Data We Trust—But What Data?

Jennifer Golbeck

Dr. Jennifer Golbeck *(jgolbeck@umd.edu) is Associate Professor, College of Information Studies, Affiliate Associate Professor, Computer Science, Affiliate Associate Professor, Journalism, ADVANCE Professor, and Director of the Social Intelligence Lab at the University of Maryland, College Park.*

Reference & User Services Quarterly, vol. 57, no. 3, pp. 196–99

The Obama administration's time saw massive amounts of government data shifting online. It can be hard to remember the landscape back in 2008, when very few people had smartphones, and Facebook had fewer than 150 million users—less than 10 percent of its current size.[1] We were just starting to grapple with all the data that was becoming available. The administration embraced the trend. They launched data.gov, a project designed to serve as a repository of important data sets from the federal government. Agencies followed suit, uploading their data or creating their own repositories. Databases, websites, and all sorts of content became accessible online. It appeared we were entering a golden age of open data, where citizens would have access to the raw data that their tax dollars funded, that fueled policy decisions, and that affected their lives. The movement of government data to the web improved transparency and fueled research to complement official sources.

With the shift in administrations from Obama to Trump, the climate of open government data has shifted as well. There were serious fears that the Trump administration would remove vast amounts of data from government websites. Academic groups, libraries, and nonprofits began archiving open data sets and government web pages. Up to now, however, there has not been a massive removal of government data. Most of the data.gov data sets are still present, and there has been no order to delete these records en masse. But does that mean that the current administration is committed to open data like the Obama administration was? No. We have seen information, data, and websites from government agencies hidden, pushed aside, or suppressed when it does not align with administration policies.

On top of that, the administration has allowed new, less trustworthy information to invade publicly accessible sources. Bots—automated programs that post content and interact with existing content—have corrupted public processes on social media such as Twitter and in government systems. While the bot comments are not official government information, they provide complementary optics to the suppression of information; they may illegitimately make it look like there is public support for or interaction around an issue.

How does an information seeker determine what information from government websites is trustworthy and what is not? At this point it is often

a matter of thinking about what should and should not be there rather than the data itself.

UNDERSTANDING WHAT'S MISSING

Suppression has been the tactic of choice for the current administration when government websites have politically inconvenient information. The first of these steps was the very visible Inauguration Day ban on the National Park Service using social media,[2] after they shared photos comparing the crowds on the National Mall during the Trump and Obama inaugurations. The Badlands National Park account responded by "going rogue" and tweeting facts about climate change in the subsequent days.[3] The mere fact that there was controversy around a national park account sharing scientific facts about the environment signaled how dramatically the landscape had changed for government information sharing.

That social media ban was followed by an order to the Environmental Protection Agency (EPA) and the Departments of Transportation, Agriculture, and the Interior that banned any communication with the media.[4] Within the EPA, the term climate change has been systematically removed from many pages.[5] A subsite that was called "Climate and Energy Resources for State, Local, and Tribal Governments" was removed and eventually reappeared without the "climate" part—shortened to just "Energy Resources for State, Local, and Tribal Governments."[6] The entire climate change section of the site at epa.gov/climatechange was taken down; for months, it has simply said it was being "updated." Though an archive of the old site is available, the EPA is clearly done updating their climate information for the foreseeable future. Other government websites have seen sections that are out of step with the Trump administration's priorities hidden or totally removed.[7]

The absence of information on a government website sends a message. If there is almost no mention of climate change on the EPA website, does that mean it is no longer an issue of serious concern? Of course not. However, for citizens looking for information about the topic, the lack of mention may communicate that climate change is not important in the United States. That is a failure of government websites to provide trustworthy information about the state of the world.

On social media, concerns have also arisen regarding access to Donald Trump's Twitter account. Members of the administration have claimed that tweets on Trump's account @realDonaldTrump are official policy statements.[8] If that's true, it is official information, and it would be considered a government publication that any and every citizen should have access to; however, Trump has taken to blocking people who criticize him. This prevents the blocked accounts from viewing Trump's posts. Some of those blocked users are now part of a lawsuit against Trump.[9]

All of these actions raise questions about the trustworthiness of government data. While there has not been large-scale manipulation of the content of data sets, there has been significant suppression of government information. Agencies have been prevented from sharing content that would have been part of their normal business under most administrations. Websites and information have been hidden. Individuals have been blocked from accessing some data. Does this mean the data you can access cannot be trusted?

It seems so far that government data sets are still accurate. In that sense, if you are looking for census numbers and you download them from census.gov, you can trust that those numbers are accurate. We will look at more ways of assessing those data sets later in this article.

It is worth noting that this is certainly not the first time political interference has had this effect. Consider gun control. Gun violence is a serious public health concern. The Centers for Disease Control and Prevention (CDC) is the government agency tasked with protecting public health. They study not just diseases but also causes of injury such as automobile accidents. Gun violence is a leading cause of injury and death in the United States, but it is barely studied by the CDC. This is not because they fail to understand the magnitude of the problem; it is entirely political. Since 1996, the Dickey Amendment to the government spending bill prohibited the CDC from using any funds to "advocate or promote gun control." This essentially prevented any research because if conclusions from a study found that gun control would improve public health outcomes, the CDC could be seen as advocating for or promoting it. Thus, there is very little government-funded gun-control research; politics has suppressed its visibility despite the fact that it is a major public health issue. If someone wants public health data about gun control, they need to look elsewhere.

WHEN THERE'S LESS THAN MEETS THE EYE

The flipside of the problem of suppressed information is that illegitimate information is making its way into government information sources, sometimes even official records.

While no one would seriously consider social media comments as a reliable source of information about a topic, social media is a cornerstone of the current administration's public communication strategy. This opens official statements to commentary, likes, and shares from anyone operating on those platforms.

Consider this tweet from Donald Trump (see figure 1). It has close to 160,000 likes. Does a tweet with 160,000 likes indicate there is broad public support for the idea Trump shared in the tweet? What if only two people had liked that tweet? Even if those likes aren't official government information, the volume of likes sends a message.

Now what if I told you that 159,998 of the likes were fake, generated automatically by Russian computer programs with fraudulent Twitter accounts, and only two likes came from real human Twitter users? That sends a message, too. Unfortunately, we don't really know how many likes come from bots, but research has shown that pro-Trump bots

Figure 1. Sample Tweet

overwhelmed Twitter with posts and likes to the point where it may have affected the outcome of the election. Researchers have identified many of Trump's followers and accounts that like his tweets as bots. When there is uncertainty about the validity of public interaction with government, it is important that the volume of interactions not be given weight.

Unfortunately, these problems have seeped from social media into official records. The debate over Internet neutrality rules—regulations that require Internet service providers (ISPs) to treat all data online the same, without blocking, slowing, or speeding up certain content—has been ongoing for years. Internet service providers argue they can be more innovative without regulation. The vast majority of Americans want net neutrality and do not want their ISPs manipulating their online experiences. From April 27 to August 30, 2017, the Federal Communications Commission (FCC) collected public comments on their plans to repeal net neutrality regulations and give ISPs control over the way Internet traffic is treated. Millions of comments were submitted. An analysis found over a million of these comments were generated by bots that used artificial intelligence to create comments that were posted under the names of Americans who knew nothing about it and never intended to submit comments. Many were posted from Russian accounts. These comments have not been removed from the record; the FCC has kept them as part of the legitimate set of public comments. The fraud has been so bad, including the specter of foreign influence over an American regulatory process, that New York Attorney General Eric Schneiderman has been investigating the comments.[10] Despite his requests, the FCC has refused to cooperate with the investigation. The fraudulent comments align with the administration's political goals, which decreases any incentive to correct the record.

There are many ways to interpret the FCC's actions, but one message is clear from the perspective of trustworthiness: the presence of public comments on government proposals cannot be trusted as representative of the public's feelings. Certainly, some people will always try to manipulate things to their advantage, but when an agency refuses to support an investigation into improper actions within their own system, you know there is not a vested interest in accurately reflecting public opinion through the process.

NOW WHAT?

So, in this situation, how does one find and analyze trustworthy information sources? Here are some guidelines that may be useful:

Before using a data source, check on its status. Many watchdog groups are monitoring documents, websites, and data sets for changes. You can check with groups like the Sunlight Foundation to see if your data set has been flagged. It may be that information has been changed or removed. When looking for groups to verify your data, look for nonpartisan organizations, academic groups from well-known universities (be wary of private schools that have an ideology to push), or professional societies that represent large groups of working professionals in a field.

The absence of government data means nothing. If a government website does not discuss an issue or provide data on a topic, that does not mean the government or society at large is unconcerned with that issue. Data that you know once existed of may disappear, whether it is a tweet or an entire topic like climate change. The watchdog groups mentioned above may also track disappearing data. Many are archiving data sets outside the United States, so you can download copies of the originals.

Be wary of considering interactions with the public. Whether it is interaction statistics or actual comments, we are in a period where parties outside the United States are using automated techniques to completely corrupt any public interactions surrounding governmental or political discourse. It can be tempting to consider the volume of interactions as meaningful, but these can be easily falsified on a massive scale.

Look for other sources. If you are looking for certain types data, governmental sources outside the United States may be a good resource. Canada has an outstanding open government data program, and there are many excellent resources in the EU as well. For things like scientific data, these may be more reliable and complete sources. For US-centric data, it is again worth looking at professional societies and nonpartisan nonprofits. These groups will typically be focused on the particular issue you care about (e.g., gun control, immigration, etc.), but for non-biased data, be sure they do not have an advocacy agenda.

In the last year, we have not seen a massive removal of government data. We have seen targeted suppression and a general lack of concern for having government data sources reflect objective truth. Fortunately, many organizations are monitoring, archiving, and analyzing changes to official data. They can help users assess the data they see, recognize the content that is missing, and access data that has been lost. In a shifting environment of data reliability, such resources are likely to grow in value and importance.

Reference

1. Josh Constine, "Facebook Now Has 2 Billion Monthly Users . . . and Responsibility," Tech Crunch, June 27, 2017, https://techcrunch.com/2017/06/27/facebook-2-billion-users/.
2. Dan Merica and Dana Bash, "Trump Admin Tells National Park Service to Halt Tweets," *CNN Politics*, January 23, 2017, http://www.cnn.com/2017/01/21/politics/trump-national-park-service-tweets/index.html.
3. Katie Reilly, "A Rogue National Park Is Tweeting Out Climate Change Facts in Defiance of Donald Trump," *TIME*, January 24, 2017, http://time.com/4645927/badlands-national-park-climate-change-tweets/.
4. Michael Biesecker and John Flesher, "President Trump Institutes Media Blackout at EPA," *Boston Globe*, January 24, 2017, https://www.bostonglobe.com/news/politics/2017/01/24/trump-bans-epa-employees-from-updating-public-via-press-social-media/Anr90pkwhavC2kzK8pwsyK/story.html.
5. Madison Park, "EPA Removes Climate Change References from Website, Report Says," CNN, December 8, 2017, http://www.cnn.com/2017/12/08/politics/epa-climate-change-references/index.html.
6. "Energy Resources for State, Local, and Tribal Governments," United States Environmental Protection Agency, accessed December 12, 2017, https://www.epa.gov/statelocalenergy.
7. "Tracking U.S. Government Data Removed from the Internet during the Trump Administration," Sunlight Foundation, accessed December 12, 2017, https://sunlightfoundation.com/tracking-u-s-government-data-removed-from-the-internet-during-the-trump-administration/.
8. Elizabeth Landers, "White House: Trump's Tweets Are 'Official Statements,'" *CNN*, June 6, 2017, http://www.cnn.com/2017/06/06/politics/trump-tweets-official-statements/index.html.
9. Charlie Savage, "Twitter Users Blocked by Trump File Lawsuit," *New York Times*, July 11, 2017, https://www.nytimes.com/2017/07/11/us/politics/trump-twitter-users-lawsuit.html?_r=0.
10. "A.G. Schneiderman Releases New Details On Investigation Into Fake Net Neutrality Comments," press release, New York State Office of the Attorney General, December 13, 2017, https://ag.ny.gov/press-release/ag-schneiderman-releases-new-details-investigation-fake-net-neutrality-comments.