'What Data?' Records and Data Policy Coordination During Presidential Transitions

Kristin B. Cornelius and Irene V. Pasquetto[(✉)]

University of California, Los Angeles, Los Angeles, USA
`irenepasquetto@ucla.edu`

Abstract. The presidential transition in the United States takes place over the course of several years and involves the efforts of many different agencies and organizations. While it is standard practice for an incoming administration to change the content on government agencies' websites, the Trump administration pushed this practice beyond convention, even to alter the official narrative on climate change. Almost immediately after the inauguration, the official White House website deleted nearly all references to the phrase 'climate change,' and all online mentions of climate change on federal and government websites had been excised in the following months. Even if government data cannot be deleted completely, the manner in which they are preserved and made accessible, or hidden and obscured, is vitally important to the researchers and public that rely on this information. This project argues for the coordination of controls on this information: the policies, standards, and directives that regulate *both* the content accessed (e.g. the datasets) and the access points themselves, including the government agencies' websites that act as information sources and portals to the databases and repositories of publically funded research.

Keywords: Information policy · Open data · Record keeping
Knowledge preservation

1 Introduction

On Earth Day 2017, Saturday, April 22nd, 2017, more than 400 coordinated protests and marches took place across the United States under the banner "March for Science." These protests were a reaction, in part, to the recently elected Trump administration's agenda to support certain policies that seem to defy the findings of scientific data in relation to climate change research. Notably, President Trump publicly denied climate change science on several occasions, and, only few days after the new administration's inauguration, the head of the Environmental Protection Agency (EPA) was replaced with Scott Pruitt who previously stated that the debate on climate change is "far from settled" [14]. As a reflection of the Trump administration's attitude toward environmental research and transparency more generally, early in the transition process the official White House website deleted nearly all references to the phrase 'climate change,' and, in the following months, all online mentions of climate change on federal and government websites had been excised, buried or stripped of any importance [3, 13]. The EPA website was among the most targeted websites, with the entire climate

change section removed, only to be replaced by a static holding page (the section is still missing as of September 11th, 2017) [18]. These actions alarmed US environmental groups, archivists communities, and scientists, who worried not only for the change in the narrative by the science community, which has reached 97% consensus on agreement of manmade climate change [22], but also for the science data that is made available on those same websites. The activists' fear was confirmed by the fact that within the first few weeks of the inauguration, the Trump Administration removed dozens of data sets from openwhitehouse.gov [11].

The change from one presidential administration to another in the United States takes place over the course of several years and involves the efforts of many different agencies and organizations. It begins at least six months prior to the election, in accordance with the recently passed Presidential Transitions Improvements Act of 2015 that requires presidential candidates set up a team with agency leaders to smooth transition efforts [21]. While it is standard practice, to some extent, for an incoming administration to change the government agencies' output (e.g. content of websites) during a transition to reflect their own worldview and politics, generally, in addition to political ideology, transparency is the guiding principle, if not an achievable ideal, and public accessibility the aim. Alex Howard of the Sunlight Foundation, a think tank that influenced Obama's National Archives and Records Administration's (NARA) 2012 directive, referred to as the Managing Government Records Directive (M-12-18) [26], which supports public knowledge and transparency of government as its goal, for instance. Howard explains Obama's position: "They approached it from the default position that government data belongs to the public, and it should be (1) easily accessible and (2) machine-readable. They also tried to (3) put the data in formats that would easily allow software developers and researchers to use and analyze" [11]. In this observation, one can detect some trust placed in the ability of technology to make public information more accessible; however, even with the best intentions and clearest language, this goal requires much thought and coordination among several already existing processes and policies, as well as an evolving understanding of digital records. In explaining the need for a more efficient process to achieve a higher standard of accountability for users of data from publically funded research, former Chief Data Scientist for the White House, Dr. D.J. Patil notes how infrastructure changes can have an effect on the retrieval of datasets that are important to the public and how asking for datasets by FOIA is "an incredibly inefficient use of taxpayer dollars" [15]. Patil describes how the Obama administration dedicated "a lot of people" and "a tremendous amount of time […] to streamline the process of requesting government data" [15]. And, how the motivation for this was "allowing what all administrations have typically provided, and that is transparency into who is using the White House and other datasets that people have a right to see" [17]. In other words, even prior to Trump taking office and promoting a radical overhaul of the information on and accessed from these sites, the processes by which these data were retrieved was complicated and already in need of revising. And now that it seems the current administration is not continuing to work toward transparency, but rather favoring "alternative facts" [1] and nondisclosure agreements [10]. These actions are *testing* our institutional procedures to see if the

information provided to the public on these sites can uphold a certain amount of integrity through these transitions, and therefore it is time to look critically at the processes and policies whose job it is to do this vital work.

Even with the aim of being the primary archival institution for government records and public knowledge, NARA's strategies take coordinated effort and funding, which are two aspects that are subject to an administration's whims. For instance, reporters found that positions created under the Obama Administration to curate and protect data such as "chief digital officer" were eliminated [17]. Part of the response to this chain of events was the formation of the "Data Refuge" initiative, which saw the participation of data archivists, scientists, and volunteers from multiple North American universities, including, to mention just a few, UPenn, University of Toronto, UCLA, UCSD, MIT, and Harvard. Between December 2016 and April 2017, data archivists and information professionals selected, downloaded, and stored climate change-related datasets hosted on federal agencies' websites that were collected with public funding and in danger of disappearing or being deleted under the current administration's transition [7]. Thanks to the work of these activists and volunteers, these highly valuable "rescued datasets" are now safely stored, properly curated, and available to the scientific community, as well as to the general public [25]. The main outcome of these interventions is probably the foundation of the Environmental Data and Governance Initiative (EDGI), a non-profit international organization that has as a mission to preserve publicly accessible and potentially vulnerable scientific data and archive web pages from multiple agencies' websites, such as the Environmental Protection Agency (EPA), the Department Of Energy (DOE), the National Oceanic and Atmospheric Administration (NOAA), and the National Aeronautics and Space Administration (NASA).

However, much of this data rescue effort was more complicated than necessary. This was partly due to insufficient oversight over an administration that works tirelessly to subvert policies designed to ensure the integrity of information disseminated from government agencies. In particular, while much attention was paid to the data policies (e.g. open access, preservation, curation best practices), the policies that regulate the spaces where information is accessed, such as federal agencies' websites, failed to protect the preservation, access, and consistency of this data. This project argues that the policies (e.g. M-12-18 directive) that govern the platforms where government data is accessed should be coordinated with the open data policies (e.g. NSF/NIH data sharing policies for gov. 2013/2012, FAIR principles, codata, RDA) in order for there to be sufficient protection of and access to this information during presidential transitions.

2 Background

The background of this project looks at both policies and principles that govern the way data is presented to the public on government agency sites, as well as the directives and publishing regulations that control the way information is preserved in the case of change due to presidential transitions.

2.1   Data Policies

More than 30 DataRescue events, also called "archive-a-thons," occurred at cities across the U.S. and Canada during the first quarter of the Trump organization [25]. These events were mainly organized and coordinated by activists with the help of the EDGI, the Internet Archive, and the DataRefuge initiatives. Between Fall 2016 and Spring 2017, the activists archived over 200 TB of government websites and data. This includes over 100 TB of public websites and over 100 TB of public data from Federal FTP file servers totaling, together, over 350 million URLs/files. This includes over 70 million html pages, over 40 million PDFs and, towards the other end of the spectrum and, for semantic web aficionados, 8 files of the text/turtle mime type. The EDGI volunteer tech team of over 30 contributors has built open source and freely available tools and projects for grassroots archiving, and made all available to the public on their GitHub account. EDGI's efforts to archive, preserve, curate and make publicly available scientific data can be seen in the light of the expanding and increasingly influential "open data movement" in science [8]. Making scientific data open is a science policy priority in the US, Europe, and elsewhere. Typically, with the expression "open data" commentators refer to publicly funded research data that have been made openly available in digital repositories, archives, or databases. Openness is generally defined as "access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination" [19]. Multiple, but related, rationales for making science data open exist [2]. For instance, for the scientists, accessing and reusing each other's data can lead to faster discoveries and knowledge integration. Often, open data initiatives symbolize a reaction against the view of scientific knowledge production as an esoteric, technical, and overspecialized process – instead promoting the idea that scientific knowledge can and should be investigated as a whole. Finally, for policy makers, reusing data is a matter of return on investment, promoting economic innovation and enabling knowledge transfer to the industry. American funding agencies such as the National Science Foundation (NSF) and the National Institute of Health (NIH) commonly require scientists to deposit their data in open repositories as a condition of receiving funding [20]. A number of tools, standards, and conceptual models have been designed to enable scientists to work in "open science frameworks." Institutions and organizations worldwide are investing in infrastructures and policies to promote the centralization, access, and integration of scientific data. A promising new development to address the vagaries of open data is the FAIR standards – Findable, Accessible, Interoperable, and Reusable data. These standards apply to the repositories in which data are deposited. The FAIR standards were enacted by a set of stakeholders to enable open science, and they incorporate all parts of the "research object," from code, to data, to tools for interpretation [24].

2.2   Website Policies

Certain protections in the form of regulations are in place to preserve government information and records and to maintain the country's cultural record. The face of the agency's website provides information about both the agency's inner-activities, in

addition to providing information about their work in the form of synthesized research data and official statements. EPA's website, for instance, is a place to find research data, but the website itself publishes information about both the environment and the work of the EPA. There are different standards and regulations that govern each of these types of information, in the form of record keeping practices and preservation strategies (like those outlined in [26]) and publishing regulations.

An awareness of the importance of record keeping in transparency efforts proved characteristic of the Obama administration's records policies; for instance, in 2012, a centralized records management program (M-12-18) [26] was put into place by a memorandum sent out from President Obama, Jeffrey D. Zients (Acting Director, Office of Management and Budget), and David S. Ferriero (Archivist, NARA) that aimed to ultimately increase transparency and accountability for the government by preserving records and making them accessible for the public. This was the stated goal of the memo, titled "Managing Government Records Directive," released on August 24, 2012:

> Records are the foundation of open government, supporting the principles of transparency, participation, and collaboration. Well-managed records can be used to assess the impact of programs, to improve business processes, and to share knowledge across the Government. Records protect the rights and interests of people, and hold officials accountable for their actions. Permanent records document our nation's history *(Managing Government Records Directive)* [26].

Here, Obama's strategy seemingly embraces the values of openness promoted by archivists and the open data community, even so far as to include a nod toward "collaboration," which can be used to subvert top-down policies. Certain mechanisms and practices began to be implemented as a result of this memo, including its request for each agency to assign a Senior Agency Official (SAO) to oversee their records management processes [26]. Agency SAOs continue to hold responsibility for records management practices such as M-12-18 makes it their charge to be "responsible for protecting the integrity of agency programs and trustworthiness of agency information" as they have "statutory responsibility" for the agency's records management program [5]. Further, the NARA "Guidance on Managing Web Records," released in January 2005, describes how the treatment of information on Federal agency websites is dictated by the head of the agency and that agencies cannot delete web records "related to the operation of Federal websites" without permission from NARA [15]. This leaves their entire official 'record' status up to NARA, placing a significant amount of trust in a system that is not yet entirely functional. Alex Howard, for instance, says he has only "low to moderate confidence" in the "completeness of the NARA archive," and, to compound these issues, the links pointing to the tools on NARA's White House portal were "simply broken" and the NARA couldn't guarantee API access would work for all the datasets [17].

As described in previous sections, certain mechanisms prevent the Trump administration from directly "go[ing] out and delet[ing] decades worth of information" entirely [17]. Howard notes how there is a difference between "retaining scientific data within the agencies and keeping things on the website" [17]. Spokeswoman Miriam Kleinman, for instance, said her agency primarily focuses on agencies' ability to preserve records, not whether they are "making them available to the public" [6]. Her

full statement on this issue: "NARA's records management guidance mainly focuses on records creation, retention, and eventual deletion or transfer to NARA for permanent preservation. NARA has not issued specific guidance about large data sets being taken down from publicly-available websites" [6]. Therefore, while these policies are intended to ensure accountability, there are paths to circumvent them, some of which were utilized during the recent transition. For instance, the determination of which web records need to be retained as part of 'Federal agency operations' and for how long is ultimately left up to each agency's SAO's (who are put in place by the agency head appointed by the president) judgment—essentially, most agency records, including web records, need to be kept for a certain period to mitigate risk (in the business sense), and only records who hold "long-term historical value" should be transferred to NARA (at NARA's identification and discretion) [16]. Thus, in most cases, anything published on EPA's website besides datasets and records related to operations can be deleted once "old or when superseded, obsolete," and, minor upkeep and changes can be made without notification [16]. Even though several details are given as to how web records should be archived, including a suggestion to include contextual information, screenshots, code, and even website maps and all other documentation, the live site can change at the whim of the SAO under the guise of "standard practice" and normal upkeep; this was one of the explanations of the changes made by the Trump administration [23]. Moreover, at any time the current president could release a new directive to NARA that would overturn M-12-18 and its initiatives, and thus all of these regulations are only in place as long as this directive holds.

3   Discussion and Conclusion

The Trump administration's removal of information from its agencies' websites the weeks following its inauguration, including federal climate plans created under former President Obama, tribal assistance programs, and references to international cooperation on climate change efforts, reveals a shift in epistemological position. Transparency is not only *not* a priority for this administration, it might also be argued that secrecy is encouraged, as evidenced by the non-disclosure agreements asked to sign of all agency employees [10].

NARA's lack of distinction and the Trump administration's sense of 'comfort' in deleting this information could be from dealing with digitized documents (i.e. webpages), that seemingly possess immaterial qualities, rather than physical or paper documents, which could be seen as more difficult to alter or destroy. In other words, the ability to replace information on the websites because the very institutions that support their authority (i.e. NARA) seem to also support their immaterial, digital status, eases the path towards alteration and being ignored. Scholars have commented on the change from paper to electronic, including Drucker's [4] caution against the text's "(mis) perceived condition as immaterial in the electronic environment," or its ability to seem an "idea that appears to consciousness as a form but without materiality." In this view, and in line with NARA's lack of attention toward the display of information on website pages, it seems as though the pages of a website almost lose their status as document completely; this supports Levy's argument that digitized documents can "fail to register

in a social space, to fail to have social identity, and thereby fail *to be* a document" [9]. This quality of 'immateriality' that exists in a digital space and attaches itself to electronic documents has been the concern of several activists, including Jefferson Bailey of the *Internet Archive,* who stated that with the recent transition, he was worried about "politically-driven ephemerality," or, in plain speak, "pages being shut down" [6]. He elaborates: "There's a lot more dynamic content on the web than there was four or eight years ago. Some of that is challenging to capture […] There are sometimes FTP servers or other directories that a crawler might not discover because they're hidden […] Subdirectories are very hard to find" [6]. Additionally, the authority of an institution, such as a federal agency, can imbue subversive acts like the removal of data with the qualities of benign, bureaucratic 'housekeeping'—such as combining the removal of climate science data and associated programs with the standard, basic functions of just 'updating' the websites. The authority of the data on these websites relies heavily, then, on their ability to perform as a record of evidence—not simply as a digital, immaterial document, or standard practice.

While current open data policies and principles that guided the Data Rescue efforts largely focus on identifying and describing what is necessary for ensuring the long-term preservation of the data themselves, little attention has been paid to the spaces in which the data are hosted and made available. Open access principles and policies apply to the datasets, to the databases that host and organize the data for retrieval, and, increasingly, to the code or software used to collect and analyze the raw data. The data archiving efforts from EDGI, for instance, brought attention to a new aspect of the open access challenge—one that, maybe, we never considered before: we can curate and make publicly available science data, databases, and code, but all these efforts are in vain if the federal websites that host, or point to, these resources can be taken down by new administrations at any time. Of course, by taking down EPA webpages, research data on climate change do not disappear. The science community is well equipped to ensure this would be a very unlikely outcome. In addition to being uploaded to agencies' repositories and NARA, research datasets exist in multiple copies, and whenever made available, these copies are safely stored in publicly funded data centers around the country. A recent white paper written by National Center for Atmospheric Research (NCAR) researchers rightfully pointed out that national research data centers have ad-hoc plans for data storage, migration, and rescue in place [12]. In the paper, the researchers also explained that research data are in need to be rescued all year around, not only during transition period, especially when data become obsolete (compared to the technologies available to analyze or manage them), or when there are types of data (e.g. small-scale datasets) that are not required to be shared by the scientists, but still of extreme importance to the science community at large, as well as to industries that rely on it for critical infrastructure projects. As the very history of our country relies on the effective implementation records keeping practices, and considering Orwell's ever-pervasive perception that "who controls the present controls the past," it is vital to consider how different administrations' policies and practices shape and control records as to maintain their version of truth, such as scientific research and data policies that affect how the story of climate change is told to the public and supported (or otherwise) by scientific research.

References

1. Blake, A.: Kellyanne Conway says Donald Trump's team has "alternative facts". Which pretty much says it all. Washington Post (2017)
2. Borgman, C.L.: Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press, Cambridge (2015)
3. Davenport, C.: With Trump in charge, climate change references purged from website. The New York Times (2017)
4. Drucker, J.: Performative materiality and theoretical approaches to interface. Digit. Humanit. Q. **7**, 1 (2013)
5. Duranti, L.: Luciana Duranti, Trusting Records and Archives in the Era of Alternative Facts and Misinformation, ILS Colloquium Speaker. IU Calendar (2017)
6. Gerstein, J.: Fears rise of Trump-era "memory hole" in federal data. POLITICO (2016)
7. Harmon, A., Fountain, H.: In age of Trump, scientists show signs of a political pulse. The New York Times (2017)
8. Leonelli, S.: Why the current insistence on open access to scientific data? Big data, knowledge production, and the political economy of contemporary biology. Bull. Sci. Technol. Soci. **33**, 1–2, 6–11 (2013). https://doi.org/10.1177/0270467613496768
9. Levy, D.M.: Scrolling Forward, Second Edition: Making Sense of Documents in the Digital Age. Skyhorse Publishing, Inc., New York (2016)
10. Marcotte, A.: Donald Trump is already trying to muzzle leakers—but government employees have broad free speech rights (2017). Salon.com
11. Maurielllo, T.: Data disappeared from Obama administration site promoting transparency. Pittsburgh Post-Gazette
12. Mayernik, M., et al.: Stronger together: the case for cross-sector collaboration in identifying and preserving at-risk data—ESIP (2017)
13. Milman, O., Morris, S.: Trump is deleting climate change, one site at a time. The Guardian (2017)
14. Mooney, C., et al.: Trump names Scott Pruitt, Oklahoma attorney general suing EPA on climate change, to head the EPA. Washington Post (2016)
15. NARA Guidance on Managing Web Records (2016). https://www.archives.gov/records-mgmt/policy/managing-web-records-index.html. Accessed 18 Sept 2017
16. NARA Guidance on Scheduling Web Records (2016). https://www.archives.gov/records-mgmt/policy/managing-web-records-scheduling.html. Accessed 18 Sept 2017
17. O'Brien, T.: Trump's quiet war on data begins. Engadget
18. Page being updated: https://www.epa.gov/sites/files/production/signpost/index.html. Accessed 17 Sept 2017
19. Pilat, D., Fukasaku, Y.: OECD principles and guidelines for access to research data from public funding. Data Sci. J. **6**, OD4-OD11 (2007). https://doi.org/10.2481/dsj.6.OD4
20. Ray, J.M.: Research Data Management: Practical Strategies for Information Professionals. Purdue University Press, West Lafayette (2014)
21. S.1172 - 114th Congress (2015–2016): Edward "Ted" Kaufman and Michael Leavitt Presidential Transitions Improvements Act of 2015 (2016). https://www.congress.gov/bill/114th-congress/senate-bill/1172. Accessed 17 Sept 2017
22. Scientific consensus: Earth's climate is warming: https://climate.nasa.gov/scientific-consensus. Accessed 17 Sept 2017
23. Sohn, T.: Leaked emails show what is wrong at the EPA 2017

24. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data. **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
25. Archiving Data. EDGI
26. Memorandum for the Head of Executive Departments and Agencies and Independent Agencies (2012)