

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



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



About me



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, identifiable, available**

Open citations: characteristics

The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles

PeerJ

View 433 tweets

Related research

Heather Piwowar *, Jason Priem *, Vincent Larivière , Juan Pablo Alperin , Lisa Matthias , Bree Norlander , Ashley Farley , Jevin West , Stefanie Haustein 

Published February 13, 2018

 Note that a Preprint of this article also exists, first published August 2, 2017.

PubMed 29456894

- Author and article information
- Abstract

Joined

References

- ▼ Björk BC, Laakso M, Welling P, Paetav P. 2014. Anatomy of green open access. *Journal of the Association for Information Science and Technology* 65(2):237–250.

Anderson. 2017b. The forbidden forecast: thinking about open access and library subscriptions. The Scholarly Kitchen. <https://scholarlykitchen.sspnet.org/2017/07/> (accessed 15 July 2017)

Antelman K. 2017. Leveraging the growth of open access in library collection decision making. In: Proceeding from ACRL 2017: at the helm: leading transformation.

Archambault É, Amyot D, Deschamps P, Nicol A, Provencher F, Rebout L, Roberge G. 2013. Proportion of open access peer-reviewed papers at the European and world levels—2004–2011. European Commission, Brussels

Archambault É, Amyot D, Deschamps P, Nicol AF, Provencher F, Rebout L, Roberge G. 2014. Proportion of open access papers published in peer-reviewed journals at the European and world levels—1996–2013. European Commission

Archambault É, Côté G, Struck B, Voorons M. 2016. Research impact of paywalled versus open access papers.

Unstructured



Closed

"reference": [{"issue": "2", "key": "10.7717/peerj.4375/ref-11"}, Structured (JSON Format)

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

Available
(E.g. via HTTP)

```
"article-title": "Anatomy of green open access",
"volume": "65",
"author": "Björk",
"year": "2014",
"journal-title": "Journal of the Association for Information Science and Technology",  
},  
...
```

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>

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

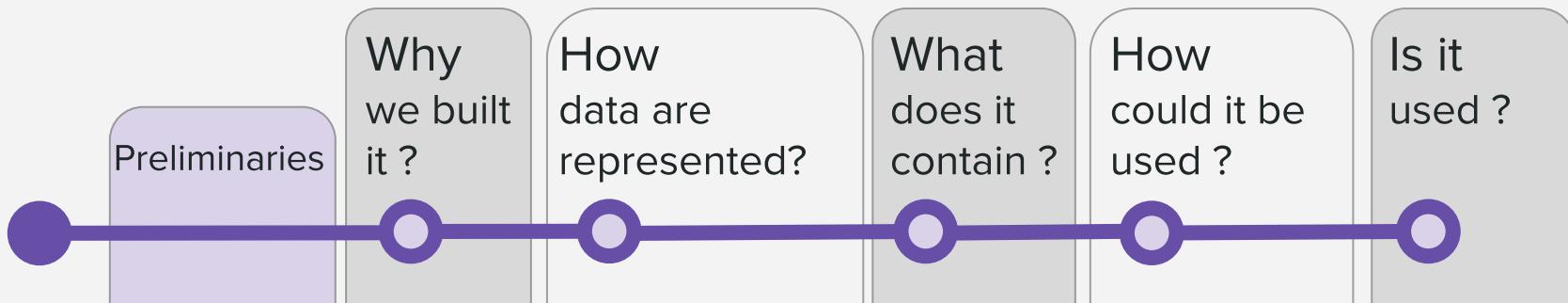
- **OpenCitations Corpus** (OCC, <https://opencitations.net/corpus>): new instance was set up at the University of Bologna in early July 2016, and currently contains **~14M citation links** to over **7.5M cited resources**
- **OpenCitations Indexes** (<https://opencitations.net/index>, launched in July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

Citing Article

Setting our bibliographic references free: towards open citation data

Silvio Peroni, Alexander Dutton, Tanya Gray, David Shotton
Journal of Documentation
ISSN: 0022-0418
Publication date: 9 March 2015

Abstract Purpose

Citation data needs to be recognised as a part of the Commons – those works that are freely and legally available for sharing – and placed in an open repository. The paper aims to discuss this issue.

Design/methodology/approach

The Open Citation Corpus is a new open repository of scholarly citation data, made available under a Creative Commons CC0 1.0 public domain dedication and encoded as Open Linked Data using the SPAR Ontologies.

cites

Cited Article

Information 2010, Volume 40, Issue 2, pp 239-244 | DOI:10.1108/00220411011034462

A macro study of self-citation

Authors: Silvio N. Peroni

Article | 4 | LIn | 386 | Author biography | Author information

Abstract

This study investigates the role of self-citation in the scientific production of Norway (1990–1999). More than 45,000 publications have been analysed. Using a three-year citation window we find that, on average, all citations are self-citations. However, the share of self-citations decreases as decreasing when citations are traced for longer periods. We find the highest share of self-citations in the field of engineering and technology. The share of self-citations increases as increasing the number of self-citations and the number of authors of the publications. Still, only a minor part of the overall increase in the number of self-citations can be found for the publications in the field of social sciences. Also, the share of self-citation shows significant variations among different scientific disciplines. The results are relevant for the discussion concerning use of citation indicators in research assessments.

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

has citing article

Setting our bibliographic references free: towards open citation data

Silvio Peroni, Alexander Dutton, Tanya Gray, David Shotton
Journal of Documentation
ISSN: 0022-0418
Publication date: 9 March 2015

Abstract

Purpose
Citation data needs to be recognised as a part of the Commons – those works that are freely and legally available for sharing – and placed in an open repository. The paper aims to discuss this issue.

Design/methodology/approach

The Open Citation Corpus is a new open repository of scholarly citation data, made available under a Creative Commons CC0 1.0 public domain dedication and encoded as Open Linked Data using the SPAR Ontologies.

The Citation

has cited article

Information 2010, Volume 40, Issue 2, pp 239-244 | DOI:10.1108/00220411011034462

A macro study of self-citation

Authors: Silvio N. Peroni

Article | 4 | LIn | 386 | Author biography | Author information

Abstract

This study investigates the role of self-citation in the scientific production of Norway (1990–1999). More than 45,000 publications have been analysed. Using a three-year citation window we find that, on average, all citations are self-citations. However, the share of self-citations decreases as decreasing when citations are traced for longer periods. We find the highest share of self-citations in the field of engineering and technology. The share of self-citations increases as increasing the number of self-citations and the number of authors of the publications. Still, only a minor part of the overall increase in the number of self-citations can be found for the publications in the field of social sciences. Also, the share of self-citation shows significant variations among different scientific disciplines. The results are relevant for the discussion concerning use of citation indicators in research assessments.

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:

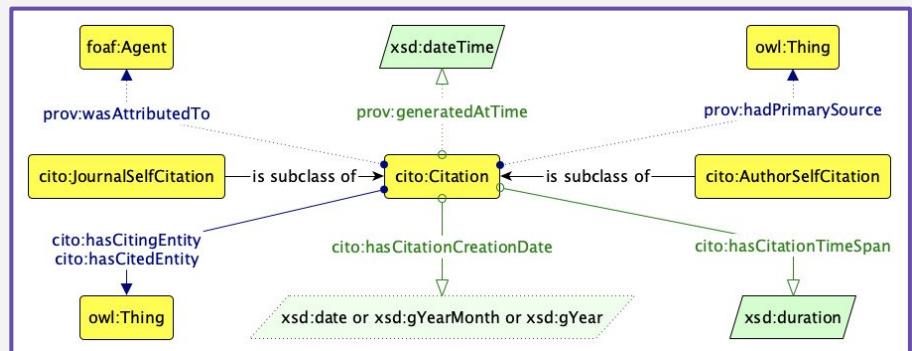
- `oci:01027931310-01022252312` (citation in Wikidata, identified by “010”)
- `oci:02001010806360107050663080702026306630509-02001010806360107050663080702026305630301` (citation in Crossref, identified by “020”)
- `oci:0302544384-0307295288` (citation in the OCC, identified by the “030”)

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



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
<i>Institute of Electrical and Electronics Engineers (IEEE)</i>	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
<i>Ovid Technologies (Wolters Kluwer Health)</i>	9,971,274.000	12,840,293.000
Oxford University Press (OUP)	9,891,000.000	11,466,659.000
<i>Elsevier</i>	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**.
[\(<https://opencitations.net/download#cocci>\)](https://opencitations.net/download#cocci)

<https://archive.org/details/cocci-triplestore-2018-10-03>



Provenance:

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

Two rectangular cards, each labeled "DATASET" with a small icon above it. The left card is titled "COCI CSV dataset of the provenance information of all the citation data" and the right card is titled "COCI N-Triples dataset of the provenance information of all the c...". Both cards include the text "OpenCitations" and the date "19/11/2018" at the bottom.

Data:

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

Two rectangular cards, each labeled "DATASET" with a small icon above it. The left card is titled "COCI CSV dataset of all the citation data" and the right card is titled "COCI N-Triples dataset of all the citation data". Both cards include the text "OpenCitations" and the date "19/11/2018" at the bottom.

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.

The screenshot shows a search results page with the following details:

Number of rows per page: 10 Export results Sort: None

Limit to 53/53 results

< Fewer More >

All Show only Exclude

Select Creation ▾

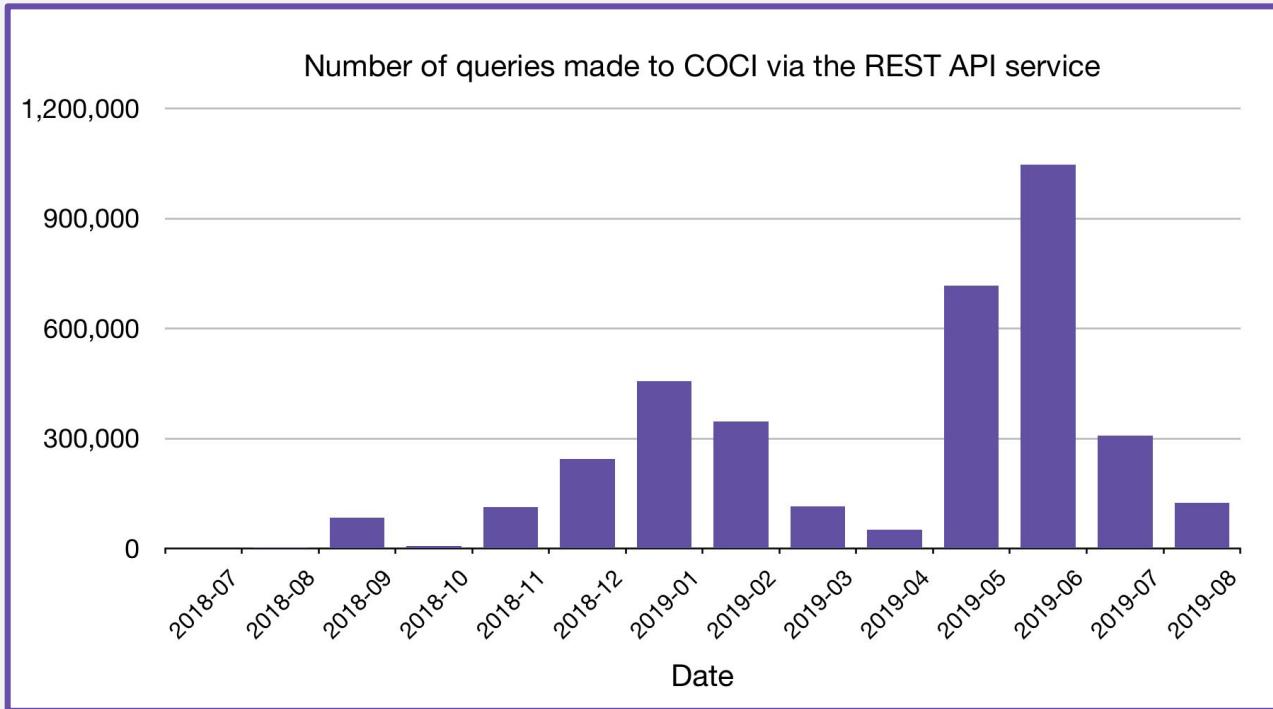
2013 (53)

Select Timespan ... ▾

OCI	Citing DOI	Citing reference	Cited DOI	Cited reference	Creation	Timespan (days)
020010108	10.1186/1756-8722-6-59	Akinleye, A., Chen, Y., Mukhi, N., Song, Y., & Liu, D. (2013). Ibrutinib and novel BTK inhibitors in clinical development. <i>Journal of Hematology & Oncology</i> , 6(1), 59. https://doi.org/10.1186/1756-8722-6-59	10.1002/ajh.23433		2013	0
020263066		Ibrutinib and novel BTK inhibitors in clinical development.				
30509-		Journal of Hematology & Oncology, 6(1), 59. https://doi.org/10.1186/1756-8722-6-59				
020010000						
023610191						
737020304						
0303						
020010108	10.1186/1756-8722-6-59	Akinleye, A., Chen, Y., Mukhi, N., Song, Y., & Liu, D. (2013). Ibrutinib and novel BTK inhibitors in clinical development. <i>Journal of Hematology & Oncology</i> , 6(1), 59. https://doi.org/10.1186/1756-8722-6-59	10.1002/cmd.c.200600221		2013	1825
020263066		Ibrutinib and novel BTK inhibitors in clinical development.				
30509-		Journal of Hematology & Oncology, 6(1), 59. https://doi.org/10.1186/1756-8722-6-59				
020010000						
023612221						
312370200						
000600000						
20201						

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

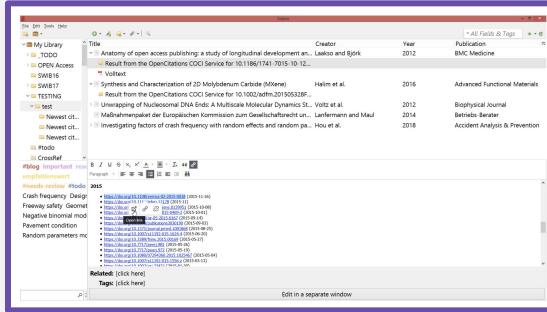
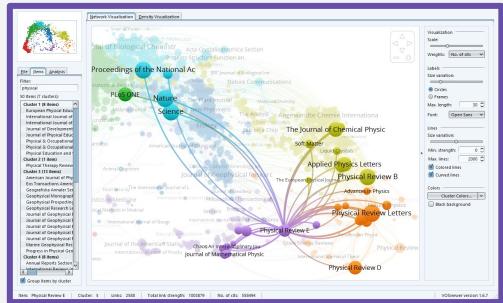
Is it used ? (quantitative analysis)



- 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/y6ll7fky>); 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://ivanheib.github.io) – <https://ivanhb.github.io>



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

