Full paper

STI 2022 Conference Proceedings

Proceedings of the 26th International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Proceeding Editors

Nicolas Robinson-Garcia
Daniel Torres-Salinas
Wenceslao Arroyo-Machado

Citation: Arroyo-Machado, W., Díaz-Faes, A. A., & Costas, R. (2022). New insights on social media metrics: examining the relationship between universities' academic reputation and Wikipedia attention. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti22159). https://doi.org/10.5281/zenodo.6962442

Collection: https://zenodo.org/communities/sti2022grx/

New insights on social media metrics: examining the relationship between universities' academic reputation and Wikipedia attention

Wenceslao Arroyo-Machado[*], Adrián A. Díaz-Faes[**] and Rodrigo Costas[***]

[*]*wences@ugr.es*
Department of Information and Communication Sciences, University of Granada, Spain

[**]*diazfaes@ingenio.upv.es*
INGENIO (CSIC-UPV), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain

[***]*rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands
DST-NRF SciSTIP, Stellenbosch University, South Africa

Introduction

Since the publication of the Altmetric Manifesto (Priem et al., 2010), interest in social media metrics has escalated as potential indicators of scientific performance (Haustein et al., 2014), with most analysis focused on translating scientometric models to the social media (Robinson-Garcia, van Leeuwen, and Ràfols, 2018). However, not only no significant relationships have been found between social media metrics and citations analysis[1], but the capability of the former to directly gauge social impact has been seriously contested (Sugimoto, Work, Larivière, & Haustein, 2017). A recent stream of literature has suggested that social media metrics should be conceptualized through the lens of science-society interactions rather than as direct sources of impact and recognition (Díaz-Faes, Bowman, & Costas, 2019; Wouters, Zahedi, & Costas, 2019), giving rise to new methodological proposals more focused on the science-society interactions that can be captured on altmetric data (Costas, de Rijcke, & Marres, 2020; Arroyo-Machado, Torres-Salinas, & Robinson-Garcia, 2021).

Among the social media platforms, the collaborative encyclopedia Wikipedia, with more than 250 million page views per day on its English version[2], stands out as one of the most appealing sources to explore the generation and circulation of knowledge across broader communities (Arroyo-Machado, Torres-Salinas, & Robinson-Garcia, 2021; Zagorova et al., 2022). Within the realm of altmetrics, a recurrent approach has been to examine whether universities' social attention captured through Wikipedia is somehow related to their academic reputation (Li, Li, & Li, 2019). Previous studies mostly have built Wikipedia university rankings by counting the number of page views and analyzing the network of pages connected by hyperlinks, which were

[1] The only notable exception being Mendeley readers, which show a moderate positive correlation with citations (Thelwall, 2018; Zahedi, Costas, & Wouters, 2014), which should not come as a surprise given its stronger academic focus.

[2] https://pageviews.wmcloud.org/siteviews/?platform=all-access&source=pageviews&agent=user&start=2022-04-09&end=2022-04-29&sites=en.wikipedia.org

then compared with scientometric based rankings. For instance, Katz and Rokach (2017) drew on page links, page views, and alumni's Wikipedia pages and compared the resulting ranking with the Academic Rating of World Universities (ARWU), Times Higher Education (THE) and Webometrics rankings. Coquidé, Lages, and Shepelyansky (2019) proposed a ranking based on Wikipedia's page links using social network analysis and contrasted it with ARWU; whereas Babkina et al. (2021) used Wikipedia's alumni pages to measure the impact of universities against the QS World University Rankings and ARWU.

The studies cited above point to potential relationships between social media attention and academic reputation; however, more fine-grained analyses are needed. It is necessary not only to consider a wider variety of Wikipedia and scientific performance indicators —previous works merely relied on the position within the rankings, overlooking the actual indicators—, but also to account for a number of factors that may influence the relationship between scientific reputation of universities and the social media attention they receive.

Objectives

This main objective of this paper is to explore whether the academic impact of universities may be related to their online social attention on Wikipedia. We address this objective as follows: first, we analyze the relationships between a diverse set of scientific performance indicators and Wikipedia metrics and, second, we explore how geographical and contextual factors may affect such relationships.

Methodology

Data processing

We adopted a two-step procedure to build our dataset. First, we drew on the CWTS Leiden Ranking 2021[3] to identify universities worldwide and their scientific performance measured with scientometric indicators. The Leiden Ranking has been noted as one of the most reliable and reproducible rankings to assess universities' academic reputation (Vernon, Balas, & Momani, 2018). The 1,225 universities included in the ranking were extracted together with the following key performance indicators for the period 2016-2019: *total number of publications*, number of publications in the top 10 percent most cited (hereafter referred to as *top publications*), *percentage of top publications*, number of *publications in collaboration* with another institution/s, and number of *open access publications*. These data were matched with GRID[4], which allowed us to retrieve each university's Wikipedia page (English edition) and its founding date[5]. Note that 10 universities from the Leiden Ranking could not be retrieved through GRID and were excluded from the analysis. Second, we used the Wikipedia Knowledge Graph dataset, a comprehensive and curated dataset of the English edition of Wikipedia (Arroyo-Machado, Torres-Salinas, & Costas, 2022). The English edition is the largest and most accessed edition. We extracted the following indicators from universities' Wikipedia pages until July 2021: *page views* (number of visits to the Wikipedia page in the past 3 months), *edits* (number of edits to the page in the past 3 months), *length* (in bytes), *page age* (number of years since the creation of the English version of the Wikipedia page of the university), and number of versions in other language editions (hereafter *language links*)[6]. Note that in the case of page views, data correspond to April 1 to June 31, 2021, whereas the number of language links was captured on April 27, 2022. Since GRID links to Wikipedia pages were

[3] https://www.leidenranking.com/ranking/2021/list (Accessed as of April 13, 2022)
[4] https://www.grid.ac (Accessed as of April 13, 2022)
[5] There are 108 universities that do not include this data
[6] Language links were retrieved through the Wikipedia API: https://www.mediawiki.org/wiki/API:Langlinks

not updated for some universities 141 redirect links were identified and amended thanks to the Wikipedia API.

Data analysis

Spearman's rank-order correlations were applied to measure the strength and direction of the associations between each pair of indicators (they are not normally distributed, and we expect a monotonic relationship). To examine factors that influence the social attention universities' Wikipedia pages receive, we employed Negative Binomial regression estimations, since our dependent variable is the total number of pages views. We have selected this indicator since it clearly reflects users' interest and attention to Wikipedia pages, by assuming that page relevance is reflected in the number of page views (Katz & Rokach, 2017). We discarded a Poisson model based on a significant degree of overdispersion (deviance goodness-of-fit = 22,000,000, $p < 0.000$). As explanatory variables that account for universities' academic reputation, we selected the percentage of top publications and the total number of publications. Since both are clearly related, the latter was divided into three categories (based on percentiles): low (ref. category), medium, and large. To account for potential cultural and geographical determinants of Wikipedia page views, we control whether universities are located in an Anglo-Saxon country[7] (dummy = 1 if universities is in an Anglo country) and the continent (Africa, Asia, Europe, North America, South America —ref. category—, Oceania). Data processing and analysis were conducted through Python and R. The scripts and resulting data can be found on GitHub (doi:10.5281/zenodo.6508254).

Results

Comparison between scientific performance and social attention indicators

Table 1 presents, grouped by country, the average scientific performance and Wikipedia indicators for the 1,215 universities finally examined from the Leiden Ranking. When comparing them, we find notable differences. While universities in the Netherlands show the highest average values for academic impact indicators, universities in the United States and the United Kingdom offer the highest values for page views on their Wikipedia pages. However, in the other indicators of attention on Wikipedia we find differences, for example, the pages of universities in the Netherlands and Sweden have more versions in other language editions, with the pages of the latter being the oldest created and those of France the most recent.

[7] We labelled as Anglo-Saxon the main core of English-speaking countries: Australia, Canada, Ireland, New Zealand, United Kingdom, and United States (https://en.wikipedia.org/wiki/Anglosphere)

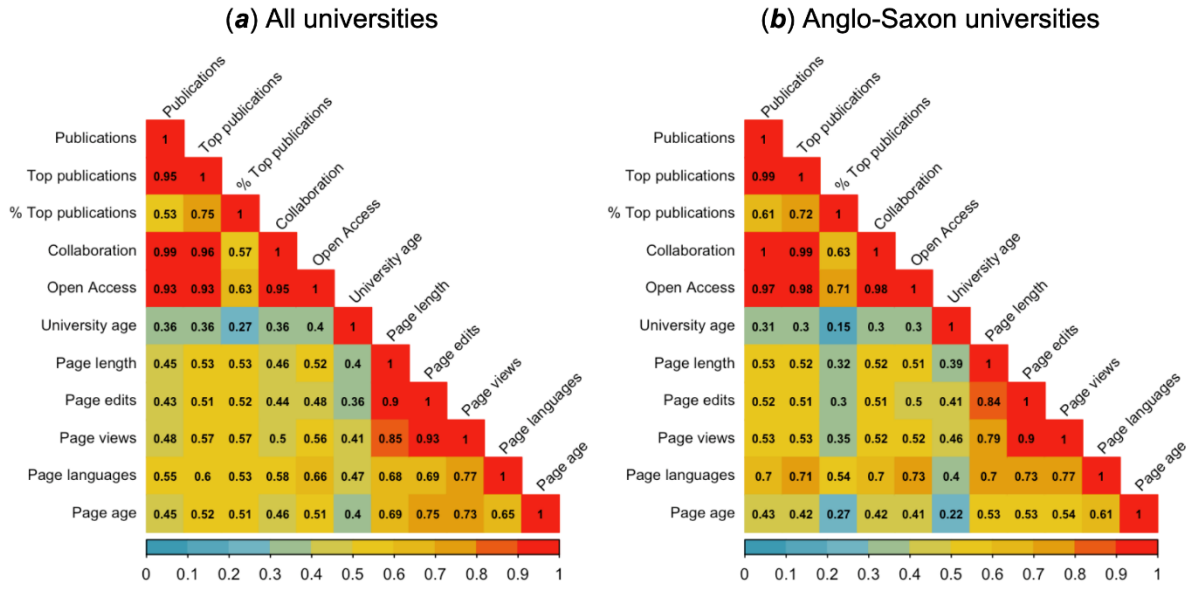
Table 1. Average values of the academic impact and Wikipedia social attention indicators for Leiden Ranking universities by country.

	Total universities	Average university age	Leiden Ranking					Wikipedia				
			Avg. publications	Avg. top publications	Avg. % top publications	Avg. collaboration	Avg. open access	Avg. length	Avg. page edits	Avg. page views	Avg. language links	Avg. page age
China	216	88.86	8023.71	917.93	10.51	6019.07	2904.72	14094.83	263.13	2974.02	7.16	14.99
United States	200	147.25	10694.89	1796.38	14.86	8847.54	6722.96	93359.49	3174.71	60016.35	30.12	18.23
United Kingdom	61	162.66	9448.30	1679.28	16.14	8141.03	8049.08	77672.08	2372.92	40454.49	36.41	18.21
Germany	54	213.44	8322.22	1217.31	13.68	7074.72	4768.24	27439.41	466.98	12613.96	35.26	16.69
Japan	54	103.56	5765.85	494.61	7.58	4704.39	3055.39	17885.28	237.04	4550.37	15.35	15.52
South Korea	46	87.16	6492.70	532.02	7.69	5070.63	2794.59	18167.41	461.61	9687.89	11.50	15.43
Italy	42	330.25	7596.41	980.45	12.81	6558.95	4115.83	16770.43	295.83	6682.62	24.48	15.74
Spain	42	240.73	5305.10	608.90	10.9	4433.10	3014.40	14007.88	232.95	3788.60	20.60	15.83
India	38	74.10	3349.71	278.18	8.35	2120.24	835.63	30989.74	1578.24	32343.34	14.00	16.32
Iran	36	60.47	3797.61	333.08	8.68	2711.19	804.33	17601.14	325.92	2060.89	7.06	15.36
Australia	32	74.30	10643.63	1631.78	14.76	9185.06	5537.78	49965.62	1408.22	14721.84	23.56	18.09
Brazil	31	71.11	6191.32	501.06	7.62	5164.90	2859.94	19383.32	232.97	1555.74	16.77	15.06
Poland	31	117.20	3139.29	244.06	7.37	2266.48	1681.06	14299.81	222.97	2885.65	17.29	16.26
Turkey	31	81.20	2722.29	190.10	6.68	1979.42	976.65	15363.48	317.68	4403.68	13.81	15.42
Canada	30	120.39	10716.37	1538.93	12.97	8591.17	5343.33	78472.07	2177.90	32369.90	29.80	18.80
France	27	258.61	9409.67	1323.11	13.47	8747.48	6012.78	17679.30	245.07	5604.70	24.26	13.22
Taiwan	21	75.60	5128.38	408.19	7.43	4236.33	2341.86	15869.05	380.10	3217.10	12.10	16.05
Netherlands	13	189.1	15213.77	2604.08	16.58	13297.08	10758.77	39058.46	740.77	14724.15	37.92	18.15
Austria	12	308.18	4813.42	685.83	13.78	4086.50	3120.33	13096.42	233.67	5396.58	23.75	16.92
Sweden	12	162.67	11142.75	1651.75	14.29	9594.92	7303.50	25034.17	579.92	10510.75	37.42	19.08
Universities < 10	186	138.96	6418.89	807.11	11.04	5346.75	3465.20	31176.05	623.92	10164.36	26.53	16.63

Note: countries with less than 10 universities in the Leiden Ranking are grouped under the category 'Universities < 10'.

The correlation matrixes in Figure 1 show that the strength of the association between university' scientific impact and Wikipedia attention is positive and from moderate to strong. When looking at all universities (Figure 1a), the correlations between Wikipedia page views and the number of top publications (r_s = 0.57), percentage of top publications (r_s = 0.57), and open access publications (r_s = 0.56) suggest that both constructs are clearly related. Similar values are observed for Wikipedia page length, number of edits or page age with regard to scientific performance. The highest value between both constructs occurs between the number of open access publications and the number of Wikipedia language links of a university (r_s = 0.66). Interestingly, when only Anglo-Saxon universities (Figure 1b) are examined, we find that, with the exception of language links, the correlations between Wikipedia indicators and top publications are reduced (especially with the percentage of top publications, r_s = 0.35), and in general there is an increase in the correlation of Wikipedia indicators with the total number of publications, suggesting a relationship between Wikipedia attention and the overall output of the universities in the Anglo-Saxon context. Worth mentioning is also the considerable increase of the correlation between the Wikipedia language links and the number of publications in collaboration (from r_s = 0.58 to r_s = 0.71) and with the number of top publications (from r_s = 0.60 to r_s = 0.71).

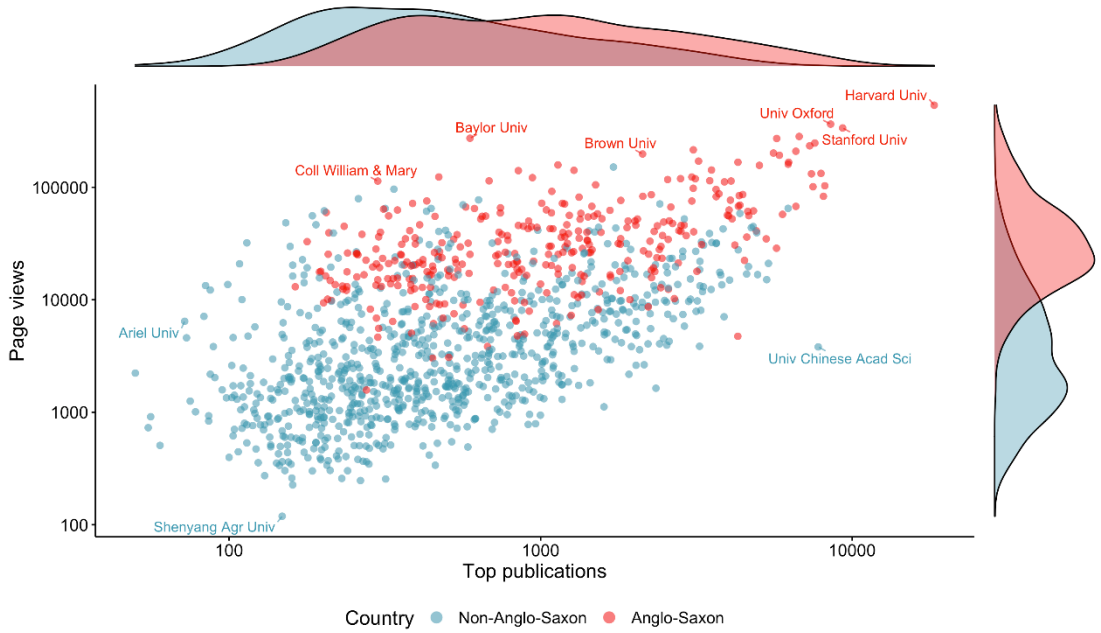
Figure 1. Spearman's Rho correlations between scientific performance and Wikipedia indicators: (a) all universities (b) only Anglo-Saxon universities.



Note: all p-values are below 0.001.

The effect of taking into account whether universities are located in Anglo-Saxon countries or not, is clearly apparent when plotting the two. Figure 2 shows that most universities from Anglo-Saxon countries gather more Wikipedia attention regardless of their academic reputation. Particularly outstanding is the performance of Harvard, Oxford and Stanford universities. For non-Anglo Saxon universities, although this relationship holds true, they clearly receive much less page views despite achieving a high number of highly cited publications, with the University of Chinese Academy of Sciences as the most extreme case. On the contrary, Baylor University. achieve a large number of visits with less than 1,000 highly cited publications.

Figure 2. Scatter and density plots (log scale) of top publications and Wikipedia page views of universities included in the Leiden Ranking.



The main results from the Negative binomial regressions are reported in Table 2. Model 1 tests the effect of the academic reputation on universities' Wikipedia pages views. We see that the estimated coefficient of both the percentage of top publications (β = 18.05, p = 0.000) and the number of publications (medium: β = 0.253, p = 0.002; large: β = 0.490, p = 0.000) have a positive and significant effect. Model 2 adds the two contextual variables that control for geographical determinants. Universities' size in terms of publications remains positive and highly significant (medium: β = 0.351, p = 0.000; large: β = 0.822, p = 0.000). In contrast, the percentage of top publications is only significant at 10% level (β = 5.157, p = 0.093), which evidence a fewer association with social attention when geographical factors come into play.

Table 2. Results for Negative Binomial Regression. Dependent variable: Number of universities Wikipedia pages views (N = 1,107).

	Model 1		Model 2	
	β / SE	P-value	β / SE	P-value
% Top publications	18.051 (1.032)	0.000	5.157 (3.073)	0.093
Number of publications				
Medium	0.253 (0.083)	0.002	0.351 (0.039)	0.000
Large	0.490 (0.095)	0.000	0.822 (0.137)	0.000
University age	0.001 (0.001)	0.000	0.002 (0.001)	0.005
Anglosphere			1.591 (0.133)	0.000
Geographical area				
Africa			1.809 (0.052)	0.000
Asia			1.191 (0.031)	0.000
Europe			0.672 (0.107)	0.000
North America			1.137 (0.063)	0.000
Oceania			-0.044 (0.075)	0.555
Constant	7.118 (0.115)	0.000	6.599 (0.226)	0.000
Nagelkerke R2	0.417			0.619

Notes: In Model 2, robust standard errors (SE) are clustered by the geographical area. P-values in bold font indicate p < 0.05.

Concluding remarks

In the context of novel developments in the altmetric realm, in which more interactive and network perspectives are set forth (see for example the "heterogeneous couplings" framework – Costas, de Rijcke & Marres, 2020) to study science-science interactions via social media and altmetric data, and also building on the idea that altmetrics and social media metrics do not need to be confined to the mere study of the impact of scientific publications on social media (see Haustein, Bowman & Costas, 2016), this paper illustrates the possibility and relevance of studying the academic-media relationships of research universities by exploring the relations between their overall scientometric indicators of academic performance (as captured in the Leiden Ranking) and their overall Wikipedia attention.

This paper explores how universities academic reputation as measured by scientometric indicators is to some extent related to social attention captured in Wikipedia, and tries to identify the main variables involved in this relationship. We find that the publication size and highly cited publications have a positive association with social attention on Wikipedia. However, when variables that account for geographical and cultural factors come into play, the latter's

influence is reduced significantly. This suggests that in the translation of academic reputation to the social media realm, different factors than just the academic performance of universities may also be playing important roles, including geographical, cultural, and linguistic aspects, which are usually not consider in most altmetric research. These results suggest important directions for future research, which should inevitably consider contextual factors to properly disentangle the relationship between academic reputation and Wikipedia social attention, including among others the Wikipedia pages of universities in languages other than English. Finally, it is also important to remark that the approach developed in this paper for Wikipedia could easily be also extrapolated to other social media platforms, like Twitter, Facebook or even more local platforms like WeChat, enabling more advanced studies of the relationships between the academic reputation and social media impact of scientific entities.

References

Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022). Wikipedia Knowledge Graph dataset. Zenodo. Retrieved March 23, 2022, from https://doi.org/10.5281/zenodo.6346900

Arroyo-Machado, W., Torres-Salinas, D., & Robinson-Garcia, N. (2021). Identifying and characterizing social media communities: A socio-semantic network approach to altmetrics. *Scientometrics*, *126*(11), 9267–9289.

Babkina, T. K., Goiko, V., Khomutenko, V., Palkin, R., Mundrievskaya, Y., Myagkov, M., Sukhareva, M., et al. (2021). Measuring University Impact: Wikipedia Approach. *2021 3rd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)* (pp. 625–632).

Coquidé, C., Lages, J., & Shepelyansky, D. L. (2019). World influence and interactions of universities from Wikipedia networks. *The European Physical Journal B*, *92*(1), 3.

Costas, R., de Rijcke, S., & Marres, N. (2020). "Heterogeneous couplings": Operationalizing network perspectives to study science-society interactions through social media metrics. *Journal of the Association for Information Science and Technology*, *72*(5), 595–610. John Wiley & Sons, Ltd.

Díaz-Faes, A. A., Bowman, T. D., & Costas, R. (2019). Towards a second generation of 'social media metrics': Characterizing Twitter communities of attention around science. *PLOS ONE*, *14*(5), e0216408.

Haustein, S., Bowman, T. D., & Costas, R. (2016). Interpreting 'Altmetrics': Viewing Acts on Social Media through the Lens of Citation and Social Theories. In C. R. Sugimoto (Ed.), *Theories of Informetrics and Scholarly Communication* (pp. 372–406). Berlin, Boston: De Gruyter Saur.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics*, *101*(2), 1145–1163.

Katz, G., & Rokach, L. (2017). Wikiometrics: A Wikipedia based ranking system. *World Wide Web*, *20*(6), 1153–1177.

Li, Z., Li, C., & Li, X. (2019). Mining the rank of universities with Wikipedia. *Science China Information Sciences*, *62*(10), 209202.

Robinson-Garcia, N., van Leeuwen, T. N., & Ràfols, I. (2018). Using altmetrics for contextualised mapping of societal impact: From hits to networks. *Science and Public Policy*, *45*(6), 815–826.

Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. *Scientometrics*, *115*(3), 1231–1240.

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, *68*(9), 2037–2062. John Wiley & Sons, Ltd.

Vernon, M. M., Balas, E. A., & Momani, S. (2018). Are university rankings useful to improve research? A systematic review. *PLOS ONE*, *13*(3), e0193762. Public Library of Science.

Wouters, P., Zahedi, Z., & Costas, R. (2019). Social Media Metrics for New Research Evaluation. In W. Glänzel, M. Henk F, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*, Springer Handbooks (pp. 687–713). Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-030-02511-3_26

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*, *101*(2), 1491–1513.

Zagorova, O., Ulloa, R., Weller, K., & Flöck, F. (2022). "I updated the <ref>": The evolution of references in the English Wikipedia and the implications for altmetrics. *Quantitative Science Studies*, *3*(1), 147–173.