Scholarly Communication and Documentary Fragmentations in the Public Space: a Functional Citation Study

Fidelia Ibekwe, Lucie Loubère

Scholarly Communication and Documentary Fragmentations in the Public Space: a Functional Citation Study

Fidelia Ibekwe

Lucie Loubère

1. Introduction

Open Edition (OE) is a multilingual repository hosting social sciences and humanities literature in several languages and from all over the world. It comprises four platforms each hosting a specific type of document—books, journals, blogs and calendar events—totalling approximately 500,000 documents. OE has been championing open access to science although a few of the titles hosted require subscription to access their full text.

A multi-disciplinary consortium was formed in order to investigate how the scientific literature published on OE platforms are re-appropriated and reused in the public arena. This gave rise to the "*Open Knowledge Appropriations*" (OKA) project funded by the Aix-Marseille University in France, through its Initiative of Excellence foundation, a.k.a. AMIDEX. This paper reports on work from the OKA project aimed at analysing the digital traces left by citations of contents published on the OE platforms with the aim to better understand how members of the public appropriate and reuse academic knowledge in contexts that are not scientific.

2. Document collection

Open Edition (OE) comprises four platforms, namely:

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

The first stage of our study was to collect the logs of visits made to scholarly contents published on these four platforms. OE 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 and the page visited on its platforms. 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 OE pages but no citation context of the visited document. We also excluded password protected sites and intranets of corporate or individual organisations since we could not access the precise pages containing the link to an OE 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 OE. This yielded 33,667 citing web pages for the one-year period considered (between 01/01/2017 and 01/01/2018). We also collected tweets

containing a link to OE from 2013 to 2017. This resulted in 830,085 tweets including 375,994 original tweets.

3. Automatic document processing suite

After collecting the corpus of web pages containing citations to OE documents, we had to find a way to delimit the context of citation on the citing pages. Indeed considering the entire page on which the citation to OE appeared would make the context too diffuse because web pages can by nature address several themes. We therefore sought to reduce the window to the paragraph containing the citation to an OE content. For this, we used the Texttilling tool (Hearst, 1997) which allows, after tokenization and elimination of stop words, to calculate for each paragraph the lexical proximity that it has with the surrounding ones. Once the citation contexts had been automatically extracted, we performed a surface morphological tagging followed by noun phrase extraction using the Unitex/GramLab package (https://unitexgramlab.org/), "an open source, cross-platform, multilingual lexicon and grammar-based corpus processing suite."

These noun phrases become the input into the further stages of mapping and visualisation of citation contexts. To do this, we generate a co-occurrence matrix for each citation context, the noun phrases (NPs) present in them and the OE resource cited. Using the Gephi software (Bastian, Heymann, & Jacomy, 2009), we generated dynamic graphs with the application sigma.js (Jacomy & Plique, 2017) in order to explore the citation contexts of OE content via NPs and additional data such as the types of citing pages, the citation functions and the time elapsed before the citation took place. The entire data processing procedure is shown in Figure 1. Figure 2 shows an example of a graph generated to explore the citation contexts of OE contents on the web.

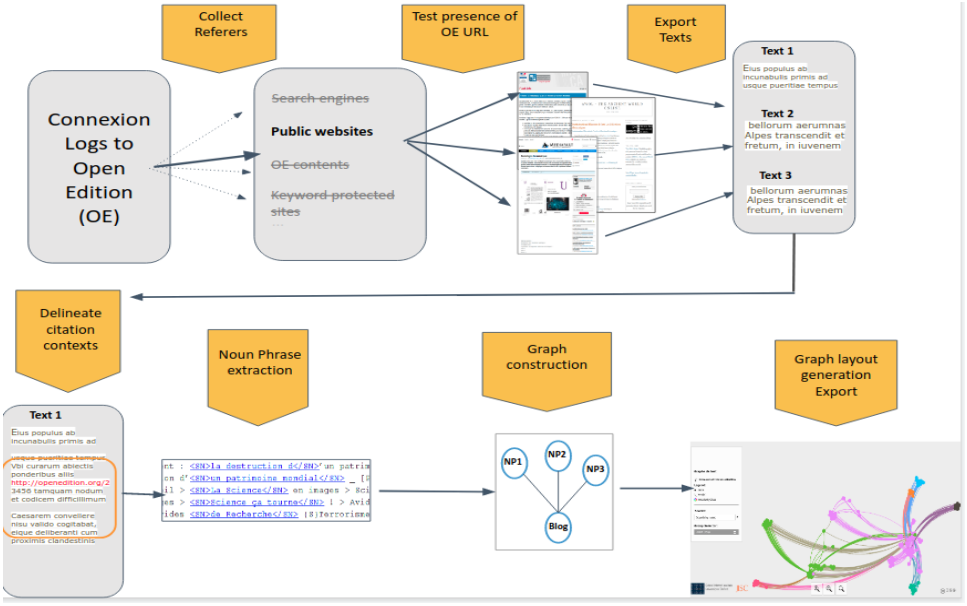

Figure 1. Overview of the data processing suite.

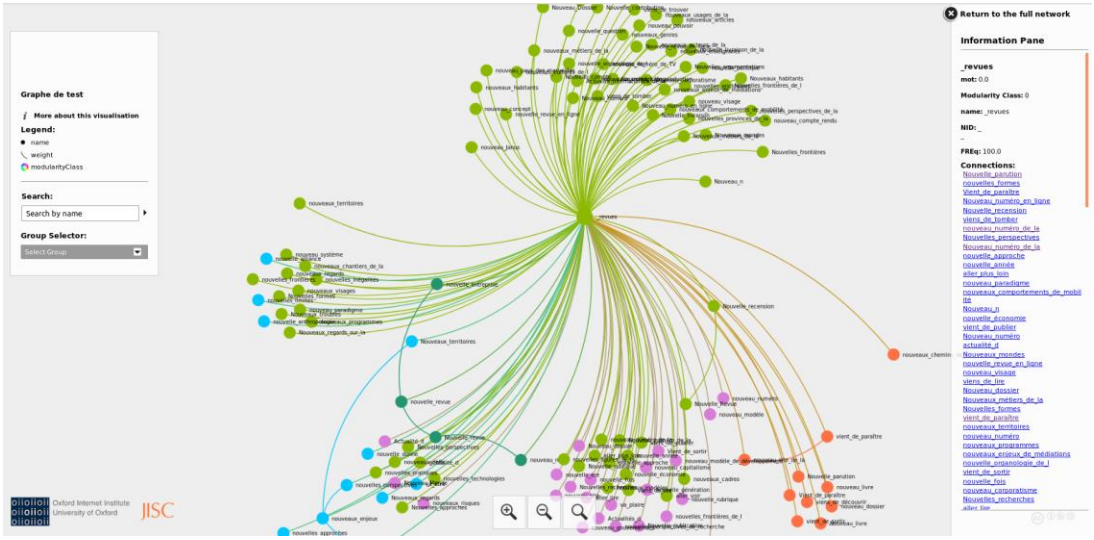

Figure 2. An example of the generated graph for exploring citation contexts of Open Edition contents.

4. Research questions

In this paper, we will address the following research questions.

1. What motivates members of the public to reuse and cite academic publication literature? In other words, what are the citations functions of academic content?

2. What are the documentary forms of the citing documents and what can this tell us about the documentary mutations of the initial cited academic document as it journeys through the cyberspace?
3. What is the average time lapse between the publication of a scientific literature and its citation and appropriation by members of the public?

To investigate the above research questions, we manually analysed the citation contexts (the surrounding texts) of some highly cited OE contents. We will illustrate our preliminary findings on two OE documents. The first is the "Digital Humanities Manifesto" (http://tcp.hypotheses.org/318) first published in French on 26th March 2011 and updated on 25th January 2012. The second is the content page of the journal *Ethique et Publique*, an international journal of public and social ethics (http://journals.openedition.org/ethiquepublique/2723).

4.1 Citation functions

All the examples we manually examined tended to show that academic content are cited in a positive manner, either to recommend content to read to others ("see also" relation in thesaural parlance) or to buttress the argumentation of the authors, hence as scientific. The graph in Figure 3 below shows the web pages citing articles published in *Ethique et Publique* (in the centre of the graph).

The citing pages are coloured according to a typology of citation functions detailed in the key. We identified 10 functions, namely: research news (*actualité recherche*), argument, source, cv, recommendation, environmental scanning (*veille*), deepening knowledge (*approfondissement*), openarchive, promotion, and recommendation.

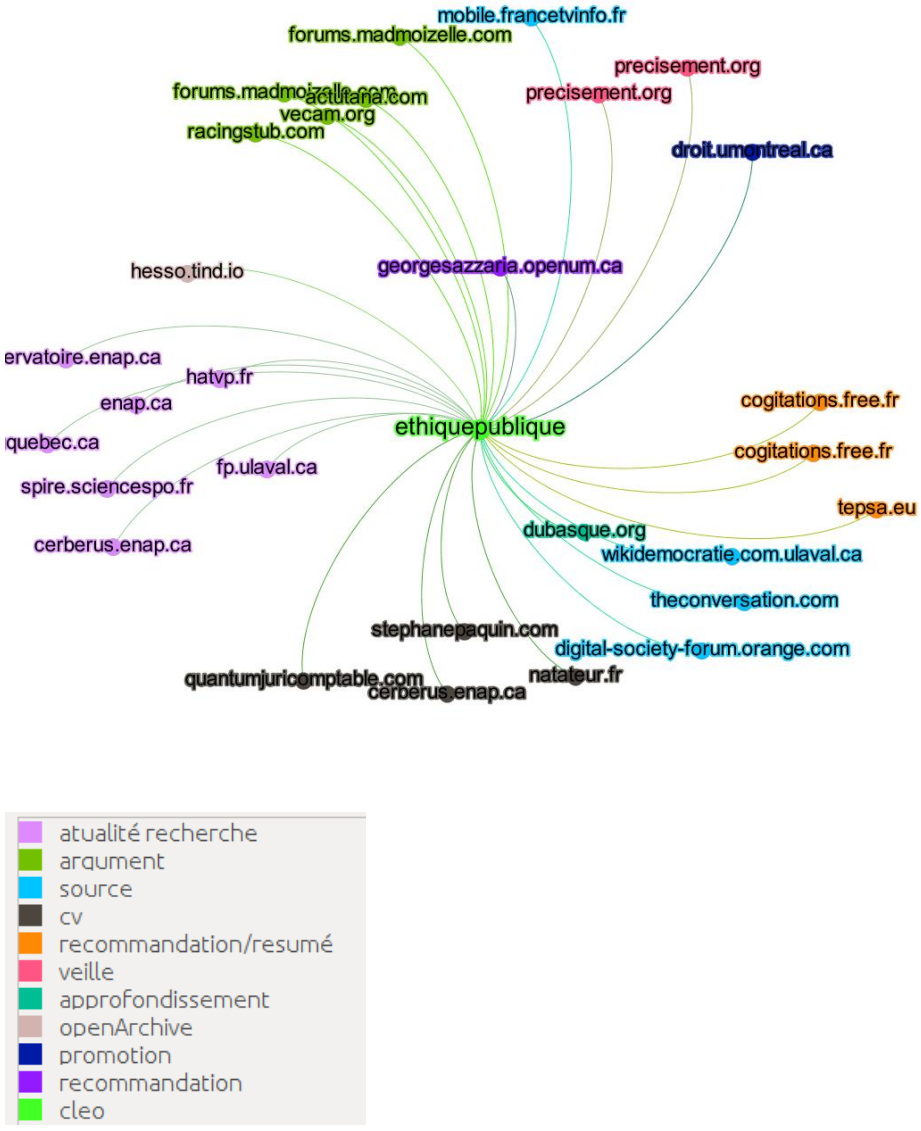

Figure 3. Graph showing citations to articles published by *Ethique publique*.

4.2. Types of pages citing academic content on the web

Our second research question was to know if there were specific types of web pages that cited academic content on OE. Our preliminary results tend to suggest that there was no predominant pattern in the origins of the citing pages and that academic content can be cited by all sorts of web pages including blogs, press

releases, book reviews, newspaper articles, scientific articles, discussion forums, library catalogs, institutional websites such as government agencies or ministries.

For instance, organisations like the Scéren[1] or Canopé associated with the French Ministry of Education often cited publications from OE platforms in its discussions in order to give recommendations of readings to its users on education and pedagogy. Thus, academic content on OE platforms form part of an environmental watch and scanning strategy (alert system) of these organisations in order that they may keep abreast of the evolution of trends on the topics they are surveying.

Social media platforms such as Twitter also cited content published on OE. The tweets data showed that this platform was used used to announce the publication of books, blogs, journal articles or scientific events on OE platforms. Figure 4 shows a co-occurrence graph of words from tweets announcing research news in the field of Education science. This usage of Twitter to disseminate news and discoveries has been observed in other contexts.

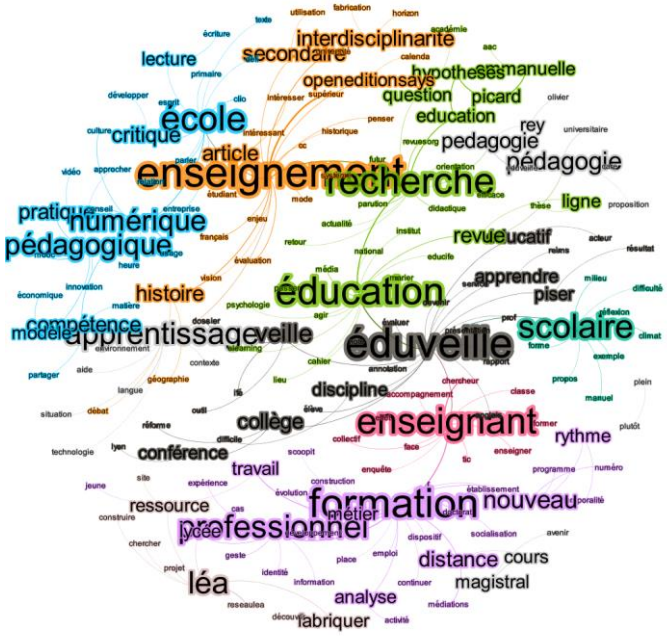

Figure 4. Graph of co-occurring words in tweets used to disseminate research published on OE in the field of Education.

4.3. Time lapse before citation of an academic content by members of the public

Our third research question sought to determine how much time elapses between the publication of academic contents and their citation on public web sites. For this,

[1] http://www2.cndp.fr/accueil/accueil.htm

we first identified the publication date of OE contents and that of web pages citing them and calculated the difference. We then generated a dynamic graph that shows the chronological appearance of citing pages to an OE content.

Two videos accompany this article (https://doi.org/10.35492/docam/6/1/7). The first shows the time lapse between the 23 articles published by the *Ethique et Publique* journal between January 01, 2001, and June 13, 2018, and their 29 citation contexts which appeared between September 1, 2013, and November 24, 2018. The time lapse is shown in number of years. The second video shows the same information for citations to the blog article "The Digital Humanities Manifesto," first published on March 26, 2011. This article generated 19 citations on web pages which appeared between July 22, 2010 and April 7, 2018.

Our findings corroborate the patterns of citation of scientific literature already observed within the scientific community. Some academic contents receive attention in the days or weeks following their publication while others can be dormant for months or even years before receiving citations owing to current affairs in the public space. The four citation patterns observed in the literature are:

- *sleeping beauties*: these are publications that are ignored or little cited for years or even decades, and then suddenly begin receiving a lot of citations owing to an event or a discovery (Raan, 2004);
- *long tail*: this model popularised by Chris Anderson corresponds to the well-known Saint Matthews effect n in citation studies whereby a few titles/publications (the heavy weights) receive 95% of citations while the majority (the tail) receive few citations or sales (the remaining 5%. However, Andersen's theory is a commercial version of the well-known power law found in economics (Pareto's law) or in the field of bibliometric studies: Lotka's and Zipf's 1926 laws modelling the distribution of bibliographic units in scientific publications. Lotka studied the distribution of authors in a field while Zipf modeled the distribution of words in the same document;
- *unexpected reader*: this citation behaviour pattern was proposed by Phil Bourne (Smith, 2011) in the context of open access advocacy with the argument that the latter enables untargeted audience to access to scientific publications. This is of particular interest to our study as it could reveal how scientific literature can be repurposed in contexts and for ends that the authors of the initial publication had not intended.
- *silent conversation*: Dacos (2010) suggested this reading pattern to illustrate the fact that some readers will consume content without writing about it. They are therefore a difficult audience to identify.

Conclusion and perspectives

We manually analysed the citation functions, types of citing web pages and the time elapsed between the publication of academic content on the web and their

appropriation in the public arena. Our long-term goal is to formalise then automate the processes involved. In order to automate the categorisation of types of visiting web pages (research question 2), we tried the following procedures:

1. the 45,190 citing pages to OE documents were subjected to a list of 5,320 categorised visiting web pages already done by this organisation, this enabled the automatic categorisation of 14,950 (33%) citing pages;
2. the remaining 30,240 citing pages were then subjected to a manually curated filter we created by scanning the URL of pages citing OE for cue words such as "*blog, forum, univ., uni., université, academia*" which points to their type. Using this list which currently contains 132 cue phrases as boot strap, we were able to automatically categorise an additional 11,331 (25%) of citing pages.

Further studies are needed to find parameters to categorise the remaining 18 909 sites. This will probably involve an incremental acquisition procedure of more cue words with which to expand our current list.

Another point needing further investigation is the fragmentation space of the original document in the cyberspace. By this, we aim to track out how the original scientific document cited on OE evolves from a documentary point of view across the cyberspace.

References

Andersen, C. (2008). *The long tail: Why the future of business is selling less of more*, Hachette Books.

Dacos, M., & Mounier P. (2010). Les carnets de recherches en ligne, espace d'une conversation scientifique décentrée. In C. Jacob (Ed.), *Lieux de savoir* (vol. 2). Paris: Albin Michel.

Hearst, M. A. (1997), TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, *23*(1), 33–64. Retrieved from http://www.aclweb.org/anthology/J97-1003

Ibekwe-SanJuan, F. (2010). Semantic metadata annotation. Tagging Medline abstracts for enhanced information access. *Aslib Proceedings journal*, *62*(4/5), 476–488.

Raan, A. F. J. van. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 467–72.

Smith, K. (2011). The unexpected reader, *Scholarly Communications @ Duke*. Retrieved from http://blogs.library.duke.edu/scholcomm/2011/11/15/the-unexpected-reader/

DOI: 10.35492/docam/6/1/7