Does Wikipedia Cover the Relevant Literature on Major Innovations Timely?

An Exploratory Case Study of CRISPR/Cas9

Marion Schmidt,[1] Wolfgang Kircheis,[2] Arno Simons,[3] Martin Potthast,[2] and Benno Stein[4]

[1]German Center for Higher Education Research and Science Studies (DZHW), schmidt@dzhw.eu
[2]Leipzig University, wolfgang.kircheis@uni-leipzig.de, martin.potthast@uni-leipzig.de
[3]DZHW, Humboldt Universität zu Berlin, arno.simons@posteo.de
[4]Bauhaus-Universität Weimar, benno.stein@uni-weimar.de

Abstract

This *research-in-progress* paper analyzes Wikipedia's representation of the Nobel Prize winning CRISPR/Cas9 technology to explore to what extent and with what temporal dynamics Wikipedia cites the most relevant and visible scientific literature on this topic. We use both verbatim and fuzzy matching heuristics to match publications cited from a selection of secondary formats—like reviews—as well as field-delineated highly cited publications with the central Wikipedia article on CRISPR. Our methodical results confirm that a combination of verbatim searches by title, DOI, and PMID is sufficient. Initial evidence also shows that the Wikipedia article references a substantial amount of articles that are well acknowledged by experts and highly cited, as well as literature that is not strictly scientific or less visible. Delays in coverage on Wikipedia compared to the publication years show a dependence on the dynamics of both the field and the Wikipedia article itself.

Introduction

The Nobel Prize winning development of the CRISPR/Cas9 mechanism and technique, the first steps of which date back more than twenty years, has left a long paper trail of scientific publications: the CRISPR/Cas9 literature and its citation network. To understand outstanding innovation processes such as CRISPR/Cas9 and for e.g. science policy to act upon it, the identification of the most relevant publications is of key importance. One way this can be done is by measuring certain properties of the citation network and taking them as proxies for relevance. Another option is to turn to the secondary literature (like reviews) which tries to answer the question of relevance with its specific methodologies, depending on the systematic or narrative format. A problem with the latter approach is time: Even if a review accurately reflects the relevant literature at the time of its publication, the review may soon become outdated—a problem even more pronounced in the recent development of so-called Living Reviews (Elliot et al. 2014).

In this paper we introduce and explore a third approach to determining the relevant literature of innovation processes like CRISPR/Cas9: the analysis of Wikipedia. The online encyclopedia is a tertiary literature format whose extensive network of articles on scientific subjects is comparable to the concept of a Living Review and hence may be utilized as such for specific domains. Wikipedia is characterized by the fact that it addresses the general public, that articles are constantly revised and updated; thus reflecting dynamics to a greater or lesser extent, as well as by its unique reviewing process. The evolution of references to academic literature on Wikipedia has increased significantly over the years, suggesting that Wikipedia could even be harnessed for altmetrics (Zagovora et al., 2020). Giles (2005), Reavley et al. (2012), Estevez & Cukierman (2012) and Garcia del Valle et al. (2018) suggest that Wikipedia's reflection of scientific knowledge is accurate, while Teplitskiy et al. (2015) as well as Jemielniak et al. (2019) show that top impact journals are most referenced in Wikipedia (medical) articles. For these top biomedical journal papers, it takes (on average) only about three months to be referenced in Wikipedia. But from the opposite perspective of a single research field, Benjakob & Aviram (2018) determined a medium citation latency time of five years. Seeing as innovation processes are characterized by transferring scientific knowledge into application contexts with implications on commercial, political, ethical, and legal aspects or questions, connecting wider societal spheres and public perception, these characteristics suggest Wikipedia as an interesting format for representing and mapping innovation processes.

In our paper we use the CRISPR/Cas9 innovation as an in-depth case study to shed light on Wikipedia's referencing patterns. Based on reference corpora matched to the central Wikipedia article on CRISPR we assess when references to scientific articles are picked up. Furthermore we analyze to which extent the reference structure in the Wikipedia article can be explained by two field perspectives, focusing on secondary accounts and on citation impact, respectively. This paper is part of a larger pilot study on Wikipedia as a lens into innovation-in-the-making (cf. Simons et al. 2021).

Materials and Methodology

We provide two corpora of publications on CRISPR, the CRISPR Accounts Corpus and the CRISPR WoS Corpus, each representing a different perspective on the CRISPR innovation, and we compare them with the revision history of the central Wikipedia article on CRISPR.[1,2] Both corpora are exploratory, and, to a certain extent, pragmatically defined, as patterns for literature referencing in Wikipedia in the specific case of innovations are as of yet unknown. The CRISPR Accounts Corpus comprises publications referenced in predominantly secondary literature formats, such as reviews and short communications. These accounts present and discuss the development of CRISPR, thus conveying experts' and stakeholders' perspectives on which publications have been relevant for the innovation process—collectively called 'accounts'. To create this corpus, we searched for the phrases "crispr history", "crispr development", and "crispr discovery" on Google Scholar, and, based on Google's relevance ranking, reviewed the paginated search result pages until no more relevant publications were identified in a row of consecutive pages. 12 publications, despite containing history sections, were discarded for not focussing on the history of CRISPR, one for not being listed in WoS. The resulting 29 sources have been complemented by three web resources presenting CRISPR timelines, which were obtained by searching for "crispr timeline" on Google, and reviewing the search results pages accordingly. We extracted all references from the 29 sources through WoS and extracted references manually in case of the timeline documents.

The CRISPR WoS Corpus is based on a bibliometric field delineation of CRISPR in the Web of Science (WoS). Publications containing "crispr"—as a highly distinctive term—in the title are defined as the core field. As a second layer, we add publications containing "crispr" in their abstracts. For a third layer, representing influences and effects, we delineate publications for which the proportion of references or citations to the core field to total references or citations is higher than 30 percent, thus representing a substantial connection. The number of publications in this corpus amounts to 20,532.[3] In order to represent an impact perspective in contrast to that of the history-oriented accounts, we sort this corpus by the absolute citation counts and cut off at 500 publications. This number is roughly in line with the magnitude of the references in the article (see below). The reason for not using citation windows here is rooted in the specific dynamics of an innovation, where early publications typically accumulate significant citation numbers over time[4].

For both corpora we downloaded the metadata from WoS, resulting in **1,186** and **500** unique publications, respectively, to be matched against all revisions of the English CRISPR article. All revisions of the CRISPR article were downloaded using the MediaWiki API, collecting the HTML versions of each revision's page (2,072 revisions as of February 28, 2021). Each revision has a unique revision timestamp, text sections, such as headings, paragraphs, captions, tables, and lists, the reference sections *References* and *Further Reading*, as well as other metadata. While more recent revisions of articles usually apply Vancouver style to format references, Wikipedia does not have a single house style but expects the editors to adopt a consistent style within articles,[5] potentially causing shifts of citation styles over the course of an article's history. This ambiguity together with the fact that editors easily introduce

[1] https://en.wikipedia.org/wiki/CRISPR

[2] There are more thematically related Wikipedia articles, like *Cas9* and *CRISPR Gene Editing*, but we limit our study to the central one. Our current work focuses on devising new approaches to field delineation on Wikipedia.

[3] On a frozen version of the WoS raw database from April 2020, to ensure reproducibility.

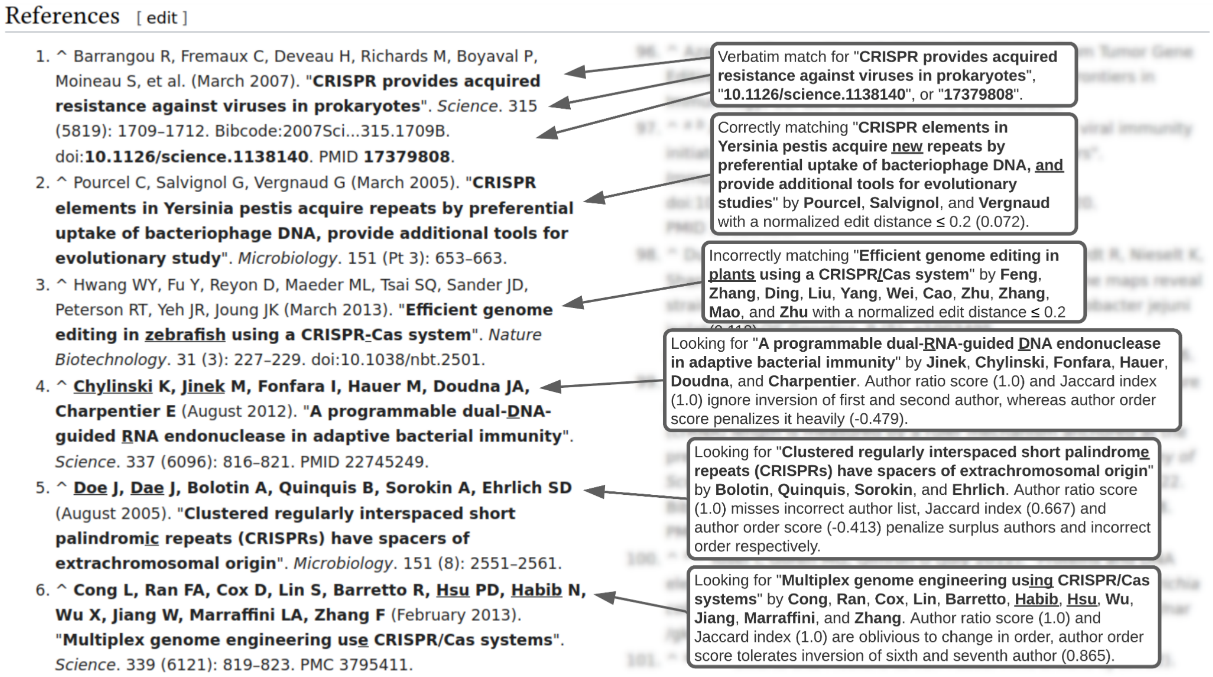
[4] We also opted against field-normalization due to biases in case of high-impact multidisciplinary journals

[5] https://en.wikipedia.org/wiki/Wikipedia:Citingsources#Citationstyle

typos when manually adding a reference to an article necessitates the development of a fuzzy reference matching approach—at least we thought so at the outset.

Our reference matching approach implements heuristics with various degrees of precision, which can be divided into verbatim heuristics and fuzzy heuristics. As verbatim heuristics, we match titles, DOIs, and PMIDs of the publications against the entire article text including all references. All strings are converted to lowercase ASCII, with title matching additionally utilizing alpha-numerical normalization. As fuzzy matching heuristics, we match publications' titles against extracted references and allow for a normalized edit distances of 0.2, 0.3, and 0.4 and combine the latter with three author matching strategies. We use the publication-to-reference-author ratio, the Jaccard Index of publication and reference authors, and an author order score: Each author of the publication is assigned a gain equal to its position in the inverted list of authors divided by its actual position, with the sum of all values being the ideal score. The authors of a given reference are then evaluated in turn, winning the same value if matching the author of the publication in the respective position and losing that value if not. The author order score is calculated by dividing the sum of these values by the ideal score. Figure 1 shows all matching heuristics at work.

For the time being we refrain from calculating exact recall values, since it requires to manually review all 2,072 revisions manually for possible matches. However, for an estimate of what might have been missed, we manually checked and deduplicated all titles, DOIs and PMIDs that had been extracted from all references throughout the article's entire revision history, resulting in 324 DOIs, 302 PMIDs, and 465 titles. The latter boil down to around 370 titles, which result in 331 unique WoS items. When mapped to our corpora, only two publications were missed by the reverse procedure, thus being almost completely in line with our matching.[6] Around 40 items not indexed in WoS are mainly articles in popular scientific or technological journals and blogs, as well as clinical trials and patents.



References [edit]

1. ^ Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. (March 2007). "**CRISPR provides acquired resistance against viruses in prokaryotes**". *Science*. 315 (5819): 1709–1712. Bibcode:2007Sci...315.1709B. doi:**10.1126/science.1138140**. PMID **17379808**.
2. ^ Pourcel C, Salvignol G, Vergnaud G (March 2005). "**CRISPR elements in Yersinia pestis acquire repeats by preferential uptake of bacteriophage DNA, provide additional tools for evolutionary study**". *Microbiology*. 151 (Pt 3): 653–663.
3. ^ Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh JR, Joung JK (March 2013). "**Efficient genome editing in zebrafish using a CRISPR-Cas system**". *Nature Biotechnology*. 31 (3): 227–229. doi:10.1038/nbt.2501.
4. ^ **Chylinski K, Jinek M, Fonfara I, Hauer M, Doudna JA, Charpentier E** (August 2012). "**A programmable dual-DNA-guided RNA endonuclease in adaptive bacterial immunity**". *Science*. 337 (6096): 816–821. PMID 22745249.
5. ^ **Doe J, Dae J, Bolotin A, Quinquis B, Sorokin A, Ehrlich SD** (August 2005). "**Clustered regularly interspaced short palindromic repeats (CRISPRs) have spacers of extrachromosomal origin**". *Microbiology*. 151 (8): 2551–2561.
6. ^ **Cong L, Ran FA, Cox D, Lin S, Barretto R, Hsu PD, Habib N, Wu X, Jiang W, Marraffini LA, Zhang F** (February 2013). "**Multiplex genome engineering use CRISPR/Cas systems**". *Science*. 339 (6121): 819–823. PMC 3795411.

Verbatim match for "**CRISPR provides acquired resistance against viruses in prokaryotes**", "**10.1126/science.1138140**", or "**17379808**".

Correctly matching "**CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies**" by **Pourcel**, **Salvignol**, and **Vergnaud** with a normalized edit distance ≤ 0.2 (0.072).

Incorrectly matching "**Efficient genome editing in plants using a CRISPR/Cas system**" by **Feng**, **Zhang**, **Ding**, **Liu**, **Yang**, **Wei**, **Cao**, **Zhu**, **Zhang**, **Mao**, and **Zhu** with a normalized edit distance ≤ 0.2 (0.118).

Looking for "**A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity**" by **Jinek**, **Chylinski**, **Fonfara**, **Hauer**, **Doudna**, and **Charpentier**. Author ratio score (1.0) and Jaccard index (1.0) ignore inversion of first and second author, whereas author order score penalizes it heavily (-0.479).

Looking for "**Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin**" by **Bolotin**, **Quinquis**, **Sorokin**, and **Ehrlich**. Author ratio score (1.0) misses incorrect author list, Jaccard index (0.667) and author order score (-0.413) penalize surplus authors and incorrect order respectively.

Looking for "**Multiplex genome engineering using CRISPR/Cas systems**" by **Cong**, **Ran**, **Cox**, **Lin**, **Barretto**, **Habib**, **Hsu**, **Wu**, **Jiang**, **Marraffini**, and **Zhang**. Author ratio score (1.0) and Jaccard index (1.0) are oblivious to change in order, author order score tolerates inversion of sixth and seventh author (0.865).

Figure 1: Examples showcasing the verbatim and fuzzy heuristics; divergent data underlined.

Results

Table 1 shows for both publication corpora the absolute numbers of matched publications, the relative numbers in relation to the sample corpora, and the precision of each method, calculated by manually checking the respective publications and matched references. The fuzzy matching heuristics D, E, and F generally identify more publications than the verbatim ones; at the cost of precision. The comparably

[6] Two items are not recovered; one of them because of an inconsistency between WoS online and raw database.

smaller number of matches of the fuzzy heuristics G, H, and I may be a result of the fact that author lists in the Wikipedia article are somewhat less well-maintained than identifiers and titles.

Table 1: Evaluation of the reference matching heuristics applied to the revision history of the CRISPR article, dependent on the two corpora of relevant CRISPR-related publications.

Reference Matching Heuristic	CRISPR Accounts Corpus			CRISPR WoS Corpus		
	Absolute	Relative	Precision	Absolute	Relative	Precision
Verbatim matching heuristics						
A Title	173	14.59%	1.000	130	26.00%	1.000
B DOI	178	15.01%	1.000	135	27.00%	1.000
C PMID	177	14.92%	1.000	135	27.00%	1.000
Fuzzy matching heuristics						
D `Title edit distance ≤ 0.2`	182	15.35%	0.984	137	27.40%	0.985
E `Title edit distance ≤ 0.3`	186	15.68%	0.952	140	28.00%	0.957
F `Title edit distance ≤ 0.4`	203	17.12%	0.847	152	30.40%	0.849
G `Title edit distance ≤ 0.4 + author ratio score = 1.0`	166	14.00%	0.976	126	25.20%	0.984
H `Title edit distance ≤ 0.4 + author jac-card index ≥ 0.8`	152	12.82%	0.987	118	23.60%	1.000
I `Title edit distance ≤ 0.4 + author order score ≥ 0.8`	162	13.66%	0.988	124	24.80%	1.000

Table 2: Left: Delay in days for each heuristic in relation to the earliest correct match. Right: Number of times a publication (rows) is matched earlier by another heuristic (columns).

Reference Matching Heuristic	Mean Relative Delay in Days		Median Relative Delay in Days		Comparison of Reference Matching Heuristics									
	Accounts	WoS	Accounts	WoS		A	B	C	D	E	F	G	H	I
A Title	69	70	0.0	0.0	A		10	47	1	1	2	1	0	0
B DOI	101	112	0.0	0.0	B	16		62	18	18	18	9	8	9
C PMID	40	51	0.0	0.0	C	19	26		20	20	21	12	4	5
D `Title edit distance ≤ 0.2`	67	67	0.0	0.0	D	2	9	49		0	1	0	0	0
E `Title edit distance ≤ 0.3`	68	67	0.0	0.0	E	2	9	49	0		1	0	0	0
F `Title edit distance ≤ 0.4`	67	67	0.0	0.0	F	2	8	47	0	0		0	0	0
G `Title edit distance ≤ 0.4 + …`	161	185	0.0	0.0	G	25	28	62	25	25	24		2	4
H `Title edit distance ≤ 0.4 + …`	160	173	36.0	32.5	H	52	56	81	53	53	50	34		2
I `Title edit distance ≤ 0.4 + …`	145	157	36.0	32.5	I	52	57	85	52	52	50	34	0	

A revision cites a publication if any of the matching heuristics correctly flags the publication in the revision. The number of uniquely matched publications resulting from both corpora is 201.

Table 2 (left) shows the mean and median delays for each method relative to the timestamp of the earliest correctly matched revision for both corpora. In 184 out of 1,186 and 138 out of 500 cases, respectively, the earliest match can successfully be identified using the three verbatim heuristics alone. Only in one case a publication is identified earlier using a fuzzy heuristic. In this case, it was due to a divergence between ‘^{2}-’ to ‘*beta-*’. While there is very little discrepancy in the number of matches between the verbatim methods, the delays in Table 2 indicate that the methods sometimes match at different times, e.g. PMIDs earlier than titles and DOIs. For the majority of entries, however, this does not matter, as can be seen from the median.

The matrix on the right side of Table 2 indicates for each matching heuristic (rows) how often another heuristic (columns) correctly identifies a publication in an earlier revision. The matrix depicts the results based only on the CRISPR Accounts Corpus and only takes correct matches into account. Since all matching is aimed at eliciting the earliest revision, a relaxed method might incorrectly select a revision to be the first one for a specific publication, which leads to a lower overall recall if cleaned for correctness. The title edit distance heuristic with the most relaxed threshold of 0.4 therefore seems more sensible than the more stringent methods using thresholds of 0.2 and 0.3, being superseded by the PMID methods in only 47 rather than 49 cases. This, however, is an expected consequence of the former's overall weaker precision of 172 correctly identified publications, as compared to 179 and 177 for the latter methods. Overall, the results of the combination of verbatim methods can hardly be improved by the relaxed methods, even with regard to delays.

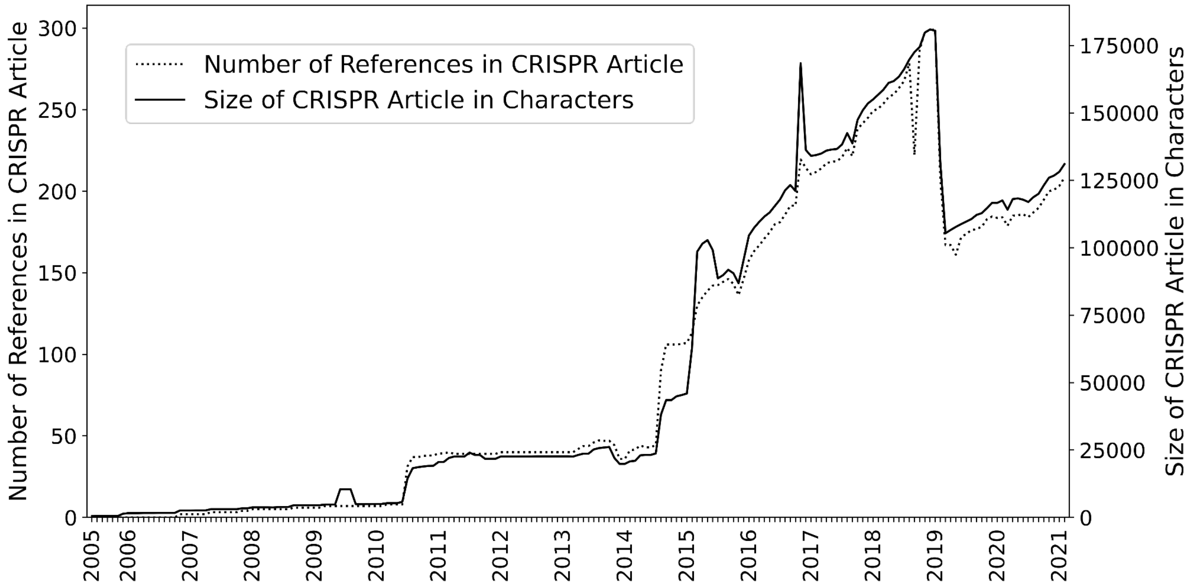


Figure 2: Smoothed dynamics of the text growth of the CRISPR Wikipedia article to the growth of the references.

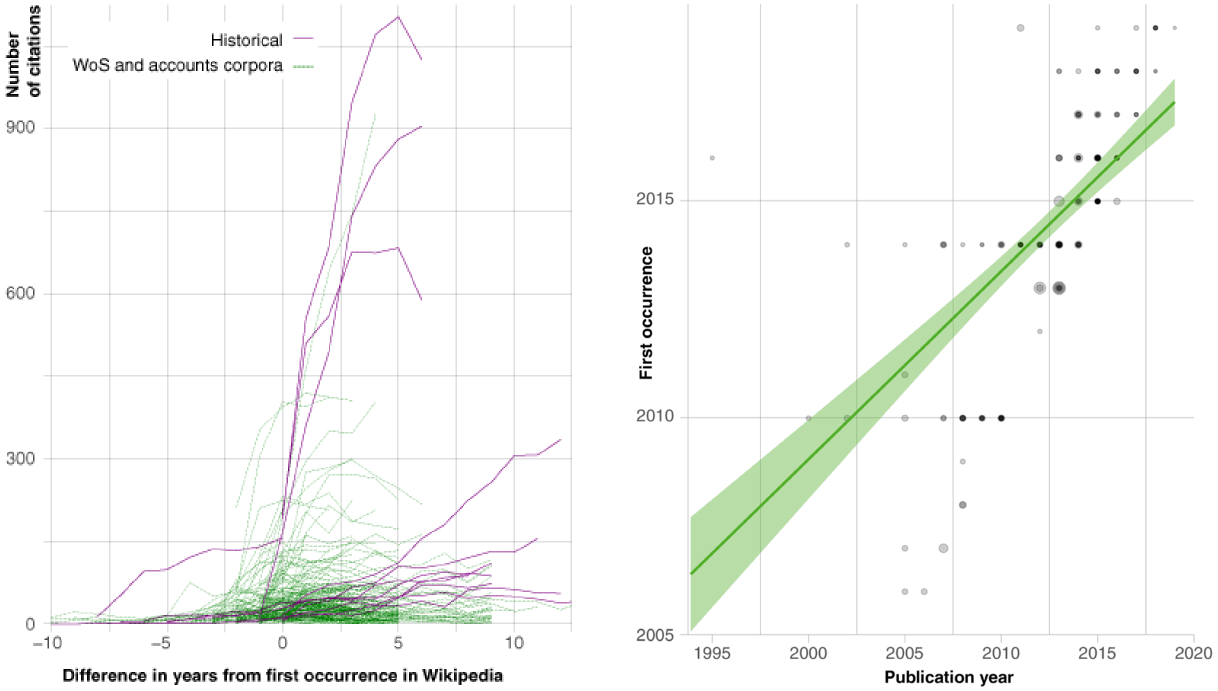


Figure 3: Citation distributions in relation to the occurrence in Wikipedia. *Left:* The matched publications (201) in yearly citation counts, with the date of first occurrence in the CRISPR article as baseline (so that citations before this date appear with negative numbers). Some prominent

cases—selected on the basis of citations by secondary accounts, as well as the references' occurrence in the current version of the History section—are marked by solid lines. *Right:* Publication years over years of first occurrence. Point sizes correspond to the number of citations in absolute numbers, overlapping points are displayed darker.

Both graphs in Figure 2 show a step in 2010. This step coincides with the introduction of a 'History' section in the article. Key publications result from the years 2011 to 2013, which is reflected in the growth of text and references from 2014 on-wards. In 2019, the application-related section is moved to a new article on CRISPR gene editing, resulting in a decrease of text and references. Similarly, Figure 3 on the right side shows that a number of publications dating from 2000 to 2010 were added in 2010 (possibly due to the newly created History section), and a similar phenomenon can be observed in 2014, corresponding to the dynamics in Figure 2. These patterns suggest that the delays with which publications are referenced on the Wikipedia page in our case may correspond to the general dynamics of Wikipedia's editors' take on this topic. The graph on the right side of Figure 3, however, shows that in a substantial amount of cases the first occurrence happens before the peak of the respective citation distribution.

Conclusion

We explored Wikipedia's usage of scientific literature in Wikipedia's article on CRISPR and the timeliness of its referencing patterns in order to gauge its relevance and adequacy as a medium for the representation and tracing of scientific innovations. More specifically, we proposed matching procedures to map from WoS publication corpora to all revisions of Wikipedia's central article on CRISPR. The results are promising: Initial evidence suggests that substantial portions of the CRISPR/Cas9 literature referenced in Wikipedia are highly cited or have been acknowledged by experts in the field. We observe that the currency of referencing improves over time, and that article editing dynamics is captured as well. For the CRISPR/Cas9 case we can give evidence that a combination of verbatim matching heuristics yields sufficient accuracy, thus making Wikipedia an interesting object for analyses of science communication in addition to standard bibliometric sources.

References

Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P. T., Mavergames, C., & Gruen, R. L. (2014). Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLoS Medicine*, *11*(2), e1001603.

Benjakob, O., & Aviram, R. (2018). A Clockwork Wikipedia: From a Broad Perspective to a Case Study. *Journal of Biological Rhythms*, *33*(3), 233–244.

Estevez, B., & Cukierman, H. (2012). The climate change controversy through 15 articles of Portuguese Wikipedia. Wikipedia Academy.

Garcia del Valle, E. P., Lagunes Garcia, G., Prieto Santamaria, L., Zanin, M., Menasalvas Ruiz, E., & Rodriguez Gonzalez, A. (2018). Evaluating Wikipedia as a Source of Information for Disease Understanding. 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), 399–404.

Giles, J. (2005). Internet encyclopaedias go head to head. Nature, 438(7070), 900–901.

Jemielniak, D., Masukume, G., & Wilamowski, M. (2019). The Most Influential Medical Journals According to Wikipedia: Quantitative Analysis. Journal of Medical Internet Research, 21(1).

Reavley, N. J., Mackinnon, A. J., Morgan, A. J., Alvarez-Jimenez, M., Hetrick, S. E., Killackey, E., Nelson, B., Purcell, R., Yap, M. B. H., & Jorm, A. F. (2012). Quality of information sources about mental disorders: A comparison of Wikipedia with centrally controlled web and printed sources. Psychological Medicine, 42(8), 1753–1762.

Simons, A., Kircheis, W., Schmidt, M., Potthast, M., Stein, B. (2021). Who are the Heroes of CRISPR? Priority Disputes on Wikipedia. Manuscript under review.

Zagovora, O., Ulloa, R., Weller, K., & Flöck, F. (2020). "I Updated the <ref>": The Evolution of References in the English Wikipedia and the Implications for Altmetrics. arXiv:2010.03083