# Visual Analysis of a Research Group's Performance thanks to Linked Open Data

Oscar Peña, Jon Lázaro, Aitor Almeida, Pablo Orduña, Unai Aguilera, Diego López-de-Ipiña

Deusto Institute of Technology - DeustoTech, University of Deusto
Avda. Universidades 24, 48007, Bilbao, Spain
{oscar.pena, jlazaro, aitor.almeida, pablo.orduna, unai.aguilera, dipina}@deusto.es

**Abstract.** Managing data within a research unit is not a trivial task due to the high number of entities to deal with: projects, researchers, publications, attended events, etc. When all these data are exposed on a public website, the need to have it updated is fundamental to avoid getting an incorrect impression of the group's performance. As research centres websites are usually quite static, external documents are generated by managers, resulting in data redundancy and out-of-date records. In this paper, we show our efforts to manage all these data using Labman, a web framework that deals with all the data, links entities and publishes them as Linked Open Data, allowing to get insightful information about the group's productivity using visual analytics and interactive charts.

## 1 Introduction

Managing metadata effectively within a research unit is an ambitious goal, as information systems need to deal with the relationships among the entities that form the organization's data model: projects, publications, researchers and project managers, topics, etc. Most research groups expose their data using a well known Content Management System (CMS) such as Joomla![1], WordPress[2] or Drupal[3]. Nonetheless, in order to extract valuable knowledge from all those data, external tools are needed to perform data analysis techniques. Exporting data in easy to handle formats from the CMS's databases usually leads to the creation of external documents which store data that will later be analysed.

This common situation has the following drawbacks: external documents (e.g., CSV, spreadsheets, text files, etc.) cause data redundancy, resulting in data quality, completeness and updating issues. This gets worse when investigators have their own personal pages (outside the system) where they show the achievements of their researching careers, funding data is managed by the accounting department and so on. When data needs to be updated in different

[1] http://joomla.org/

[2] http://wordpress.com/

[3] http://drupal.org/

systems, the expected outcome is that at some point data is going to be outdated somewhere, thus leading to errors when trying to get the whole picture of a research unit's performance.

Therefore, we present our efforts towards managing our research group's data, avoiding redundancy, improving quality and sharing the data in a standardized and interoperable way. Labman (Laboratory Management) is presented as a tool to manage all these data, publishing them as Linked Open Data. Linked Data allows to uniquely identify each entity instance with an URI, encouraging the creation of relationships among instances in order to discover patterns and insights in a dataset. Labman is a web application developed in Python using Django[4], and is Open Sourced on its Github's repository page[5], where it can be downloaded and contributed to. Labman is developed to substitute a previous Joomla! plugin developed at the research unit to publish publication data as RDF [1], thus overtaking the previously mentioned limitations.

This paper is structure as follows: First, we discuss similar efforts in section 2. Next, section 3 elaborates on the benefits of publishing information as Linked Data. Section 4 exhibits how patterns and knowledge can be extracted thanks to visualization techniques. Finally, conclusions and future work are addressed in section 5.

## 2 Related work

Even though some plugins have been developed to publish data stored within CMS systems as RDF (Resource Description Framework) files and RDFa metadata[6], they lack the ability to both make it accesible through a SPARQL endpoint (not allowing complex queries from external entities) and the advantages of publishing them following the Linked Data principles.

Research metadata visualization has also been studied by works such as [2] and [3], where authors use techniques from the visual analytics area to extract insights of research evolution in the studied cases. However, these works do not take the interlinking advantages of semantic descriptions, working with static dumps of database data.

The efforts of iMinds Multimedia Lab [4] demonstrates the potential insights that visual analytics provide when analysing research status on a country-level basis (i.e., applied to the whole research system of Belgium), publishing more than 400 million triples. Whereas users can get a full picture of the nation's research status, it does not substitute the information systems of the individual research centres.

ResearchGate[7] is a social networking site for scientist and researchers to share their work, providing metrics ands statistics to show their performance.

[4] https://djangoproject.com/

[5] https://github.com/OscarPDR/labman_ud

[6] http://rdfa.info/

[7] http://www.researchgate.net/

The focus is set on individuals to promote their work, whereas our proposal focuses on providing information on a research unit level basis.

Linked Universities, according to the definition on their website[8] "*is an alliance of european universities engaged into exposing their public data as linked data*". Specially focused on sharing educational data (e.g., courses, educational materials, teachers information, etc.), it also promotes the publishing of research and publication-related data. Linked Universities highlights the needs to have common shared vocabularies and tools to allow interoperability among how people access information about different institutions. Labman takes the recommendations from this alliance at its own core to avoid loosing the benefits provided by shared standards and vocabularies.

Finally, VIVO[9] is a huge project that provides an Open Source semantic web application to enable the discovery of researchers across institutions. VIVO allows any university to manage their data, and publish it using the VIVO ontology. VIVO is specially used among American universities.

## 3 Publication of resources as Linked Open Data

Linked Data (LD) is a series of principles and best practices to publish data in a structured way, encouraged by Tim Berners-Lee and the W3C [5]. LD is built over web standards such as HTTP, RDF and URIs, in order to provide information in a machine readable format. Every resource has its own URI, becoming a unique identifier for the data entity through all the system, thus avoiding data redundancy. Should somebody decide to extend the description of a given resource in its own dataset, both resources can be linked using the *rdfs:seeAlso* property, addressing that both resources refer to the same conceptual entity. The use of *rdfs:seeAlso* over *owl:sameAs* is preferred due to the semantic meaning difference between these properties: the former links two resources which refer to the same entity (maybe through different vocabularies), whereas the later connects two resources described in quite a similar way in different datasets.

The implicit linkage between resources in LD also allows to interconnect resources among them, e.g., a research project with the descriptions of people working on it and the related articles published as the outcomes of the study. Although this feature can also be achieved through plain relational database models, LD allows to connect references to external datasets, so complex queries can be performed in SPARQL, avoiding the potential headaches of joining consecutive SQL sentences and the need of having all the data in our system.

The "*Open*" term in Linked Open Data indicates that is freely available to everyone to use and republish data as they wish, without copyright and patent restrictions. All the information published on Labman is of public domain by default, making it freely consumable through its SPARQL endpoint. However, there is an option to mark a certain's project funding as *private*. If marked, this financial information will be used for the generation of top level visualizations

[8] `http://linkeduniversities.org/`

[9] `http://www.vivoweb.org/`

(those which give a full view of the unit's performance), but no funding charts will be rendered for that specific project and the funding amounts triples will not be generated.

### 3.1 Managing data within Labman

To encourage the adoption of Labman among the Semantic Web community, we have used well known vocabularies to describe the data of the different entities in our data model. The Semantic Web for Research Communities (SWRC) [6] ontology has been extended to provide financial information about research projects, together with some missing properties to link resources in our model. SWRC-FE (SWRC Funding Extension) is available for any semantic enthusiast to be used in their descriptions[10]. Researchers are mainly described using FOAF[11], while publications are defined thanks to the complete BIBO ontology[12]. Actually research topics are published using the Modular Unified Tagging Ontology (MUTO)[13], but we are considering to reference external topic datasets in the near future.

Labman stores data both in a relational database and as RDF triples (the relational database is used to increase performance and to allow non-semantic erudits to work with relational dumps). When an instance of any model is saved in Labman, a call is triggered to publish the instance and its attributes as RDF, generating or updating the referenced resource and its associated triples thanks to the rules of mapping specified for each model. Those triples are loaded into an Open Link Virtuoso[14] instance to be later on accessible through the dedicated SPARQL endpoint[15]. Semantics can be enabled/disabled on demand for a full deploy of Labman through general settings (useful when installing a local instance of Labman to get a taste of the system and easing the transition from a legacy relational database model). A single management command allows to make a full dump of the relational database and publish it as RDF triples. The list of available extra commands within labman can be consulted through the *–help* modifier of Django's *manage.py* command line feature.

To help with publications data adquisition, a Zotero[16] parser has been developed in order to extract all publication-related data and import it in Labman's system, publishing it as Linked Open Data using the previous described ontologies. Thanks to Zotero and the browser plugins, all metadata regarding a publication is extracted from well known publication indexing databases such as Web of Science, CiteSeer, Springer, Elsevier and so forth.

As the same authors may appear under slightly different names on different sites, Labman implements an author alias algorithm to perform term disam-

[10] `http://www.morelab.deusto.es/ontologies/swrcfe`

[11] `http://www.foaf-project.org/`

[12] `http://bibliontology.com/`

[13] `http://muto.socialtagging.org/core/v1.html`

[14] `http://virtuoso.openlinksw.com/`

[15] `http://www.morelab.deusto.es/labman/sparql`

[16] `https://www.zotero.org/`

biguation and apply the corresponding substitutions. This simple algorithm uses Python's difflib library[17] to compare strings (e.g., author full names), taking as input string pairs of all the author names present in Labman, and returning a similarity ratio between them (as shown in table 1. If the ratio is greater than the given threshold, both strings are sent to Labman's administrators for dissambiguation checking. If the match is approved, the *incorrect* author name from the pair is assigned as an alias of the valid name. A periodic background task unlinks all the referenced triples to the invalid alias, and assigns them to the correct author resource.

**Table 1.** Character sequence similarity

| Sequence #1 | Sequence #2 | Similarity ratio |
|---|---|---|
| Diego López de Ipiña | D. Lopez-de-Ipiña | 0.700 |
| E. Fernández | F. Hernández | 0.879 |
| Oscar Pena | Oscar Peña del Rio | 0.621 |

## 4 Understanding research data through visualizations

The adage "*A picture is worth a thousand words*" fits perfectly when visualizing research related info, as the huge amounts of data related to projects, funding agencies, organizations, researchers and so forth makes them perfect candidates to be rendered using visual representations, instead of displaying all the information in text form without highlighting underlying connections.

In order to access and interact with the visualizations through any web browser, web graphics libraries such as Google Charts[18], d3js[19] and sigma.js[20] have been used. Charts and graphs are rendered on the screen using JavaScript, with data extracted from Labman using Python.

Due to the interlinked nature of Linked Open Data, most visualizations showing linked entities are rendered as graphs and linked nodes. Visualizations are available on the *Charts* section[21] and on the extended information subsections of the webpage.

### 4.1 Funding

Project managers and principal investigators usually depend on funds to continue their research work. In a transparency effort, Labman allows to consult how much money is gathered from public entities, as displayed in figure 1.

[17] https://docs.python.org/2/library/difflib.html

[18] https://developers.google.com/chart/

[19] http://d3js.org/

[20] http://sigmajs.org/

[21] http://www.morelab.deusto.es/labman/charts/

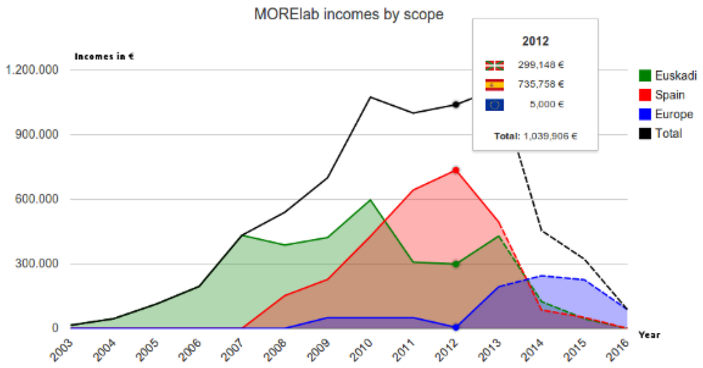

**Fig. 1.** Gathered funding timeline, by geographical scope.

Funds are provided by public administrations and organizations, usually under a named funding call. Labman also takes this information into account and allows to compare different calls' performances. For example, the principal investigator can view historical records from european FP7 and spanish INNPACTO funding calls to design the new budget strategy for the forthcoming years. Geographical scopes can be defined and related to Geoname's[22] feature classes, to classify funding call *levels* according to their effect area.

### 4.2 Project and publication collaborations

Research would not be possible without the collaborations of different researchers working together to generate new knowledge. Being able to detect research networks is a fundamental insight to have always present, together with the communities of practice our unit takes part in and the evolution and the interactions with members of external disciplines.

In figure 2, a force directed graph is selected to represent project collaborations present in the system. When hovering over an element, only the node's community is visible, allowing to consult who each person is related with. Node's size is calculated using Eigenvector centrality, a value which increases if the connection with other central nodes of the graph is relevant, and the color of each node indicates the community it belongs to. Community belonging is calculated using modularity, and a different color does not mean they do not work for the same organization, but that their connections make them beloging to a different group of interconnected people. The calculations for generating these graphs are further explained in [7]. Link weights take into account the strenght of the collaboration. For example, in projects, the more time spent working with a colleague, the stronger the connection will be, whereas the number of co-authored publications is a strong indicator of the preferences of publishing together. Collaboration edges create different triples, more accurate than the *foaf:knows* relation to analyse the relationship between two researchers.

[22] http://www.geonames.org/

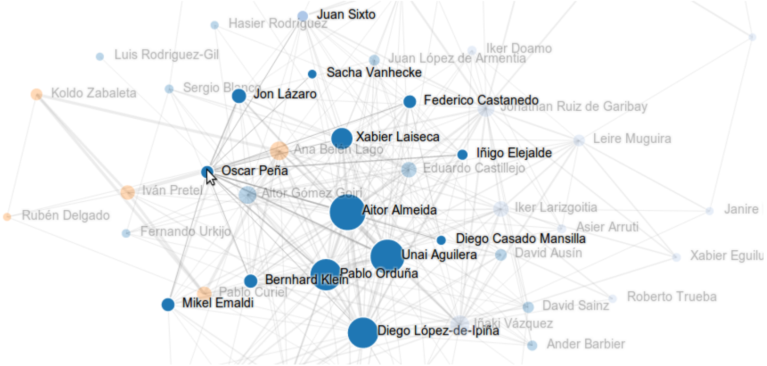

**Fig. 2.** Project collaborations of a researcher.

Figure 3 shows the egonetwork [8] of one of our researcher's (highlighted with a dotted circle), which represents the actual publication collaborations the researcher has with other members of the system. The stronger the links between two authors, the more publications they have produced together.

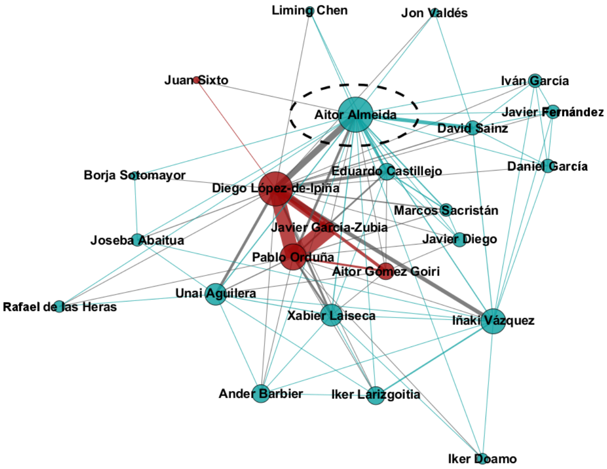

**Fig. 3.** Publications egonetwork of one of our researchers.

Sometimes relations between researchers are not explicit (e.g., two researchers have not worked together in a project or co-authored the same paper, but both of them work in the same knowledge area). In order to identify these relations we have implemented a similarity coefficient using the tags of the papers of each researcher. To ascertain the similarity of one researcher with another we have devised the following formula:

$$coef = \frac{|B \cap A|}{|A|}$$

Where A is the set of tags belonging to the base researcher and B is the set of tags belonging to the researcher which we want to compare the base researcher with. It must be taken into account that this similarity coefficient is not symmetrical. The reason is the topic similarity for a given researcher is considered within the whole of its tags, without taking into account the whole topics a related researcher works in. This situation is common for novel PhD students with a few published articles, who share all their topics with their advisors (due to their co-authorship), but senior researchers will have a broader set of areas they have worked on because of ther large research trajectory.

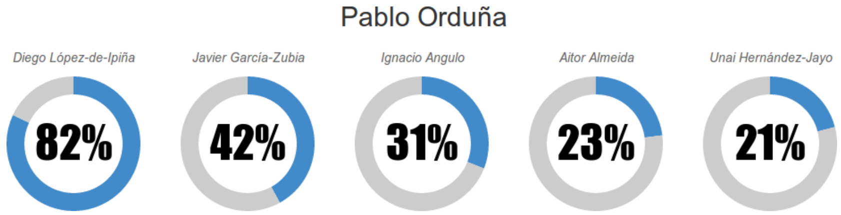

**Fig. 4.** Similar researchers to a given one using the previous normalized coeficient.

### 4.3 Topic insights

Together with the identification of research networks, knowing which topics those networks and the involved researchers are working in is fundamental to understand the most relevant areas the group is focusing on. Projects and publications are tagged with concepts in Labman, published as *dc:subject* triples using the *muto:Tag* ontology. The first obvious visual representation is to generate weighted lists (also known as tag or word clouds) of the topics used by a researcher. Figure 5 displays the topics used by one of our researchers, being the size of the tag representative of its weight (i.e., the bigger the tag, the more prolific in that area).

Research topic evolutions are also a good indicator to detect which areas the group is focused on. The historical evolution helps understanding which topics are no longer *hot* amongst researchers, and which topics have *died* to evolve into new research areas (e.g., from *Ubiquitous computing* to *Internet of Things* to *Weareable computing*). ArnetMiner [9] generates similar visualizations using automatically extracted metadata from the papers it finds for a researcher. However, many papers are not gathered, making those visualizations not to show the real status of the research.

Eventually, establishing a robust taxonomy of topics leads to the identification of interest groups and expertise hubs around topics, allowing to relate

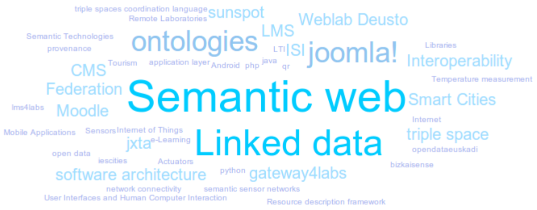

**Fig. 5.** Weighted list of the topics related to a person.

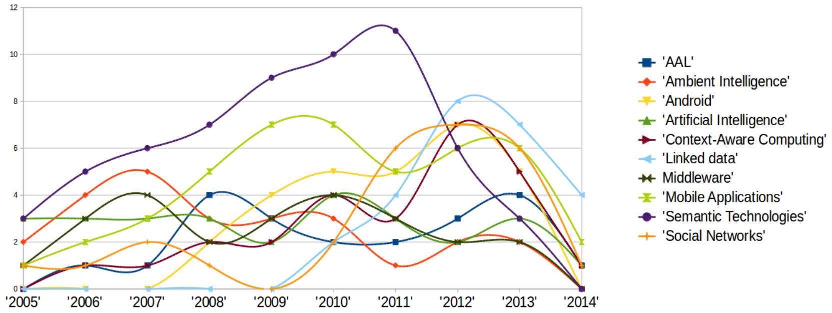

**Fig. 6.** Topic evolution of the research centre.

researchers, projects and publications automatically where no previous obvious hints were available to connect them. Describing resources using shared vocabularies and connecting to external knowledge datasets allows to create these links in a global space, opening new doors to knowledge discovery thanks to the use of Linked Data principles.

## 5 Conclusions and future work

Using dynamic visualizations of research information published as Linked Open Data, helps end-users (i.e., those consulting the information either from the research centre or visitors) to discover real connections between its members and the different entities modelled in the system. Visually representing which topics the researchers work on, who can be consulted about a certain area, and the historical collaborations within the group members can be of great help to guide the development of new project proposals, discover non-obvious potentials and address the key entities.

As future work, we will continue working on the generation of new visualizations to provide opportunities to improve the performance and strategic vision of a research centre. There is a strong continuous commitment to connect entities with external datasets, in order to evolve from a four-star to a full Linked Open Data data space, making Labman able to answer complex queries with data not present in our system, including other descriptions to the same resources available on the Web of Data. Finally, a fine detail level when well defining and describing topics will allow for deeper analysis of data, taking into consideration the evolution of topics through time and how research areas are hierarquically structured. Actually, topics are cleaned and reviewed automatically on a regular basis to improve how resources are tagged. Better data completeness will lead to more enlightening reports, so automatizing even further the data adquisition stage will benefit all users.

## 6 Acknowledgements

The research activities described in this paper are partially funded by DeustoTech, Deusto Institute of Technology, a research institute within the University of Deusto and the Basque Government's Universities and Research department, under grant *PRE_2013_1_848*.

## References

1. Mikel Emaldi, David Buján, and Diego López-de Ipina. Towards the integration of a research group website into the web of data. In *CAEPIA - Conferencia de la Asociación Española para la Inteligencia Artificial*, 2011.
2. Weimao Ke, Katy Borner, and Lalitha Viswanath. Major information visualization authors, papers and topics in the acm library. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, page r1–r1. IEEE, 2004.
3. Kevin W Boyack, Richard Klavans, and Katy Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, 2005.
4. Anastasia Dimou, Laurens De Vocht, Geert Van Grootel, Leen Van Campe, Jeroen Latour, Erik Mannens, and Rik Van de Walle. Visualizing the information of a linked open data enabled research information system. euroCRIS, May 2014. Delivered at the CRIS2014 Conference in Rome; to appear in the Procedia online CRIS2014 procs on ScienceDirect.
5. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22, 2009.
6. York Sure, Stephan Bloehdorn, Peter Haase, Jens Hartmann, and Daniel Oberle. The SWRC ontology – semantic web for research communities. In Carlos Bento, Amílcar Cardoso, and Gaël Dias, editors, *Progress in Artificial Intelligence*, number 3808 in Lecture Notes in Computer Science, pages 218–231. Springer Berlin Heidelberg, January 2005.
7. Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008. arXiv: 0803.0476.
8. Martin Everett and Stephen P. Borgatti. Ego network betweenness. *Social Networks*, 27(1):31–38, January 2005.
9. Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 990–998. ACM, 2008.