Wikipedia as a tool for contemporary history of science: A case study on CRISPR

Rona Aviram[1,2], Omer Benjakob[1,2]
(1) Université Paris Cité, Inserm U1284, System Engineering and Evolution Dynamics, F-75004 Paris, France
(2) Learning Planet Institute, F-75004 Paris, France

Keywords

History of Science, CRISPR, Scientometrics, Digital Humanities, Science of Science.

Abstract: Rapid developments and methodological divides hinder the study of how scientific knowledge accumulates, consolidates and transfers to the public sphere. Our work proposes using Wikipedia as a historiographical source for contemporary science, through a case study on CRISPR. Using a mixed method approach, we qualitatively and quantitatively analyzed the text, sections and references, of 50 English-language CRISPR-linked articles. These, we found, documented CRISPR's maturation from a scientific discovery to a biotechnological revolution with vast social implications. Our method, as well as open tools we developed, can support additional case studies.

Overview: Wikipedia's science articles top search results, making it a key node in the transference of academic knowledge to the public.[1] Expanding on past case studies[2], we propose a method for using Wikipedia as a source of historical knowledge.[3] Our method combines traditional history of science[4], e.g. bibliometrics[5] and thick description[6], with computational approaches from digital humanities[7], and is demonstrated through a case study of the CRISPR field.
CRISPR-based gene editing has been labeled the scientific breakthrough of the 21st century, earning widespread notoriety and sparking public debate. In a relatively short period of time, CRISPR-based gene-editing tools have been labeled the scientific breakthrough of the 21st century.[8] While CRISPRs were identified in the 1980's, they came to prominence after 2005 and 2012. Thus, it is a prime example of a scientific field that has undergone massive growth during Wikipedia's lifespan.

Methodology: Our approach combines quantitative and qualitative analyses to harness both data- and content-dependent historical value. We also provide automated tools[9] that utilize

[1] Teplitskiy, M., Lu, G. & Duede, E. Amplifying the impact of open access: Wikipedia and the diffusion of science. J. Assoc. Inf. Sci. Technol. 68, 2116–2127 (2017).
[2] Benjakob, Omer, and Rona Aviram. 2018. "A Clockwork Wikipedia: From a Broad Perspective to a Case Study." Journal of Biological Rhythms 33 (3): 233–44. https://doi.org/10.1177/0748730418768120.
[3] Benjakob, O., Aviram, R. & Sobel, J. A. Citation needed? Wikipedia bibliometrics during the first wave of the COVID-19 pandemic. GigaScience 11, (2022).
[4] Rowlands Bruce. Grounded in Practice: Using Interpretive Research to Build Theory. Electron. J. Bus. Res. Methodol. 3, 81–92 (2005).
[5] Mostafa, M. M. Two decades of Wikipedia research: a PubMed bibliometric network analysis. Glob. Knowl. Mem. Commun. (2021) doi:10.1108/GKMC-03-2021-0056.
[6] Elliott, Richard, and Avi Shankar. New paths to thick descriptions: innovativeness in data collection and interpretation. Vol. 8. No. 2. Emerald Group Publishing, 2005.
[7] Su, F. & Zhang, Y. Research output, intellectual structures and contributors of digital humanities research: a longitudinal analysis 2005–2020. J. Doc. 78, 673–695 (2022).
[8] Cohen, J. A cut above: pair that developed CRISPR earns historic award. Science 370, 271–272 (2020).
[9] See Github https://github.com/RonaTheBrave/WikiCorpusBuilder/blob/main/README.md

Wikipedia's data - its articles, their edit histories and their references - for this end. We employ a stepwise strategy: first, once a "term of interest" is selected (e.g., "CRISPR"), our "corpus" building tool searches Wikipedia's English articles to identify and delimit a corpus of relevant articles. Next, this body of articles is filtered and the tool retains only those with the term in their *title* or *section title* (thus excluding articles with only minor or incidental usage). With a well defined corpus, a number of automatic processes are conducted: Our tool scraps the articles for their metadata, such as date of creation, creator ID, edit counts, article size, and references type and counts. These support a number of quantitative analyses, but are also combined with in-depth semantic reading of the revision history of "anchor article(s)" (e.g., "CRISPR"), which is then addressed through a "thick description" approach[10], facilitated by the "view history" tool. Moreover, two bibliometric measures were developed - the "SciScore", which gauges the ratio of academic to non-academic sources within an article; and "Latency", which calculates the duration between an academic paper's publication and its first reference on Wikipedia. Our tool creates a timeline of all the articles in a corpus to provide an overview of the historical dynamics at the corpus level, such as the creation of the different articles at different times.

Results: We analyzed 50 CRISPR-related Wikipedia articles and their references. Our findings suggest that Wikipedia can serve as a tool in the digital history of contemporary science. By reviewing the CRISPR article's history, we saw that the article started off describing the "basic science" behind CRISPR, and was updated in the wake of the publication of canonical works in the field. These processes were reflected quantitatively and qualitatively. For example, shifts in our SciScore corresponded with the emergence of public interest in the term (and thus the appearance of non-academic sources). Collectively, these highlight how bibliometric shifts are reflective of substantive changes in the article's texts, which in turn are reflective of real-world developments in the field.
Over time, the article and the corpus grew, with the emergence of gene editing technology and the forking off of the main article into a number of affiliated ones with a more narrow focus, while the original CRISPR article offered a consolidated overview of the scientific narrative on CRISPR in bacterial systems. Using our 'first mention' analysis, we could also observe CRISPR's interface with other scientific fields. For example, the two oldest articles in the corpus, "Wheat" and "Antibiotic", were opened in 2001, and were late to adopt "CRISPR" some twenty years later, only after it was consolidated as a field and its interface with other bodies of knowledge became clear. The articles, their text and their different citations, we found, served as a rich record of the growth not just of academic knowledge, but also social aspects of science. For example, the legal battles regarding CRISPR and the academic credit wars over what the journal Science called the "CRISPR Craze"[11] were documented in a blow-by-blow manner.

Contribution: These findings join a growing body of research using Wikipedia for historical ends - including past studies we published on COVID-19 (2022) and circadian clocks (2018). However, as of yet, no method has been put forward to this end. Here, we introduce and implement such a method, define the scope of future work and provide open automated tools to allow others to realize them (Github). These are but initial steps towards the creation of more case studies into how scientific knowledge is represented on Wikipedia.

[10] Jemielniak, D. Thick big data: doing digital social sciences. (Oxford University Press, 2020).
[11] Pennisi, E. The CRISPR Craze. Science 341, 833–836 (2013).