cLODg - Conference Linked Open Data Generator

Anna Lisa Gentile[1]* and Andrea Giovanni Nuzzolese[2]

[1] Department of Computer Science, University of Sheffield, UK
[2] Semantic Technology Lab, ISTC-CNR. Italy
a.gentile@sheffield.ac.uk, andrea.nuzzolese@istc.cnr.it

Abstract. In this paper we describe cLODg (conference Linked Open Data generator), a set of tools to collect, refine and produce Linked Data to describe a scientific conference and its publications, participants and events. cLODg is an open source project, which has the aim to encourage conference metadata publication and foster collaborative efforts in this direction between researchers and publishers.

1 Introduction

A good practise in the semantic Web community is to encourage the publication "eating our own dog food" [7]. The main example is the *Semantic Web Dog Food*[3] (SWDF), a corpus that collects linked data about papers, people, organizations, and events related to academic conferences. Currently, all main semantic Web conferences and related events publish their data as linked data on SWDF, but for many other conferences, events and publication venues information is still not available in a structured and linked form.

There are two main challenges to pursue this task: (i) the knowledge of available vocabularies to represent the data and (ii) the availability of tools to ease the task of data acquisition, conversion, augmentation, verification and finally publication.

In this work we present cLODg (conference Linked Open Data generator), a tool that provides a formalized process for the conference metadata publication workflow. cLODg is an Open Source solution currently under development[4], which has been used to gather and publish metadata for ESWC2014[5] and ESWC2015[6].

The main contributions of cLODg is a formalized, open source workflow which provides: (i) Facilities to gather/convert conference data from different source formats. (ii) The possibility to represent such data using different vocabularies.

* Part of this research has been sponsored by the EPSRC funded project LODIE: Linked Open Data for Information Extraction, EP/J019488/1

[3] SWDF: http://data.semanticweb.org

[4] https://github.com/AnLiGentile/cLODg

[5] http://2014.eswc-conferences.org/

[6] http://2015.eswc-conferences.org/

(iii) Facilities to involve conference participants in the loop and allow the collection of additional information and the verification of automatically generated data. (iv) Facilities to represent additional and non-conventional information, not currently captured by existing vocabularies, using Ontology Patterns[7]. (v) The serialization of data in different output formats, including efficient representations for mobile app consumption.

The main advantage is the open source and modular nature of the work, with the primary goal to encourage the collaboration between researchers and publishers towards increasing the availability of structured scholarly data.

2 State of the art

The first considerable effort to offer comprehensive semantic descriptions of conference events is represented by the metadata projects at ESWC 2006 and ISWC 2006 conferences [10], with the Semantic Web Conference Ontology [9] being the vocabulary of choice to represent such data.

Increasing number of initiatives are pursuing the publication about conferences data as Linked Data, mainly promoted by publishers such as Springer[8] or Elsevier[9] amongst many others. For example, the knowledge management of scholarly products is an emerging research area in the Semantic Web field known as Semantic Publishing [11]. Semantic Publishing aims at providing access to semantic enhanced scholarly products with the aim of enabling a variety of semantically oriented tasks, such as knowledge discovery, knowledge exploration and data integration. The Semantic Publishing challenge [8] is a breakthrough in this direction. Its objective is assessing the quality of systems that extract meaningful metadata from scholarly articles and represent them as RDF. Similarly to the Semantic Publishing challenge, the *Jailbreaking the PDF* initiative [5] is aimed at creating a formal flexible infrastructure to extract semantic information from PDF documents by combining existing solutions and tools in order to extract data from raw PDFs and convert data to domain-specific annotations.

Despite these continuous efforts, it has been argued that lots of information about academic conferences is still missing or spread across several sources in a largely chaotic and non-structured way and a viable solution is a strong cooperation between researchers and publishers [4].

3 The cLODg tool - publishing Conference Semantic Data

Conference metadata collected from different unstructured and semi-structured resources must be expressed with appropriate vocabularies to be exposed as linked data. cLODg currently implements two data representations: one that

[7] `ontologydesignpatterns.org`

[8] `http://lod.springer.com/wiki/bin/view/Linked+Open+Data/About`

[9] `http://data.elsevier.com/documentation/index.html`

strictly follows the Semantic Web Conference ontology [9] and one which enriches the representations with the SPAR ontologies[10] and novel ad-hoc ontology patterns to also capture social data[11]. Nevertheless, cLODg architecture allows the addition of other representations in the future. The Semantic Web Conference ontology [9] is one of the vocabulary of choice to describe academic conferences. The SWC ontology extends and combines existing widely accepted vocabularies (i.e., FOAF [3], SIOC [2], Dublin Core [1]) to provide a reference model to describe typical actors in an academic conference, such as accepted papers, authors, their affiliations, organizing committee and all other roles involved. Choosing the SWC ontology as reference vocabulary makes sure that data is homogeneous with the SWDF corpus. The additional concepts, properties and axioms that we introduce are further described in [6] and the resulting ontology is available on-line[12].
The cLODg workflow consists of four main steps:

1. Data acquisition
2. Data conversion and integration
3. Data augmentation and verification
4. Linked Data Publication

Data acquisition is currently supported from (i) proprietary XML data obtained through the easychair conference management system[13], (ii) html based input (iii) csv files (iv) ics files for calendar events.

Data conversion is implemented via XSLT transformations [14] and integrated with existing LOD (e.g. we check if some of the conference participants are already present in the SWDF corpus). Each person in the graph is identified by a URI. For each person we generate a transparent URI of the form `http://data.semanticweb.org/person/<firstname>-<surname>`. The convention to generate a URIs for a person is to use `http://data.semanticweb.org/person/` as prefix and concatenate any firstname, middle names and surnames, separeted by the dash character. This procedure should make sure that if a person is already present in the SWDF corpus, the same URI is reused. This can in practice cause problems due to misspelling, noise and ambiguity, for which we implemented a naive disambiguation procedure, described in [6].

Produced data is used to pre-populate on-line forms[15], which are submitted to conference participants in order to (ii) verify correctness and (ii) collect additional information such as their photos and twitter accounts.

Resulting LOD data about the conference is sent as dump file to be integrated in the SWDF corpus. Currently this step is performed manually by the SWDF corpus administrator.

[10] `http://sempublishing.sourceforge.net/`

[11] `ontologydesignpatterns.org/ont/eswc/ontology.owl`

[12] The ontology is available at `ontologydesignpatterns.org/ont/eswc/ontology.owl`

[13] `http://www.easychair.org/`

[14] `http://www.w3.org/TR/xslt`

[15] Example of form: `http://wit.istc.cnr.it/conference-live/data`

4 Conclusions and future work

This paper describes cLODg, a set of tools to collect, refine and produce Linked Data to describe scientific conferences and their publications, participants and events. cLODg is an answer to the need of open tools for metadata generation (in the spedific case in the domain of scientific conferences). It has the ambition to foster a synergy between publishers and researches and to provide a possible way forward to combine the efforts between the two. The main contribution of this work is an open source tool to support the production of metadata for conferences and scholarly data.

References

1. D. Beckett, E. Miller, and D. Brickey. Expressing simple dublin core in rdf/xml. retrievable on line at `http://dublincore.org/documents/dcmes-xml/`, 2002.
2. D. Berrueta, D. Brickley, S. Decker, S. Fernández, C. Görn, A. Harth, T. Heath, K. Idehen, K. Kjernsmo, A. Miles, A. Passant, A. Polleres, L. Polo, and M. Sintek. SIOC Core Ontology Specification. W3c member submission, W3C, June 2007.
3. D. Brickley and L. Miller. FOAF Vocabulary Specification. Technical report, FOAF project, May 2007. Published online on May 24th, 2007 at `http://xmlns.com/foaf/spec/20070524.html`.
4. V. Bryl, A. Birukou, K. Eckert, and M. Kessler. What is in the proceedings? combining publishers and researchers perspectives. In *4th Workshop on Semantic Publishing (SePublica 2014), Anissaras, Greece, May 25th, 2014*, 2014.
5. A. Garcia, P. Murray-Rust, G. Burns, R. Stevens, D. Tkaczyk, C. McLaughlin, A. Belin, A. Di Iorio11, L. García, C. Gruson-Daniel, et al. Pdfjailbreak–a communal architecture for making biomedical pdfs semantic. *Proceedings of BioLINK SIG 2013*, page 13, 2013.
6. A. L. Gentile, M. Acosta, L. Costabello, A. G. Nuzzolese, V. Presutti, and D. Reforgiato Recupero. Conference live: Accessible and sociable conference semantic data. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 1007–1012. International World Wide Web Conferences Steering Committee, 2015.
7. W. Harrison. Eating your own dog food. *Industrial and Organizational Psychology*, (June):5–7, 2011.
8. C. Lange and A. Di Iorio. Semantic publishing challenge - assessing the quality of scientific output. In *Semantic Web Evaluation Challenge*, volume 475 of *Communications in Computer and Information Science*, pages 61–76. Springer International Publishing, 2014.
9. K. Möller, S. Bechofer, and T. Heath. Semantic web conference ontology. retrievable on line at `http://data.semanticweb.org/ns/swc/swc_2009-05-09.html`, 2009.
10. K. Möller, T. Heath, S. Handschuh, and J. Domingue. Recipes for semantic web dog food: The eswc and iswc metadata projects. In *Proc. of ISWC'07/ASWC'07*, pages 802–815, Berlin, Heidelberg, 2007. Springer-Verlag.
11. D. Shotton. Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing*, 22(2):85–94, 2009.