




Uncovering alternative metrics: data mining Wikipedia for evidence of public engagement and impact

Peter Neish

Program Manager, Stewardship and Open Research
The University of Melbourne

peter.neish@unimelb.edu.au

 <https://orcid.org/0000-0002-7171-2334>

Sally Tape

Open Research Support Specialist
The University of Melbourne


stape@unimelb.edu.au

 <https://orcid.org/0000-0003-4477-0050>

Lars Alvik

Digital Curation Technical Specialist
The University of Melbourne

lars.alvik@unimelb.edu.au

 <https://orcid.org/0009-0003-9942-6297>

Abstract:

As the world's largest reference website, Wikipedia is one of the main ways knowledge is disseminated outside of traditional publications and news reporting. In this paper, we investigate how GLAM institutions have engaged and measured impact with Wikipedia and some of the tools used. We look at the structure of content in Wikipedia and ways of accessing data of various qualities and we describe some preliminary coding experiments that test data mining in Wikipedia to examine how research is being cited and referenced in Wikipedia.

First published 9 July 2024



This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International Licence](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Introduction

Wikipedia

“Imagine a world in which every single person on the planet is given free access to the sum of all human knowledge. That's what we're doing” Jim Wales (‘Wikipedia: Prime objective’ 2023)

English Wikipedia is one of the most popular websites in the world, with around 10 billion visits per month (Semrush 2024) and is the largest reference website in the world (Figure 1). When searching for a known person or event, Wikipedia is frequently at the top of Google searches and information from Wikipedia is integrated into sites such as Facebook, YouTube, Google and Twitter (X). In contrast to more traditional scholarly writing or news reporting Wikipedia pages potentially have greater reach and impact because they are written in summary style and in layman’s terms (‘Wikipedia: Writing better articles’ 2024). Wikipedia follows the FAIR data principles (Wilkinson et al. 2016) with content that is Findable (as demonstrated by Wikipedia search results), Accessible (Wikipedia content can be readily accessed by anyone for free in multiple languages), Interoperable (Wikipedia content is created in standard markup; projects such as Wikidata allow powerful querying and integration), and Reusable (Wikipedia is released under a Creative Commons Attribution-ShareAlike 4.0 International License for anyone to reuse).

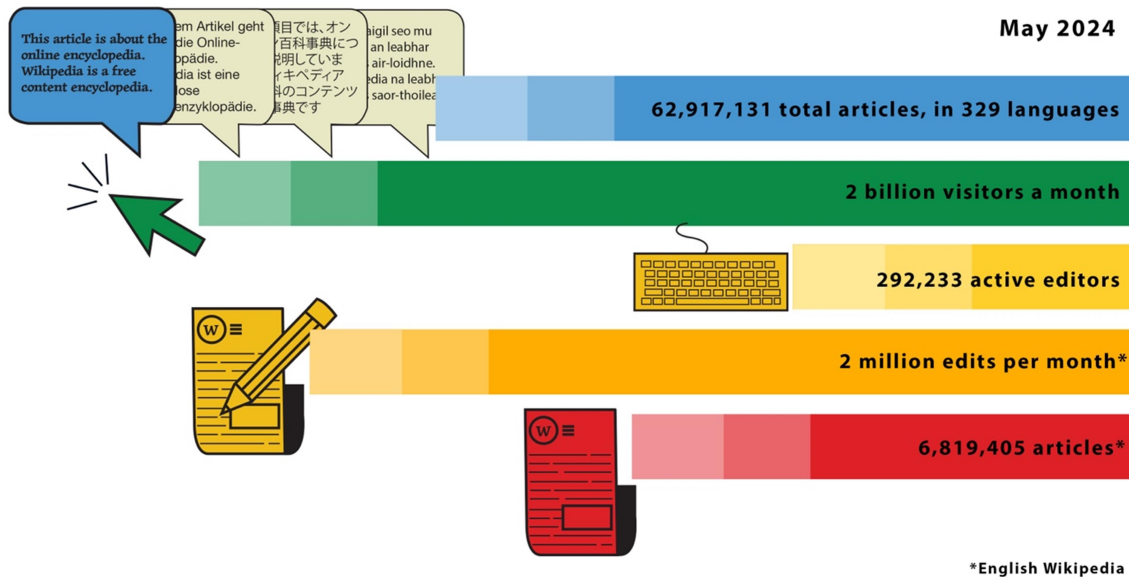


Figure 1: Numbers in Wikipedia (from https://meta.wikimedia.org/wiki/List_of_Wikipedias & https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia)

Over the years, trust in Wikipedia has grown and Wikipedia is now widely perceived as trustworthy (Bruckman 2022; Barnett 19 February 2018). Steinsson (2024) goes further and calls Wikipedia a “proactive debunker, fact-checker and identifier of fringe discourse”. Many studies have been conducted that compare Wikipedia content to traditional encyclopaedias or other references (‘Reliability of Wikipedia’ 2024) and usually compare articles from a particular domain. For example, Reavley et al. (2012 p. 1753) examined the quality of mental health articles compared to a standard

psychiatry textbook and the *Encyclopaedia Britannica* and found that Wikipedia was “as good as or better than *Britannica* and the standard text book”. Another study found that pharmacology articles were 99.7% accurate when compared with a standard pharmacology textbook, with completeness of such information at 83.8% (Kräenbring et al. 2014).

Wikipedia is not without its problems¹. Editors in English Wikipedia are historically self-selected and are predominately white males with a bias towards western perspectives (Oeberst and Ridderbecks 2024). Initiatives have made some progress on addressing these issues (‘Wikipedia: WikiProject Women in Red’ 2024; Bjork-James 2021).

Trustworthiness in Wikipedia is boosted through reliable secondary sources that are used as references or citations to verify facts. We see Wikipedia as an enticing prospect for a researcher or institution seeking to generate engagement and impact from their research activities. Experts can contribute their knowledge for the common good and there is often a strong alignment with institutions moving to be more open with data and research outputs. Institutions are often interested in content that is related to notable people or subjects connected the institution. References supporting Wikipedia content demonstrate the value of disseminating research more widely. But how is research being used in Wikipedia? Are references in Wikipedia comparable to published citations in their impact? How can we get a high-level picture of engagement happening through Wikipedia? This paper will examine methods to address these questions. Firstly, however, we will give a quick overview of the current metrics landscape.

Metrics and Alternative Metrics

For many years, the success of research and a researcher’s career has been measured by the number of publications “published in peer-review, indexed, high impact journals” (Butler et al. 2017 p. 164). The traditional way of measuring these metrics has been to count citation impact data or other bibliometrics as a measure of impact for research. “Bibliometric indicators can be derived from different levels of aggregation: single publications, the publications of a researcher, a research institution, a scientific journal, or a whole country” (Bornmann et al. 2016 p. 42).

The idea of measuring the successful outcome of a research project based on engagement with peers and often intimate academic circles has, in recent years, been challenged with the introduction of alternative metrics. Alternative metrics measure the societal impact of research. Whereas citation impact data only allows the measurement of the impact of research on research itself, alternative metrics allow a measurement of (public) engagement with research output (Bornmann et al. 2016).

Alternative metrics are generated when a research output receives a ‘mention’ online. Mentions are tracked using persistent identifiers (PIDs) for both traditional and non-traditional outputs. Non-traditional research outputs (NTROs) are generally considered to be artistic, creative, and practice-based works, as well as other assets produced during a research project such as reports, presentations, media etc. Code, software, and data are sometimes also considered NTROs (Pearce et al. 2023).

Alternative metrics demonstrate public interest and engagement in research activity. This allows the measurement of research activity to move beyond academic networks (peer-to-peer citation) to be more inclusive of community, government and industry opinions and interests. Alternative metrics can be applied to a range of different asset types and outputs, both traditional and non-traditional. Sharing assets and outputs from right across a research project means there are more outputs that can be shared online, cited, and referenced, leading to more engagement and research impact.

As a side note, the term alternative metrics is often shorted to altmetrics. This can be confusing, as it is the name of a commercial provider *Altmetric*, which provides a dashboard and graphical displays of alternative metrics through their *Altmetric Explorer* product. While mostly tracking blog posts, social media and news reporting, some Wikipedia metrics are available through *Altmetric Explorer*ⁱⁱ.

Pitfalls of Metrics and Alternative Metrics

We have seen that standard metrics are based on the publication of research outcomes as determined by a traditional publishing process that often includes peer review. Adding alternative metrics as a tool to measure the engagement and distribution of research allows the measurement of research to extend to include public opinion and usage.

Both traditional and non-traditional methods for measuring the impact of research outputs are not without their problems. The peer review and journal publishing process often delays the release of research outcomes. This can slow the impact and benefits research outcomes can have for communities and governments. Further to this, the publication of research outputs alone does not guarantee immediate engagement, and research may take 'more than five years to become popular in the scientific' (Bornmann et al. 2016) and other communities once it has been published.

Through social media, alternative metrics are delivered in a real time environment, meaning the metrics generate a more immediate measure of engagement and impact. Alternative metrics are measured using an attention score. As Elmore (2018, p 254) states "the score is helpful to rank research outputs based on attention from various sources, but it can't tell you anything about the quality of the article itself. It simply tracks attention, and attention can be good or bad". The like, swipe, post culture of social media makes it easy for information to spread rapidly. This spread does not necessarily mean the information is true, real or has been adequately researched and may simply be the result of a topic that is trending or an algorithm. In contrast, references in Wikipedia are used to verify facts and have been peer reviewed by other Wikipedia editors. Statements and references in Wikipedia articles persist through the consensus of Wikipedia editors.

While online and social behaviours can influence the number of engagement-metrics being generated, the functional design of a platform can also impact the data being collected. In 2018, a study looking into the most cited sources across all Wikipedia's language editions was completed (Matsakis 2018). This study shows the most cited reference on all of Wikipedia to be the journal article *Updated world map of the Köppen-Geiger climate classification*.ⁱⁱⁱ Interestingly, the authors of the paper had no

idea that their article was being cited to this extent. At the time of the study (2018), the article had received 2,830,341 citations across Wikipedia. As of April 2024, the *Altmetric Explorer* database^{iv} shows this article as having received a total of 13,113 citations on 13,014 Wikipedia pages. A clear discrepancy in citation numbers is evident. The discrepancy is explained by *Altmetric* as possibly being a limitation to the number of languages tracked by the *Altmetric Explorer* database, citations only being traced if they appear in the reference section of a page and a citation only being counted once even if it appears multiple times on a page. Whether it is 2 million or 13,000, Wikipedia's citation metrics for this publication demonstrate high impact in comparison to traditional publication citations for the same article currently sitting at around 7,500.

Data Mining Wikipedia

Wikipedia is supported by multiple Application Programming Interfaces (APIs) that allow data to be queried and extracted. The related project, Wikidata, provides additional structured data that supports Wikipedia. These systems provide the ability to interrogate aspects of Wikipedia to gain insights into the references being used to verify information. This presents an opportunity for institutions to get a better understanding of the Wikipedia ecosystem, and particularly the role of references and how they underpin the reliability of information presented in Wikipedia. Ultimately, an institution may be able to uncover some of the most impactful engagement happening via the Wikipedia platform.

Wikidata

Wikidata provides structured data that supports all the different language Wikipedias^v. Data in Wikidata can be reused multiple times across Wikipedia, for example, for info boxes and references. Every Wikipedia page has a corresponding Wikidata item that contains the structured data about the topic of the page. All Wikipedia pages, regardless of the language, use the same Wikidata item. Like Wikipedia, anyone can edit Wikidata and many 'bots' have automatically created Wikidata items and connections that help semantically link these entities to each other and to data sources across the internet (see Figure 2). Wikidata can act as a source of truth for an item and include alternative spellings and languages as well as connections to any persistent identifiers applicable for that item.



provides images

provides structured data



Wikimedia Foundation, CC BY-SA 3.0, via Wikimedia Commons

Figure 2: Structured Data (from Wikimedia Foundation, CC BY-SA 3.0, via Wikimedia Commons. <https://commons.wikimedia.org/wiki/File:Commons-logo.svg> & <https://commons.wikimedia.org/wiki/File:Wikidata-logo-en.svg>)

Wikimedia Commons

Reuse of images is another alternative metric that is not currently captured by existing systems. Wikimedia Commons (or Commons for short) is a repository for media (images, sound, and video) that can be reused across Wikipedia and related projects. Like Wikipedia, content in Commons is available under a licence that allows for reuse. Many institutions have found that use of their digital content increases enormously if they make it available through Commons with an open licence and this usage can be easily measured.

A highly successful example was the State Library of Queensland, which released 50,000 of its images on Wikimedia Commons in 2010. Since then, these images have been used in over 6,000 Wikipedia articles (Wikimedia Australia 2023b) and the subject of much related publicity. Since then the library has been an active partner with Wikipedia, including four Wikipedian in Residence Programs, the most recent being the First Nations Wikipedian in Residence Bianca Valentino (Wikimedia Australia 2023a). Other organisations have done the same. For example, the National Library of Norway images have been used across over 14,000 articles^{vi}.

The University of São Paulo uploaded a modest set of around 650 images from their Museum of Veterinary Anatomy. While this is a relatively small number of images, they now collectively receive on average 1.7 million page views a month^{vii}.

Existing Tools

Not surprisingly, tools already exist that help track the usage of Wikipedia in different ways – the GLAM Wiki project provides a list of tools on their project page^{viii}. The GLAM Wiki project supports GLAM institutions to engage with Wikipedia in different ways including uploading digital content and adding content to Wikipedia.

While these tools give an insight into overall usage within Wikipedia, and especially usage of Commons files, it is harder to understand how external research is being

cited in Wikipedia. This is something we wanted to explore and document using code, as we describe below.

Data mining code

The huge number of references and citations in Wikipedia represents a good opportunity for large-scale data mining. However, there are some pitfalls. Because of the organic growth of Wikipedia, with a range of different contributors, there have been many ways of adding references and citations with different templates and reusing references within the same article. This makes the job of accurately mining reference data from Wikipedia more challenging.

Wikipedia also has many ways of retrieving data. There is an API with several relevant methods, for example *linksearch*,^{ix} which returns all external links from a Wikipedia article. This works regardless of how the references were set up by the contributors. It is also possible to bulk download all of Wikipedia, either the source code behind the article (wikitext) or as parsed HTML pages (which would have a standardised output). These different ways of accessing the data come with their own drawbacks; for example, bulk downloads are large (several hundreds of GB) and going through all pages using the API is a laborious task that would probably take several weeks.

Our first approach was to try to use the *linksearch* method of the API to find all references to DOIs and record those. One of the downsides of this approach is that even if there are multiple references to the same DOI within an article, the API will only count them as one single link. This could have been used with other methods to refine the results; however, the sheer number of DOI links within Wikipedia seems to have overwhelmed the API, which failed to return any links.

The approach we settled on (see Figure 3) was to use the category system to narrow down the number of articles and search one category of interest at a time. We wrote a script that searches through and returns all articles recursively within the category Biotechnology and its subcategories. We can use that list as the basis for a script to download the parsed HTML for each article and search for DOIs within the text. This approach picks up all DOIs, regardless of whether they were used in a footnote or as a literature list at the end of the article.

We then use the captured DOIs to query the Crossref API^x to extract more information about the DOIs in question. For example, the affiliation of the researchers, type of publication and publication date.

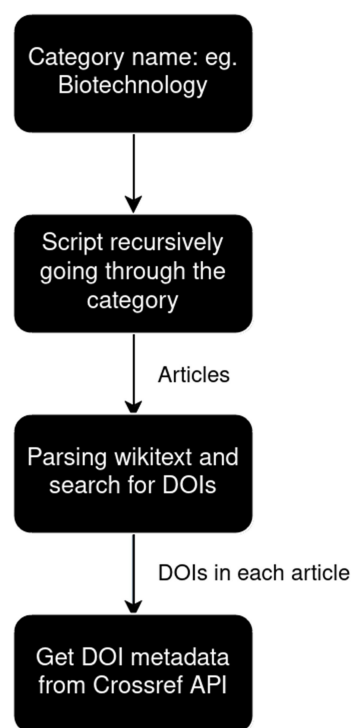


Figure 3:
Our data mining workflow

Results

The code we have developed as part of this project has started to give us an insight into the use of references in Wikipedia. We ran searches on individual categories (including 3 levels of sub-categories) as a way of limiting the results to a manageable subset. The references with DOIs extracted from each of these searches can be seen in Table 1. In addition we were able to use the Crossref and Datacite APIs to extract authors, affiliations and ORCIDs where they available.

Table 1: Reference details extracted from Wikipedia pages for different categories

Category	Wikipedia pages	Pages with DOIs	Authors listed in DOIs	Authors with affiliations	Authors with ORCIDs
History of Australia	1255	243 (19%)	1526	446 (29%)	108 (7%)
Geology of Australia	888	283 (32%)	3649	980 (27%)	368 (10%)
Politics of Australia	2768	279 (10%)	1228	290 (24%)	82 (7%)
Biota of Australia	5961	2641 (44%)	15267	2691 (18%)	954 (6%)
Total	10872	3446 (32%)	21670	4407 (20%)	1512 (7%)

In our small sample, only about one in three Wikipedia pages contains a DOI-based reference. This was not uniform across the different categories we chose to examine. Australian politics has the lowest, perhaps because politics is mostly referenced by current news articles, whereas the Biota of Australia had the most, possibly reflecting the work that has been done connecting information about organisms to Wikipedia and Wikidata - see for example Page (2022).

Obtaining a DOI is the first step but is not a guarantee that useful metadata can be extracted. When we subsequently examined the author information and affiliation details, only one in five authors had an affiliation listed, and in many cases the affiliation was incorrect. There is also much inconsistency in the affiliation, which is a free text field rather than a link to a persistent identifier. ORCIDs did not assist much here either as only 7% of authors listed an ORCID in the metadata.

We wanted to provide some tools to complement the existing community tools that tend to be necessarily narrow in scope and have limited application and tend to focus only on data within Wikipedia itself. We have tried to expand on the usefulness of these tools by connecting to registries of published information, such as Crossref and DataCite. Our work is ongoing; for the latest version please see our code repository^{xi}.

Discussion and Conclusion

Investigating something as large and as complex as Wikipedia is always going to be challenging. The free-text nature of the collaboratively built site presents challenges in extracting text and persistent identifiers. APIs go some way to helping, but the multitude of ways links can be used and the limitations of the *linksearch* API meant that calls to this API were not particularly successful. In Wikidata, the proportion of articles with DOIs present is insufficient. The lack of uniform ways of citing from Wikidata is also limiting the usefulness. Despite this, we persisted in extracting a

subset of data from Wikipedia, and have undertaken some preliminary analysis and visualisation of results that will be further expanded at our code repository.

Despite the difficulty of extracting data about references, the benefits for GLAM institutions in releasing their material through Wikimedia Commons are clear. Nearly 1 in 5 images uploaded to Wikimedia Commons by the State Library of Queensland now appear in Wikipedia articles^{xii}, and the images from the Museum of Veterinary Anatomy from the University of São Paulo were viewed on average 1.7 million times per month^{xiii}.

Our goal over the coming months is to expand on our code base to enable extraction and analysis of citation data from Wikipedia. Trust in Wikipedia is enhanced through citations of reliable sources and the more we can get a handle on citation data, the more we can understand how individual institutions contribute to building of the largest open knowledge base of our time.

References

- Barnett D (19 February 2018) 'Can we trust Wikipedia? 1.4 billion people can't be wrong | The Independent', *The Independent*, accessed 17 April 2024, https://www.independent.co.uk/news/long_reads/wikipedia-explained-what-is-it-trustworthy-how-work-wikimedia-2030-a8213446.html, accessed 17 April 2024.
- Bjork-James C (2021) 'New maps for an inclusive Wikipedia: decolonial scholarship and strategies to counter systemic bias', *New Review of Hypermedia and Multimedia*, 27(3):207–228, <https://doi.org/10.1080/13614568.2020.1865463> .
- Bornmann L, Marx W and Haunschild R (2016) 'Calculating journal rankings: Peer review, bibliometrics, and alternative metrics?', in C Sugrue and S Mertkan (eds) *Publishing and the Academic World*, Routledge.
- Bruckman AS (2022) *Should You Believe Wikipedia?: Online Communities and the Construction of Knowledge*, Cambridge University Press, Cambridge, <https://doi.org/10.1017/9781108780704> .
- Butler JS, Sebastian AS, Kaye ID, Wagner SC, Morrissey PB, Schroeder GD, Kepler CK and Vaccaro AR (2017) 'Understanding Traditional Research Impact Metrics', *Clinical Spine Surgery*, 30(4):164–166, <https://doi.org/10.1097/BSD.0000000000000530>.
- Elmore SA (2018) 'The Altmetric Attention Score: What Does It Mean and Why Should I Care?', *Toxicologic Pathology*, 46(3):252–255, <https://doi.org/10.1177/0192623318758294>.
- Kräenbring J, Penza TM, Gutmann J, Muehlich S, Zolk O, Wojnowski L, Maas R, Engelhardt S and Sarikas A (2014) 'Accuracy and Completeness of Drug Information in Wikipedia: A Comparison with Standard Textbooks of Pharmacology', *PLOS ONE*, 9(9):e106930, <https://doi.org/10.1371/journal.pone.0106930>.
- Matsakis L (2018) *The Authors of Wikipedia's Most-Cited Source Had No Idea*, *WIRED*, <https://www.wired.com/story/wikipedia-most-cited-authors-no-idea/>, accessed 30 April 2024.
- Oeberst A and Ridderbecks T (2024) 'How article category in Wikipedia determines the heterogeneity of its editors', *Scientific Reports*, 14(1):740, <https://doi.org/10.1038/s41598-023-50448-y>.
- Page RDM (2022) 'Wikidata and the bibliography of life' *PeerJ* 10:e13712 <https://doi.org/10.7717/peerj.13712>
- Pearce G, Clift J, Neish P, Ratana P and Reeson T (2023) 'FAIR and Open Non-Traditional Research Outputs Project Report', <https://zenodo.org/records/8429350>, accessed 3 May 2024.
- Reavley NJ, Mackinnon AJ, Morgan AJ, Alvarez-Jimenez M, Hetrick SE, Killackey E, Nelson B, Purcell R, Yap MBH and Jorm AF (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, <https://doi.org/10.1017/S003329171100287X>.

'Reliability of Wikipedia' (2024) *Wikipedia*, https://en.wikipedia.org/w/index.php?title=Reliability_of_Wikipedia&oldid=1218576118, accessed 17 April 2024.

Semrush (2024) *Top Websites*, <https://www.semrush.com/trending-websites/global/all>, accessed 5 April 2024.

Steinsson S (2024) 'Rule Ambiguity, Institutional Clashes, and Population Loss: How Wikipedia Became the Last Good Place on the Internet', *American Political Science Review*, 118(1):235–251, <https://doi.org/10.1017/S0003055423000138>.

Wikimedia Australia (2023a) *Announcing our First Nations Wikipedian In Residence*, *Wikimedia Australia*, https://wikimedia.org.au/wiki/Announcing_our_First_Nations_Wikipedian_In_Residence, accessed 26 April 2024.

— (2023b) *State Library of Queensland*, *Wikimedia Australia*, https://wikimedia.org.au/wiki/State_Library_of_Queensland, accessed 26 April 2024.

'Wikipedia: Prime objective' (2023) *Wikipedia*, https://en.wikipedia.org/w/index.php?title=Wikipedia:Prime_objective&oldid=1159856445#cite_note-1, accessed 17 April 2024.

'Wikipedia: WikiProject Women in Red' (2024) *Wikipedia*, https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_Women_in_Red&oldid=1217240425, accessed 5 April 2024.

'Wikipedia: Writing better articles' (2024) *Wikipedia*, https://en.wikipedia.org/w/index.php?title=Wikipedia:Writing_better_articles&oldid=1217144521, accessed 5 April 2024.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hoofstede R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J and Mons B (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3(1):160018, <https://doi.org/10.1038/sdata.2016.18>.

Endnotes

- i https://en.wikipedia.org/wiki/Criticism_of_Wikipedia
- ii <https://www.altmetric.com/solutions/altmetric-explorer/>
- iii <https://doi.org/10.5194/hess-11-1633-2007>
- iv https://www.altmetric.com/explorer/highlights?identifier=10.5194%2Fhess-11-1633-2007&scope=institution&show_details=3112735
- v https://en.wikipedia.org/wiki/List_of_Wikipedias
- vi [https://glamtools.toolforge.org/glamorous.php?doit=1&category=Nasionalbiblioteket&use_globalusage=1&depth=3&projects\[wikipedia\]=1](https://glamtools.toolforge.org/glamorous.php?doit=1&category=Nasionalbiblioteket&use_globalusage=1&depth=3&projects[wikipedia]=1)
- vii <https://glamwikidashboard.wmcloud.org/MAV>
- viii <https://outreach.wikimedia.org/wiki/GLAM/Resources/Tools>
- ix <https://www.mediawiki.org/wiki/API:Exturlusage>
- x <https://api.crossref.org/swagger-ui/index.html>
- xi <https://doi.org/10.26188/25727712>
- xii <https://glamwikidashboard.wmcloud.org/SloQ/usage>
- xiii <https://glamwikidashboard.wmcloud.org/MAV>