



University of Pennsylvania  
**ScholarlyCommons**

---

Scholarship at Penn Libraries

Penn Libraries


---

2019

## The Data Refuge Project for Protecting Federal Data in the United States

Margaret M. Janz  
University of Pennsylvania, [mjanz@upenn.edu](mailto:mjanz@upenn.edu)

Follow this and additional works at: [https://repository.upenn.edu/library\\_papers](https://repository.upenn.edu/library_papers)

 Part of the [Library and Information Science Commons](#), and the [Other Environmental Sciences Commons](#)

---

### Recommended Citation

Janz, M. M. (2019). The Data Refuge Project for Protecting Federal Data in the United States. *Liberte de la recherche: Conflits pratiques horizons*, 145-152. Retrieved from [https://repository.upenn.edu/library\\_papers/114](https://repository.upenn.edu/library_papers/114)

Work written in English and translated into French for publication in *Liberte de la recherche: Conflits, pratiques, horizons*, compiled by Melanie Duclos and Ander Fjeld, Editions Kime: Paris: 2019. Republished in Scholarly Commons by permission.

This paper is posted at ScholarlyCommons. [https://repository.upenn.edu/library\\_papers/114](https://repository.upenn.edu/library_papers/114)  
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).

---

## The Data Refuge Project for Protecting Federal Data in the United States

### Keywords

data refuge, data rescue, Freedom of Research Forum

### Disciplines

Library and Information Science | Other Environmental Sciences

### Comments

Work written in English and translated into French for publication in *Liberte de la recherche: Conflits, pratiques, horizons*, compiled by Melanie Duclos and Ander Fjeld, Editions Kime: Paris: 2019. Republished in Scholarly Commons by permission.

## Introduction

After the 2016 United States presidential election, there was concern about what would happen to federal climate and environmental data under an administration that was openly hostile towards science and that denies the realities of climate change. The fellows from the Penn Program for Environmental Humanities (PPEH) were one group feeling that concern. They had heard about people in their field losing access to data under the George W. Bush administration and the incoming administration seemed considerably more of a threat. The PPEH fellows brought their concern to the program director, Bethany Wiggin, who enlisted Laurie Allen and Margaret Janz from the University of Pennsylvania Libraries. The three of them with PPEH fellows coordinator Patricia Kim and support from the PPEH fellows and other librarians began thinking through different methods to create refuge for these data.

## Precedence

Some people felt the concern being expressed by the PPEH fellows and other groups was extreme and unnecessary<sup>1</sup> but there was precedence for these reactions. The Stephen Harper administration had only recently been replaced in Canada and Americans saw how that administration had closed off access to colleagues there<sup>2</sup> and eliminated information about the environment<sup>3</sup>. In our own country, the administration of George W. Bush had shown us first-hand how our research and information could be censored and made more inaccessible<sup>4</sup>. In Australia, Tony Abbott declined to appoint a minister of Science, opting to divide those responsibilities to the Education and Industry ministries<sup>5</sup>. Abbot also folded “33 climate change programs run by seven departments and eight agencies” into two departments<sup>6</sup>. In Brazil funding for science was cut so much they couldn’t pay their electric bills<sup>7</sup>, let alone those for the internet to make their discoveries known outside their institutions. The incoming Trump administration gave every indication that it would be at least as extreme as these examples<sup>8</sup>.

---

<sup>1</sup> Megan Molteni, “Old-guard archivists keep federal data safer than you think,” *Wired*, 2017, <https://www.wired.com/2017/02/army-old-guard-archivers-federal-data-safer-think/>, accessed 27 November 2018.

<sup>2</sup> Lesley Evans Ogden, “Nine Years of Censorship,” *Nature*, no 7601, vol. 533, 2016.

<sup>3</sup> “Research library’s closure shows Harper government targets science ‘at every turn,’ union says,” *CBC News*, 21 Aug 2015.

<sup>4</sup> “Voices of federal scientists,” Union of Concerned Scientists, 2009, [https://www.ucsusa.org/sites/default/files/legacy/assets/documents/scientific\\_integrity/Voices\\_of\\_Federal\\_Scientists.pdf](https://www.ucsusa.org/sites/default/files/legacy/assets/documents/scientific_integrity/Voices_of_Federal_Scientists.pdf), accessed 12 November 2018.

<sup>5</sup> Ariel Bogle, “Australia’s new PM nixes science minister post at worst possible moment,” *Slate*, 2013, <https://slate.com/technology/2013/09/tony-abbott-nixes-australia-s-science-minister-post-cuts-climate-change-research.html>, accessed 27 November 2018; Graham Readfearn, “Australia, where is your science minister?,” *The Guardian*, 17 Sept 2013.

<sup>6</sup> Bogle, “Australia’s new PM nixes science minister post at worst possible moment.”

<sup>7</sup> Herton Escobar, “In Brazil, researchers struggle to fend off deepening budget cuts,” *Science*, 20 Oct 2017, doi:10.1126/science.aar2805, accessed 27 November 2018.

<sup>8</sup> Oliver Milman, “Donald Trump presidency a ‘disaster for the planet’, warn climate scientists,” *The Guardian*, 11 Nov 2016, Michael Greshko, “The global dangers of Trump’s climate denial,” *National Geographic*, 9 Nov 2016.

## Concerns

The notion that the incoming administration could remove or redact any information that didn't support their policies from public access<sup>9</sup>, or destroy it outright, was a concern many people were having following the election. A more pressing threat through our eyes was funding. Almost all the data produced by the United States government is stewarded by government employees and held in government buildings on government systems. If an agency, such as the Environmental Protection Agency, loses its funding, the employees who maintain the data could lose their jobs, the systems that house the data might have to be used for other priorities or fall into disrepair without proper upkeep<sup>10</sup>. Budget cuts and low morale can also cause an agency to lose personnel, and with them irreplaceable institutional knowledge<sup>11</sup>.

Funding is also an issue for the continuation of the research within agencies. These agencies are collecting most of these data, and with funding cuts the research they do will become constrained. Several agencies also fund a great deal of academic research. In 2017, 53.5% of university science and engineering funding came from federal sources<sup>12</sup>. With funding for research cut within and outside of the agencies, discontinuation of longitudinal studies is a real risk. While Data Refuge would not be able to prevent gaps in research in this scenario, we wanted to prevent the loss of the results of previous research.

The data from these scientific studies are used to understand our environment. Tracking the data over time is deeply important to climate modeling. It can be used to study how an intervention can mitigate or aggravate climate change. But more than that, data are used to help our communities determine where we are and where we should go. On a day to day basis, data from the US Census, the most thorough collection of data about the population of the country, is used by communities to decide where public services, such as schools and roads, need to be built or close and to give people indications where they might open a new business. Data from the Bureau of Labor Statistics give essential information about which sectors are seeing jobs growth, which can help college and universities prepare students better<sup>13</sup>. Data about the weather allow people to make mundane decisions about what to wear or how to travel, but can go further and warn about dangerous weather such as snowstorms, excessive heat, or hurricanes. After disasters, data are needed to help communities rebuild, as in the case of Hurricane Katrina in New

---

<sup>9</sup> Gretchen T. Goldman, "Five things we've learned from surveys of government scientists," *Union of Concerned Scientists* (blog), 2018, <https://blog.ucsusa.org/gretchen-goldman/five-things-weve-learned-from-surveys-of-government-scientists>, accessed 27 November 2018.

<sup>10</sup> David Rosenthal, "Talk for 'RDF vocabulary preservation' at IPres2013," *DSHR's Blog* (blog), 2013, <https://blog.dshr.org/2013/09/talk-for-rdf-vocabulary-preservation-at.html>, accessed 27 November 2018; Andre Perry and Katherine Guyot, "Threats to government data are threats to democracy," *Excellence in Government* (blog), 2018, <http://www.govexec.com/excellence/management-matters/2018/02/threats-government-data-are-threats-democracy/145806/>, accessed 27 November 2018.

<sup>11</sup> Lisa Friedman, Marina Affo, and Derek Kravitz, "Brain drain at the EPA," *ProPublica*, 2017, <https://www.propublica.org/article/brain-drain-at-the-epa>, accessed 27 November 2018.

<sup>12</sup> "University S&E R&D funding by source, 1990-2017," American Association for the Advancement of Science, 2018, <https://www.aaas.org/programs/r-d-budget-and-policy/historical-trends-federal-rd>, accessed 27 November 2018.

<sup>13</sup> Robin Young, Wind turbine technician blows away competition as country's fastest-growing job, *Here & Now*, 31 Jan 2017.

Orleans<sup>14</sup> and gives rescue workers vital information about the location of roads and buildings where people may be stranded, as the volunteer mapping groups did for Puerto Rico after Hurricane Maria.<sup>15</sup> Federal data make an enormous impact on individuals and losing access would cost us all quite a lot.

## **Paths to Solving This Problem**

### *Data Rescue*

The initial motivation for Data Refuge was to make sure the data would remain available. We decided to have a hackathon-style event, called a data rescue, to talk about the importance of federal data and to make copies of what we could. As we thought about the problem of making copies, though, we realized that copies of the data would not be enough. Having a copy is a good thing, but if the original goes away, there is no longer a way to verify that the copy is the same. We spent a lot of time trying to find a way to instill trust in the copies of data we would be collected and decided a documented chain of custody<sup>16</sup> would be the best way to achieve this given the circumstances. We developed a workflow<sup>17</sup>, which evolved as time went on, that would ensure this documentation.

The idea of this kind of data rescue gained a great deal of attention fairly early on<sup>18</sup>, with many individuals, universities, and organizations reaching out hoping to help or host their own events. From January to June 2017, over 50 data rescue events took place around the United States<sup>19</sup>. Volunteers at events gathered around 400 datasets into our Data Refuge repository<sup>20</sup>.

### *Beyond Data Rescue*

The problem that we intended to address with Data Refuge is clearly quite a complex one with far reaching impacts. The data rescue events we had been supporting were bringing much needed attention to the issues and engaging communities in exciting new ways. However, by summer 2017, two things became clear: Enthusiasm and support for these events could not last forever and using volunteers to download data from federal websites was not the most sustainable way to ensure continued access to them. We had learned through these activities that the data collected and shared by the federal government is enormous in size and quantity, and is not well inventoried. This makes it impossible to collect all of it and equally impossible to have any sense of what amount of the whole you have collected and what to prioritize. Another great challenge is that data continues to be collected, updated, and shared. These data rescue events were taking

---

<sup>14</sup> Denice Ross, "Ten years after Katrina: New Orleans' recovery, and what data had to do with it," *Medium*, 2015, <https://medium.com/@ObamaWhiteHouse/ten-years-after-katrina-new-orleans-recovery-and-what-data-had-to-do-with-it-3df0bb2467e9>, accessed 28 November 2018.

<sup>15</sup> Corinne Segal, "Volunteers are helping Puerto Rico from home, with a map anyone can edit," *PBS News Hour*, 1 Oct 2017.

<sup>16</sup> Laurie Allen, "Data Refuge rests on a clear chain of custody," *PPEH Fellows Blog* (blog), 2017, <http://ppehlab.squarespace.com/blogposts/2017/2/1/data-refuge-rests-on-a-clear-chain-of-custody>, accessed 28 November 2018.

<sup>17</sup> "DataRescue workflow," n.d., <https://datarefuge.github.io/workflow/>, accessed 27 November 2018.

<sup>18</sup> "DataRefuge and related press," PPEH Lab, n.d., <http://ppehlab.squarespace.com/datarefugenews>, accessed 28 November 2018.

<sup>19</sup> "Data Rescue events," PPEH Lab, n.d., <http://ppehlab.squarespace.com/datarescue-events>, accessed 28 November 2018.

<sup>20</sup> "Data Refuge datasets," Data Refuge, n.d., <https://www.datarefuge.org/dataset>, accessed 28 November 2018.

on one aspect of the problem –awareness– but could not ensure continued access to these valuable data.

One path to a solution that we envisioned for a more sustainable solution was an updated version of the Federal Depository Library Program (FDLP) – a program that distributed copies of many government publications to libraries around the country<sup>21</sup>. This was a very successful way to distribute government information to the public when most of that information was in print. The program is not designed to work with data and other born-digital information, though. As this has become the most common form in which government information is published, agencies and libraries need to work together to find new ways libraries can serve as backup stewards of the information.

As the project had progressed though, we had spoken to more and more people working across different fields who work on different aspects of this problem. Librarians and archivists were thinking about how to incorporate digital-born data into their government document and data collections. People working for the federal government within agencies and in dedicated data centers have been employing thorough backup strategies, improving their metadata, and streamlining workflows in efforts to make data more accessible<sup>22</sup>. People in the civic technologies world were creating tools to download and work with federal data more efficiently. Open government and open data advocates were working with public officials to draft policies to encourage openness and sharing of resources. Journalists and humanists were telling stories about how these data affect us. Somehow, these and still more parties were not often working together, or indeed aware of all the work being done outside their own fields. It became clear that the only way to solve this problem would be to get these different groups to work together on sustainable solutions.

We organized a meeting of many of the people we'd been talking to throughout this project. We wanted to share ideas about solutions to this problem as well as come together to get a better understanding of the complexity of the problem and create new partnerships that would try new solutions. We called this meeting the Libraries+ Network meeting<sup>23</sup>.

### *Partnerships*

Libraries+ Network meeting was a great start to the conversation. The problem of having most federal data stored and shared only on federal servers and websites persists, but there are many groups continuing to work on it. The Preserving Electronic Government Information project has spent two years doing interviews and other research to continue to understand the problem the perspectives of those connected to it<sup>24</sup>. Data Refuge has transitioned with PPEH to Data Refuge

---

<sup>21</sup> "FDLP Basics," Federal Library Depository Program, 2014, <https://www.fdlp.gov/fdlp-basics>, accessed 29 November 2018.

<sup>22</sup> Matthew S. Mayernik et al., "Stronger together: the case for cross-sector collaboration in identifying and preserving at-risk data," 2017, <https://www.esipfed.org/press-releases/stronger-together>, accessed 28 November 2018; Edward J. Kerns, "On the preservation of and access to NOAA's open data," *Libraries+ Network* (blog), 2017, <https://libraries.network/blog/2017/4/30/on-the-preservation-of-and-access-to-noaas-open-data>, accessed 27 November 2018.

<sup>23</sup> "Libraries+ Network Meeting," accessed November 27, 2018, <https://libraries.network/may-meeting/>.

<sup>24</sup> "Preserving Electronic Government Information," n.d., <https://www.pegiproject.org/>, accessed 28 November 2018.

Stories<sup>25</sup>, focusing on the work of connecting data to the communities who rely on them. Some university libraries experimented with ingesting government information in different forms and some continue to try new things<sup>26</sup>. A variety of public/private partnerships have grown as a result of bringing attention to these issues<sup>27</sup>. The Civic Switchboard<sup>28</sup> project is looking at this problem through the local lens and works to connect local government data producers with libraries and other partners to find ways to make data more accessible and better preserved. Of course, important work that has been foundational to Data Refuge continues as it always has: data.gov<sup>29</sup>, the US open data catalog continues to work with agencies to make data discoverable, and the End of Term Harvest project<sup>30</sup> will continue to archive government websites ahead of governmental transitions.

## Conclusion

While this chapter, and the Data Refuge project, is very focused on the US problem of having government data only available through government websites, this problem is not unique to the US. The concerns we had were based on the experiences faced by our own past administrations as well as those mentioned in Canada, Australia, and Brazil. Australia could be encountering the same kinds of concerns again with Scott Morrison<sup>31</sup> and Brazil will continue to face these challenges under Jair Bolsonaro<sup>32</sup>. Even in countries and communities with leadership that encourages open data and scientific research, these cases provide ample evidence that we should not take our born-digital information for granted. Discovering your country's policies and protections for data and other information is a good first step to ensuring access to this important information will not be lost to the public. There's much to learn from each other in this space and a great deal more to do. The Data Refuge team encourages you and your institutions to experiment with ways you can help advocate for and protect your countries' information.

---

<sup>25</sup> "Data Refuge," n.d., <https://www.datarefugestories.org/>, accessed 28 November 2018.

<sup>26</sup> Laurie Allen, Claire Stewart, and Stephanie Wright, "Strategic open data preservation," *College & Research Libraries News*, 2017; Andrew Battista and Stephen Balogh, "The challenge of rescuing federal data: thoughts and lessons," *Libraries+ Network* (blog), 2017, <https://libraries.network/blog/2017/5/5/the-challenge-of-rescuing-federal-data-thoughts-and-lessons>, accessed 28 November 2018; Anna E Kijas, "Engaging in small data rescue," *Libraries+ Network* (blog), 2017, <https://libraries.network/blog/2017/6/16/engaging-in-small-data-rescue>, accessed 28 November 2018.

<sup>27</sup> "The ease of discovering NOAA data," News, National Centers for Environmental Information, 2018, <https://www.ncei.noaa.gov/news/ease-discovering-noaa-data>, accessed 28 November 2018; Sayeed Choudhury, "An op-ed piece about data rescue and libraries by Sayeed Choudhury," *Data Conservancy* (blog), 2017, <https://dataconservancy.org/an-op-ed-piece-about-data-rescue-and-libraries-by-sayeed-choudhury/>, accessed 28 November 2018.

<sup>28</sup> "Civic Switchboard Introductory Blogpost," *Civic Switchboard Updates* (blog), 2017, [https://civic-switchboard.github.io/updates/post\\_1](https://civic-switchboard.github.io/updates/post_1), accessed 28 November 2018.

<sup>29</sup> "About data.gov," Data.gov, n.d., <https://www.data.gov/about>, accessed 28 November 2018.

<sup>30</sup> "Project Background," End of Term Web Archive, n.d., <http://eotarchive.cdlib.org/background.html>, accessed 27 November 2018.

<sup>31</sup> Nicole Hasham, "Architect of Paris Climate Accord says Morrison government's Emissions stance is 'anti-science,'" *The Sydney Morning Herald*, 3 Oct 2018.

<sup>32</sup> Herton Escobar, "'We are headed for a very dark period.' Brazil's researchers fear election of far-right presidential candidate," *Science*, 18 Oct 2018, doi:10.1126/science.aav7518, accessed 27 November 2018.