BIG DATA & SOCIETY

Commentary

Big Data & Society
July–December 2015: 1–5
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/2053951715614332
bds.sagepub.com
SAGE

Wikipedia, sociology, and the promise and pitfalls of Big Data

Julia Adams[1] and Hannah Brückner[2]

Abstract

Wikipedia is an important instance of "Big Data," both because it shapes people's frames of reference and because it is a window into the construction—including via crowd-sourcing—of new bodies of knowledge. Based on our own research as well as others' critical and ethnographic work, we take as an instance Wikipedia's evolving representation of the field of sociology and sociologists, including such gendered aspects as male and female scholars and topics associated with masculinity and femininity. Both the gender-specific dynamics surrounding what counts as "notability" on the online encyclopedia and Wikipedia's relative categorical incoherence are discussed. If "Big Data" can be said to construct its own object, it is, in this instance, a curious and lop-sided one, exemplifying pitfalls as well as promise with respect to more accurate and democratic forms of knowledge.

Keywords

Wikipedia, sociology, gender, Big Data, knowledge, encyclopedia

Since its inception in 2001, Wikipedia has grown into one of the most frequently visited online platforms in the world. The English version alone attracts over 8 million views per hour.[1] Wikipedia includes some 32 million articles in 287 languages (as of August 2014). The current incarnation of Wikipedia (without revisions and talk pages) takes up 44 GB; the complete version "expands to multiple terabytes of text."[2] Aside from its size, which exceeds the venerable Britannica and Brockhaus by an order of magnitude, and its ambitions for exhaustiveness, Wikipedia is not like other encyclopedias: everyone can contribute to its content—presumably resulting in a grassroots, democratic documentation of knowledge, accessible to anyone with an internet connection, for free. Hence it promises access—to such power as knowledge may grant—to those who were traditionally excluded from it. Conversely, Wikipedia is a potential window into knowledge produced at the margins, in places far removed from the established centers of learning in the Western world, but perhaps for that reason even more enlightening. In addition, new insights and new information about the world were to be available practically immediately, without gate-keeping by experts employed by the business to guard the integrity of content and lengthy publishing processes; instead, quality control is conceptualized to emerge from deliberation among contributors.

Both production and consumption of Wikipedia content are intimately connected to questions of democracy. Good democratic decision-making requires informed citizens; and in order to make good decisions, those citizens require accessible and complete information that includes and respects all societal groups and interests. Wikipedia has the potential to generate and deliver such information—a scholar's "real utopia" (Wright, 2011)—but it exemplifies the pitfalls as well as the promise of the impact of technology on society and culture.[3] Contributing to Wikipedia requires some technical wherewithal that falls outside the skill set of

[1]Yale University, New Haven, CT, USA
[2]NYU-Abu Dhabi, Abu Dhabi, United Arab Emirates

Corresponding author:
Julia Adams, Yale University, 493 College Street, New Haven, CT 06511, USA.
Email: julia.adams@yale.edu

the average Internet user. What is more, beyond the technical skills, negotiating the interaction of editors on Wikipedia requires mastery of a particular jargon and rules of conduct that have evolved in online communities, skills that are similarly not easily acquired by newcomers (Jemielniak, 2014). In fact, the typical Wikipedia contributor is a computer-savvy white male in his thirties, residing in the US or Europe (Wikimedia Foundation, 2011).

Aside from questions of access, the literature on the English Wikipedia clearly shows that decisions about content and form are now made top-down and not bottom-up and that new users (and some Wikipedia veterans as well) may encounter a hostile environment that is not always open to their contributions (Baker, 2008; Schneider et al., 2014). These exclusionary practices evolved with Wikipedia itself, in part, grounded in good intentions that included quality control and content regulation. They are also the outcome of struggles among Wikipedia editors over status and influence. Over time, Wikipedia's policies have generated a maze of contradictory rules that are not transparent to newcomers and are enforced by experienced editors often furthering their own interests and agendas (Jemielniak, 2014). Our personal favorite is the ''professor test'' that contains guidelines for notability of academics.[4] These guidelines read in part like a faculty handbook for promotion to tenure, and are similarly often unhelpful in assessing the notability of a person once contested.

Wikipedia, for all of its quirks, is an important topic of study. It forms people's frames of reference, and the younger they are, the more it does so (these days in tacit collaboration with primary and secondary school teachers) and the more they are influenced. It is therefore increasingly important to understand how it works. For academics, more generally, Wikipedia is an entree into better understanding how new bodies of knowledge emerge and are recognized as such. An innovative way of forming knowledge—crowdsourcing—applied to encyclopedia content is a dramatic departure. Yet unlike software engineering, the crowdsourcing of encyclopedic content has no built-in quality control—such as compilation of code into a functioning piece of software—that ensures that contributions aggregate into a complete, coherent and high quality end product. Concerns over the quality of Wikipedia articles have inspired numerous attempts to provide scholarly assessment, precisely because Wikipedia's promise hinges on the extent to which it provides unbiased and error-free content (Flekova et al., 2014; Gilles, 2005; Halfaker et al., 2011).

Equally interesting and important, recent scholarship about the English-language Wikipedia has focused on policies and technologies that emerged in response to the need to resolve questions of quality control, as well as controversies among contributors and users about content (Elvebakk, 2008). Such controversies are almost never resolved through voting among the users. Instead, deliberations continue until a resolution is found or imposed. Wikipedia developed both a rich set of policies that govern the scope of contributors' activities and a hierarchy of contributors—ranging from powerful administrators and arbitrators to newbies who are limited to commenting and suggesting changes to controversial content. Protracted controversies may eventually be resolved by the persistence of the users who have the greatest stakes rather than by those with the best arguments; and the history of deliberations preserved in ''editing history'' and ''talk'' pages may result in great quantities of fragmented text that are hard for less experienced users to access and comprehend (Jemielniak, 2014).

These dynamics limit the extent to which knowledge documentation on Wikipedia can be said to be democratic. But they also pose severe problems for the success of quality control. Higher positions in the Wikipedia user hierarchy are achieved through building a reputation for one's contributions—but as always, quantity is more easily measured than quality, and thus users who are technologically savvy and build software that helps them to make large numbers of small automatic or semi-automated edits are more likely to succeed than those who focus instead on the scholarly aspect of editing. Furthermore, editors with the greatest stakes in a topic may not always be those who get it right; and compromises may be made for articles that are awkwardly written and incoherent. Finally, Wikipedia is increasingly a combined formation of knowledge, even explicitly including imported old-style encyclopedia materials that may improve coverage but may sit unfinished and uneasily in the midst of user-grown categories and content.

Our overall project, in a nutshell, examines the way that Wikipedia represents male and female scholars, and topics associated with masculinity and femininity, in the social sciences, humanities, and sciences. In this short piece, however, we focus not only on gender but also on the generally curious shape of one discipline, sociology, as it has taken shape on the English-language Wikipedia. What is the form and substance of knowledge about sociology and sociologists that is forming on Wikipedia, then? As might be expected based on the preceding remarks about crowd-sourced encyclopedic content, it is a somewhat incoherent and amorphous creature. The Wikiproject sociology,[5] formed in response to just such a diagnosis in 2004 with support from the American Sociological Association, identifies around 5600 Wikipedia pages as pertaining to sociology, and ranks them according to importance and quality. A look at the titles of the 76 Wiki pages ranked ''top importance'' yields a heterogeneous bunch that not only includes Karl Marx, Max Weber, social class, "man," and culture, but also boyfriend, evil, murder, jihad, and, mysteriously, infection in child care.

The sociology portal[6] has, among many other features, a list of "branches of sociology" that contains crowd psychology, criminology and demography, as well as sociobiology and socio-musicology, while references to gender appear under sociobiology (subtopic "biology of gender"), under social problems ("gender inequality") and under "sociology stubs," subtopic "Gender studies journal stubs." Right next to socialization, one finds "sociological genres of music" with a well-developed list of categories pertaining to "marching bands" such as pep, pipe, scramble bands, and so forth. This jumble resembles Jorge Luis Borges' fantastic list of animal categories, attributed to a "certain Chinese encyclopedia," The Celestial Emporium of Benevolent Knowledge, or the wildly "heterogene subdivisions" by which Borges notes that the Bibliographic Institute of Brussels "exerts chaos."[7]

At the least, these examples illustrate the difficulty of generating and maintaining a coherent system of categories in the Wikipedia context as a consequence of crowdsourcing. Traditional encyclopedias do not provide coherent systems of categories either, of course, and desiring coherent category mapping in a context of Wikipedia's form of crowdsourcing may seem like wanting to have one's cake and eat it too. On the other hand, a technology that enables mapping and maps of knowledge through extensive linking of related content is one of the promising features of Wikipedia, which make it that much more useful than the traditional alphabetically ordered encyclopedic list of terms, and our project departed from the assumption that the specific mapping that occurs on Wikipedia is sociologically meaningful and interesting with respect to the formation of knowledge. It remains to be seen what, if any, sociological insights might emerge from exploring the history of a project like the sociology portal.

Our own analysis of the 452 living "American Sociologists" listed on Wikipedia in August 2014 yields a mixture of notable academics (about 60%; although quite a few of them are actually active in adjacent disciplines), of social activists and social workers, the occasional motivational speaker, and a surprising number of not particularly notable sociologists of religion who teach in a very few schools in the United States. Surprising here is not so much that someone creates a page for people and objects that have little connection to the discipline; rather, that these creations seem to escape the scrutiny of the crowd, or pass the professor test uncontested.

Others fare worse, such as the unfortunate Denise Donelly, whose Wikipedia page became embroiled in a controversy about the term "involuntary celibacy" that some people wanted to write about, while others deemed it nonsense. Consider the following excerpt from a discussion about deleting a Wikipedia page devoted to her:[8]

> Strong delete. Article was created against previous consensus and is based on zero independent, secondary biographical sources. The reason is simple: she is not notable.
>
> Keep—per DGG. A topic receiving a tenured position at a major university is not generally a fringe theory.
>
> Comment. I have to agree with other editors that the work done by the subject is not "Fringe". The problem is that notability according to WP: Prof is not yet achieved: too early.

While in other passages on this page some substantive arguments can be found, the notability discussion is largely based on assertions.[9] Similarly, a discussion of two editors about the idea that the work of Donelly is "fringe", and therefore not fit to include in Wikipedia, rapidly devolves into a skirmish devoid of substantive arguments:

> Nah, given the WP: Fringe guideline, I can't at all see how the topic of involuntary celibacy is not fringe.
>
> It's not fringe because it doesn't "depart significantly from" (i.e. contradict in any way) the views of mainstream academics. It's far more controversial among Wikipedia editors than it is in academia. If you asked 100 sociologists whether the topic deserves to be studied, wouldn't you honestly expect 90+ of them to say yes?
>
> Nah, it's WP: Fringe; plain as day, that it is.
>
> Oh, well if it's plain as day, then I apologize. Your mere assertion was one thing, but your reassertion really leaves no room for doubt.
>
> Yep. All thanks to my several years of Wikipedia experience. Apology accepted.
>
> I may not be a person of your incomparable eminence in these hallowed servers, but even I know that Wikipedia is governed by consensus, and refusing to engage in dialog isn't the best way to achieve it.

While one might find some of these encounters on Wikipedia amusing, there is more and more evidence that the outcome of such contests aggregates into an online world that is structured in a way that favors the tastes of some over the inclusion of all. Our own preliminary findings suggest that women sociologists are relatively underrepresented on Wikipedia; other sources show that the digital public sphere is largely dominated by men (Martin, 2015; Pierson, 2015).[10] Furthermore, reports of online harassment of women are piling up; it is suspected that the underrepresentation of women is partially caused by unfettered online misogyny (Buni and Chemaly, 2014), which drives at least some women offline altogether. Some suggest a similar pattern of exclusion and harassment of people of color. Although the extent to which women suffer harassment disproportionately is controversial, and not much systematic data is available, one might generalize cautiously to say that the cyberworld favors a mode of interaction that exposes many, especially young people, to abusive and threatening online interactions (Pew Research Center, 2014). Recent work on Wikipedia also suggests that the number of editors is on the decline, and that new editors are driven away by the hostility of the established crowd.

What, then, does all this suggest about knowledge formation online, and about the vices and virtues of "Big Data"? First of all, by all means, "Big Data" is certainly sociologically interesting, in part because of its very indeterminacy. While we do not know where Wikipedia is ultimately heading, we do know that the representation of at least one academic discipline, sociology, is a peculiar one, both in terms of its rendition of male and female scholars and its version of scholarship more broadly. It more closely resembles the wild categorical systems in "The Analytical Language of John Wilkins," the Borges story cited above (Borges, [2000] 1937), than it does Borges' image of the scientifically precise map designed by mythical Cartographers' Guilds who "struck a Map of the Empire whose size was that of the Empire, and which coincided point by point for it" (Borges and Casares, 1998).[11] Do these peculiarities dovetail with the more general tendencies of development we have identified? We do not know yet. But our reading of the literature and our data suggest that a critical analysis of Big Data directed toward social interaction online, and focused particularly on patterns of exclusion, will be more illuminating than an uncritical stance that conceptualizes the cyber world as an unfiltered expression of real world social trends.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors gratefully acknowledge support for this research from the National Science Foundation (grant #1322971).

Notes

1. As of 31 December 2014, according to http://stats.wikimedia.org/EN/Sitemap.htm (accessed 3 February 2015).
2. See http://en.wikipedia.org/wiki/Wikipedia:Database_download (accessed 11 August 2014).
3. Wikipedia itself, according to https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not#Wikipedia_is_not_a_democracy, is not a democracy, and much of the discussion below confirms this self-assessment. But note the concept of democracy used by Wikipedians here refers only to voting (rather than reaching a consensus) as a mechanism for decision-making about content. The passage is silent on bottom-up formation of content.
4. https://en.wikipedia.org/wiki/Wikipedia:Notability_(academics) (last accessed 19 August 2015).
5. http://en.wikipedia.org/wiki/WikiProject_Sociology (accessed 19 August 2015).
6. https://en.wikipedia.org/wiki/Portal:Sociology (accessed 19 August 2015).
7. Both are from Borges' "The Analytical Language of John Wilkins." The Chinese Encyclopedia famously features in Michel Foucault's preface to *The Order of Things: An Archaeology of the Human Sciences* (1994 [1970]). Regarding the chaotic hand of The Bibliographic Institute of Brussels, thus Borges: "it has divided the universe into 1000 subdivisions, from which number 262 is the pope; number 282, the Roman Catholic Church; 263, the Day of the Lord; 268 Sunday schools; 298, mormonism; and number 294, brahmanism, buddhism, shintoism and taoism. It doesn't reject heterogene subdivisions as, for example, 179: "Cruelty towards animals. Animals protection. Duel and suicide seen through moral values. Various vices and disadvantages. Advantages and various qualities."
8. https://en.wikipedia.org/wiki/Wikipedia:Articles_for_deletion/Denise_Donnelly (accessed 19 August 2015).
9. We cite this passage not to illustrate that Wikipedia's notability standards are problematic (as noted above, they are not more problematic than the standards used by the average tenure and promotion review committee); rather, the point is that the discussion is not based in arguments that engage the substance of the scholar(ship) under scrutiny (perhaps not unlike some discussions in tenure review committees).
10. It should be emphasized that traditional encyclopedias may not fare much better in this respect (and in other issues of content quality) than Wikipedia, which, after all, has the great advantage of being freely accessible to anyone in its entirety.
11. Incidentally, the map of the world as seen through Wikipedia may be largely terra incognita. See http://www.theguardian.com/technology/2009/dec/02/wikipedia-known-unknowns-geotagging-knowledge (accessed 24 September 2015).

References

Baker N (2008) The charms of Wikipedia. *The New York Review of Books* 55(4): 1–10.

Borges JL and Casares AB (1998) On exactitude in science. *Collected Fictions,* trans. Hurley A. New York, NY: Penguin.

Borges JL (2000 [1937]) The analytical language of John Wilkins. In: Borges JL (ed) *Selected Non-Fictions*, trans. Weinberger E. London: Penguin, pp. 229–232.

Buni C and Chemaly S (2014) The unsafety net: How social media turned against women. Available at: http://www.theatlantic.com/technology/archive/2014/10/the-unsafety-net-how-social-media-turned-against-women/381261/ (accessed 4 August 2015).

Elvebakk B (2008) Philosophy democratized? *First Monday* 13(2). Available at: http://firstmonday.org/ojs/index.php/fm/article/view/2091 (accessed 22 October 2015).

Flekova L, Ferschke O and Gurevych I (2014) What makes a good biography? Multidimensional quality analysis based on Wikipedia article feedback data. In: *Proceedings of the 23rd international conference on World Wide Web*, Seoul, Korea, 7–11 April, pp. 855–866.

Foucault M (1994 [1970]) *The Order of Things: An Archaeology of the Human Sciences*. New York, NY: Vintage Books.

Gilles J (2005) Internet encyclopaedias go head to head. *Nature* 438: 15.

Halfaker A, Kittur A and Riedl J (2011) Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. In: *Proceedings of the 7th international symposium on wikis and open collaboration*, New York, pp. 163–172.

Jemielniak D (2014) *Common Knowledge? An Ethnography of Wikipedia*. Stanford, CA: Stanford University Press.

Martin F (2015) Getting my two cents worth in. *ISOJ* 5(1). Available at: https://isojjournal.wordpress.com/2015/04/15/getting-my-two-cents-worth-in-access-interaction-participation-and-social-inclusion-in-online-news-commenting/ (accessed 19 August 2015).

Pew Research Center (2014) Online harassment. Available at: http://www.pewinternet.org/2014/10/22/online-harassment/ (accessed 19 August 2015).

Pierson E (2015) Outnumbered but well-spoken: Female commenters in the New York Times. *CSCW'15*, 14–18 March. Available at: http://dx.doi.org/10.1145/2675133.2675134.

Schneider J, Gelley BS and Halfaker A (2014) Accept, decline, postpone: How newcomer productivity is reduced in English Wikipedia by pre-publication review. In: *Proceedings of the international symposium on open collaboration*, Berlin, Germany, 27–29 August, p. 26.

Sewell WH (1992) A theory of structure: Duality, agency, and transformation. *American Journal of Sociology* 98(1): 1–29.

Swidler A and Arditi J (1994) The new sociology of knowledge. *Annual Review of Sociology* 20: 305–329.

Wikimedia Foundation (2011) *Wikipedia Editors Study: Results From the Editor Survey*. Available at: http://upload.wikimedia.org/wikipedia/commons/7/76/Editor_Survey_Report_-_April_2011.pdf (accessed 1 February 2015).

Wright EO (2011) A call to duty: ASA and the Wikipedia initiative. *ASA Footnootes* 39(8): 1–20.

This article is part of a special theme on *Colloquium: Assumptions of Sociality*. To see a full list of all articles in this special theme, please click here: http://bds.sagepub.com/content/colloquium-assumptions-sociality.