What is in the proceedings? Combining publisher's and researcher's perspectives

Volha Bryl[1], Aliaksandr Birukou[2], Kai Eckert[1], and Mirjam Kessler[2]

[1] University of Mannheim, Germany
{volha|kai}@informatik.uni-mannheim.de
[2] Springer-Verlag, Heidelberg, Germany
{aliaksandr.birukou|mirjam.kessler}@springer.com

Abstract. Despite many efforts for making data about scholarly publications available on the Web of Data, lots of information about academic conferences is still contained in (at best) free-text format. When available in a structured format, these data would provide an essential input for the decisions researchers, libraries, publishers, funding and evaluation bodies take every day. In this paper we present a vision for having such data available as Linked Open Data (LOD), and we argue that this is only possible – and for the mutual benefit – in cooperation between researchers and publishers. We also present a pilot project aimed at publishing data about 8,500 computer science conferences as LOD.

Keywords: Linked Open Data, linked science, research evaluation

1 Motivation: why we need more data

Data on scientific publications, authors, institutions, and conferences are widely and publicly available on the web. Moreover, there have been many initiatives aimed at publishing these data as linked and open: for example, DBLP data on publications[1] or bibliographic information on books and authors provided by the German National Library[2]. Both examples provide trusted data on publications with high coverage. A few applications have been developed to browse and query these data [1, 3], with a focus on authors, publications and research topics.

However, making sense of data about *conference proceedings* is still an issue. The conference series in which research results are published appears to be a crucial provenance dimension – along with the author metadata – based on which the research results are evaluated and trusted. The problem becomes even more complex when one takes into account the fact that conferences change name over time, and recent developments of "predatory publishing" and "fake conferences" phenomena[3]. Here are some examples of trust-related questions about conferences that various stakeholders in the academic world regularly face:

[1] D2R Server for DBLP data – http://dblp.l3s.de/d2r/

[2] http://datahub.io/dataset/dnb-gemeinsame-normdatei

[3] http://scholarlyoa.com/publishers/

– Shall I submit a paper to this conference? How good and relevant is it? What are the alternatives? *(younger researcher)*
– Shall I accept a PC membership invitation, give an invited talk, send a workshop proposal to this conference? *(more senior researcher)*
– Shall we publish the proceedings of this conference? *(publisher)*
– Is it worth sponsoring this conference? *(sponsor)*

What do you do when you face these questions? You google, read many documents and webpages, ask people, and you are never sure whether you have found all relevant data and numbers.

Data about conferences are spread across several sources in a largely chaotic and non-structured way, being duplicated multiple times. Let us take the example of the PC (Program Committee) membership: being involved in paper reviewing and other activities related to a conference organization is hard work that should be credited [2]. The data on the conference organizers and PC is also essential for a publisher when evaluating a new conference proposal. On one hand, Semantic Web Conference Ontology[4] provides a way to describe the roles of scientists in conference organization, such as “chair”, “PC member”. Any conference management system (CMS), e.g. EasyChair, contains the list of PC members. On the other hand, hardly any conference reuses such PC data through the conference lifecycle. Instead, the PC membership information is copied to appear at a conference webpage, in the call for papers (on WikiCfP, Eventseer or mailing lists), in the preface of the proceedings. Moreover, traces of such PC data are also present at author webpages and in CVs. Obviously, changes in one system (e.g. reject of a PC member to assume their role via a CMS) are not necessarily be reflected in other data sources (CfP, conference website).

The key issues to address here are data exchange between various systems involved in conference organizations and the lack of *trusted sustainable*[5] *large-scale data sources* providing detailed conference data. Currently, the LOD cloud includes several resources that contain conference metadata. Semantic Web Conference Corpus[6] includes information on major Semantic Web conferences (37) and workshops (235), therefore, providing high quality but low coverage data. Another example is COLINDA [4][7] that contains information about 15,000 conferences in a 2003–2013 time span, with main data sources being WikiCfP and Eventseer, which aggregate information from the call for papers, meaning that there is no guarantee that the events (especially workshops) actually happened and had formal proceedings.

In the following, we show how the reliability, coverage and sustainability of such data can be improved by cooperation between publishers and researchers.

[4] `http://data.semanticweb.org/ns/swc/ontology`

[5] Not many resources and tools outlive the research projects they originate from.

[6] `http://data.semanticweb.org/`

[7] `http://www.colinda.org/`

2 Filling the gap: linked open conference data

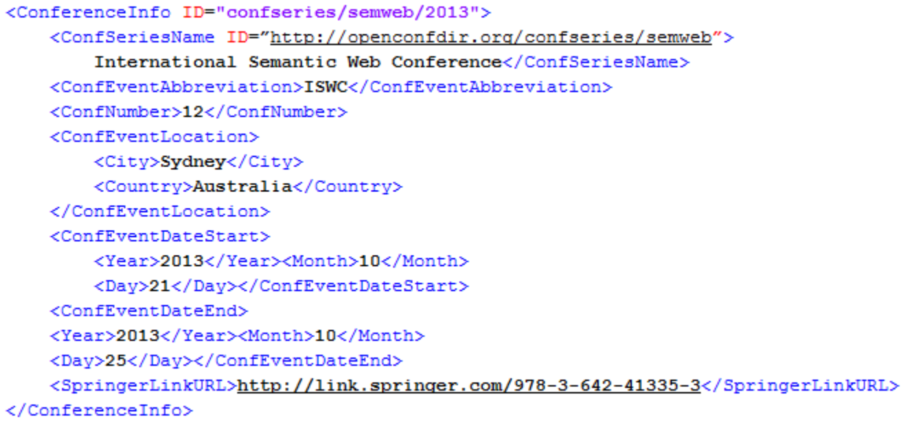
The issues outlined above motivated the launch of the Springer LOD pilot, which aims at publishing data about Computer Science conferences as a linked open dataset. The availability of such a dataset will contribute to the broader goals of publishing the scholarly data as LOD:

- *accessible science*: data about publications, authors, topics, and conferences should be easy to explore;
- *transparent science*: the data on productivity and impact of authors, research institutions, and conferences should be open and easy to analyze.

But these goal are only marginally relevant for publishers, whose primary goal is, not surprisingly, commercial benefits. So, how do the interests of publishers and researches align?

Publishing conference data as LOD would allow Springer to enrich bibliographic data provided via data services to libraries, data agencies and aggregators. This would also allow linking to other data, thus increasing the visibility of the proceedings in SpringerLink digital library. This would provide benefits for conference community (i.e. researchers): more readers, more downloads, more citations, conference submissions and participants. Moreover, Springer sees this as a way of collaborating with the research community and other stakeholders (libraries, indexing services, conference-related systems) to get new insights on the data. Also, the data would allow detecting trends in the conference business, and plan accordingly: knowing that many conferences go to Russia or China, publishers need to establish agreements with local printers, take into account customs regulations, etc. As with any LOD resource, sustainability is crucial: and in our opinion, it is directly related to the economic value the data brings. Moreover, the benefits of boosting the content usage and discoverability, and data enrichment via linking, outweigh potential profits from selling these data.

The pilot has started in 2013 and is ongoing. In the conference dataset that will be made available as a result of the pilot (later in 2014), for each conference the following information is provided: conference series name and ID; conference ID, acronym, and number in the series; city country, start and end dates. See Figure 1 for an internal XML representation. The starting point are the conference data that are present in the subtitles of the proceedings, i.e. in a free-text format: e.g. *"12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I"*. In the pilot, the Springer internal conference data management system was extended with a module that extracts and structures this information from the subtitles. Then, the quality of the extracted data is manually assured with the help of an interactive GUI, following the same philosophy for the data quality standards as the one of DBLP. The resulting conference data are stored in a database, which makes their conversion to RDF straightforward. As the example shows, the data contains some fields, e.g., conference acronym, number, city, country, link to the proceedings, which are either not available in COLINDA or the Semantic Web Conference Corpus or available there in free text form.



```
<ConferenceInfo ID="confseries/semweb/2013">
    <ConfSeriesName ID="http://openconfdir.org/confseries/semweb">
        International Semantic Web Conference</ConfSeriesName>
    <ConfEventAbbreviation>ISWC</ConfEventAbbreviation>
    <ConfNumber>12</ConfNumber>
    <ConfEventLocation>
        <City>Sydney</City>
        <Country>Australia</Country>
    </ConfEventLocation>
    <ConfEventDateStart>
        <Year>2013</Year><Month>10</Month>
        <Day>21</Day></ConfEventDateStart>
    <ConfEventDateEnd>
    <Year>2013</Year><Month>10</Month>
    <Day>25</Day></ConfEventDateEnd>
    <SpringerLinkURL>http://link.springer.com/978-3-642-41335-3</SpringerLinkURL>
</ConferenceInfo>
```

Fig. 1. Data about conferences: example

Currently, the data for 8,500 conferences (which correspond to around 2,000 conference series) published in LNCS, LNAI, LNBI, LNBIP, CCIS, IFIP-AICT and LNICST series since 1973 was processed following the above procedure. As every year 650 new conferences are published in these series, the information about them will be added to the system, structured and exported into RDF. The data are curated at publisher's end, using well-established (over the course of the last 40 years) processes: the process of producing metadata for SpringerLink was augmented with an additional step, during which the conference metadata is extracted and its quality is assured. Services such as scholarly search engines will be able to use the conference data directly from SpringerLink. The very same conference metadata will be then published as LOD, in RDF format. Such separation of data from formats allows for adding third party LOD conference data (for conferences not published by Springer) in the future.

In the future we plan to provide richer metadata that includes the number of submitted and accepted papers, acceptance rates, information on the best paper awards, PC and chairs, co-located workshops, links to CORE rankings[8], etc. Moreover, in the future the data would go beyond the computer science scope, extending to approximately 350 conferences published annually by Springer in other disciplines.

According to the internal Springer statistics, the 8,500 conferences contain almost 300,000 articles published in the proceedings, and slightly over 300,000 distinct authors contributing to the papers. Making publication and author metadata available is not the focus of the current stage of the pilot, but such information can be provided in the future by linking to other datasets, such as DBLP. Linking to citation figures (e.g. from CrossRef[9]) and ORCIDs will further enrich the data.

[8] http://core.edu.au/index.php/categories/conference\%20rankings

[9] http://www.crossref.org/

3 How to move further?

The result of this initial data publishing stage is a well-structured carefully maintained conference dataset, which can be interlinked with other datasets (DBLP or national libraries' data, GeoNames and DBpedia for locations, etc.) and used in applications. However, the initial data publishing stage will hardly go any further unless both researchers and publishers actively participate in providing more data, linking them, and developing new applications supporting the questions we posed in the introduction.

One example of application is based on the Rexplore [3] tool with its focus on *sensemaking* tasks: Rexplore combines statistical analysis, semantic technologies and visual analytics, and allows answering complex queries to make sense of scholarly data. Fetching a conference dataset into Rexplore and linking it with the publication datasets and the topic ontology the tool uses, would allow analyzing how the focus and main topics of a specific conference series were changing over time, how "good" the conference is in terms of citations, top researchers publishing there or involved in its organization, etc.

Another application is using the conference data during the conference lifecycle. Once entered in a CMS, the data about PC membership could be exported[10] to become part of LOD cloud and then displayed on the website in one of n standard ways (e.g. using specific plugins), or be included in the preface of the proceedings, various conference apps, etc. Such coordination between researchers and publishers would prevent data duplication and enable data reuse.

4 Acknowledgments

This work has been supported by the LOD2 and DM2E EU FP7 projects. We thank Max Schmachtenberg, University of Mannheim, for providing the metadata on the scholarly domain in LOD, and Markus Richter for developing the Springer conference data management tools.

References

1. Diederich, J., Balke, W.T., Thaden, U.: Demonstrating the semantic GrowBag: Automatically creating topic facets for FacetedDBLP. In: JCDL'07. pp. 505–505. ACM (2007)
2. Ley, M.: DBLP – some lessons learned. PVLDB 2(2), 1493–1500 (2009)
3. Osborne, F., Motta, E., Mulholland, P.: Exploring scholarly data with Rexplore. In: International Semantic Web Conference (1). pp. 460–477 (2013)
4. Softic, S., Vocht, L.D., Mannens, E., de Walle, R.V.: COLINDA – conference linked data. Submitted to Semantic Web Journal (2013), available at http://semantic-web-journal.net/content/colinda-conference-linked-data

[10] In fact, in the OCS (Online Conference Service) conference management system – `http://ocs.cs.uni-dortmund.de` – such an export already exists.