Impact of Open Access scholarly publications on the web

Lucie Loubère, Fidelia Ibekwe

Impact of Open Access scholarly publications on the web.

Loubère Lucie
lucie.loubere@univ-amu.fr

Ibekwe Fidelia
fidelia.ibekwe@univ-amu.fr

Aix Marseille Univ, Université de Toulon,
IMSIC, Marseille, France

Introduction. *We present a methodology and tools for studying the impact of scholarly publications published on an Open Access (OA) repository on the web.*
Method*. From the logs to Open Edition pages, a platform that hosts mainly OA scholarly publications in the Social Sciences and Humanities, and the digital traces left on the web by the visitors, we identified web pages that cited publications from the Open Edition platform and delineated the citation contexts using the TextTiling thematic segmentation library. With Unitex, a Natural Language Processing engine, we identified noun phrases from the citing web pages and built a co-occurrence graph linking the cited Open Edition publications to the phrases found in their citation contexts on the web.*
Analysis*. We generated a dynamic and interactive graph interface which enables the user to navigate between the cited publications from Open Edition and their contexts of citation on the web.*
Results*. The exploration of the graphs showed how OA scholarly contents were being reused on the web by various stakeholders. In particular, our method enables the user to visualise what the citing web pages say about the cited scholarly publications, who the citers are (organisations and persons, their reasons for citing the scholarly publication and the sentiments (positive or negative) expressed by the citer(s) about the cited OA publication.*
Conclusion*. We observed a strong lexical coherence between the cited scholarly publications from Open Edition and the citing web pages. Also, much of the sentiments expressed by the citers were favourable (positive), thus upholding the widespread belief that the scientific enterprise is generally well regarded by members of the public at large.*

1. Introduction

Studying the impact of scholarly publications is the goal of the metrics field (biblio/scientometrics) (Filliatreau, 2008; Hicks, Wouters, Waltman, Rijcke and Rafols, 2015; Robinson-Garcia, Sugimoto, Murray, Yegros-Yegros, Larivière and Costas, 2019). However, these fields focus solely on the impact of scientific productions within the academic sphere using mathematical and statistical models that yield quantitative indicators. Also, scientometrics and bibliometrics studies tend to focus almost exclusively on (co)-citation counts of authors, journals or keywords. While this has enabled the production of apparently 'neutral numeric indicators' such as citation counts, impact factors or the H-index, the reasons for the citations remain obscure. The algorithms and visualisation packages for exploring scholarly content are based on a quantitative analysis of the logs to scholarly publication platforms, hence their output continues the same trend as bibliometrics indicators.

The availability of scholarly publications in open access online repositories has increased the chances of dissemination and thus of re-use not only by the scientific community but also by the general public. The diversification of online publication media outside of the traditional scholarly publication channels (journal, conference or books) has led to a diversity of the audience liable to access and thus reuse scholarly content on the web. We are witnessing a proliferation in the means of disseminating scholarly publications via academic blogs, scientific magazines destined to a wider audience. Scientific publications and discoveries are now being announced via social media (Twitter in particular) before being amplified by the mainstream media. All this has led to the emergence of less scholarly and more "social" metrics. Grouped under the name of 'altmetrics' (O'Neill, 2016), these more recent indicators of scholarly impact are a testimony to the current porosity between the academic community

and the society at large. However, "altmetrics" has continued in the same tradition as its predecessors, biblio/scientometrics, producing quantitative indicators based on the number of downloads, posts, likes, tweets or retweets that a scholarly publication receives on the web. None of the above indicators provide information on the actual use of the scholarly publication cited nor on the reasons for their citation.

Our study aims to bridge this gap by proposing a combined approach, symbolic and numerical, in order to study the impact of scholarly publications on the web. Our study is part of the *Open Knowledge Appropriation* project (OKA) funded by the Initiative of Excellence of Aix-Marseille University in France.
The research questions posed by this project is to determine how scholarly publications hosted on the Open Edition platforms (https://www.openedition.org/?lang=en) are appropriated by the general public, i.e. what traces of these publications are found on the web, in what form and for what purposes? Open Edition comprises four distinct platforms that host books, journals, blogs, scientific events and news in the Social Sciences and Humanities (SSH). Most of its content is open access while others require a subscription.

Our specific research question is on the written forms of appropriation of scholarly literature in the public arena, i.e., we wish to study the traces people left on the web about scholarly contents outside of *bona fide* scholarly communication channels (journal articles, conferences, books, etc). Our focus is therefore not on generating numerical indicators of impact such as count citations received by a given scholarly publication nor to count the number of visits (logs or traffic to those publications). Rather, we aim to shine a light on what the citing web pages (the citers) actually say about the cited scholarly publications. This requires looking at their citation contexts. This has always been a bottleneck issue in citation studies because it requires knowledge intensive approaches and methods amongst which Natural Language Processing (NLP) is an obvious choice.
Our study there departs from traditional biblio/scientometrics in that we designed a methodology and tools to track the traces of appropriation of scholarly publication in non-scholarly contexts based on the citation contexts. Our research is therefore concerned with how scientific ideas circulate amongst the layman/woman, thus with science popularisation in the society.

2. Methodology

Our methodology brings together symbolic and numerical approaches. The symbolic approach draws on Natural Language Processing (NLP) methods and tools while the numerical approach deploys a top-down hierarchical clustering algorithm coupled with graph generation and information visualisation methods and tools. Figure 1 hereafter is an overview of our corpus processing suite. The sections hereafter describe in more details the different stages from corpus collection to information visualisation of the citation contexts of OA publications.

2.1 Corpus collection

As stated previously, Open Edition comprises four platforms, namely:

- Open Edition Books (https://books.openedition.org/)
- Open Edition Journals (https://www.openedition.org/catalogue-journals)
- Open Edition research blogs (https://www.openedition.org/catalogue-notebooks)
- Calenda: announcements of scientific events (https://calenda.org/)

Our study begins by collecting the logs data of visits made to scholarly contents published on these four platforms. Open Edition conserves in its logs, the identifiers of the session, the

visitor's ID, the time of connection, its duration, the URL from where the visit came (the referrer) and the page visited on its platforms. Figure 2 hereafter shows an extract of the logs.

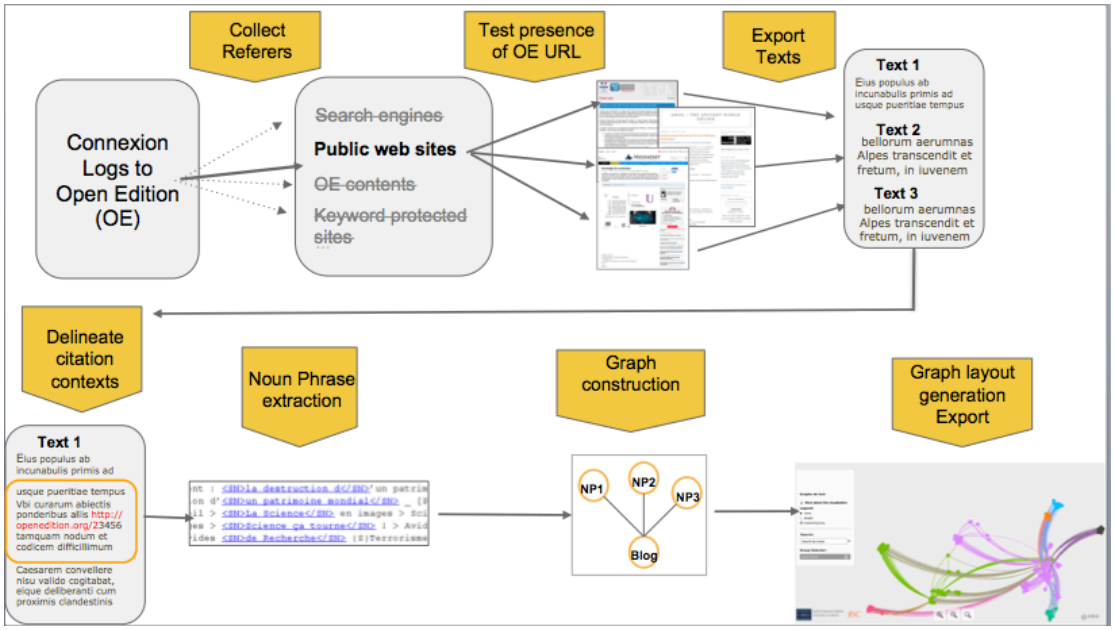

Figure 1: Overview of our corpus processing suite.

idvisit	idvisitor	visit_last_action_time	visit_first_action_time	referer_url	name
177909	1.51571811112E+19	2017-01-01 00:00:14	2017-01-01 00:00:00	https://www.google.fr/	rha.revues.org/6885
177909	1.51571811112E+19	2017-01-01 00:00:14	2017-01-01 00:00:00	https://www.google.fr/	rha.revues.org/7255
177914	1.48617546873E+19	2017-01-01 00:00:25	2017-01-01 00:00:15	https://www.google.it/	romatevere.hypotheses.org/297

Figure 2: Examples of logs to pages hosted on Open Edition.org.

In this project, logs were collected over a 1-year period, between 01/01/2017 to 01/01/2018.

2.2 Identification of visiting web pages

Using the log collection tool Matomo (https://matomo.org/), we eliminated logs that came from robots and search engines because they only provide the queries entered by visitors to get to Open Edition pages but no citation context of the visited document. We also excluded password protected sites such as social media platforms (Facebook, Snapchat, ...) and intranets of corporate or individual organisations since we could not access to the precise pages containing the link to an Open Edition publication. Since the application of the European General Data Protection Regulation (GDPR), some servers no longer specify the complete URL of the referrer page. Thus, visits coming from Wikipedia only provide the generic link 'http://wikipedia.fr' regardless of the actual Wikipedia page from the visit was made. Hence, we could not exploit citation contexts coming from Wikipedia.
After removal of these domain names, we then extracted the textual contents of the visiting web pages and identified the position of links to publications hosted on Open Edition. This yielded 33, 667 citing web pages for the 1-year period considered.

The corpus we collected is heterogeneous on several levels. The first level of heterogeneity lies with the hosted content on Open Edition platforms. The four OE platforms cover the whole spectrum of social sciences humanities. Publications on neuropsychology sit alongside those on literature, philosophy, history, sociology or law. The disciplinary diversity mentioned above is linked to the thematic diversities of social sciences and humanities titles published on Open Edition. A sociology journal can deal with a multitude of subjects which

themselves can be of interest to several fields. A second level of heterogeneity is observed in the diversity of re-appropriating media. Amongst the citing web pages, we find discussion forums, blogs of researchers, government sites, web sites of NGOs, articles by the mainstream media and sites and blogs by the layman/woman. This wealth of content, far from being a limitation, constitutes in itself a first level of exploration of our corpus.

2.3 Delineating citation contexts

The next step is to delineate the citation contexts to Open Edition documents in the citing web pages content. To this end, we first extracted the textual contents of the 33, 667 citing pages. However, we are not concerned with the whole article on every citing page but with the immediate context of the link to Open Edition. For this, we turned to thematic segmentation approaches and after literature review, we chose the Textiling library (Hearst, 1997). This tool segments a text into lexically cohesive thematic segments (http://people.ischool.berkeley.edu/~hearst/research/tiling.html). For the web pages where the citation to an Open Edition document contains little or no surrounding text (isolated links), we performed a manual selection of what could constitute a citation context.

2.4 Extraction of noun phrases from citation contexts

Next, we extracted the noun phrases (NPs) contained in the delineated citation contexts using Unitex, an open access Natural language Processing (NLP) package (https://unitexgramlab.org/fr). Unitex is based on Intex®, a proprietary NLP engine (http://www.nyu.edu/pages/linguistics/intex/). Unitex performs several NLP tasks such as morphological analysis, tagging, identification of textual concordances or regular expressions and corpus alignment (in different languages). Its graphic interface enables the user to draw the sequence of morpho-lexical categories that s/he wishes to extract from a corpus as Finite State Automata (FST) which are then applied to the corpus. The sequences identified by these FST are highlighted in the corpus and can then be extracted.
We wrote several finite state automata to identify phrases comprising at least 2 sequences of one or more noun(s), article(s), adverb(s) and adjective(s) which are searched for recursively. Figure 3 hereafter is an example.

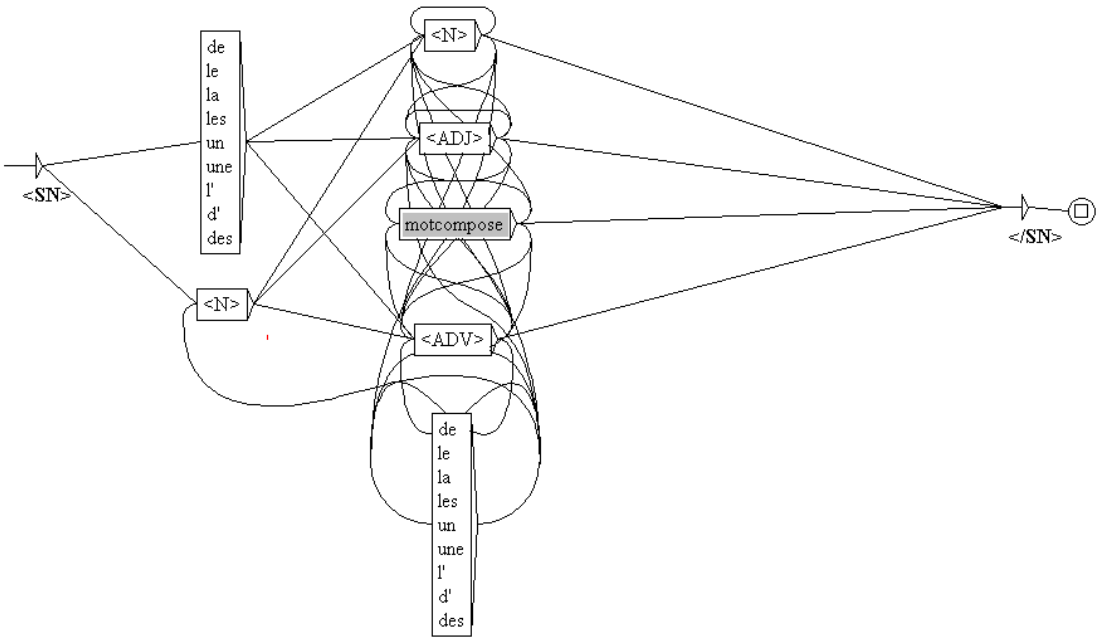

Figure 3: A finite state automaton that identifies noun phrases contained in the citation contexts to Open Edition documents.

Figure 4 is an example of NPs identified which then serve as input for the next phases of processing (clustering, graph generation and visualisation).

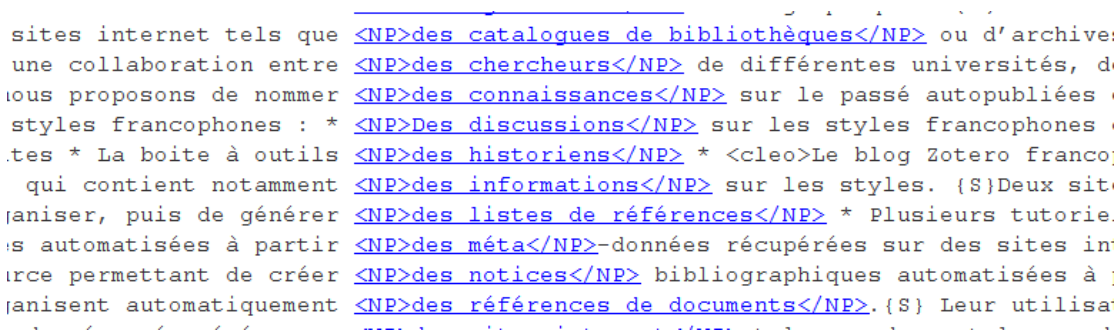

Figure 4: Examples of noun phrases identified from the citation contexts.

2.5 Building graphs of co-cited contexts and of cited documents

The next step is to build a co-occurrence graph connecting noun phrases from the citation contexts to Open Edition publications they cited. The graph generation algorithm used here is the Leuven method (Blondel, Guillaume, Lambiotte, and Lefebvre, 2008) available in the Gephi package (Bastian, Heymann, and Jacomy, 2009). It proceeds in two phases: first, the algorithm creates a cluster for every node, then each node is absorbed into the cluster closest to it. If this structure allows a gain of modularity, it is conserved, if not, it is cancelled and the process is reiterated until it converges or it reaches optimal modularity, i.e., no change improves it. The second phase reproduces the same pattern by considering the communities built previously as distinct nodes, the edges connecting the nodes are calculated as the sum of the weights of the links between the individuals constituting the two communities. These graphs of similarity thus relate nodes— the cited Open Edition page and the noun phrases found in the citation contexts, by edges (the number of co-occurrence of the same phrases in the same contexts citations).

2.6 The interactive graph navigation interface

The graph thus generated becomes a navigable interface only in its dynamic export. We used the sigma.js (http://sigmajs.org/) library developed by Alexis Jacomy for the Oxford Internet Institute which is used to generate interactive graphs on websites. This program also allows us to enrich this visualisation with additional elements such as links to the citation contexts and the citing web pages. For the sake of legibility, we assigned colours to the clusters formed. The final navigation tool is divided into three parts: a search component, the main graph and an information component (see Figure 5 below).

Figure 5: The navigation interface showing cited documents on Open Edition and their contexts of citation.

The left pane contains a search tool which can be used to rapidly locate parts of or an entire NP or to select a cluster by its colour. The central pane contains the graph which can be navigated. A click on a node zooms the graph to its neighbours and displays contextual information on the right pane showing the links to the noun phrases in the citation contexts as well as connections with other phrases and the cited document.

3. Results

To illustrate the results of our corpus processing suite, we will analyse the citation contexts of three blogs hosted on Open Edition. For reasons of legibility, we added a root "ed_" to the names of these blogs:

- ThatCamp Paris (ed_tcp)
- Monde sociaux (ed_sms)
- Criminocorpus (ed_criminocorpus)

Figure 6 hereafter shows the immediate citation context of these three blogs with their names highlighted as prominent nodes and around them, the noun phrases extracted from their citation contexts on the citing web pages.

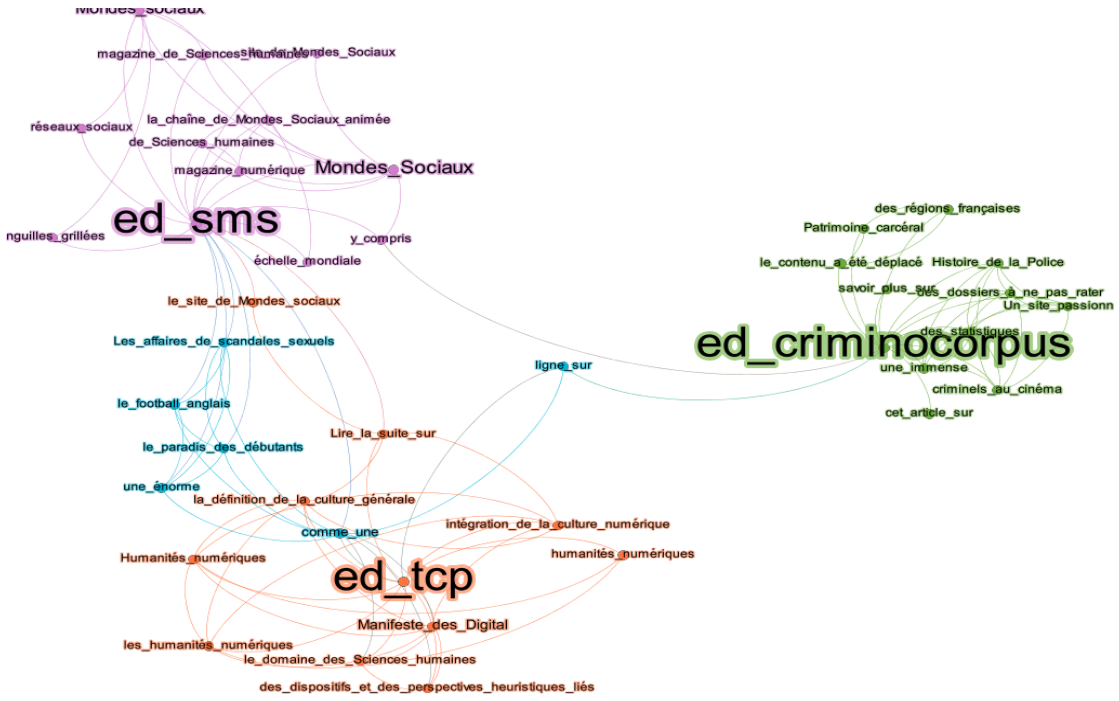

Figure 6: Overall graph showing the citation contexts of three blogs: '*ThatCamp Paris (ed_tcp), Monde Sociaux (ed_sms), Criminocorpus*.'

3.1 Citation Contexts of the blog 'ThatCamp Paris'

Created in 2009 for the preparation of the first unconference on Digital Humanities, the ThatCamp blog (http://tcp.hypotheses.org/318) published a manifesto to promote informal discussions and collaborations on Digital Humanities. It offered a launching pad for gathering ideas, participations and proposals of interested persons to organise various events such as workshops, data sprints, etc. Although the last article of the blog was dated 2015, visits to this blog continues and we were able to identify 38 citing web pages between January 2017 to January 2018. The phrases extracted in citation contexts to this blog are coloured in the orange zone in Figure 6. A further splitting of this cluster brought several themes to the fore (see Figure 7 here below).

As shown by its nice shape, this very homogeneous graph recalls the foundation of Digital Humanities which is strongly embedded in the social sciences and humanities. We also see the presence of a node labelled "*Manifeste_des_Digital*", a reference to the collective document from the first ThatCamp. Recurrent phrases in the citation contexts mostly echoed the subject of the blog, namely:

- *the field of digital humanities, (le champ des Humanités Numériques)*
- *digital humanities (humanités numériques)*
- *Manifesto of the digital humanities*

Other phrases found in the citation contexts of this blog revealed actors active on this topic, be they partners of the *ThatCamp* movement, actors of digital humanities, doctoral schools, universities or scholarly associations. All the actors appear to be united by their citations and reliance on articles published on this blog. The exploration of this wider audience allowed us to identify the announcement for a day's workshop organised by a doctoral school in a French University (http://www.univ-paris3.fr/ed268-2017-2018-conference-du-samedi-a-quoi-bon-les-humanites-numeriques-pour-les-sciences-du-langage--466360.kjsp?RH=1263513068245).

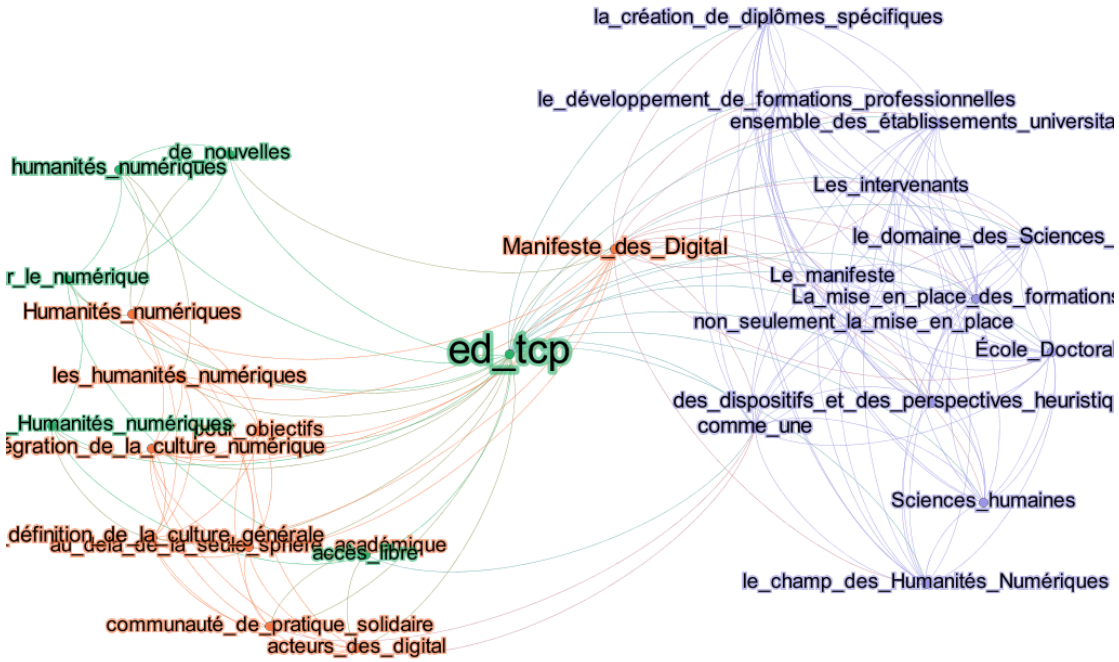

Figure 7: Sub-graph showing the citation contexts of the *'ThatCamp Paris'* blog.

3.2 The citation context of the Criminocorpus blog

The second blog *Criminocorpus* (https://criminocorpus.hypotheses.org/) is managed by CLAMOR (UMR 3726), a CNRS (*Centre National de Recherche Scientifique*) research laboratory as part of an online Francophone scientific publication platform on the history of justice, crimes and punishment accessible here (https://journals.openedition.org/criminocorpus/). This publication platform hosts a virtual museum, a journal and a blog and it is the latter whose external citation contexts we analysed. With a strong publishing activity (several articles per day), it is also the blog that received the most citations on the web (94 citation contexts). The noun phrases extracted from these citation contexts are coloured in green in Figure 6. A more detailed view of this cluster (Figure 8) brought to the fore several topics.

A significant proportion of noun phrases in the citation contexts of this blog echoed its theme (*criminal phenomenon, criminals in the cinema, statistics', the museum, the first digitally native museum)* while others echoed the other activities of this publication platform, in particular its virtual museum as shown in noun phrases such as '*thematic exhibitions, permanent exhibitions, cultural events*'. Phrases such as '*ordinary violence, the criminal controversy, birth of the scientific police*' reflected book reviews published by the *Criminocorpus* journal. Yet other noun phrases reflected how the publication was regarded by readers with phrases such as "*un site passionnant*" (*an exciting site*), "*dossiers à ne pas rater*" (*dossiers not to be missed*), "*objets constituant des sources particulièrement rares*" (*objects constituting relatively rare sources*), thus showing the positive recommendations of the blog's citers.

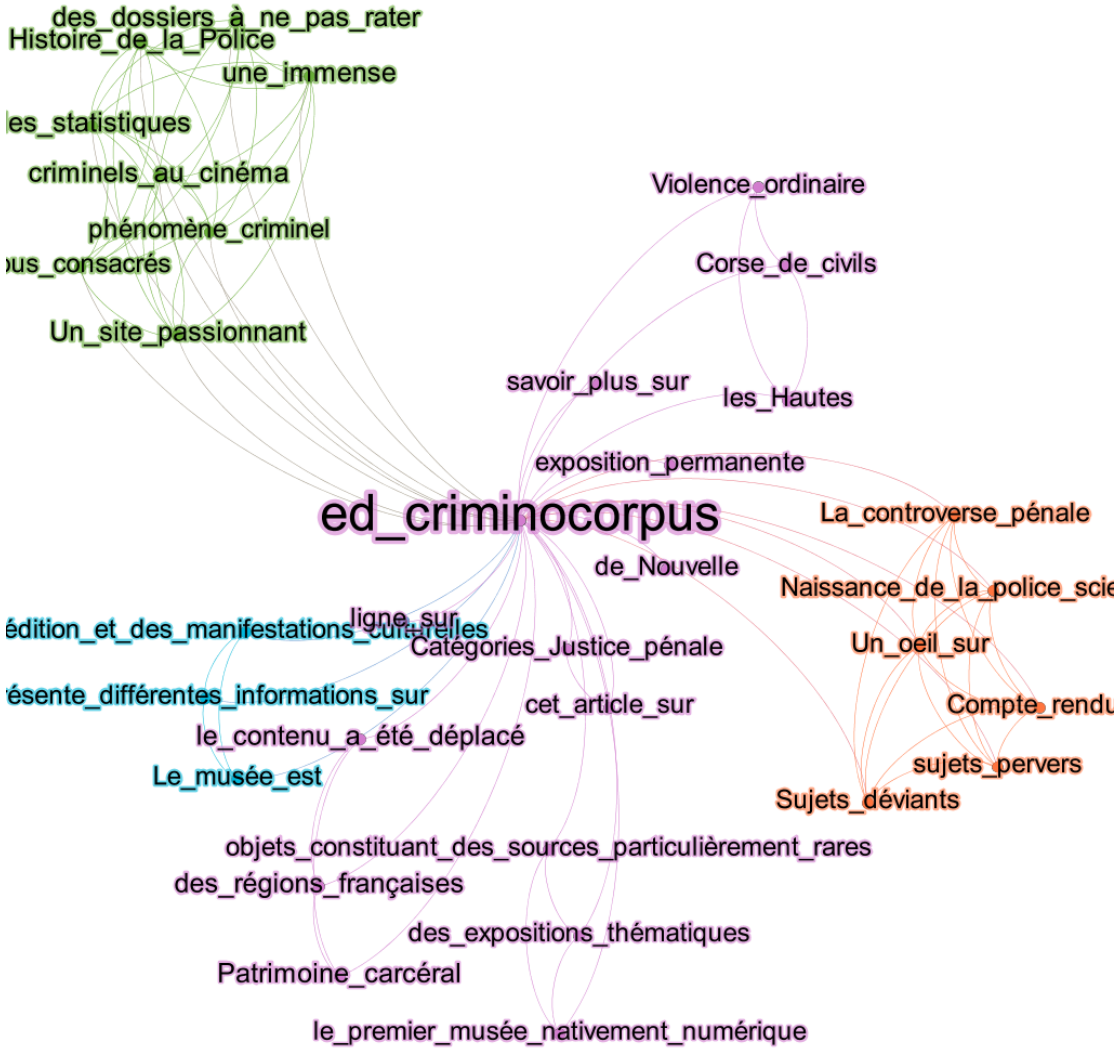

Figure 8: Sub-graph showing the citation contexts of the '*Criminocorpus*' blog.

3.3 Citation contexts of the Mondes Sociaux blog

The third blog, '*Structuration des Mondes Sociaux*' (Structuring Social Worlds) is edited by a French Excellence research laboratory called "Labex SMS" (https://sms.hypotheses.org/) based at the University of Toulouse. The blog is described as a French-language open access digital magazine promoting the sharing and the circulation of scientific knowledge to a wider audience through writing short articles on already published scientific articles. As a bi-monthly magazine, it yielded 78 exploitable citation contexts found in the purple cluster (*ed_sms*) in the main graph (Figure 6). Further splitting of this cluster revealed a very dense structure (see Figure 9 hereafter) that brought to the fore more specific topics.

Mainly, the phrases found in the citation contexts of this blog are a reflection of its goal and theme: to popularise scientific knowledge to a wider audience through publishing short and accessible pieces. The orange zone in Figure 9 (bottom left) reflected this scientific mediation role of the blog through phrases such as "*scientific mediation, of the public, humanities, social media, is a digital magazine*" (*médiation scientifique, du public, sciences humaines, médias sociaux*, *est un magazine numérique*). The green cluster (upper right) contain phrases that identified the blog as a scientific publication platform by French researchers "*french researchers, research work, visibility in front of the public, Mondes sociax site*" (*les chercheurs français, les travaux de recherche, visibilité auprès du public, site des Mondes Sociaux*). Other phrases reflect the trust awarded to this blog by members of the public such

as "*recognition of the researcher by a*" (*la reconnaissance de chercheur par un*). The other sub-clusters echoed the different themes covered by the publications of this blog as seen in the phrases at the top of this graph "*sexual affairs scandals, english football, players*" (*les affaires de scandales sexuels, football anglais, les joueurs, tourisme sexuel*).

Figure 9. Sub-graph of showing the citation contexts of the '*Structuration des Mondes Sociaux*' blog.

Overall, exploration of the citation contexts of these three blogs on web pages shows some regularities in the aspects of the information contained in the cited documents which are reused on the web. By focusing on noun phrases as content bearing elements in linguistic utterances, we were able to highlight the strong lexical proximity between cited scholarly publications on Open Edition platforms and their contexts of re-use and citation in the public space (web pages).

The technical characteristics of the cited documents are expressed in a descriptive way by phrases evoking the numerical formats, the modality of access (open access, free) and the editorial committees. Citing pages also emphasised the scientific identity of the cited documents and thus the trust in which members of the public continue to hold scientists and scholars. Finally, phrases conveying the positive evaluations and recommendations given to the scientific publications reinforce the feeling of trust and respect the public has for all things scholarly or scientific.

Concluding remarks

Our study has shown how scientific literature from a largely open access hosting platform is re-appropriated and repurposed in the public arena. Our study has also shown that it is possible to expand citation studies in a qualitative direction by focusing not only on counting their numbers but on their real impact by studying their contexts of reuse. We also showed how symbolic (Natural Language Processing) and numerical approaches (clustering, graph generation) can be fruitfully combined to perform a knowledge intensive Human Language Technology task.

The work presented here is still in progress. For instance, our NPs extraction graphs require refining as they sometimes led to identifying syntactically invalid NPs. For instance, one of the NPs extracted "*est un magazine numérique"* (*is a digital magazine*) is due to lexical and morphological ambiguities. The form "*est*" in French can mean either the "east" or the present tense of the verb "*to be*" (*être*) and in this case, it is the latter, so this phrase should not have been extracted with the verbal element.

While the interactivity of the information visualisation interface we built is an asset for exploring complex undirected graphs, it can also constitute a source of bias. A possible bias may lie in the variations in the number of citation contexts for each cited document. The three publications (blogs) we used as illustration yielded 28, 94 and 79 citation contexts respectively. The graph generation process, by its iterative nature, can penalise publications with few citation contexts. A second bias linked to a variation in the lexical density of the citation contexts may occur. For instance, the *Structuration des Mondes Sociaux* blog had fewer citing pages although its citation contexts returned more noun phrases because of a higher lexical density of these contexts. On the other hand, the *Criminocorpus* blog had more citing pages but whose citation contexts yielded fewer noun phrases. This was because the web pages citing the *Criminocorpus* blog repeated the same sentences from one site to another and also because of the presence of many discussion forums where the immediate citation context contained few sentences. These biases remind us that while corpus processing based on quantitative aspects enable us to synthesise and navigate the information in a user-friendly manner, it is not sufficient for a fine-grained exploration and interpretation of the phenomenon under study.

Acknowledgements

The project during which this publication was prepared received funding from the Excellence Initiative of Aix-Marseille University - A*MIDEX, a French "Investissements d'Avenir" program.

References

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsm*, *8*(2009), 361-362.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Filliatreau, G. (2008). Bibliométrie et évaluation en sciences humaines et sociales : une brève introduction. *Revue dhistoire moderne contemporaine*, *n° 55-4bis*(5), 61-66.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, *23*(1), 33-64.

Hicks, D., Wouters, P., Waltman, L., Rijcke, S. de, & Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature Publishing Group*, *520*(7548), 429-431.

Marchand, P., & Ratinaud, P. (2012). L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT*, *2012*, 687-699.

O'Neill, J. (2016). NISO recommended practice: Outputs of the alternative assessment metrics project. *Collaborative Librarianship*, *8*(3), 4.

Robinson-Garcia, N., Sugimoto, C. R., Murray, D., Yegros-Yegros, A., Larivière, V., & Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics*, *13*(1), 50-63.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX. Masson.*