
¿Es demasiado anticuada la investigación en astronomía para el público? Citas de la Wikipedia a artículos de revistas de astronomía y astrofísica 1996-2014

Mike Thelwall

Mike Thelwall is the head of the Statistical Cybermetrics Research Group at the University of Wolverhampton, UK. He has developed a wide range of software for gathering and analysing web data, including hyperlink analysis, sentiment analysis and content analysis for Twitter, YouTube, MySpace, blogs and the Web in general.

http://orcid.org/0000-0001-6065-205X

Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton
Wulfruna Street, Wolverhampton WV1 1LY, UK
m.thelwall@wlv.ac.uk

Abstract

Astronomy is a natural science attracting substantial public interest. On a human scale, most individual celestial objects are essentially unchanging but is the same true for interest in astronomy research? This article uses the popular online encyclopedia Wikipedia as a proxy for public interest in academic research and assesses the extent to which it cites astronomy and astrophysics articles published between 1996 and 2014. Automatic Bing searches in Webometric Analyst were used to count the number of citations to astronomy and astrophysics articles from Wikipedia. The results show that older papers from before 2008 are increasingly less likely to be cited. This is true overall and in most of the major language versions of Wikipedia, although it may reflect editors’ interests rather than the public’s interests. This is consistent with a moderate tendency towards obsolescence in public interest in research, although it is probably affected by the dates on which most Wikipedia content on the topic was created. Papers may become obsolete if they report evidence that are later superseded by improved data or if they propose a model that is later replaced.

Keywords

Wikipedia; Astronomy; Scientometrics; Citation analysis.

Resumen

La astronomía es una ciencia natural que atrae un importante interés del público. A una escala humana, la mayoría de los objetos celestes individuales son esencialmente inmutables, pero ¿ocurre lo mismo con el interés por la investigación astronómica? Este artículo utiliza la popular enciclopedia online Wikipedia como un proxy del interés público en la investigación académica y evalúa el grado en que se citan artículos de astronomía y astrofísica publicados entre 1996 y 2014. Se hicieron búsquedas automáticas en Webometric Analyst a través de Bing para contar el número de citas a artículos de astronomía y astrofísica desde la Wikipedia. Los resultados muestran que los papers más antiguos, de antes de 2008, tienen menos probabilidades de ser citados. Esto es cierto en general y en la mayoría de las principales versiones lingüísticas de Wikipedia, aunque puede reflejar los intereses de los editores en lugar de los intereses del público. Esto concuerda con una tendencia moderada hacia la obsolescencia en el interés público en la investigación, aunque es probable que el fenómeno esté influido por las fechas en las que se creó la mayoría del contenido de Wikipedia sobre ese tema. Los artículos pueden llegar a ser obsoletos si presentan la evidencia de que más tarde fueron sustituidos por otros con datos mejorados o si proponen un modelo que luego se reemplazó.

Palabras clave

Wikipedia; Astronomía; Cienciometría; Análisis de citas.
1. Introduction

Astronomy interests large sections of the public and is supported by news and wider media coverage, including the long running UK TV show *The Sky at Night* (Moore, 2010). This attention is due to intrinsic interest in the topic (e.g., Stephens, 1990), the relative simplicity of key components and the provision of attractive images (Heck; Madsen, 2003, pp. vii; Meadows, 2000). Astronomy research, like all natural sciences research, might be expected to have permanent value because it deals with a mostly unchanging topic (at least within human lifetimes). Within academia, however, it is common for more recent scientific articles to attract more attention, as evidenced through citations (Brookes, 1970). This may be due to changes in the state of the art or changes in topics that are actively researched (Griffith et al., 1979) and does not imply that older research has become obsolete. Since about 1975, the average age of the literature cited in astronomy and astrophysics publications has increased, which is consistent with older articles retaining their value over time to some extent, even though cited articles tend to be relatively young, at about 9 years (Larivière; Archambault; Gingras, 2008). The same may not be true, for example, of space science engineering research because this can be expected to become obsolete as technology progresses. If astronomy research had lasting public value in this way, then this would help to justify the substantial resources devoted to space exploration and astronomical telescopes.

"Wikipedia is a source of evidence about the long term value of astronomy research to the public because it is a widely used source of knowledge that attempts to be reasonably comprehensive and extensively cites astronomy research"

Public interest in topics can be gauged to some extent by media coverage (e.g., Schäfer, 2008) and social media postings (Cody et al., 2015) or web engagement with science-related content (e.g., Haran; Poliakoff, 2012; Sugimoto; Thelwall, 2013). These tend to be driven by specific topics or news events and often do not cite academic sources and so are not ideal for analysing overall public interest in academic research. Surveys have also been used (e.g., Kennedy; Funk, 2015) but these are slow and expensive to do well. An alternative source of evidence about the long term value of astronomy research to the public is *Wikipedia* because it is a widely used source of knowledge that attempts to be reasonably comprehensive (e.g., Mesgari et al., 2015) and extensively cites astronomy research (Kousha; Thelwall, in press). This is an indirect source of evidence because it is mediated by *Wikipedia* editors but can be expected to reflect public interest to some extent, or at least what the editors believe to be public interest. *Wikipedia* citations to academic research are an alternative web impact indicator (Thelwall; Kousha, 2015), but in the current article are used to track interest in a discipline rather than the impact of individual articles.

*Wikipedia* was founded in 2001, had become the 37th most visited website by October 2005, according to Alexa.com (Giles, 2005) and was the 6th most visited website in August 2016, also according to Alexa.com. As reported by the Google Trends tool, the proportion of Google searches for *Wikipedia*, grew exponentially from January 2004 to November 2006, then stabilised until May 2010, before gradually reducing (roughly in line with its growth: Suh et al., 2009). The amount of editing work conducted on the main English language version of *Wikipedia* (there is also a simple *English* variant) exhibits a similar pattern. The total number of hours devoted by registered editors grew exponentially from 2001 to 2007 and then declined slowly to about two thirds of its peak value by the middle of 2012 (Geiger; Halfaker, 2013). If this trend has continued, then current main English language *Wikipedia* editing activity in 2016 may be at about half of its peak 2007 value.

The total number of hours devoted by *Wikipedia* registered authors grew exponentially from 2001 to 2007 and then declined slowly to about two thirds of its peak value by the middle of 2012.

"Wikipedia articles are both read and written by researchers. A survey in 2005 found that 17% of *Nature* authors consulted it weekly and 10% sometimes edited it. These proportions have presumably increased substantially since 2005. *Wikipedia* is also a standard resource for many students (Lim, 2009). Although only 13 scientific articles were within 270 of the most visited *Wikipedia* pages September 2006 to January 2007, four of these were from *Astronomy: Extra-solar planet; Pluto; 136199 Eris; Black hole* (Spoerri, 2007). An early study of citations in *Wikipedia* to academic journal articles found that the journals that receive the most academic citations are also the most cited in *Wikipedia*, but found four astronomy journals to be unusually highly cited (Nielsen, 2007). *Wikipedia* is also studied by researchers, both as a social phenomenon and as a data source (Bar-Ilan; Aharony, 2014)

*Wikipedia* editors mediate between public interest and the contents of pages, introducing an element of bias for attempts to use *Wikipedia* as an indicator of public interest. Nevertheless, editors are expected to curate existing knowledge rather than to create new knowledge and ci-
Does Astronomy research become too dated for the public? Wikipedia citations to Astronomy and Astrophysics journal articles 1996-2014

This article assesses the longevity of published astronomy and astrophysics research from the perspective of public interest, using Wikipedia coverage as a proxy for public interest and Wikipedia citations to published journal articles as the main source of evidence. As discussed above, whilst Wikipedia has a huge user base and is therefore a logical source of this type of information, it has biases due to the nature of its users and the mediating effect of a small number of active editors. The research questions are:

- RQ1: How does the proportion of Astronomy and Astrophysics articles cited in Wikipedia vary depend on their publication year?
- RQ2: Does the answer to the above question vary by language version of Wikipedia?

2. Methods

Articles from the Scopus Astronomy and Astrophysics category were downloaded for every even year from 1996 to 2014 using the query submain(3103) AND DOCTYPE(ar) AND SRCTYPE(j). The start year was chosen to be 1996 because Scopus indexing of journals changes in this year (Archambault et al., 2009), even though its coverage increased steadily from 1965 (Jacs, 2005), and so 1996 is a natural cut-off point for a long term analysis. Even years were chosen to reduce the amount of data needed. The most recent 5,000 and the oldest 5,000 articles for any query can be downloaded from Scopus. These two sets were combined to create a complete set of journal articles for most years in Scopus but for more recent years, not enough articles could be extracted. In these cases, the most recent and oldest 5000 articles were combined and used instead because these form a balanced set, in terms of publication date, and so it is reasonable to compare them to the complete sets (Table 1).

A Wikipedia-specific Bing query was created for each article by combining the author last name of the first (up to) three authors, the first (up to) six words of the article title as a phrase search, the publication year, and the site-specific advanced search term site:wikipedia.org, as in the following example:


This method has been shown to be accurate and reasonably comprehensive (Kousha; Thelwall, in press). Articles with titles having less than three words were discarded as likely to generate false matches. All cases where two different articles generated the same query were also removed. From the remaining queries, a random sample of up to 5,000 (Table 1), was submitted to Bing through its API in Webometric Analyst.

http://lexiurl.wlv.ac.uk

This number was judged to be sufficient to reveal differences between years. The queries were submitted August 6-7, 2016. Manual checking of the results confirmed that the Bing matches were accurate, although it is likely that they were not comprehensive since search engines do not index the entire web (Lawrence; Giles, 1999; Vaughan; Thelwall, 2004) and so Bing presumably does not index all Wikipedia pages.

Some queries returned both the mobile Wikipedia page and an equivalent standard Wikipedia page. Such pages have identical contents but different formats and so whenever both URLs was present, the mobile equivalent URL was removed.

The results were analysed overall and also separately for different language versions of Wikipedia in order to detect different linguistic or national trends. Step by step instructions for data collection and analysis are available online.

http://lexiurl.wlv.ac.uk/reports

The average number of Wikipedia citations per article was calculated separately for each year to show temporal trends. The geometric mean was used instead of the arithmetic mean because citation-like data is often highly skewed (Price, 1976; Thelwall; Wilson, 2016) and the arithmetic mean is inappropriate for this type of data. Confidence intervals were calculated using the normal distribution formula on the log-transformed data during the geometric mean calculation process (Thelwall, 2016). The percentage of articles cited by Wikipedia was also calculated to give additional insights. Confidence intervals were obtained for this with a standard formula for proportions, Wilson’s score interval (Wilson, 1927).

Table 1. The number of Astronomy and Astrophysics Wikipedia article queries submitted, by year.

<table>
<thead>
<tr>
<th>Year</th>
<th>Scopus articles</th>
<th>Scopus articles identified</th>
<th>Usable Scopus articles</th>
<th>Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>5,404</td>
<td>5,404</td>
<td>5,322</td>
<td>5,000</td>
</tr>
<tr>
<td>1998</td>
<td>5,959</td>
<td>5,959</td>
<td>5,865</td>
<td>5,000</td>
</tr>
<tr>
<td>2000</td>
<td>6,746</td>
<td>6,746</td>
<td>6,669</td>
<td>5,000</td>
</tr>
<tr>
<td>2002</td>
<td>5,282</td>
<td>5,282</td>
<td>4,753</td>
<td>4,753</td>
</tr>
<tr>
<td>2004</td>
<td>4,981</td>
<td>4,981</td>
<td>4,931</td>
<td>4,931</td>
</tr>
<tr>
<td>2006</td>
<td>5,487</td>
<td>5,487</td>
<td>5,453</td>
<td>5,000</td>
</tr>
<tr>
<td>2008</td>
<td>8,032</td>
<td>8,032</td>
<td>7,883</td>
<td>5,000</td>
</tr>
<tr>
<td>2010</td>
<td>12,354</td>
<td>10,000</td>
<td>9,917</td>
<td>5,000</td>
</tr>
<tr>
<td>2012</td>
<td>13,029</td>
<td>9,999</td>
<td>9,941</td>
<td>5,000</td>
</tr>
<tr>
<td>2014</td>
<td>12,872</td>
<td>10,000</td>
<td>9,914</td>
<td>5,000</td>
</tr>
</tbody>
</table>
3. Results and discussion

For Wikipedia overall and the language versions with at least 500 citations, the trend is that 2008 is the peak year for Astronomy and Astrophysics articles to be cited, with a gradual tendency for older article to be less cited (Figures 1, 2). This is true both for the average (geometric mean) number of citations per article (Figure 1) and the proportion of articles cited (Figure 2). Thus, the results give clear evidence that older Astronomy and Astrophysics articles are less likely to be cited in Wikipedia.

Comparing Figure 1 with Figure 3, it is clear that the trend is different to that for citations. Within Figure 3, there is an increasing trend 1996-2000, but since this observation is based on only three data points, it is not clear that this trend would continue. It is possible that very old Astronomy and Astrophysics are less cited than moderately old Astronomy and Astrophysics articles because the database coverage of Scopus has expanded over time and so, in earlier years of its coverage it may have included a lower proportion of the citing journals and therefore captured a lower percentage of an articles citations within the first few years after publication. Since articles tend to be cited in the few years after publication, this would affect older articles more than younger ones. This issue does not affect Wikipedia and so cannot explain the decreasing trend in the encyclopedia. The comparison between Figure 1 and Figure 3 suggests that the increasing trend until 2008 in Figure 1 is unlikely to be due to Scopus tending to cover lower impact Astronomy and Astrophysics research during earlier years.

The decreasing trend from 2008 to 2014 in Figure 1 and Figure 2 seems likely to be at least partly due to new articles needing time to be recognised and incorporated into Wikipedia. Although some editors may monitor new academic research and incorporate it into existing or new Wikipedia pages as soon as it is published, they may have other strategies. For example, they may attempt to give comprehensive coverage to an area, seeking out academic research to back up new content. Authors of academic papers, or their colleagues, may also add new articles judged important to Wikipedia. Alternatively, editors may extract the citations from secondary sources, such as textbooks or review articles, when adding or updating content. Thus, a degree of time lag between the publication of a paper and its incorporation into Wikipedia seems reasonable, although 8 years (from 2008 to 2016, the data collection year) seems like a long time. Alternatively, it is also possible that Wikipedia’s astronomy content matured 6-8 years ago and has stabilised since then, with less need to be updated and newer research therefore tending to be overlooked. For the English language version of...
Wikipedia there has been a gradual decrease in total editing hours by registered editors, giving an estimated total reduction of 50% by 2016 (see the discussion in the Introduction based on: Geiger; Halfaker, 2013). Thus, Wikipedia as a whole may have matured to the extent that less work is needed to maintain it.

The increasing trend from 1996 to 2008 in Figure 1 and Figure 2 could have multiple explanations. Older articles may end to cover topics that have become obsolete, may be subsumed within future research that cites them, or may be subsumed within review articles or textbooks that are cited instead. Thus, older references may be pruned if the citing page is judged to be no longer relevant to the encyclopedia or if they are replaced by a newer citation that incorporates, refutes or modifies their knowledge. An alternative explanation is that the peak period 2008-2010 for Wikipedia citations to Astronomy and Astrophysics research could have also been the peak period in editing Astronomy and Astrophysics articles in Wikipedia, with the editors at that time being more aware of recent research and therefore citing it more. This broadly fits with the decrease in editing of Wikipedia by registered users since 2007 (Geiger; Halfaker, 2013).

The trend is that 2008 is the peak year for Astronomy and Astrophysics articles to be cited, with a gradual tendency for older article to be less cited

An investigation of articles from 1996 that were cited in Wikipedia suggested that they mostly contained timeless information. The most cited, A catalog of parameters for globular clusters in the Milky Way, was drawn upon by several Milky Way articles. These were mainly in the Czech version of Wikipedia. This is consistent with the astronomy coverage of Wikipedia being internationally uneven and driven in each language by one or a few highly active editors, presumably with an amateur or professional interest in astronomy. Some popular articles had apparently time-dependent information, including Further evidence for the existence of additional small satellites of Saturn, which might become irrelevant when better evidence is found for its claims. The same is true for some of the uncited articles, such as Proposed reference model for middle atmosphere water vapor, which may become obsolete when a better model is found. Wikipedia also contains some articles about academics (Samoilenko; Yasseri, 2014) that might cite their work irrespective of its current value. These articles might be pruned when the scholars are no longer active and this would affect the authors of older articles the most.

For the English language version of Wikipedia there has been a gradual decrease in total editing hours by registered editors, giving an estimated total reduction of 50% by 2016

3.1. Language differences

The main language versions of Wikipedia, in the sense of citing Astronomy and Astrophysics articles at least 500 times, all display a broadly similar pattern of citing both older and younger research less than research from 2006-2010 (Figure 4). This is particularly evident for English, French, Russian and Spanish. The Chinese version seemed to peak in 2002 (perhaps partly blocked by government actions after this) and the Italian version gives similar coverage to a wider range of years, 2000-2008. An inspection of articles cited by Italian version in 2000 suggested that they tended to be about individual planetary bodies, such as the yellow dwarf star HD 202206, for which a brown dwarf companion was discovered in 2000. Another example, a more general article on asteroid families (Famiglia di asteroidi), included a “biography” section that seemed to have been written in 2002, listing recent research that was related to the topic, even though not specifically described. One of the listings was an article entitled, 9 Metis and 113 Amalthea: A genetic asteroid pair, despite neither asteroid being mentioned by name in the Wikipedia page. Thus, some articles may be added to Wikipedia pages as examples of recent relevant research rather than as citations to inform the contents of the page.
A different trend is evident for half of the six language versions of Wikipedia that cite 200-499 astronomy and astrophysics articles, Portuguese, Japanese, and Korean (Figure 5). In these language versions the trend is broadly similar except for a spike in 2006. An examination of the citing pages did not find a clear theme of a topic for this year, but public interest in Astronomy may have peaked in this year due to the widely publicised agreement by the International Astronomical Union that Pluto was not a planet: https://www.iau.org/public/themes/pluto
This may have helped to recruit, or activate, Wikipedia editors with an interest in astronomy research in 2006.

4. Limitations and conclusions
The results suggest, but do not prove, that older Astronomy and Astrophysics articles tend to carry information that is of less current interest to the public than that of newer articles. The evidence of this is from Wikipedia, which may reflect the concerns of a small number of researchers and amateurs that are enthusiastic international Wikipedia editors rather than the general public. The role of active astronomy editors is particular limitation from the perspective of the research question because these mediate between public interest and academic research. Most importantly, their collective activity on the English language version peaked in 2007 and they may work less on established areas of the encyclopaedia, allowing core Wikipedia astronomy pages to stabilise. The results may also be affected by the changing content of Scopus categories and of the astronomy and astrophysics journals covered. The pattern seems to be similar across languages, although perhaps affected by popular astronomy events triggering a burst of public interest and Wikipedia editing.

The apparent lower interest in older articles published before 2008 may be due, at least in part, to the existence of some articles with research that can be superseded by better evidence or models. If a follow-up study can address the issue in a few years, then it would be useful to compare the results to see whether there is a tendency for older citations to be eliminated. It would also be useful to exploit Wikipedia’s recorded history of edit changes to examine when references are added, updated or pruned. This would allow some of the tentative conclusions in this article to be made more definite.

Finally, the method of using Wikipedia citations as a proxy for public interest, as introduced in this article, is limited due to the intermediating role of Wikipedia editors, although the ability to compare different language versions of Wikipedia can help to check overall trends. Astronomy may be a best case scenario for public interest in the natural sciences, but the method may also be useful in other areas with public
interest, such as the health sciences (Kennedy; Funk, 2015), arts and humanities. It may also be useful in the social sciences due to the degree of professional interest in academic research and the need for professionals, such as teachers, to keep updated with useful theories and strategies.

Public interest in Astronomy may have peaked in 2006 due to the widely publicised agreement by the International Astronomical Union that Pluto was not a planet.

5. References


Cody, Emily; Reagan, Andrew; Mitchell, Lewis; Dodds, Peter; Danforth, Christopher (2015). “Climate change sentiment on Twitter: an unsolicited public opinion poll”. PloS one, v. 10, n. 8, e0136092. https://doi.org/10.1371/journal.pone.0136092


Kousha, Kayvan; Thelwall, Mike (in press). “Are Wikipedia citations important evidence of the impact of scholarly articles and books?”. Journal of the Association for Information Science and Technology. https://doi.org/10.1002/asi.23694


