One year of the OpenCitations Corpus
Releasing RDF-based scholarly citation data into the Public Domain

Paper (DOI): https://doi.org/10.1007/978-3-319-68204-4_19
Slides (DOI): https://doi.org/10.6084/m9.figshare.5526967

Silvio Peroni • David Shotton • Fabio Vitali
16th International Semantic Web Conference (ISWC 2017), 21-25 October 2017, Vienna, Austria

Creative Commons Attribution 4.0 International License
OpenCitations

Citations are the links that knit together our scientific and cultural knowledge.

However often they are enclosed in closed citation indexes thought only for human consumption and/or accessible only by paying significant fees.

The OpenCitations Corpus (OCC, http://opencitations.net) is a LOD repository of CC0 citation data described with SPAR Ontologies (http://www.sparontologies.net).

It provides >11M citation links from ~260,000 citing articles to ~6M cited resources + provenance information.
Ingestion workflow

We developed several scripts for implementing the ingestion workflow that populates the OpenCitations Corpus.

All the software is available on the OpenCitations GitHub repository and released as open source code with the ISC License.

These scripts implement a live and iterative process:

- **Live**: it's working while I'm speaking. It doesn't sleep, never. It's like a sentient, relentless, fast zombie – watch out!
- **Iterative**: the ingestion workflow continuously calls several external APIs to obtain new reference lists and clean metadata of the citing and cited papers.
External APIs

At present, all the reference lists are taken by processing the XML sources of the papers in the PubMed Central Open Access subset

Europe PubMed Central API for retrieving the XML sources

- We ask for the most recent papers first
- Citing papers includes articles published in 2016 and 2017
- There are 1.75M OA articles available in PubMed, according to their API. We have harvested 15% so far...

Crossref APIs to obtain additional information (title, authors, venues, etc.) about citing/cited papers, and then call the ORCID APIs to obtain ORCID(s) of the authors
Data model

Available at https://doi.org/10.6084/m9.figshare.3443876 and implemented in the OpenCitations Ontology
Resources in the OCC

<table>
<thead>
<tr>
<th>Entity type</th>
<th>What it describes</th>
<th>Count in the OCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliographic resource (br)</td>
<td>Conference papers, book chapters, journal articles, academic proceedings, books, journals, etc.</td>
<td>~7.3M</td>
</tr>
<tr>
<td>Resource embodiment (re)</td>
<td>Digital vs. print, first and ending pages, etc.</td>
<td>~4.3M</td>
</tr>
<tr>
<td>Bibliographic entry (be)</td>
<td>Textual content of a reference in a reference list</td>
<td>~10.5M</td>
</tr>
<tr>
<td>Responsible agent (ra)</td>
<td>Given name, family name and ORCID of the agent involved</td>
<td>~22.5M</td>
</tr>
<tr>
<td>Agent role (ar)</td>
<td>Author, publisher, etc.</td>
<td>~28.6M</td>
</tr>
<tr>
<td>Identifier (id)</td>
<td>DOI, PubMed ID, PubMed Central ID, ORCID, ISSN, etc.</td>
<td>~15.1M</td>
</tr>
</tbody>
</table>

Data statements: ~0.45B
Provenance statements: ~1.5B
FAIR citation data

Accessing OCC data

- Direct access to bibliographic resources by means of their HTTP URIs (via content negotiation, e.g. https://w3id.org/oc/corpus/br/1)
- SPARQL endpoint (https://w3id.org/oc/sparql)
- Monthly dumps (http://opencitations.net/download, stored in Figshare)

Dump views and downloads

![Bar chart showing views and downloads by month from September 2016 to September 2017. The chart indicates a peak in March 2017 for downloads.]
Usage of the OCC

Some known adopters: Wikidata, OpenAIRE, LOC-DB, eLife, OntoText, independent researchers (Anna Kamińska, Daniel Himmelstein, Thiago Nunes and Daniel Schwabe)

Website accesses

Use by country: United States (33.7%), France (31.3%), Bulgaria (14.3%)
OpenCitations Blog on Wordpress at https://opencitations.wordpress.com
Twitter

OpenCitations Twitter account at https://twitter.com/opencitations
Current and future infrastructure

DELL PowerEdge R730xd V4, 512GB RAM, 22TB HD + 30 Raspberry Pi 3

It's a virtual machine! It doesn't exist

We have recently received a small grant from the Sloan Foundation for one year's salary for a postdoc to develop new user interfaces, and new hardware to enhance the OCC performance - from 8M citations per year to 240M
Data are coming
END

One year of the OpenCitations Corpus

Silvio Peroni ● David Shotton ● Fabio Vitali

16th International Semantic Web Conference (ISWC 2017), 21-25 October 2017, Vienna, Austria

Creative Commons Attribution 4.0 International License