Discovering and charting the black matter of citations. Matilda as a new source for Open Citations

Didier Torny[1]

[1] *Centre de sociologie de l'innovation, (CNRS, I3 UMR9217), MinesParis, PSL University*

Purpose

The open citations initiative originally focused on freeing existing citations, and on publishers using Crossref DOIs. That strategy was successful in building the COCI index [1] and then extended to other sources (Datacite, OpenAire, …) by OpenCitations. However, there is a limit to that strategy: it presumes that the information linking the cited text to the citing text already exists in digital form and simply needs to be circulated under a CCO license. In many cases, particularly for older documents or for publishers with limited resources, this information does not currently exist in a manipulable format. The risk is then to limit ourselves to citations as open as available and as closed as publishers wish or can. Therefore, we will present the operationalisation of a complementary strategy to reconstruct these missing links and share them via the Open Citations infrastructure.

Extraction and alignment methods in Matilda

The Matilda platform aims to provide bibliographic search services for the entire literature on an open data basis in its sources and results, thus enabling the widest possible sharing and re-use [2]. Matilda was publicly released in September 2023; it supports the approach designed by the Initiative for Open Citations (I4OC) and is now pushed by the Barcelona Declaration on Open Research Information. It is continuously available to all at the following url: https://matilda.science/?l=en . Citations are an important part of the infrastructure, as citation tracking (from authors or works) is available both on the Web interface and through RSS streams.

In May 2025, Matilda displayed 148 million works based on the exploitation of ArXiV, Crossref, Pubmed and HAL, the French National archive. Daily, a mean count of 35K documents is being added and enriched through ORCID and Unpaywall. In addition, Matilda indexes the XML full text of legally accessible PDFs in the sources, around 100K daily for a total of 22M PDFs by the end of May 2025. This operation is carried out using Grobid [3] and enables specific information to be extracted on authors, abstract, keywords and, for the central object of this presentation, referential links.

The goal is then to process these references extracted by Grobid to index them as links in the Matilda graph. As we wish to limit errors and are still in an experimental phase, Matilda only uses DOI, ArXiv ID and Pubmed ID as a pivot to operate. In other words, all references found without these PIDs or with PIDs that do not match in Matilda are not included in the graph (but are kept for later runs). Matilda's ontology is based on the reconciliation and deduplication of identical texts, eg. a preprint present in ArXiv, a version of record in Crossref and an archived

version in HAL. The citation/reference from a given version to another version is traced, but also attributed to all the versions, again with links deduplicated. An order of priority of the versions makes it possible to treat and preserve only one indexed full text. The current process leads to the creation of around 950K reference links from sources and 830K from PDF extraction every day.

OpenCitations export

The OpenCitations ontology is slightly different from that of Matilda, which implied that a specific procedure for preparing data for export had to be defined and coded. Notably, versions of different texts are taken into account. Everything is available for OC in the form of a batch file, which is also daily fed and contains both the reference links found and the series of metadata required by the Open Citations model for texts in Matilda. Moreover, we are also making available all the referential links (PID to PID) not found in Matilda, in particular because the systematic collection of sources has only been carried out for texts dating from after 31 December 2018. The whole process was made easier by the existence of a validator on the Open Citations side. This import will enable Matilda to become a new source of referential links available and maintained in Open Citations, and to measure the additional contribution it makes to all existing sources.

Results & perspectives

The extraction of referential links and their processing was put in place in the spring of 2024, and we are now operating at scale, making it possible to perpetuate this alternative to citation data from open sources. The introduction of such procedures is all the more crucial for the future of open citations because openness is not a permanent property: large commercial publishers have stopped making available a large proportion of the abstracts of their publications because they have become a source of profit in themselves [4], and there is nothing to prevent this happening in the future for citation links in their new publications. However, this procedure is biased in favor of publications whose version is itself open. As part of the development of Matilda, the aim is to use the legal and contractual provisions of the archive funds purchased from the major publishers to extract the reference links from paywalled publications as well. Finally, the increase in the number of sources and the use of historical data, particularly from Crossref, should make it possible to extend this process of discovering currently unknown referential links as far as possible. Not only will Matilda be strengthened in terms of bibliographic research, but it will also provide various communities with a comprehensive working basis for the collection and analysis of referential links.

References

[1] Heibi, I., Peroni, S., & Shotton, D. (2019). Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. Scientometrics, 121(2), 1213-1228.

[2] Torny, D., Capelli, L., Danjean, L., & Pouyllau, S. (2019, June). Matilda: Building a bibliographic/metric tool for open citations and open science. ELPUB 2019 23rd edition of the International Conference on Electronic Publishing, Jun 2019, Marseille, France. 10.4000/proceedings.elpub.2019.22ff. hal-02141839f

[3] Lopez, P. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13 (pp. 473-474). Springer Berlin Heidelberg.

[4] Kramer, P. (2024). More open abstracts? Comparing abstract coverage in Crossref and OpenAlex. https://bmkramer.github.io/SesameOpenScience_site/thought/202411_open_abstracts/