COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations

Heibi, I., Peroni, S., & Shotton, D. (2019). COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics*. https://doi.org/10.1007/s11192-019-03217-6

Ivan Heibi

Digital Humanities Advanced Research Centre (DHARC),
Department of Classical Philology and Italian Studies,
University of Bologna, Bologna (Italy)
ivan.heibi2@unibo.it

OpenCitations

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA

CC BY

About me

... - 2017

Bachelor and Master degree in **Computer Science** at the University of Bologna, Italy

2017-18

Research fellow under the supervision of **Silvio Peroni** on the "**OpenCitations** Enhancement Project", funded by the Alfred P. Sloan Foundation

2018 - ...

Ph.D. student at the Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy;
Ph.D. thesis title: ***"The application of Semantic Publishing technologies in the Science of Science research domain for the Humanities field"***

A Citation

- A **bibliographic citation** is a conceptual directional link from a **citing entity** to a **cited entity.**

- The **citation data** related to a particular citation must include:
  - The **representation** of such a **conceptual directional link**
  - The basic **metadata** of the **citing entity and the cited entity**, i.e. sufficient information to create or retrieve textual bibliographic references
- A bibliographic citation is an **open citation** when the data needed to define the citation are: **structured, separate, open;** and the **citing and cited entities must be identifiable, available**

Peroni, Shotton (2018): Open Citation: Definition. Figshare. DOI: https://doi.org/10.6084/m9.figshare.6683855

Open citations: characteristics

Joined

Unstructured

Structured (JSON Format)

```
"reference":[{
    "issue":"2",
    "key":"10.7717/peerj.4375/ref-11",
    "article-title":"Anatomy of green open access",
    "volume":"65",
    "author":"Björk",
    "year":"2014",
    "journal-title":"Journal of the Association for
},
...
```

Identifiable: "DOI":"10.1002/asi.22963"

Available (E.g. via HTTP)

Separate (e.g. via REST calls): https://api.crossref.org/works/10.7717/peerj.4375

Open

"No claims of ownership to individual items of bibliographic metadata" https://api.crossref.org

Closed

"Estimation of WOS costs is about $100,000 per year for large organizations [...] the cost of Scopus database is about 85-95% of the cost of WOS for the same organizations" https://doi.org/10.5539/ass.v9n5p18

OpenCitations

OpenCitations (https://opencitations.net) is a **scholarly infrastructure organization,** and one of the **founders of the Initiative for Open Citations (I4OC)**

It works on:

- **advocacy for open citations**
- the **publication of open bibliographic and citation data** by the use of Semantic Web technologies, and **RDF for its description**

It provides:

- **Data models**: the OpenCitations Data Model (based on SPAR Ontologies)
- **Datasets (in CC0)**: OpenCitations Corpus, and Citation Indexes
- **Software**: GitHub repository released with open source licenses
- **Online services**: dumps, REST APIs, SPARQL endpoints, and interfaces

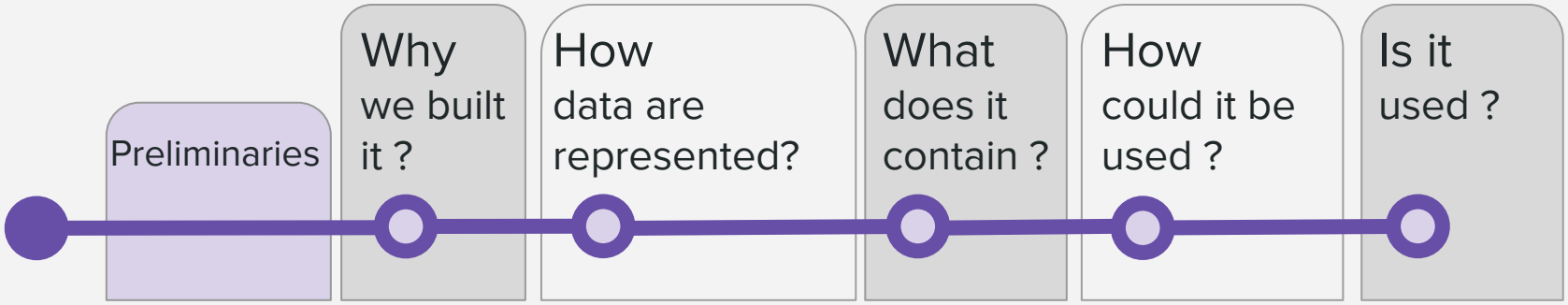
About the datasets

- OpenCitations **Corpus** (OCC, https://opencitations.net/corpus): new instance was set up at the University of Bologna in early July 2016, and currently contains **~14M citation links** to over **7.5M cited resources**
- OpenCitations **Indexes** (https://opencitations.net/index, launched in July 2018): contain **~445M citations** between **~46M bibliographic entities**

COCI, is the main dataset here

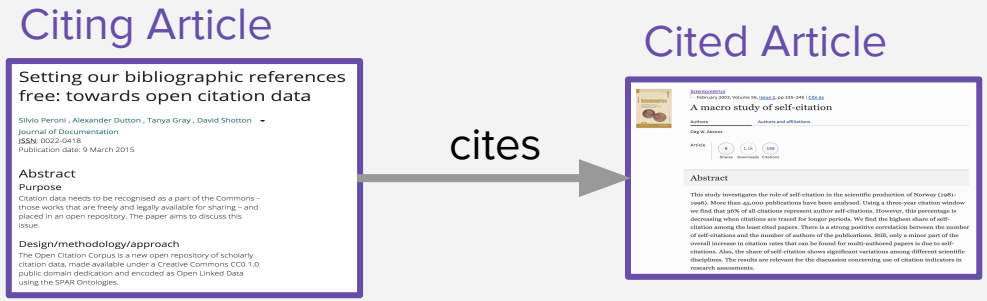
COCI

The OpenCitations Index of Crossref open DOI-to-DOI citations, is the **first of the indexes** proposed by OpenCitations, in which **citations are exposed as first-class data entities with accompanying properties** (https://opencitations.net/index/coci)

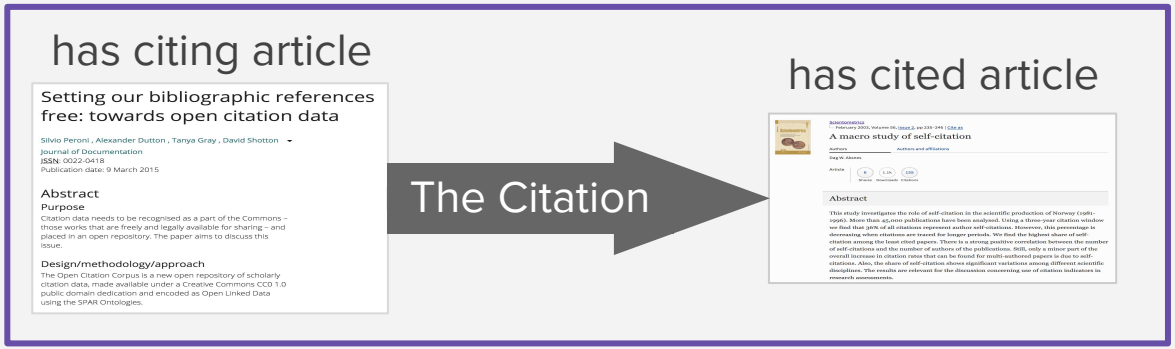


Citations as first-class data entities

Citations are normally treated simply as the links between published entities



Alternative richer view is to regard a **citation as a data entity** in its own right



More info at https://opencitations.wordpress.com/2018/02/19/citations-as-first-class-data-entities-introduction/

Open Citation Identifier (OCI)

We defined the **Open Citation Identifier (OCI)**, a persistent identifier scheme for citations contained in bibliographic databases

Structure: **oci:number-number,** where “oci:” is the identifier prefix

Some examples:

- *oci:01027931310-01022252312* (citation in Wikidata, identified by “010”)
- *oci:020010108063601070506630807020263066305090200101080636010705066308070202630563030*1 (citation in Crossref, identified by “020”)
- *oci:0302544384-0307295288* (citation in the OCC, identified by the “030”)

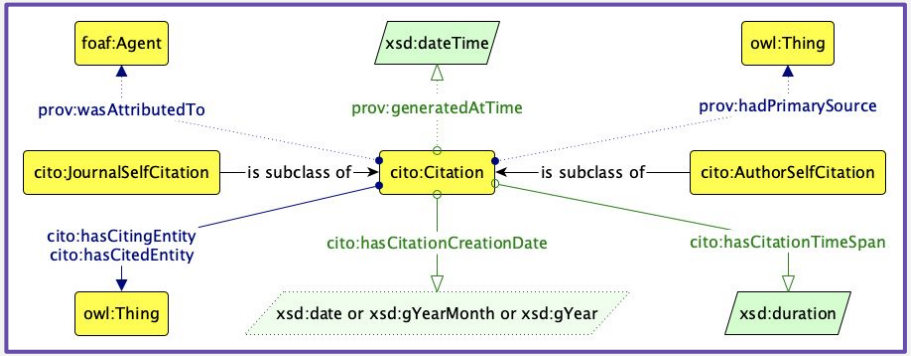
Peroni S, Shotton D (2019). Open Citation Identifier: Definition. Figshare. DOI: https://doi.org/10.6084/m9.figshare.7127816

Why we built it ?

- **Citations have a significant value** to the academic community and the general public, from different perspectives:
  - **Topologically:** through the definition of citing-cited graph evolution over time;
  - **Sociologically:** identifying researchers behaviours, or elitist access paths;
  - **Quantitatively:** creating citation-based metrics for impact evaluation;
  - **Financially:** defining the researcher's scholarly “value” within their communities.
- A large dataset embedding citations with multiple characteristics/metadata will enforce the above benefits, and **its representation in RDF will help us query and apply novel methodologies for different discoveries.**

How data are represented ? (model)

- The OpenCitations **data representation is based on OCDM** (OpenCitations Data Model). It has been defined for **encoding scholarly bibliographic and citation data in RDF.** The **SPAR Ontologies** are the core elements needed to establish a semantical meaning for the data entities, and to define the relations between them.
- **To define citations as first-class entities**, and their provenance, the **OCDM** uses the **Citation Typing Ontology** (CiTO, http://purl.org/spar/cito), which is part of the SPAR Ontologies.
- A new version of the OCDM will be released soon



Peroni, S., & Shotton, D. (2018). The OpenCitations Data Model. Figshare. https://doi.org/10.6084/m9.figshare.3443876

How data are represented ? (citation characteristics)

Characteristic	Description	In COCI
citing entity	The bibliographic entity which acts as source for the citation.	DOI (e.g. 10.1108/JD-12-2013-0166)
cited entity	The bibliographic entity which acts as target for the citation.	DOI (e.g. 10.1001/jama.295.1.90)
citation creation date	The date on which the citation was created. This has the same numerical value as the publication date of the citing bibliographic resource, but is a property of the citation itself. When combined with the citation time span, it permits that citation to be located in history.	A date in yyyy-mm-dd format (e.g. 2018-03-15)
citation timespan	The temporal characteristic of a citation, namely the interval between the publication date of the cited entity and the publication date of the citing entity.	Duration in PnYnMnD format, such that: nY: number of years; nM: number of months; nD: number of days. (e.g. P4Y3M)
type	A classification of the citation according to particular dimensions, e.g. whether or not it is a self-citation.	Check if it is a journal self citation or an author self citation

What does it contain ?

- COCI was first created and **released on July 4, 2018**. The most recent update to it, has been made on **November 2018**, and it contains **445,826,118 citations** between **46,534,705 bibliographic entities**. These are stored by means of **2,259,134,894 RDF statements** (around 5 per citation)
- An upcoming new extended version of COCI is planned to be released in the following weeks.

Publisher	Outgoing citations	Incoming citations
Springer Nature	79,860,827.000	52,257,862.000
Wiley	76,819,685.000	48,174,542.000
Informa UK Limited	41,433,917.000	14,975,989.000
Institute of Electrical and Electronics Engineers (IEEE)	30,114,985.000	20,940,703.000
SAGE Publications	15,933,805.000	7,915,082.000
American Physical Society (APS)	15,729,297.000	16,065,862.000
AIP Publishing	10,130,022.000	8,455,097.000
Ovid Technologies (Wolters Kluwer Health)	9,971,274.000	12,840,293.000
Oxford University Press (OUP)	9,891,000.000	11,466,659.000
Elsevier	2,853,739.000	96,310,027.000

How could it be used ? (REST API Service)

- A **REST API Service** implemented by means of RAMOSE, the Restful API Manager Over SPARQL Endpoints (https://opencitations.net/index/coci/api/v1)

  Users can easily retrieve:

  - **Citations and References** of a specified bibliographic item identified by a DOI
  - **The citation data** for a precise citation identified by an OCI
  - **Metadata** of the bibliographic items identified by specific DOIs

Usage example:

To get the list of citations received by the article identified by the DOI= 10.1002/adfm.201505328.

https://w3id.org/oc/index/coci/api/v1/citations/10.1002/adfm.201505328

***Note:** results format, CSV or JSON, could be specified with the "?format" parameter

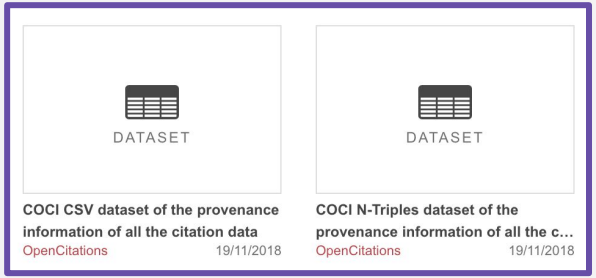

```
[
  {
    "oci": "020010000023619-020010000023610",
    "timespan": "P11M20D",
    "citing": "10.1002/jrs.5087",
    "creation": "2017-02-06",
    "author_sc": "no",
    "journal_sc": "no",
    "cited": "10.1002/adfm.201505328"
  },
  ...
]
```

How could it be used ? (data dumps)

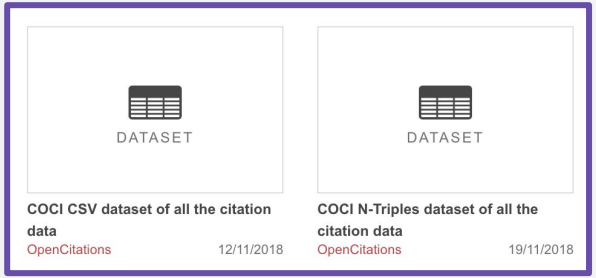
- All the citation data and provenance information of COCI are available as dumps stored in **Figshare** in both **CSV** and **N-Triples** (for RDF graphs) formats, while a dump of **the whole triplestore** is available on **The Internet Archive.** (https://opencitations.net/download#coci)

https://archive.org/details/coci-triplestore-2018-10-03

Provenance:
https://doi.org/10.6084/m9.figshare.6741431.v3
https://doi.org/10.6084/m9.figshare.6741446.v3

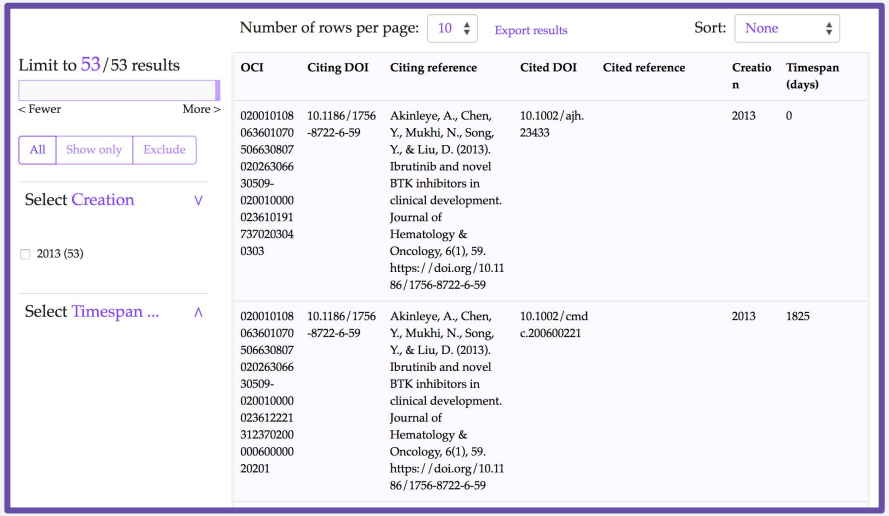


Data:
https://doi.org/10.6084/m9.figshare.6741422.v3
https://doi.org/10.6084/m9.figshare.6741425.v3



How could it be used ? (interfaces)

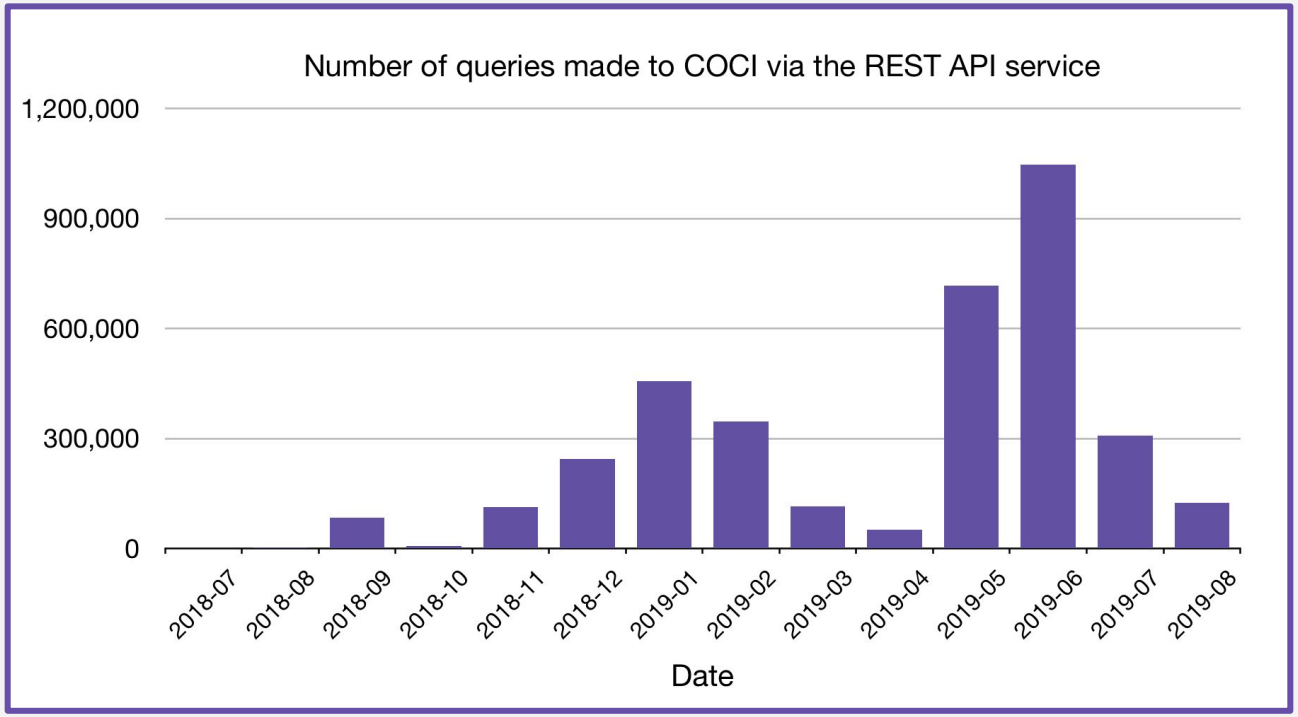
- **Searching and browsing interfaces** (https://opencitations.net/index/search)
  Two interfaces have been developed by means of OSCAR, the OpenCitations RDF Search Application, and LUCINDA, the OpenCitations RDF Resource Browser.



- **Open Citation Index SPARQL endpoint** (https://w3id.org/oc/index/sparql)

Heibi, I., Peroni, S., & Shotton, D. Enabling text search on SPARQL endpoints through OSCAR. Data Science, 1-23. https://doi.org/10.3233/DS-190016

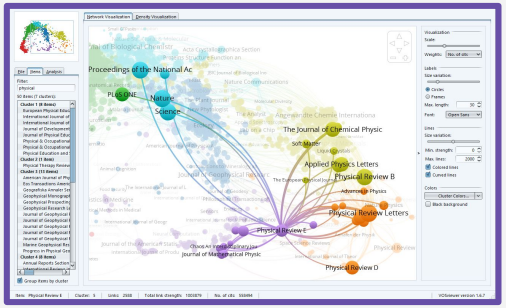
Is it used ? (quantitative analysis)



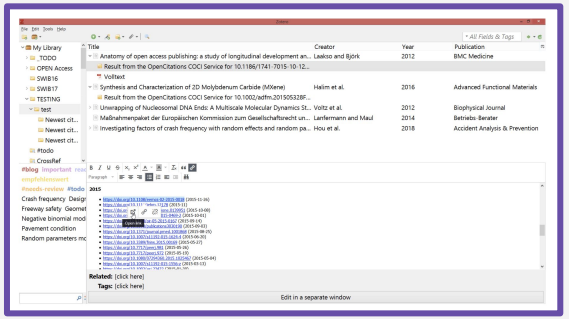
- The Figshare COCI dumps register: **682 downloads for the CSV data dump,** and **128 for the N-Triples data dump.**

Is it used ? (community uptake)

- VOSviewer: a tool for constructing and visualizing bibliometric networks. (https://www.vosviewer.com/)
  - Developed by CWTS, Leiden University, The Netherlands (https://www.cwts.nl/)



- Zotero: a free, easy-to-use tool to help users collect, organize, cite, and share research. The Open Citations Plugin with COCI has been released, (https://github.com/zuphilip/zotero-open-citations)



- Citation Gecko (http://citationgecko.com)
- OCI Graphe (https://tinyurl.com/y3lkzqjq)
- citecorp: (https://github.com/ropenscilabs/citecorp)

- **Used by several studies:** Stephen Pearson presentation (https://tinyurl.com/y6ll7fkv); Di Iorio et al. (2019) (https://arxiv.org/abs/1902.03287); Thomas Donoghue (https://doi.org/10.21105/joss.01674); Chun-Kai (Karl) Huang et al. (https://doi.org/10.1101/750075)

Thank you for your attention

COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations

Ivan Heibi

Digital Humanities Advanced Research Centre (DHARC),
Department of Classical Philology and Italian Studies,
University of Bologna, Bologna (Italy)
ivan.heibi2@unibo.it – @ivanheib – https://ivanhb.github.io

OpenCitations

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

CC BY