Against the Grain

Volume 29 | Issue 6 Article 11

December 2017

Maintaining Access to Public Data-Lessons from Data Refuge

Margaret Janz
University of Pennsylvania, mjanz@upenn.edu

End of Term 2016 Presidential Web Archive
from page 29

Finally, the team suggested that the most helpful activity for volunteers was to nominate the URLs of the items that they believed most at risk via the URL Nomination Tool. This influx of nominations helped identify a wide range of content from websites to individual PDFs and datasets. It was a great help, and it allowed people to contribute in a way that they found meaningful. It also exposed a problem with the project: the team needed a better web presence to communicate with the public. Currently, the team has a **Twitter** account that was active during the project, but that is clearly not enough, as it is difficult to use as the only primary news and information outlet. In addition, the EOT project's interface, which is hosted by the **California Digital Library**, wasn't designed to have a section that listed new content, so updating the public via this resource simply wasn't possible. Now, one of the major goals for the 2020 EOT project is to have a better news and information platform for communicating with those who are interested, including information about the project and how people can help.

Conclusion

The End of Term projects in 2008, 2012, and 2016 were volunteer efforts by a number of institutions across the U.S. The time, effort, and infrastructure are all donated by the participating organizations. The individuals from these institutions are the ones that moved the project forward and made it successful. The 2016 election cycle offered new challenges and opportunities in relation to project management, channeling user interest, fielding media requests, and gathering and sharing the harvested content. While there were challenges, they were insignificant in comparison to the overall benefit of the project, as well as the accomplishments of the project and its project team.

References

Data Refuge – *https://www.datarefuge.org/*

End of Term Archive Website – *http://eotarchive.cdlib.org/*

End of Term Publications Collection – *https://digital.library.unt.edu/explore/collections/EOT/*

National Archives and Records Administration (2008). *Web Harvest Background Information.* Available from *https://www.archives.gov/files/records-mgmt/pdf/nwm13-2008-brief.pdf*.

Presidential Term 2004 Web Archive – *https://www.webharvest.gov/*

Social Feed Manager – *https://gwu-libraries.github.io/sfm-ui/*

URL Nomination Tool EOT 2016 – *http://digital2.library.unt.edu/nomination/eth2016/*

URL Nomination Tool EOT 2016 Bulk – *http://digital2.library.unt.edu/nomination/eth2016_bulk/*

U.S. Digital Registry – *https://www.digitalgov.gov/services/u-s-digital-registry/*

Maintaining Access to Public Data: Lessons from Data Refuge

by **Margaret Janz** (Scholarly Communications and Data Curation Librarian, University of Pennsylvania, Philadelphia, PA) <mjanz@upenn.edu>

An Abbreviated History of Data Refuge

The **Data Refuge** project began in December 2016 after fellows in the **Penn Program for Environmental Humanities (PPEH)** grew concerned about how the incoming presidential administration might find ways to limit access to federal climate and environmental data. These concerns stemmed from a public denial of climate change from key figures within the administration, and its stated intent to dismantle the **Environmental Protection Agency (EPA)**. Previous administrations had taken actions to limit these data, including that of **George W. Bush**.[1] There have also been similar actions taken abroad. Canada's **Stephen Harper**, for example, closed governmental libraries of environmental information[2] and made rules to prevent governmental scientists from communicating with the public.[3]

With these precedents in mind, the **PPEH** fellows, the **PPEH** program director **Bethany Wiggin**, **PPEH** coordinator **Patricia Kim**, and librarians from **Penn Libraries** wanted to create a refuge for these federal data by holding what we called "data rescue" events. We quickly got to work planning **DataRescue Philly**, which would feature a teach-in, a panel discussion, and a day of data archiving, which would be informed by a similar event held in Toronto[4] roughly a month before our event.

As the fellows started preparing for the teach-in and panel discussion, **Wiggin**, **Kim**, and the librarians — primarily **Laurie Allen** and myself — began discussing how to go about backing up these data locally. **Wiggin** reached out to **Mark Phillips** at the **University of North Texas** who works on the End of Term (EOT) Harvest, a project that aims to archive government websites ahead of presidential administration changes. **Phillips** told us that one limitation of the project is that the web crawler it employs only goes a few layers deep into the pages. We could provide support by seeding more lower-level URLs to the EOT project and we began thinking about the ways this could be done.

Seeding the EOT project was a great way to have **DataRescue Philly** attendees participate, particularly those who are less tech savvy, but the web crawlers used by EOT are unable to capture all types of digital information. Large data files, complex databases, and embedded and interactive data interfaces are not picked up by most web crawlers and need to be scraped or downloaded some other way. We had been in touch with a group called **Climate Mirror** that was working on doing just that. At the time, the volunteers with **Climate Mirror** were downloading federal data and hosting it on their own servers around the world. We worked with them to help set priorities and avoid duplication. While we were impressed with the tireless efforts of **Climate Mirror** volunteers, as librarians and academics we were concerned about how researchers using these data in the future could have confidence in the copies. It's easy enough to take the copied version and compare it to the original. However, if the original is taken away, it's much more difficult for someone to trust that the copy is the same. This became the challenge our team focused on ahead of **DataRescue Philly**.

We decided that one way to instill some amount of trust would be to require multiple quality checks before data would be archived in **Data Refuge's** cloud storage, and cataloged in our datarefuge.org open data catalog. Additionally, we required that anyone performing the checks would need to sign off on their assessment by including their name in the data's metadata. If the participant preferred to stay anonymous, a registered username could be

continued on page 32

used in place of their real name. This was not the optimal solution to the question of trust, but we felt it was a sufficient solution for our purposes.

An Event Becomes a Movement

In the meantime, our work in this area caught the attention of the media. We were fielding a large number of interviews, some in high-profile outlets. We started hearing from other institutions and individuals who wanted to help however they could: share storage space and technical skills, share their stories, or host their own **DataRescue** events. The response was beautiful and overwhelming. **DataRescue** events started being planned all over the country — and a few abroad — over the next several months. Many of the events were held at universities, and they were often planned by graduate students, civic tech groups, and small groups of librarians. During **DataRescue Philly**, we, along with incredible partners, notably **Justin Schell** from **University of Michigan**, **Ben Goldman** from **Penn State**, and **Rachel Appel** and **Delphine Khanna** from **Temple University**, developed a workflow for data archiving that we were able to share with these events. Members of the **Environmental Data Governance Initiative (EDGI)**, an organization that shared our concerns and with which we'd worked closely, also developed a workflow for seeding the EOT that they introduced at **DataRescue Philly**. We shared these workflows with other **DataRescue** event organizers, and those of us who were most familiar with the details helped organizers prepare and then troubleshoot issues remotely during their events. By June 2017, about fifty individual **DataRescue** events had taken place, thousands of URLs had been seeded to the EOT, and over 400 datasets had been uploaded into *datarefuge.org*.

Lessons

The workflows developed in January that most events used were a great response to our concerns, but we knew this plan of action was not a long-term, sustainable way to ensure continued access to these data. We are so proud of the work volunteers did at the many **DataRescue** events throughout the first half of 2017 and we learned so much from them and the other amazing people we spoke to during this time. These lessons would serve as the cornerstones in moving the project to the next phase.

One important lesson learned by **DataRescue** event attendees who worked on seeding URLs to the EOT was how government websites are organized. At first blush, government data and information appears to be a rabbit hole of disorganized fragments. The more time we spent with it and the more we spoke with data creators within the agencies, the more we understood that the information they provide is designed to serve the public's various and specific needs for short-term or immediate access. They're quite successful at achieving this goal, but the nonlinear organization makes it very difficult to keep track of what exists so it can be captured and preserved.

Not knowing what or how much data the government creates is a major obstacle for efforts to back up and maintain access to them. *Data.gov* is one attempt at keeping track of and cataloging federal data. *Data.gov* is an overarching catalog of open federal data. The small *data.gov* team has done an amazing job working with agencies to easily and incrementally make an inventory process simple, more inclusive, and largely automatic. An agency works with *data.gov* to set up an account and learn the workflow, and then the agency can create metadata files that *data.gov* can automatically read and import into their catalog. This is a fairly low effort addition to an agency's workflow. After learning more about how *data.gov* works, we at **Penn** think libraries could support and adapt the process in order to catalog the federal, state, and local data that matter to their researchers.

Another lesson we learned about federal data is that they share the various vulnerabilities of all born-digital information. Different technical vulnerabilities put born-digital information at risk. For example, proprietary file formats become outdated. Hardware breaks down over time, as does the information itself as bits corrode and files become corrupted. A lack of description, context, or sufficient documentation also renders data useless.

Political factors are another potential risk for these data. Not only might an administration actively attempt to limit access to data, more passive measures such as cutting budgets is another way to lose curatorial staff and fail to meet maintenance priorities. There may be legal protections for some data otherwise vulnerable to political risks, but the enforceability of those protections may or may not be apparent. Weighing the risks inherent to specific datasets to assess their vulnerability is an important part of prioritizing our work. We spoke to a number of the stewards who work with these data within agencies and in affiliated data centers, and their intimate knowledge about the data puts them among the best suited to make these assessments. Their expertise is integral to protecting access to these data.

A lesson we set out to impart through **DataRescue Philly** and other events was that federal data are more than products of specific research projects and legislatively-mandated administrative functions. It was important to us to have a path at our event that focused on telling the stories of how these data are used by local organizations and professionals to make decisions that impact the community on a daily basis. City planners, architects, real estate developers, and social service providers are just a few examples of groups that rely on these data to improve life for citizens in their cities and towns. Raising awareness that data aren't only used for scientific study, and connecting data to humans makes the issue more pressing for a much larger group. To quote **Eric Holthaus**, a climate journalist and friend to **Data Refuge**, "We are all part of this story. This is our story, we are shaping it every day."[5]

The most significant lesson that came out of **Data Refuge** concerns the nature of the problem we sought to solve. From the very beginning of the project, many people generously offered to provide storage space and technical skills for our efforts. Technical solutions are all important for working in this problem space, but we found as we dug deeper that technology is only one part of the problem; many technical solutions have been attempted or considered by various stakeholders at various points in history. The more complicated problem is one of culture and communication. All of the professionals who work with these data have established workflows to meet their own internal needs. While many groups have overlapping goals, it's rare that one group's workflow works nicely with another's. Getting any group, in any scenario, to alter its workflow to benefit a different group is enormously challenging. These changes also require excellent, reciprocal communication, which is in itself very difficult. *Data.gov's* simple metadata file creation is one great example of how these challenges can be overcome.

Moving to the Libraries+ Network

Throughout spring 2017 we continued to connect with a wide variety of people who work directly and indirectly with federal data. We spoke to many librarians hosting **DataRescue** events and started thinking that a network of libraries working to backup and archive these data could be a solution. This was similar to an idea articulated by **Jim Jacobs** and **James Jacobs** in their work with **Free Government Information** (*https://freegovinfo.info/*): a sort of reboot of the **Federal Depository Library Program (FDLP)** oriented toward the collective distributed management of federal digital content.[6]

We also talked to city planners, people in the open data community, researchers in federal agencies, data managers and curators, journalists, and archivists. Just like the librarians we'd spoken with, all of these knowledgeable stakeholders have been thinking about how to make these important data and other born-digital resources available for the long haul in one way or another. Each group had been doing great things in their own communities, but no one group had solved the problem. No one group had identified all of the challenges; blind spots existed for everyone. As we pieced together the work being done, we could tell that even with all the pieces, there were still blind spots. This problem can't be solved by a network that consists solely of libraries; we need a network with all these key partners working together. We decided the best thing to do would be to connect these groups and get these brilliant people to talk to each other, identify the challenges they face, and try to define the problem space so that we can all start experimenting with long term solutions.

Libraries+ Network May Meeting

On May 8-9, 2017, we did just that. Together with the **Association of Research Libraries (ARL)** and the **Mozilla Foundation**, we held a meeting with many of these stakeholder groups

continued on page 33

Maintaining Access to Public ...
from page 32

in Washington DC at **New America,** a think tank that focuses on technology and policy. One outcome was the mapping of the problem space (*http://libraries.network/problem-space*), which serves as a helpful reminder of what we're working towards, and that there will be neither a single nor a simple solution.

The meeting also got the group talking about the work that's been done so far and where we'd like to be in 2020. Some projects started to emerge by the end of the two-day meeting and attendees left with some ideas about paths forward. The meeting was dense and brought to light many challenges and opportunities. Many who are tackling their pieces of this endeavor are still in planning mode, but updates will continue to come forth.

Our team at **Penn** has only just begun to think about how to continue these efforts and support the overarching goals, and more interested organizations continue to reach out to us. The storytelling project continues to grow and expand with **Wiggin** and others. As we rethink our repository services at **Penn**, we're discussing instituting a catalog of data being created or used by our researchers and employing other lessons from **Data Refuge**. Regionally, we think there's great promise in the project that the **University of Pittsburgh** and the **Carnegie Library of Pittsburgh** are doing with the **Western Pennsylvania Regional Data Center** and the **Urban Institute**. On the national level, we're watching the **Code for Science and Society** as they work to pilot a mirror of *data.gov* that inventories federal datasets that are already being archived at research institutions. We're also really excited about the work being done by the **Preservation of Electronic Government Information (PEGI) project** and the Government Records Transparency group of the **Digital Library Federation**.

Stay Involved, Y'all

We know there are many paths to reach this goal. The workflow we used initially with **DataRescue** events has been retired, but we still have a number of other ideas for hosting events to engage your community on our website: http://www.ppehlab.org/datarescueworkflow. People also frequently ask us what their institutions should do to help our efforts. Our answer is always the same: Something. Anything. Figure out what's important to your communities. Consider your capacity for doing something. Experiment. Then — and this is key — report back so we can learn from and build off each other. We can only solve this problem together.

Endnotes

1. **Cheryl Hogue**, "Bush's Legacy at EPA," *Chemical & Engineering News* 86 (51): 27-31. *http://pubs.acs.org/cen/email/html/cen_86_i51_8651gov1.html*

2. **CBCNews**, "Research library's closure shows Harper government targets science 'at every turn,' union says," last modified August 21, 2015, *http://www.cbc.ca/news/canada/calgary/research-library-s-closure-shows-harper-government-targets-science-at-every-turn-union-says-1.3199761*.

3. **Lesley Evans Ogden**, "Nine Years of Censorship." *Nature* 533 (7601): 26-8, last modified May 3, 2016, *http://www.nature.com/news/nine-years-of-censorship-1.19842*.

4. **Kathleen O'Brien**, "U of T Preserving Environmental Websites in Response to Trump Presidency," *U of T News,* last modified December 14, 2016, *https://www.utoronto.ca/news/u-t-preserving-environmental-websites-response-trump-presidency*.

5. **Eric Holthaus**, "Final thoughts that NYMag Story." *Today in Weather & Climate*, last modified July 17, 2017, *https://tinyletter.com/sciencebyericholthaus/letters/today-in-weather-climate-final-thoughts-that-nymag-story-edition-monday-july-17th*.

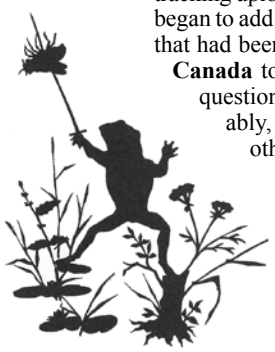
6. **James A. Jacobs** and **James R. Jacobs**, "A Long-Term Goal For Creating A Digital Government-Information Library Infrastructure," *Libraries+ Network*, last modified February 27, 2017, *https://libraries.network/blog/2017/3/7/a-long-term-goal-for-creating-a-digital-government-information-library-infrastructure*.

Documentation as Data Rescue: Restoring a Collection of Canadian Health Survey Files

by **Kristi Thompson** (Data Librarian, Leddy Library, University of Windsor) <kristi.thompson@uwindsor.ca>

Background

In Canada, most nationally representative survey data is collected by **Statistics Canada**, our national statistical agency. **Statistics Canada** data are generally considered to be of high quality, and the agency has long been the primary source for nationally representative surveys of the Canadian population. In American terms, **Statistics Canada** — which takes the straightforward, if acronym-limiting, Canadian standard for naming federal agencies with a guiding noun followed by "Canada" — roughly takes the place of the **Census Bureau**, the **Bureau of Labor Statistics**, the **National Center for Health Statistics**, and the **Center for Education Statistics**, as well as collecting data on behalf of a number of other departments and agencies. Once collected, data are published through several outlets including the **Data Liberation Initiative**, a program in which data files are processed by **Statistics Canada** into formats suitable for use by researchers and students, and then released to a country-wide network of librarians and library representatives for distribution at their respective academic institutions. However, as a single agency with a broad mandate in a very large country with a relatively small population base, they are not able to collect, process, and release nearly as much survey data as researchers might wish. In addition, other government agencies also maintain large primarily administrative data collections to support their own operations. These collections generally do not make it into the Statistics Canada-to-university data pipeline and at one point were largely inaccessible.



In 2011, the Government of Canada launched an open data pilot, a move that was applauded by data librarians and researchers across Canada as well as internationally. An open data portal soon provided access to thousands of geospatial and economic datasets, and in 2012 the pilot became a permanent program.[1] In 2014, the Canadian *Directive on Open Government* came into effect, requiring that data be "released in accessible and reusable formats."[2] Soon departments ranging from Agriculture and Agri-Food Canada to Veterans Canada began uploading data collections to the portal.

The Collection

One department adding data to the portal was **Health Canada**, the national public health agency. Although the portal lacks a system for tracking upload dates, it is apparent that at some point the agency quietly began to add to the portal a collection of public opinion research studies that had been conducted by various survey firms on behalf of **Health Canada** to assess opinions and behavior on policy-relevant health questions. These surveys were quite unknown except, presumably, to people who peruse internal **Health Canada** reports. In other words, this was a treasure trove of unmined, nationally representative survey data on Canada. In 2015, the author accidentally came across this data collection and realized that it was likely to be of great value to researchers if the data were to be made available in appropriate forms for research use. Unfortunately, the files as released were difficult, and in some cases impossible, to use.

continued on page 34