CEUR-WS.org/Vol-3220/keynote.pdf

# OpenCitations: a short introduction

Silvio Peroni[1,2]

[1]Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

[2]Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

### Abstract

In this paper, I introduce a brief history of open citations, their main characteristics and use in the context of OpenCitations, a scholarly infrastructure organisation dedicated to open scholarship and the publication of open bibliographic and citation data using Semantic Web technologies.

### Keywords

OpenCitations, open citation data, open bibliographic metadata, Semantic Web

## 1. The origins of open citations

The concept of *open citations* [1] is strongly tied with that of the Web. Since 1989, the Web has drastically changed how we think about academic publishing and science. Publishers have adopted Web Standards to create and deliver their products quickly and to a broader audience. Standards, guidelines, and services based on Web technologies have been proposed in the past 30 years to increase the discoverability of academic products and publications, improve research practices and allow reusability of scholarly data in different applicative contexts. Open citations are no exception.

Even if the definition of open citations has been introduced recently, past works implicitly started to highlight their main characteristics. As far as I know, the first embryonal description of open citations is in Robert Cameron's visionary article published in 1997 [2]. In this article, he speculated about the existence of a decentralised and freely available *Universal Citation Database*. Such a database would have had daily updates and links to every scholarly work, providing information for all types of publications (from journal articles to technical reports, datasets, and other publication types) and being equally visible and accessible to all.

From this initial Web age, things have started to develop. In the same year of Cameron's article, *CiteSeer* was established [3], a service that crawled citations from PostScript documents available on the Web. Along the same lines, a few years later, *CiteBase* was created in the context of the *OpCit* project [4]. In 2004, *Google Scholar* (https://scholar.google.com) was launched to provide one of the first open Web interfaces for looking at a scientist's paper and the citations

*ULITE-ws: Understanding Literature References in Academic Full Text at JCDL 2022*

silvio.peroni@unibo.it (S. Peroni)
https://essepuntato.it (S. Peroni)
0000-0003-0530-4305 (S. Peroni)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

that this paper received (even if the data are not openly accessible). A few years later, *CiteSeerX* [5] was proposed as an evolution to CiteSeer to address some problems of its predecessor.

However, the tipping point for open citations was when, in 2009, David Shotton introduced the concept of *semantic publishing* [6], which concerns the use of Semantic Web technologies applied to the scholarly publishing domain to make journal articles and other scholarly publications more discoverable and reusable. This idea led him to the JISC OpenCitations project in 2010 (https://opencitations.wordpress.com/2010/07/15/jisc-open-citations-aims-objectives-and-final-outputs/), a year-long project (with a subsequent extension) that aimed at creating the first corpus of open citation data entirely available on the Web by using URLs to identify resources and RDF to expose these data to the public.

The idea of providing open citations was spread to the scholarly and publishing community in the following years in two different editions of the Annual Conference of Open Access Publishers (OASPA, https://oaspa.org/conference/). Both David Shotton's talk at OASPA 2013 and Dario Taraborelli's speech at OASPA 2016 highlighted the essential need to release citation data as soon as possible for the whole scholarly community.

Since 2016, everything has started to change on a large scale. The importance of open citations got a broader audience and led to the introduction of *OpenCitations* [7] and *WikiCite* as testimonials of communities providing open citation data and services to allow their programmatic access. After these first technical implementations, in 2017, the *Initiative for Open Citations* (I4OC, https://i4oc.org) was launched to convince publishers to make their reference lists free and openly available on *Crossref* (https://crossref.org) [8]. In the following years, other international events and scholarly initiatives helped increase the interest in open citations and related technical infrastructures. The successful movement toward public domain citation data is now more strong than ever and "improve the transparency and robustness of scientific portfolio analysis, improve science policy decision-making, stimulate downstream commercial activity, and increase the discoverability of scientific articles" [9].

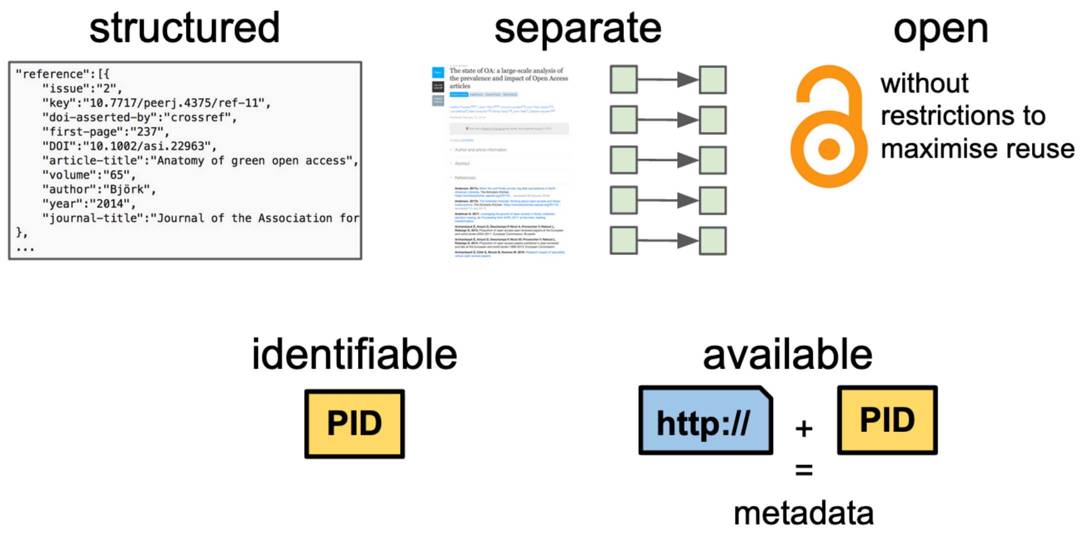
## 2. What is an open citation

In the previous section, I used the concept of open citations several times. However, I have not clarified what it is about and the characteristics a citation must have to be claimed as open. However, first, it is necessary to explain what I refer to when I mention the word *citation*.

A bibliographic citation is a conceptual directional link from a citing entity to a cited entity to acknowledge or ascribe credit for the contribution made by the authors of the cited entity. This link is defined using particular textual devices such as a bibliographic reference in the reference list, denoted by an in-text reference pointer – e.g. "[3]" or "(Doe et al., 2013)" – within the body of the citing entity.

The citation data related to a particular citation must include the representation of such a conceptual directional link and the basic metadata of the citing entity and the cited entity, i.e. sufficient information to create or retrieve textual bibliographic references for each of the entities involved in the citation (i.e. the citing entity and the cited entity).

A bibliographic citation is an *open citation* when the data needed to define the citation are compliant with the following principles [10]:



```
"reference":[{
    "issue":"2",
    "key":"10.7717/peerj.4375/ref-11",
    "doi-asserted-by":"crossref",
    "first-page":"237",
    "DOI":"10.1002/asi.22963",
    "article-title":"Anatomy of green open access",
    "volume":"65",
    "author":"Björk",
    "year":"2014",
    "journal-title":"Journal of the Association for
},
...
```

structured
separate
The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles
open
without restrictions to maximise reuse
identifiable
PID
available
http://
+
PID
=
metadata

**Figure 1:** The five principles citation data must comply with to talk about an *open* citation.

- **structured** – citation data must be expressed in one or more machine-readable formats such as JSON or RDF;
- **separate** – citation data must be available without the need to access the source bibliographic entity (e.g. the article or book) in which the citation is defined, which can be even behind a paywall;
- **open** – citation data must be freely accessible and reusable without restrictions, for example, by publication under the CC0 1.0 Universal waiver/license;
- **identifiable** and **available** – citing and cited entities must be identified by using a specific persistent identifier scheme (e.g. a DOI) or a URL. In addition, by resolving the identifiers of the citing and cited entities, it must be possible to obtain the basic metadata of both entities, sufficient to create or retrieve textual bibliographic references for each of them. Such basic entity metadata must also be structured, separate and open.

These principles have been thoroughly followed in the technical developments of OpenCitations, introduced in the following section.

## 3. OpenCitations, a scholarly infrastructure organisation

OpenCitations (https://opencitations.net) [7], of which I am proudly one of its directors, is a scholarly infrastructure organisation dedicated to open scholarship and the publication of open bibliographic and citation data using Semantic Web technologies. We also undertake advocacy for open scholarly metadata, mainly via the *Initiative for Open Citations* (I4OC, https://i4oc.org) and the *Initiative for Open Abstracts* (I4OA, https://i4oa.org). Our goal is to provide open metadata with a scope, depth, accuracy and provenance surpassing commercial sources.

We provide the OpenCitations Data Model [11] that we use to describe all the bibliographic metadata and citation data OpenCitations provides. Of course, we also provide bibliographic and citation data (all released using the CC0 waiver to maximise their reuse), available in different collections, including the OpenCitations Indexes, our primary collection. In addition, all the software we developed to gather and expose these data is available in our GitHub repository (https://github.com/opencitations) and released with open source licenses. Finally, all the data are available online: full dumps of OpenCitations data can be downloaded and accessed programmatically via REST APIs, SPARQL endpoints, and other Web interfaces.

Our primary database, COCI (the OpenCitations Index of Crossref open DOI-to-DOI citations) [12], currently hosts more than 1.29 billion citations. All these citations have been made available in Linked Open Data. They can be accessed programmatically using our REST API by specifying either publication' DOI or the Open Citation Identifier (OCI) [13] identifying the complete citation, i.e. the relation *entity A cites entity B*.

Since 2020, OpenCitations has significantly benefited from crowdfunding from the scholarly community, which has resulted from (a) the Global Sustainability Coalition for Open Science Services's (SCOSS, https://scoss.org) selection of OpenCitations as a scholarly infrastructure worthy of support, and (b) its involvement in international projects, such as the OpenAIRE-Nexus project (https://www.openaire.eu/openaire-nexus-project) and RISIS project (https://www.risis2.eu/).

OpenCitations espouses the UNESCO principles of Open Science [14], the Principle of Open Scholarly Infrastructures [15], the FAIR data principles that data should be Findable, Accessible, Interoperable, and Reusable [16], and the I4OC principles that citation data should be Structured, Separable, and Open (https://i4oc.org/#goals). In compliance with these values, one of OpenCitations' main priorities is to keep its services, software, and data always without charge under open licenses (CC0 for data and ISC for software) to foster their maximum reuse.

## 4. Conclusions and future directions

This undeniable aspect of keeping all OpenCitations data and services free leads to an acknowledged sustainability issue, principally in terms of salaries and technical infrastructure costs. OpenCitations can rely on an international network of generous supporters that apply for membership and donation programmes. We are grateful to the institutions that believe in our mission and values. However, we are already far from being a fully financially sustained infrastructure, and we still need help from the global scholarly community to keep open bibliographic and citation data and related services available for many years and to reach the following goals:

- to provide high-quality metadata with full provenance relating to scholarly publications and the citations that link them, including those in areas such as the humanities and social sciences, the global south, and non-English publications;
- to expand our coverage into the 'grey literature' of reports, patents, datasets, software, etc.;
- to surpass in terms of coverage and quality – and thereby provide an open and free alternative to – the major commercial citation indexes;

- to provide the open data crucial for research in bibliometrics and scientometrics, and the creation of transparent and reproducible metrics for research assessment;
- to continue developing and making public free and open-source software (FOSS) with relevant functionality and our open services built over our data.

We (scholars, institutions, founders, etc.) can make a difference and create an open, inclusive future for science and research. OpenCitations is a plural: together, we are OpenCitations.

## Acknowledgments

I want to thank David Shotton and Chiara Di Giambattista, who provided insightful discussions, materials and feedback which has been included in this paper, and all the people working for and supporting OpenCitations. This work has been funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 101017452 (OpenAIRE-Nexus).

## References

[1] D. Shotton, Open citations, Nature 502 (2013) 295–297. doi:10.1038/502295a.

[2] R. D. Cameron, A Universal Citation Database As a Catalyst For Reform In Scholarly Communication, First Monday 2 (1997). doi:10.5210/fm.v2i4.522.

[3] C. L. Giles, K. D. Bollacker, S. Lawrence, CiteSeer: an automatic citation indexing system, in: Proceedings of the third ACM conference on Digital libraries - DL '98, ACM Press, Pittsburgh, Pennsylvania, United States, 1998, pp. 89–98. doi:10.1145/276675.276685.

[4] T. Brody, S. Harnad, L. Carr, Earlier Web usage statistics as predictors of later citation impact, Journal of the American Society for Information Science and Technology 57 (2006) 1060–1072. doi:10.1002/asi.20373.

[5] H. Li, I. Councill, W.-C. Lee, C. L. Giles, CiteSeerx: an architecture and web service design for an academic document search engine, in: Proceedings of the 15th international conference on World Wide Web - WWW '06, ACM Press, Edinburgh, Scotland, 2006, p. 883. doi:10.1145/1135777.1135926.

[6] D. Shotton, Semantic publishing: the coming revolution in scientific journal publishing, Learned Publishing 22 (2009) 85–94. doi:10.1087/2009202.

[7] S. Peroni, D. Shotton, OpenCitations, an infrastructure organization for open scholarship, Quantitative Science Studies 1 (2020) 428–444. doi:10.1162/qss_a_00023.

[8] G. Hendricks, D. Tkaczyk, J. Lin, P. Feeney, Crossref: The sustainable source of community-owned scholarly metadata, Quantitative Science Studies 1 (2020) 414–427. doi:10.1162/qss_a_00022.

[9] B. I. Hutchins, A tipping point for open citation data, Quantitative Science Studies 2 (2021) 433–437. doi:10.1162/qss_c_00138.

[10] S. Peroni, D. Shotton, Open Citation: Definition, 2018. URL: https://doi.org/10.6084/m9.figshare.6683855, version 1.0.

[11] M. Daquino, S. Peroni, D. Shotton, G. Colavizza, B. Ghavimi, A. Lauscher, P. Mayr, M. Romanello, P. Zumstein, The OpenCitations Data Model, in: The Semantic Web – ISWC

2020, volume 12507 of *Lecture Notes in Computer Science*, Springer, Cham, Switzerland, 2020, pp. 447–463. doi:10.1007/978-3-030-62466-8_28.

[12] I. Heibi, S. Peroni, D. Shotton, Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations, Scientometrics 121 (2019) 1213–1228. doi:10.1007/s11192-019-03217-6.

[13] S. Peroni, D. Shotton, Open Citation Identifier: Definition, 2019. URL: https://doi.org/10.6084/m9.figshare.7127816.

[14] UNESCO, UNESCO Recommendation on Open Science, Programme and meeting document SC-PCB-SPP/2021/OS/UROS, 2021. URL: https://unesdoc.unesco.org/ark:/48223/pf0000379949.

[15] G. Bilder, J. Lin, C. Neylon, The Principles of Open Scholarly Infrastructure, 2020. URL: https://doi.org/10.24343/C34W2H.

[16] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. doi:10.1038/sdata.2016.18.