DE GRUYTER
OPEN

Research Paper

Can Crossref Citations Replace Web of Science for Research Evaluation? The Share of Open Citations

Tomáš Chudlarský[1,2], Jan Dvořák[1,2†]

[1]Czech Technical University in Prague, Computing and Information Centre, Jugoslávských partyzánů 3, CZ-16000 Praha 6, Czech Republic
[2]Charles University, Institute of Information Studies and Librarianship, Na Příkopě 29, CZ-11000 Praha 1, Czech Republic

Citation: Chudlarský, Tomáš, and Jan Dvořák. "Can crossref citations replace web of science for research evaluation? The share of open citations." *Journal of Data and Information Science* (2020). https://doi.org/10.2478/jdis-2020-0037

Received: Feb. 8, 2020
Revised: Jun. 19, 2020; Jul. 29, 2020
Accepted: Aug. 6, 2020

Abstract

Purpose: We study the proportion of Web of Science (WoS) citation links that are represented in the Crossref Open Citation Index (COCI), with the possible aim of using COCI in research evaluation instead of the WoS, if the level of coverage was sufficient.

Design/methodology/approach: We calculate the proportion on citation links where both publications have a WoS accession number and a DOI simultaneously, and where the cited publications have had at least one author from our institution, the Czech Technical University in Prague. We attempt to look up each such citation link in COCI.

Findings: We find that 53.7% of WoS citation links are present in the COCI. The proportion varies largely by discipline. The total figures differ significantly from 40% in the large-scale study by Van Eck, Waltman, Larivière, and Sugimoto (blog 2018, https://www.cwts.nl/blog?article=n-r2s234).

Research limitations: The sample does not cover all science areas uniformly; it is heavily focused on Engineering and Technology, and only some disciplines of Natural Sciences are present. However, this reflects the real scientific orientation and publication profile of our institution.

Practical implications: The current level of coverage is not sufficient for the WoS to be replaced by COCI for research evaluation.

Originality/value: The present study illustrates a COCI vs WoS comparison on the scale of a larger technical university in Central Europe.

Keywords Open citations; Crossref Open Citation Index; Web of Science; Current Research Information System

† Corresponding author: Jan Dvořák (E-mail: jan.dvorak@ff.cuni.cz).

Research Paper

1 Introduction

The adoption of the Digital Object Identifiers (DOIs, see the DOI Handbook) by publishers of scholarly works is advancing. DOIs are persistent identifiers with a resolution service and a set of metadata about the referenced resources. Scholarly publishing DOI registration is almost exclusively operated by the Crossref DOI registration agency (Crossref). An important part of the metadata that is deposited with Crossref is the list of references, which can be aggregated as the network of citation links between scholarly works. The COCI project (OpenCitations, 2018) makes openly available the citation links from Crossref that are marked as open. This presents an open alternative to commercial citation databases such as Web of Science (WoS, by Clarivate Analytics) which only offer citation data limited by restrictive and fee-based licenses.

The ISSI Open Citations Letter (ISSI, 2017) calls for citation metadata to become openly available for scientometrics, both for research in the field and for its applications that support science policy and research evaluation, the latter having a large impact on the scientific community. The lack of transparency and reproducibility implied by the vendor paywalls around citation data inhibit sound practices in the field of scientometrics. Crossref, the only named candidate organization in the open letter, appears to be the best positioned for fulfilling the role of an open citation infrastructure, as it (1) is existing and operational, (2) already contains a sizeable proportion of the required metadata, and (3) makes its metadata openly available.

The proportion of open citations in Crossref is increasing. More than half of the citations in Crossref were classified as open (Shotton, 2017). Van Eck et al. (2018) show that while 77.1% of citations in the Web of Science (WoS) are present in Crossref, only 39.7% are classified as open. Efforts towards open scientometric data sources, documented by events such as the workshop reported on by Fraumann and Van Eck (2019), promise the advent of "open scientometrics" where citation data need not be sourced from commercial providers. The prerequisite for that is that Crossref covers and openly provides a sufficiently large part of citations from the WoS, today's de-facto standard citation database for most fields of science. We study whether this prerequisite is satisfied in the context of the Czech Technical University in Prague (CTU), Czech Republic i.e. we investigate the level of coverage of the WoS citation database by the openly available citation links from the COCI project (OpenCitations, 2018) on the sample where the cited publications are those we track in our institution's Current Research Information System (CRIS). We provide a breakdown to individual faculties, fields and where possible, also subfields

in two different discipline classifications: the OECD Fields of Research and Development classification and the Czech national discipline classification.

The Czech Technical University is the largest technical university in the country (and the oldest one as well, established in 1707) and is comparable to many technical universities in Central Europe. We expect our results to be relevant to other institutions of similar profiles in the region.

This article extends the work presented at the ISSI 2019 conference (Chudlarský & Dvořák, 2019).

2 Data sources and method

The Czech Technical University in Prague has a long tradition of running an in-house built institutional CRIS. The CRIS integrates our records and those harvested from the WoS web service interface, including the citations of our authors' works. This is one of the many integrations of the CRIS, for a detailed description see Dvořák, Chudlarský, and Špaček (2019).

We limit ourselves to publications from the period 2013–2017 which have both (1) a WoS accession number with a valid record in WoS, and (2) a DOI that is registered in Crossref. For checking the second condition we consult the DOIBoost dataset described in (La Bruzzo, Manghi, & Mannocci, 2019) or perform an API call to Crossref. We exclude those publications that have differing DOI values in the CRIS itself and in the WoS record. This gives the sample of 12,796 publications for which we look up the citations in both the WoS and Crossref: the citing and the cited publication are both present in both WoS and Crossref.

The November 2018 release of the Crossref Open Citations corpus (OpenCitations, 2018) was used. The "cited" side of the linking relationships is of very diverse quality. Some multiline values need to be straightened up. Some values seem to contain several DOIs concatenated, separated by spaces. To rectify these most severe errors we developed a script; its application made the data load possible and even slightly raised the number of citations to 449,843,367 (by 2,864 from the original 449,840,503). However, removing duplicate DOI pairs from the dataset leaves only 445,827,638 unique citation links (by 4,015,729 less). Some of the cited "DOIs" are still unsatisfactory: they contain internal spaces or illegal characters, end in an extra full stop, have superfluous parts in their contents or are incomplete. There clearly is room for further investigation and improvements which we are undertaking in a different thread of activity and plan to report on separately. Data quality problems on the side of Crossref citations clearly have a lowering effect on the recall of our study.

3 Findings

We found that 53.7% of WoS are present in the COCI dump of the open citation network.

This is significantly more than the approximate 40% coverage measured by Van Eck et al. (2018) for four out of five broad main fields (in the CWTS Leiden Ranking classification). Note that the remaining main field of Social Sciences and Humanities is marginal in our sample, given the research profile of a technical university.

We found important differences in the coverage among faculties (ranging from 63% down to 28%) – see Figure 1 and the supporting Table 1.

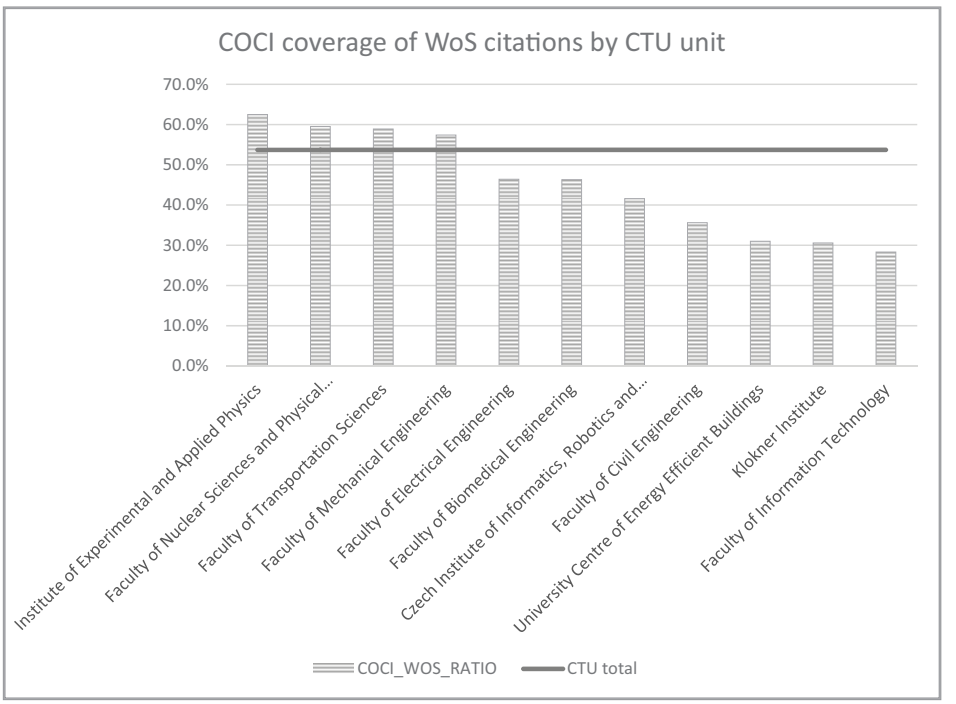

Figure 1. Coverage of WoS citations in COCI by CTU unit. COCI_WOS_RATIO denotes the proportion of Web of Science citations that are found in Crossref as open citations.

Also, the coverage significantly differs among disciplines (ranging from 78% to 25%)—see Figure 2 and the supporting Table 2. Only the disciplines with more than one hundred publications are listed. The field of Physical sciences is the most populous one and lends itself to a useful subdivision; the subfields of Astronomy (at 78% coverage) on one side and Optics (with 35%) on the other side illustrate the variance even within the single field. The second most populous field of

"Electrical engineering, Electronic engineering, Information engineering" is dominated by Electronic engineering in the context of the Czech Technical University, so no useful subdivision is possible there.

Table 1. Coverage of WoS citations in COCI by the unit of the Czech Technical University.

Faculty or University Institute	WoS publications	WoS citations	Of which in COCI	Coverage
Institute of Experimental and Applied Physics	1,122	24,348	15,225	62.5%
Faculty of Nuclear Sciences and Physical Engineering	4,225	54,470	32,398	59.5%
Faculty of Transportation Sciences	567	15,830	9,329	58.9%
Faculty of Mechanical Engineering	1,778	26,114	14,999	57.4%
Czech Technical University (whole)	**12,796**	**90,675**	**48,707**	**53.7%**
Faculty of Electrical Engineering	3,959	16,726	7,768	46.4%
Faculty of Biomedical Engineering	478	2,050	950	46.3%
Czech Institute of Informatics, Robotics and Cybernetics	219	459	191	41.6%
Faculty of Civil Engineering	1,727	7,131	2,539	35.6%
University Centre of Energy Efficient Buildings	114	232	72	31.0%
Klokner Institute	126	255	78	30.6%
Faculty of Information Technology	347	654	185	28.3%

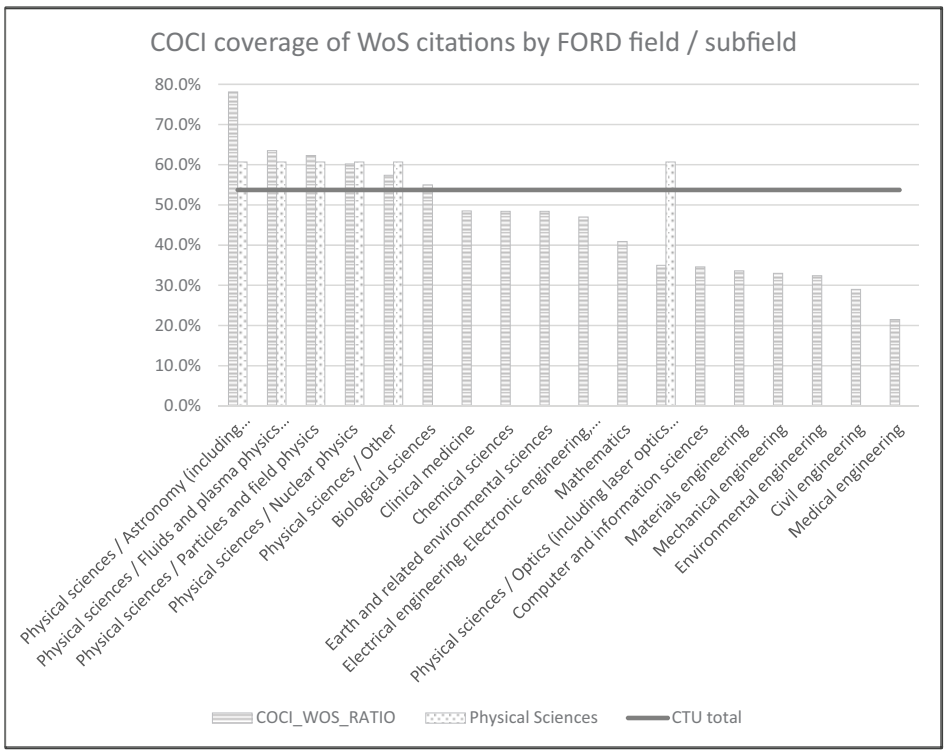

Figure 2. Coverage of WoS citations in COCI by discipline (the OECD FORD classification). COCI_WOS_RATIO denotes the proportion of Web of Science citations that are found in Crossref as open citations. The constant column Physical Sciences represents the average value for the equally named FORD field.

Table 3 lists information similar to Table 2 aggregated in the original Czech national discipline classification. Similar fields in both classifications have very similar levels of coverage, e.g. Astronomy, Particle physics, Nuclear physics, Optics, Mathematics, Electrical and electronic engineering, and Civil engineering. The discipline classification system that is used does not to affect the end result too much.

Table 2. Coverage of WoS citations in COCI by discipline (the OECD FORD classification).

Field (/ Subfield)	WoS publications	WoS citations	Of which in COCI	Coverage
- Physical sciences / Astronomy (including astrophysics, space science)	*117*	*1,028*	*803*	*78.1%*
- Physical sciences / Fluids and plasma physics (including surface physics)	*521*	*2,444*	*1,552*	*63.5%*
- Physical sciences / Particles and field physics	*1,426*	*35,838*	*22,320*	*62.3%*
Physical sciences (whole)	4,307	57,877	35,152	60.7%
- Physical sciences / Nuclear physics	*868*	*12,604*	*7,585*	*60.2%*
- Physical sciences / Other	*788*	*3,810*	*2,187*	*57.4%*
Biological sciences	114	991	545	55.0%
Czech Technical University (whole)	**12,796**	**90,675**	**48,707**	**53.7%**
Clinical medicine	131	652	316	48.5%
Chemical sciences	200	1,083	524	48.4%
Earth and related environmental sciences	252	1,468	711	48.4%
Electrical engineering, Electronic engineering, Information engineering	2,834	10,523	4,951	47.0%
Mathematics	820	2,303	942	40.9%
- Physical sciences / Optics (including laser optics and quantum optics)	590	2,253	789	35.0%
Computer and information sciences	1,000	3,097	1,071	34.6%
Materials engineering	745	4,184	1,404	33.6%
Mechanical engineering	542	1,562	516	33.0%
Environmental engineering	223	617	200	32.4%
Civil engineering	942	2,555	740	29.0%
Medical engineering	103	177	38	21.5%

4 Discussion & conclusion

The significant difference of our results from those of Van Eck et al. (2018) may be caused by the specific discipline profile of our institution and by the specific publisher choice patterns of our authors, and also the fact that the 5-year window of our sample (2013–2017) is one year later than that of the referenced work (2012–2016). These differences all deserve further research in the future.

The open citations network in Crossref is not yet ready to replace the Web of Science citations. The observed levels of coverage of citations are not yet sufficient for Crossref to be used as the source for citation analyses in research evaluation at the university and/or faculty levels. Note also that while scholarly publications without a DOI are increasingly rare, they still exist.

Table 3. Coverage of WoS citations in COCI by discipline (the original Czech national discipline classification).

Discipline	WoS publications	WoS citations	Of which in COCI	Coverage
Astronomy, Celestial Mechanics, Astrophysics	114	1,025	803	78.3%
Plasma and Gas Discharge Physics	376	1,986	1,389	69.9%
Theoretical Physics	375	1,957	1,353	69.1%
Elementary Particles and High Energy Physics	1,398	35,792	22,308	62.3%
Nuclear, Atomic and Molecular Physics, Colliders	934	12,720	7,635	60.0%
Czech Technical University (whole)	**12,796**	**90,675**	**48,707**	**53.7%**
Nuclear & Quantum Chemistry	101	463	241	52.1%
Sensors, Measurement, Regulation	377	1,139	572	50.2%
Computer Applications, Robotics	530	3,807	1,868	49.1%
Solid Matter Physics & Magnetism	294	1,466	670	45.7%
Electronics & Optoelectronics, Electrical Engineering	1,385	3,149	1,416	45.0%
Computer Hardware & Software	636	2,611	1,168	44.7%
General Mathematics	730	1,993	802	40.2%
Other Materials	152	977	355	36.3%
Fluid Dynamics	161	489	172	35.2%
Optics, Masers, Lasers	584	2,247	787	35.0%
Control Systems Theory	324	1,528	528	34.6%
Informatics, Computer Science	577	1,362	462	33.9%
Non-nuclear Energetics, Energy Consumption & Use	212	535	175	32.7%
Composite Materials	281	2,000	641	32.0%
Civil Engineering	633	1,761	521	29.6%
Building Engineering	256	682	195	28.6%
Metallurgy	166	658	188	28.6%
Nuclear Energetics	119	247	62	25.1%

References

La Bruzzo, S., Manghi, P., & Mannocci, A. (2019). OpenAIRE's DOIBoost—Boosting Crossref for Research. In: Manghi, P., Candela, L., Silvello, G. (eds) Digital Libraries: Supporting Open Science. IRCDL 2019. Communications in Computer and Information Science, vol 988. Springer, Cham, DOI 10.1007/978-3-030-11226-4_11

Chudlarský, T., & Dvořák, J. (2019). Can Crossref Citations Replace Web of Science for Research Evaluation? The Share of Open Citations. [poster abstract] In Proceedings of the 17th Conference of the International Society for Scientometrics and Informetrics. (pp. 2551–2552). Rome, Italy: Edizioni Efesto. ISBN 978-88-3381-118-5.

Crossref. Retrieved from https://www.crossref.org

Dvořák, J., Chudlarský, T., & Špaček, J. (2019). Practical CRIS Interoperability. In 14th International Conference on Current Research Information Systems: Practical CRIS Interoperability. Amsterdam: Elsevier B.V., pp. 256–264. Procedia Computer Science. vol. 146. ISSN 1877-0509. DOI 10.1016/j.procs.2019.01.077

van Eck, N.J., Waltman, L., Larivière, V., & Sugimoto, C. (2018). Crossref as a new source of citation data: A comparison with Web of Science and Scopus. Retrieved from https://www.cwts.nl/blog?article=n-r2s234

Fraumann, G., & Waltman, L. (2019). The 2019 Workshop on Open Scientometric Data Infrastructures at Leiden University. Retrieved from https://www.cwts.nl/blog?article=n-r2x274&title=the-2019-workshop-on-open-scientometric-data-infrastructures-at-leiden-university

ISSI (2017). Open citations: A letter from the scientometric community to scholarly publishers. Retrieved from http://issi-society.org/open-citations-letter/

OpenCitations. (2018). Open Citation Indexes: COCI, the OpenCitations Index of Crossref open DOI-to-DOI references. Dump, the "Citation data (CSV)" file. DOI 10.6084/m9.figshare.6741422.v3

Shotton, D. (2017). Milestone for I4OC—open references at Crossref exceed 50%. Retrieved from https://opencitations.wordpress.com/2017/11/24/milestone-for-i4oc-open-references-at-crossref-exceed-50/

The DOI Handbook. (2012). The DOI Foundation, 2012. DOI 10.1000/182