Wikipedia and gender: The deleted, the marked, and the unpolluted biographies

David Ramírez-Ordóñez
Universitat Oberta de Catalunya
Barcelona, Spain
davidramirez@uoc.edu

Núria Ferran-Ferrer
Universitat de Barcelona
Barcelona, Catalonia
nferranf@ub.edu

Julio Meneses
Universitat Oberta de Catalunya
Barcelona, Spain
jmenesesn@uoc.edu

ABSTRACT

Wikipedia is the self named *free encyclopedia*, available in more than 300 languages and one of the most popular websites on the Internet [12]. Despite its mission of collecting the sum of all knowledge, one of Wikipedia's struggles is its gender bias [24]. In this paper we present a proposal of the corpus for analysis of the generation of biographies, written in the English Wikipedia, in order to identify the gender bias in the creation of new content to reflect the new *valid knowledge* of all human beings.

First we identify a mechanism to access a corpus of deleted biographies and those which have been accepted into the category *Articles for Deletion*, where editors vote to keep, merge, redirect or delete content in an online debate. Then we access a different set of data, a second corpus from the category *Scientist by field* in which we have chosen biographies marked as content to be improved due to its lack of bibliographic references and those which have never been marked for improvement. To do so, we focused on the area of science, in the first case, with the category *Articles for Deletion* we selected scientists, and in the second case, with the category *Scientists by field* we selected STEM scientists, in order to compare how gender affects the development of content in Wikipedia. Lastly we propose a path to understanding the generation of the gender gap in the collaborative creation of shared content, this entails a close up look at the policies and guidelines of the digital encyclopedia, such as *notability* and *reliable sources*, created by the community of editors to shape the type of content accepted as valid knowledge.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative interaction**; **Collaborative content creation**.

KEYWORDS

Gender gap, Gender bias, Equality, Content production, Digital inequality, Open content, Biographies by gender, Biographies by field, Notability, Reliable sources

Wiki Workshop 2022, April 25-26, 2022, Online, hosted by London, UK

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
David Ramírez-Ordóñez, Núria Ferran-Ferrer, and Julio Meneses. 2022. Wikipedia and gender: The deleted, the marked, and the unpolluted biographies. In *Wiki Workshop 2022: A forum bringing together researchers exploring all aspects of Wikimedia projects. Held virtually at The Web Conference 2022, April 25-26, 2022, Online, hosted by London, UK*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Wikipedia is aware of its gender bias problem. Jimmy Walles (one of its creators) proposed to solve it several years ago and then admitted to have failed to do so [31]. Sue Gardner exposed nine reasons that explain why women don't edit Wikipedia, such as lack of time or the avoidance of the competitive culture experienced at the debates [19]. This is not only a problem of the English Wikipedia, it has also been observed in other languages as well. For instance, the percentage of female editors of the Spanish Wikipedia' represents a mere 11.6% of the total editors [25].

The gender bias in Wikipedia presents as a problem of three different kinds: unequal participation of its editors, limited content related to women available, i.e. relatively few biographies on Wikipedia are about women and topics of interest to women are less well-covered; and lower female readership. Editing Wikipedia is a "Boy's Club" matter [21, 25] Its contents underrepresent women's biographies [30] and the gender gap in content may also include different kinds of biases like race, class, sexual orientation or ethnicity [10], in other words, overlaps known as intersectionalities [16]. Regarding readership, the gender bias among Wikipedia e-readers happens because two-thirds of them on any given day are men [20]. In this paper we focus on the gender content bias, specifically in the content creation and deletion process, which is part of the editing process, that determines what is valid knowledge and what is not.

2 BACKGROUND

The creation of new content in Wikipedia is ruled by its policies and guidelines [1]. Interestingly, the five basic pillars result in a large number of rules that create what Italo Calvino refers to as an "anti-language". That is to say, a technical jargon used by and for experts [15]. One of Wikipedia' s five basic pillars states that Wikipedia should be written from a neutral point of view. The *neutral point of view* is also one of the three core content policies, the others being *verifiability* and *no original research*. Of these core content policies we are particularly interested in two guidelines that explain what these policies mean: *notability* and *reliable sources*. However, the difference between policies and guidelines, "is obscure", as Wikipedia itself states [9].

Another of the basic pillars is that Wikipedia's content is free and anyone can use, edit and distribute it. In regards to that ethos, , Ford and Wajcman wrote an interesting analysis on how Wikipedia is rooted in a male dominant culture of the encyclopedic tradition and the free software community [18]. The final of the five basic pillars says that Wikipedia has no firm rules. We can get a sense of the complexity in Wikipedia, but we'll focus on the *notability* and the *reliable sources guidelines* in order to limit the scope of this study.

2.1 Notability

One of Wikipedia's rules that determine the decision of what is published as knowledge in the encyclopedia is the Notability guideline. This guideline refers to the topics "that have gained sufficiently significant attention by the world at large and over a period of time" and take evidence from reliable and independent sources. Notability avoids the indiscriminate inclusion of topics and archives topics that are "worthy of notice" [5] known in Wikipedia jargon -antilanguage for Calvino- as *WP:N*. The problem stands with who defines it and what are its limits.

There are contents clearly identified for deletion, such as copyright infringements [32] . But when editors are unsure about immediate deletion, for example, with regards to an article's notability, they can use the *Articles for Deletion* category to decide if it should be included in the encyclopedia [27, 33]. This category sends the articles for evaluation over a period of at least 7 days, during which the community of editors may vote to keep the article, delete, redirect or merge it with another(among other kinds of votes) and are required to provide arguments to explain their decision. Once that time frame is over, the administrators, and rarely other editors, [26] end the debate or choose to extend it for a further 7 days by relisting it -*WP:RELIST*- [3]. The debate within this category is aimed at achieving community consensus (*WP:CON*) which is not unanimity, nor the sum of votes [8]. The deletion process via *Articles for Deletion* can be different for each Wikipedia Language. For example, in the English Wikipedia, anonymous (unregistered) and new users are permitted to vote [3] meanwhile, in the Catalan Wikipedia, a vote is valid only from editors that have registered for more than 30 days and that have authored at least 100 editions within the main space of Wikipedia within the last 3 months [7].

Previous studies show that 69,5% of discussions and 91% of comments refer to just four factors: notability, sources, maintenance and bias [27] and that notability is the main reason for deletion, up to 28% of which is especially for the articles of newcoming editors [28]. In studies focused on this deletion process and gender, the findings present that biographies about women are more frequently considered non-notable compared to men, and that individuals identified as non-binary or trans are frequently classified as non-notable [29].

2.2 Reliable sources

Defined as a content guideline in Wikipedia, "if no reliable sources can be found on a topic, Wikipedia should not have an article on it" [6], meaning Wikipedia is built on what others said elsewhere about a certain topic or person. The guideline discusses the reliability of sources - or *WP:RS* in antilanguage- meaning there are


The following discussion is an archived debate of the proposed deletion of the article below. ***Please do not modify it.*** *Subsequent comments should be made on the appropriate discussion page (such as the article's talk page or in a deletion review). No further edits should be made to this page.*

The result was **delete**. Yunshui 雲水 07:43, 19 March 2020 (UTC)

John Tauras [edit]

John Tauras (edit | talk | history | links | watch | logs | views) – (View AfD • Stats)
(*Find sources:* Google (books • news • newspapers • scholar • free images • WP refs) • FENS • JSTOR • *NYT* • TWL)

AfDs for this article:
- Articles for deletion/John Tauras
- Articles for deletion/John Tauras (2nd nomination)

Seems to be a redux of a non-notable musician who doesn't meet enwiki's notability standards...
TJH2018talk 22:34, 11 March 2020 (UTC)

Note: This discussion has been included in the list of Actors and filmmakers-related deletion discussions. **CAPTAIN RAJU**(T) 22:40, 11 March 2020 (UTC)
Note: This discussion has been included in the list of Lithuania-related deletion discussions. **CAPTAIN RAJU**(T) 22:40, 11 March 2020 (UTC)
Note: This discussion has been included in the list of England-related deletion discussions. **CAPTAIN RAJU**(T) 22:41, 11 March 2020 (UTC)
Note: This discussion has been included in the list of Television-related deletion discussions. **CAPTAIN RAJU**(T) 22:41, 11 March 2020 (UTC)
Note: This discussion has been included in the list of California-related deletion discussions. **CAPTAIN RAJU**(T) 22:41, 11 March 2020 (UTC)

- **Speedy Delete and WP:SALT** no notability whatsoever ands serves as little more than a vanity page. Best, GPL93 (talk) 23:01, 11 March 2020 (UTC)
- **Delete** and salt, looks like an advert and does not meet notability standards. dibbydib 💬/📝 23:17, 11 March 2020 (UTC)
- **Delete orSpeedy Delete and of course WP:SALT** per GPL93. MistyGraceWhite (talk) 04:11, 12 March 2020 (UTC)
- **Delete** the article was already deleted once this year. Nothing has changed to merit keeping it now.John Pack Lambert (talk) 15:22, 12 March 2020 (UTC)
- **Delete** Non notable entertainer. **~EDDY** (talk/contribs)~ 20:03, 12 March 2020 (UTC)
- **Delete**,Not notable, Nothing changed since it was once deleted. Alex-h (talk) 15:05, 15 March 2020 (UTC)

The above discussion is preserved as an archive of the debate. ***Please do not modify it.*** *Subsequent comments should be made on the appropriate discussion page (such as the article's talk page or in a deletion review). No further edits should be made to this page.*

Figure 1: Lithuanian musician John Tauras' *Article for Deletion* debate. The result of this debate was for the article to be deleted because, according to the editor, it didn't meet the English Wikipedia notability standards. It had 6 votes

certain sources that are recognised as notable, but this causes certain sources to be excluded. Within the Wikipedia Community, researchers from Art+Feminism studied the English, French and Spanish Wikipedia pages and found that there is no clear definition of what "reliability" means, creating a systematic bias [14].

The effect of a guideline such as *Reliable sources*, is that The Media has a primary role in what content is available in Wikipedia. The interaction between The Media and Wikipedia can be illustrated with a Twitter exchange. The press questioned Wikipedia with regards to its gender bias, and the former executive director of the Wikimedia Foundation, Katherine Maher, exposed on Twitter that the encyclopedia is a direct reflection of what The media focus is on [23]. Wikipedia is affected by the bibliographic universe of a topic [22] and the gender gap of the information environment is not the exception. In practice, it means that the more that is written in The Media and other resources about a person, the easier it is for editors to write a biography that can demonstrate its notability through reliable sources by including those mentions.

An editor can tag articles if they consider that they can be improved or enhanced [4]. The studies regarding notable sources in Wikipedia that have inspired us to analize the biographies in the encyclopedia are of three kinds: those which analize the distribution of the bibliographical references within the different sections of a biography, such as the studies of the biographies of UK polititians [11]. Others focus on the types of documents referenced in the Wikipedia articles, whether they are journals, textbooks, guidelines, newspapers or websites [13] and others review if the references are primary, secondary or tertiary and from which country the sources cited are from, with findings such as that 56% of the sources are from North America versus 0.3% from Africa [17], an inequality record for of all human knowledge. Even if these works are not directly focused on gender bias, they can lead us to a methodology with which to analyze the management of information on Wikipedia, to which we can add the gender perspective to gain a better understanding of content development within the digital encyclopedia.

The tags and cleanup templates can lead us to a corpus of articles questioned by their perception of quality, but that exceeds the scope of this study.

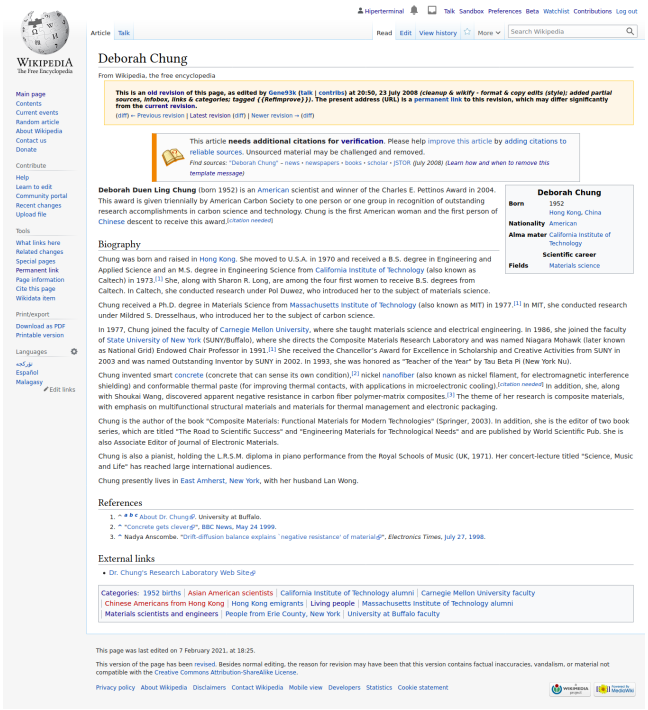

Figure 2: Deborah Chung's article, an MIT scientist, tagged to "add more reliable sources", above in orange

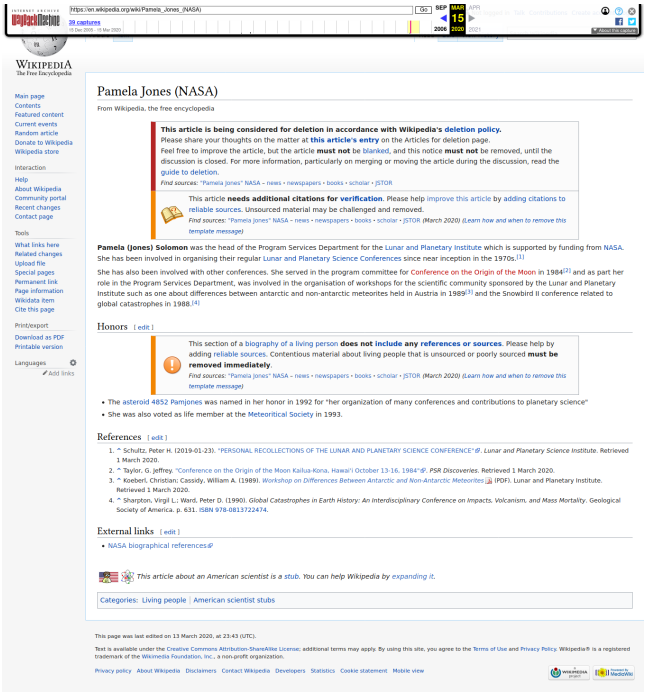

Figure 3: Pamela Jones's biography, a NASA scientist nominated to Articles for Deletion. Her biography is available on the Internet Archive's Wayback Machine, not in English Wikipedia

3 A CORPUS TO COVER THE SPECTRUM OF CREATED BIOGRAPHIES

In Wikipedia we can read the articles available once they have completed the evaluation process, but those which fail are invisible and unaccessible unless you are a Wikipedia administrator. That's why we propose the analysis of a corpus that includes deleted biographies from the *Articles for Deletion* process. To cover a second segment, we propose a corpus of biographies that are not in risk of being deleted but that need maintenance, from the reliable sources tagging process. This corpus should include biographies without the maintenance mark too, in other words, unpolluted biographies. This spectrum covers different kinds of biographies found in Wikipedia.

We propose a gender analysis of two corpus as shown in figure 4: a first corpus of biographies in the *Articles for Deletion* category, to cover the deleted and kept biographies. A second corpus includes biographies tagged to include reliable sources and biographies without marks. All data retrieved is limited by topic: "scientist"; and a time frame. For scientists we use the category *Scientist by field* [2] and identify tagged or not tagged biographies.

This selection is used in our search for "scientist" in Wikipedia but can be applied to any other profession. We are using Wikipedia in English because it is the largest in number of articles created, but it can be applied to any other language in Wikipedia. The bibliography related to *Articles for Deletion* and *Reliable sources* does not necessarily cover the gender gap and if it is covered it is often from a binary perspective. We hope this can be useful to identify the gender gap not just in terms of men and women, but

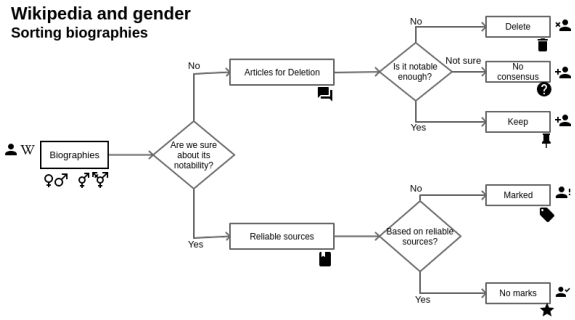

Figure 4: Flow diagram to sort the biographies of the corpus

being more inclusive to make visible those who are not in the sum of all human knowledge.

We consider that in order to solve the gender bias within Wikipedia we need to understand the logic of the evaluation of biographies regardless of the number of biographies created. If we don't take this into account, despite that more articles are created, the rate of deletion or tagging may still maintain the imbalance and the gap will continue to persist.

ACKNOWLEDGMENTS

We would like to thank Aina Vidal, Xandre Pichel and Júlia Ojeda for sharing their ideas regarding this corpus and the Wikipedia female editor's community "Dones and Wikipedia".

REFERENCES

[1] 2019. Wikipedia:List of Policies and Guidelines. *Wikipedia* (Oct. 2019).
[2] 2020. Category:Scientists by Field. *Wikipedia* (Jan. 2020).
[3] 2020. Wikipedia:Articles for Deletion. *Wikipedia* (Feb. 2020).
[4] 2020. Wikipedia:Template Index/Cleanup. *Wikipedia* (Sept. 2020).
[5] 2021. Wikipedia:Notability. *Wikipedia* (Nov. 2021).
[6] 2021. Wikipedia:Reliable Sources. *Wikipedia* (Dec. 2021).
[7] 2022. Viquipèdia:Esborrar pàgines/Propostes. *Viquipèdia, l'enciclopèdia lliure* (March 2022).
[8] 2022. Wikipedia:Consensus. *Wikipedia* (Feb. 2022).
[9] 2022. Wikipedia:The Difference between Policies, Guidelines and Essays. *Wikipedia* (Jan. 2022).
[10] Julia Adams, Hannah Brückner, and Cambria Naslund. 2019. Who Counts as a Notable Sociologist on Wikipedia? Gender, Race, and the "Professor Test". *Socius: Sociological Research for a Dynamic World* 5 (Jan. 2019), 237802311882394. https://doi.org/10.1177/2378023118823946
[11] Pushkal Agarwal, Miriam Redi, Nishanth Sastry, Edward Wood, and Andrew Blick. 2020. Wikipedia and Westminster: Quality and Dynamics of Wikipedia Pages about UK Politicians. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20)*. Association for Computing Machinery, New York, NY, USA, 161–166. https://doi.org/10.1145/3372923.3404817
[12] Alexa Internet. [n.d.]. Top Sites. https://www.alexa.com/topsites.
[13] S.A. Azer, N.M. AlSwaidan, L.A. Alshwairikh, and J.M. AlShammari. 2015. Accuracy and Readability of Cardiovascular Entries on Wikipedia: Are They Reliable Learning Resources for Medical Students? *BMJ Open* 5, 10 (2015). https://doi.org/10.1136/bmjopen-2015-008187
[14] Amber Berson, Monika Sengul-Jones, and Melissa Tamani. 2021. Unreliable Guidelines: Reliable Sources and Marginalized Communities in French, English and Spanish Wikipedias.
[15] Italo Calvino. [n.d.]. La antilengua.
[16] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* The MIT Press, Cambridge, Massachusetts.
[17] Heather Ford, Shilad Sen, David R. Musicant, and Nathaniel Miller. 2013. Getting to the Source: Where Does Wikipedia Get Its Information From?. In *Proceedings of the 9th International Symposium on Open Collaboration (WikiSym '13)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2491055.2491064
[18] Heather Ford and Judy Wajcman. 2017. 'Anyone Can Edit', Not Everyone Does: Wikipedia's Infrastructure and the Gender Gap. *Social Studies of Science* 47, 4 (Aug. 2017), 511–527. https://doi.org/10.1177/0306312717692172
[19] Sue Gardner. 2011. Nine Reasons Women Don't Edit Wikipedia (in Their Own Words).
[20] Isaac Johnson, Florian Lemmerich, Diego Sáez-Trumper, Robert West, Markus Strohmaier, and Leila Zia. 2020. Global Gender Differences in Wikipedia Readership. *arXiv:2007.10403 [cs]* (July 2020). arXiv:cs/2007.10403
[21] S.K. Lam, A. Uduwage, Z. Dong, S. Sen, D.R. Musicant, L. Terveen, and J. Riedl. 2011. WP:Clubhouse? An Exploration of Wikipedia's Gender Imbalance. In *Computer Science & Engineering User Home Pages_ University of Minnesota (WikiSym 2011 Conference Proceedings - 7th Annual International Symposium on Wikis and Open Collaboration)*. 1–10. https://doi.org/10.1145/2038558.2038560
[22] Brendan Luyt. 2021. Representation and the Problem of Bibliographic Imagination on Wikipedia. *Journal of Documentation* ahead-of-print, ahead-of-print (Jan. 2021). https://doi.org/10.1108/JD-08-2021-0153
[23] Katherine Maher. 2020. Katherine Maher En Twitter: "Next Time Someone Tells Me a Woman Isn't Notable Enough to Be on @Wikipedia Perhaps I Shall Point Them to This Gent, Whose Spare and Dubious Achievements Have Languished Unmolested in Their Thinly Cited Glory for 806 Days and Counting. https://t.co/MhRokFXbGl" / Twitter. https://twitter.com/krmaher/status/1213542455334694913.
[24] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. 2015. "The Sum of All Human Knowledge": A Systematic Review of Scholarly Research on the Content of Wikipedia: "The Sum of All Human Knowledge": A Systematic Review of Scholarly Research on the Content of Wikipedia. *Journal of the Association for Information Science and Technology* 66, 2 (Feb. 2015), 219–245. https://doi.org/10.1002/asi.23172
[25] Julià Minguillón, Julio Meneses, Eduard Aibar, Núria Ferran-Ferrer, and Sergi Fàbregues Feijóo. 2021. Exploring the Gender Gap in the Spanish Wikipedia: Differences in Engagement and Editing Practices. *PLOS ONE* (2021). https://doi.org/10.1371/journal.pone.0246702
[26] Jacqueline C Pike, Elisabeth W Joyce, and Brian S Butler. 2017. Overcoming Transience and Flux: Routines in Community-Governed Mass Collaborations. *Information Technology & People* 30, 2 (2017), 449–472. https://doi.org/10.1108/ITP-08-2015-0194
[27] Jodi Schneider, Alexandre Passant, and Stefan Decker. 2012. Deletion Discussions in Wikipedia: Decision Factors and Outcomes. In *WikiSym 2012 Conference Proceedings - 8th Annual International Symposium on Wikis and Open Collaboration*. https://doi.org/10.1145/2462932.2462955
[28] Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. 2013. Arguments about Deletion: How Experience Improves the Acceptability of Arguments in Ad-Hoc Online Task Groups. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, 1069–1080.
[29] Francesca Tripodi. 2021. Ms. Categorized: Gender, Notability, and Inequality on Wikipedia. *New Media & Society* (June 2021), 146144482110237. https://doi.org/10.1177/14614448211023772
[30] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *arXiv:1501.06307 [cs]* (March 2015), 454–463. arXiv:cs/1501.06307
[31] Jimmy Wales. 2014. Wikipedia 'completely Failed' to Fix Gender Imbalance.
[32] Andrew G. West and Insup Lee. 2011. What Wikipedia Deletes: Characterizing Dangerous Collaborative Content. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*. Association for Computing Machinery, New York, NY, USA, 25–28. https://doi.org/10.1145/2038558.2038563
[33] Shing-Chung Jonathan Yam. 2016. Negotiating Boundaries of Knowledge: Discourse Analysis of Wikipedia's Articles for Deletion (AfD) Discussion. *Communication and Critical-Cultural Studies* 13, 3 (Sept. 2016), 305–323. https://doi.org/10.1080/14791420.2015.1137334