# Matilda: Building a bibliographic/metric tool for open citations and open science

Didier Torny, Laurent Capelli, Lydie Danjean, Stéphane Pouyllau

# Matilda

## Building a bibliographic/metric tool for open citations & open science

Didier Torny, Laurent Capelli, Lydie Danjean and Stéphane Pouyllau

1 The Open Access movement has long insisted on the availability and reusability of academic texts as a goal to achieve knowledge dissemination, without putting specific attention to the question of metadata. Indeed, OAI-MPH has been the norm made for them, and as subparts of the texts, they are as much concerned by the efforts against privatization of knowledge. Yet, the fact that no reference was made to metadata in the main OA declarations (Budapest, Berlin, Bethesda) has led to a paradoxical situation. The more publication as a process became accessible and reusable, the more its content was searched and found through privately-owned and often costly bibliographic/metric tools by research communities, if they could afford it. And it was through the use of the Web of Science that OA advocates were eager to show how much accessibility led to a citation advantage compared to paywalled articles (Eyenbach 2006; Norris, Oppenheim and Rowland 2008).

2 Our ongoing project to build a bibliographic/metric tool for open science, as other recent initiatives, aim at changing the situation of reference information being the forgotten child of open science, with two main objectives. First, by doing so, we do not want to just "open" the existing closed information, but wish to give back a fair place to the whole academic content that has been excluded from such tools, in a "all texts are born equal" fashion. Second, we want to give academic communities and researchers the most possible control on the way they search for text and metadata information, as we consider the closed design of current tools often encapsulates an objectifying view of search processes rather than relying on user-based technologies.

## Bibliographic/metric tools and their users

3 Contemporary bibliometric tools are characterized by a double division, in terms of their design and use. These divisions reflect the ambiguity of bibliometric proposals from the outset, between the desire to predict the directions that science has taken and would

take, tracing what would be "important" in different disciplines and strong practical limits on the information actually available (Wouters 1999). On the database design side, there is a strong opposition between two models: on the one hand, the Web of Science and Scopus offer lists of specific outlets, for the vast majority journals, supposedly prestigious or important, but voluntarily limited; on the other hand, Google Scholar, which operates in an open electronic world containing books, reports, preprints, articles, presentations, but totally opaque about its limits even if its technical inclusion rules[1] are public. In the current situation, users therefore seem to be able to choose only between a sample of the whole scientific production, which is more or less controlled but very strongly biased, and a delegation to a generalist search engine, without the possibility of opening it or setting it up.

4 On the usage side, the situation seems to be even more diverse: we shall first state that, to our knowledge, there is almost no literature on the actual uses of bibliometrics tools. Beyond very limited quantitative surveys on use/non-use (e.g. Bar-Ilan, Peritz and Wolman 2001; Dukic 2013), detailed user surveys, ethnographic studies or other form of academic studies on what colleagues do with existing tools seem to be non-existant. So, the following is more a set of hypotheses than certified knowledge about usage, based both on actual features of these tools and their marketing on one side, and one limited paper (David, Minel and Pouyllau 2011), non-disclosable information and (critical) narratives of their growing popularity within some academic communities and institutions.

5 We may analytically distinguish four uses of current bibliographic/metric tools. First, the most trivial one, is the one of a *browser*: to look for specific texts, already known by their title/DOI/authors. As for other browsers, this use is purely instrumental in order to find and read the abstract and/or full text. As these tools in their web form have included links to full documents - contrary to their pre-web predecessors mostly limited to metadata - they can be used as a library catalog. Second, there is a massive but limited use of functionalities by a large number of researchers as a *search engine*: a bibliographic search and citation tracking in order to find relevant texts for their subject matter/ research/state of literature, to scan their abstract and, potentially, read them through a link to their full text, whether it is to publish an article, designing research programs or teaching (Okiki 2012). Dating at least back to Current Contents and other abstract services, it is then used an information retrieval service, intermediary between users and the existing literature.

6 Third, they can be used as a *mirror*, in order to perform some uses that Wouters and Costas (2012) named "narcissistic technologies", through which authors can assess their own production and impact, typically Google Scholar profiles created in 2012. Bibliographic/metric tools are then processed to present a public or private view of oneself or one's group in a favorable way, peek at competitors' positions, been ranked among peers or promoting one's merits in a research assessment. As such, it takes part in a long history of scholarly valuation technologies, which probably dated back to the invention of the "list of publications" as a CV at the end of the 19th century (Csiszar 2017).

7 Fourth, there are used by limited users, but extensively, as a (sophisticated) *counting device*, by some bibliometric labs, which results can be read in the large scientometrics litterature, but also ranking organizations (Times Higher Education, Shanghai AWRU...) and finally heads of research and university organizations. The latter one focus on

tracing the production of a scientific field, members of an institution, a network of laboratories, while ordinary researchers are not interested in it. Though these uses could be subsumed under the term Evaluative Bibliometrics (Narin 1976), there are profound divisions not only between academic-oriented and management-oriented uses, but also within the academia as competing definition of bibliometrics coexist, not only from an epistemological point of view, but also an economic one (Pontille and Torny 2013).

8 Of course, these have to be seen as prototypes of "pure" uses, while they are often mixed in actual uses. For example, people would use their Google Scholar author profile as an academic mirror, but the citing articles pointing at their own texts would be browsed, as they would have interest in looking at their results' use (to look at the competition, to build new collaborations...)

## Towards which open citations?

9 This already complex landscape is under transformation with the cumulating effect of open access and open science movements. In fact, massive bibliographic and bibliometric data sources may now be considered as open: this is the case of all texts under certain creative commons licenses, which are blooming both on repositories and preprint platforms. There also specific disciplinary diapositives, one canonic example being Citec[2], based on the economics distributed preprint service Repec, with currently more than 1.3 million working papers and 38 million references.

10 On a multidisciplinary basis, it is the outcome of the "open citations" movement: originally, in 2010, it was a reference data corpus, the Open Citation Corpus (Pironi et al. 2015), before these remarkable precursors were joined by various organizations demanding the release of Crossref citation data to publishers. The I4OC[3] collective has consequently obtained the availability, under a CC0 license, of a large part of the CrossRef database, which has led to the use of this data in a number of tools, including the VOS Viewer[4] developed by Leiden University. Beyond their common interest in "freeing" metadata and citation data, there is no shared agenda: they hope that other actors would take them and build services on this new shared resource. Like the OCC database, they often presuppose professional users, either experts in API manipulation or interested in very advanced bibliometric developments. For example, the very intriguing proposition of crowdsourced open citations (Heidi, Peroni and Shutton 2019a) designs an incredibly expert user, able and willing to code DOIs and lines into a specific format in order to fill a shared database up. Though an elegant proposal, it would need the very participation of a large part of the global academic community in order to "free" their authored data, and almost all of them won't neither have the skills nor the time to do it.

11 On a separate front, facing the resistance of some big publishers (IEEE, ACS...) to "free" their metadata in Crossref, some scientometricians have used one of the OA movement most common tactics. First, they have made a public letter and petition in December 2017, while they negotiated with Elsevier, the publisher and title of *Journal of Informetrics*, in order to obtain new contractual conditions between the editorial board and the publisher, including a CC0 license for the journal metadata. As it was refused, the board resigned[5] at the beginning of 2019 and founded a new OA journal, *Quantitative Science Studies*, published by MIT Press. As Vincent Lariviere, member of the board, stated: "Journals should serve the research community—not the other way around." Nevertheless, this spectacular move to "declare the independence" of their journal, as

SPARC coined the term in 2001 and a long list of journals[6] have done since, remains limited in its global effects for open citations.

12 It is on the legal side, with the intense lobbying of some scientific organizations like Science Europe,[7] that things could change on a massive scale. In fact, the very recent vote of the European copyright directive by the European Parliament would lead to a wider and easier road for open citations. A mandatory exception to copyright for TDM is included in the current text, meaning that bibliographic data could soon be scrapped, stored and manipulated independently of its intellectual property protection, especially database protection. It order to grasp the potential effect of such a law, we can take the example of the very recent paper of the OCC team (Heidi, Peroni and Shutton 2019b), which has created the COCI RDF data with 445 million DOI-to-DOI citation links, based on 43 million Crossref documents. Based on a linear distribution of references, that means around 500 million links are still hidden from the public because of proprietary policies.

## Strategies to make an open bibliographic/metric tool

13 In the short term future, bibliographic data will thus become a public good or a common, at least from within the borders of the European Union. But what should we do with Open Citations? Until now, and especially within OCI and OCC frameworks, the answer seems to be: to build neat and clean databases that are available for API developers and other expert users. In this paper and in our current development, we give a different and complementary one, that is building a tool precisely not designed for scientometricians and experts in semantic web or scholarly communication, but for ordinary researchers. As we have seen, they do perform at least the three first kind of tasks (browser, search engine, mirror) with their current tools, so we have begun the construction of a bibliographic/metric tool, intended for lay users looking for documents and using basic citation-tracking, following a dual development strategy.

14 On the one hand, we have started by gathering theoretically available corpus, would it be in dumps or through dedicated APIs. That includes REPEC, PubMed, ArXiv and Crossref, all for metadata and some for full texts when available. For PubMed and Crossref, we have focused on its more recent part, in order to maximise the quality of metadata and the presence of references in the dedicated field at this stage of development. In fact, as it has been noted before, the Crossref data is highly variable in quality (Van Eyck et al. 2018), as it does not depend on a centralized production and checking process, but on the good will and expertise of each publisher member of Crossref. Data curation is an ongoing long process, especially on the references field, as a lot of them are unstructured, don't include identifiers (DOIs or other ones) or contain incorrect ones. Compared to the already mentioned COCI corpus, we intend to systematically reallocate DOIs to references missing them, through the use of GROBID developed at INRIA (Riondet and Foppiano 2017) and BILBO (developed by OpenEdition[8] and adapted to variable forms of SEO). When this process succeeds, we harvest the metadata of these documents and if an open version of the document exists—directly available or found through Unpaywall[9] it is harvested and indexed. All the results produced are then enriched by a similar APIs to the one already developed in ISIDORE (for example, the language of the document, the disciplinary field, etc.)

15 On the other hand, we create a user interface that can be configured by users based on the APIs and data already produced via the ISIDORE search engine (Dumouchel 2018). It

will include full-text search, contrary to legacy databases (Dimensions recently introduced that for commercial tools), and author pages. Its entire design aims at interoperability with existing tools—for example Unpaywall—or tools under development, which share the same open science vision, in order to place Matilda in the existing ecology. Above and beyond technical developments, which are essential for a tool, we would like in the final part of this communication to insist on two principles, which are for us quintessential to an open science tool.

## All texts should be born equal

16 The first principle of an open bibliometrics tool should be inclusivity, at the exact opposite of the foundation of previous generations bibliometrics tools. Since the Journal Impact Factor has become successful in some disciplines, the database on which it is calculated, owned by the Institute for Scientific Information (ISI), then Thomson Reuters, have acted as a triage for scientific quality of outputs in some communities, and by a rippling effect, on evaluation and careers (McKiernan e al. 2019). Conversely, journals and articles that were excluded have become invisible in some parts of academia. As an example, in the 1980s, an international movement tried to have a better representation of journals published in the Global South in ISI indexes, with no success (Moravscik 1985), as scarcity was at the heart of its constitution. Thirty years later, even if the marketisation of Scopus has led to a less selective regime, the biases towards certain disciplines, english-language sources and US-produced science are still powerful (Mongeon and Paul-Hus 2016), and the divide between STS on one side, Humanities and Social Sciences on the other remains abyssal (Archambault et al. 2006), especially if non-journals outputs are considered.

17 In 1993, Margaret Rossiter published an article entitled *The ~~Matthew~~ Matilda Effect in Science*, in which she discusses the systematic reduction of women's contributions to science. This reduction is the symmetrical and corollary of the Matthieu effect described by Merton (1968), which leads to the accumulation of scientific credit by a small number of researchers, usually men. She concluded by naming this effect by the first name of a 19th century American feminist figure, suffragette but also the first one to have identified this phenomenon of invisibilisation of women in science. We have chosen this same name for our project, observing a similar invisibilisation process of a large number of scientific productions, because they are written in a language other than English, because they come from publishers with limited economic weight, because they are supposedly non-scientific or, at least, uncertified types of production. Matilda's promise is to remedy this situation by treating all available or identifiable texts and their metadata with equal dignity.

18 The inclusive movement has already started in the tools made by newcomers: for example, Dimensions has included some pre-prints servers in its sources (BiorXiv, SSRN... ), as they attribute a DOI and thus are available through Crossref in its data. Some scientometricians[10] have started to consider Crossref data a basic source for their work, rather than WoS or Scopus. But it could only be completed if it is made as a political statement embodied within a tool. Matilda, as an open tool, should always aim at including more sources, in order to give the largest possible choices to its users, by not stigmatizing or, even more, excluding working papers, preprints or "south" journals. As the theme of "predatory journals" grew in the last decade, some critics underlined the

messiness of Google Scholar sources, the ability to game its results through fake papers (López-Cózar, Robinson-Garcia, and Torres-Salinas 2014) and its absence of selection[11] towards these "journals". But in a time where the most "prestigious" journals—i.e. high impact factor ones—are also the ones at the top of the "retraction index" (Fang & Casadevall 2011), on which ground should we build a black list of outputs?

19 It is much preferable to consider all sources, harvest them and then give users the complete control to select/exclude with any available metadata on "quality" (DOAJ, DOAB...), non peer-reviewed/peer reviewed text, etc. Inclusion is the first condition of openness, but it should not only concern texts, but also users as true part of the tool.

## The beauty should be in the eyes of the users

20 Bibliometric databases have no theory, they are purely empirical data, cleaned and rawified (Denis and Goëta 2017). This view is commonly shared and certainly amongst most designers and users of evaluative bibliometrics. Though they have often searched for a theory of citation, scientometricians have failed in according on one (see for example Cronin 1981; Leydesdorff 1998). The diversity of citation usage, though a very challenging research and practical question, only becomes a problem when adding citations to count them, compare texts or journals on this basis, would give an objectified view while meshing in fact very different communication and social processes (Erikson and Erlandson 2014). Matilda is built, in a sense, for non-counting but reading users, which would actually explore the meaning of these links between texts, if they are interested.

21 More generally, while the database of source has to be as inclusive as possible, it is by no way objective for our users, because we wish to offer them the most possibilities to personalize their usage. More precisely, as Daston and Galison would have said it (2007), they don't search for mechanical objectivity, but a tool to help them to perform their "trained judgment" on what matters to them. It is remarkable that the literature on recommendation algorithms has almost been absent of the bibliographic/metric topic, while it has been blooming on cultural products (Shay & Pinch 2006), coming mainly from computer science, management science and sociology. Rather than seeing the tool as an objectifying object, we aim at subjectification for Matilda: for us, academic texts are much more like one of their kind cultural products than standard commodities for their users. Just as journals are often considered as unique commodities for scientific commodities, hence the strong market power of publishers to determine their prices, texts should rather be considered as only valuable for specific audiences.

22 Yet, it is true that some coded information, including the name of the journal, keywords attached by authors and, in some disciplines, object taxonomies, are currently the main semantic landmarks in information retrieval. And, of course, citation tracking has become a standard as, in a way or another, the process of referencing is, in a way or another, a trace of influence (Zuckerman 1987). Hence, like others, Matilda will include faceted engines, sort by citation or date, but also author pages and personalized alerts. But more importantly, we think it should go on further away by proposing recommendations. Bibliometrics tools have been reluctant at using recommendation, leaving the selection process to users after an « objective view » of their search was given to them. Google Scholar only introduced recommended articles in 2012, linked to author profiles, and Microsoft Academic did the same on its reboot in 2016. Among other

personalization systems, Matilda will build on the suggestions engine already built for Isidore, and import methods from the recommendation algorithms community and give users the maximum control on them. On top of that, when our proof of concept is ready, we will open accounts to invite users in order to conceive the interface and services that academic users need.

## Redifining bibliometrics?

23 Until now, the Open Citations movement has been dominated by visions of building information systems for users making APIs or advanced code manipulators, with a strong connection to the tradition of bibliometrics. Matilda, though resting on the same kind of information, will enable search, recommendation and as often as possible reading, directly available instead of a giving access to a database. Matilda is therefore part of a twofold heritage of conception and usage with, on the one hand, Google Scholar, which has popularized free access to citation tracking, outside the circuit of libraries and publisher sites, to metadata of scientific content, intervention of users as enrichers and curators. On the other hand, it inherits from ISIDORE, Base and other search engines based on OAI-MPH harvesting, with full-text search, blindness to prestige or selection process. The objective is therefore to complete/replace these various uses, as well as those of commercial databases (WoS/Scopus), which are widely used in certain disciplines.

24 By doing so, we wish to redefine bibliometrics tools as a technology. From the start, it has been ambiguous between description and evaluation, showing the past and predicting the future, aimed at funders, ordinary researchers or those studying the science of science, mapping its evolutions. It has become popularized in the 21st century through the web and its hyperlink core, the circulation of some in-fashion indexes well outside scientometrics circles (from Journal Impact Factor to h-index) and the large use of tools as narcissistic machines or search engines. It is time that these last uses, which will deeply benefit from inclusivity, become the basis for a bibliographic/metric tool, rather than diverting ones made for other purposes.

---

BIBLIOGRAPHY

## References

Archambault, Éric, Étienne Vignola-Gagné, Grégoire Côté, Vincent Larivière, and Yves Gingras. 2006. “Benchmarking scientific output in the social sciences and humanities: The limits of existing databases.” *Scientometrics* 68 (3): 329–342.

Bar-Ilan, Judit, Bluma C. Peritz, and Yecheskel Wolman. 2003. “A survey on the use of electronic databases and electronic journals accessed through the web by the academic staff of Israeli

universities." *The Journal of Academic Librarianship* 29 (6): 346–361. https://doi.org/10.1016/j.jal.2003.08.002

Cronin, Blaise. 1981. "The need for a theory of citing." *Journal of documentation* 37(1): 16–24.

Csiszar, Alex. 2017. "How lives became lists and scientific papers became data: cataloguing authorship during the nineteenth century." *The British Journal for the History of Science* 50, (1): 23–60. https://search.proquest.com/docview/1882435264

Daston, Lorraine, and Peter Galison. 2007. *Objectivity*. New York: Zone.

David, Sophie, Jean-Luc Minel, and Stéphane Pouyllau. 2011. "Documenting some Uses of the Isidore Platform." https://isidore.science/document/10670/1.lbc7dv

Denis, Jérôme, and Samuel Goëta. 2017. "Rawification and the careful generation of open government data." *Social studies of science* 47 (5): 604–629.

Dukić, Darko. 2013. "Online databases as research support and the role of librarians in their promotion: The case of Croatia." *Library Collections, Acquisitions, and Technical Services* 37 (1–2): 56–65. https://doi.org/10.1016/j.lcats.2013.09.005

Dumouchel, Suzanne. 2018. "The EOSC as a knowledge marketplace: the example of ISIDORE." Poster presented at the EUDAT conference: Putting the EOSC vision into practice, Porto, Portugal, January. https://isidore.science/document/10670/1.a2kkhg

Erikson, Martin G., and Peter Erlandson. 2014. "A taxonomy of motives to cite." *Social Studies of Science* 44 (4): 625-637.

Eysenbach, Gunther. 2006. "Citation advantage of open access articles." *PLoS biology* 4 (5): e157.

Fang, Ferric C., and Arturo Casadevall. 2011. "Retracted Science and the Retraction Index." *Infection and Immunity* 79 (10): 3855. https://doi.org/10.1128/IAI.05661-11

Heibi, Ivan, Silvio Peroni, and David Shotton. 2019. "Crowdsourcing open citations with CROCI-An analysis of the current status of open citations, and a proposal." *arXiv preprint* arXiv:1902.02534. https://arxiv.org/abs/1902.02534

Heibi, Ivan, Silvio Peroni, and David Shotton. 2019. "COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations" *arXiv preprint* arXiv: 1904.06052. https://arxiv.org/abs/1904.06052

Leydesdorff, Loet. 1998."Theories of citation?" *Scientometrics* 43 (1): 5–25.

López-Cózar, Emilio Delgado, Nicolas Robinson-Garcia, and Daniel Torres-Salinas. 2012. "Manipulating Google Scholar citations and Google Scholar metrics: Simple, easy and tempting." *arXiv preprint* arXiv:1212.0638.

McKiernan, Erin C., Lesley A. Schimanski, Carol Muñoz Nieves, Lisa Matthias, Meredith T. Niles, et Juan Pablo Alperin. 2019. "Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations". *PeerJ Preprints*, 7:e27638v2 (April). https://doi.org/10.7287/peerj.preprints.27638v2.

Merton, Robert K. 1968. "The Matthew effect in science: The reward and communication systems of science are considered." *Science* 159 (3810): 56-63.

Mongeon, Philippe, and Adèle Paul-Hus. 2016. "The journal coverage of Web of Science and Scopus: a comparative analysis." *Scientometrics* 106 (1): 213-228.

Moravcsik, M. J. 1985. *Strengthening the coverage of third world science*. Eugene, OR: The University of Oregon.

Narin, Francis. 1976. *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill, NJ: Computer Horizons.

Norris, Michael, Charles Oppenheim, and Fytton Rowland. 2008. "The citation advantage of open-access articles." *Journal of the American Society for Information Science and Technology* 59 (12): 1963–1972. https://doi.org/10.1002/asi.20898

Okiki, Olatokunbo Christopher. 2012. "Electronic information resources awareness, attitude and use by academic staff members of University of Lagos, Nigeria." http://hdl.handle.net/123456789/507

Peroni, Silvio, Alexander Dutton, Tanya Gray, and David Shotton. 2015. "Setting our bibliographic references free: towards open citation data." *Journal of Documentation* 71 (2): 253–277. http://dx.doi.org/10.1108/JD-12-2013-0166

Pontille, David, and Didier Torny. 2013. "La manufacture de l'évaluation scientifique." *Réseaux* 1:23–61. https://hal-mines-paristech.archives-ouvertes.fr/hal-00821956

Riondet, Charles, and Luca Foppiano. 2017. "GROBID for Humanities When engineering meets History." In *Text as a Resource. Text Mining in Historical Science*. Paris: Institut Historique Allemand. http://hdl.handle.net/10670/1.7dlpbi

Rossiter, Margaret W. 1993. "The Matthew Matilda effect in science." *Social studies of science* 23 (2): 325–341.

Shay, David, and Trevor Pinch. 2006. "Six degrees of reputation: The use and abuse of online review and recommendation systems." *First Monday* 11 (3).

Van Eck, Nees Jan, Ludo Waltman, Vincent Larivière, and Cassidy Sugimoto. 2018. "Crossref as a new source of citation data: a comparison with Web of Science and Scopus." *CWTS* (blog), January 17. https://www.cwts.nl/blog?article=n-r2s234

Wouters, Paul. 1999. "The citation culture." Ph.D. dissertation, Universiteit van Amsterdam. garfield.library.upenn.edu/wouters/wouters.pdf

Wouters, Paul, and Rodrigo Costas. 2012. *Users, narcissism and control: tracking the impact of scholarly publications in the 21st century*. Utrecht: SURFfoundation.

Zuckerman, Harriet. 1987. "Citation analysis and the complex problem of intellectual influence." *Scientometrics* 12 (5–6): 329–338.

## NOTES

**1.** https://scholar.google.com/intl/fr/scholar/inclusion.html%23overview
**2.** http://citec.repec.org/
**3.** https://i4oc.org/
**4.** www.vosviewer.com
**5.** http://issi-society.org/media/1380/resignation_final.pdf
**6.** http://oad.simmons.edu/oadwiki/Journal_declarations_of_independence
**7.** https://www.scienceeurope.org/legislation/activities/directive-on-copyright/
**8.** https://lab.hypotheses.org/1437
**9.** https://unpaywall.org/
**10.** https://fr.slideshare.net/LudoWaltman/comparing-bibliographic-data-sources
**11.** http://blogs.cc.umanitoba.ca/mhiknet/2014/11/17/predatory-journals-in-google-scholar/

## ABSTRACT

Although bibliometrics and library science are older, bibliometric tools were really born about 50 years ago and were only made available to a large audience with the widespread use of the Internet. Although their concrete forms have been largely modified, they are still based today on epistemic and computer foundations decided at the time. Three important characteristics of these tools can be identified: first, they are proprietary, i.e. users not only have to pay for access to the data but it is also difficult to manipulate and verify; second, in the name of a principle of scarcity or quality, tool creators assume to rely only on a selection of accessible scientific documents; thirdly, this choice of a small sample is, moreover, very marked by a historical irreversibility that makes invisible in particular some types of documents (books, conferences, preprints) and written documents in the vast majority of languages other than English. However, over the last twenty years, there has been a progressive liberation of scientific texts through the existence of different disciplinary (ArXiv, PubMedCentral, REPEC) and institutional (HAL, universities archive...) open archival systems, and publication models allowing the harvesting of texts and/or metadata - including the references cited. It is in the continuation of this movement that the construction of a real tool, Matilda, is taking into account all available sources and user personalization, in order to serve as an elementary brick for bibliographic and bibliometric research in the age of open science.



## INDEX

**Keywords:** bibliographic tool, citation tracking, open citations, open science, search engine


## AUTHORS


### DIDIER TORNY

CSI-I3 (UMR 9217, CNRS)
didier.torny@mines-paristech.fr

### LAURENT CAPELLI

TGIR Huma-Num (UMR 3598, CNRS)

### LYDIE DANJEAN

TGIR Huma-Num (UMR 3598, CNRS)

### STÉPHANE POUYLLAU

TGIR Huma-Num (UMR 3598, CNRS)