Freedom for bibliographic references: OpenCitations arise

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, Oxofrd, UK
silvio.peroni@unibo.it, david.shotton@oerc.ox.ac.uk,
fabio.vitali@unibo.it

Abstract. Scholarly citations from one publication to another, expressed as reference lists within academic articles, are the core elements of scholarly communication. Unfortunately they usually can be accessed *en masse* only by paying significant fees to commercial organizations. Those few services that do made them available for free impose strict limitation on their reuse. In this paper we provide an overview of the OpenCitations project undertaken to address this issue, and of its main product, the OpenCitations Corpus, which is an open repository of accurate bibliographic citation data harvested from the scholarly literature, made available under a Creative Commons public domain dedication in RDF.
RASH version: http://bit.ly/29Cy92i

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

1 Introduction

Databases of citation data are among the most attractive and used artefacts in the Scholarly Communication domain. They are the main tool used by researchers for gaining knowledge about a particular topic, and by scientists in Bibliometrics, Informetrics, and Scientometrics for analysing the complex relationships that exist within huge networks of citations of scholarly works. They also serve institutional goals, since they are one of the main mechanisms for the assessment of the quality of research by means of (sometimes questionable) metrics and indicators calculated from such citation databases. While some of these , e.g. Microsoft Academic Graph[3] and Google Scholar[4], are freely accessible (but not downloadable), those considered the most authoritative by institutions worldwide, namely Scopus[5] and Web of Science[6], can be accessed only by paying

[3] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

[4] https://scholar.google.com/

[5] https://www.scopus.com/

[6] http://webofscience.com/

significant annual access fees, which may amount to tens of thousands of pounds databases [7].

Reference lists within academic articles are 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. In the current age where Open Access is considered a necessary practice in research, it is a scandal that reference lists from scholarly publications (conference papers, books, journal articles, etc.) are not readily and freely available for use by all scholars. As we have already stated in a previous work [7]:

> 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 (OC) project [7] [13], which is creating an open repository of scholarly citation data – the OpenCitations Corpus (OCC) – made available under a Creative Commons public domain dedication[7], which provides in RDF accurate citation information (bibliographic references) harvested from the scholarly literature. Since the beginning of July, the OCC has started to ingest and process the reference lists of scholarly papers available in Europe PubMed Central[8]. In this paper we provide a brief overview of the OCC's main components that make possible the extraction and description of such reference lists in RDF.

The rest of the paper is organised as follows. In Section 2 we recall the story of the OpenCitations project since its beginning in 2010. In Section 3 we describe all the materials that have been recently developed within the OpenCitations project for the creation of a new and improved instantiation of the OCC. In Section 4 we briefly introduce some recent works about the creation of open (and RDF-based) repository of scholarly document metadata. Finally, in Section 5, we sketch out some future works.

2 The story so far

The OpenCitations (OC) has formally started in 2010 as a one-year project funded by JISC[9] (and extended for an additional half year), with David Shotton as director, who at that time was working in the Department of Zoology at the University of Oxford. The project was global in scope, and was designed to change the face of scientific publishing and scholarly communication, since it aimed to publish bibliographic citation information in RDF and to make citation

[7] `https://creativecommons.org/publicdomain/zero/1.0/legalcode`

[8] `http://europepmc.org/`

[9] `http://www.jisc.ac.uk/whatwedo/programmes/inf11/jiscexpo/jiscopencitation.aspx`

links as easy to traverse as Web links. The main deliverable of the project, among several outcomes[10], was the release of an open repository of scholarly citation data described using the SPAR (Semantic Publishing and Referencing) Ontologies[11], namely the OpenCitations Corpus, initially populated with the citations from journal articles within the Open Access Subset of PubMed Central[12] [13].

In May 2014, OC was adopted by the Infrastructure Services for Open Access (IS4OA)[13] as one of its academic Open Access services. IS4OA is UK-based not-for-profit charitable company that aims to provide benefit to the global community of research information users by acting as an umbrella organisation that supports openly accessible information and discovery services relating to academic information, research results and scholarly publications, by providing business structure and expertise and a means of channelling financial support to these services.

At the end of 2015 Silvio Peroni joined the OC project as co-director, with the aim of setting up a new instantiation of the Corpus based on a new metadata schema and employing several new technologies to automate the ingestion of fresh citation metadata from authoritative sources. The current instantiation of the OCC is hosted by the Department of Computer Science and Engineering (DISI) at the University of Bologna, and since the beginning of July has started to ingest, process and publish reference lists of scholarly papers available in Europe PubMed Central[14], as described in the following section.

3 The new instantiation of the OpenCitations Corpus

The OpenCitations project (`http://opencitations.net`) is currently creating an open citation database with an integrated SPARQL endpoint and a browsing interface to support data consumers. Its main output is the Open Citations Corpus (OCC), an open repository of scholarly citation data made available under a Creative Commons public domain dedication (CC0), which provides accurate bibliographic references harvested from the scholarly literature, described using the SPAR Ontologies [6] according to the OCC metadata document [11], that others may freely build upon, enhance and reuse for any purpose, without restriction under copyright or database law.

3.1 The model

The metadata model used for the data stored in the OCC, available at [11] and briefly summarised in Fig. 1, is explicitly aligned with the SPAR Ontologies and other standard vocabularies. In particular:

[10] `https://opencitations.wordpress.com/2011/07/01/jisc-open-citations-project---final-project-blog-post/`

[11] `http://www.sparontologies.net/`

[12] `http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`

[13] `https://is4oa.org/services/open-citations-corpus/`

[14] `http://europepmc.org/`

- the FRBR-aligned Bibliographic Ontology (FaBiO)[15] [8] is used for providing a description of all the metadata of citing/cited resources (conference papers, book chapters, journal articles, etc.), their related container resources (academic proceedings, books, journals, etc.), as well as metadata about the particular formats in which they have been embodied (digital vs. print, first and ending pages, etc.);
- the Publishing Roles Ontology (PRO)[16] [12] is used for describing the roles of agents (author, editor, publisher, etc.) related to bibliographic resource – while the order among roles, e.g. the list of authors of a paper, is handled by extending PRO with an additional property, i.e. `oco:hasNext`;
- the Bibliographic Reference Ontology (BiRO)[17] and the Citation Counting and Context Characterization Ontology (C4O)[18] [4] are used for describing the textual content of each reference in the reference list of a citing bibliographic resource;
- finally, the DataCite Ontology[19] is used for defining all the identifiers (e.g. DOI, PubMed ID, PubMed Central ID, ORCID, ISSN, etc.) for bibliographic resources and all the agents involved – the Friend Of A Friend (FOAF)[20] ontology is used for defining additional data (e.g. names) about agents.

All the terms from the aforementioned ontologies are collected within a new ontology called OpenCitations Ontology (OCO)[21]. This is not yet another bibliographic ontology, rather just a place where existing complementary ontological entities from several other ontologies are grouped together for the purpose of providing descriptive metadata for the OCC.

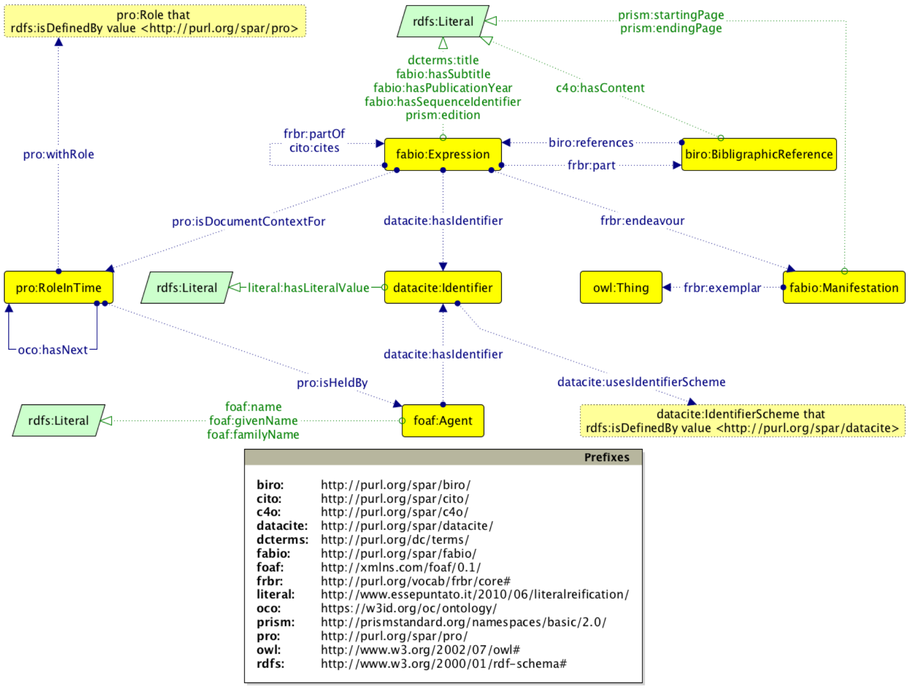
3.2 The data

The OCC stores metadata relevant to these citations in RDF, encoded as JSON-LD. In the near future, all the data will be also available as downloadable datasets – in the meantime, two exemplar dataset, compliant with the OCC metadata model introduced in Section 3.1, have been made available starting from article metadata provided by Springer Nature (available at [10]) and gathered via Europe PubMed Central (available at [9]).

The OCC (as well as the aforementioned exemplar datasets) includes information about six different kinds of bibliographic entity:

- bibliographic resources (br), class `fabio:Expression` – resources that either cites or are cited by other bibliographic resources (e.g. journal articles), or that contain such citing/cited resources (e.g. journals);

[15] `http://purl.org/spar/fabio`
[16] `http://purl.org/spar/pro`
[17] `http://purl.org/spar/biro`
[18] `http://purl.org/spar/c4o`
[19] `http://purl.org/spar/datacite`
[20] `http://xmlns.com/foaf/spec/`
[21] `https://w3id.org/oc/ontology`



Prefixes

biro:	http://purl.org/spar/biro/
cito:	http://purl.org/spar/cito/
c4o:	http://purl.org/spar/c4o/
datacite:	http://purl.org/spar/datacite/
dcterms:	http://purl.org/dc/terms/
fabio:	http://purl.org/spar/fabio/
foaf:	http://xmlns.com/foaf/0.1/
frbr:	http://purl.org/vocab/frbr/core#
literal:	http://www.essepuntato.it/2010/06/literalreification/
oco:	https://w3id.org/oc/ontology/
prism:	http://prismstandard.org/namespaces/basic/2.0/
pro:	http://purl.org/spar/pro/
owl:	http://www.w3.org/2002/07/owl#
rdfs:	http://www.w3.org/2000/01/rdf-schema#

Fig. 1. The Graffoo diagram of the main ontological entities described by the OCC metadata model.

– resource embodiments (re), class `fabio:Manifestation` – details of the physical or digital forms in which the bibliographic resources are made available by their publishers;
– bibliographic entries (be), class `biro:BibliographicReference` – the literal textual bibliographic entries occurring in the reference lists within bibliographic resources, that reference other bibliographic resources;
– responsible agents (ra), class `foaf:Agent` – names of agents having certain roles with respect to bibliographic resources (i.e. names of authors, editors, publishers, etc.);
– agent roles (ar), class `pro:RoleInTime` – roles held by agents with respect to bibliographic resources (e.g. author, editor, publisher);
– identifiers (id) (class `datacite:Identifier`) – 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[22] and the VoID Vocabulary[23], and a SPARQL endpoint[24] is made available for all the entities included in the entire OCC.

Upon initial curation into the OCC, a URL is assigned to each entity within each sub-dataset, which can be accessed in different formats (HTML, RDF/XML, Turtle, and JSON-LD) via URL dereferencing. 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 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. The final part of such URL, i.e. the two-letter short name for the class of items plus "/" plus the number ("be/537" in the example), is called *corpus identifier*, since it allows the unique identification of any entity within the OCC.

Each of these entities has associated metadata describing its provenance by means of PROV-O[25] and its PROV-DC extension[26] (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.

All these RDF data are stored in BibJSON[27] encoded as JSON-LD, defined through an appropriate JSON-LD context[28] which hides the complexity of the model in Fig. 1 behind natural language keywords. For instance, the following excerpt is the JSON-LD linearisation of the aforementioned "be/537" entity:

```
{
  "iri": "gbe:537",
  "a": "entry",
  "label": "bibliographic entry 537 [be/537]",
  "content": "Chang, KY, Unanue, ER. Prediction of HLA-DQ8beta cell peptidome
       using a computational program and its relationship to autoreactive T cells,
        Int Immunol, 2009, 21, 6, 705, 13, DOI: 10.1093/intimm/dxp039, PMID:
       19461125",
  "crossref": "gbr:1682"
}
```

In this excerpt, "iri" defines the URL of the resource in consideration (where "gbe:" is a prefix for "https://w3id.org/oc/corpus/be/"), while "a", "entry", "label", "content" and "crossref" stand for rdf:type, biro:BibliographicReference, rdfs:label, c4o:hasContent and biro:references respectively (where "gbr:" is a prefix for "https://w3id.org/oc/corpus/br/").

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 [11].

[22] https://www.w3.org/TR/vocab-dcat/

[23] https://www.w3.org/TR/void/

[24] http://w3id.org/oc/sparql

[25] https://www.w3.org/TR/prov-o/

[26] https://www.w3.org/TR/prov-dc/

[27] http://okfnlabs.org/bibjson/

[28] https://w3id.org/oc/corpus/context.json

3.3 The ingestion workflow

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[29].

BEE As shown Fig. 2, BEE is responsible for the creation of JSON files containing information about the articles in the OA subset of PubMed Central (retrieved by using the Europe PubMed Central API[30]). Each of these JSON files is created by asking Europe PubMed Central about all the metadata of the articles it stores that have available the source XML file. Once identified, BEE processes all the XML sources so as to extract the complete reference list of the paper in consideration, and includes all the data in the final JSON file. An excerpt of one of those JSON files is introduced as follows:

```
{
  "doi": "10.1007/s11892-016-0752-4",
  "pmid": "27168063",
  "pmcid": "PMC4863913",
  "localid": "MED-27168063",
  "curator": "BEE EuropeanPubMedCentralProcessor",
  "source": "http://www.ebi.ac.uk/europepmc/webservices/rest/PMC4863913/
      fullTextXML",
  "source_provider": "Europe PubMed Central",
  "references": [
    ...
    {
      "bibentry": "Chang, KY, Unanue, ER. Prediction of HLA-DQ8beta cell peptidome
           using a computational program and its relationship to autoreactive T
          cells, Int Immunol, 2009, 21, 6, 705, 13, DOI: 10.1093/intimm/dxp039,
          PMID: 19461125",
      "pmid": "19461125",
      "doi": "10.1093/intimm/dxp039",
      "pmcid": "PMC2686615",
      "process_entry": "True"
    },
    ...
  ]
}
```

In particular, for each articles retrieved by means of the Europe PubMed Central API, BEE stores all the possible identifiers (in the example, "doi", "pmid", "pmcid", and "localid") and all the textual references, enriched by their own related identifiers if they are available. In addition, the JSON file also includes provenance information about the source, its provider and the curator (i.e. the particular BEE Python class responsible for the extraction of these metadata from the source).

We have done a few tests so as to understand the performances of BEE in generating these JSON files. In particular, we queried Europe PubMed Central for the metadata of articles while running BEE for 30 minutes on a MacBook Pro, with 2 GHz Intel Core i7 processor, 8 GB DDR3 1600 MHz, OS X 10.11.3. In that time, we were able to create 185 JSON files containing all the aforementioned

[29] https://github.com/essepuntato/opencitations

[30] https://europepmc.org/RestfulWebService

metadata, giving a rate of about 6 new JSON files per minute – which will be processed, independently, by the tool presented in the next section.

SPACIN Starting from the output provided by BEE, SPACIN processes each JSON file, retrieving metadata information about all the citing/cited articles described in it by querying the Crossref API[31] and the ORCID API[32]. These API are also used to disambiguate bibliographic resources and agents by means of the identifiers retrieved (e.g., DOI, ISSN, ISBN, ORCID, URL, and Crossref member URL). Once SPACIN has retrieved all these metadata, appropriate RDF resource are created (or reused, if they have been already added in the past) and stored in the file system in JSON-LD format (as shown in Section 3.2) and additionally within the OCC triplestore. It is worth noting that, for space and performance reasons, the triplestore includes all the data about the curated entities , but does not store their provenance data nor the descriptions of the datasets themselves, that are accessible only via HTTP.

The SPACIN workflow introduced in Fig. 2 is a process that runs until no more JSON files are available from BEE. Thus, the current instance of the OCC is evolving dynamically in time, and can be easily extended beyond Europe Pubmed Central by reconfiguring it to interact with additional REST APIs from different sources, so as to gather new article metadata and their related references.

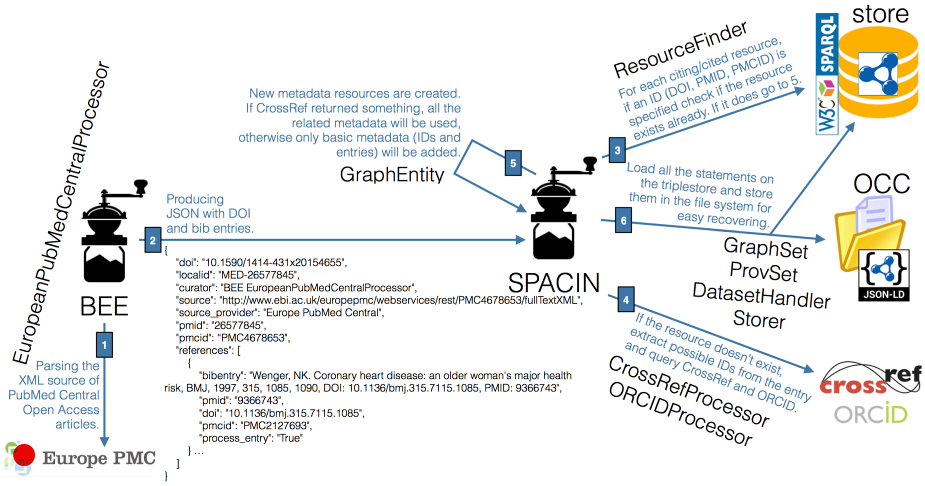
Each new resource recorded within the OCC by SPACIN occupies between 0.3 and 4 kb, plus an additional 32 kb dedicated to storage of its provenance data. Each day the workflow adds about 2 million triples to the corpus, describing more than 20,000 new citing/cited bibliographic resources and about 100,000 new authors, 5% of whom are disambiguated through their ORCID ids.

We have done a few tests so as to understand the performances of SPACIN in processing the JSON files generated by BEE and produce new resources for the OCC. In particular, we run SPACIN on on two subsets of JSON file: 67 JSON files describing all the papers included in the Proceedings of ISWC 2015, and the first 67 JSON files produced by BEE as the outcome of the experiment described in Section 3.3. We use the same configuration as before, i.e. a MacBook Pro, with 2 GHz Intel Core i7 processor, 8 GB DDR3 1600 MHz, OS X 10.11.3.

ISWC 2015 dataset. SPACIN took 45 minutes to process all the papers in the ISWC 2015 Proceedings, and the outcomes have been published in [10]. Each citing paper contained about 23 references on average, and SPACIN produced 1,441 new citing/cited resources, for a total of 1,531 citation links. These resources are contained in 411 different container resources (e.g. journals, proceedings, books), published by 42 distinct publishers. The total number of authors is 3,076, 157 of whom (5.1%) have been disambiguated through their ORCID. The total number of RDF statements created is 69,995 (which, as explained, excludes provenance data and datasets information).

[31] `http://api.crossref.org/`

[32] `http://members.orcid.org/api/`


EuropeanPubMedCentralProcessor
BEE
1
Parsing the XML source of PubMed Central Open Access articles.
Europe PMC
2
Producing JSON with DOI and bib entries.
New metadata resources are created. If CrossRef returned something, all the related metadata will be used, otherwise only basic metadata (IDs and entries) will be added.
GraphEntity
5
SPACIN
3
ResourceFinder
For each citing/cited resource, if an ID (DOI, PMID, PMCID) is specified check if the resource exists already. If it does go to 5.
store
W3C SPARQL
6
Load all the statements on the triplestore and store them in the file system for easy recovering.
GraphSet
ProvSet
DatasetHandler
Storer
OCC
JSON-LD
4
If the resource doesn't exist, extract possible IDs from the entry and query CrossRef and ORCID.
CrossRefProcessor
ORCIDProcessor
crossref
ORCID

```
{
    "doi": "10.1590/1414-431x20154655",
    "localid": "MED-26577845",
    "curator": "BEE EuropeanPubMedCentralProcessor",
    "source": "http://www.ebi.ac.uk/europepmc/webservices/rest/PMC4678653/fullTextXML",
    "source_provider": "Europe PubMed Central",
    "pmid": "26577845",
    "pmcid": "PMC4678653",
    "references": [
        {
            "bibentry": "Wenger, NK. Coronary heart disease: an older woman's major health
risk, BMJ, 1997, 315, 1085, 1090, DOI: 10.1136/bmj.315.7115.1085, PMID: 9366743",
            "pmid": "9366743",
            "doi": "10.1136/bmj.315.7115.1085",
            "pmcid": "PMC2127693",
            "process_entry": "True"
        } ...
    ]
}
```

Fig. 2. The steps involving BEE and SPACIN, and their related Python classes, in the production of the OpenCitations Corpus.

Europe PubMed Central dataset. SPACIN took 2.5 hours to process the 67 papers from Europe PubMed Central, and the outcomes have been published in [9]. Each citing paper contained about 50 references on average, and SPACIN produced 3,391 new citing/cited resources, for a total amount of 3,337 citation links. These resources are contained in 1,047 different container resources (e.g. journals, proceedings, books), published by 137 distinct publishers. The total number of authors is 21,658, 957 of whom (4.4%) have been disambiguated through their ORCID. The total number of RDF statements created is 377,237 (excluding, as before, provenance data and datasets information).

In Table 1 we summarise some metrics related to the resources included in the aforementioned exemplar datasets. While these data are far from having a full coverage, they provide interesting snapshots of these two communities. On the one hand, the community of ISWC 2015 is a rather small Computer Science group of Semantic Web experts. Even if the average number of references per paper is quite small (average 23, the paper with the most references having 47 citation links), there are several papers that were cited more than one time (the most cited one received 7 citations). Many of the citations are to resources for which Crossref was not able to return any metadata. This is understandable, since many citations in these Semantic Web papers are to Web documents (e.g. W3C Recommendations) and to workshop papers not indexed by Crossref (e.g. CEUR Workshop Series[33]). Some of these non-Crossref-indexed publications are well known and well cited within this community (the most cited one has 4 citations within these 67 ISWC papers).

[33] http://ceur-ws.org/

Table 1. Some aggregated data of the two exemplar datasets produced by SPACIN.

Property	ISWC 2015	Europe PubMed Central
Max. number of citations within a paper	47	320
Max. number of citations received by a paper within this sample	7	3
Percentage of cited resources for which Crossref did not return any metadata	44%	13%
Max. number of citations received by a cited resource for which Crossref did not return any metadata	4	1

On the other hand, we see a quite different citation behaviour in the Europe PubMed Central papers. In this case, as expected, the number of average references is higher (average 50, with one paper having 320 citation links). The paper within this small sample that has been cited most received only 3 citations from the other 66 papers, and this is clearly due to the dimension of the Biomedical and Life Science community to which PubMed Central relates, which is clearly bigger and more sparse than the ISWC one. Again, these papers usually cites others published in journals, to which proper identifiers (e.g. DOI) have been properly assigned, explaining the lower percentage of citations to resources that are not indexed by Crossref[34].

4 Related works

In recent years we have seen a growing interest within the Semantic Web community for creating and making available RDF dataset concerning bibliographic metadata of scholarly documents. While the list of such works is quite extensive, in this section we would like to describe four of the most important contributions in the area.

The Semantic Lancet[36] Project [2] aims at building a Linked Open Dataset of scholarly publication metadata starting from the articles published by Elsevier. In particular, the current dataset contains SPAR-based [6] metadata about several papers published in the Journal of Web Semantics[37], including citation links marked with the motivations justifying them. It has several graphical interfaces that allow browsing and sense-making of these data.

[34] However, as of July 12, 2016, the most cited resource in the OCC (with 45 citations) is `https://w3id.org/oc/corpus/br/26550`, which refers to the R Project[35], a language and environment for statistical computing – see the bibliographic entry at `https://w3id.org/oc/corpus/be/9976`. Crossref did not return any metadata for it (thus, no title and other metadata are not explicitly specified in “br/26550”), since it is part of the grey literature.

[36] `http://semanticlancet.eu/`

[37] `http://www.journals.elsevier.com/journal-of-web-semantics/`

Springer LOD[38] [3] is an RDF dataset made available by Springer Nature that publishes Springer metadata about conferences as Linked Open Data (LOD). Its main focus in on proceedings volumes and the related conferences, but it does not contain metadata describing the articles contained in such proceedings.

OpenAIRE[39] [1] is an Horizon 2020 project which publishes metadata of more than 14 millions of publications and thousands of datasets. It makes available mechanism for searching, discovering and monitoring scientific outputs.

Finally, Scholarly Data[40] [5] is a new project that refactors the Semantic Web Dog Food[41] so as to keep the dataset growing in good health, and adopts the new Conference Ontology[42] (aligned with other existing models, e.g. SPAR [6]) for describing the data.

5 Conclusions

In this paper we have introduced the OpenCitations project, which is 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 is already populated with data describing 136,189 references (as of July 12, 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 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 the end of August 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 and Springer LOD. 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.

Acknowledgements. All the scripts used in OpenCitations have been developed as outcomes of several personal communications with people responsible for external services. We would like to thank leading people in Europe PubMed Central (in particular Johanna McEntyre and Vid Vartak), Crossref (in particular Ed Pentz, Geoffrey Bilder, and Karl Ward), and ORCID (in particular Josh Brown and Laurel Hack) for their help. We would also like to thank Alfred Hofmann and Aliaksandr Birukou (Springer Nature) for allowing us to publish in Figshare the OCC metadata concerning the Proceedings of ISWC 2015, which they kindly provided us in XML.

[38] `http://lod.springer.com/`

[39] `https://www.openaire.eu/`

[40] `http://www.scholarlydata.org/`

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

[42] `https://w3id.org/scholarlydata/ontology/conference-ontology.owl`

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. Bryl, V., Birukou, A., Eckert, K., Kessler, M. (2014). What's in the proceedings? Combining publisher's and researcher's perspectives. In Proceedings of SePublica 2014. http://ceur-ws.org/Vol-1155/paper-01.pdf
4. Di Iorio, A., Nuzzolese, A. G., Peroni, S., Shotton, D., Vitali, F. (2014). Describing bibliographic references in RDF. In Proceedings of SePublica 2014. http://ceur-ws.org/Vol-1155/paper-05.pdf
5. 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.
6. 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
7. 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
8. Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. In Journal of Web Semantics, 17: 33-43. http://dx.doi.org/10.1016/j.websem.2012.08.001
9. Peroni, S., Shotton, D. (2016). Exemplar OCC dataset from Europe PubMed Central metadata. Figshare. https://dx.doi.org/10.6084/m9.figshare.3481922
10. Peroni, S., Shotton, D. (2016). Exemplar OCC dataset from Springer Nature metadata. Figshare. https://dx.doi.org/10.6084/m9.figshare.3481949
11. Peroni, S., Shotton, D. (2016). Metadata for the OpenCitations Corpus. Figshare. https://dx.doi.org/10.6084/m9.figshare.3443876
12. Peroni, S., Shotton, D., Vitali, F. (2012). Scholarly publishing and the Linked Data: describing roles, statuses, temporal and contextual extents. In Proceedings of i-Semantics 2012: 9–16. http://dx.doi.org/10.1145/2362499.2362502
13. Shotton, D. (2013). Open Citations. Nature, 502 (7471): 295–297. http://dx.doi.org/10.1038/502295a