text analysis.

4. Visualizing Information as Linked Open Data

In order to navigate through the available published data, we implemented a set of visualizations using the published and semantically enriched data in the frame of the "*EWI Open Data: an experimental platform*" project[10]. The visualizations are available at http://ewi.mmlab.be/academic. These visualizations provide a mean to non-expert users to discover the published data and gain primary insights regarding the available information.

The challenge is twofold:

(i)   *Implementation-wise* we aimed to contribute to the development of a suite of data/schema-agnostic visualizations which can be easily applicable and transferable across different datasets. The visualizations offer an interactive interface on top of the published Linked Open Data which allows users unfamiliar with the Semantic Web technologies to explore the data and their links.

(ii)   *Context-wise* we aimed to take advantage of the semantically enriched research data and try to reveal the potential of research Linked Open Data visualizations to offer new analysis opportunities and alternative scientometrics on top of Research Information Systems relying on Semantic Web technologies.

4.1. Linked Open Data Visualizations architecture

We used the LOD Visualization Suite (LOD/VizSuite) to create the visualizations on top of the RILOD infrastructure. It is a complete stack of in-house technologies which are used to reach an end-to-end solution for creating visualizations on top of research data published as Linked Data. The visualizations are the result of a set of Linked Open Data retrieved by a triplestore and displayed to the client. Overall, LOD/VizSuite is easily customized and parameterized to be applied to different datasets.

---

[6] http://www.eurocris.org/ontologies/cerif/

[7] http://ewilod.be/ewilod/lod/0.1/ontology

[8] http://goo.gl/XviYVN

[9] http://d2rq.org

[10] http://ewi.mmlab.be

The LOD Visualization Suite consists of a number of JavaScript libraries created, around D3[11], to support the display of the different types of visualizations. Each time a user asks for a certain visualization to be displayed, a request is fired from the client application that hosts the visualizations. The request is directed to the intermediate Java-based implementation that handles the request and retrieves the corresponding data using a dedicated DataTank instance.
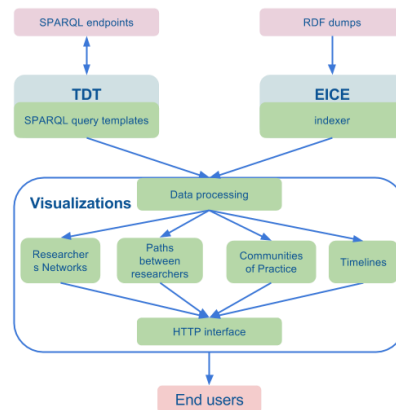
Fig. 1. The LOD/VizSuite architecture.

The DataTank[12] is a RESTful (Linked) Open Data management system which publishes data on the Web by transforming them in an HTTP API. The DataTank reads the data out of these structures and publishes them on the Web using a URI as an identifier. It can then provide these data in any format a user wants, no matter what the original data-structure was. Additionally, the DataTank describes its datasets using DCAT[13] and VoID[14]. One of the source types that the DataTank supports is the SPARQL templates. Parameters can be passed to the SPARQL template at request time, which replace placeholders to construct a valid SPARQL query .This way, a data publisher can easily publish triples, constructed by a SPARQL query, as a single URI.

In our case, different SPARQL queries were published as SPARQL templates at the dedicated DataTank instance. The Java-based backend implementation, based on the received request, instantiates the corresponding SPARQL template's URI parameters and fires a request to retrieve the data from the triplestore. Then the Java-based backend processes the returned data and returns the resulting data to be displayed.

5. Research Information visualizations

Visualizations of research information published as Linked Open Data can support non-expert users (e.g., government, enterprises, researchers, and the public)in discovering the data, identifying the links, analysing the information and gaining insights. They can show the existence and the nature of research networks, the collaborations, the emerging communities of practice within scientific disciplines, the evolution of research over time and the corresponding research trends.

In this case, The LOD/VizSuite is used to provide different novel visualizations implemented on top of the RILOD dataset. In the remaining of this section, the different possible visualizations are presented.

----

[11] http://d3js.org/

[12] http://theDataTank.com

[13] http://www.w3.org/TR/vocab-dcat/

[14] http://www.w3.org/TR/void/

5.1. Visualizing research networks

The visualizations that represent the research network around a researcher relies on the researchers' publications and his/her co-authors. The network is represented using weighted graphs with center where the center is the researcher whose research network is examined, as in Fig. 2. The size of the persons' nodes depends on the total number of their publications while the papers' nodes and the thickness of the links does not change in this case.
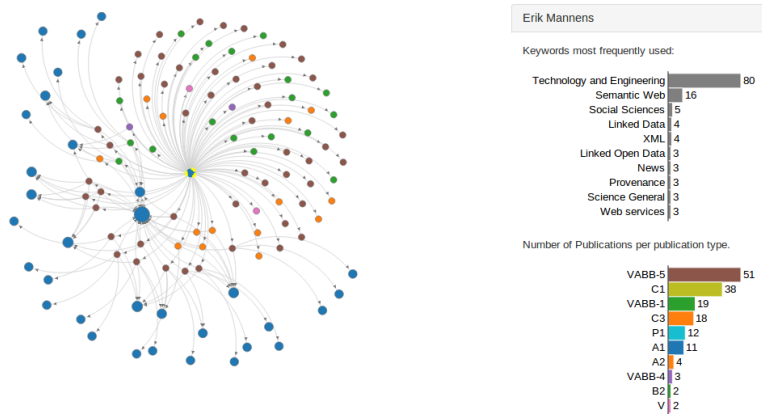

Fig. 2. The research network around Erik Mannens.
Nodes in blue color represent his co-authors, nodes in other colors represent papers he co-authored.

Beyond the aforementioned research networks around a certain researcher, the LOD/VizSuite of RILOD visualizes the collaborations between a research group and other groups, as demonstrated in Fig. 3. This type of visualizations is represented using weighted graphs with center, too. However, the nodes' size and the links' thickness are determined differently. In this type of visualizations, the more projects a research group has, the stronger its presence on the visualization, thus it has a bigger node. The stronger the collaboration the research group has with another research group, the stronger their link is, thus they have thicker edge to interlink them.
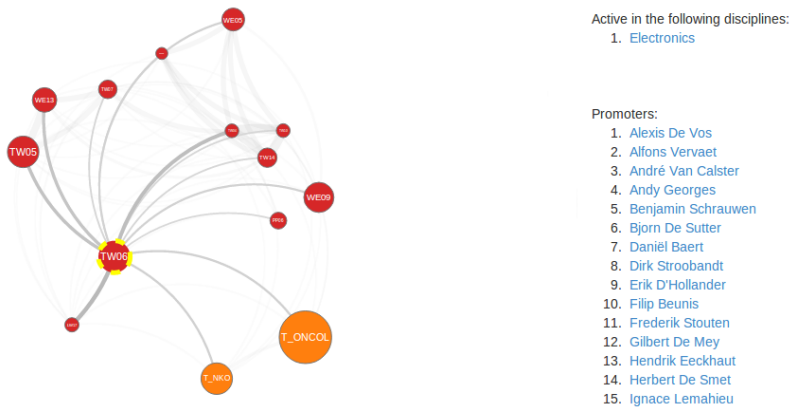

Fig. 3. The research network around the department of Electronic Engineering, its promoters and the discipline this research group is active.

5.2. Visualizing communities of practice

The visualizations that represent communities of practice around a discipline rely on the combination of the researchers' publications and the projects they participate, taking also into consideration the discipline they are

active in and are represented using weighted graphs. To the best of our knowledge, there are no previous attempts to visualize the performance of a research group in terms of publications and projects together in a single graph.
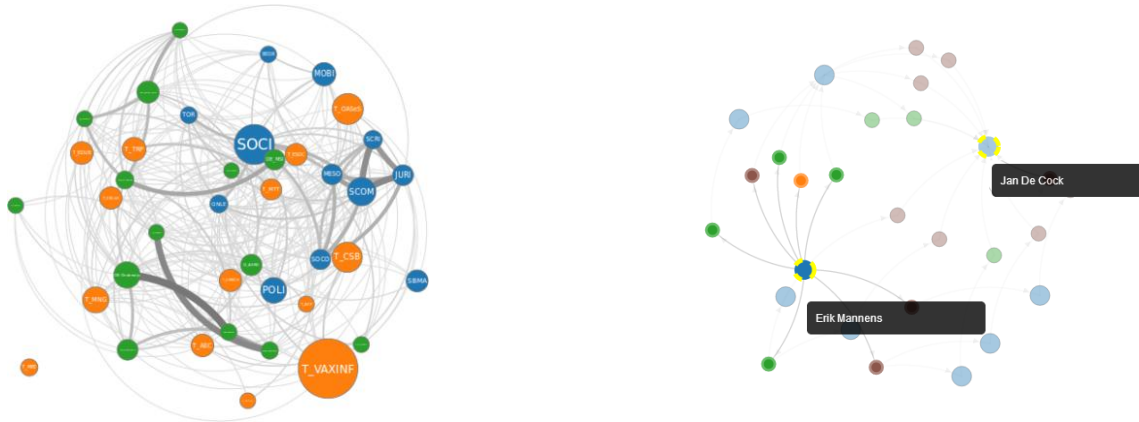

Fig. 4. (a) Visualization of the Social science community of practice;
(b) visualization of the alternative paths between the researchers Erik Mannens and Jan De Cock.

In more detail, the available visualizations capture the research groups that are active in a certain discipline, the links between these research groups, and the strengths of their collaborations, as in Fig. 4 (a). The number of projects in a certain discipline of a certain research group determines the presence of a research group; the bigger the node size, the more projects the research group has. On the other hand, the strength of the collaborations in terms of co-publications between the different research groups is depicted by the strength of the edge which connects the two research groups; the thicker the line, the more co-publications the two research groups have.

5.3. Visualizing paths between researchers

Additionally to the aforementioned research networks, the LOD/VizSuite provides another type of visualizations that display alternative paths that interlink two researchers considering their publications and their co-authors as intermediate steps. In order to get those paths, the LOD/VizSuite uses the Everything Is Connected Engine EICE [3] whose function relies on an optimized path-finding algorithm [7]. Fig 4(b) presents such visualization.

5.4. Visualizing timelines

Finally, the last type of available visualizations monitors the evolution of research over time based on corresponding timelines. The timelines present the number of publications over year per university or publication type for instance or any other parameter.
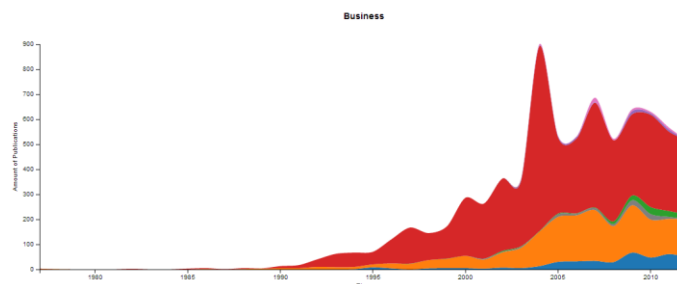

Fig. 5. Timeline visualization demonstrating the number of papers per university in the Business discipline.

6. Evaluation

We evaluated the LOD Visualization Suite at iMinds The Conference 2013 in Brussels[15] a yearly gathering for people active in several aspects of research and digital innovation. Thus, 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). Test users evaluated the tool in two ways: a user test and a questionnaire. Sixteen (16) test users were selected to participate to the observation and they received no information about the LOD/VizSuite in advance. They were asked to execute two assignments and to fill in a questionnaire afterwards. The questionnaire was completed also by another twenty (20) additional test users who only attended a short demo presentation.

6.1. Observation's results

We analyzed how the users explored and perceived the visualizations by measuring the Positive Predictive Value, how often users perceived a distinctively evident node or edge compared to the others displayed, and True Positive Rate, how good the displayed visualization is at prominently representing the prevalent vertices and edges.

The users were asked to complete two assignments and interpret the displayed visualizations considering the differences of the nodes and links size: In assignment 1, a network of research units which are active in a certain discipline was displayed as a weighted graph. All test users perceived spontaneously the differentiation of the vertices size in assignment 1. In assignment 2, the collaborations of a research unit with others were displayed in the form of a weighted star graph. 12/16 (75%) of the test users understood spontaneously the evidently prevalent vertices and 9/16 (56%) the evidently prevalent edges in the second assignment.
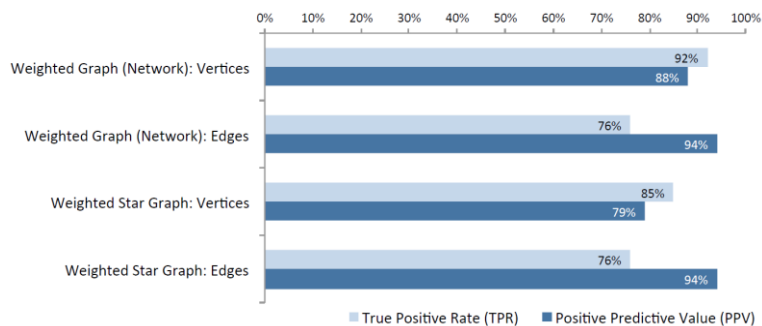

Fig. 6. True Positive Rate and Positive Predictive Value of the visualizations.

As shown in Fig. 6, 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 1st assignment was 88% for the vertices and 94% for the edges. In the 2nd assignment the positive predictive value was exactly the same for the vertices and almost the same (85%) for the edges. Thus, 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 1st assignment and equal to 79% in the case of the 2nd assignment. Additionally, the users better interpret the scaling in the cases of more dense graphs as in the 1st assignment (average density 0.30), compared to the 2nd assignment (average density 0.14). Overall, the visualizations proved to better display the prevalent nodes (the true positive rate is equal to 92% for the 1st

_____

assignment and 85% for the 2nd assignment), compared to the prevalent edges (the true positive rate is equal to 0.76 in both assignments).

6.2. Survey's results

In total the respondents found relevant insights about the researchers they were looking for (medians=4). Almost all of them agree or strongly agree that after a learning period, users should benefit of the visualizations (median = 4, agree) and many of the respondents agreed that they can learn quickly to interpret the visualizations (median = 4, agree). Finally, the test users agree on the fact that they would use the workflow to explore opportunities for collaborations (median = 4, agree) and many of the respondents agree that the visualizations facilitate the exploration of the published Linked Open Data (median = 4, agree).

7. Conclusions

These visualizations aim to offer some primal insights of the available data published as Linked Open Data to the RILOD triplestore and allows non-expert users to discover opportunities raised combining the integrated and semantically annotated data. Therefore, the usefulness of Linked Open Data and Semantic Web technologies is revealed to non-expert users. In this use case, it is highlighted how publishing and semantically annotating the research information helps in improving the quality of the data, the diversity of the information and the integration of knowledge. Considering these visualizations, the potential of Linked Open Data to offer new analysis methods in order to monitor, analyze and evaluate the research activity in a larger scale is revealed.

Acknowledgements

References

1. Debruyne C, De Leenheer P, Spyns P, Van Grootel G, Christiaens S. Publishing open data and services for the Flemish research information space. In: ER Workshops.; 2011. p. 389-394.
2. Spyns P, Van Grootel G: Realising the Flanders research information space. In: Proceedings of the 2011th Confederated International Conference on On the Move to Meaningful Internet Systems. OTM'11, Berlin, Heidelberg, Springer-Verlag ; 2011. p. 138-141.
3. Vander Sande M., Colpaert P, Van Deursen D, Mannens E, Van de Walle R: The datatank: an open data adapter with semantic output. In: 21st International Conference on World Wide Web, Proceedings; 2012.
4. Hu Y, Janowicz K, McKenzie G, Sengupta K, Hitzler P: A linked-data-driven and semantically-enabled journal portal for scientometrics. In Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J., Aroyo, L., Noy, N., Welty, C., Janowicz, K., eds.: The Semantic Web ISWC 2013. Volume 8219 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013. p. 114-129.
5. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z: Arnetminer: Extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '08, New York, NY, USA, ACM; 2008. p. 990-998.
6. De Vocht L, Mannens E, Van de Walle R, Softic S, Ebner M: 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, CEUR-WS; 2013. p. 21-24.
7. De Vocht L, Coppens S, Verborgh R, Vander Sande M, Mannens E, Van de Walle R: Discovering meaningful connections between resources in the Web of Data. In: Proceedings of the 6th Workshop on Linked Data on the Web; 2013.