Generating Conference Linked Open Data in One Click

Andrea Giovanni Nuzzolese[1], Anna Lisa Gentile[2], Valentina Presutti[1], and Aldo Gangemi[1]

[1] Semantic Technology Lab, ISTC-CNR. Italy
[2] University of Mannheim
andrea.nuzzolese@istc.cnr.it, annalisa@informatik.uni-mannheim.de, valentina.presutti@cnr.it, aldo.gangemi@cnr.it

Abstract. In this paper we describe cLODg2 (conference Linked Open Data generator - version 2), a tool to collect, refine and produce Linked Data about scientific conferences with their associated publications, participants and events. Conference metadata collected from different unstructured and semi-structured resources must be expressed with appropriate vocabularies to be exposed as Linked Data. cLODg2 facilitates this task by providing a one-click workflow to generate data which is ready to be integrated in the ScholarlyData.org dataset. cLODg2 is an open source project, which has the aim to foster the publication of scholarly Linked Open Data and encourage collaborative efforts in this direction between researchers and publishers.

1 Introduction

Scholarlydata [4] is the evolution of the *Semantic Web Dog Food* (SWDF) dataset[3]. The SWDF corpus was the first considerable effort to offer comprehensive semantic descriptions of conference events [3], collecting linked data about papers, people, organizations, and events related to academic conferences.

A comprehensive description of Scholarlydata can be found in [4], while in this paper we provide technical details about cLODg2, the Open Source tool[4] that supports data generation for Scholarlydata. cLODg2 (conference Linked Open Data generator - version 2) provides a one click process for the conference metadata publication workflow. cLODg2 has been used to refactor the SWDF dataset and to gather and publish new conference metadata[5]. The tool provides an easy process to generate Linked Data which can be directly added to the ScholarlyData dataset.

[3] SWDF: http://data.semanticweb.org
[4] https://github.com/anuzzolese/cLODg2
[5] Amongst other it has been used for ESWC conference since 2014 http://2016.eswc-conferences.org

2 cLODg2 - publishing Conference Semantic Data

The main goal of cLODg2 is to facilitate the generation of conference Linked Data which can be readily integrated in the Scholarlydata[6] dataset. Scholarlydata [4] is the evolution of the SWDF dataset [3] based on an improvement of the Semantic Web Conference (SWC) Ontology[7], the Conference Ontology[8][5], which improves SWC adopting best ontology design practices. The necessary steps to add conference data to Scholarlydata are: (i) Data acquisition, (ii) Linked Data generation, (iii) Linked Data enrichment and (iv) Linked Data Publication.

The Data acquisition step, to be done by the user, consists of acquiring metadata about the conference, generally exported from a conference management system. We currently support data acquisition from CSV files[9]. Additionally, Linked Data represented with the SWC ontology can be used as initial input[10].

Starting from provided input cLODg2 performs two sequential steps: Linked Data generation and data enrichment. Figure 1 shows the system architecture, including all accessed services and technologies, modelled as an UML activity diagram. The initialisation step merely consists of configuring a property file to point to (i) the collected CSV files containing the input data and (ii) the D2RQ mapping that will serve for converting CSV files to RDF. A D2RQ mapping for dealing with easychair data is provided by default, but expert users can change this to import ad hoc CSV files.

The Linked Data generation activity is composed of the following steps:

- **Data gathering**. This action merely represents the system fetching data from the specified location. We remark that for the sake of simplicity we fix the easychair model for the input data, but that this can be easily configured for multiple data gathering support.
- **RDB population**. This action aims at populating a relational database (RDB) from the CSV files gathered from the previous action. The RDB is based on HyperSQL (HSQLDB)[11], which is a lightweight open-source Java database;
- **D2R conversion**. The previous action, i.e., RDB population, is preparatory to this step. In fact, cLODg2 relies on the D2R framework [1] to perform the conversion of a non-RDF source to RDF. The conversion is guided by the mapping provided as input. This mapping is described by using the D2RQ

[6] `http://w3id.org/scholarlydata`

[7] `http://data.semanticweb.org/ns/swc/swc_2009-05-09.html`

[8] Refer to `http://w3id.org/scholarlydata/ontology/conference-ontology.owl` to obtain the OWL source code and to `http://goo.gl/4lOHSk` to obtain the HTML documentation of the Conference Ontology.

[9] A simplified example of such data, exported from `easychair.org` can be found at `https://github.com/anuzzolese/cLODg2/tree/master/csv_samples`

[10] Example dump at `https://github.com/AnLiGentile/cLODg/tree/master/resources/swdf_samples`

[11] `http://hsqldb.org`

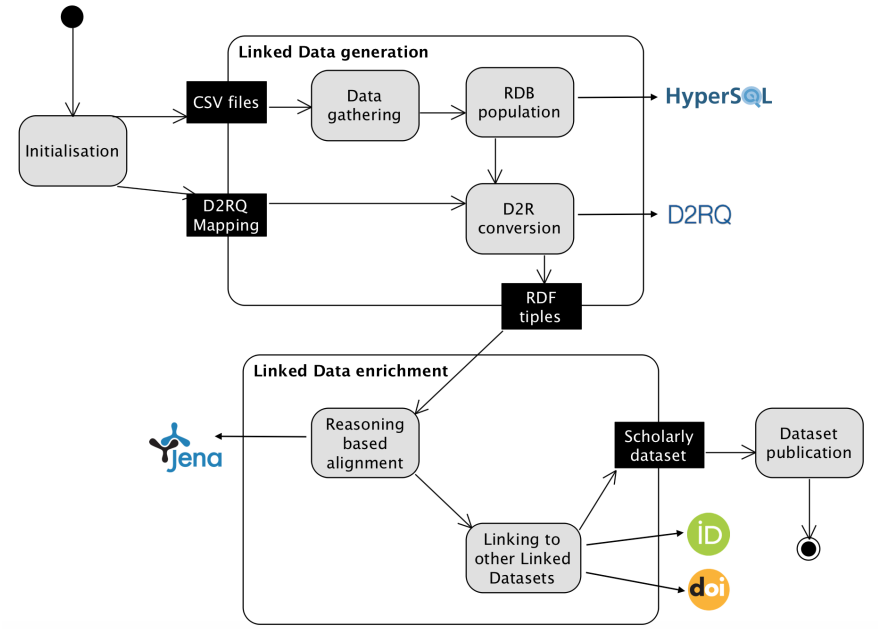

Fig. 1. cLODg2 architecture represented as an UML activity diagram.

mapping language [2]. cLODg2 is released along with a default mapping for easychair data and targets two distinct alternative datasets: the SWDF and Scholarlydata.

The Linked Data enrichment activity is composed of the following actions:

- **Reasoning-based alignment.** Input of this action are the RDF triples produced by the Linked Data generation activity. The output is the materialisation of a set of RDF triples that enable the alignment to other ontologies and vocabularies, i.e., the SWDF ontology, SPAR[12], Dolce D0[13], the Organization Ontology[14], FOAF, SKOS, icatzd , and the Collections Ontology[15]. The alignment triples are materialised by means of OWL-DL reasoning, which is enabled by the Apache Jena inference layer;
- **Linking to other Linked Datasets.** This action is aimed at producing instance level alignments, expressed via `owl:sameAs` axioms. The target linked datasets are ORCID[16] and DOI[17]. ORCID provides persistent digital identifiers for scientific researchers and academic authors. A digital object identifier

[12] `http://www.sparontologies.net`
[13] `http://www.ontologydesignpatterns.org/ont/dul/d0.owl`
[14] `https://www.w3.org/TR/vocab-org`
[15] `http://purl.org/co`
[16] `http://orcid.org`
[17] `https://www.doi.org`

(DOI) is a serial code used to uniquely identify digital objects, particularly used for electronic documents. The alignments to ORCID are produced by relying on the public API provided by ORCID[18]. The references to DOI are produced by relying on the API provided by Crossref[19], performing a search on each article title.

The Linked Data Publication step, which is the last action in the cLODg2 workflow, has to be done by the user and consists of submitting produced data to Scholarlydata.org.

3 Conclusions

This paper describes cLODg2, a tool to collect, refine and produce Linked Data to describe scientific conferences and their publications, participants and events. The main contribution of this work is an open source tool to support the production of metadata for conferences and scholarly data which is ready to be integrate in the ScholarlyData dataset, with minimal user effort. Future work will be mainly focused at addressing data quality and reduce duplications and misspelling in the data.

References

1. C. Bizer and R. Cyganiak. D2R Server - Publishing Relational Databases on the Semantic Web. In *Proc. of ISWC2006 Poster&Demo*, 2006.
2. C. Bizer and A. Seaborne. D2RQ - Treating Non-RDF Databases as Virtual RDF Graphs. In *Proc. of ISWC2004 posters*, 2004.
3. 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.
4. A. G. Nuzzolese, A. L. Gentile, V. Presutti, and A. Gangemi. Conference Linked Data Our Web Dog Food has gone gourmet. In *Proc. of ISWC2016 Resource Track*, page to appear, 2016.
5. A. G. Nuzzolese, A. L. Gentile, V. Presutti, and A. Gangemi. Semantic web conference ontology - a refactoring solution. In *The Semantic Web: ESWC 2016 Satellite Events*, page to appear. Springer, 2016.

[18] `http://members.orcid.org/api/introduction-orcid-public-api`
[19] `http://www.crossref.org/guestquery`