Wikipedia as a Data Source for Political Scientists: Accuracy and Completeness of Coverage

Adam R. Brown, Brigham Young University

ABSTRACT In only 10 years, Wikipedia has risen from obscurity to become the dominant information source for an entire generation. However, any visitor can edit any page on Wikipedia, which hardly fosters confidence in its accuracy. In this article, I review thousands of Wikipedia articles about candidates, elections, and officeholders to assess both the accuracy and the thoroughness of Wikipedia’s coverage. I find that Wikipedia is almost always accurate when a relevant article exists, but errors of omission are extremely frequent. These errors of omission follow a predictable pattern. Wikipedia’s political coverage is often very good for recent or prominent topics but is lacking on older or more obscure topics.

Since its 2001 launch, Wikipedia has ballooned into the world’s largest and most popular reference site on the Internet, eclipsing all competitors and putting Microsoft’s Encarta out of business entirely.1 Over 12 million visitors have registered as volunteer Wikipedia editors, with 150,000 active editors as of May 2010.2 This anonymous army has written every word on the site. At the same time, Wikipedia’s “crowd-sourcing” approach to content creation produces little confidence in its accuracy. One of Wikipedia’s more infamous scandals emerged in 2007, when prolific editor “Essjay” was found to have used false credentials to persuade other volunteers to accept his edits on thousands of articles. For years, nobody knew that Essjay was not the tenured professor of theology and canon law he claimed to be—in fact, he was a community college dropout with no qualifications whatsoever. Wikipedia founder Jimmy Wales acknowledges that episodes like this one have hurt his site’s credibility, noting that “people do need to be aware of how [Wikipedia] is created and edited so they can treat it with the appropriate caution” (Time 2007).

Surprisingly, studies of Wikipedia’s accuracy have generally found worries about its credibility to be overblown. Most famously, Nature commissioned blind, expert reviews of 50 articles on chemistry, physics, biology, and other “hard” scientific fields from Wikipedia and Britannica and found that both sources had similarly low rates of “serious errors” (Giles 2005). Later reviews have affirmed Wikipedia’s accuracy on subjects in pharmacology (Clau- son et al. 2008), medicine (Devgan et al. 2007; Nicholson 2006; Pender et al. 2008), history (Rector 2008), philosophy (Bragues 2007), and library science (Royal and Kapila 2009). Contrary to what we might expect, then, Wikipedia seems to be roughly as accurate as established reference sources.

To date, however, nobody has reviewed Wikipedia’s accuracy about politics, which, by nature, deals with contentious people, institutions, and policies. Wikipedia’s democratic approach to content creation would seem more prone to abuse when dealing with candidates than chemistry. In fact, some of Wikipedia’s biggest scandals have arisen from politically motivated editing. In 2006, Wikipedia’s administrators found that congressional staffers had been whitewashing their bosses’ biographies in Wikipedia and inserting negative information about their rivals. And, in the 24 hours before John McCain announced Sarah Palin as his running mate, a McCain staffer rewrote Palin’s Wikipedia profile (Cohen 2008).3 Given the unusually strong incentives for misrepresentation in politically relevant articles, we should not assume that Wikipedia’s political coverage is as accurate as its coverage of other fields.

In this article, I review the quality of Wikipedia’s political data. I begin by describing my assessment methodology. Applying this method, I find that even for political topics, Wikipedia is generally accurate, although its spotty coverage produces many errors of omission. This first finding begs a follow-up: If omissions are Wikipedia’s biggest problem, which political topics receive the most thorough coverage? I find that Wikipedia is heavily skewed toward recent, prominent political topics and provides extremely poor coverage of events that occurred prior to its 2001 creation. Finally, I find that individual politicians receive better coverage if they are technologically savvy enough to edit their own profiles.

ASSESSING WIKIPEDIA’S ACCURACY All of the assessments of Wikipedia cited previously have adopted Nature’s approach: The reviewers select a small number of seminal topics within a field and then check the accuracy of every
ACCURACY OF CANDIDATE BIOGRAPHIES

Using this approach, I begin by reviewing Wikipedia’s articles about every major-party gubernatorial candidate who ran between 1998 and 2008, examining entries on both prominent candidates (e.g., Arnold Schwarzenegger, Jon Corzine, Eliot Spitzer) and obscure ones (e.g., Jay Lucas, Doug Gross, Mark O’Keefe). One hundred fifty-five gubernatorial elections took place between 1998 and 2008, contested by 246 distinct general election candidates. Of these 246 candidates, 230 (93%) were the subject of a dedicated Wikipedia article. For each of these 230 articles, I checked a specific fact: whether Wikipedia accurately characterized the candidate’s previous political experience.5 I found no errors in these articles at all. Furthermore, this record is even better than it sounds. The experience level of many of the candidates who ran for governor more than once during this period changed between their candidacies. For example, Bobby Jindal ran for Louisiana governor as a newcomer in 2003 and then ran for governor again as a U.S. Representative in 2007. Likewise, Linda Lingle ran for Hawaii governor as Maui’s mayor in 1998 and then ran for governor again as chair of Hawaii’s Republican Party in 2002. Overall, the sample included 50 occasions in which a previous candidate ran for office again, but with a change in political experience.6 In every case, Wikipedia provided sufficient detail in its biographies to identify every candidate’s political experience at the time of each distinct candidacy.

For the 230 candidates with a dedicated article, then, Wikipedia contained no errors. For the remaining 16 candidates, however, Wikipedia contained no information at all. These 16 candidates were mentioned by name in other relevant articles, particularly those articles on their opponents, but no information other than their names was offered. Wikipedia simply did not contain any information about these candidates. Thus, it was accurate regarding candidates’ previous experience, but included no information on 7% of candidates. I will further explore these errors of omission later in the article.

ACCURACY OF ELECTION RESULTS

I also checked the accuracy of Wikipedia’s reported election results. For each election year back to 1976, Wikipedia had an article such as “United States Gubernatorial Elections, 2000”; each of these entries recorded the results of all general elections held that year. For each state, I compared the major-party candidates’ vote margins to each state’s official results, rounded to one-tenth of a percent. Wikipedia’s reported margin was exactly accurate in 61% of the 155 races held between 1998 and 2008. In another 27% of races, Wikipedia was accurate within rounding error.7 In only four (2.6%) gubernatorial elections was Wikipedia’s margin off by more than one percentage point. Wikipedia’s most significant error related to the 2002 New Mexico race, accurately reporting that Bill Richardson won 55.5% of the vote but inaccurately claiming that John Sanchez won 44.9% of the vote, instead of his official 39.0%. Thus, Wikipedia overstated the correct margin by +4.1 percentage points. Wikipedia’s next largest errors were for the New Hampshire 1998 (-2.2), New York 2006 (+1.2), and New Hampshire 2006 (-1.2) gubernatorial races. On the whole, though, Wikipedia’s election results were mostly accurate. There were no errors of omission in these data; Wikipedia had data for every gubernatorial election back to 1998. A statistical analysis based on Wikipedia’s reported election results would return essentially the same results as an analysis relying on official data.

WIKIPEDIA’S ERRORS OF OMISSION

It seems, then, that Wikipedia suffers less from inaccuracies than omissions. Scholars in other fields have come to the same conclusion. Pharmacologists report, for example, that “no factual errors were found in Wikipedia, [but] errors of omission were [frequent]” (Clauson et al. 2008). Computer scientists have found that an article’s length is the best predictor of its overall quality (Blumenstock 2008), implying that omissions rather than inaccuracies are Wikipedia’s greater flaw. As an online encyclopedia that seeks to contain “the sum of all human knowledge,”8 those omissions do not result from space constraints; rather, they reflect the limited expertise and interests of contributors. Therefore, we might profitably ask whether any general variables can predict which topics are likely to receive the best coverage. To a large degree, we can probably expect that the same forces that make political topics “salient” to public opinion (cf. Zaller 1992) also make them likely to be addressed in Wikipedia. That is, we should expect Wikipedia to have greater coverage of political topics that are either recent or prominent.

Wikipedia’s bias toward recent events is apparent in table 1, which depicts the percentage of midterm-year gubernatorial elections with a dedicated Wikipedia article (such as “California
Wikipedia's Coverage of Historical Gubernatorial Elections

<table>
<thead>
<tr>
<th>YEAR</th>
<th>NUMBER OF GUBERNATORIAL RACES</th>
<th>RACES WITH A DEDICATED ARTICLE</th>
<th>PERCENT WITH A DEDICATED ARTICLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>37</td>
<td>37</td>
<td>100</td>
</tr>
<tr>
<td>2006</td>
<td>36</td>
<td>36</td>
<td>100</td>
</tr>
<tr>
<td>2002</td>
<td>36</td>
<td>18</td>
<td>50</td>
</tr>
<tr>
<td>1998</td>
<td>36</td>
<td>13</td>
<td>36</td>
</tr>
<tr>
<td>1994</td>
<td>36</td>
<td>10</td>
<td>28</td>
</tr>
<tr>
<td>1990</td>
<td>36</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>1986</td>
<td>36</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>1982</td>
<td>36</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>1978</td>
<td>36</td>
<td>5</td>
<td>14</td>
</tr>
</tbody>
</table>

Gubernatorial Election, 2010). Wikipedia offered a separate article about every gubernatorial election that has been held since the website's rise to prominence in the mid-2000s, as well as articles about many of the gubernatorial elections that occurred during its infancy. However, coverage of elections that occurred prior to its creation in 2001 was extremely poor. Even for elections held as recently as 1998, only 36% had an article. Prior to 1990, coverage dries up completely, with the exception of four states (California, Maine, Minnesota, and South Carolina), for which dedicated volunteers have produced articles going back decades. However, there were virtually no articles about gubernatorial elections in other states prior to the mid-1990s. Consider also Wikipedia's coverage of the U.S. House. All 435 members who served in 2009 had an article, as did all 325 who served in 1887. However, the median article about a member of the 111th Congress was five times longer than the median article about a member of the 50th.

A better way of describing this “recentness” bias is to say that Wikipedia's volunteer editors are more likely to write about political topics about which they are actively thinking. An event will be more prominent in their minds if it occurred recently than far in the past, of course. More generally, a topic that has received more attention from the public and political observers will be more prominent. For example, the U.S. president receives far more media attention than governors, who receive far more attention than state legislators. Fittingly, then, Wikipedia had an article on every U.S. presidential election since 1788 describing the relevant candidates, issues, and election results. Moving down a notch, the website offered decent coverage of gubernatorial elections of recent years but very little coverage of elections held prior to 1978, as noted previously. Moving down another notch, state legislative elections received almost no coverage at all.

For articles about events (such as an election) or ideas (such as direct democracy), we might expect this sort of “prominence” to be the dominant predictor of Wikipedia's political coverage. For articles about people, however, we might also expect coverage to increase when the subject of the article possesses enough technological savvy to edit his or her own profile. Wikipedia's users are younger and more educated than the general population. In a 2007 study, only 26% of people over the age of 65 claimed to use Wikipedia, but 38% of respondents between the ages of 30 and 49 and 44% of those between the ages of 18 and 29 reported using it. Likewise, only 22% of people with a high school diploma use Wikipedia, but 50% of people with a college degree or higher do so (Rainie and Tancer 2007). As such, we might expect coverage to improve for young or highly educated politicians, since these individuals are more likely to contribute to their own Wikipedia profiles.

Thus, we might expect to find two general biases: a bias toward prominent political topics (for all articles) and a bias toward younger, educated politicians (for articles about living people). To quantify the strength of these two biases, I gathered data on 7,361 state legislators serving in late 2009. Of these, 3,512 (48%) were the subject of a Wikipedia article. The median article contained 2,026 characters (1,099 after formatting data and category labels were removed). The shortest article contained a mere 218 characters (77 characters after cleanup). The longest, for New York's Representative Greg Ball, contained 57,978 characters (41,663 after cleanup).

Legislators vary widely in both their prominence and their technological savvy. Within a state, legislators can rise in prominence through leadership experience, tenure, or upper chamber service. Across states, legislators are more prominent in states where average district sizes are larger. New Hampshire’s representatives have only about 4,000 constituents each who might write about them on Wikipedia, whereas California’s state senators enjoy almost one million constituents each. By a similar logic, legislators may also be more prominent if their chamber is smaller. Alaska’s political junkies only need follow 20 senators, but New Hampshire’s junkies must track 400 representatives. These various indicators of prominence are outlined along with other variables in table 2.

Using probit analysis, Model 1 predicts whether each legislator will have an article in Wikipedia. Legislative leadership has a strong effect on the likelihood of an article existing; a leader’s probability of having a dedicated article is 0.20 higher than an otherwise identical non-leader’s (see table 3). Logged district population has a similarly strong effect; a one-standard-deviation increase results in a 0.23 increase in the predicted probability that a legislator will have an article. As an illustration, Ohio’s 33 senators have the fifth-largest districts in the nation and all 33 have separate Wikipedia articles. Ohio’s 99 representatives, by contrast, come from average-sized districts, and only 40% have articles. Political experience also has a significant but more modest effect on the likelihood of an article existing; a leader’s article will be 120% toward technologically savvy politicians. At the margin, for example, OLS predicts that a legislative leader’s article will be 120% longer than an otherwise identical non-leader’s; in addition, a 1% increase in district population is associated with a 1.355% increase in article length. The tobit estimates in Model 3 are even larger.
suitable as a sole source for students, who ought to be consulting other sources, longtime politicians, and legislators with larger constituencies. Of course, even if Wikipedia is perfectly accurate, it is not predictable pattern: coverage is best on topics that are more recent and impactful, but when it does contain information, its accuracy is much better than the 50.1–49.9 margin, rounding error would allow for 0.2 percentage points.

### DISCUSSION

Like reviewers from non-political fields, I find that Wikipedia’s political data does appear to be accurate, although it is marked by severe errors of omission. Wikipedia’s coverage may not be perfect, but when it does contain information, its accuracy is much better than we might suppose. Wikipedia’s omissions follow a predictable pattern: coverage is best on topics that are more recent or prominent. Using state legislators as an example, I find that the depth of Wikipedia’s coverage improves for legislative leaders, longtime politicians, and legislators with larger constituencies. Of course, even if Wikipedia is perfectly accurate, it is not suitable as a sole source for students, who ought to be consulting better resources than encyclopedias. As with any encyclopedia, Wikipedia’s usefulness lies in readers’ ability to gain a quick feel for a subject; serious research should always be based on reliable primary and secondary sources, not on a tertiary (at best) source like Wikipedia.

At the same time, Wikipedia’s surprising accuracy may allow it to play a useful role for researchers with limited resources. Suppose that a political scientist proposes a hypothesis that a certain variable influences election results. For some quick preliminary analysis, he or she might save some time by pulling election results from Wikipedia and entering them into a spreadsheet. A preliminary analysis based on these data might even be acceptable in a conference paper. Indeed, some existing conference papers have used Wikipedia to temporarily fill in missing data (e.g., Meredith 2010)—an approach that may well improve on interpolation. If our hypothetical political scientist decided to pursue his or her project further or publish results, he or she would of course need to locate more reliable data; no serious political scientist should base published research on Wikipedia’s data. Still, for political scientists with limited time and research assistance, Wikipedia may be just accurate enough to permit its use in preliminary work. ■

### NOTES

1. Two major analytics providers, Alexa and Compete, rank Wikipedia among the top ten websites overall. In a typical month, Compete reports that Wikipedia receives 30 times more visitors than its nearest competitor, Britannica. Microsoft has not officially blamed Wikipedia for Encarta’s 2009 demise, but see Gralla (2009) and McDougall (2009).

2. “Active” means that a user had logged in and made at least one edit to a Wikipedia article within the past 30 days.


4. For example, Wikipedia identified Mendeleev as the fourteenth instead of thirteenth child in his family, misspelled “pyrocollodion,” and was unclear about the time of Mendeleev’s wedding. These errors are irrelevant to understanding Mendeleev’s scientific contributions.

5. To compile official data for comparison, I employed a team of research assistants to search through local news archives using Lexis-Nexis, as well as through volumes of Who’s Who in American Politics, which sends out biographical questionnaires to elected officials.

6. In most of these cases, the change was simply to identify the governor seeking reelection as the incumbent.

7. If the official result was 50.1–49.9, a 0.2 margin, rounding error would allow for a 50.2–49.8 report, a 0.4 margin. Hence, rounding error in the vote margin allows for 0.2 percentage points.

Table 2

<table>
<thead>
<tr>
<th>Predicting Wikipedia’s Coverage of State Legislators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prominence</td>
</tr>
<tr>
<td>Legislative Leader</td>
</tr>
<tr>
<td>Years in Politics (Logged)</td>
</tr>
<tr>
<td>Upper Chamber</td>
</tr>
<tr>
<td>District Population (1,000s, Logged)</td>
</tr>
<tr>
<td>Number of Districts</td>
</tr>
<tr>
<td>Technological Savvy</td>
</tr>
<tr>
<td>Age (Logged)</td>
</tr>
<tr>
<td>High School or Less</td>
</tr>
<tr>
<td>Graduate Degree</td>
</tr>
<tr>
<td>Other Variables</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Republican</td>
</tr>
<tr>
<td>Constant</td>
</tr>
<tr>
<td>Model-Specific Statistics</td>
</tr>
<tr>
<td>Modal: 51.7%</td>
</tr>
<tr>
<td>Predicted: 71.6%</td>
</tr>
<tr>
<td>N</td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>Predicted Probabilities of Wikipedia Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHANGE IN X</td>
</tr>
<tr>
<td>Legislative Leader</td>
</tr>
<tr>
<td>Years in Politics (Logged)</td>
</tr>
<tr>
<td>District Population (1,000s, Logged)</td>
</tr>
<tr>
<td>Age (Logged)</td>
</tr>
<tr>
<td>Graduate Degree</td>
</tr>
</tbody>
</table>

Note: Other variables are held at their mean or mode. A 1 SD change in X refers to a shift from 0.5 standard deviations below the mean to 0.5 standard deviations above it.

Notes: I thank Katrina Smith Cammack and Matt Ipson for their expert research assistance. Any faults that remain are my own.

6. In most of these cases, the change was simply to identify the governor seeking reelection as the incumbent.

7. If the official result was 50.1–49.9, a 0.2 margin, rounding error would allow for a 50.2–49.8 report, a 0.4 margin. Hence, rounding error in the vote margin allows for 0.2 percentage points.
8. The quotation is from Wikipedia founder Jimmy Wales; see http://

9. The article read: “John L. Scott Jr. is a member of the South Carolina House
of Representatives.”

10. Legislator characteristics are drawn from Project Vote Smart. “Years in Poli-
tics” measures the time since the legislator’s first political experience (e.g., as a
candidate, appointee, local government official, etc.). Age and education are
available for only 4,813 legislators, but table 2 would have similar coefficients
and significance levels even if these variables were omitted (producing $N$
= 7,009).

11. That is, an increase from a one-half standard deviation below the logged
mean (35,928 average district population) to a one-half standard deviation
above it (104,946 average district population).

12. In the words of Wikipedia founder Jimmy Wales, “You’re in college; don’t cite
the encyclopedia” (Young 2006).

REFERENCES
Blumenstock, Joshue E. 2008. “Size Matters: Word Count as a Measure of Quality
on Wikipedia.” Paper presented at the 17th International Conference on the
World Wide Web, Beijing, China, April 21–25.

Bragues, George. 2007. “Wiki-Philosophizing in a Marketplace of Ideas: Evaluat-
ssrn.com/abstract=975877.

Clauson, Kevin A., Hyla H. Polen, Maged N. Kamel Boulos, and Joan H. Dzeno-
wajis. 2008. “Scope, Completeness, and Accuracy of Drug Information in

Times, August 31.

Devgan, Lara, Neil Powe, Brittney Blakely, and Martin Makary. 2007. “Wiki-
Surgery? Internal Validity of Wikipedia as a Medical and Surgical Reference.”
Journal of the American College of Surgeons 205 (suppl): S76–S77.

Giles, Jim. 2005. “Special Report: Internet Encyclopedias Go Head to Head.” Na-
ture 438:900–01.

March 31. http://www.pcworld.com/article/162320/what_was_encarta_
look_it_up_on_wikipedia.html.

McDougall, Paul. 2009. “Microsoft Encarta is Web 2.0’s Latest Victim.” Informa-
showArticle.jhtml?articleID=216491974.

Meredith, Marc. 2010. “Exploiting Friends and Neighbors: An Instrumental Vari-
able Approach to Estimating Coattail Effects.” Paper presented at the 10th

Information on Wikipedia.” Master’s capstone, Oregon Health and Science
University School of Medicine.

Pender, Michael, Kaye Lasserre, Lisa Kruesi, Christopher Del Mar, and Satyamur-
presented at to the Special Libraries Association Annual Conference, Seattle,
June 16.


Rector, Lucy Holman. 2008. “Comparison of Wikipedia and Other Encyclopedias
for Accuracy, Breadth, and Depth in Historical Articles.” Reference Services

Royal, Cindy, and Deepina Kapila. 2009. “What’s on Wikipedia and What’s

time/business/article/0,8599,1601491,00.html.

Young, Jeff. 2006. “Wikipedia Founder Discourages Academic Use of His Cre-
wirecampus/wikipedia-founder-discourages-academic-use-of-his-creation/
2305.

University Press.
Education About Asia

A unique and innovative magazine, Education About Asia is a practical teaching resource for secondary school, college, and university instructors, as well as an invaluable source of information for all students who have an interest in Asia. To review subscription information, the table of contents for all issues, selected articles, as well as information on the Association for Asian Studies, please visit www.asian-studies.org.

FOCUS ON ASIA: JAPAN’S DEMOCRACY
two-part series (spring and winter 2011)

Government and political science instructors who want to improve their student’s understanding of democratic government and encourage global perspectives on civic literacy can effectively utilize the study of Japan, Asia’s oldest democracy, in their classrooms. The spring and winter 2011 issues of Education About Asia will include a two-part “Focus on Asia: Japan’s Democracy” series. ‘Japan’s Democracy” will offer feature articles and resource essays that should provide instructors illumination about a variety of aspects of Japanese democracy. Much of the content will also include a US-Japan comparative focus.

Funding generously provided by the United States-Japan Foundation.

Subscribe today at www.asian-studies.org