

January 01, 2011

Gender bias in Wikipedia and Britannica

Joseph M. Reagle
Northeastern University

Lauren Rhue
New York University

Recommended Citation

Reagle, Joseph M. and Rhue, Lauren, "Gender bias in Wikipedia and Britannica" (2011). *Communication Studies Faculty Publications*. Paper 1. <http://hdl.handle.net/2047/d20002078>

This work is available open access, hosted by Northeastern University.

Gender Bias in Wikipedia and Britannica

JOSEPH REAGLE¹
Northeastern University

LAUREN RHUE
New York University

Is there a bias against women's representation in Wikipedia biographies? Thousands of biographical subjects from six sources are compared against the English-language Wikipedia and the online *Encyclopædia Britannica* with respect to coverage, gender representation, and article length. We conclude that Wikipedia provides better coverage and longer articles, and that it typically has more articles on women than *Britannica* in absolute terms, but we also find that Wikipedia articles on women are more likely to be missing than are articles on men relative to *Britannica*. For both reference works, article length did not consistently differ by gender.

Introduction

Wikipedia's self-description as "the encyclopedia that anyone can edit" is evidence of two key influences. Obviously, Wikipedia is an encyclopedia. While reference works are often thought of as bland tomes, they are sometimes objects of contention in larger cultural debates about what deserves to be recognized as knowledge (Einbinder, 1964; Morton, 1994). Also, Wikipedia is a wiki, a Web-based editing platform made popular by those who prefer simple and open collaboration (Cunningham, 2004). Indeed, Wikipedia was inspired by the growing strength of the larger free/libre and open-source software (FLOSS) movement whereby software is shared and developed in the open. However, instead of software code, Wikipedians produce encyclopedic prose.

Yet, Wikipedia also (seemingly) inherits an unfortunate trait from its reference work and technical ancestors: In the realm of reference work production, women's representation as contributors and subjects has been slight. In her study of the *Encyclopædia Britannica*, historian Gillian Thomas notes that, as contributors, women were relegated to matters of "social and purely feminine affairs"—as stated by an

¹ We thank Axel Boldt, Benjamin Mako Hill, Kaldari, Jelena Karanovic, Felipe Ortega, Siebrand, and the anonymous reviewers for their feedback on this work.

Joseph Reagle: joseph.2011@reagle.org

Lauren Rhue: lrhue@stern.nyu.edu

Date submitted: 2010-03-18

early 20th century editor—and were “perceived as acting as pedantic handmaidens to the wide-ranging sweep of male intelligence” (Thomas, [1992](#), pp. 18, 26). As subjects, women were often little more than addendums to male biographies (e.g., Marie Curie as the husband of Pierre Curie). In the realm of technology, the gender imbalance found in computer-related fields is exacerbated rather than mitigated. That is, despite a voluntary, egalitarian and meritocratic ethos, women are only a tiny fraction of participants, as low as 1.1% in an oft-cited FLOSS survey (Ghosh, Glott, Krieger, & Robles, [2002](#)). One would hope Wikipedia would be more balanced, yet surveys indicate that women constitute around 13% of Wikipedians (Glott, Schmidt, & Ghosh, [2010](#)).

Such imbalances in participation and representation prompt us to look for indications of “systemic bias” in the biographical coverage of women at the English-language [Wikipedia](#) and online [Britannica](#). In this article, we introduce Wikipedia, discuss gender bias in reference works and technical communities, and examine prior work on comparative analyses of Wikipedia. Subsequently, we describe a method that allowed us to obtain thousands of biographies from Wikipedia and *Britannica*. Our analysis and findings are then presented via tables of percentages (of men and women whom each work covers) and complementary logistical regressions.

Background

Wikipedia

Wikipedia’s name is a portmanteau of “wiki,” an online collaborative editing tool, and “encyclopedia,” itself a contraction of the Greek *enkyklios* and *paidei*, referring to the “circle of learning” of the classical liberal arts. Furthermore, “wiki wiki” means “super fast” in the Hawaiian language, and Ward Cunningham chose the name in 1995 for his collaborative WikiWikiWeb software to indicate the ease with which one could edit pages. (He learned of the word during his first visit to Hawaii when he was initially confused by the direction to take the “Wiki Wiki Bus,” the Honolulu airport shuttle (Cunningham, [2003](#)). In a sense, the term *wiki* captures the original conception of the World Wide Web as both a browsing and editing medium; the latter capability was largely forgotten when the Web began its precipitous growth and the most popular clients (i.e., “browsers”) did not permit users to edit Web pages.

The wiki changed this asymmetry by placing the editing functionality on the website itself. Using the Wikipedia syntax, one types “#” this provides a link to “[[Ward Cunningham]]” to add a numbered list item with a link to the “Ward Cunningham” article. The Wikipedia software translates this into the appropriate HTML and hypertext links for users to view. To create a new page, one simply creates a link to it, which remains red until someone actually adds content to its target destination. These capabilities are central to and representative of wikis.

Women, FLOSS, and Reference Works

The well-known gender imbalance in computer-related fields—approximately 27% female—is further exacerbated in the FLOSS community, with women making up about 1.1% of participants (Ghosh

et al., 2002; NCWIT, 2007). Likely hindrances to participation include a lack of mentors and role models, language usage, a male-dominated competitive world view, lack of women-centered perspectives, other life demands (e.g., personal time and work), scope of interests, and the identity or construal of women in the community (Karanović, 2008; Lin, 2005).

The history of reference work production includes notable examples of chauvinism. Robert Cawdrey's *Table Alphabetical* can't help but be read today as patronizing, given that it was "gathered for the benefit & helpe of Ladies, Gentlewomen, or any other unskillful persons" (Cawdrey, 1604/1997). Centuries later, privileged women had greater access to learning and were lauded for supposed feminine virtues, but otherwise chauvinism persisted. John Ruskin (1910), Victorian art critic and social commentator, advised that women should be permitted knowledge only such that it "may enable her to understand, and even to aid the work of men." That is, "a woman ought to know the same language, or science, only so far as may enable her to sympathize with her husband's pleasures, and those of his friends" (pp. 154–155). Thomas (1992) argued that this attitude was very much reflected in *Britannica's* articles and in the work environment of its female contributors and staff. Thomas notes that at the beginning of the 20th century the "Women" article can be characterized by its "evasiveness about the suffrage issue, but praise for women's special moral qualities and 'demeanor,' the apparent credit given to pioneer women professionals, while at the same time describing them as 'invaders' of their chosen fields of work" (p. 39). In 1942, Mary Ritter Beard led "A Study of the *Encyclopædia Britannica* in Relation to Its Treatment of Women." She and her colleagues found many biases and omissions. For example, their report noted that the article on "Song" gives the impression that "No women sang in Europe, it appears from this review. The contributions of nuns, in choir composition and singing, is [sic] not recognized at all" (Beard, Edinger, Selig, & White, 1977, p. 220).

Furthermore, few women prominently appear in the historical record of reference works. Beyond those in Thomas' work, the few exceptions are in the domain of librarians and documentalists, such as Suzanne Briet (Maack, 2004). Unfortunately, even Melvil Dewey's advocacy for women in the library profession is marred by alleged discrimination and personal scandal (Wiegand, 1996). Thomas reported that of the some 1,500 authors contributing to the 11th *Britannica*, 35 of them were women (about 2%), with no woman listed among the 49 editorial advisors (1992, p. 18). Today, at least, women are visible in everyday tasks and positions of authority at Wikipedia, though they continue to be a minority.

Topical Comparisons and Systemic Bias

While there have been no large-scale comparisons of biographical coverage and gender in reference works, there are quality and topical coverage comparisons. For example, in a report from the prestigious science journal *Nature*, Wikipedia articles were found to contain more errors than did *Britannica* articles (Giles, 2005). With respect to topical coverage, George Bragues (2007) compared the biographical articles of seven prominent philosophers with authoritative reference works and found that Wikipedia only covers 52% on average (56% median) of topics commonly found in the other reference works. (No errors were found, though there were significant omissions.) Alexander Halavais and Derek Lackaff (2008) undertook two analyses of Wikipedia topical coverage and found that Wikipedia does well with respect to general knowledge (because of its size) and technical issues (likely because of Wikipedia's

contributors), but it is weak on law and medicine ("the purview of licensed experts" (p. 438)). Furthermore, comparisons of content between different language versions of Wikipedia reveal divergences in interlanguage links, information boxes, and topics addressed (Adar, Skinner, & Weld, [2009](#); Hecht & Gergle, [2009](#), [2010](#)).

Among Wikipedians, the likelihood of imbalanced topical coverage has long been acknowledged. The "Countering Systemic Bias" WikiProject began in the fall of 2004 with discussion of how the interests and the demographics of its contributors affected its topical coverage (e.g., a "Western" and "geeky" focus). The project page notes that:

The Wikipedia project suffers systemic bias that naturally grows from its contributors' demographic groups, manifesting an imbalanced coverage of a subject, thereby discriminating against the less represented demographic groups. This project aims to control and (possibly) eliminate the cultural perspective gaps made by the systemic bias, consciously focusing upon subjects and points of view neglected by the encyclopedia as a whole. (Wikipedia, [2004](#), [2009c](#))

Countering systemic gender bias specifically has also been a concern of the "WikiProject Gender Studies," though its subproject on gender bias does not appear to be very active (Wikipedia, [2008](#), [2009d](#)).

Recent informal studies of gender imbalance indicate a Wikipedia bias. A report by John Limey ([2010](#)) compared the percentage of women in a sample of Wikipedia biographies (19.3%) with in an online biographical database (29.9%); the 10-point difference is "highly suggestive" of bias. Similarly, an analysis of the gender balance of people appearing on Wikipedia's front page finds "Nine men to every one woman on a portal that represents the greatest easily accessible store of knowledge is outrageously disproportionate and unacceptable" (RMJ, [2010](#)).

Method

The preceding literature prompted us to ask the following: (1) What kind of gender bias can one find in existing lists of notable persons; and (2) supposing incomplete encyclopedic coverage of those sources, how do *Britannica* and Wikipedia fair? That is, if both works only had biographies for 20% of a source's listings, can one discern a further bias in which gender is focused upon by the reference work? For example, assume a source list has 100 notable persons split evenly between the genders. Because each reference work only has 20 articles among the 100 subjects, either work could focus on either gender exclusively in its coverage (i.e., 0–100%), revealing a bias.

To answer this question, a Python program was used to find, "crawl," and compare Web pages related to biographical subjects in the targeted reference works. (Unless specified otherwise, these results are from June 2010.) Biographical subjects were chosen from six sources: The National Women's History Project (NWHHP), *The Atlantic's* 100 most influential figures in American history, *TIME Magazine's* list of 2008's most influential people, *Chambers Biographical Dictionary*, American National Biography Online,

and Wikipedia itself. Such a task is challenging in that there are often ambiguities inherent in the naming of subjects, including people with the same name, a single person with different names (e.g., indigenous and colonial), nicknames, differing transliterations, order, honors, syntactical conventions, and use of diacritics. Additionally, sources for such names may have their own errors in addition to such variations, including typos. So as to fairly and accurately test topical coverage of *Britannica* and Wikipedia, we made use of the Google search engine, which has indexed both sites.

Google provides a convenient search API to its search engine that has expansive coverage and an effective search algorithm for natural language queries. Therefore, after converting a biographical source list into a standard format, the Google API was queried for its top four results for a given name. Queries were restricted to the content of britannica.com/EBchecked/topic/ and of en.wikipedia.org (excluding doc, pdf, jpg, gif, and png results). From the query, the results' titles were normalized (e.g., titles were lowercased, and hyphens and diacritics were removed). They were then compared with the source name using a "fuzzy comparison" algorithm (i.e., Python's `diff` Levenshtein-Distance ratio). These comparisons were augmented with hand-tuned heuristics (e.g., sorting long names that have many honorifics) to maximize correct matches and minimize false matches. The length of an article is determined by counting the words of article content (sans markup) and does not include external links, notes, citation information, and other miscellany.

Gender was primarily guessed via the balance of gendered pronouns used in a biography (i.e., he/his and she/her). If the difference between instances of masculine and feminine pronouns is less than 25% of their sum, gender is unknown. This method was tested against the approximately 400 biographies from the *TIME* and NWHP lists. For those biographies for which gender was guessed, we manually confirmed that all guesses were correct. (Anita Hill's biography was the edge case, coming close to being labeled as "unknown" given its extensive discussion of Clarence Thomas, that is: she = 37, he = 21; $(37-21)/(37+21) = 28\% = \text{female}$). Those few subjects whose gender could not be guessed by this method were largely the result of neither reference work having a biographical article. But these subjects are still amendable to guesses based on honorifics (e.g., Count vs. Countess) and given names. Honorifics were unambiguous. Given name guesses were against the 1990 U.S. census; gender was guessed if the frequency of the name occurring in one gender was at least four times that in the other. "Joseph" was reliably guessed as predominantly male (1.404% male; 0.005% female) but "Pat" was considered unknown (0.040% female to 0.022% male). In the case of the 18,495 subjects from *Chambers*, all but 1,210 (~7%) are assessed a gender via pronoun counts, and only 431 (2%) of subjects remain unknown after name-based guesses.

To consider an example in full, Claudine Alexandrine Guérin de Tencin, (French novelist, socialite, and mother of encyclopedist Jean le Rond d'Alembert), was selected from *Chambers Biographical Dictionary*. This name was queried via Google against the English Wikipedia and *Britannica* websites. Wikipedia's article on "Claudine Guérin de Tencin" and *Britannica's* article on "Claudine-Alexandrine Guerin de Tencin" were returned as possible matches. The source name and both titles were then transformed by lowercasing, removing hyphens and diacritics, and then sorting. After transformation, Wikipedia's article title (i.e., "claudine de guerin tencin") and that of *Britannica* (i.e., "alexandrine claudine de guerin tencin") were considered appropriate matches against the source (i.e., "alexandrine claudine de guerin tencin")

because the words in both transformed titles are subsets of the transformed source name. The subject was assessed as female, given that the Wikipedia article has 30 feminine pronouns (i.e., "her" and "she") and no masculine pronouns (i.e., "his" and "he"). The encyclopedic content of the Wikipedia article includes 389 words to *Britannica's* 196.

Wikipedia and *Britannica* each have a rare article variant that we include in the analysis. At Wikipedia, similar subjects often have a "disambiguation page" that provides links to more specific articles. For example, its article for "Mary Anderson" in the NWHP list is a disambiguation page that includes links to nine biographies. If "(disambiguation)" is within the URL of the results returned by Google, we skip this result, as the next result is possibly the appropriate biography. Yet, not all disambiguation pages are identified via the URL. However, such biographical pages are tagged with the `{dmbox}` or `{setindexbox}` templates, and we choose the first article listed. (Such disambiguation pages are rare; they are encountered in only about 1% of the source lists we tested, for example, the *TIME*, *Atlantic*, NWHP, and Wikipedia lists.) On the other hand, *Britannica* includes pages wherein a person is related to, but not the subject of, a complete article. For example, the composer "Buffy Sainte-Marie" has a very short article that is only a paragraph excerpt from the "Billy Williams (British cinematographer)" article. Hence, inclusion of these fragments may improve *Britannica's* coverage figures, but also may decrease article length figures. While Wikipedia does permit "targeted redirects" in which a subject's article may redirect to the section of another article (such as the "Malia Obama" article redirecting to a section of the "Family of Barack Obama"), we never encountered an instance of this in manually examined data (*TIME*, *Atlantic* and NWHP); if any occurred, the (theoretical) impact would be to lessen Wikipedia's coverage.

The data used in our analyses are available online (Reagle, [2010](#)).

Findings

Biographical Coverage Analysis by Percentages

What kind of gender imbalance exists in biographical sources, and how do Wikipedia and *Britannica* compare? We begin by looking at lists of biographical persons and the proportions of gender balance in covered articles in Wikipedia and *Britannica*. We then focus on the missing articles to see if we can discern what factors contribute to their absence.

Comparison of Gender Balance

The NWHP maintains short biographies of women of import. Some are historical, some contemporary, some widely known (e.g., Susan B. Anthony), but many are not. Hence it provides a diverse challenge to encyclopedic coverage. Out of an initial test selection of 174 subjects Wikipedia lacked 23 biographical articles, much less than the 74 articles missing from *Britannica*. In September 2009 these results were posted and reviewed by interested Wikipedians who helped identify bugs in our source lists and results. Also, within a day, the "WikiProject Gender Studies/Feminism Task Force" set about

providing biographies for those missing in the preliminary analysis (Wikipedia, [2009a](#)). By November, 6 of the 23 original missing biographies were at least started. As of June 2010, analysis of all 268 women revealed 77 biographies were missing from Wikipedia and 155 entries from *Britannica*. (While the NWHP data is consequently tainted, it makes no differences to our findings and is not included in the regression analyses in any case.)

However, this might only indicate that Wikipedia coverage is greater than that of *Britannica* and does not address whether Wikipedia biographies are disproportionately male focused. Yet, because Wikipedia (and other reference works) are documenting a biased world, it can be difficult to settle upon a figure for “normal” female representation. Repeating Limey’s ([2010](#)) query of biographical resources for those born after 1909, we approximated his finding of 29.9% female coverage at the Gale Biographical Resource Center with 28.7% of queried persons being female (i.e., 347,874 female; 865,068 male biographies). Furthermore, we queried two other sources: Wilson’s Current Biography Illustrated (CBI) yielded 24.5% female biographies (2,715 female; 8,350 male), and American National Biography Online offered up 15% (289 female; 1,644 men).

In any case, a more detailed comparative exploration required actual lists of notable men and women, as is the case in the following analyses.

Table 1. Topical Coverage and Gender Representation.

List	Size	Female	WP Female	WP Male	EB Female	EB Male
<i>ANBO</i>	1000	163 (17%)	113 (14%)	673(86%)	60 (19%)	254 (80%)
<i>Atlantic</i>	100	10 (10%)	10 (10%)	90 (90%)	10 (10%)	89 (90%)
<i>Chambers</i>	18,495	2,257 (12%)	2,010 (12%)	14,981 (88%)	1,265 (10%)	11,243(90%)
<i>NWHP</i>	269	269 (100%)	192 (100%)	0 (0%)	114 (100%)	0 (0%)
<i>TIME 2008</i>	105	24 (23%)	24 (23%)	79 (75%)	14 (22%)	51 (78%)
<i>Wikipedia</i>	1000	160 (16%)	159 (16%)	806(82%)	12 (14%)	75 (85%)

In 2006, *The Atlantic* enumerated its top 100 most influential figures in American history, all of which had biographies in Wikipedia and *Britannica*. Women are 10% of this population. However, a different list of contemporary influential people shows a significant increase in female representation. *TIME*’s list of 2008’s most influential people includes 105 persons (sometimes people are recognized in pairs), 24 of whom are female (see Table 1). However, since Wikipedia has almost perfect coverage of these two sources (as seen in Table 2), there is little room for its coverage to deviate from these two sources.

Therefore, we sought a larger collection of people that spanned a greater historical period. An online version of *Chambers Biographical Dictionary* lists 18,495 names that can be used to generate unique queries (e.g., not including collisions between those with the same name, but with ambiguous birth

dates). While this source list does contain relatively contemporary figures, including those who died in 2007, it also includes the whole breadth of history, and favors those born in the early 20th and the 19th centuries (roughly 12,000 subjects). It shares many naming conventions with *Britannica* and also contains a fair amount of Barons, Lords, Sirs, and Dames. Performing the analysis on the 18,495 entries, 2,257 are surmised to be female, 15,775 as male, and the program could not determine the gender of 463 persons. Females are 12% of the source population, 12% of the Wikipedia population, but only 10% of the *Britannica* population.

Because *Chambers* focuses on the 18th and 19th centuries, we wondered if some of our findings are influenced by this particular list or by the period in which the subjects lived. So, we looked for yet another large source from which to take a sample. We obtained a list of 1,000 random persons from the American National Biography Online (ANBO) and found that, with respect to imbalance, Wikipedia actually fared much worse than *Britannica*, with only 14% of its coverage dedicated to females relative to *Britannica*'s 19%. Interestingly, while Wikipedia had nearly twice the number of female biographies than did *Britannica* (113 to 60), it had over two and a half times the number of male biographies (673 to 254).

Gender Percentages of Missing Articles

Given that Wikipedia and *Britannica* roughly follow the biases of existing works, one might conclude that they do not contain much gender bias; however, we can extend the analysis by focusing upon the gender balance of the *missing* articles.

Table 2. Missing Articles Totals and Percentages.

List	WP missing Total	WP missing Female	WP missing Male	EB missing Total	EB missing Female	EB missing Male
<i>ANBO</i>	213 (21%)	50 (31%)	136 (17%)	686 (67%)	103 (63%)	555 (69%)
<i>Atlantic</i>	0 (0%)	0 (0%)	0 (0%)	1 (1%)	0 (0%)	1 (1%)
<i>Chambers</i>	1,475 (8%)	247 (11%)	824(5%)	5,973 (32%)	992(44%)	4,562 (29%)
<i>NWHP</i>	77 (29%)	66 (26%)	N/A	155 (58%)	144(56%)	N/A
<i>TIME 2008</i>	2 (2%)	0 (0%)	1 (1%)	40 (38%)	10(42%)	19 (36%)
<i>Wikipedia</i>	13 (1%)	1 (1%)	4 (0%)	914 (91%)	148 (93%)	735 (91%)
OVERALL	1,780 (8%)	364 (13%)	965 (5%)	7,767 (37%)	1397 (49%)	5,872 (33%)

WP=Wikipedia EB=*Encyclopedia Britannica*

The columns in Table 2 display the number and percentage of missing articles for each reference work and gender combination. For example, in the second column, Wikipedia is missing 247 females from *Chambers*, or 11% of *Chambers*' female population. The missing rates for unidentified gender subjects are generally high (and omitted from this table) because the gender identification relies upon the occurrences of pronouns in an article, if it exists, and on name-based guesses.

Over all six source lists, *Britannica* misses 7,767 names, much more than Wikipedia's 1,780 missing subjects. Also, Wikipedia consistently misses fewer biographies than does *Britannica* for each source list, confirming that Wikipedia has more exhaustive coverage than does *Britannica*.

Wikipedia's exceptional overall coverage does not address systematic differences in *which* subjects are covered. Gender could influence which subjects are overlooked in both Wikipedia and *Britannica*. Toward understanding potential gender bias, we examine missing articles for males and females separately, and we find that Wikipedia and *Britannica* both cover women less comprehensively than they do males. Of the 2,287 women in the source population, Wikipedia misses 364 females and *Britannica* misses 1,397, 13% and 49%, respectively. Of the 17,594 men, Wikipedia misses 965 subjects and *Britannica* misses 5,872 subjects, 5% and 33%, respectively. Although the absolute number of missing men exceeds the number of missing women in each reference work, women fare worse in their proportion of missing articles. Relative to men, across all source lists, women have a 2.6 (13/5) greater odds of omission in Wikipedia and a 1.48 (49/33) greater odds of omission in *Britannica*.

Finally, how do these reference works compare when we use Wikipedia as a source? A list of biographies from Wikipedia in 2007 included 445,966 subjects (Wikipedia, [2009e](#)). Unlike other sources, this included music bands. It also included pages that were listed as "(page does not exist)." Excluding these two subjects, we randomly sampled 1,000 subjects and found that *Britannica* fared poorly, missing 914 entries. Wikipedia missed 14, meaning they've since been deleted, perhaps for lack of notability. Of all the entries, 159 are female, 810 are male, and 31 are of unknown gender. That is, women are 16% of the gender-known population of the Wikipedia list and its analyzed articles. Women comprise only 14% of the *Britannica* articles.

Coverage Analysis via Logistic Regression

The preceding analysis of gender balance reveals that the reference works have more comprehensive coverage of male biographies. To better isolate the correlation between subject gender and coverage, if any, we analyze the relationship between subject coverage and reference work, gender, and source list. The dependent variable is *Coverage*, which is coded as 1 if a subject's biography is included in the reference work and as 0 if the subject is missing. There are four independent variables related to gender: *Male*, *Unknown Gender*, *Male in Wikipedia*, and *Unknown in Wikipedia*. *Male* and *Unknown Gender* are indicator variables: *Male* is coded as 1 for male subjects and as 0 for other subjects, and *Unknown Gender* is coded as 1 for unknown gender subjects and as 0 otherwise. Females are coded as 0 for *Male* and *Unknown Gender*. With this specification, the coefficient on the *Male* variable becomes a comparison in the relative change in coverage for male subjects relative to female subjects. Based on earlier analysis, our hypothesis is that men receive better coverage than women, so the *Male* coefficient is expected to be positive, indicating a correlation between male subjects and increased coverage.

The effect of gender may not be consistent across both reference works. For example, *Britannica* has more comprehensive coverage of men than it does of women from most source lists, but it also has better coverage of women from ANBO. To identify the interaction between gender and reference work, the

model includes another independent variable: *Male in Wikipedia*, which is coded as 1 for a male subject in a Wikipedia entry and as 0 otherwise. The *Male in Wikipedia* coefficient estimates the effect of gender for subjects in Wikipedia as distinct from either Wikipedia's overall coverage or from gender bias in reference works. The *Unknown in Wikipedia* is coded similarly and included for control purposes.

We used logistic regression to estimate the effects of the independent variables on *Coverage*. We restricted the regression dataset to names from *TIME*, *The Atlantic*, the 1,000 ANBO random samples, and a random sample of 1,000 subjects from *Chambers*. We sampled from our larger *Chambers* dataset to prevent idiosyncrasies in *Chambers* from dominating the regression results. The effect of gender is stronger using the entire set of names from *Chambers*, possibly due to the historical nature of the biographies. We attempted to control for historical period by gathering birth/death years for subjects; however, accurate period information was missing for more than 3,000 subjects, so those controls were not included in the regression. The sample of 1,000 names was chosen from the *Chambers* source to match the number of names from the ANBO source. NWHP was excluded because it is composed of only female names, precluding comparison between the coverage of men and women. Wikipedia as a source of names was omitted because the coverage is naturally skewed toward Wikipedia and could confound analysis of gender bias. The revised dataset for empirical analysis contains 2,205 subjects with 333 females, 1,824 males, and 48 subjects without an identifiable gender.

The three source list variables are included to control for differences among the source lists. There are variables for three of the four source lists in the dataset: *Chambers*, *ANBO*, and *TIME*. For example, the *TIME* variable is coded as 1 for every name that originates from the *TIME* source list and as 0 otherwise. Subjects from *The Atlantic* are coded as 0 for the three source list variables, allowing for a comparison between the coverage rates for the other source lists and *The Atlantic*. Because Wikipedia and *Britannica* contain all except one name from *The Atlantic*, the coefficients on the source list variables are all expected to be negative.

Finally, the independent variable *Wikipedia* is coded as 1 if the biography is from Wikipedia and as 0 if from *Britannica*. Thus, the *Wikipedia* coefficient indicates Wikipedia's coverage relative to that of *Britannica*. We expect a positive coefficient on *Wikipedia* to confirm the earlier analysis.

Comparison of *Britannica* and Wikipedia Bias

These regression results support many of our earlier findings. Wikipedia's superior coverage, regardless of gender, is highlighted in the significant positive coefficient on Wikipedia. Furthermore, the negative coefficients on *Chambers*, *TIME*, and *ANBO* confirm that both reference works are missing more subjects from these source lists than they are from *The Atlantic*, the baseline.

Table 3. Gender Coverage Analysis.

Variable	Estimate	Std. Error	Pr(> z)
(Intercept)	4.5621	1.0107	6.37e-06 ***
Male	0.1618	0.1304	0.214568
Unknown	-15.522	197.5019	0.937357
Wikipedia	1.6860	0.1897	< 2e-16 ***
Chambers	-3.8641	1.0061	0.000123 ***
TIME	-3.9654	1.0215	0.000104 ***
ANBO	-5.4119	1.0058	7.43e-08 ***
Male in Wikipedia	0.5522	0.2115	0.009018 **
Unknown in Wikipedia	10.6430	197.5033	0.957025
Null deviance: 5387.8 on 4409 degrees of freedom			
Model deviance: 3965.8 on 4401 degrees of freedom			
Likelihood ratio: 1422 on 6 degrees of freedom			
AIC: 3986.8			
Significant codes: 0 "****" 0.001 "***" 0.01 "**" 0.05 "." 0.1 " " 1			

Interestingly, two of the gender variables, *Male* and *Unknown*, have non-significant coefficients, suggesting that the influence of gender may not be consistent across both reference works. In contrast, the *Male in Wikipedia* coefficient is significant, providing evidence that gender contributes to the subject's degree of coverage on Wikipedia. This finding supports the earlier result that Wikipedia's female missing rates were consistently higher across source lists than were those of males.

The logistic regression results allow us to ascertain the probabilities of a subject's coverage based on the source list of the names, the subject's gender, and the reference work. The probability of a subject's coverage is given by the inverse logit of the sum of the coefficients and attributes. For example, to find the probability of Wikipedia covering a man from the *Chambers* source list, we first assign values to his attributes: 1 for *Male*, 1 for *Wikipedia*, 1 for *Chambers*, 1 for *Male in Wikipedia*, and 0 for all other variables. Then we use the significant coefficients in the regression results to assess the probability of inclusion: $Pr[\text{Coverage} = 1] = \frac{1}{1 + \exp(4.5621 + 1.686*1 + -3.8641*1 + -3.9654*0 + -5.4119*0 + 0.5522*1)} = 95\%$. The expected 95% coverage rate aligns with the empirical coverage rates shown in Table 2, which shows that for *Chambers*, Wikipedia misses 5% of its males.

To assess the goodness-of-fit,² we use a likelihood ratio test to compare the deviance of our model to the deviance of the model without any variables, or the null deviance. The difference between the null deviance and the residual deviance, 1422, is evaluated according to a chi-squared distribution with 6 degrees of freedom. Because the difference of 1422 is larger than the critical value of 16.81, the model fits the data reasonably well.

The finding of this analysis is contingent on a couple of factors. First, these results are dependent upon data gathered from the Web and on our approaches to determining whether an article exists, and if it does, its word length and gender. These approaches and our assessment of their reliability are described in the Methods section. To assess the effect of the subjects with unidentified gender on the findings noted previously, we reran the model, classifying all the unidentified subjects as male and then as female. Our results are similar, suggesting that the inclusion of the unknown gender does not influence the findings. Second, this reduced-form model identifies a correlation between gender and coverage, but cannot establish causation. If a correlation exists between gender and an unobserved variable like time period, then our results would indicate gender bias. For example, if the men in a listing were from the 20th century and the women from the 19th century, then the differences in coverage could be caused by Wikipedia's strength on current noteworthy figures, but the results would identify a link between men and better coverage. We investigated this possibility in the *Chambers* data and reran the regressions with controls for the historical era of the subjects and still found a significant gender bias in Wikipedia coverage that is absent in *Britannica*. We therefore believe that our findings are robust to alternative explanations.

Article Length Analysis

For our source lists of biographies, Wikipedia has greater female biographical coverage than does *Britannica* in absolute terms. However, we have also found that articles on women are more likely to be missing than are articles on men in Wikipedia relative to *Britannica*. What happens when we go beyond article coverage and consider article length?

A small number of subjects are covered with extremely long articles, leading to a highly right-skewed distribution of the article length. To offset the large outliers, like the extensive article on Jesus in *Britannica*, we compare the medians of article lengths between Wikipedia and *Britannica*.

² The degree to which the observed frequencies of occurrence of events in an experiment correspond to the probabilities in a model of the experiment. Also known as best fit.

Table 4. Median Values of Article Length.

Source	Wikipedia		<i>Britannica</i>		% Lift WP over EB	
	Female	Male	Female	Male	Female	Male
<i>ANBO</i>	585.5	906	311.5	257	88%	253%
<i>Atlantic</i>	3,689.5	4,611	619.5	1,598	496%	189%
<i>Chambers</i>	1,260.5	1,109.5	317	286	298%	288%
<i>NWHP</i>	1,411	N/A	405	N/A	248%	
<i>TIME</i>	3,099	3,529	442.5	514	600%	587%
<i>Wikipedia</i>	712	851	339.5	199	110%	328%

As shown in Table 4, Wikipedia articles are consistently larger than those of *Britannica* across all source lists. The median article length in Wikipedia ranges from 88% longer to 600% longer than that of *Britannica* for same-gender, same-source subjects. The longest articles in both Wikipedia and *Britannica* are for subjects from *The Atlantic*, an expected result because those names constitute an attempt to identify the 100 most influential people. Interestingly, no consistent difference in article length exists between males and females. For example, the median number of words in Wikipedia articles on female subjects from *Chambers* is higher than that for male subjects. At *Britannica*, median article lengths for men are larger for biographies from *The Atlantic* and *TIME*, but not for the others. Therefore, the median analysis does not imply a consistent relationship between article length and gender.

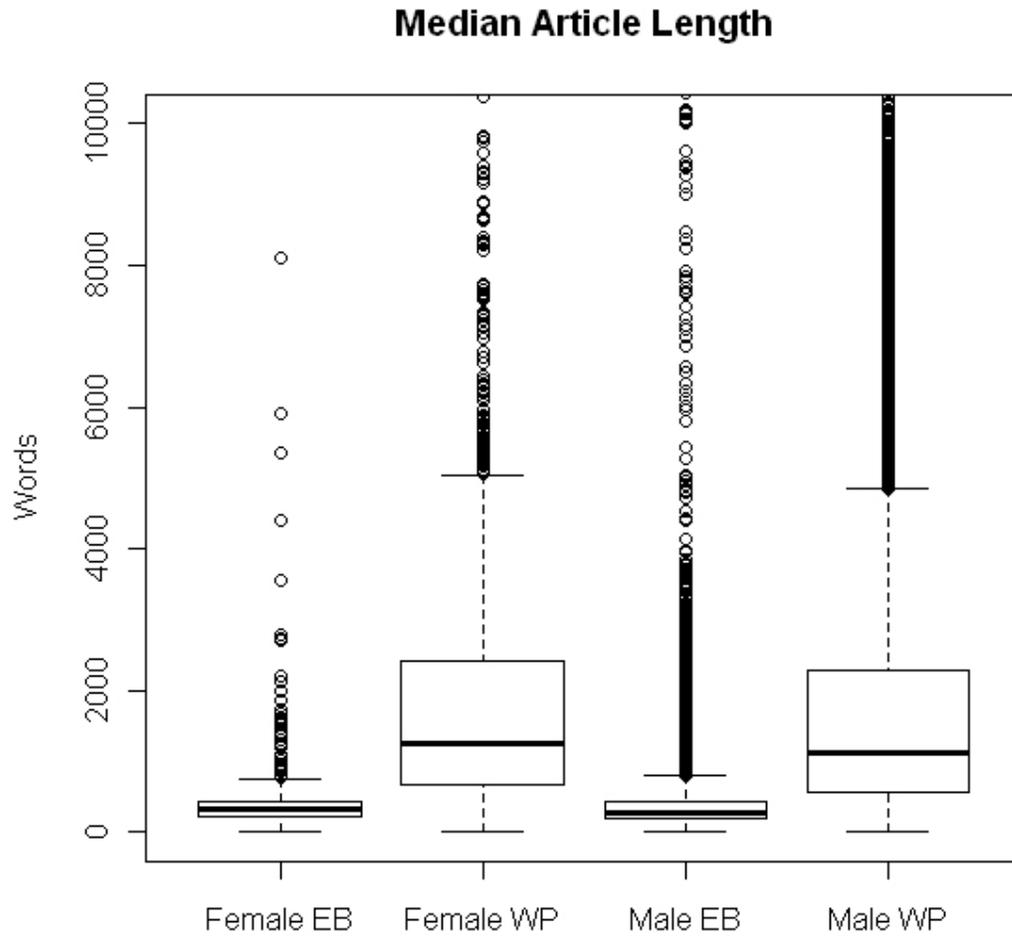


Figure 1. Boxplots for Length Distributions.

Comparison of Wikipedia and *Britannica* Bias

We can also use a regression to test the relationship between gender and the article length. The dependent variable is the number of words per article. The independent variables are similar to those in the coverage analysis: *Male*, *Chambers*, *ANBO*, *TIME*, *Wikipedia*, and *Male in Wikipedia*. These variables control for differences among the source lists, but the analysis is primarily focused on the *Male in Wikipedia* coefficient.

The dataset for the article length analysis is subject to the same restrictions as is the one used for the earlier logistic regression—a reduced sample of names from *Chambers* and exclusion of names from the NWHP and Wikipedia source lists. In addition, subjects missing from one or both works are eliminated from the data to reduce any bias towards Wikipedia's better coverage. The new data contain 1,142 subjects from the four lists with 158 female subjects and 984 male subjects. There are no subjects with unknown gender, so that variable is dropped from the analysis.

Quantile regression is used to estimate the effects of gender bias at different places in the article length distribution. Longer articles could indicate subjects with more societal importance, and the treatment of noteworthy figures may systematically differ by gender. For example, biographies of lower profile subjects with a lower word count may be similar in length irrespective of gender. Longer articles, covering high profile subjects, may differ in length systematically by gender. Quantile regression would uncover these differences in the relationship between article length and gender. For this reason, quantile regression is more robust to the large outliers present in the distribution. Based on the median analysis, we would not expect a systematic gender difference in article length at the median (50th quantile).

Quantile Regression Results: Gender and Article Length

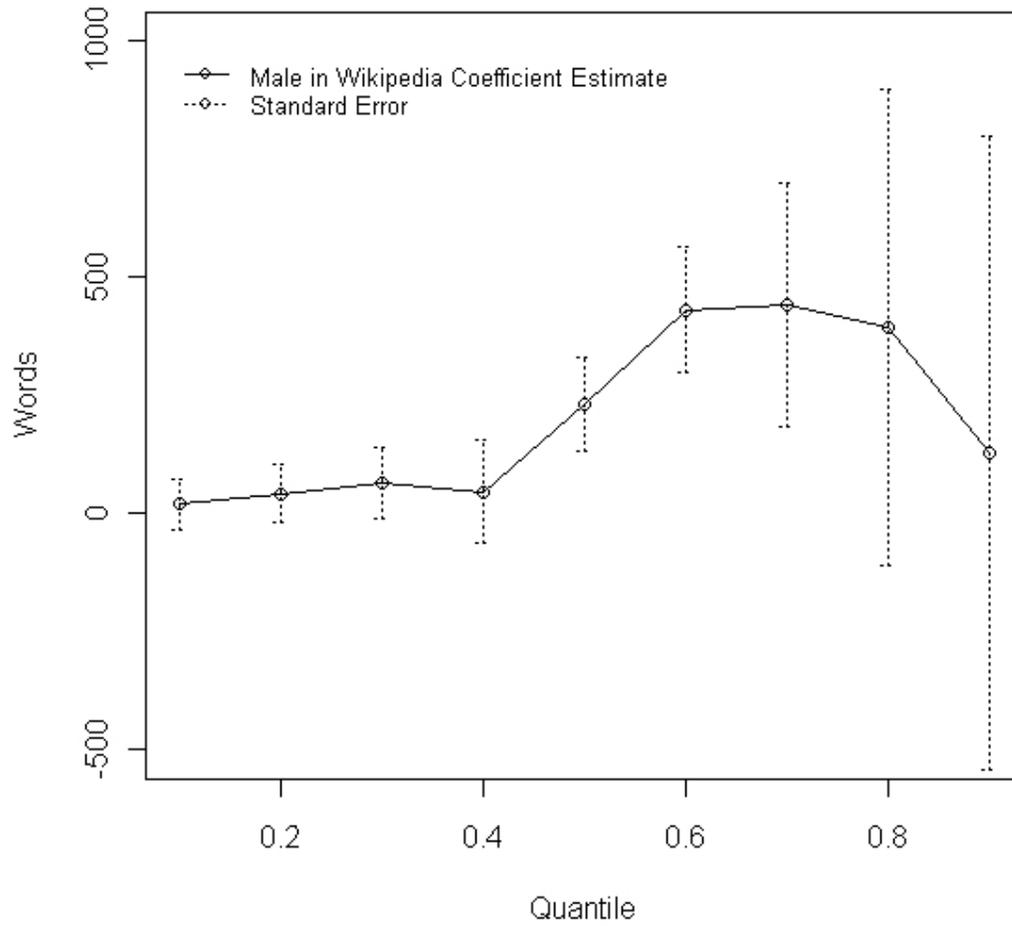


Figure 2. Influence of Gender in Wikipedia Article Length.

Figure 2 depicts the *Male in Wikipedia* coefficient estimates and standard errors for quantile regressions on each decile of the article length. Similar to the preceding section, the *Male in Wikipedia* coefficient measures the relative increase or decrease in words associated with being a male subject in a Wikipedia entry. The coefficient is significant for only 6th and 7th deciles; however, there is no significant effect in the lower or higher quantiles of the distribution.

Although not the primary focus, the source list coefficients are significant for each decile and confirmed our earlier findings about differences in median article length among different source lists. The *Male* coefficient is significant for the 2nd quantile with p-values of $p < 0.05$, but it is not significant at other deciles. The Wikipedia coefficient is significant across all deciles, confirming that Wikipedia articles are consistently longer than are *Britannica* articles.

The quantile regression results are not conclusive to gender bias in Wikipedia or *Britannica's* article lengths. The gender variables' coefficients are not significant for the majority of the deciles, including the extremes. There is not a straightforward explanation for the significant increase of male article length for the 6th and 7th deciles, but there are alternative explanations to gender bias in Wikipedia. Perhaps many of the moderately noteworthy figures share a common unobserved element, like being professional athletes who naturally skew male. Without more definite results or the ability to eliminate the alternative explanations, gender bias can be only one possible explanation for the significant difference at the 6th and 7th deciles on the *Male in Wikipedia* coefficient and at the 2nd decile of the *Male* coefficient.

Limitations and Further Research

The present work is limited in that it takes place in a sociohistorical context of existing bias. Hence, the presented method is a comparison of Wikipedia coverage relative to specific sources for biographies and to *Britannica's* coverage. Also, this work is wholly focused on article coverage and length; we do not speak to the character or quality of article content. While we were motivated by reports of gender imbalance in contributors, this study does not address a causal relationship between contributors and content. First, we do not know the gender balance of contributors to other lists and works. (And what we know of Wikipedia is based on surveys. Anecdotally, a participant in the Wikipedia Feminism Task Force noted that most of their articles are not written by women.) Second, other factors such as pre-existing historical bias are present and may be dominant. Even so, with respect to quality, one possible avenue open to the present method would be to look at coverage according to Wikipedia-designated quality categories, for example, "featured" and "good" articles (Wikipedia, 2009b). Finally, it would be useful to repeat the last analysis—where we tested *Britannica* against a sample of all Wikipedia biographies—using *Britannica*, if a sample of all online *Britannica* biographic articles could be obtained.

Conclusion

Our efforts to collect and compare data across thousands of articles at Wikipedia and *Britannica* permit us to report a number of novel findings. First, with respect to article coverage and length, we conclude that Wikipedia has significantly greater coverage than does *Britannica* because the percent of missing articles is higher for *Britannica* than it is for Wikipedia; *Britannica* has only about 10% of the articles from a random sample of Wikipedia biographies (Table 2). Also, Wikipedia articles are significantly longer than *Britannica* articles (Table 4). The median article length for Wikipedia is larger than that of *Britannica* for every source list.

Second, we take a couple of approaches in attempting to describe and compare gender imbalances in biographical coverage, each with their own finding. In looking at the gender imbalance in listings of notable persons, queries to three large comprehensive biographical databases yield 15%, 24.5%, and 28.7% female coverage. From the biographical source lists we used, it seems that contemporary lists of notable persons are more likely to have more women than are older lists or those with more historic figures. With respect to the percentage of women in a reference work, Wikipedia and *Britannica* both roughly mirror the bias of the source list, with Wikipedia performing slightly better than does *Britannica* in most cases (Table 1). While Wikipedia has more biographies of women than does *Britannica* in absolute terms (Table 1), Wikipedia tends to be less balanced in whom it misses than is *Britannica* as seen in the percentages of missing articles (Table 2) and the positive and significant *Male* coefficient in the logistic regression (Table 3). However, gender is inconsistently correlated with article length (Figure 2). Although the coefficients on male subjects are significantly different from zero for certain deciles in the article length distribution, gender does not influence article length in the majority of deciles, suggesting that gender bias may not be a strong factor for article length. That is, if a subject is deemed notable enough to warrant inclusion in Wikipedia and *Britannica*, then the subjects, regardless of gender, may be treated similarly by the contributors.

Overall, we find evidence of gender bias in Wikipedia coverage of biographies. While Wikipedia's massive reach in coverage means one is more likely to find a biography of a woman there than in *Britannica*, evidence of gender bias surfaces from a deeper analysis of those articles each reference work misses. That Wikipedia's missing articles are disproportionately female relative to those of *Britannica* is seen in a comparison of the ratio of female to male subjects in each work and in the related logistic regression.

References

- Adar, E., Skinner, M., & Weld, D. S. (2009, February 9–12). Information arbitrage across multi-lingual Wikipedia. Paper presented at the Second ACM International Conference on Web Search and Data Mining (WSDM'09), Barcelona, Spain.
- Beard, M. R., Edinger, D., Selig, J. A., & White, M. (1977). A study of the *Encyclopaedia Britannica* in relation to its treatment of women. In A. J. Lane (Ed.), *Mary Ritter Beard: A sourcebook—studies in the life of women* (pp. 215–224). New York: Schocken Books.
- Bragues, G. (2007, April). Wiki-philosophizing in a marketplace of ideas: Evaluating Wikipedia's entries on seven great minds. Retrieved from Social Science Research Network at <http://ssrn.com/abstract=978177>
- Cawdrey, R. (1997). *A table alphabetical of hard usual English words* (I. Lancashire, Ed.). Web Development Group, University of Toronto Library. (Original work published 1604). Retrieved from <http://www.library.utoronto.ca/utel/ret/cawdrey/cawdrey0.html>
- Cunningham, W. (2003, November). Correspondence on the etymology of Wiki. Retrieved from <http://c2.com/doc/etymology.html>
- Cunningham, W. (2004). Wiki design principles. Retrieved from <http://c2.com/cgi/wiki?WikiDesignPrinciples>
- Einbinder, H. (1964). *The myth of the Britannica*. New York: Grove Press.
- Ghosh, R. A., Glott, R., Krieger, B., & Robles, G. (2002, June). *Free/libre and open source software: Survey and study*. The Netherlands: International Institute of Infonomics University of Maastricht. Retrieved from http://flossproject.org/report/FLOSS_Final4.pdf
- Giles, J. (2005, December 14). Internet encyclopaedias go head to head. *Nature*, 438, 900–901. Retrieved from <http://www.nature.com/nature/journal/v438/n7070/full/438900a.html>
- Glott, R., Schmidt, P., & Ghosh, R. A. (2010, March). Wikipedia survey—Overview of results. Retrieved from UNI-MERIT at http://www.wikipediaurvey.org/docs/Wikipedia_Overview_15March2010-FINAL.pdf
- Halavais, A., & Lackaff, D. (2008). An analysis of topical coverage of Wikipedia. *Journal of Computer-Mediated Communication*, 13, 429–440.
- Hecht, B., & Gergle, D. (2009, June 25–27). Measuring self-focus bias in community-maintained knowledge repositories. *Proceedings of the fourth international conference on communities and*

- technologies*. University Park, PA: Penn State University. Retrieved from http://www.brenthecht.com/papers/bhecht_CommAndTech2009.pdf
- Hecht, B., & Gergle, D. (2010, April 10–15). *The Tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context*. Paper presented at the CHI 2010 (ACM Conference on Human Factors in Computing Systems, Atlanta, Georgia). Retrieved from http://www.brenthecht.com/papers/bhecht_chi2010_towerofbabel.pdf
- Karanović, J. (2008). *Sharing publics: Democracy, cooperation, and free software advocacy in France* (Chapter 1, pp. 74–116). Early draft of unpublished doctoral dissertation chapter with different pagination, New York University.
- Limey, J. (2010, January 14). Who's on Wikipedia? Part 2: Gender and nationality. Retrieved from On Wikipedia at <http://onwikipedia.blogspot.com/2010/01/whos-on-wikipedia-part-2-gender-and.html>
- Lin, Y. (2005, November 23). Gender dimensions of FLOSS development. *Mute Magazine*, 3(1). Retrieved from <http://www.metamute.org/en/Gender-Dimensions-of-Floss-Development>
- Maack, M. N. (2004, March 22). The lady and the antelope: Suzanne Briet's contribution to the French documentation movement. *Library Trends*, 52(4), 719–747. Online preprint retrieved from <http://www.gseis.ucla.edu/faculty/maack/BrietPrePress.htm>
- Morton, H. C. (1994). *The story of Webster's Third: Philip Gove's controversial dictionary and its critics*. New York: Cambridge University Press. Retrieved from <http://books.google.com/books?id=1dKJrIRXhFgC>
- NCWIT. (2007). NCWIT Scorecard 2007: A report on the status of women in information technology. Retrieved from National Center for Women and Information Technology at http://ncwit.org/pdf/2007_Scorecard_Web.pdf
- Nov, O. (2007, November). What motivates Wikipedians? *Communications of the ACM*, 50(11), 60–64. Retrieved from <http://portal.acm.org/citation.cfm?id=1297798>
- Reagle, J. (2010, August). *Results of reference work analysis*. Retrieved from <http://reagle.org/joseph/2010/06/qender/results.html>
- RMJ. (2010, September 2). Wikipedia's main page mentions nine men for every one woman. Retrieved from Deeply Problematic: <http://www.deeplyproblematic.com/2010/09/wikipedias-main-page-mentions-nine-men.html>

- Ruskin, J. (1910). Sesame and lilies: Lecture II - Lillies: Of Queens' Gardens. In C. W. Elliot (Ed.), *Essays, English and American: With introductions and notes* (Vol. 28, pp. 139–168). New York: P. F. Collier & Son. Retrieved from <http://books.google.com/books?id=xowEAAAAYAAJ>
- Thomas, G. (1992). *A position to command respect: Women and the Eleventh Britannica*. Metuchen, NJ: The Scarecrow Press.
- Wiegand, W. A. (1996). *A biography of Melvil Dewey: Irrepressible reformer*. Chicago: American Library Association.
- Wikipedia. (2004, October 4). Wikipedia:WikiProject countering systemic bias. Retrieved from Wikipedia at <http://en.wikipedia.org/?oldid=6332105>
- Wikipedia. (2008, January 11). Wikipedia: WikiProject countering systemic gender bias. Retrieved from Wikipedia at <http://en.wikipedia.org/?oldid=183541656>
- Wikipedia. (2009a, December 21). Wikipedia talk: WikiProject gender studies/feminism task force/archive 2. Retrieved from Wikipedia at <http://en.wikipedia.org/?oldid=333015320>
- Wikipedia. (2009b, May 14). Wikipedia: Version 1.0 editorial team/assessment. Retrieved from Wikipedia at <http://en.wikipedia.org/?oldid=289898436>
- Wikipedia. (2009c, December 12). Wikipedia: WikiProject countering systemic bias. Retrieved from Wikipedia at <http://en.wikipedia.org/?oldid=331177310>
- Wikipedia. (2009d, March 3). Wikipedia: WikiProject gender studies/countering systemic gender bias. Retrieved from Wikipedia at <http://en.wikipedia.org/?oldid=274610658>
- Wikipedia. (2009e, March 6). Wikipedia: WikiProject persondata/list of biographies. Retrieved from Wikipedia at <http://en.wikipedia.org/?oldid=275324859>