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 a unique open and transparent journal management process during which each submitted article is available online together with the full timestamped history of its successive decision statuses, assigned editors, solicited and voluntary reviewers, their full text reviews, comments, and in many cases also the authors' response letters. Combined with typical bibliographic data such as authors, abstracts, page numbers, issues, paper categories, and so forth, this creates a powerful, data-rich, and publicly available Linked Dataset. This dataset can be explored to learn about researchers, research fields, trending topics, and popular paper categories, but also to study the Semantic Web as a research field as well as the quality of the scientific review process in general. Where are authors and visitors coming from, what is the average yearly review load per reviewer, who is editing papers on a given topic, how many papers are accepted with minor revision during the first round of review, is the average review length a good predictor for the assigned decision, by how many degrees are reviewers and authors typically connected within the Semantic Web academic network? These and many more questions can be answered by querying the journal's publicly available SPARQL endpoint as well as using the Linked Data-driven and semantically-enabled scientometrics system developed based on the SWJ 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 related topics. It has been established in 2010 and is published by IOS Press. SWJ accepts multiple paper types such as regular research papers, surveys, dataset and ontology descriptions, as well as system and application reports.

Most importantly, the SWJ adopts an open and transparent review process [1], in which the names of the reviewers and editors are visible to the public. The full text reviews, the authors' response letters, multiple versions of the revised manuscripts, as well as the editor decisions are also publicly available on the journal's web page. This rich dataset creates a valuable timeline that does not only document the history of a paper but also of the research field as such. The Semantic Web journal exposes these data as Linked Data and makes them available through a public SPARQL endpoint and a Linked Scientometrics portal [2].

[1] http://www.semantic-web-journal.net/

In this paper, we describe this SWJ dataset and how it can be used to learn about researchers, publications, trending topics, and popular paper categories. The dataset can also be employed to study the evolution of the Semantic Web as a field and the scientific review process in general. The rest of the paper is organized as follows. In section 2, we discuss the novelty of the dataset, i.e., how the SWJ data distinguishes themselves from other general bibliographic datasets, and why the SWJ data are of interest to the Semantic Web community. In section 3, we present the design details including how we reused and extended existing ontologies, and how we followed Linked Data principles [3] to publish the data. In section 4, we describe the availability of the SWJ data, as well as a Linked-Data-driven journal portal which was developed on top of the SWJ data to perform scientometric analysis. Finally, section 5 summarizes our work.

2 Novelty and Relevance of the SWJ Dataset

This section describes how the SWJ dataset differs from existing bibliographic data and why it is relevant for the Semantic Web community and beyond.

2.1 Novelty

There are a number of bibliographic datasets available on the Linked Open Data (LOD) cloud. Two prominent examples are DBLP[2] and CiteSeer[3], both of which contain structured information about paper titles, authors, affiliations, journal (or conference) names, years, volume numbers, and so forth. While such data can be used to explore, say, collaborations between researchers (e.g., via co-authorship networks), they lack the important full text data, partially because of copyright limitations. The Semantic Web Dog Food (SWDF) [4] is a large structured dataset that focuses on publications from the Semantic Web community. It contains not only common bibliographic information, but also the academic roles (such as program committee membership) played by researchers in conferences.

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

2.2 Relevance

One important application of the SWJ dataset is in *scientometrics*, which is the science of measuring and analyzing of science [5]. In recent years, scientometrics has also begun to examine the performance of individual researchers [6–8]. A popular example is Google Scholar which provides metrics, such as the *h-index* and

[2] http://datahub.io/dataset/fu-berlin-dblp
[3] http://thedatahub.org/dataset/rkb-explorer-citeseer

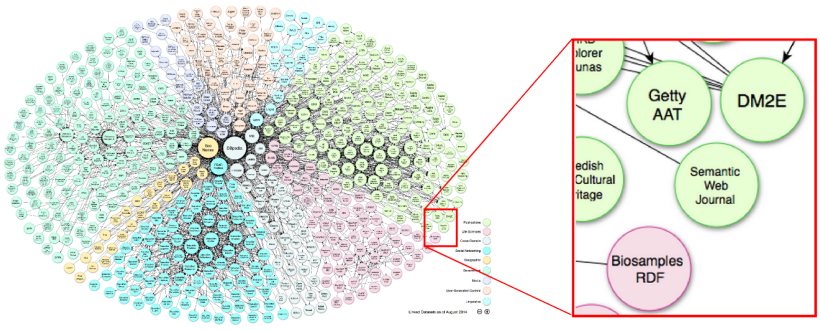


Fig. 1: The SWJ dataset in the 2014 version of the LOD cloud.

i10-index, for quantifying the productivity of individual researchers. However, the data of Google Scholar is not available as Linked Open Data. Another related work is *spatial@linkedscience* which explores interactions between researchers and academic conferences in the field of Geographic Information Science [9].

The SWJ data is useful and relevent for multiple reasons. First, the full text data contain a rich amount of information which enables natural-language-based research topic analysis. In previous work [2], we reported on using Latent Dirichlet allocation (LDA) to mine topics from paper full texts to discover trending topics. Second, 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. Third, multiple revisions of the submitted manuscripts allow us to examine how a paper is improving through the review process. The data also enable researchers to evaluate the open and transparent review process. Finally, the links among researchers (based on their co-author relations) can also be used in many applications. For example, editors can leverage those links to search for suitable reviewers for a paper while avoiding conflicts of interest. In fact, we have developed a reviewer recommendation program in another previous work [10], which recommends reviewers based on high topic similarity and far co-authorship network distance.

The SWJ dataset has been included in LOD diagram in 2014; see Fig. 1.

3 Design

Here we briefly explain design decisions and the (re)used ontologies.

3.1 Publishing the SWJ Dataset Following Standards

To follow the established Linked Data design principles [3], we first minted URIs for the entities in the SWJ dataset. The main entity types include *paper*, *paper version*, and *person* (who can be an author, a reviewer, or an editor). We designed URIs by using the namespace of the 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

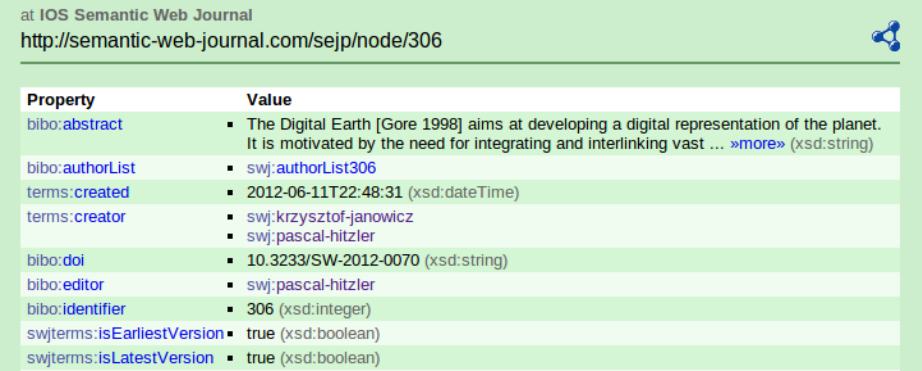
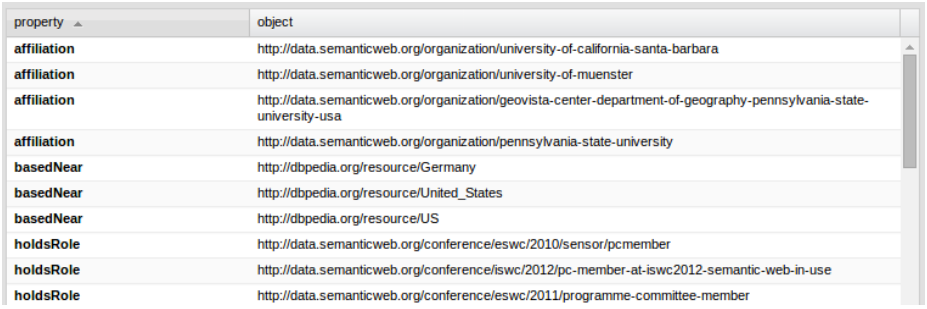


Fig. 2: Information provided at the URI of an entity, in this case a paper version.

process, reviewers can still choose to remain anonymous. For these reviewers, we use *salted MD5* hashes to protect their privacy while still assigning individual URIs to them. To slightly reduce the length of the URIs (as some papers have long titles), we also removed the stop words, punctuation marks, spaces, and special characters. Below are four examples of the designed URIs.

- A paper author whose name is *Marta Sabou*
  http://semantic-web-journal.com/sejp/page/marta-sabou
- A paper whose title is *TourMISLOD: a Tourism Linked Data Set*
  http://semantic-web-journal.com/sejp/page/tourmislod-tourism-linked-data-set
- A version of the paper above
  http://semantic-web-journal.com/sejp/page/node/273
- An anonymous reviewer:
  http://semantic-web-journal.com/sejp/page/AnonymousReviewere6fd64b4172acfd5a2f615c9bf7a5228

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). Such relevant information can be downloaded in the form of RDF, and therefore can be retrieved in a machine-readable and understandable form. Figure 2 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). Currently, a simple string matching method based on the names of researchers has been used to establish external links. In the near future, we will exploit the co-author network as an additional heuristic to improve disambiguation. In addition to the SWDF, we also link the SWJ dataset to DBpedia based on the affiliations of researchers. Figure 3 shows the external information linked to a researcher.



property	object
affiliation	http://data.semanticweb.org/organization/university-of-california-santa-barbara
affiliation	http://data.semanticweb.org/organization/university-of-muenster
affiliation	http://data.semanticweb.org/organization/geovista-center-department-of-geography-pennsylvania-state-university-usa
affiliation	http://data.semanticweb.org/organization/pennsylvania-state-university
basedNear	http://dbpedia.org/resource/Germany
basedNear	http://dbpedia.org/resource/United_States
basedNear	http://dbpedia.org/resource/US
holdsRole	http://data.semanticweb.org/conference/eswc/2010/sensor/pcmember
holdsRole	http://data.semanticweb.org/conference/iswc/2012/pc-member-at-iswc2012-semantic-web-in-use
holdsRole	http://data.semanticweb.org/conference/eswc/2011/programme-committee-member

Fig. 3: External links from a researcher to the data on SWDF and DBpedia (To see this information, go to http://semantic-web-journal.com/SWJPortal, and click the "Details" button in "People in the Semantic Web Journal" window).

3.2 Reusing and Extending Existing Ontologies

Three existing ontologies have been reused to organize and annotate the SWJ dataset, which are the Bibliographic Ontology (BIBO)[4], the Dublin Core Metadata Initiative ontology (DCMI)[5], and the Friend of a Friend ontology (FOAF)[6].

We use the BIBO ontology to annotate information related to papers, such as the *abstract*, *author list*, *issue number*, *DOI*, and so forth. While BIBO is sufficient to capture the medadata of published papers, it cannot model the publishing process which often involves the first submission, revisions, and resubmissions. To overcome this limitation, we extended the BIBO ontology with a class *AcademicArticleVersion*. This class is defined as a subclass of the *AcademicArticle* class, and is connected to the main article using *hasVersion* and *isVersionOf* from the DCMI. We also create two object relations, *hasPreviousVersion* and *hasNextVersion*, to indicate the sequence of the different versions of a paper. Figure 4 (a) provides a schema diagram to illustrate the links between the *AcademicArticle* and the extended *AcademicArticleVersion*.

The FOAF ontology was used to annotate information about researchers, such as their names. For a given paper, there is often an author order which specifies the sequence of the authors. To model such information, we employ the *rdf:Seq* class from RDFS, and use membership properties, such as *rdf:_1*, *rdf:_2*, and *rdf:_3*, to represent the author order. In addition, we also use the *creator* property from DCMI to connect each paper with all of its authors. Such a design allows users to search all papers published by a researcher without having to know the specific author order beforehand. Figure 4 (b) shows the relations between authors and papers.

Finally, we added new relations and classes to represent reviewers, their reviews, and the SWJ paper types.

[4] http://bibliontology.com/
[5] http://dublincore.org/documents/dcmi-terms/
[6] http://xmlns.com/foaf/spec/

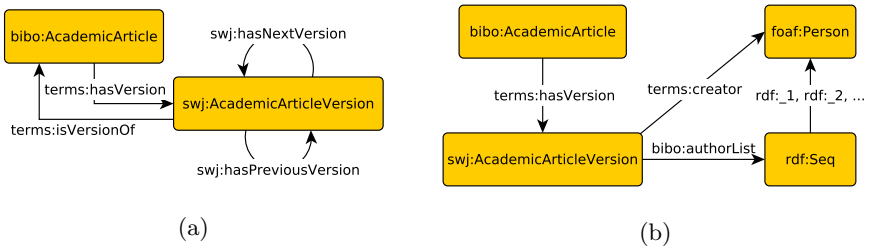


Fig. 4: Schema diagrams for (a) the relations among the *AcademicArticle* and the extended *AcademicArticleVersion* class; (b) the relations among authors, papers, and author orders.

4 Availability

Tim Berners-Lee has proposed a 5-star ranking system to evaluate the quality and availability of open data [11]. In this ranking system, data, which have been simply published online (in any format) with an open license, are considered as one star (the lowest rank among the five). On the contrary, data, which have been structured using W3C standards and have been linked to external datasets, are ranked as five star (the highest rank). According to this ranking system, we consider our SWJ dataset as five star since we have satisfied the requirements.

Moreover, we made the SWJ dataset available through a variety of open channels, which are listed below:

- 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/

The SWJ dataset also powers the journal's Linked Scientometrics portal at http://semantic-web-journal.com/SWJPortal. The portal provides more than 20 scientometric modules including overall statistical information about the journal (e.g., the acceptance rate), total number of authors, total number of papers, trending topics in different time periods, number of papers in different categories, as well as many other visualization and analysis modules. Examples for such modules include co-author networks, citation maps, and so forth. More details about this Linked-Data-driven and semantically-enabled journal portal can be found in a previous paper [2]. A screenshot of the system is shown in Figure 5.

Feedback and suggestions can be submitted through the journal's feedback page at: http://www.semantic-web-journal.net/feedback [7]. Besides, users can

[7] Login using a registered account is required.

Fig. 5: A screenshot of the Linked Scientometrics platform.

send questions to the SWJ email account, contact@semantic-web-journal.net. To ensure the sustainability of the data, the SWJ journal office is maintaining the dataset as well as the scientometrics portal. A server-side script keeps the journal and the Linked Dataset in our SPARQL endpoint in sync.

To demonstrate some interesting queries that can be submitted to the SPARQL endpoint, we provide two examples as below:

Query 1: given one paper version, finding all other related versions.

```
SELECT distinct ?paperVersion
WHERE { {swj:exampleVersion swj:hasNextVersion* ?paperVersion }
UNION {swj:exampleVersion swj:hasPreviousVersion* ?paperVersion } }
```

This query retrieves a history of a submitted paper, and can help analyze the content improvement during the review process.

Query 2: finding the top 10 reivewers who have contributed most to the journal.

```
SELECT ?reviewer (COUNT(?paperVersion) as ?count)
WHERE{ ?paperVersion swj:reviewer ?reviewer . }
GROUP BY ?reviewer ORDER BY DESC(?count) LIMIT 10
```

This query counts the number of paper versions reviewed by each reviewer. This approach is different from counting simply the number of papers, since some manuscripts may require multiple revisions.

5 Conclusions

This paper describes the Semantic Web journal's Linked Dataset which was generated from its unique open and transparent review process. Compared with other bibliographic datasets, the SWJ dataset distinguishes itself through the availability of the full manuscript texts, full revision and decisions history, as well as through the avalability of the reviews and information about reviewers

and editors. Such a rich dataset, combined with its structured and machine-understandable format, offers new opportunities for analyzing trending topics, collaboration relations among researchers, contributions from authors and reviewers, and many other interesting scientometrics. Most importantly, it offers a unique opportunity to study the scientific review process as such; something that was not possible to date. This dataset has been exposed through a variety of open channels, including a SPARQL endpoint and simple bulk download. It has also been added as dataset to the LOD cloud, and has been registered on *datahub.io*. Community discussions and feedback can be provided through the journal system and email. A synchronization server-side script has been developed to regularly update the SWJ dataset. Currently, we link out to the Semantic Web Dog Food dataset and will provide more links in the future. Finally, to ease data exploration, we offer more than 20 scientometrics modules to answer common questions about papers, topics, authors, reviewers, and editors.

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. 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
3. 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
4. Möller, K., Heath, T., Handschuh, S., Domingue, J.: Recipes for semantic web dog food – The ESWC and ISWC metadata projects. Springer (2007)
5. Hood, W.W., Wilson, C.S.: The literature of bibliometrics, scientometrics, and informetrics. Scientometrics **52**(2) (2001) 291–314
6. Braun, T., Glänzel, W., Schubert, A.: A Hirsch-type index for journals. Scientometrics **69**(1) (2006) 169–173
7. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences of the United States of America **102**(46) (2005) 16569
8. 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
9. Keßler, C., Janowicz, K., Kauppinen, T.: spatial@linkedscience – Exploring the Research Field of GIScience with Linked Data. In Xiao, N., Kwan, M.P., Goodchild, M., Shekhar, S., eds.: Geographic Information Science. Volume 7478 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 102–115
10. Hu, Y., McKenzie, G., Yang, J.A., Gao, S., Abdalla, A., Janowicz, K.: A linked-data-driven web portal for learning analytics: Data enrichment, interactive visualization, and knowledge discovery. In: LAK Workshops. (2014)
11. Berners-Lee, T.: Linked data – design issues. (2006) Available at http://www.w3.org/DesignIssues/LinkedData.html.