Online Appendix to "Reference Rot: An Emerging Threat to Transparency in Political Science"

Aaron L. Gertler and John G. Bullock

October 14, 2016

In this appendix, we provide further information about the analyses in our article, describe related analyses that space constraints precluded us from presenting in the article, and discuss several related topics.

Data and Research Design

We created a database with a single record for each article-URL pair. (The same URL therefore appears in the database multiple times if it appears in multiple articles, but it earns only one entry in our database per article.) We did not record links to personal websites if they appeared only in notes that provided authors' contact information.

In November 2014, we examined each article and coded each URL as having one of three purposes:

- *Reproducibility* URLs refer to data, code, software, or other resources that readers need to reproduce the authors' results.
- *Database* URLs refer to repositories of data—e.g., papers, articles, or datasets—that the authors do not suggest are needed to reproduce the results. For example, several authors referred to http://www.census.gov without making any suggestion that the data needed to reproduce their results could be found at that URL.

- *Bibliographic source* URLs supported inherently irreproducible claims. They were used mainly to support characterizations of historical events or to link to sources from which the authors quoted.

We also distinguished between two types of resources to which URLs refer. *Institutional resources* are hosted by corporations, universities, or other organizations and are not hosted on personal websites, i.e., sites devoted to the research or interests of a single person.[1] *Personal-site resources* are located on personal websites, whether or not those sites are in turn hosted on institutional servers. Thus, resources hosted on personal websites qualify as personal-site resources even if those sites are hosted by universities.[2]

Finally, we followed ("clicked on") each URL and examined the result to determine whether the link was working. Links were classified as working if and only if they led to the intended resource. All other links were classified as broken.[3]

The set of broken links includes links that are "too vague." For example, the suggestion that a specific dataset can be found at http://www.census.gov would

---

[1] Official campaign websites—for example, http://www.dolekemp96.org—were categorized as institutional resources.

[2] As we created our dataset, we distinguished between personal-site resources that were associated with .edu domains (e.g., http://gking.harvard.edu) and those that were not (e.g., http://johnbullock.org). But across 14 years of APSR articles, we found only 28 links to personal sites that were not associated with .edu domains, and we made no use of this distinction in our analysis.

[3] We also distinguished between several types of broken links. "Page not found" links are those that lead to a blank page, redirect to the page of a domain-name seller or registrar, or produce a 404 HTTP response code. "Page found but information unavailable" links lead to the intended page (i.e., there is no redirection to a different URL), but the information to which the author referred is unavailable. This result is most likely to arise because the content of the site has changed since the author visited it. Finally, "page redirection and information unavailable" links redirect to a page that does not contain the information to which the author referred. These distinctions play no role in our analysis, but our dataset does include this finer coding of broken links.

probably lead to a broken-link classification: the dataset may exist somewhere within the census.gov website, but unless the dataset is available from http://www.census.gov itself (that is, from the Census Bureau's "home page"), the link does not lead to the intended resource and is classified as broken.

In May 2016, we re-examined each URL to determine whether it was working. As before, links were classified as working if and only if they led to the intended resource.

Characterizing the Population of URLs in the APSR

Counting URLs no more than once per article, we recorded 1,135 URLs in the 56 issues of the APSR that were published from 2000 through 2013. Of these URLs, 1,055 were unique. Figure A1 classifies the URLs by their referent type (bibliographic, database, reproducibility) and by the type of site (institutional or personal) to which they linked. Thirty-six percent of all URLs were reproducibility URLs: links that pointed to information that readers would need to reproduce the authors' findings.

Unsurprisingly, the publication of URLs in APSR articles has increased over time, suggesting the discipline's increasing reliance on this practice to foster transparency. The left-hand panel of Figure A2 illustrates the trend: the mean number of URLs per issue has ranged from seven in 2002 to 45 in 2013, the last year that our dataset covers. Somewhat more surprisingly, the pace of the increase has not been steady. The use of URLs declined from year to year four times in the period that we study. And the average number of URLs appearing in each issue changed little from 2000 (8.0 URLs per issue) through 2008 (12.0). But since 2008, the use of URLs has increased dramatically.

Importantly, these trends have nothing to do with over-time change in the number of articles that are published in each APSR issue. We can control for this

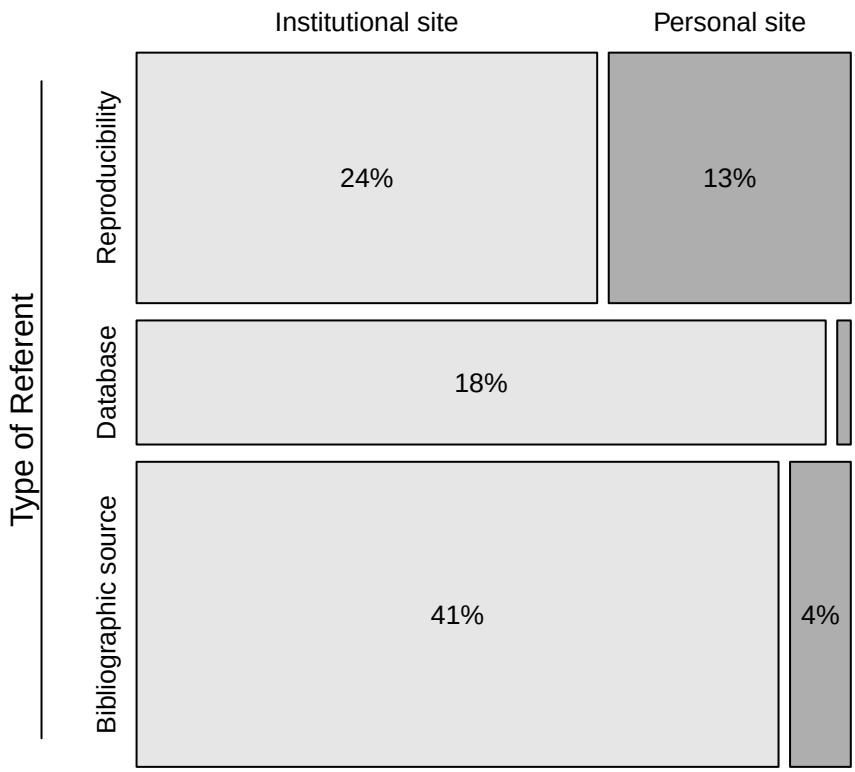

Figure A1: URLs by type of site and type of referent. We classified 1,135 URLs that appeared in the *American Political Science Review* between 2000 and 2013.

possibility by examining the average number of URLs published in each article (rather than each issue). The right-hand panel of Figure A2 reports this alternative measure. It is quite similar, and the two measures of URL use are correlated at $r = .98$.

Comparison of Broken-Link Rates in Personal and Institutional Sites

One may imagine that the high rates of reference rot in the APSR are due to authors linking to resources that are stored on personal sites rather than institutional sites. By this reasoning, personal sites—typically sites that authors maintain to make their research available to the public—are more likely than institutional sites to change in

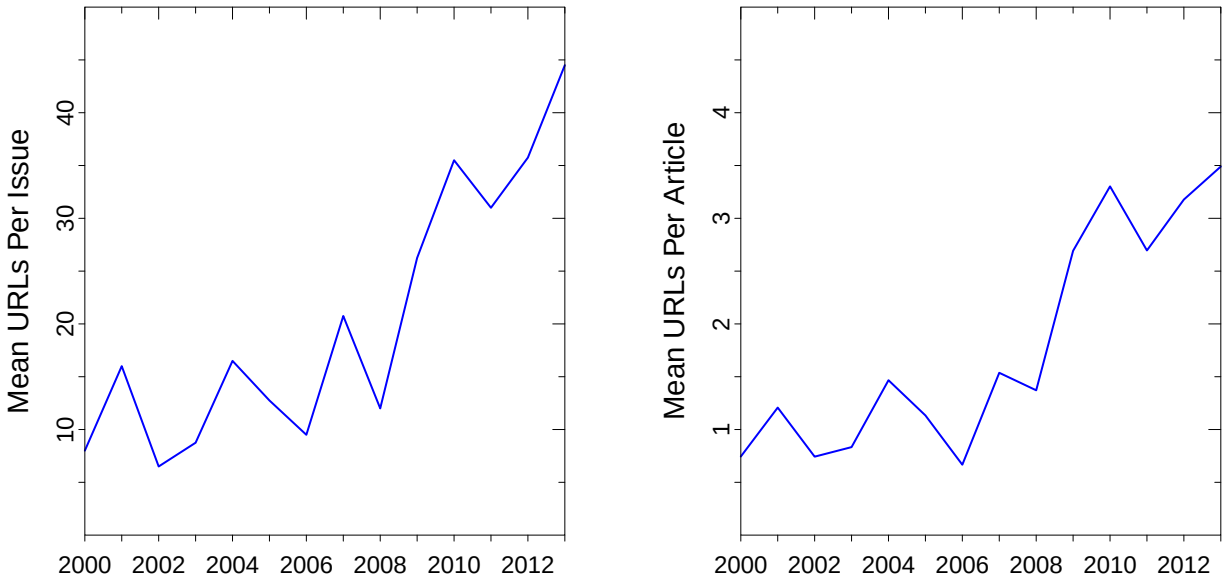

Figure A2: Increasing use of URLs in the American Political Science Review, 2000-2013. Our canvas of every APSR issue from 2000 through 2013 produced a dataset of 1,135 URLs, ranging from 26 in 2002 to 178 in 2013. The left-hand panel illustrates the mean number of URLs per issue in each year from 2000 through 2013. The right-hand panel illustrates the mean number of URLs per article (rather than per issue) over the same period. The two variables are correlated at $r = .98$.

ways that break URLs. This is a plausible explanation, but it is not correct. To see why, begin by noting that, as Figure A1 shows, few URLs published in the APSR refer to personal websites. In all, only 192 of our 1,135 URLs (17%) refer to personal sites. Even if personal-site URLs were more fragile than institutional-site URLs, their small numbers would make them unlikely to have a large impact on overall rates of reference rot.

Still, one might imagine that links to personal sites are more likely to be broken than links to institutional sites. To our surprise, the opposite is true. The left-hand panel of Figure A3 shows that in every year from 2000 through 2010, personal-site URLs were less likely to be broken than institutional-site URLs.

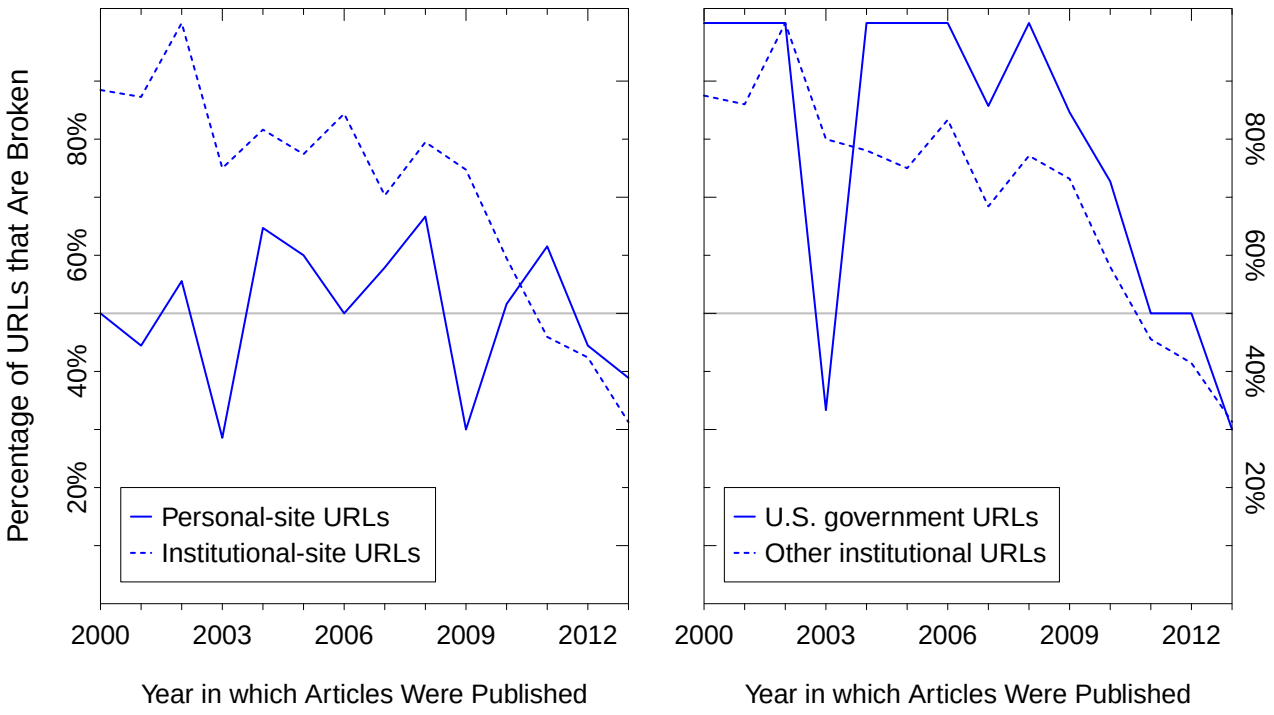

Figure A3: Broken links in the American Political Science Review, 2000-2013: personal vs. institutional sites and U.S. government sites vs. other institutional sites. The left-hand panel illustrates broken-link rates among links to personal sites (N = 192) and to institutional sites (N = 943). The right-hand panel focuses exclusively on institutional sites, distinguishing between links to U.S. government sites (N = 97) and all other sites (N = 846).

Figure A3 does not lead us to infer that linking to personal websites is a good practice. Note that even personal-site URLs are broken at high rates: in our dataset, 52% of all such links (99 of 192) are broken. This rate is better than the rate for institutional sites, for which 60% of all links (567 of 943) are broken, but it is not good. The proper inference from Figure A3 is not that personal sites are reliable, but that the hosting of resources on institutional sites is no guarantee that URLs will continue to work.

Comparison of Broken-Link Rates in Different Kinds of Institutional Sites

Of course, not all institutional sites are alike. Our definition of an institutional site encompasses sites run by both massive international organizations (e.g., the World Bank) and far smaller organizations that consist of only a handful of people (e.g., Project Vote Smart). Perhaps large and robust institutions are better able to maintain stable, functioning websites than institutions that are run on a shoestring. With this distinction in mind, we report one further analysis: we compare results from URLs associated with the U.S. federal government (e.g., URLs pointing to census.gov or fbi.gov) to institutional URLs that are not associated with the U.S. federal government.

The results are in the right-hand panel of Figure A3. We must be circumspect about drawing inferences from these data, inasmuch as only a small number of URLs in our database—97—refer to U.S. government sites.[4] Still, these data suggest that there is no large difference in reliability between links to sites hosted by the U.S. federal government and links to other institutional sites. Across all 14 years, 72% of links (70 of 97) to U.S. government sites are broken, as opposed to 59% of links (497 of 846) to other institutional sites. This evidence is only suggestive, but what it suggests is that a connection to an institution with massive resources does not make URLs more reliable.

URL Shortening

Some URLs that incorporate persistent digital identifiers may seem too long to be displayed in print or online. For example, one may be unwilling to insert a URL as long as http://doi.org/10.3886/ICPSR06987.v1 in the middle of one's article. Various URL-shortening services can solve this problem by condensing the

[4] The number of links to U.S. government sites in any particular year is smaller still. For example, the right-hand panel shows that there are seven different APSR volumes in which 100% of the links to U.S. government sites are broken. There were no more than eight links to U.S. government sites in any one of those years.

URL into a much smaller one, but most of these services cannot be recommended, because the corporations that run them make no guarantee that they will exist in the future. However, the International DOI Foundation has created a shortening service specifically for DOIs: http://shortdoi.org. It can be used to create short URLs that are suitable for display, and it guarantees the persistence of these short URLs, just as it guarantees the persistence of DOIs themselves. These short URLs are equivalent to the conventional forms: for example, http://doi.org/10.3886/ICPSR06987.v1 is equivalent to http://doi.org/ckgrjt.

Implications of Copyright for Digital Archiving

One may wonder about the implications of copyright for the archiving of digital resources. These implications vary from country to country and from case to case. So far as the United States is concerned, the best discussion of these issues that we have found is Besek (2003), a report commissioned by the Library of Congress. Besek suggests that, in the United States, the archiving of published, fact-based work for academic purposes is often protected by "fair use" exceptions to copyright (Besek 2003, 5, 17).

References

Besek, June M. 2003. "Copyright Issues Relevant to the Creation of a Digital Archive: A Preliminary Assessment." http://www.clir.org/pubs/reports/pub112/pub112.pdf (accessed August 23, 2015). Archived at http://perma.cc/YN5G-KE33.