The Semantic Web Journal as Linked Data

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

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

Abstract. The Semantic Web journal implements an open and transparent review process which creates a unique bibliographic dataset. In addition to traditional publication data such as author names and paper titles, each paper in this dataset is also accompanied with a fully timestamped history of its successive decision statuses, assigned editors, solicited and voluntary reviewers, full text reviews, comments, and in many cases also the authors' response letters. This dataset presents a rich and valuable resource for a variety of studies, such as understanding the collaboration networks of scholars as well as exploring the trending topics in the field of Semantic Web. This dataset is now publicly available online as Linked Data. In this short paper, we report the availability, novelty, as well as some design considerations of this dataset.

1 Introduction

The Semantic Web journal (SWJ)[1] is an international journal focusing on research topics related to the Semantic Web, Linked Data, ontology engineering, and so forth. A unique feature of the SWJ is that it adopts an open and transparent review process [1], in which the reviews from reviewers, the authors' response letters, multiple versions of the revised manuscripts, as well as the editor's decisions are publicly available on the journal's web page. This process creates a rich dataset that can be used in a variety of studies. For example, it can be used to learn about researchers, publications, trending topics, and popular paper categories.

This paper provides a formal documentation of this SWJ dataset. Particularly, we describe the novelty of this dataset, i.e., how it distinguishes itself from other general bibliographic datasets, and why it is of interest to the Semantic Web community. The availability of this dataset, including its SPARQL endpoint, bulk download URL, as well as a scientometrics portal have been provided. We also discuss some design considerations of the dataset in this paper.

2 Novelty

A number of bibliographic datasets are available on the Linked Open Data (LOD) cloud. Two prominent examples are DBLP[2] and CiteSeer[3], both of which

---

[1] http://www.semantic-web-journal.net/
[2] http://datahub.io/dataset/fu-berlin-dblp
[3] http://thedatahub.org/dataset/rkb-explorer-citeseer

contain structured information about paper titles, authors, affiliations, journal (or conference) names, years, volume numbers, and so forth. While such data are already useful, they lack the important full text data, partially because of copyright limitations. The Semantic Web Dog Food (SWDF) [2] is another large structured dataset that focuses on publications from the Semantic Web community. It contains not only common bibliographic information, but also the academic roles, and some conference data browsers, such as Conference Live, have also been developed based on SWDF data.

The novelty of the SWJ dataset compared with existing bibliographic datasets is three-fold. First, it provides URL access to the full texts of the manuscripts in multiple revised versions, as well as the full texts of many reviews and response letters. Second, unlike most datasets that only contain information about the final version of a paper, the SWJ dataset provides a timeline for each submitted article. Finally, information about reviewers and editors is also openly available.

These novel features of the SWJ dataset allow new applications in *scientometrics* [3, 4]. For example, the full text data of manuscripts enable natural-language-based research topic analysis. In previous work [5], we reported on using Latent Dirichlet Allocation (LDA) to mine topics from papers. In addition, the full text of the review comments can be used to study the quality of reviews and the review process. This can help quantify and credit the contributions of reviewers. Besides, multiple revisions of the submitted manuscripts also allow us to examine how a paper is improving through the review process.

3   Design

The SWJ dataset is formalized and published following the four established Linked Data design principles [6]. The main entity types in the dataset include *paper*, *paper version*, and *person* (who can be an author, a reviewer, or an editor). The ontology organizing these entities can be accessed at: http://semantic-web-journal.com/ontology. We design URIs by using the namespace of the journal's server (http://semantic-web-journal.com/sejp/) and adding the names of the entities (e.g., a person's name, a paper's title, or a paper version's id). Although the SWJ adopts an open review process, reviewers can still choose to remain anonymous. For these reviewers, we use *salted MD5* hashes to protect their privacy. Below are four examples of the designed URIs.

- A paper author whose name is *Karel Kremer*:
  http://semantic-web-journal.com/sejp/page/karel-kremer
- A paper whose title is *Supporting the Linked Data publication process with the LOD2 Statistical Workbench*:
  http://semantic-web-journal.com/sejp/page/supporting-linked-data-publication-process-lod2-statistical-workbench
- A version of the paper above:
  http://semantic-web-journal.com/sejp/page/node/591
- An anonymous reviewer:
  http://semantic-web-journal.com/sejp/page/AnonymousReviewere6fd64b41 72acfd5a2f615c9bf7a5228

Fig. 1: Information provided at the URI of an entity, in this case a paper version.

Fig. 2: External links from a researcher to the data on SWDF and DBpedia.

The design of URIs satisfies the Linked Data principles 1) and 2). We also provide relevant information for the URI of each entity, and this implementation helps satisfy the principle 3). Figure 1 shows a fragment of the detailed information provided at the URI of a paper version. For principle 4), we link the researchers in the SWJ dataset to their information on the Semantic Web Dog Food, such as their roles in important Semantic Web conferences (e.g., ISWC and ESWC). Figure 2 shows the external information linked to a researcher.

4 Availability

Tim Berners-Lee has proposed a 5-star ranking system to evaluate the quality and availability of open data [7]. In this ranking system, data that have been simply published online (in any format) with an open license are considered as one star. On the contrary, data, which have been structured using W3C standards and have been linked to external datasets, are ranked as five star. According to this ranking system, we consider the SWJ dataset as five star since it satisfies the requirements.

Since 2014, the SWJ dataset has been officially included in the LOD diagram (Fig. 3). Moreover, we made the SWJ dataset available through a variety of open channels as listed below. A synchronization program has also been developed, and runs every 3 days to insert data about newly submitted articles into the SPARQL endpoint.
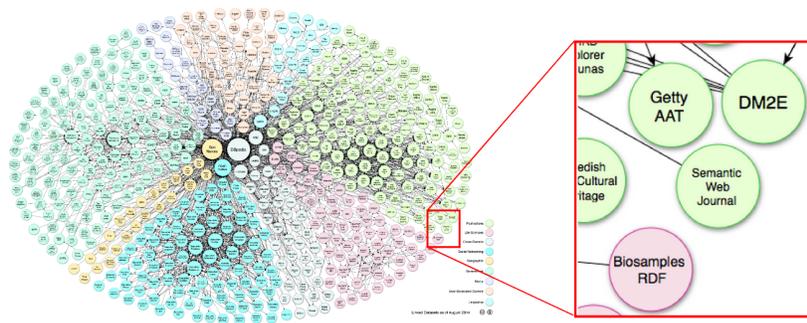
Fig. 3: The SWJ dataset in the 2014 version of the LOD cloud.

– Data registration on *datahub.io*
http://datahub.io/dataset/semantic-web-journal
– SPARQL endpoint supporting queries
http://semantic-web-journal.com:3030
– URL for simple bulk download of all triples
http://semantic-web-journal.com/SWJData/SWJ.rdf
– Linked Scientometrics Portal for non-technical end users
http://semantic-web-journal.com/SWJPortal/

5   Conclusions

This short paper documents the SWJ dataset generated from the journal's open and transparent review process. Compared with other bibliographic datasets, this dataset provides access to full manuscript texts, full revisions and decisions history, as well as the information about reviewers and editors. This dataset could be enriched with topics and named entities in future. We hope this dataset could contribute to analyzing and understanding the field of Semantic Web.

References

1. Janowicz, K., Hitzler, P.: Open and transparent: the review process of the Semantic Web journal. Learned Publishing **25**(1) (2012) 48–55
2. Möller, K., Heath, T., Handschuh, S., Domingue, J.: Recipes for semantic web dog food – The ESWC and ISWC metadata projects. Springer (2007)
3. Hood, W.W., Wilson, C.S.: The literature of bibliometrics, scientometrics, and informetrics. Scientometrics **52**(2) (2001) 291–314
4. Gao, S., Hu, Y., Janowicz, K., McKenzie, G.: A spatiotemporal scientometrics framework for exploring the citation impact of publications and scientists. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM (2013) 204–213
5. Hu, Y., Janowicz, K., McKenzie, G., Sengupta, K., Hitzler, P.: A linked-data-driven and semantically-enabled journal portal for scientometrics. In: The Semantic Web–ISWC 2013. Springer (2013) 114–129
6. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems **5**(3) (2009) 1–22
7. Berners-Lee, T.: Linked data – design issues. (2006) Available at http://www.w3.org/DesignIssues/LinkedData.html.