Linked Scientometrics: Designing Interactive Scientometrics with Linked Data and Semantic Web Reasoning

Grant McKenzie[1], Krzysztof Janowicz[1], Yingjie Hu[1], Kunal Sengupta[2], and Pascal Hitzler[2]

[1] University of California, Santa Barbara, CA, USA
[2] Wright State University, Dayton, OH, USA

Abstract. In this demo paper we introduce a Linked Data-driven, **S**emantically-**e**nabled **J**ournal **P**ortal (SEJP) that offers a variety of interactive scientometrics modules. SEJP allows editors, reviewers, authors, and readers to explore and analyze (meta)data published by a journal. Besides Linked Data created from the journal's internal data, SEJP also links out to other sources and includes them to develop more powerful modules. These modules range from simple descriptive statistics, over the spatial analysis of visitors and authors, to topic trending modules. While SEJP will be available for multiple journals, this paper shows its deployment to the Semantic Web journal by IOS Press. Due to its open & transparent review process, SWJ offers a wide variety of additional information, e.g., about reviewers, editors, paper decisions, and so forth.

1 Introduction

Scientometrics are playing an increasingly important role in facilitating the understanding of different research fields as well as the research topics within them. In combination with Semantic Web reasoning, Linked Data provides the principles to structure and interlink data in a way that facilitates data integration and knowledge discovery and can therefore enhance scientometric analysis. In this paper, we present a Linked Data-driven, semantically-enabled journal portal (SEJP) that currently supports over 20 different interactive analysis modules. These modules range from simple descriptive statistics such as the acceptance rate of a journal to more complex modules that provide spatial analysis and topic modeling. For instance, SEJP allows editors to visualize authors together with keywords reflecting their expertise to better select reviewers for a new paper submission. While SEJP will be deployed to other IOS Press journals in the future, this paper showcases the SEJP functionality by showing its application to the structured data from the Semantic Web journal (SWJ).[3] SWJ adopts a unique open and transparent review process. This provides a rich amount of data related to the review and publication process, including not only papers' contents, but also reviewers and their reviews, assigned editors, paper decisions, resubmission time lines, and so forth.

[3] The current SEJP version can be used at *http://sejp.geog.ucsb.edu/SWJPortal*

2 The Linked Data Portal for SWJ

2.1 Structuring and Publishing Data

SWJ employs a highly customized version of the popular Drupal content management system (CMS)[1]. All submissions, reviews, notifications, and feedbacks are contributed through the CMS, storing content in a relational database management system. The first step in developing the portal was to export all data from the database, and convert it to the Resource Description Framework (RDF) format, making use of the bibliographic ontology BIBO [2]. While most of the data accessed from SWJ can be modeled by BIBO, the ontology was extended to include aspects such as the versioning of articles (*AcademicArticleVersion*). Once the data was organized and relationships were defined (e.g., Article *hasAuthor*), a custom Java converter was constructed using the OWL API and published online via Apache Jena's SPARQL server *Fuseki*; see [3] for details.

2.2 User Interface

Once the back-end data was organized, structured, and published to the Web, a modular user interface was developed to allow visual analysis of the SWJ data. A modular approach was taken with a *plug and play* mentality, allowing the analysis modules to be separated and configured based on the particular requirements of the applications. Built through HTML5, CSS, JavaScript, D3, and ExtJS, the front-end interface for the application is light-weight and compatible with any modern W3-compatible browser. Given the separation between the back-end data and the front-end analysis modules, SEJP is able to integrate data from other SPARQL endpoints and APIs; in our case the Semantic Web Dog Food portal and Microsoft Academic Search.

3 Modules

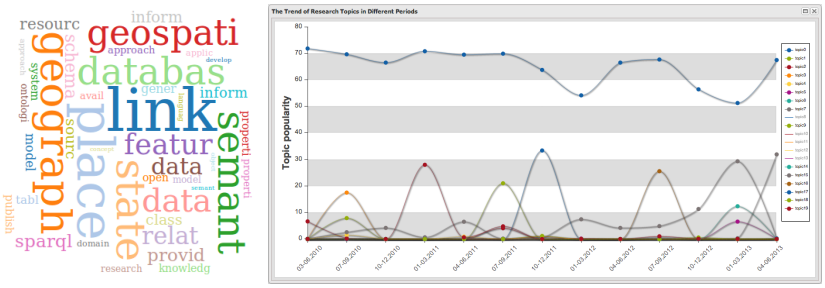
A variety of scientometric modules were developed for analyzing the SWJ data. Visual-analytic tools range from pie charts showing paper submission types to Cartograms of website visitors to edge-node graphs showing links between collaborating authors. Two of the more unique *trend* modules are discussed here.

3.1 Research Topic Trends

This module shows how the research topics contained in the SWJ trend over time. In order to construct this module, a topic modeling approach was taken to extract latent topics in papers submitted to the SWJ. First, the text for all original submissions between March 2010 and April 2013 were accessed, cleaned of standard English stop-words and non-alpha numeric characters and stemmed[4]. The submissions were then grouped by time periods (3 month are considered as

[4] Using the Snowball stemmer - http://snowball.tartarus.org

one period), combining text from all articles within this period in to one single document. This produced a total of 13 documents. Latent Dirichlet allocation (LDA) was then applied to the documents with the purpose of extracting a set number of latent topics. LDA is an unsupervised, generative probabilistic model used to infer latent topics in a textual corpus [4]. In this case, LDA is applied across the set of 13 documents and topics are discovered, represented as a multinomial distribution over words. Based on the co-occurrence of words in the corpus and a numerical value for the resulting topics, LDA produces probability values for each word in each topic and for each topic in each document. The LDA model was tested with 50, 20 and 10 topics, and 20 topics produce the most human comprehensible results for this module. An example of one of these topics is shown in Figure 1a with font-size indicating relative probability of the word existing in the topic.



(a) Word cloud showing an example topic.

(b) Research topic trending module showing how topics change over time

Fig. 1: Research topic trending module

The topics are then displayed through the user interface via an interactive line graph constructed with the JavaScript D3 charting library (Figure 1b). LDA defines each document (publications grouped by time period) as a distribution over all topics with the total probability across all topics summing to 1. Visually this is represented with time period shown on the X-axis and probabilities for each topic (multiplied by 100) shown on the Y-axis. Initially the 20 topics are color coded and shown on the chart with the option to show or hide each topic through a click-able legend on the right. Hovering one's mouse over a line produces a pop-up bubble that informs the user of the topic strength as well as the top ten words most probable to that topic.

3.2 Author-paper-keyword Hive Chart

The author-paper-keyword hive chart module (Figure 2) is a unique interactive visualization showing the relationship between authors, papers, and keywords. This module has the capability to help users to discover the possibly hidden relations among the three distinct types of data.

Hovering over a node on the authors axis (orange) produces a one-to-many relationship on the keyword axis (green) showing all the keywords mentioned by

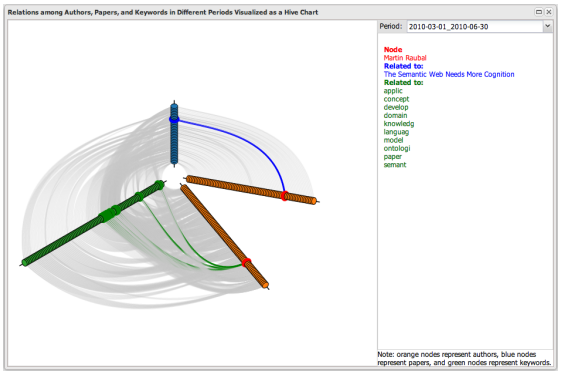


Fig. 2: Hive chart showing relationships between authors, papers and keywords.

a specific author. Additionally, there is a one-to-many relationship between the selected author and the papers (blue axis) that he or she has contributed to the SWJ. The same relationships are true of selecting any node in either the paper or the keywords axis, allowing for exploration of the data from any node. This module allows users to find authors who are concerned with similar research topics, and can also help visually discover all the coauthors of a researcher. Editors can use the module to find suitable reviewers.

4 Conclusions

This demo paper presents a Linked Data-driven, semantically-enabled journal portal for scientometrics and deploys it to the Semantic Web journal. SEJP uses a journal's internal data and also connects to other (Linked Data) sources to includes them in the analysis. Two scientometric analysis modules were discussed in the prior sections focusing on the changing of topics over time as well as the relations between authors, papers and keywords. These modules, however, are only a small subset of a suite of interactive modules developed for the portal. As development continues to progress, new modules and tools will be added, further advancing the portal's capability for scientometrics. In the near future, SEJP will be deployed to other journals as well.

References

1. Hitzler, P., Janowicz, K., Sengupta, K.: The new manuscript review system for the semantic web journal. Semantic Web **4**(2) (2013) 117–117
2. D'Arcus, B., Giasson, F.: Bibliographic ontology specification. Online: http://bibliontology.com/specification (November 2009) Last accessed 2013-5-12.
3. Hu, Y., Janowicz, K., McKenzie, G., Sengupta, K., Hitzler, P.: A linked data-driven semantically-enabled journal portal for scientometrics. International Semantic Web Conference (October 2013)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3** (2003) 993–1022