

Amplifying the Impact of Open Access: Wikipedia and the Diffusion of Science

Misha Teplitskiy

Dept. of Sociology and
KnowledgeLab
University of Chicago
misha@uchicago.edu
(214) 574-9845
5735 South Ellis Avenue
Chicago, Illinois 60637

Grace Lu

KnowledgeLab
University of Chicago
gracelu@uchicago.edu
(773) 834-4787
5735 South Ellis Avenue
Chicago, Illinois 60637

Eamon Duede

Computation Institute and
KnowledgeLab
University of Chicago
eduede@uchicago.edu
(773) 834-4787
5735 South Ellis Avenue
Chicago, Illinois 60637

Abstract

With the rise of Wikipedia as a first-stop source for scientific knowledge, it is important to compare its representation of that knowledge to that of the academic literature. This article approaches such a comparison through academic references made within the world's 50 largest Wikipedias. Previous studies have raised concerns that Wikipedia editors may simply use the most easily accessible academic sources rather than sources of the highest quality or highest status. We test this claim by identifying the 250 most heavily used journals in each of 26 research fields (4,721 journals, 19.4MM articles in total) indexed by the *Scopus* database, and modeling whether topic, academic status, and accessibility make articles from these journals more or less likely to be referenced on Wikipedia. We find that, controlling for field and impact factor, the odds that an open access journal is referenced on the English Wikipedia are 47% higher compared to closed access journals. Moreover, in most of the world's Wikipedias, a journal's high status (impact factor) and accessibility (open access policy) both greatly increase the probability of referencing. Among the implications of this study is that the chief effect of open access policies may be to significantly amplify the diffusion of science, through an intermediary like Wikipedia, to a broad public audience.

Word count: 7894

Introduction

Wikipedia, one of the most visited websites in the world¹, has become a destination for information of all kinds, including information about science (Laurent & Vickers, 2009; Okoli, Mehdi, Mesgari, Nielsen, & Lanamäki, 2014; Spoerri, 2007). Given that so many people rely on Wikipedia for scientific information, it is important to ask whether and to what extent Wikipedia's coverage of science is a balanced and high quality representation of the knowledge within the academic literature. One lens through which to ask this question is through references used in Wikipedia articles. Wikipedia requires all claims to be substantiated by reliable references², but what, in practice, are "reliable references"?

An intuitive approach is to examine whether the sources Wikipedia editors use correspond to the sources scientists value most. In particular, within the scientific literature, a journal's status is often associated, albeit problematically (Seglen, 1997), with its impact factor. If status within the academic literature is taken as a "gold standard," Wikipedia's failure to cite high impact journals of certain fields would constitute a failure of coverage (Samoilenko & Yasseri, 2014), while a high correlation between journals' impact factors and citations in Wikipedia would indicate that Wikipedia does indeed use reputable sources (P. Evans & Krauthammer, 2011; Nielsen, 2007; Shuai, Jiang, Liu, & Bollen, 2013).

Yet high impact journals tend to require expensive subscriptions (Björk & Solomon, 2012). The costs are, in fact, so prohibitive that even Harvard University has urged its faculty to "resign from publications that keep articles behind paywalls" because the library "can no longer afford the price hikes imposed by many large journal publishers" (Sample, 2012). Curiously, much of the discussion of open access focuses on the consequences of open access for the scientific community (Van Noorden, 2013). Yet, within wealthy countries, it is precisely the scientific community that is least affected by expensive subscriptions (Davis & Walters, 2011; J. A. Evans & Reimer, 2009). Indeed, recent experimental and cross-sectional evidence has not shown any lasting effects of open access *on the scientific literature* (Davis, 2011; Davis, Lewenstein, Simon, Booth, & Connolly, 2008; Gaulé & Maystre, 2011; Moed, 2007). The

seemingly minor consequence Victoria University economist John Haughton mentions in (Van Noorden, 2013) is “that [with a fully open literature] researchers would save [time] when trying to access or read papers that were no longer lodged behind paywalls,” may actually be the main one.

Contrary to their limited impact on the scientific literature, open access journals may have a tremendous impact on the diffusion of scientific knowledge *beyond* this literature. To date, this potential of open access has been a matter of speculation only (Trench, 2008). Previous research has found that open access articles are downloaded from publishers’ websites more often and by more people than closed access articles (Davis, 2010, 2011), but it is currently unknown whether open access affects the use of science by the *general public* (Davis & Walters, 2011). We hypothesize that Wikipedia, with more than 8.5 million page views *per hour*³, diffuses scientific knowledge to unprecedented distances and that diffusion of science through it may relate to accessibility in two ways. By referencing findings from paywall journals, Wikipedia distills and “bootlegs” these findings to the general public. It thereby undermines claims that open access democratizes science, since even paywall science becomes freely (albeit indirectly) accessible. On the other hand, Wikipedia editors may be unable to access expensive paywall journals, and consequently reference the easily accessible articles instead. For example, Luyt and Tan’s (Luyt & Tan, 2010) study found accessibility to drive the selection of references in a sample of Wikipedia’s history articles. In this case Wikipedia “amplifies” open access science by broadcasting its (already freely accessible) findings to millions. This “amplifier” effect, a chief motivation for open access but hitherto unverified, may truly overshadow all previously identified effects of open access.

Correspondence between academic and Wikipedia statuses

This article tests both the bootlegger and amplifier hypotheses by evaluating which references Wikipedia editors around the world use and do not use. In particular we study the correspondence between journals’ status within the scientific community (impact factor) and their accessibility (open access policy) with their status within Wikipedia (percent of a journal’s articles referenced in Wikipedia). It is im-

portant to note that an observed correspondence may be affected by a variety of mechanisms besides the aforementioned accessibility. First, the status ordering of academic journals as measured with impact factors may have only a tenuous correspondence with the importance and notability -- considerations of special relevance to Wikipedia⁴ -- of the published research. Citations, and therefore impact factors, are in part a function of the research field (Seglen, 1997), and may be affected by factors as circumstantial as whether a paper's title contains a colon (Jamali & Nikzad, 2011; Seglen, 1997). Second, the academic status ordering results from the objectives of millions of scientists and institutions, and may be irrelevant to the unique objectives of Wikipedia. Wikipedia's key objective is to serve as an encyclopedia, not a medium through which scientists communicate original research⁵. Relative to the decentralization of the scientific literature, Wikipedia is governed by explicit, if flexible, policies and a hierarchical power structure (Butler, Joyce, & Pike, 2008; Shaw & Hill, 2014). Apart from a remark that review papers serve Wikipedia's objectives better than primary research articles, Wikipedia's policies generally pass no judgment over which items within the scientific literature constitute "the best" evidence in support of a claim⁶. Wikipedia's objectives and explicit, centrally accessible, policies differ from the decentralized decisions that produce the scientific literature status ordering and do not imply that the two status orderings should correspond. Indeed, if editors are not scientists themselves they need not even be aware that journal impact factors exist. On the other hand, despite the well-worn caveats, prestigious, high-impact journals may publish findings that are more important to both academics and Wikipedia's audience. In fact, a Wikipedia editor's *expectation* that the truly important research resides within high-impact journals may be enough to predispose them to want reference such journals. Second, little is known about editors of science-related articles (West, Weber, & Castillo, 2012); they may be professional scientists with access to these high-impact journals, resulting in both the *motivation* and *opportunity* to reference them.

Previous research

Wikipedia references and academic status

The first large-scale study of Wikipedia's scientific references was performed by Finn Arup Nielsen (2007). Nielsen found that the number of Wikipedia references to the top 160 journals, extracted using the *cite-journal* template, correlated modestly with the journal's *Journal Citation Reports* impact factor. This implication that Wikipedia preferentially cites high impact journals is delicate in part because the data used in the study included only a subset of journals with references that appear in Wikipedia, not journals that were and *were not* referenced. It is possible, albeit unlikely, that an even larger number of prestigious journals, made invisible by the methodology, are never referenced on Wikipedia at all, weakening the correlation to an unknown degree, or that the referenced journals are simply those that publish the most articles (see Nielsen 2007: Fig. 1). Shuai, Jiang, Liu, and Bollen (2013) also found modest correlations when they investigated a possible correspondence between the academic rank of computer science papers, authors, and topics and their Wikipedia rank. Adopting the perspective of alt-metrics, Evans and Krauthammer (P. Evans & Krauthammer, 2011) examined the use of Wikipedia as an alternative measure of the scholarly impact of biomedical research. The authors correlated scholarly metrics of biomedical articles, journals, and topics with Wikipedia citations and, in contrast to other studies, included in some of their analyses a random sample of journals never referenced on Wikipedia. The authors also recorded a journal's open access policy but, unfortunately, do not appear to have used this information in analyses.

Open access and the Web

Despite the promise of open access for disseminating scientific information to the world at large (Trench, 2008), the rather voluminous literature on open access policies has focused primarily on their impact on the academic literature⁷. By all accounts the effects on scientists from developed nations are relatively minor (Davis & Walters, 2011; J. A. Evans & Reimer, 2009). Yet, more and more of the world

at large seeks scientific information on the Web. For instance, as early as 1999 a full 20% of American adults sought medical and science information online (Miller, 2001). What's more, one who actively seeks such information within the academic literature will quickly discover that, despite the paywalls, the most important and impactful research articles can be easily located, unadulterated and for free in not so distant corners of the internet (Wren, 2005). This is to say nothing of the fact that science may also be disseminated through distillation of its findings into venues like Wikipedia or science centric websites and blogs so that, here too, the impact of open access may be limited. While full texts of the most impactful literature are, at least nominally, behind a paywall (Björk & Solomon, 2012), do Wikipedia's editors consult these texts? If they cite them in Wikipedia, have they consulted the full texts beyond a freely available abstract before referencing? If the academic literature is any guide, referenced material is sometimes consulted rather carelessly (Broadus, 1983; Rekdal, 2014). In short, the current understanding of the relationship between open access and the general public in the literature is limited at best (Davis & Walters, 2011).

Shortcomings and our contribution

In addition to the role of accessibility, a number of substantive and methodological shortcomings remain. First, it is unclear if professional scientists edit Wikipedia's science articles. As we will show below, a preponderance of paywall references would suggest, albeit indirectly, this to be the case. The scant existing evidence indicates that science articles are edited by people with general expertise, relative to the more narrow experts of popular culture articles (West et al., 2012). Second, all of the aforementioned studies have examined the English-language Wikipedia only. Non-English Wikipedias have been noticeably neglected by the research community (Mesgari, Okoli, Mehdi, Nielsen, & Lanamäki, 2015; Schroeder & Taylor, 2015). It is unclear whether any of the findings extend to the millions of articles in non-English Wikipedias.

We find methodological shortcomings in the broader literature. For instance, most previous studies have completely ignored the articles that are never referenced on Wikipedia, thus sampling on the dependent variable. The only notable exception, (P. Evans & Krauthammer, 2011), treated the unreferenced articles outside the main analytic framework. While the framework treated (referenced) articles or journals as the unit of analysis, the unreferenced articles and journals were treated as a group. The central tension underlying methodological choices of how to deal with unreferenced items stems from the belief that some of these items are candidates for being referenced, but some are not, perhaps because they address topics irrelevant for Wikipedia. Authors have resisted incorporating all articles in their analyses and struggled with how to effectively identify which unreferenced work belongs to which group.

This study extends existing work in four chief ways. First, it models the role of accessibility (open access) on referencing. Second, it covers *all* major research areas of science by observing rates at which Wikipedia references nearly 5,000 journals, accounting for nearly 20 million articles. Third, it performs these analyses not only for the English-language Wikipedia, but also Wikipedias of 49 other languages. Lastly, it does all of this within a unified analytic framework, which does not ignore unreferenced journals.

Data and Methods

Empirical Strategy

Certainly not all findings published in the academic literature belong on Wikipedia. Only small subsets of published findings are important and notable enough to be referenced in Wikipedia. Ideally, studies of how Wikipedia editors reference sources should explain which items in this smaller subset are and are not referenced. Nevertheless, previous studies have struggled to distinguish the candidate articles that are at risk for reference from those that do not belong on Wikipedia. Yet, to model referencing decisions

with *all* articles – including the dozens of millions of articles never referenced on Wikipedia – is likely to result in a model that predicts that no article will ever be referenced. Consequently most studies have voluntarily hobbled themselves by simply modeling only on the subset of referenced articles.

Here we propose a compromise strategy based on “demonstrated demand.” The idea is simple: articles are at risk for reference if *other* articles on the same topic are referenced. Topical reference indicates that there is demand from Wikipedia editors for literature on the topic and that an article’s characteristics (e.g. accessibility, status) may determine which of the candidate articles an editor finds and references. Conversely, if articles on a given topic are never referenced, it is likely that Wikipedia editors do not “demand” literature on this topic, no matter the accessibility or status of the supply. Demonstrated demand exists at the level of topics and, like accessibility and status, we identify an article’s topic at the journal level. Demonstrable demand is also a language-dependent metric: some Wikipedias may lack editors with expertise or interest in, for example, dentistry, thereby consigning all dentistry journals to irrelevance with regards to referencing decisions (but not irrelevant for analysis of coverage, of course). To calculate demonstrated demand we identify for each journal its topical “neighbors” and assign demand of 0 if none of these neighbors are ever referenced in a particular Wikipedia.

Data sample

Journal data

Our analysis uses journal-level data from thousands of journals indexed by *Scopus*. Indexing over 21,000 peer-reviewed journals and with more than 2,800 classified as open access, *Scopus* is the world’s largest database of scientific literature⁸. We obtained information on the 250 highest-impact journals within each of the following 26 major subject areas⁹: *Agricultural Sciences, Arts and Humanities, Biochemistry and General Microbiology, Business Management and Accounting, Chemical Engineering, Chemistry, Computer Science, Decision Sciences, Earth and Planetary Sciences, Economics and Fi-*

nance, Energy Sciences, Engineering, Environmental Sciences, Immunology and Microbiology, Materials Sciences, Mathematics, Medicine, Neurosciences, Nursing, Pharmacology, Physics, Psychology, Social Science, Veterinary Science, Dental, Health Professions. Assignment of journals to these broad subject areas is not exclusive; many journals fall into more than one category. As a result of cross listing, the list of candidate journals was less than 6500. The final data consisted of 4721 unique journals, 335 (7.1%) of which are categorized by the Directory of Open Access Journals as “open access.”

Journals were also categorized more narrowly using the more than 300 “All Science Journal Classification” (ASJC) subject codes¹⁰, e.g. Animal Science and Zoology, Biophysics, etc. These narrow codes were used to identify journals that address similar topics and thus indicate whether the journal is at risk for reference *vis-à-vis* demonstrated demand. Journals with at least one narrow subject code in common were considered “neighbors” and if at least one of these neighboring journals has been referenced the original “ego” journal was considered to be at risk for reference as well. Journals with no demonstrable demand were excluded from analysis. As an example, consider the journal *Science*. It is listed under (ASJC) subject code 1000 – general science. Other journals with this code – the “neighbors” of *Science* – are *Nature*, *PNAS*, and *Language Awareness*. *Language Awareness* is cross-listed under 5 others subject codes.

Impact factor was measured by the 2013 SCImago Journal Rank (SJR) impact factor. SJR correlates highly with the more conventional impact factor but takes into account self-citations and the diverse prestige of citing journals (González-Pereira, Guerrero-Bote, & Moya-Anegón, 2010; Leydesdorff, 2009). Table 1 displays the 15 highest SCImago impact journals, calculated with citations data available up to 2013.

| Journal | Impact factor (SCImago2013) |
|--------------------------------------|-----------------------------|
| CA - A Cancer Journal for Clinicians | 45.894 |

| | |
|--|--------|
| Reviews of Modern Physics | 34.830 |
| Annual Review of Immunology | 32.612 |
| Cell | 28.272 |
| Annual Review of Biochemistry | 27.902 |
| Quarterly Journal of Economics | 25.168 |
| Nature Genetics | 24.052 |
| Nature Reviews Genetics | 23.813 |
| Nature Reviews Molecular Cell Biology | 23.593 |
| Chemical Reviews | 23.543 |
| Nature | 21.323 |
| Acta Crystallographica Section D: Biological Crystallography | 20.717 |
| Advances in Physics | 20.349 |
| Annual Review of Cell and Developmental Biology | 19.686 |
| Annual Review of Neuroscience | 19.662 |

Table 1. 15 highest-impact journals within Scopus according to SCImago impact factor (2013).

English Wikipedia data

References in the English Wikipedia were extracted from the 2014-11-15 database dump of all articles.

We parsed every page, and following (Nielsen, 2007) extracted all references that use Wikipedia’s *cite journal* template. This template is the convention followed by editors when they reference scientific sources in Wikipedia because it allows editors to easily include inline references that are then automatically rendered into an end-of-article bibliography¹¹. In all, there were 311,947 “cite-journal” tags in the English Wikipedia.

Non-English Wikipedia data

In addition to the English Wikipedia, we followed a similar procedure for obtaining and matching scientific references from the 49 largest non-English Wikipedias. Database dumps of these Wikipedias were downloaded on 2015-05-10. For each of these, we extracted lines containing “journal” or “doi”. Thus the process for obtaining scientific references in non-English Wikipedias did not take into account language-specific tags. Non-English Wikipedias may also reference domestic scientific journals that are not

indexed by *Scopus*. Thus, it is plausible that our approach of searching for “journal” or “doi” keywords and *Scopus* journals undercounts scientific references to non-English Wikipedias.

The English Wikipedia referenced by far the greatest number of unique articles. Figure 1 displays the number of unique articles referenced in other Wikipedias, sorted by size (total articles).

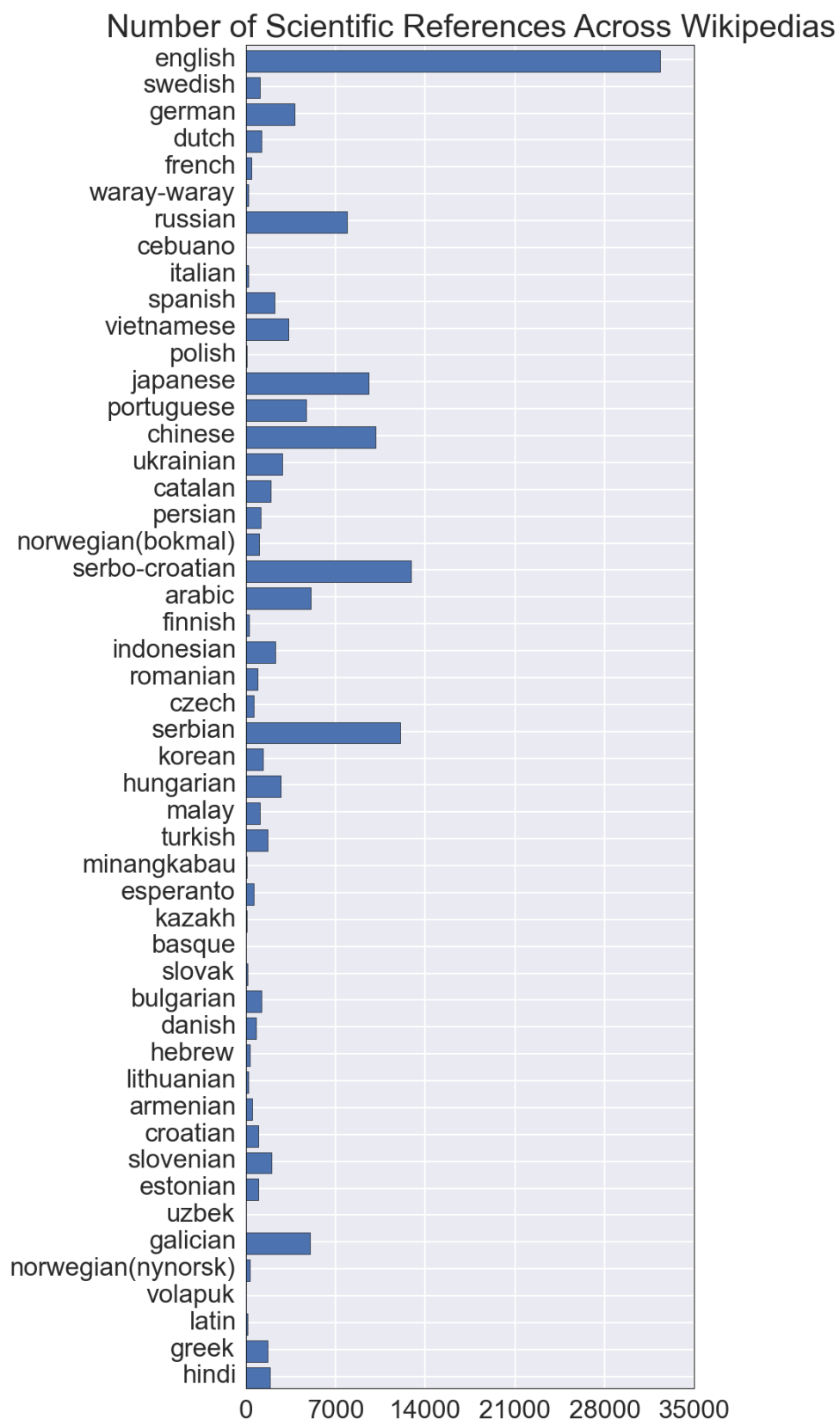


Figure 1. Number of unique scientific articles referenced on the 50 largest Wikipedias.

It is important to note that, given our audience as well as the skewed distribution of references, the analyses that we present in the Results section of this paper focus on the English Wikipedia. Nevertheless, the strategy of modelling referencing decisions only on journals with demonstrable demand adequately deals even with Wikipedias containing only a handful of scientific references.

Matching Scopus journals to Wikipedia references

We checked each of the referenced journal names on Wikipedia against a list of *Scopus*-indexed journal names and common *ISI* journal name abbreviations. Of the 311,947 *cite-journal* tags, 203,536 cited journals could be linked to journals indexed by *Scopus*. However, many of these references were non-unique, whereas our outcome of interest is whether articles from a journal are referenced on Wikipedia at all, not how many times. Therefore, to ensure that the counts for each journal included only unique articles, we distinguished articles by their DOIs and, if a given DOI was not available, we used the article's title. *Scopus'* coverage of the output of various journals varies widely; our counts included only those articles published within the years of *Scopus* coverage.

In the end we matched 32,361 unique articles (and 55,262 total references) to our subset of the top *Scopus* journals. 2,005 of the top *Scopus* journals are never referenced on the English Wikipedia. In most cases observed “journal names” that did not match to *Scopus* were not academic journals but newspapers and magazines. Table A1 in the Appendix displays the 20 most frequently referenced sources that we were unable to link to *Scopus*. The top 3 non-science sources are *Billboard*, *National Park Service*, and *Royal Gazette*. However, efforts to match Wikipedia references to *Scopus* were imperfect, and the list also includes a handful of academic journals, including *The Lancet*.

Journal vs. article level unit of analysis

We chose to take journals instead of individual articles as our unit of analysis for several reasons. First and most importantly, accessibility of articles, a focal point of this study, was measured at the journal

level by whether the journal is or is not open access. Switching to the article as the unit of analysis would have simply assigned the same value of accessibility to all articles from a particular journal. Second, while article-level citations are an attractive, finely grained metric, a journal's impact factor captures the same information albeit more coarsely. The general topic of any given article is also well captured by the host journal's *Scopus*-assigned topic(s). Lastly, the matching of Wikipedia journal title strings to *Scopus* required some manual matching and such matching is simply more pragmatic at the level of thousands of journals instead of hundreds of thousands of articles.

percent_cited and Other Variables

We present some of our results in terms of *percent_cited* -- the *percent of a journal's articles that are referenced on Wikipedia*. An equivalent interpretation of this journal-level metric is the *probability that a given article from a journal is referenced on Wikipedia*. Figure 2 illustrates the distribution (kde) of *percent_cited*.

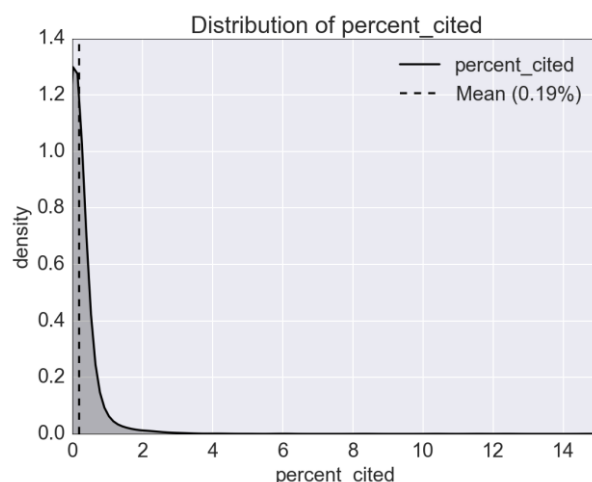


Figure 2. *Distribution (kde) of percent_cited of 4774 journals.*

As figure 2 demonstrates, the vast majority of journals that scientists use are referenced on the English Wikipedia very little: on average 0.19% of a journal's articles are referenced¹². Non-English Wikipedias reference substantially fewer articles (see Figure 1 above) and have a correspondingly low average *percent_cited*.

As mentioned above, journals' academic status was measured using (SCImago) impact factors. To limit the influence of the few journals with uncommonly high impact factors the impact factor variable was (natural) log-transformed when used in the models. Figure 3 displays the distribution of impact factor and log-impact factor; to aid visualization only journals with impact factor ≤ 15 are shown.

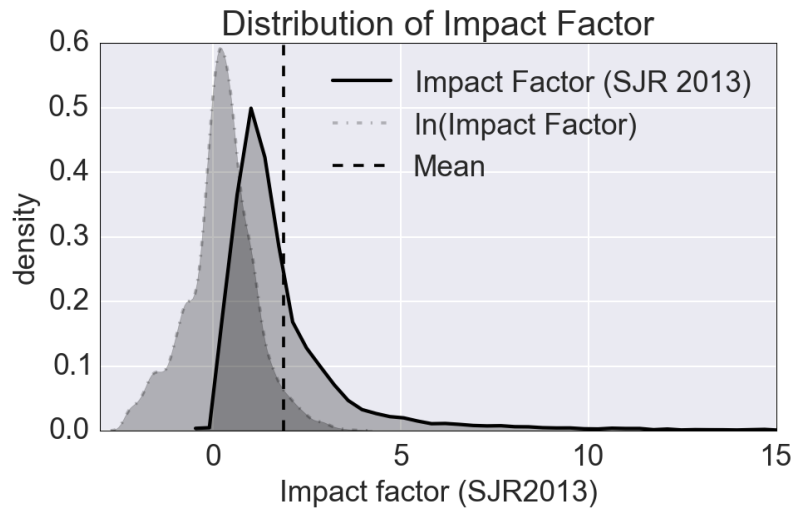


Figure 3. Distribution (kde) of impact factor and $\ln(\text{impact factor})$. To aid visualization, impact factor > 15 is not displayed.

We calculate demonstrated demand for a journal through its topical neighbors, which are defined as other journals that share at least one narrow (ASJC) subject code. Only 1 journal, *Prevenzione & assistenza dentale*, had no neighbors while the mean neighborhood size was 144.8. Figure 4 displays the distribution (kde) of neighborhood size.

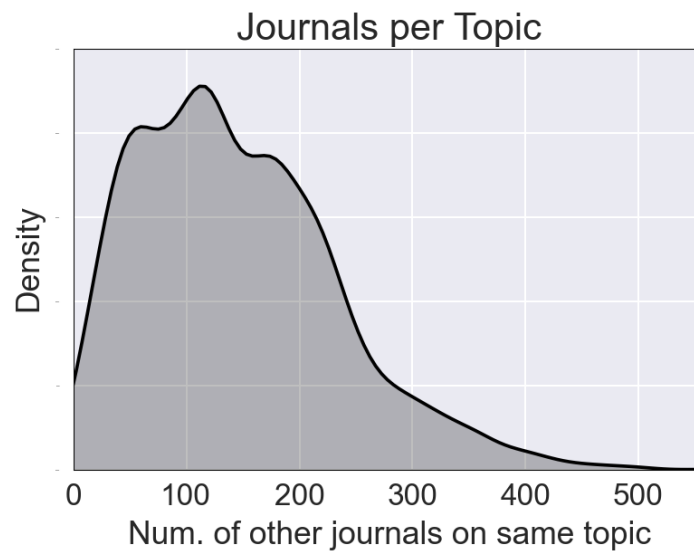


Figure 4. Distribution of the topical neighborhood sizes of journals. On average journals had 144.8 other journals that addressed the same topic(s).

Table 2 presents the summary statistics for key variables: *percent_cited*, *impact factor*, *ln(impact factor)*, and *open access*. Additionally analyses use dummies for the 26 subject categories, *e.g.* psychology (0 or 1).

| <i>Variable name</i> | <i>Mean</i> | <i>Std.</i> | <i>Min</i> | <i>Max</i> |
|--------------------------|-------------|-------------|------------|------------|
| percent_cited | 0.193% | 0.545 | 0% | 14.7% |
| impact factor | 1.89 | 2.47 | 0.100 | 45.9 |
| ln(impact factor) | 0.212 | 0.909 | -2.30 | 3.83 |
| open access | 7.1% O.A. | ---- | 0 | 1 |
| neighborhood size | 144.8 | 88.6 | 2 | 559 |

Table 2. Descriptive statistics of key variables.

Lastly, Figure 5 displays a scatter plot of the key dependent variable, *percent_cited*, versus impact factor and open access.



Figure 5. Scatter plot of journals' `percent_cited` vs. impact factor and open access policy. Dots are scaled by the total amount of articles published by each journal (and indexed by Scopus). Open access journals are shown with red dots.

The scatter plots appear to show a modest relationship between a journal's impact factor and `percent_cited`, the percent of its articles referenced on Wikipedia, especially when considering journal size (dot size). The next section analyzes these relationships statistically.

Results

We first present results of English Wikipedia's coverage. We ask, does Wikipedia draw equally on all branches of science? Next we focus on the role played by a journal's status and accessibility in predicting Wikipedia references. Lastly we extend the analysis to non-English Wikipedias and demonstrate the trade-off between the status and accessibility of references.

Coverage

Figure 6 below summarizes which branches of the scientific literature the English Wikipedia draws upon. The left panel shows the number of articles published by the top 250 journals in each field. The right panel shows the percent of those articles that are referenced at least once in the English Wikipedia.

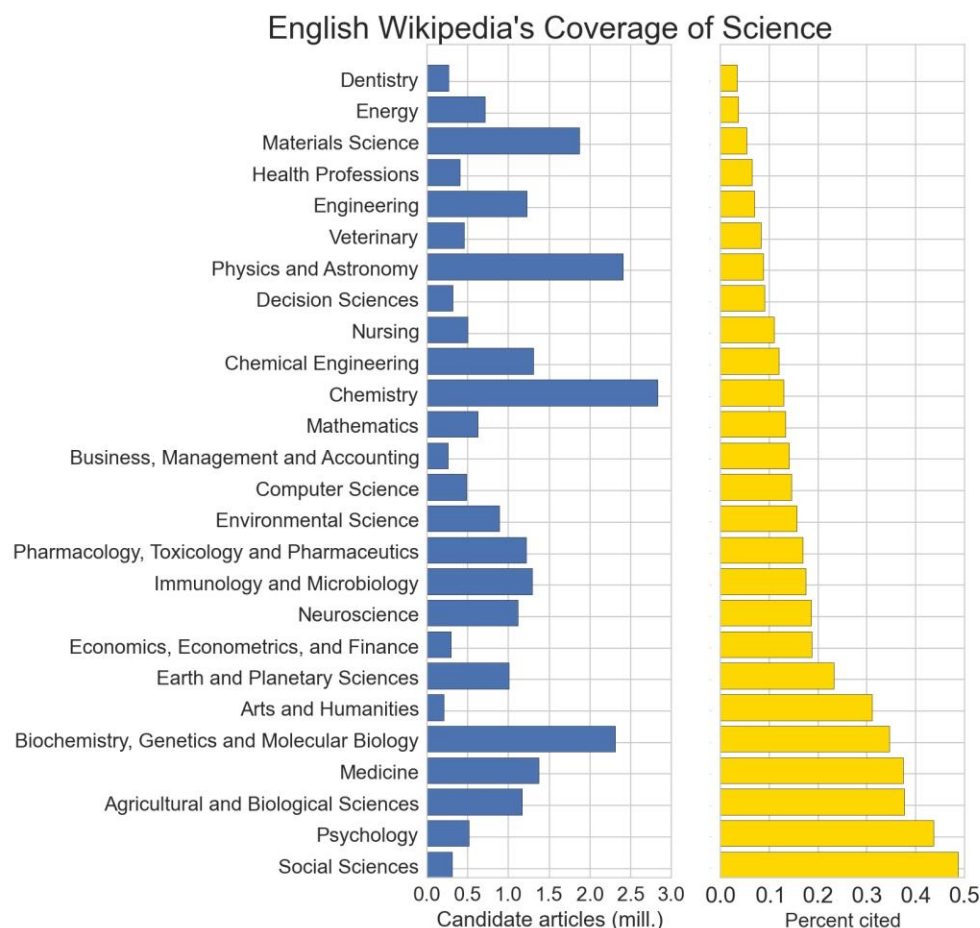


Figure 6. English-language Wikipedia's coverage of academic research. Each field's candidate literature (left) consists of the articles published by the 250-highest impact journals within the field and indexed by Scopus. The right panel shows the percent of these articles referenced on Wikipedia.

Figure 6 indicates that the coverage of science, as measured by the use of references, is very uneven and limited across scientific fields (Samoilenko & Yasseri, 2014). The social sciences represent a relatively small published literature but a relatively large portion of this literature is referenced on the English

Wikipedia (0.4 – 0.5%). At the other end of the spectrum, dentistry, also a relatively small literature, is rarely referenced (< 0.05%). The ordering of disciplines by *percent_cited* does not engender a simple explanation. For example, such an ordering does not appear correlated with traditional distinctions like hard vs. soft science, or basic vs. applied. This finding is echoed by Nielsen (2007), who found that “computer and Internet–related journals do not get as many [references] as one would expect if Wikipedia showed bias towards fields for the ‘Internet–savvy’”.

Status and accessibility

We now present results from a statistical model that predicts the probability p that an article from a given journal will be cited given that journal’s characteristics. The data-generating process is assumed to be a binomial process: each journal i publishes n_i articles and each of these articles is at risk p_i of being referenced in Wikipedia, where p_i depends upon journal characteristics \mathbf{x}_i such as impact factor. The number of articles referenced from journal i , y_i , then follows the binomial distribution

$$Pr(y_i = k \mid n_i, p_i) \sim \binom{n_i}{k} p_i^k (1 - p_i)^{n_i - k}$$

with expected value $n_i p_i$. The expected value of the fraction of a journal’s total articles referenced on

Wikipedia is then $\frac{n_i p_i}{n_i} = p_i$ and p_i is assumed to be related to \mathbf{x}_i ’s as $\ln\left(\frac{p}{1-p}\right) = \mathbf{x}_i \beta$, where β are the parameters to be estimated. The model described is a generalized linear model (GLM) with the binomial error distribution and the logit link function (Papke & Wooldridge, 1996) and is fit to the data $(\mathbf{x}_i, p_i = k_i/n_i)$ using maximum likelihood. This *journal-level* model may be viewed as a conventional logistic regression at the *article-level*; each journal produces n_i separate observations, k_i 1’s (Wikipedia references) and $(n_i - k_i)$ 0’s (unreferenced articles).

Table 3 below displays estimates from this model fitted to the English Wikipedia.

| Variable | Odds ratio | Std err of coeff | P-value |
|-------------------------|------------|------------------|--------------|
| open_access | 1.471 | 0.023 | 0.000 |
| log_sjr2013 | 1.878 | 0.007 | 0.000 |
| ag_bio_sciences | 2.291 | 0.018 | 0.000 |
| arts_hum | 1.836 | 0.043 | 0.000 |
| biochem_gen_mbio | 1.059 | 0.014 | 0.000 |
| bus_man_acct | 0.714 | 0.057 | 0.000 |
| chem | 1.004 | 0.022 | 0.863 |
| cheme | 0.968 | 0.030 | 0.282 |
| cs | 0.991 | 0.041 | 0.831 |
| decision_sci | 0.957 | 0.064 | 0.489 |
| dental | 0.520 | 0.107 | 0.000 |
| earth_plan_sci | 1.515 | 0.024 | 0.000 |
| econ_fin | 1.106 | 0.046 | 0.030 |
| energy | 0.551 | 0.064 | 0.000 |
| engineering | 0.507 | 0.037 | 0.000 |
| envi_sci | 0.743 | 0.029 | 0.000 |
| healthpro | 0.787 | 0.063 | 0.000 |
| immu_micro_bio | 1.114 | 0.023 | 0.000 |
| materials | 0.640 | 0.035 | 0.000 |
| math | 0.716 | 0.039 | 0.000 |
| medicine | 0.660 | 0.014 | 0.000 |
| neuro | 1.033 | 0.023 | 0.168 |
| nursing | 1.206 | 0.045 | 0.000 |
| pharm | 1.481 | 0.025 | 0.000 |
| phys | 0.629 | 0.025 | 0.000 |
| psyc | 2.628 | 0.025 | 0.000 |
| socialsci | 1.357 | 0.029 | 0.000 |
| vet | 0.898 | 0.054 | 0.048 |

Table 3. Estimates from the GLM estimated on English Wikipedia reference data. Variables with statistically insignificant odds ratios are not bolded.

The column of odds ratios indicates how the odds of referencing change with unit changes in the independent variables. For indicator variables, e.g. open access, these ratios are interpreted as the increase in odds when the indicator is true. For example, the odds that an article is referenced on Wikipedia increase by 47% if the article is published in an open access journal.

To interpret these results in terms of probabilities rather than odds ratios we must evaluate the model at particular values of the variables. Figure 7 displays the observed and predicted references for a range

of values of impact factor and *open_access*. The indicator variables designating particular disciplines are set at their modes (0).

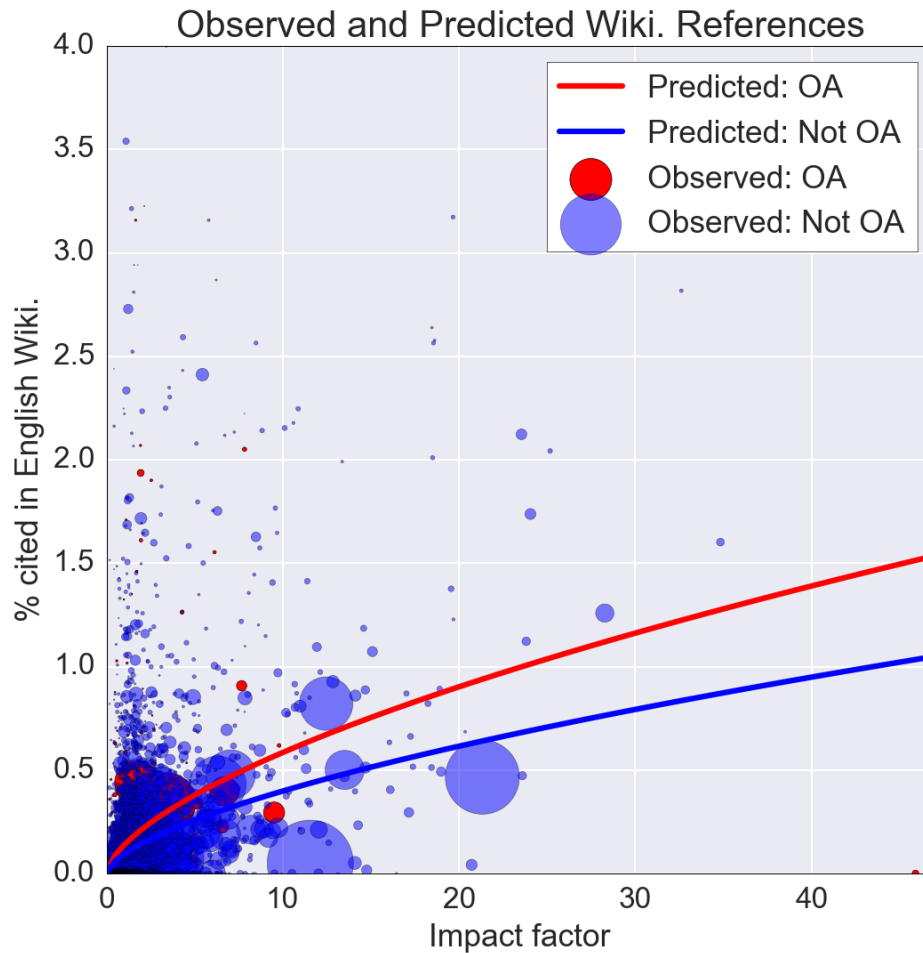


Figure 7. Observed (dots) and predicted (solid lines) English Wikipedia references. Red dots designate open access journals. The dot size is proportional to the number of articles the journal published.

The figure demonstrates that a journal's impact factor has positive and asymptotic effect on the percent of its contents referenced in the English Wikipedia (*percent_cited*). Open access journals (red dots) are relatively uncommon, but these journals are referenced more often than paywall journals of similar impact factor. For example, in our sample of psychology journals, open access journals have an average impact factor of 1.59, while closed access journals have an average impact factor of 1.77. Yet in the

English Wikipedia, editors reference an average of 0.49% of open access journals' articles but only 0.35% of closed access journals' articles, despite the higher impact factors.

Other languages

We estimated the same model on scientific references from 49 non-English Wikipedias. The weighting (described previously) ignores journals on subjects for which there was no demonstrable demand, and consequently various portions of the data across languages were excluded. Table A2 in the Appendix contains the percentages of journals that were excluded from estimating models in each language. This percentage varies widely. For example, only 0.17% of journals were not used for the English Wikipedia model, 49.87% for Slovak, and a 100% for Volapuk. These numbers correspond directly to demonstrable demand for various research literatures by the editors of each Wikipedia. While the English Wikipedia references ~32,000 articles from top journals, the Slovak Wikipedia references only 108 and Volapuk references 0.

From each Wikipedia's model, two parameters are of focal interest: the odds ratio (of probability of referencing) when open access is True, and odds ratio when (log of) impact factor increases by 1 unit. Table A3 shows these odds ratios and associated p-values for each Wikipedia. Figure 8 displays these odds ratios for those Wikipedias in which both estimates of both estimates were statistically significant at the 0.05 level.



Figure 8. The effect (odds ratio) of a journal’s impact factor vs. open access policy on referencing in several Wikipedias. The red line is the line of best fit (weighted least squares), and highlights the inverse relationship between a journal’s status and accessibility. The dot size (weight) is proportional to the number of scientific references. Only those languages are shown for which both effects are statistically significant.

While earlier results showed that both accessibility and status increase the odds that a journal will be referenced in the English Wikipedia, the relative strength of these effects varies across languages. Figure 8 shows that across 50 Wikipedias, there is an inverse relationship between the effects of accessibility and status on referencing. Some Wikipedias, like the Turkish, prioritize a journal’s academic status over accessibility; the odds of referencing high status journals are nearly 200% higher than lower status journals. Other Wikipedias, like the Serbian, prioritize accessibility over status; the odds of referencing an open access journals are ~275% higher than a paywall journal.

Intuition and previous work suggests poorer countries rely on open access literature more (J. A. Evans & Reimer, 2009), yet this pattern is not apparent in Figure 8. For example, India and Ukraine, relatively poor countries naturally associated with the Hindi and Ukrainian Wikipedias, actually exhibit a small

preference against open access literature, while a relatively wealthy country like Sweden has a Wikipedia that exhibits a huge preference for open access literature. The unexpected patterns may be due to the influence of bots (Steiner, 2014). For example, about a third of all articles on the Swedish Wikipedia were created by a bot (Jervell, 2014). Idiosyncrasies of the small number of human and non-human entities that edit science in non-English Wikipedias may thus play a larger role than gross cross-national patterns.

Conclusion

This article examined in unprecedented detail and scale how the world's 50 largest Wikipedias reference the scientific literature. Of central interest were the relationship between an article's academic status and accessibility on its probability of being referenced in a Wikipedia. Previous studies have focused only on the role of academic status on referencing in the English Wikipedia and have often ignored unreferenced articles. In contrast, we began by identifying an enormous (~19.4MM articles) corpus of scientific literature that scientists routinely cite, found a subset of this literature in which each of the 50 largest Wikipedias demonstrate demand, and estimated a statistical model for each Wikipedia to identify the features of journals that predict referencing.

We found that across languages, a journal's academic status (impact factor) routinely predicts its appearance on Wikipedia. We also demonstrated, for the first time, that a journal's accessibility (open access policy) generally increases probability of referencing on Wikipedia as well, albeit less consistently than its impact factor. The odds that an open access journal is referenced on the English Wikipedia are about 47% higher compared to closed access, paywall journals. Moreover, of closed access journals, those with high impact factors are also significantly more likely to appear in the English Wikipedia.

Therefore, editors of the English Wikipedia act as "bootleggers" of high quality science by distilling and distributing otherwise closed access knowledge to a broad public audience, free of charge. Moreo-

ver, the English Wikipedia, as a platform, acts as an “amplifier” for the (already freely available) open access literature by preferentially broadcasting its findings to millions. There is a trade-off between academic status and accessibility evident on Wikipedias around the world.

Limitations and directions for future research

Our findings are not without limitations. First and foremost, it bears emphasis that this study did not investigate the nature of Wikipedias’ sources as a whole (see Ford et al., 2013 for an excellent examination of sources). Only a fraction of Wikipedia’s references use the scientific literature, and this is the subset on which we focused. Consequently the present study cannot address the concern expressed by others, e.g. (Luyt & Tan, 2010), that sources outside the scientific literature are used too heavily in scientific articles. Second, the study was cross-sectional in nature; it is conceivable that open access articles differ from closed access, paywall articles in their relevance to Wikipedia. Future work can test the potential confounding factor of unmeasured relevance by observing reference rates for articles which have been experimentally assigned to open and closed-access statuses, as has been done by some psychology journals (Davis et al., 2008). Third, the study’s estimate of the impact of open access may be a conservative one. Wikipedia editors may simply use an article’s abstract rather than its full text to reference, and abstracts for both closed and open access journals are generally freely available. Furthermore, free copies of high impact articles from closed access journals may often be easily found online. Lastly, our attempt at studying the transmission of science to non-English Wikipedias is but a preliminary one. Scientific references from these Wikipedias can and should be extracted more carefully. The counter-intuitive patterns in referencing across Wikipedias also call for explanation. Future research should address whether scientific content found in Wikipedia across languages is differentially embedded in or husbanded by local scientific communities.

The impact of open access science across the world

The chief finding of this study bears emphasis. We believe the existing discussion of open access has focused too narrowly on the academic literature. Early results showing that open access improves scientific outcomes such as citations have been tempered by newer experimental evidence showing small to null causal effects. Perhaps the limited role of open access on the literature should not be surprising in light of the social stratification of research. The most productive scientists are often those with the most resources, and these resources include expensive journal subscriptions. On the other hand, our research suggests that open access policies have a tremendous impact on the diffusion of science to the broader general public through an intermediary like Wikipedia. The present article takes a major step in showing that this hypothesis, previously a matter of speculation only, has empirical support. As millions of people search for scientific information on Wikipedia, the science they find distilled and bottled into its articles consists of a disproportionate quantity of open access science.

Notes

1. <http://www.alexa.com/siteinfo/wikipedia.org>
2. <https://en.wikipedia.org/wiki/Wikipedia:Verifiability>. Accessed 2015-06-15.
3. <http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm>. Accessed 2015-06-16.
4. <http://en.wikipedia.org/wiki/Wikipedia:Notability>. Accessed 2015-06-11.
5. http://en.wikipedia.org/wiki/Wikipedia:Five_pillars. Accessed 2015-05-29.
6. http://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources#Scholarship. Accessed 2015-05-29.
7. This literature has grown to thousands of items and is impossible to summarize fully. See (Craig, Plume, McVeigh, Pringle, & Amin, 2007; Davis & Walters, 2011) for two reviews of parts of this literature.
8. <http://www.elsevier.com/online-tools/scopus/content-overview>. Accessed 2015-01-24.
9. The subject area “general” was excluded because it contained only four journals, all of which were cross-listed with other top-level categories.
10. <http://info.sciencedirect.com/scopus/scopus-in-detail/content-coverage-guide/journalclassification>. Accessed 2015-06-03.
11. Editors may also reference articles in other ways, for example by providing in-line links. We focus on the “cite-journal” template for three reasons. First, it shows clear intent to reference. Second, it has been used in previous research. Lastly, Ford and colleagues (Ford, Sen, Musicant, & Miller, 2013) found that “<ref>” tags were used most often to reference sources, and the “cite-journal” templates on which we focus are nested within such <ref> tags.

12. 2,005 (out of 4,721) journals are never referenced at all (*percent_cited* = 0).

Acknowledgements

This research was enabled by grant 39147 to the Metaknowledge Network by the John Templeton Foundation.

References

- Björk, B.-C., & Solomon, D. (2012). Open access versus subscription journals: a comparison of scientific impact. *BMC Medicine*, 10(1), 73. <http://doi.org/10.1186/1741-7015-10-73>
- Broadus, R. N. (1983). An investigation of the validity of bibliographic citations. *Journal of the American Society for Information Science*, 34(2), 132–135. <http://doi.org/10.1002/asi.4630340206>
- Butler, B., Joyce, E., & Pike, J. (2008). Don'T Look Now, but We'Ve Created a Bureaucracy: The Nature and Roles of Policies and Rules in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1101–1110). New York, NY, USA: ACM. <http://doi.org/10.1145/1357054.1357227>
- Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact?: A critical review of the literature. *Journal of Informetrics*, 1(3), 239–248. <http://doi.org/10.1016/j.joi.2007.04.001>
- Davis, P. M. (2010). Does open access lead to increased readership and citations? A randomized controlled trial of articles published in APS journals. *The Physiologist*, 53(6), 197, 200–201.
- Davis, P. M. (2011). Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *The FASEB Journal*, 25(7), 2129–2134. <http://doi.org/10.1096/fj.11-183988>
- Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G., & Connolly, M. J. L. (2008). Open access publishing, article downloads, and citations: randomised controlled trial. *BMJ*, 337, a568. <http://doi.org/10.1136/bmj.a568>
- Davis, P. M., & Walters, W. H. (2011). The impact of free access to the scientific literature: a review of recent research. *Journal of the Medical Library Association : JMLA*, 99(3), 208–217. <http://doi.org/10.3163/1536-5050.99.3.008>
- Evans, J. A., & Reimer, J. (2009). Open Access and Global Participation in Science. *Science*, 323, 1025.
- Evans, P., & Krauthammer, M. (2011). Exploring the Use of Social Media to Measure Journal Article Impact. *AMIA Annual Symposium Proceedings, 2011*, 374–381.
- Ford, H., Sen, S., Musicant, D. R., & Miller, N. (2013). Getting to the Source: Where Does Wikipedia Get Its Information from? In *Proceedings of the 9th International Symposium on Open Collaboration* (pp. 9:1–9:10). New York, NY, USA: ACM. <http://doi.org/10.1145/2491055.2491064>
- Gaulé, P., & Maystre, N. (2011). Getting cited: Does open access help? *Research Policy*, 40(10), 1332–1338. <http://doi.org/10.1016/j.respol.2011.05.025>
- González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, 4(3), 379–391. <http://doi.org/10.1016/j.joi.2010.03.002>
- Jamali, H. R., & Nikzad, M. (2011). Article Title Type and Its Relation with the Number of Downloads and Citations. *Scientometrics*, 88(2), 653–661. <http://doi.org/10.1007/s11192-011-0412-z>
- Jervell, E. E. (2014, July 14). For This Author, 10,000 Wikipedia Articles Is a Good Day's Work. *Wall Street Journal*. Retrieved from <http://www.wsj.com/articles/for-this-author-10-000-wikipedia-articles-is-a-good-days-work-1405305001>

- Laurent, M. R., & Vickers, T. J. (2009). Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association*, 16(4), 471–479. <http://doi.org/10.1197/jamia.M3059>
- Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology*, 60(7), 1327–1336. <http://doi.org/10.1002/asi.21024>
- Luyt, B., & Tan, D. (2010). Improving Wikipedia's credibility: References and citations in a sample of history articles. *Journal of the American Society for Information Science and Technology*, 61(4), 715–722. <http://doi.org/10.1002/asi.21304>
- Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., and Lanamäki, A. (2015). "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *J Assn Inf Sci Tec* 66(2): 219-245.
- Miller, J. D. (2001). Who is Using the Web for Science and Health Information? *Science Communication*, 22(3), 256–273. <http://doi.org/10.1177/1075547001022003003>
- Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, 58(13), 2047–2054. <http://doi.org/10.1002/asi.20663>
- Nielsen, F. A. (2007). Scientific citations in Wikipedia. *First Monday*, 12(8). Retrieved from <http://firstmonday.org/ojs/index.php/fm/article/view/1997>
- Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, 65(12), 2381–2403. <http://doi.org/10.1002/asi.23162>
- Papke, L. E., & Wooldridge, J. M. (1996). Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates. *Journal of Applied Econometrics*, 11(6), 619–32.
- Rekdal, O. B. (2014). Academic urban legends. *Social Studies of Science*, 44(4), 638–654. <http://doi.org/10.1177/0306312714535679>
- Samoilenko, A., & Yasseri, T. (2014). The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. *EPJ Data Science*, 3(1). <http://doi.org/10.1140/epjds20>
- Sample, I. (2012, April 12). Harvard University says it can't afford journal publishers' prices. Retrieved June 16, 2015, from <http://www.theguardian.com/science/2012/apr/24/harvard-university-journal-publishers-prices>
- Schroeder, R., & Taylor, L. (2015). Big data and Wikipedia research: social science knowledge across disciplinary divides. *Information, Communication & Society*, 0(0), 1–18. <http://doi.org/10.1080/1369118X.2015.1008538>
- Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, 314(7079), 497. <http://doi.org/10.1136/bmj.314.7079.497>
- Shaw, A., & Hill, B. M. (2014). Laboratories of Oligarchy? How the Iron Law Extends to Peer Production. *Journal of Communication*, 64(2), 215–238. <http://doi.org/10.1111/jcom.12082>
- Shuai, X., Jiang, Z., Liu, X., & Bollen, J. (2013). A Comparative Study of Academic and Wikipedia Ranking. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 25–28). New York, NY, USA: ACM. <http://doi.org/10.1145/2467696.2467746>
- Spoerri, A. (2007). What is popular on Wikipedia and why? *First Monday*, 12(4). <http://doi.org/10.5210/fm.v12i4.1765>

- Steiner, T. (2014). Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A Global Study of Edit Activity on Wikipedia and Wikidata. In *Proceedings of The International Symposium on Open Collaboration* (pp. 25:1–25:7). New York, NY, USA: ACM. <http://doi.org/10.1145/2641580.2641613>
- Trench, B. (2008). Internet: turning science communication inside-out? In M. Bucchi & B. Trench (Eds.), *Handbook of Public Communication of Science and Technology*. London and New York: Routledge. Retrieved from <http://www.routledgesociology.com/books/Handbook-of-Public-Communication-of-Science-and-Technology-isbn9780415386173>
- Van Noorden, R. (2013). Open access: The true cost of science publishing. *Nature*, 495(7442), 426–429. <http://doi.org/10.1038/495426a>
- West, R., Weber, I., & Castillo, C. (2012). Drawing a Data-driven Portrait of Wikipedia Editors. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (pp. 3:1–3:10). New York, NY, USA: ACM. <http://doi.org/10.1145/2462932.2462937>
- Wren, J. D. (2005). Open access and openly accessible: a study of scientific publications shared via the internet. *BMJ*, 330(7500), 1128. <http://doi.org/10.1136/bmj.38422.611736.E0>

Appendix

Table A1. Most common sources referenced using the *cite journal* template that are not indexed by *Scopus*.

| Journal name | Times referenced |
|-------------------------------|------------------|
| BILLBOARD | 1630 |
| NATIONAL PARK SERVICE | 539 |
| ROYAL GAZETTE | 523 |
| BULL AMER MATH SOC | 506 |
| FLIGHT INTERNATIONAL | 455 |
| BAH NEWS | 385 |
| NEW YORK TIMES | 369 |
| ROLLING STONE | 360 |
| ENTERTAINMENT WEEKLY | 342 |
| J BOMBAY NAT HIST SOC | 314 |
| WHOS WHO | 312 |
| BIZJOURNALS.COM | 288 |
| THE GUARDIAN | 287 |
| THE LANCET | 281 |
| VARIETY MAGAZINE | 270 |
| INSIDE SOAP | 239 |
| TIME MAGAZINE | 238 |
| BIOCHIMICA ET BIOPHYSICA ACTA | 201 |
| EDMONTON JOURNAL | 201 |
| SPORTS ILLUSTRATED | 197 |

Table A2. Percent of journal data that is not used in estimates language-specific models (demonstrated demand = 0). These percentages indicate the portion of research areas for which there is no demonstrable demand from (language-specific) Wikipedia editors.

| Wikipedia language | Percent excluded (weight=0) |
|--------------------|-----------------------------|
| english | 0.17 |
| chinese | 1.28 |
| russian | 1.28 |
| japanese | 1.36 |
| arabic | 1.47 |
| vietnamese | 2.01 |
| portuguese | 2.16 |
| german | 2.30 |
| spanish | 2.62 |
| indonesian | 2.95 |
| hindi | 3.62 |
| hungarian | 3.73 |
| ukrainian | 4.17 |
| slovenian | 4.25 |
| persian | 4.40 |
| serbian | 4.92 |
| greek | 4.99 |
| turkish | 5.63 |
| serbo-croatian | 6.35 |
| malay | 6.85 |
| bulgarian | 7.12 |
| dutch | 7.31 |
| catalan | 7.88 |
| danish | 8.36 |
| swedish | 9.78 |
| romanian | 10.08 |
| korean | 10.08 |
| estonian | 11.73 |
| galician | 12.13 |
| norwegian(bokmal) | 12.36 |
| czech | 12.84 |
| french | 13.68 |
| croatian | 14.52 |
| hebrew | 20.93 |
| armenian | 22.62 |
| waray-waray | 25.32 |
| esperanto | 26.37 |
| finnish | 26.81 |
| italian | 29.16 |
| lithuanian | 34.00 |
| norwegian(nynorsk) | 40.01 |

| | |
|-------------|-------|
| latin | 43.67 |
| slovak | 49.87 |
| polish | 62.69 |
| uzbek | 78.91 |
| minangkabau | 80.88 |
| kazakh | 81.44 |
| basque | 86.30 |
| cebuano | 86.95 |
| volapuk | 99.98 |

Table A3. Odds ratios and associated p-values for *open access* and (log) *impact factor* for 50 Wikipedias. Statistically significant odds ratios are bolded.

| Wikipedia Language | open access odds ratio | open access p-value | Ln(impact factor) odds ratio | Ln(impact factor) p-value |
|--------------------|---------------------------|------------------------|---------------------------------|------------------------------|
| arabic | 0.923 | 0.258 | 2.189 | 0.000 |
| armenian | 1.052 | 0.802 | 2.669 | 0.000 |
| basque | 0.000 | 1.000 | 1.059 | 0.890 |
| bulgarian | 0.936 | 0.623 | 2.256 | 0.000 |
| catalan | 1.452 | 0.000 | 1.963 | 0.000 |
| cebuano | 0.000 | 1.000 | 1.989 | 0.093 |
| chinese | 1.337 | 0.000 | 2.257 | 0.000 |
| croatian | 0.651 | 0.009 | 2.230 | 0.000 |
| czech | 1.247 | 0.230 | 2.258 | 0.000 |
| danish | 0.722 | 0.120 | 2.190 | 0.000 |
| dutch | 1.743 | 0.000 | 2.238 | 0.000 |
| english | 1.471 | 0.000 | 1.878 | 0.000 |
| esperanto | 1.114 | 0.591 | 1.245 | 0.001 |
| estonian | 1.221 | 0.151 | 2.705 | 0.000 |
| finnish | 0.666 | 0.300 | 1.576 | 0.000 |
| french | 0.850 | 0.550 | 2.030 | 0.000 |
| galician | 1.464 | 0.000 | 2.176 | 0.000 |
| german | 1.755 | 0.000 | 2.264 | 0.000 |
| greek | 0.798 | 0.078 | 2.008 | 0.000 |
| hebrew | 1.191 | 0.531 | 1.906 | 0.000 |
| Hindi | 0.757 | 0.029 | 2.113 | 0.000 |
| hungarian | 0.749 | 0.005 | 1.804 | 0.000 |
| indonesian | 0.886 | 0.242 | 2.467 | 0.000 |
| italian | 0.638 | 0.299 | 2.072 | 0.000 |
| japanese | 2.577 | 0.000 | 1.865 | 0.000 |

| | | | | |
|--------------------|--------------|--------------|--------------|--------------|
| kazakh | 0.000 | 1.000 | 1.759 | 0.114 |
| korean | 1.246 | 0.055 | 1.944 | 0.000 |
| latin | 0.812 | 0.637 | 1.975 | 0.000 |
| lithuanian | 1.035 | 0.923 | 2.345 | 0.000 |
| malay | 1.211 | 0.157 | 2.214 | 0.000 |
| minangkabau | 2.927 | 0.314 | 1.661 | 0.224 |
| norwegian(bokmal) | 0.866 | 0.437 | 2.328 | 0.000 |
| norwegian(nynorsk) | 0.588 | 0.109 | 1.510 | 0.000 |
| persian | 0.941 | 0.678 | 2.210 | 0.000 |
| polish | 0.588 | 0.480 | 2.330 | 0.000 |
| portuguese | 1.527 | 0.000 | 2.076 | 0.000 |
| romanian | 0.903 | 0.545 | 2.178 | 0.000 |
| russian | 1.419 | 0.000 | 2.086 | 0.000 |
| serbian | 3.824 | 0.000 | 1.516 | 0.000 |
| serbo-croatian | 3.761 | 0.000 | 1.518 | 0.000 |
| slovak | 1.943 | 0.157 | 3.249 | 0.000 |
| slovenian | 0.926 | 0.487 | 2.389 | 0.000 |
| spanish | 1.913 | 0.000 | 1.698 | 0.000 |
| swedish | 3.745 | 0.000 | 2.094 | 0.000 |
| turkish | 1.262 | 0.021 | 2.956 | 0.000 |
| ukrainian | 0.818 | 0.030 | 2.566 | 0.000 |
| uzbek | 0.000 | 1.000 | 2.963 | 0.012 |
| vietnamese | 0.966 | 0.682 | 2.143 | 0.000 |
| volapuk | 0.588 | 0.480 | 2.330 | 0.000 |
| waray-waray | 2.104 | 0.017 | 2.172 | 0.000 |