Data description for cpr_example_data.csv

Janne-Tuomas Seppänen*, Hanna Värri, Irene Ylönen

Open Science Centre, University of Jyväskylä, Finland

* Corresponding author contact information:

email: janne.t.seppanen@jyu.fi

Purpose

These data have been assembled and calculated to enable analyses presented in the article "*Co-Citation Percentile Rank and JYUcite: a new network-standardized output-level citation influence metric and its implementation using Dimensions API*" (Seppänen et al 2020).

Metadata for total of 41 713 outputs from University of Jyväskylä (JYU) current research system published between 2007-2019 (all kinds, including non-peer-reviewed outputs) were assembled and evaluated, out of which 13 337 had i) at least one citer, ii) were discoverable in Dimensions by either DOI or title.

Metadata and derived metrics for those 13 337 outputs is published here.

Description of variables

name	*description*
id	row number
doi	Digital Object Identifier of the scholarly output
dimensions_id	unique identifier to find output's entry in Dimensions database
title	title of the output
type	classification code defined by Ministry of Education and Culture in Finland (2019), and assigned by information specialists at research organizations, see References
department	Discipline affiliation at JYU
date	publication date or date inserted into Dimensions, depending on which is earlier
available	number of days the output had been available for citation at the time of analysis
times_cited	number of times the output had been cited according to Dimensions metadata at time of analysis
citer_count	number of citers found when searching by reference_ids[1]
cit_rate	(times_cited / available) x 365
co_cit_size	size of the co-citation set
co_cit_counted	number of co-citations for which *times_cited* -parameter existed in Dimensions metadata
co_cit_average	average citation rate in the counted co-citation set
co_cit_median	median citation rate in the counted co-citation set

[1] Presumably due to ansynchronies in Dimensions metadata updates, occasionally the search by reference_ids returns different number of citers than what is given in output's times_cited metadata, usually 1 more. To retain comparability to co-citation set, we used the times_cited -value to calculate citation rate, except replacing it with citer_count where it was zero, to avoid zero-induced problems in statistics and graphing.

co_cit_quart_1	25th quartile of citation rates in the counted co-citation set
co_cit_quart_3	75th quartile of citation rates in the counted co-citation set
cpr	Co-citation Percentile Rank of the output's citation rate in its counted co-citation set
solve_time	number of seconds it took to retrieve the co-citation set metadata and calculate cpr
when_calculated	UNIX timestamp of the moment when cpr was calculated.

Methods

To retrieve the co-citation set and bibliometric metadata for each scholarly output, we used Dimensions Search Language (DSL) on their API:

1. get Dimensions **ID** and other data about the target, by **DOI** or title (or multiple targets: up to 400 can be queried in a single call, then looped through the queries below): `search publications where doi in [DOI] return publications [basics + extras + date_inserted]`
2. get data about outputs where the target appears in their list of references, i.e. the citing outputs, and in particular the reference_ids of outputs they cite, i.e. the co-citation set of the target : `search publications where reference_ids="ID" return publications [title+id+reference_ids]`
3. Collect the reference_ids from the result array, remove duplicates, concatenate the unique ids into comma-separated strings chunked to max 400 ids per string.
4. get metadata about the co-citation set, by **reference_ids** string (looping through if multiple chunks) : `search publications where id in ["reference_ids "] return publications [title + doi + times_cited + date + date_inserted + id]`

Once the citation counts and days since the output became citable have been assembled, these are converted to citation rate = $\frac{citation\ count}{days\ since\ first\ citable} \times 365$. We do the scaling to 365 days to aid interpretation and comparisons, but we emphasize that 365 days is not a calendar year, and that the citation rate is defined also for outputs that are less than one year old.

The co-citation set is then ordered by citation rate, and then

cpr = percentage of citation rates in the co-citation set that are less than or equal to the citation rate of the target output.

cpr thus ranges from 0 (= all outputs in the co-citation set are cited more frequently than the target) to 100 (= none of the outputs in the co-citation set are cited more frequently than the target). **cpr** is undefined for targets that have not been cited, or for which the algorithm is unable to find co-citation set metadata. The quartiles and average of the citation rates, and the size of the co-citation set are also calculated and saved.

Note on the day count used in the algorithm: when an output first becomes citable, typically as an “early online” article, Dimensions lists that date in the metadata. However, for platforms that later assign outputs to issues, that date eventually gets replaced in Dimensions metadata by a later so-called "publication date". The original, true date when the work first became citable, is not currently retained in Dimensions in these cases. For some platforms the resulting error may be just few tens of days, but in others the error may be many hundreds of days. In most cases the true date would be recoverable from metadata in CrossRef or the landing page of the output DOI, but this would impose prohibitive time performance cost on the algorithm. We hope this unfortunate and unnecessary source of error gets remedied in the future. For now, we resort to retrieving also the **date_inserted** value alongside the **date** value

from Dimensions API, and choose the earlier of these two dates, acknowledging some outputs may still enter the algorithm with unknown error in their citation rate, and that this error is more common and likely larger for some disciplines, such as Economics or Humanities.

Each output in the source data was affiliated with one or more academic departments at JYU. For the dataset here, we simplify that data by merging data from units falling under same discipline (e.g. Institute for Education Research is merged with Department of Education).

References

Ministry of Education and Culture (Finland). 2019. Publication data collection instructions for researchers 2019. https://wiki.eduuni.fi/download/attachments/39984924/Publication%20data%20collection%20instructions%20for%20researchers%202019.pdf

The source code of the algorithm is publicly available at [JYX, with its own DOI, linked from there to https://gitlab.jyu.fi/oscsolutions/jyucite/ something]