Jailbreaking your reference lists: OpenCitations strike again

Silvio Peroni[1], David Shotton[2], and Fabio Vitali[1]

[1] DASPLab, DISI, University of Bologna, Bologna, Italy
[2] Oxford e-Research Centre, University of Oxford, Oxford, UK
silvio.peroni@unibo.it, david.shotton@oerc.ox.ac.uk, fabio.vitali@unibo.it

Abstract. In this poster paper we provide an overview of the OpenCitations Project (http://opencitations.net) and of its main outcome, the OpenCitations Corpus, which is an open repository of scholarly citation data made available under a CC0 license, providing in RDF accurate citation information harvested from the scholarly literature, starting from the PubMed Central OA subset.
RASH: https://w3id.org/oc/paper/occ-poster-iswc2016.html

Keywords: Citation Database, OpenCitations, OpenCitations Corpus, Scholarly Communication, Semantic Publishing

1 Introduction

Reference lists from academic articles are the core elements of scholarly communication, since they both permit the attribution of credit and integrate our independent research endeavours. But the cruel reality is that they are not freely available. For instance, UK research universities each pay tens of thousands of pounds annually [5] for accessing to commercial sources of scholarly citations – Scopus and Web of Science. In the current age where Open Access is considered a necessary practice in research, it is a scandal that reference lists from journal articles are not readily and freely available for use by all scholars. As we have already stated in a previous work [5]:

> Citation data now needs to be recognized as a part of the Commons – those works that are freely and legally available for sharing – and placed in an open repository, where they should be stored in appropriate machine-readable formats so as to be easily reused by machines to assist people in producing novel services.

This is the main premise behind the OpenCitations Project and a number of other complementary initiatives including Semantic Lancet[3] [2], Springer LOD[4], OpenAIRE[5] [1] and Scholarly Data[6] [3]. Since the beginning of July, OpenCitations has started to ingest and process the reference lists of scholarly papers available in Europe PubMed Central[7]. In this paper we provide a brief overview of its main components that make the extraction and description of such reference lists in RDF possible.

[3] http://semanticlancet.eu/
[4] http://lod.springer.com/
[5] https://www.openaire.eu/
[6] http://www.scholarlydata.org/
[7] http://europepmc.org/

2 OpenCitations and its Corpus

The OpenCitations Project (`http://opencitations.net`) is creating an open dataset of citation data integrated with a SPARQL endpoint[8] and a very simple Web interface that shows only the data about bibliographic entities[9]. Its main output is the Open Citations Corpus (OCC), an open repository of scholarly citation data made available under a Creative Commons CC0 license[10], which provides accurate bibliographic references harvested from the scholarly literature, described using the SPAR Ontologies[11] [4] according to the OCC metadata document [6], that others may freely build upon, enhance and reuse for any purpose, without restriction under copyright or database law.

The OCC stores metadata relevant to these citations in RDF, encoded as JSON-LD, and makes them available through a SPARQL endpoint (and, in the near future, as downloadable datasets). It includes information about six different kinds of bibliographic entity:

- **bibliographic resources** (br) – resources that cite/are cited by other bibliographic resources (e.g. journal articles), or that contain such citing/cited resources (e.g. journals);
- **resource embodiments** (re) – details of the physical or digital forms in which the bibliographic resources are made available by their publishers;
- **bibliographic entries** (be) – the literal textual bibliographic entries occurring in the reference lists within bibliographic resources, that reference other bibliographic resources;
- **responsible agents** (ra) – names of agents having certain roles with respect to bibliographic resources (i.e. names of authors, editors, publishers, etc.);
- **agent roles** (ar) – roles held by agents with respect to bibliographic resources (e.g. author, editor, publisher);
- **identifiers** (id) – external identifiers (e.g. DOI, ORCID, PubMedID) associated with the bibliographic entities.

The corpus URL (`https://w3id.org/oc/corpus/`) identifies the entire OCC, which is composed of several sub-datasets, one for each of the aforementioned bibliographic entities included in the corpus. Each of these has a URL composed by suffixing the corpus URL with the two-letter short name for the class of entity (e.g. "be" for a bibliographic entry) followed by an oblique slash (e.g. `https://w3id.org/oc/corpus/be/`). Each dataset is described appropriately by means of the Data Catalog Vocabulary[12] and the VoID Vocabulary[13].

Upon initial curation into the OCC, a URL is assigned to each entity within each sub-dataset, all of which can be accessed in different formats (HTML, RDF/XML, Turtle, and JSON-LD) via content negotiation. Each entity URL is composed by suffixing the sub-dataset URL with a number assigned to each resource, unique among resources of the same type, which increments for each new

[8] `https://w3id.org/oc/sparql`
[9] Additional and more user-friendly interfaces will be available in the coming months, and will be described in the project homepage.
[10] `https://creativecommons.org/publicdomain/zero/1.0/legalcode`
[11] `http://www.sparontologies.net/`
[12] `https://www.w3.org/TR/vocab-dcat/`
[13] `https://www.w3.org/TR/void/`

entry in that resource class. For instance, the resource `https://w3id.org/oc/corpus/be/537` is the 537th bibliographic entry recorded within the OCC. Each of these entities has associated metadata describing its provenance by means of PROV-O[14] and its PROV-DC extension[15] (e.g. `https://w3id.org/oc/corpus/be/537/prov/se/1`). In particular, we keep track of the curatorial activities related to each OCC entity, the curatorial agents involved, and their roles. Additional information about OCC's handling of citation the data, and the way they are represented in RDF, are detailed in the official OCC Metadata Document [6].

The ingestion of citation data into the OCC is handled by two Python scripts called *Bibliographic Entries Extractor* (*BEE*) and the *SPAR Citation Indexer* (*SPACIN*), available in the OCC's GitHub repository[16]. As shown Fig. 1, BEE is responsible for the creation of JSON files containing information about the XML articles in the OA subset of PubMed Central (retrieved by using the Europe PubMed Central API[17] and used as input data). Each of these JSON files also includes the complete reference list of the paper under consideration extracted by means of XPath queries. Then, SPACIN processes each JSON file, retrieves metadata information about all the citing/cited articles described in it by querying the Crossref API[18] and the ORCID API[19], and stores all the generated RDF resources in the file system in JSON-LD format and within the OCC triplestore, disambiguating the resources that have been added previously by means of the retrieved identifiers (i.e. DOI, PMID, PMCID, ORCID, URL). It is worth noting that the triplestore includes all the data about the curated entities except their provenance data and the descriptions of the datasets, that are accessible only via HTTP.

The workflow introduced in Fig. 1 is a process that runs until no more JSON files can be produced by BEE. Thus, the current instance of the OCC is evolving dynamically in time, and can be easily extended so as to interact with additional REST APIs from different sources, so as to gather additional article metadata and their related references. Currently, each day the workflow adds about 2 million triples to the corpus, describing more than 20,000 new citing/cited bibliographic resources and about 100,000 new authors, about 5% of whom are disambiguated through their ORCID ids.

3 Conclusions

In this poster paper we have introduced the OpenCitations Project, which is involved in creating an open repository of accurate bibliographic references harvested from the scholarly literature: the OpenCitations Corpus (OCC). The new instance of the OCC has just been established, and already includes 728,991 citation links (as of August 30, 2016) – a number that will grow quickly over the coming months as the continuous workflow adds new data dynamically from Europe PubMed Central and other authoritative sources. The OCC SPARQL

[14] `https://www.w3.org/TR/prov-o/`
[15] `https://www.w3.org/TR/prov-dc/`
[16] `https://github.com/essepuntato/opencitations`
[17] `https://europepmc.org/RestfulWebService`
[18] `http://api.crossref.org/`
[19] `http://members.orcid.org/api/`

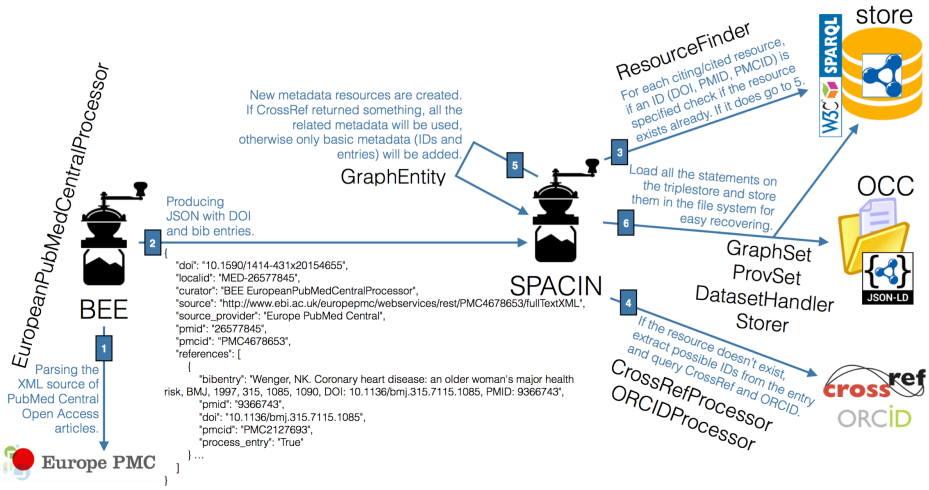

Fig. 1. The steps involving BEE and SPACIN, and their related Python classes, in the production of the OpenCitations Corpus.

endpoint is presently available for use, and distributions of the OCC will shortly be made openly available for bulk download – the first of these by early September 2016, with subsequent incremental additions.

We are currently working on two different aspects. First of all, we are developing tools for linking the resources within the OCC with those included in other datasets, e.g. Scholarly Data. In addition, we are experimenting with the use of multiple parallel instantiations of SPACIN, so as to increase the amount of new information that can be processed daily.

References

1. Alexiou, G., Vahdati, S., Lange, C., Papastefanatos G., Lohmann, S. (2016). OpenAIRE LOD services: Scholarly Communication Data as Linked Data. To appear in Proceedings of SAVE-SD 2016. `http://cs.unibo.it/save-sd/2016/papers/html/alexiou-savesd2016.html`
2. Bagnacani, A., Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., Vitali, F. (2014). The Semantic Lancet Project: A Linked Open Dataset for Scholarly Publishing. In EKAW 2014 Satellite Events: 101–105. `http://dx.doi.org/10.1007/978-3-319-17966-7_10`
3. Nuzzolese, A. G., Gentile, A. L., Presutti, V., Gangemi, A. (2016). Conference Linked Data – Our Web Dog Food has gone gourmet. To appear in Proceedings of ISWC 2016.
4. Peroni, S. (2014). The Semantic Publishing and Referencing Ontologies. In Semantic Web Technologies and Legal Scholarly Publishing: 121–193. `http://dx.doi.org/10.1007/978-3-319-04777-5_5`
5. Peroni, S., Dutton, A., Gray, T., Shotton, D. (2015). Setting our bibliographic references free: towards open citation data. Journal of Documentation, 71 (2): 253–277. `http://dx.doi.org/10.1108/JD-12-2013-0166`
6. Peroni, S., Shotton, D. (2016). Metadata for the OpenCitations Corpus. figshare. `https://dx.doi.org/10.6084/m9.figshare.3443876`