Exploring open access coverage of Wikipedia-cited research across the White Rose Universities

The popular online encyclopaedia Wikipedia is an important and influential platform that assists with the communication of science to a global audience. Using data obtained from Altmetric.com and Unpaywall, we looked at research from the White Rose Universities (Sheffield, Leeds and York) that is cited on Wikipedia. Of that research, we explored what percentage of citations were available open access (OA) and the location of those citations to ascertain whether they were hosted by publishers or within OA repositories. This article explores the importance of access to OA research within such an important and leading platform as Wikipedia and how well it supports effective scientific communication across society.

Keywords
Wikipedia; open access; altmetrics; citations; preprints; open science; encyclopaedia

ANDY TATTERSALL
Information Specialist
School of Health and Related Research (ScHARR)
The University of Sheffield

NICK SHEPPARD
Open Research Advisor
University of Leeds

THOM BLAKE
Open Research Manager
Library, archives & Learning Services
University of York

KATE O'NEILL
Scholarly Communications Librarian
University Library
The University of Sheffield

CHRISTOPHER CARROLL
Reader in Systematic Review and Evidence Synthesis
School of Health and Related Research
The University of Sheffield

Introduction

The purpose of the work undertaken was to investigate how much of the research published by the Universities of Leeds, Sheffield and York is cited in Wikipedia and what proportion of those citations are linked to an open access (OA) version. We propose that making the cited academic literature point to OA versions as standard helps support the foundations of existing and future Wikipedia entries. Increasing the number of OA citations within Wikipedia not only assists the online encyclopaedia's goal of access to transparent and evidence-based knowledge but also removes any barriers to access to research, which ultimately is good for academics. We also explored to what extent this is being achieved using a sample of three UK universities, with the further intention of exploring which areas of the three institutions had the most citations across the various disciplines. In addition, we considered which were OA and whether access to them was available via the universities' or a third-party OA repository or via a publisher's website.

We chose the White Rose Universities of Leeds, Sheffield and York due to their shared OA repository, their long history of collaboration, research focus and because they are all members of the Russell Group of universities. White Rose Research Online (WRRO) is a cross-institutional OA repository that hosts research outputs from the universities of Leeds, Sheffield and York.[1] The purpose of the repository is to provide a long-term home for research outputs from the three universities and preserve research for posterity. The overall aim of the White Rose Research Online is to:

1. Provide a long-term home for research outputs from the three Universities, preserving research for posterity.
2. Provide OA to full-text research wherever possible.
3. Provide a reliable source of information about the universities' research.
4. Make research easier to find, bringing scholarly works to new audiences inside and outside academia.

We looked at the inconsistencies within Wikipedia and its open model and the implications of those for research dissemination. Not all research is truly open for the rest of society to access, but given Wikipedia's global reach and importance, it has become fundamental that the research underpinning each entry is as open and accessible as possible.

Brief background to Wikipedia

Wikipedia has become increasingly important to the academic community as a platform for engaging with society on topics relating to their own fields of research. Its open edit model means that researchers can cite their own or other relevant articles in any of Wikipedia's millions of pages. However, there are no formal checks as to whether such citations link to an OA or paywalled version of a research article – in many cases the research article is hosted on a paywalled publisher's website. Given Wikipedia's transparent model of publishing and editing, it seems rather counterproductive to their purpose only to link to versions of articles hosted on a paywalled publisher's website. However, Wikipedia does promote the use of the OABOT tool,[2] which facilitates making links to the OA versions of publications. The OABOT Wikipedia entry states. 'Our community does not prohibit or even discourage citing paywalled sources, but there is also absolutely no prohibition on surfacing OA versions alongside those citations, as long as the link does not violate any copyrights.'

'Wikipedia does promote the use of the OABOT tool, which facilitates making links to the OA versions of publications'

Wikipedia's three principal core content policies: neutral point of view, verifiability and not original research, mean that it does not publish original thought. All material in Wikipedia must be attributable to a reliable, published source.[3] Their final core policy is the most important as Wikipedia is built upon published knowledge that is already created and hosted elsewhere. Wikipedia remains an objective platform for the sharing of knowledge rather than opinion and conjecture. Thus, it becomes increasingly important that any cited evidence within a Wikipedia entry is auditable and open for all to read.

'All material in Wikipedia must be attributable to a reliable, published source'

Wikipedia and academia

Wikipedia has progressed since its early years and the reception for it in the academic community has warmed. One of the first news features on Wikipedia was in *Nature*, suggesting that editing the platform could be an influential way of improving a researcher's visibility and communicating their work to the academic community.[4] A randomized controlled trial found, 'Wikipedia doesn't just reflect the state

of the scientific literature, it helps shape it.'[5] A subsequent piece of interactive research encouraged final-year medical students to contribute to Wikipedia articles in return for academic credit.[6] More recent research, in 2019, looked at disease-related articles on Wikipedia and found that higher-quality articles were more likely to cite a Cochrane Review from the Cochrane Library than lower-quality articles on the encyclopaedia.[7] The authors used Wikipedia's definition of 'higher-quality articles' as those that have inline citations from reliable sources. Another piece of research found that a journal's accessibility (OA policy), as well as its academic status (journal impact factor), strongly increased the probability of it being referenced on Wikipedia.[8]

'Wikipedia doesn't just reflect the state of the scientific literature, it helps shape it'

Librarians have long been actively involved in editing Wikipedia, especially given their greater understanding of information literacy and OA. One such initiative took place at Washington State University, where they hosted a public Wikipedia edit-a-thon as part of OA Week in 2014.[9]

'Librarians have long been actively involved in editing Wikipedia'

The benefits of having academic work cited in Wikipedia

Research that explored the Web of Science database to identify and examine trends in the use of Wikipedia citations in scholarly peer-reviewed publications between 2002 and 2015 found that Wikipedia citations increased over that period for both non-OA and OA research articles.[10] Citations allow Wikipedia editors to make their contributions verifiable by supporting them with trustworthy sources and enable readers to locate further information on topics of interest.[11] Thus concluding that citations in Wikipedia can be considered an indication of the transfer of scholarly output to a wider audience.[12] There is also evidence that readers do follow links to the peer-reviewed sources that are cited in Wikipedia with data from Crossref demonstrating that in 2015/2016 it was the sixth highest referrer of Digital Object Identifier (DOI) resolutions.[13]

Research on wind power found a possible citation advantage of Wikipedia.[14] It transpired that research on this topic within the Web of Science, and cited on Wikipedia, obtained proportionally far more citations than articles not mentioned. However, there is no evidence to link Wikipedia with increased citations as they might simply be the better-quality articles, with the result that they get more citations in both Wikipedia and other sources. Another piece of research found that subjects the authors considered 'controversial', such as evolution and global warming, received more edits than 'non-controversial' topics such as the standard model in physics.[15]

'The benefits of having research cited on such a prominent platform as Wikipedia is somewhat negated when the source is not universally accessible'

The benefits of Wikipedia's citation of open access versions of research

There is very little previous research that explores how much research cited in Wikipedia is linked to an OA source. Some work has been carried out in this area but only for the library and information science field, which reported it at 31.2%, with this percentage increasing for more recent literature.[16] The benefits of having research cited on such a prominent platform as Wikipedia is somewhat negated when the source is not universally accessible and is behind a publisher's paywall. To some extent, this problem was brought to wider attention after the World Health Organization (WHO) and the Wikimedia Foundation collaborated on a project to expand the public's access to the latest and most reliable information about Covid-19.[17]

Earlier research also highlighted the merits of academics engaging with Wikipedia and that by working in a free, open environment, scholars can increase their potential readership exponentially. The researchers also concluded that authors could assure themselves that access is granted to individuals who might not have the opportunity to use print journals or expensive databases, thus fulfilling their role as keepers and disseminators of knowledge.[18] Work by Teplitskiy et al. indicated that, for OA research articles, Wikipedia is an increasingly useful means of disseminating science. Taking into account the field and impact factor, they found the odds of an OA journal being referenced on the English Wikipedia is 47% higher than paywall journals and concluded that this significantly amplified the diffusion of OA science, through an intermediary like Wikipedia, to a broader audience.[19]

'Wikipedia is an increasingly useful means of disseminating science'

Methods and data collection

A data request to Altmetric.com was submitted on 16 April 2019 for entries that included authors from any of the three White Rose Universities who are cited at least once in a Wikipedia entry. Data presented by Altmetric.com were tabulated with discipline data extracted from university systems. Wikipedia page entries and embedded citations were collected by Altmetric.com using unique identifiers within the research such as a DOI, PubMed ID or ISBN, this also included the date the research was cited within a Wikipedia entry. Further bibliographic data were captured that included publication title and date. Each individual Altmetric.com page corresponding to each Wikipedia citation was also obtained.

These entries are not unique, with some pieces of research having multiple Wikipedia citations. It is important to note how Altmetric.com captures multiple citations of the same article across several Wikipedia entries. A single Wikipedia entry can cite the same research article several times, but this does not alter the altmetric score for that piece of research. Regardless of how many Wikipedia citations a piece of research receives, it only counts as one to prevent academics from gaming the system and increasing their altmetric score.

Exploring the Altmetric.com data, we found that several Wikipedia entries were edited by the same editors. The origin of these editors is unknown – we can assume that they are either academics or professionals working in that particular field or citizen scientists with a vested interest in it. Further research in this area would be useful to discover the identity of the most productive editors and what patterns of editing they exhibit. Are they exclusively citing the same article or small group of articles across a variety of Wikipedia entries, and is there a pattern that shows the same author names are appearing? The latter may offer some insight into whether these entries are self-citations by the journal article authors.

We used the data to explore the number of Wikipedia citations by discipline for each of the three institutions. The data Altmetric.com supplies are only as good as the institutional and bibliometric journal it harvests. As a result, certain fields were incomplete, and we anticipate that, based on a previous study by some of the authors of this article, a percentage of the data in relation to institutional affiliation and date of publication will be inaccurate.[20]

Using Unpaywall to check for open access compliance

DOIs of all articles appearing in the Altmetric.com data that included a Wikipedia citation were subsequently run against the Unpaywall API (application programming interface) on the same day as they were received (16 April 2019). These include publishers of articles made immediately available under the 'gold' model of OA and institutional repositories like WRRO that make research articles openly available under the 'green' model. This typically means the author's accepted manuscript (AAM) before it has been finally typeset by the publisher and is often subject to an embargo period. Unpaywall provides a number of different services, including a browser plug-in that, if a user encounters a paywall, will link to an OA version where one is available. For the purposes of this study, the primary field of interest is designated as 'is_oa', which enables us to ascertain the proportion of

articles that are available OA (is_oa = TRUE) compared to those that are not (is_oa = FALSE). It is important to note also that any repository record that is under embargo at the time of data collection will return is_oa = FALSE. Whether the OA version is gold (under a Creative Commons licence) or green (with a more restrictive or no specified licence) is also significant, as Wikipedia citations to gold articles will necessarily be OA with no further intervention, whereas Wikipedia citations to articles in subscription journals may only be accessible directly from that citation if it includes the repository link. The repository link will need to be added manually, i.e. the DOI used to automatically generate a Wikipedia citation links to a closed access publisher's version. In some cases, a published version may be freely available on the publisher's platform but without an open licence present. While these outputs might not conform to some definitions of OA, Unpaywall works on an inclusive definition — 'OA articles are free to read online, either on the publisher website or in an OA repository' — giving these more ambiguous outputs the label 'bronze' OA.[21]

The final validated data were tabulated, and descriptive statistics were produced, and the implications of the data were discussed.

Sample validation

The tools used to collect and analyse data, Unpaywall and Altmetric.com itself, are largely automated while relying on data that has been added to Wikipedia manually. Therefore, it was decided to undertake a manual check of 100 Wikipedia citations from each of the three institutional datasets to check the accuracy of the data using an online random number generator to select a random sample of 100 citations from each institutional dataset. Selected records were manually checked to ensure data accuracy, with each record checked: firstly, to confirm that the attribution of the output to the University of Leeds, York or Sheffield was correct; and secondly, to confirm that the OA status given by Unpaywall is correct.

'Selected records were manually checked to ensure data accuracy'

Attribution of outputs was checked by comparison with the output itself as available online from the publisher. It was confirmed whether one or more of the authors listed on the output had recorded their institutional affiliation as the university covered by the dataset. Of the 300 Wikipedia citations checked, the affiliation could not be confirmed for seven of the outputs as the researchers could not access the output. Two attributions were not correctly identified by Altmetric.com; in both cases, the article listed the institutions where authors had gained their qualifications, and it appears that these were being picked up as affiliations. Of the 293 sample citations where an affiliation could be validated, 291 (99.3%) had been correctly attributed.

The OA status for cited articles was checked by accessing, where possible, the output through the publisher's platform and recording the licence conditions. A web browser in private mode was used so that institutional or individual access agreements could not affect the outcome. Where the output was not openly available through the publisher's platform, Google Scholar was used to identify versions of the output available through an OA repository. These versions were then checked for public availability of the output.

In the sample of 300 Wikipedia mentions picked up by Altmetric.com, 24 (8.0%) did not include a DOI and so the OA status was not available through Unpaywall. Of the OA statuses checked, 257 (85.7%) were confirmed to be correct on validation. Of the 276 citations for which an OA status was identified (discounting the 24 citations which returned no status), Unpaywall identified the correct status for 93.1%. This is similar to the precision of 96.6% found in a study by the developers of Unpaywall.[22]

Where discrepancies were found between the Unpaywall data sample and the manually checked OA statuses, the majority — 12 out of 19 — were articles not identified as open in Unpaywall but which, on manual checking, were found to be freely available as bronze OA.

That these outputs would be revealed by manual checking but not through Unpaywall could be explained by the ambiguity in the status of these records. Bronze OA is not easily identifiable through machine-readable metadata, and it is not known how consistent Unpaywall is in picking up these outputs. Another, perhaps more likely, explanation is the time difference between the data being collected and its validation. In the intervening period, publishers may have made outputs freely available online, either permanently or for a fixed period. The Covid-19 pandemic, which struck in the middle of this study, appears to have expanded this phenomenon, with publishers temporarily removing access restrictions on pertinent research outputs. As noted above, Unpaywall uses a broad definition of OA, but it is important for research designated as 'open' to remain so in perpetuity and carry an appropriate, irrevocable licence such as Creative Commons. This is a strong argument for articles designated as bronze being excluded from OA data.

'Bronze OA is not easily identifiable through machine-readable metadata'

The time difference probably also accounts for four outputs that were not found to be open in the Unpaywall data but were subsequently identified as openly available through a repository by checking manually. Repository content is regularly deposited or released from embargoes, so these discrepancies are to be expected.

Outputs for which OA was found in the Unpaywall data but not through manual checking were less common, with only three in the sample of 300 records, as shown in Table 1. For these outputs, data is given on the OA source found by Unpaywall, making it easier to understand the difference in results. For one of these three outputs, Unpaywall identified a version of an article that had been uploaded to a departmental webpage – this would not have been picked up by manual validation as it would not have been identified as an OA repository. Again, this points to the ambiguity that can exist in OA status as a result of judgements about what should be considered a legitimate source of OA content. Academic networking sites such as ResearchGate or Academia.edu provide another example of this ambiguity. These sites have repository-like functionality and may be indexed in Google Scholar, but they are not actively curated, and it is questionable whether they should be considered as legitimate sources of OA content.

'ambiguity that can exist in OA status as a result of judgements about what should be considered a legitimate source of OA content'

It is noticeable that no errors, positive or negative, were found in the data for gold OA articles made available under an OA licence. This attests to the more permanent and unambiguous status of these outputs.

	Sheffield	Leeds	York	Total
OA status not available through Unpaywall	7	12	5	24
Unpaywall OA status matched status found in manual checking	89	81	87	257
('Open' status in Unpaywall confirmed through manual checking)	(48)	(41)	(55)	(144)
('Not Open' status in Unpaywall confirmed through manual checking)	(41)	(40)	(32)	(113)
Unpaywall OA status did not match status found in manual checking	4	7	8	19
('Open' status in Unpaywall found to be incorrect on manual checking)	(0)	(0)	(3)	(3)
('Not open' status in Unpaywall found to be incorrect on manual checking)	*(4)*	*(7)*	*(5)*	*(16)*
Total	100	100	100	300

Table 1. Results of the manual validation of the sample of Unpaywall results

The validation of Unpaywall data highlights some of the challenges in determining and classifying OA status. Outputs made permanently open under an open licence may lend themselves to a conclusive analysis, but outside of this, there can be considerable ambiguity about how open an output is, and these statuses can change over time. Overall, there were no results that raised concerns about the general reliability of the Unpaywall OA data.

Results

In total, there were 6,454 citations of White Rose Universities' research on Wikipedia in the period 1922 to April 2019. Research from the University of Sheffield had 2,523 Wikipedia citations, which was marginally more than Leeds, with 2,406 citations. The University of York had 1,525 Wikipedia citations, as highlighted in Table 2.

The total number of items in each university's Altmetric.com databases were captured, excluding datasets and clinical trial records, as these received no Wikipedia citations and represent a very small percentage of the total items produced across the White Rose institutions. We included articles, books, chapters and news stories, although there was only one record of the latter, which originated in Sheffield and was a nature column piece.

Biological Sciences and Medical and Health Sciences overwhelmingly had the highest number of Wikipedia citations for each institution, as noted in Table 2. Whilst several disciplines were comparable across the institutions, some did much better than others. For example, Physical Sciences research from the University of Sheffield received considerably more Wikipedia citations than work in this field from Leeds or York. The University of Leeds Earth Sciences and Chemical Sciences research received much higher numbers of citations than the same categories from Sheffield or York. Despite fewer citations overall across the disciplines, York had more citations in History and Archaeology compared to Sheffield and Leeds. There were 642 Wikipedia citations that were not attributed to any discipline in the sample.

'Biological Sciences and Medical and Health Sciences overwhelmingly had the highest number of Wikipedia citations'

Discipline	Sheffield	Leeds	York
Mathematical Sciences	27	29	20
Physical Sciences	343	106	45
Chemical Sciences	66	115	60
Earth Sciences	102	251	47
Environmental Sciences	102	120	84
Biological Sciences	672	586	449
Agricultural and Veterinary Sciences	4	7	3
Information and Computing Sciences	59	49	36
Engineering	70	76	19
Technology	4	11	3
Medical and Health Sciences	535	516	304
Education	3	11	2
Economics	9	32	21
Commerce, Management, Tourism and Services	9	12	3
Studies in Human Society	60	73	19
Psychology and Cognitive Sciences	82	83	76
Law and Legal Studies	6	13	0
Studies in Creative Arts and Writing	2	3	2
Language, Communication and Culture	35	21	15
History and Archaeology	83	95	98
Philosophy and Religious Studies	5	15	4
No subject field identified	245	182	215
Total	2,523	2,406	1,525

Table 2. Wikipedia citations of White Rose Universities by discipline

Results across the three institutions were similar, with a little over half of all citations available OA and York performing marginally better than Sheffield and Leeds.

Of those outputs that were OA within Wikipedia, we found a very similar pattern across the three institutions when we explored where they were hosted, as highlighted in Table 3. Around one-third of these were found to have an OA version hosted on the publisher's

own platform. The remaining two-thirds did not have an OA version available through the publisher platform but did have a version available through an OA repository or had no OA host stated. These results were fairly consistent across the three institutions, with York returning the highest proportion of outputs openly available on the publisher's platform and Leeds the highest proportion of outputs openly available through a repository.

	Sheffield	Leeds	York
OA (best OA host is publisher)	584 (36%)	523 (33%)	357 (37.8%)
OA (best OA host is a repository)	303 (18%)	303 (19%)	167 (17.7%)
Not OA (no OA host stated)	754 (46%)	768 (48%)	420 (44.5%)
Total	1,641	1,594	944

Table 3. Unpaywall records (de-duplicated)

Table 4 presents no surprises by showing that journal outputs make up the largest proportion of outputs, given that the journal article is by far the dominant format to disseminate knowledge within academia. Reference entries were identified as the second most popular genre, making up no more than 3.3% of the overall total of outputs. We were unable to capture a universal description as to what a reference entry is, as it varies according to the specific publisher, and there is no consistent taxonomy used. It may refer to encyclopaedia outputs, journal articles and forms of grey literature.

'the journal article is by far the dominant format'

Genre	Sheffield	Leeds	York
Book	11 (0.7%)	5 (0.3%)	5 (0.5%)
Book Chapter	19 (1.2%)	15 (1%)	22 (2.3%)
Journal Article	1,565 (95.4%)	1,510 (94.7%)	866 (91.7%)
Monograph	1 (0%)	1 (0%)	0 (0%)
Posted content	0 (0%)	1 (0%)	0 (0%)
Proceedings Article	11 (0.7%)	7 (0.4%)	6 (0.6%)
Reference Entry	33 (2.0%)	52 (3.3%)	44 (4.7%)
No Genre stated	1 (0%)	3 (0.2%)	1 (0%)
	1,641	1,594	944

Table 4. Genre

The percentages presented in Table 4 do not add up to 100% due to rounding up percentages to their nearest first decimal point

The information available about what licences the Wikipedia-cited research was published under was limited, see Table 5. Research published under a Creative Commons licence was most notable, and the majority of the licensed works, with 532 published outputs, had a CC BY licence. There were 55 examples of CC BY-NC licensed research across the White Rose institutions and 93 items licensed under CC BY-NC-ND. Elsevier-specific open access licences accounted for 64 research outputs, and 58 items had open access implied. Most research outputs had no licence stated, accounting for 3,327 items. This high number is probably due to, historically, green OA records in a repository not having a licence. Five of the cited outputs were identified as being 'public domain' (PD), although it is unclear how this determination was made. In this context, public domain appears to reflect a lack of clear copyright attribution and, for practical and analysis purposes, should probably be treated the same as 'No licence stated'. The oldest publication that was available open access and cited in a Wikipedia entry was from 1910,[23] whilst the oldest paywalled research article was published in 1922.[24] It is noteworthy that publication data that is tracked in Altmetric.com appears to go back to as far as 1666.[25]

Best OA Licence	Sheffield	Leeds	York
acs-specific: author choice/editor's choice usage agreement	1 (0%)	3 (0.2%)	0 (0%)
CC 0	6 (0.4%)	2 (0.1%)	0 (0%)
CC BY	207 (12.6%)	179 (11.2%)	146 (15.5%)
CC BY-NC	18 (1.1%)	20 (1.3%)	17 (1.8%)
CC BY-NC-ND	29 (1.8%)	47 (3%)	17 (1.8%)
CC BY-NC-SA	9 (0.5%)	6 (0.4%)	8 (0.9%)
CC BY-SA	0 (0%)	0 (0%)	1 (0%)
CC BY-ND	1 (0%)	1 (0%)	0 (0%)
Elsevier-specific: OA user licence	21 (1%)	25 (1.6%)	18 (1.9%)
Implied-OA	18 (1.1%)	23 (1.4%)	17 (1.8%)
PD	3 (0.2%)	1 (0%)	1 (0%)
Oxford Academic licence	4 (0.2%)	0 (0%)	0 (0%)
Publisher-specific, author manuscript[26]	0 (0%)	2 (0%)	1 (0%)
No licence stated	1,324 (80.7%)	1,285 (80.7%)	718 (76%)
Total	1,641	1,594	944

Table 5. OA Licence

The percentages presented in Table 5 do not add up to 100% due to rounding up percentages to their nearest first decimal point.

Discussion

The way the three institutions performed with regards to how much of their Wikipedia citations content was OA was very similar. York did best with 56%, compared to Sheffield with 54% and Leeds with 52% of their citations available to freely read from Wikipedia citations. This was a positive sign and an indication of how OA is gaining popularity, but it also highlighted there is some way to go before all Wikipedia citations are fully available. A current limitation of that becoming possible is how much research is available OA via publisher websites or OA repositories – itself still well below 100% OA. The date of publication will also have an effect, as we might expect more recent articles to be OA with older published content behind a paywall or only available in print format. This is an area for further study. There is conflicting data as to how much UK research is open access with Research England[27] citing over 80% in their 2018 report, whereas the Curtin Open Knowledge Initiative uses data from public sources around the world to show that even though OA adoption was climbing, in 2018 it was only just above 70%.[28]

'Less than a fifth of all the cited OA outputs linked to a repository version'

Of those that were available OA, Unpaywall includes various data to establish whether an article is available from a publisher's website – likely to be gold, though may also be bronze OA – or from a repository and likely to be green. The results were very similar across the three institutions, with at least one-third of hosts being publishers. The remaining citations were either hosted in an OA repository or not available openly. Less than a fifth of all the cited OA outputs linked to a repository version. It should be noted that an article made OA via the green route may be available from more than one repository, not only WRRO, where co-authors are based at other universities, for example, and have deposited their manuscript in their own repository. When there are multiple locations, Unpaywall determines the best OA location, based on five ascending rules, to decide which is the most current, authoritative version.[29]

As we suspected, the vast majority of content we explored was published as journal articles. This was no surprise given that journal articles are the standardized format for disseminating quality research and provide virtually all citations within that medium. It should follow that citations, in relation to academic outputs, would follow that trend given the journal article's dominance in scholarly communications.

We explored what licences the outputs had been published under but, without manual checking of all the publications, it is impossible to get an accurate number. The most frequent of the Creative Commons licence outputs were published under the CC BY licence, which is the most dominant and accessible of licences, especially used within academic publishing. Only a small percentage were evident across the three institutions, with York performing the best with 15.5%. The Creative Commons NonCommercial and NonCommercial-NoDerivatives licences had a notably small percentage. Ideally, this study would have explored what disciplines they had been assigned to.

This research has identified that, at the time of data collection, approximately 53% of Wikipedia citations from the three White Rose Universities were available OA, whether gold, green or bronze. Nearly half of cited research was therefore inaccessible without subscription access. However, only that research available via gold or bronze would be immediately accessible to a user following a link from Wikipedia. Research articles available under the green route would need further intervention from a Wikipedia editor to link the repository version. Based on a random sample, green OA accounted for 17% of records, which is likely to be an under-estimate as some of the closed access records reported by Unpaywall may well be in a repository under embargo. Furthermore, there is no guarantee that bronze records without a defined licence will remain accessible in perpetuity. Taking all of this together, we can conclude that fewer than half of research articles cited on Wikipedia are currently linked to openly accessible records.

'Nearly half of cited research was … inaccessible without subscription access'

Given Wikipedia's unique role in the information ecosystem as a bridge between informal discussion and scholarly publication,[30] this is of concern. For example, during the Covid-19 pandemic, the WHO partnered with the Wikimedia Foundation to expand public access to current and reliable information about the virus,[31] while at the start of the Covid-19 pandemic many publishers temporarily made all research on the virus freely available.[32] Much of this research is likely to revert to closed access at some unspecified point in the future. The new requirements under Plan S,[33] which came into effect in January 2021 and aims to ensure full and immediate OA, should go some way to improving the situation. One result of Plan S is likely to be that more research will be available OA from publisher's websites under the gold route, which will not require further intervention from Wikipedia editors to link to a repository version. Another condition of Plan S is that AAMs deposited into repositories via the green route are not restricted by embargo and carry a CC BY licence. However, commercial publishers are resistant to this aspect of Plan S. In any case, there will still be a role for universities and their libraries to ensure Wikipedia is properly cited and that cited research is as widely accessible as possible.

One solution that has gained some traction in recent years is the hosting of Wikipedia edit-a-thons within universities. One of the three institutions involved in this research, the University of Leeds, has hosted its own Wikipedia edit-a-thons.[34] These sessions involve academics and librarians coming together at the same place and time to edit academic entries in their field of research with guidance from Wiki.

There are some limitations to this research that need to be considered. It only considers research from three specific universities, and the pattern may be very different at other types of institutions. Given the close relationship of York, Sheffield and Leeds and their shared repository, there may also be significant duplication of citation that has not been addressed. There are also limitations associated with data collection, precisely how Altmetric.com and Unpaywall work, for example, with potential misattributed affiliation or incorrect results from Unpaywall. This problem was highlighted by the work of Tattersall and Carroll, who especially noted an issue with incorrect institutional affiliations.[35] The inherently changeable nature of Wikipedia means that this is a snapshot at a specific point in time; results may be very different if data were collected today.

Conclusion

The aim of this research was to explore the open or closed nature of research citations in Wikipedia. We found varying degrees of openness on the encyclopaedia with the result that some disciplines are well represented on the knowledge platform, others much less so. We found that almost half of our research sample was not openly available for inspection by Wikipedia users wishing to dig deeper into certain topics. Given the potential value of such citations, not just to society but also to the academics, publishers and funders behind the work, there is value in seeking to increase the accessibility of these works. This can be achieved through greater awareness regarding Wikipedia's function as an influential and popular platform for communicating science. To take full advantage of this requires greater understanding within the academic and general public communities as to the importance of citing OA works over those behind a paywall.

Data accessibility statement
The data from this work can be accessed via The University of Sheffield's ORDA repository hosted by Figshare. There is no personal data or any that requires ethical approval. The data complies with the institution's policy on access and sharing. The data can be shared openly, and the file formats are open or commonly used. Headings and units are explained in the files. https://doi.org/10.15131/shef.data.12097797.v1

Abbreviations and Acronyms
A list of the abbreviations and acronyms used in this and other *Insights* articles can be accessed here – click on the URL below and then select the 'full list of industry A&As' link: http://www.uksg.org/publications#aa.

Competing interests
The authors have declared no competing interests.

References

1. "About White Rose Research Online," https://eprints.whiterose.ac.uk/about.html (accessed 23 December 2021).
2. "Wikipedia OABOT," https://en.wikipedia.org/wiki/Wikipedia:OABOT (accessed 23 December 2021).
3. "Wikipedia's Purpose," https://en.m.wikipedia.org/wiki/Wikipedia:Purpose (accessed 1 December 2021).
4. Eugenie Samuel Reich, "Online Reputations: Best Face Forward," *Nature* 473 (2011): 138–39, https://www.nature.com/articles/473138a (accessed 18 December 2021). DOI: https://doi.org/10.1038/473138a
5. Neil Thompson and Douglas Hanley, "Science Is Shaped by Wikipedia: Evidence From a Randomized Control Trial," MIT Sloan Research Paper No. 5238-17 (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3039505 (accessed 30 November 2021).
6. Amin Azzam et al., "Why Medical Schools Should Embrace Wikipedia: Final-Year Medical Student Contributions to Wikipedia Articles for Academic Credit at One School," *Academic Medicine* 92, no. 2 (February 1, 2017): 194–200, DOI: https://doi.org/10.1097/ACM.0000000000001381 (accessed 30 November 2021).
7. Arash Joorabchi, Cailbhe Doherty, and Jennifer Dawson, "'WP2Cochrane', a Tool Linking Wikipedia to the Cochrane Library: Results of a Bibliometric Analysis Evaluating Article Quality and Importance," *Health Informatics Journal*, 26, no 3 (2019), DOI: https://doi.org/10.1177/1460458219892711 (accessed 23 December 2021).
8. Misha Teplitskiy, Grace Lu, and Eamon Duede, "Amplifying the Impact of Open Access: Wikipedia and the Diffusion of Science," *Journal of the Association for Information Science and Technology* 68, no. 9 (2017): 2116–2127, DOI: https://doi.org/10.1002/asi.23687 (accessed 30 November 2021).
9. David Free, "Wikipedia Edit-a-Thon Part of Open Access Week at WSU," *College & Research Libraries News* 75, no. 11 (December 2014): 594, https://news.wsu.edu/press-release/2014/10/16/oct-21-wikipedia-edit-a-thon-part-of-open-access-week/ (accessed 23 December 2021).
10. Robert Tomaszewski and Karen I. MacDonald, "A Study of Citations to Wikipedia in Scholarly Publications," *Science and Technology Libraries*, 35, no. 3 (2016): 246-261, DOI: https://doi.org/10.1080/0194262X.2016.1206052 (accessed 30 November 2021).
11. Aida Pooladian and Ángel Borrego, "Methodological Issues in Measuring Citations in Wikipedia: A Case Study in Library and Information Science," *Scientometrics* 113, no. 1 (2017): 455–64, DOI: https://doi.org/10.1007/s11192-017-2474-z (accessed 30 November 2021).
12. Pooladian and Borrego, "Methodological Issues."; Kayvan Kousha and Mike Thelwall, "Are Wikipedia Citations Important Evidence of the Impact of Scholarly Articles and Books?," *Journal of the Association for Information Science and Technology* 68, no. 3 (March 2017): 762–79, DOI: https://doi.org/10.1002/asi.23694 (accessed 30 November 2021).

13. Joe Wass, "Where Do DOI Clicks Come From?," *Crossref* (blog), 2016, https://www.crossref.org/blog/where-do-doi-clicks-come-from/ (accessed 30 November 2021).

14. Antonio Eleazar Serrano-López, Peter Ingwersen, and Elias Sanz-Casado, "Wind Power Research in Wikipedia: Does Wikipedia Demonstrate Direct Influence of Research Publications and Can It Be Used as Adequate Source in Research Evaluation?," *Scientometrics* 112, no. 3 (September 1, 2017): 1471–88, DOI: https://doi.org/10.1007/s11192-017-2447-2 (accessed 30 November 2021).

15. Adam M. Wilson and Gene E. Likens, "Content Volatility of Scientific Topics in Wikipedia: A Cautionary Tale," *PLoS ONE* 10, no. 8 (August 14, 2015): e0134454, DOI: https://doi.org/10.1371/journal.pone.0134454 (accessed 30 November 2021).

16. Aida Pooladian and Ángel Borrego, "Methodological Issues."

17. "The World Health Organization and Wikimedia Foundation Expand Access to Trusted Information about COVID-19 on Wikipedia," WHO, https://www.who.int/news/item/22-10-2020-the-world-health-organization-and-wikimedia-foundation-expand-access-to-trusted-information-about-covid-19-on-wikipedia (accessed 30 November 2021).

18. Erik W. Black, "Wikipedia and Academic Peer Review: Wikipedia as a Recognised Medium for Scholarly Publication?," *Online Information Review* 32, no. 1 (2008): 73–88, DOI: https://doi.org/10.1108/14684520810865994 (accessed 30 November 2021).

19. Misha Teplitskiy, "Amplifying the Impact."

20. Andy Tattersall and Chris Carroll, "What Can Altmetric.com Tell Us About Policy Citations of Research? An Analysis of Altmetric.com Data for Research Articles from the University of Sheffield," *Frontiers in Research Metrics and Analytics* 2 (2018), https://www.frontiersin.org/articles/10.3389/frma.2017.00009/full (accessed 30 November 2021). DOI: https://doi.org/10.3389/frma.2017.00009

21. Heather Piwowar et al., "The State of OA: a Large-scale Analysis of the Prevalence and Impact of Open Access Articles," *PeerJ* 6: e4375, DOI: https://doi.org/10.7717/peerj.4375 (accessed 23 December 2021).

22. Heather Piwowar et al., "The State of OA."

23. "Overview of attention for article published in Philosophical Magazine Series 7, September 1910," Altmetric.com, https://www.altmetric.com/details/3156715 (accessed 23 December 2021).

24. "Overview of attention for article published in Zoological Journal of the Linnean Society, May 2008," Altmetric.com, https://www.altmetric.com/details/794381 (accessed 23 December 2021).

25. "Overview of attention for article published in Biodiversity Heritage Library, January 1666," Altmetric.com, https://www.altmetric.com/details/55862952 (accessed 23 December 2021).

26. "Publisher-specific, Author manuscript licence," *American Physical Society*, http://link.aps.org/licenses/aps-default-accepted-manuscript-license (accessed 6 January 2022).

27. Claire Fraser et al., "Monitoring Sector Progress towards Compliance with Funder Open Access Policies," 2018, 1–72, https://re.ukri.org/documents/2018/research-england-open-access-report-pdf/ (accessed 23 December 2021).

28. COKI Open Access Dashboard, https://openknowledge.community/dashboards/coki-open-access-dashboard/ (accessed 1 December 2021).

29. Richard Orr, "How is the best OA location determined?," unpaywall support portal, https://support.unpaywall.org/support/solutions/articles/44001943223-how-is-the-best-oa-location-determined- (accessed 1 December 2021).

30. Nick Sheppard and Martin Poulter, "Wikimedia and Universities: Contributing to the Global Commons in the Age of Disinformation," *Insights the UKSG Journal* 33, no. 1 (April 29, 2020): 14, DOI: https://doi.org/10.1629/uksg.509 (accessed 1 December 2021).

31. WHO, "The World Health Organization and Wikimedia."

32. "Publishers Make Coronavirus (COVID-19) Content Freely Available and Reusable," *Wellcome Trust*, https://wellcome.org/press-release/publishers-make-coronavirus-covid-19-content-freely-available-and-reusable (accessed 1 December 2021).

33. Sheppard and Poulter, "Wikimedia and Universities."

34. Plan S, https://www.coalition-s.org/S, (accessed 1 December 2021).

35. Tattersall and Carroll, "What Can Altmetric.com Tell Us"

Corresponding author:
Andy Tattersall
Information Specialist
School of Health and Related Research (ScHARR)
The University of Sheffield, GB
E-mail: a.tattersall@shef.ac.uk
ORCID ID: https://orcid.org/0000-0002-2842-9576

Co-authors:
Nick Sheppard
ORCID ID: https://orcid.org/0000-0002-3400-0274

Thom Blake
ORCID ID: https://orcid.org/0000-0001-5507-9738

Kate O'Neill
ORCID ID: https://orcid.org/0000-0002-2013-0502

Christopher Carroll
ORCID ID: https://orcid.org/0000-0002-6361-6182

To cite this article:
Tattersall A, Sheppard N, Blake T, O'Neill K and Carroll C, "Exploring open access coverage of Wikipedia-cited research across the White Rose Universities," *Insights*, 2022, 35: 3, 1–13; DOI: https://doi.org/10.1629/uksg.559

Submitted on 17 June 2021 Accepted on 29 September 2021 Published on 02 February 2022

Published by UKSG in association with Ubiquity Press.