Studying the posts accumulation patterns of Altmetric.com data sources

Zhichao Fang[1], Rodrigo Costas[1,2]

[1]*z.fang@cwts.leidenuniv.nl; rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands

[2]DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, South Africa

Abstract
In this paper the posts accumulation patterns and altmetric post half-life of 13 Altmetric.com data sources are studied. *Created date* and *issued date* from Crossref are compared and aggregated to serve as the proxy for the first publication date of research outputs, combined with posted on date recorded by Altmetric.com, altmetric posts accumulation and half-life patterns analyses are conducted at the day time interval. Altmetric.com data sources vary in posts accumulation patterns, some altmetric posts accumulated very fast within the first few days after publication, such as Reddit, Twitter, News, and Facebook. They also hold a short altmetric post half-life in disseminating newly published research outputs. Syllabi, Policy documents, Wikipedia, Q&A, and Peer review accrued relatively more slowly, as they are not so concentrated on recent publications.

Introduction
The accumulation patterns of citations and usage metrics (views, downloads, etc.) were widely discussed in previous studies. Schlögl, et.al. (2014) reported that citations take several years until reach its maximum but most downloads accrued in the same publication year. Moed (2005) found that citations and downloads show different patterns of obsolescence and about 40% of downloads accumulated within the first 6 months after publication. From the perspective of altmetrics, Ortega (2018) made a comparison of temporal distribution at the month level among citations, views, downloads, Mendeley readership, tweets, and blog mentions recorded by PlumX, and concluded that tweets and blog mentions are the most quickly available metrics. The results based on *PeerJ* social referrals data of Wang, Fang, & Guo (2016) suggested that the number of "visits" to papers from social media (Twitter and Facebook) accumulates very quickly after publication. However, a large scale quantitative analysis on comparing the posts[1] accumulation patterns of different altmetric data sources at the micro-level time interval (day) is still missing. Our study addresses these issues and aims to answer the following research questions:

1. How are the altmetric posts accumulation patterns of various Altmetric.com data sources?
2. What is the difference of altmetric post half-life among different Altmetric.com data sources?

Data and Methods
In order to exhibit the accumulation patterns of altmetric posts of different Altmetric.com data sources at the day time interval, it is necessary to find a precise proxy for the first publication

[1] Posts collectively refer to the altmetric events on different data sources that recorded by Altmetric.com, such as tweets, blog mentions, policy documents citations, Wikipedia citations, etc.

date[2] of research outputs and post date[3] of altmetric records. As to the first publication date, Haustein, Bowman & Costas (2015) made a comparison between five kinds of publication dates that are likely to serve as the proxy for the actual first publication date: *online date* from the publishers, *Altmetric publication date*, *Altmetric first seen date*, *first tweet date* from Altmetric.com as well as *WoS indexing date*. However, according to their results, none of above dates represent a good proxy. As they suggested for future work, the first time a DOI was resolved has the potential of reflecting the first online publication date. In this study "issued date" and "created date" of DOIs collected from Crossref are combined to be used as the proxy for the first publication date, while the "posted on date" recorded by Altmetric.com for each altmetric event are collected to represent the post date of altmetric records.

Crossref dates as the first publication date

Until August 2017 Crossref has created and deposited 68,148,933 DOIs with different date information about them, Table 1 shows the description and coverage of *created date*, *issued date*, *published-print date* and *published-online date* all provided by Crossref, which have great potential for serving as proxies for the first publication date. However none of them is alone sufficient for such purpose. First, the coverage of published-print dates and published-online dates are not complete for all DOIs, especially for published-online date with only 32.46% of all the DOIs having this date recorded. Second, issued date is the integration of published-print date and published-online date by selecting the earlier one. This combination ensures the full coverage of issued date, and avoids the situation that old publications have a new online date because only the earliest one was selected as the issued date. Third, there are temporal issues related with these dates that are discussed below.

Table 1. Available Crossref publication dates

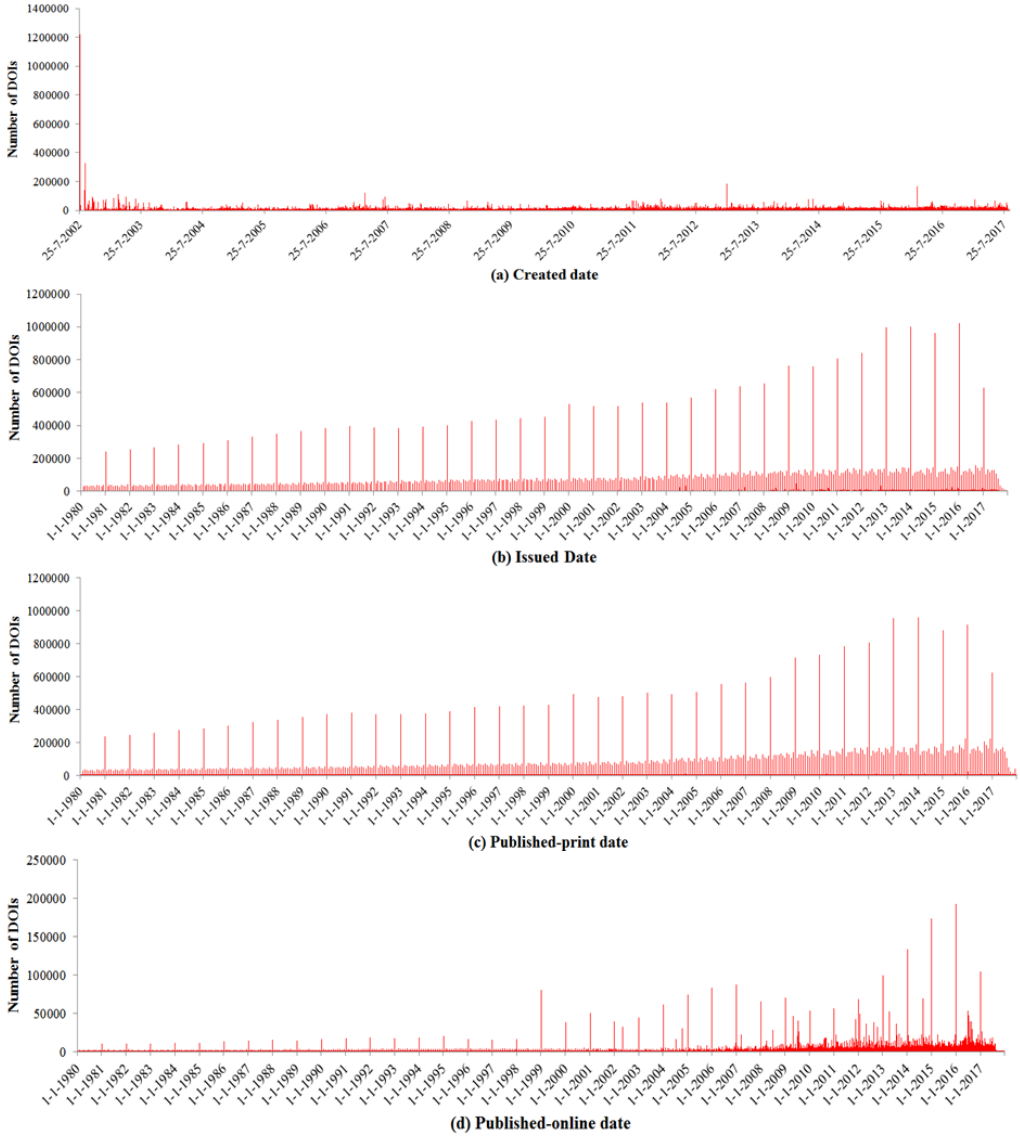
Date type	Description[4]	Number of DOIs	Coverage
Created date	Date on which the DOI was first registered.	68,148,933	100.00%
Issued date	Earliest of published-print and published-online.	68,148,933	100.00%
Published-print date	Date on which the work was published in print.	60,274,293	88.44%
Published-online date	Date on which the work was published online.	22,123,914	32.46%

The temporal distribution of these four dates discussed in Table 1 are shown in Figure 1. Issued date covers from 1980-01-01 onwards because both published-print date and published-online date start from that time. Created date begins only from 2002-07-25, while for DOIs before 2002-07-25 their created dates are later than the actual first official publication date. The pronounced peak at 2002-07-25 is mainly caused by many research outputs published before that date, but that were assigned with the created date of 2002-07-25. Considering these patterns, the created date cannot reliably be used for analytical purposes as the only publication date proxy,

[2] Date on which a publication was first formally accessible and available to the scientific community or the public.
[3] Date on which an altmetric event (e.g. tweets, blog mentions, news mentions) was posted online or published (for policy documents).
[4] https://github.com/CrossRef/rest-api-doc/blob/master/api_format.md#partial-date

particularly when publications before July 2002 are studied. Although the issued date covers a broader time span, there exists obvious peaks on every first day of a month, especially the first day of January. This is the result of the combined concentration of print dates (as the print dates of publications on journals or books) in the first day of a month, together with a large number of publications that have their online date on the first day of January of each year. In contrast, the created date is distributed more evenly and reasonably, because in the era of digital publishing, the first publication date could be any day of a month. Therefore, the date on which DOI was first registered can be regarded as a good proxy of the first formal appearance of a publication, as suggested by Haustein, Bowman & Costas (2015), but it is important to keep in mind the limitation of the creation dates of DOIs before July 2002.

Figure 1. Temporal distribution of (a) created date, (b) issued date, (c) published-print date and (d) published-online date



Based on all of the above, we decided to combine both issued date and created date as the best proxy for the first publication date of DOIs. As to publications with issued date earlier than 2002-

7-25, their issued date was selected as the proxy for the first publication date. For publications with issued date from 2002-7-25 onwards, the created date of DOIs was used to serve as the proxy since it should be relatively close to the actual first publication date of DOIs.

Altmetric.com data sources with posted on date

Table 2 presents 13 data sources with posted on date information tracked by Altmetric.com together with the date when they started their coverage[5]. Until October 2017, there are 8,157,487 Altmetric IDs (account for 99.90%) have at least one record from these data sources. In order to match with Crossref publication date through DOIs, 6,221,670 of which have DOIs are selected. However, among these Altmetric IDs with DOIs, there exists 79,761 Altmetric IDs that have preprint version (i.e. with arXiv IDs). The existence of preprint version makes research outputs available to social media before they are formally published, which may lead to the altmetric post date to be earlier than the publication date. Therefore, Altmetric IDs with arXiv IDs are excluded and the remaining 6,141,909 Altmetric IDs are matched with Crossref publication date. Finally 5,779,191 Altmetric IDs have DOIs recorded by Crossref until August 2017.

Table 2. Altmetric.com data sources with posted on date

Data source	Coverage began*
Blogs	Oct 2011
News	Oct 2011 & Dec 2015
Policy documents	Jan 2013
Reddit	Oct 2011
Twitter	Oct 2011
Facebook	Oct 2011
Google+	Oct 2011
Stack Overflow (Q&A)	Oct 2011
Faculty of 1000 Prime (F1000)	May 2013
Youtube	Apr 2013
Post-publication peer reviews	Mar 2013
Wikipedia	Jan 2015
Open Syllabus (Syllabi)	Sept 2016

*Altmetric.com has stopped collecting data from CiteULike, Sina Weibo, LinkedIn, and Pinterest. Mendeley and CiteULike, two online reference managers, lack proper post date information. Therefore, these data sources have not been included in this study.

Validity of the publication date calculation: Altmetric.com "first seen date" as benchmark

As mentioned above, except for the influence of preprint version, the first publication date of a publication should be expected to be earlier than its altmetric first seen date[6], as in theory an altmetric post cannot mention a publication before it exists. Consequently, the first seen date of all Altmetric IDs among 13 Altmetric.com data sources were aggregated to serve as the benchmark to examine whether the first publication date is reliable or not. After comparison, there are 389,818 papers (6.75%) with altmetric first seen date earlier than the first publication date. The possible reasons for the existence of these unreliable cases are the following:

[5] https://help.altmetric.com/support/solutions/articles/6000136884-when-did-altmetric-start-tracking-attention-to-each-attention-source-

[6] Date on which Altmetric.com captures the first event for a paper. Recorded for 99.9% of all the records in Altmetric.com

1. Crossref "created date" and "issued date" may contain errors and not always accurately reflecting the first publication date.
2. Publication dates may be updated by publishers due to different reasons (e.g. publisher mergers).

These Altmetric IDs with a first seen date before their best publication were excluded. As a result, all of the altmetric posts about these 5,389,373 Altmetric IDs were analysed in our study.

Results and Discussion

Coverage and temporal distribution of various altmetric posts

For 5,389,373 Altmetric IDs, all of their posts from 13 Altmetric.com data sources were extracted and analysed. Figure 2 shows the coverage of posts of each data source and the temporal distribution of altmetric posts among 8 time windows. Twitter posts have the highest coverage of Altmetric IDs, 71.40% of all Altmetric IDs had been mentioned at least once on Twitter. In contrast, the other 12 data sources show a much lower coverage. Facebook (19.10%), Policy documents (11.09%), Wikipedia (9.98%), Blog (9.13%), and News (7.78%) are relatively active in disseminating research outputs compared to some data sources with very low coverage, such as Syllabi (0.16%), Q&A (0.27%), and Youtube (0.64%). Data sources with high coverage also accumulated a large number of total posts. As Altmetric.com started to collect most data sources from 2011 onwards, we divided the posted on date into 8 periods by year to show the temporal distribution. The altmetric posts of most data sources distributed from 2011 to 2017 and increased over time. However, Altmetric.com also hold historic data for some certain data sources, for example, over half of F1000 posts (52.87%) and Policy documents citations (50.23%) happened before 2011, Wikipedia (29.09%) and Blog (9.88%) also have substantial shares of posts before 2011. Therefore, in our altmetric posts accumulation pattern analysis, all of the altmetric posts accrued until data collection date (October 2017) were taken into account without time restrictions to avoid losing historic information.

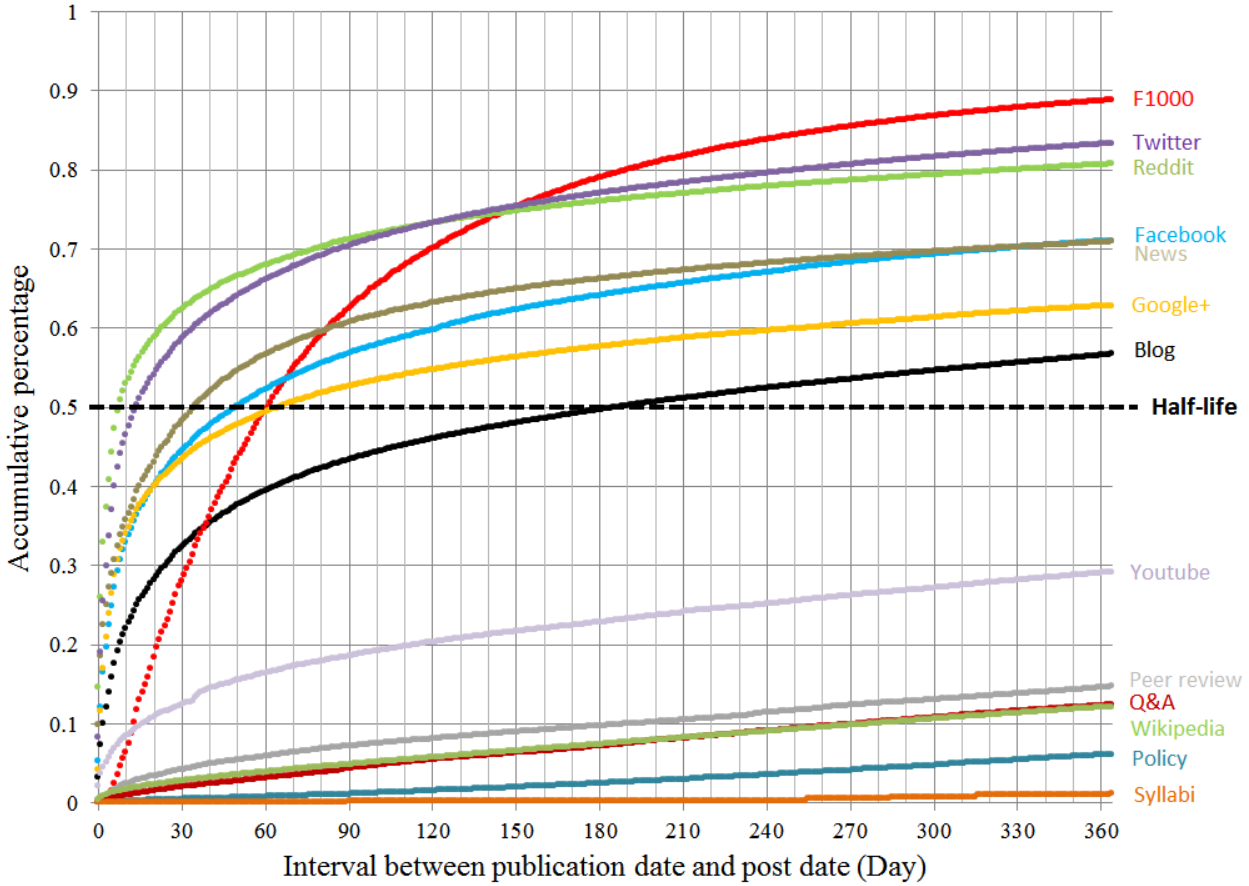
Figure 2. Coverage and temporal distribution of 13 kinds of altmetric posts

Data sources	Number of unique Altmetric IDs with posts	Coverage	Total number of posts	Posts before 2011	Posts in 2011	Posts in 2012	Posts in 2013	Posts in 2014	Posts in 2015	Posts in 2016	Posts in 2017
Blog	491,819	9.13%	835,494	9.88%	6.14%	7.92%	11.51%	14.61%	17.17%	18.81%	13.97%
News	419,044	7.78%	1,842,043	0.09%	0.28%	0.99%	5.58%	9.84%	14.58%	37.09%	31.54%
Policy	597,927	11.09%	928,372	50.23%	6.27%	6.48%	6.74%	7.01%	8.19%	10.11%	4.96%
Reddit	57,155	1.06%	80,028	2.36%	3.27%	7.14%	11.89%	11.77%	23.23%	22.76%	17.57%
Twitter	3,847,893	71.40%	27,279,926	0.00%	1.35%	5.50%	9.36%	13.61%	19.66%	24.36%	26.14%
Facebook	1,029,295	19.10%	2,587,825	0.12%	1.21%	6.23%	13.04%	19.06%	24.41%	19.41%	16.51%
Google+	151,139	2.80%	416,079	0.00%	2.17%	6.81%	12.91%	16.69%	21.05%	22.66%	17.70%
Q&A	14,592	0.27%	15,621	4.53%	12.20%	13.82%	11.55%	16.98%	20.07%	16.03%	4.83%
F1000	128,533	2.38%	163,567	52.87%	9.08%	8.51%	8.77%	5.89%	4.22%	5.96%	4.70%
Youtube	34,329	0.64%	53,038	4.76%	3.24%	6.11%	10.16%	16.42%	9.55%	17.77%	31.98%
Peer review	63,839	1.18%	94,122	0.01%	0.00%	0.00%	0.27%	11.31%	14.31%	66.90%	7.19%
Wikipedia	538,066	9.98%	762,783	29.09%	6.13%	7.52%	7.70%	10.01%	10.92%	15.21%	13.41%
Syllabi	8,655	0.16%	72,013	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%

Altmetric posts accumulation pattern

The intervals between publication dates and altmetric post dates were calculated for each data source. Thus we can investigate the altmetric posts accumulation pattern at the day time interval. Figure 3 shows the different posts accumulation patterns of the 13 data sources within one year time interval (365 days) after publication. Data sources show evidently different posts accumulation patterns. Posts to newly published research outputs on some data sources accumulated really fast, such as Reddit and Twitter, over half of posts of them accrued in the first month (31 days) after research outputs were published and over 80% of their posts happened within a year (365 days). Followed by News, Facebook, and Google+, the overall immediacy of them within the first few days after publication is relatively fast as well. By contrast, posts of Syllabi, Policy, Wikipedia, Q&A, and Peer review show different accumulation patterns. They are quite slow, as only about 1.1% of Syllabi posts, 6.1% of Policy documents posts, 12.1% of Wikipedia posts, 12.3% of Q&A posts, and 14.7% of Peer review posts accumulated within one year, most posts of these data sources happened more than a year after publication. Among these data sources, F1000 is unique. In the first month after research outputs were published, the accumulation of F1000 posts is not very fast, but it speeded up over time, with more than 88.7% of F1000 posts accrued within the first year.

Figure 3. Altmetric posts accumulation patterns of 13 data sources



Altmetric post half-life

To explore the immediacy of Altmetric posts to newly published research outputs, we defined *altmetric post half-life* as the number of days after research outputs were published that an

altmetric data source accumulates over half of posts to those publications. The dashed line at accumulative percentage of 50% in Figure 3 indicates the altmetric post half-life and Table 3 lists the altmetric post half-life of 13 data sources by ranking. Reddit ranks the first based on its altmetric post half-life of 9 days, followed by Twitter (15 days), News (36 days), and Facebook (51 days). These data sources are quite fast in disseminating newly published research outputs, they demonstrate that speed is one of the most important properties of altmetrics (Wouters & Costas, 2012; Bornmann, 2014). However, as to Peer review, Q&A, Wikipedia, and Policy documents, they spent over 2,000 days to accumulate over half of posts, they pay more attention to publications with older publication time. Especially for Syllabi, which mainly focuses on books, its altmetric post half-life is very long. These Altmetric.com data sources show remarkable time delay similar as those of citations (Schloegl & Gorraiz, 2010).

Table 3. Altmetric posts half-life of 13 data sources

Rank	Data source	Half-life (day)
1	Reddit	9
2	Twitter	15
3	News	36
4	Facebook	51
5	F1000	63
6	Google+	67
7	Blogs	188
8	Youtube	1,249
9	Peer review	2,053
10	Q&A	2,190
11	Wikipedia	2,216
12	Policy documents	2,254
13	Syllabi	5,669

Preliminary conclusions

In this study *issued date* and *created date* of DOIs provided by Crossref were introduced as the proxy for the first publication date, so that the altmetric posts accumulation pattern analysis could be advanced to the day level. As a consequence, this study provides insights on how different Altmetric.com data sources accumulated posts over time after publication. Various Altmetric.com data sources vary in their post accumulation patterns. Posts of Reddit, Twitter, News, Facebook accrued really fast within the first few days after research outputs were published, these data sources also hold short altmetric posts half-life due to their "speed". While there are also some Altmetric.com data sources exhibited a quite long altmetric posts half-life, such as Syllabi, Policy documents, Wikipedia, Q&A, and Peer review. Their posts were not so concentrated on new publications so that the posts accumulation after publication is slow too. Thus, the property of speed is not owned by all of Altmetric.com data sources, existing a relevant differentiation between the *fast sources* (e.g. Reddit, Twitter, News) and the *slow sources* (e.g. Syllabi, Policy documents, Wikipedia), which may also have implications for their analytical uses and applications.

The main limitation of this study lies in the precision of Crossref "created date" and "issued date" as proxy for the first publication of research outputs. Although altmetric first seen date was used as the benchmark to exclude some unreliable data, Crossref cannot be seen as an absolutely

precise proxy for publication dates. There might still be a small distance between the date on which DOI was created and research output was actually made publicly available, which could result in some negative influence on our results. Future research will focus on this issue as well as on the study of advanced time-based analytics of altmetric data sources.

Acknowledgements
This research is partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP), and Zhichao Fang is partially supported by the China Scholarship Council (CSC).

References

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, 8(4), 895-903.

Haustein, S., Bowman, T. D., & Costas, R. (2015). When is an article actually published? An analysis of online availability, publication, and indexation dates. *Proceeding of the 15th International Conference on Scientometrics and Informetrics (ISSI)*, (pp. 1170-1179), 29 Jun-4 July 2015, Istanbul, Turkey. https://arxiv.org/abs/1505.00796

Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the Association for Information Science and Technology*, 56(10), 1088-1097.

Ortega, J. L. (2018). The life cycle of altmetric impact: A longitudinal study of six metrics from PlumX. *Journal of Informetrics*, 12(3), 579-589.

Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: the case of oncology journals. *Scientometrics*, 82(3), 567-580.

Schlögl, C., Gorraiz, J., Gumpenberger, C., Jack, K., & Kraker, P. (2014). Comparison of downloads, citations and readership data for two information systems journals. *Scientometrics*, 101(2), 1113-1128.

Wang, X., Fang, Z., & Guo, X. (2016). Tracking the digital footprints to scholarly articles from social media. *Scientometrics*, 109(2), 1365-1376.

Wouters, P., & Costas, R. (2012). Users, narcissism and control-tracking the impact of scholarly publications in the 21st century. Utrecht: SURFfoundation. Retrieved from http://research-acumen.eu/wp-content/uploads/Users-narcissism-and-control.pdf