Laurie Allen, Claire Stewart, and Stephanie Wright

Strategic open data preservation

Roles and opportunities for broader engagement by librarians and the public

The June 2017 ACRL/SPARC Forum at the ALA Annual Conference focused on recent efforts to build cooperative programs to ensure persistent access to open data, including science data provided by the U.S. federal government. Data Rescue events, inaugurated at the University of Pennsylvania, catalyzed librarians, scientists, technologists, and other open data advocates to build a broad and resilient coalition to ensure against future data loss. Here, the three speakers from the forum reflect upon their own experiences with Data Rescue events and how they view opportunities for collective action going forward.

University of Pennsylvania

The Penn Libraries joined the Penn Program in Environmental Humanities (PPEH) in late 2016 in an effort to safeguard federal environmental and climate data as part of the Data Refuge project. The Data Refuge effort, grounded in concerns from PPEH Graduate fellows and from scientists, journalists, and others, grew into a broad and open network supporting more than 50 local events.

These Data Rescue events brought thousands of people together to support ongoing web archiving efforts, as well as downloading and harvesting many web-based datasets that are not accessible to web archiving. By the end of spring when most events were completed, many thousands of websites were archived in the Internet Archive, and terabytes of data were harvested and stored in datarefuge.org and within library repositories. The events also included teach-ins, panels, and a wide range of locally created opportunities for community learning. Storytelling programs provided opportunities for people to reflect and learn about the vast role of federal datasets in the daily lives and health of communities, and the act of harvesting, web archiving, and identifying valuable datasets was eye-opening for many of us as we learned just how deeply complex and fragile the digital information landscape can be.

Working alongside PPEH and many others in supporting these events as they gained momentum throughout the winter and into spring was deeply inspiring for many of us at the Penn Libraries. Data Refuge allowed us to learn from people across scientific communities, within open data communities, talented government data stewards, digital preservation experts, data storytellers, activists, artists, and many librarians working at institutions who saw a role for themselves in ensuring that multiple copies of vital data would remain available for our communities into the future.

Laurie Allen is assistant director for digital scholarship, at the University of Pennsylvania, email: laallen@upenn.edu, Claire Stewart is associate university librarian for Research and Learning at the University of Minnesota, email: cstewart@umn.edu, and Stephanie Wright is program lead for Mozilla Science Lab at the Mozilla Foundation, email: stephanie@mozillafoundation.org

Moving across these communities helped us see a tremendous opportunity that some libraries are already beginning to take advantage of—to reach out to federal data publishers and stewards and open data advocates in forming new partnerships.

As spring advanced, we decided to retire the bucket brigade workflow that we had collaborated in developing over the course of winter and turn our attention to absorbing and sharing the lessons of the project in other ways. We joined with ARL and Mozilla to form the Libraries+ Network[1] to support further collaboration across communities and hosted a large meeting in Washington, D.C., in May to bring together leaders from government, open data, open science, and library communities, among others.

Across that meeting and many other activities, one lesson that repeatedly emerged was the inseparable nature of data and stories. That is, the more we listened to the stories from communities about their environment, the more valuable the data became. And the more we dove into the technology, politics, and logistics of copying and storing data, the more it became clear that stories were our most vital tool for understanding the creation and stewardship of federal information, as well as its use.

To expand and enrich the effort to share stories about the uses of federal data, Bethany Wiggin and PPEH continue collecting data stories as part of the Three Stories in Our Town project with support from the National Geographic Foundation. Within the libraries, we are now pursuing and supporting new collaborations for safe and distributed approaches to federal data replication and preservation, and broadening our interest in local collaborations. We continue to look outwards towards the data needs of the wider world, and are beginning to look with fresh eyes at the data needs of the University of Pennsylvania community, rethinking our role in caring for the data that is produced locally and on campus, and for the data that is needed by members of our community now and into the future.

The University of Minnesota Libraries

The University of Minnesota (UM) Libraries, in conjunction with colleagues in Liberal Arts Technologies and Innovation Services, hosted a Twin Cities Data Rescue event on February 24 and 25, 2017. Modeled on events[2] held around the country, the event attracted 150 participants who took on roles such as selecting URLs to harvest, organizing and describing datasets, and scraping unharvestable sites. Many of the participants were new to working with data, metadata, and web harvesters. It was an eye-opening, though fulfilling, experience, though the 15GB harvested and 26 datasets "bagged" over the two days were nothing more than a drop in the ocean of valuable data. The event attracted local media interest, including a Minnesota Public Radio interview with our Government Publications and Regional Depository librarian.

As with many other events, one of the Twin Cities event's greatest benefits was creating opportunities to connect with the broader campus and local community and to identify advocates for the cause of well-preserved and accessible science data. But there were also some in our community who wondered why the effort was needed and whether the data were really at risk. This presented opportunities to share, and to reflect internally, about the libraries existing research data program, and to consider expanding its activities to include partnerships around preserving public data.

Consistent with the university's land grant mission, the UM Libraries have offered research data services for many years. In fact, some of the earliest efforts have been the preservation of critical government information through our Regional Depository of U.S. Government Documents. The libraries have been gradually increasing our investment in services to support public access to information, including data. Our regional program expanded to a three-state program in 2010. We offer robust data management education to our local research community, which has proved particularly popular with graduate students, who have overfilled every research data management boot camp we've offered. In 2015, services expanded to

include a Data Repository for the University of Minnesota. We are developing a new strategic plan for research data, exploring many new potential services.

The conversations sparked this year have also brought welcome attention to the importance of multi-institution collaboration and alignment with faculty initiatives. It has also highlighted opportunities for libraries to expand their work with local agencies, including government bodies. One such opportunity may arise through UM's work with the Big Ten Academic Alliance (BTAA) Geoportal,[3] which provides advanced discovery tools for more than 6,000 GIS data files and historical maps. The project has begun to grapple with the possibility of expanding its role to include primary preservation of these data. Some local and state agencies, under increasing financial pressure, face difficult choices between providing access to current versions of spatial information, and retaining older versions. These are truly public data at risk, and their loss would be catastrophic to researchers and policymakers. Of course, academic libraries are far from immune to financial pressures, but the BTAA Geoportal project is demonstrating that a modest shared financial commitment can yield significant returns. UM Libraries, like many academic libraries, will be seeking to balance its sense of commitment to broader data preservation with a pragmatic view of potential financial impact and a strong interest in working collectively with academic and nonacademic partners.

Mozilla Science

While well known for the Firefox browser, Mozilla is less known for the Foundation[4] behind the browser, which has a mission to keep the Internet healthy,[5] open, and accessible to all. One of the areas in which they try to do this is through the Mozilla Science Lab (MSL),[6] a program within the Foundation, focused on providing support for open data and open scholarship. The organization believes that building the capacity of researchers, librarians, educators, and developers through training programs, collaborative events, mentorships, and fellowships will lead to better adoption rates for open research and mobilize communities to advance open, data-driven science.

MSL became involved when one of our Fellows for Science,[7] Danielle Robinson, heard about the Data Refuge events taking place around the country and jumped into action to host one at the Mozilla offices in Portland, Oregon, along with another Mozilla Network member in Portland, Max Ogden. The two had several conversations with the Data Refuge team at the University of Pennsylvania and, in the end, rather than spending the evening downloading new datasets, the Portland team decided to focus on creating metadata files for the datasets already downloaded by others. In preparation for that, they developed a tutorial on how to create a metadata file in JSON and shared it on the GitHub repository created for the event,[8] where more experienced participants started. Those new to GitHub were given a quick overview of GitHub and how to use it for the project, then they moved on to the metadata creation with the others.

The event took place one evening after "normal" work hours. Pizza, drinks, and eventually coffee, were provided to keep energy levels high. This was a true community-building event and participants of all levels of experience were invited to join in. All that was required was a laptop, interest, and a GitHub account. There were people from widely varying backgrounds: beginners to laptop jockeys, academia to general public. Contrary to expectations, they didn't think metadata was boring. They all showed passion and interest toward the cause of making sure this data was available for all, long into the future. It was enough to raise the (still unanswered) question: How much are we being limited on what we can do collaboratively because we are held back by the assumptions we make about how others perceive our work?

It was this question that led MSL to pursue a partnership with Penn Libraries and the Association of Research Libraries to form a broader community of experts across libraries, archives, government, lawyers, and more, as part of a longer-term plan around the issues of long-term open access to federal data.

(continues on page 495)

DF: I know that there are many reasons to take a new job. Sometimes you simply need out of a bad or boring situation. Sometimes you need more money or a bigger challenge. Regardless of why you are looking for your next opportunity, our experiences seem to share two very important themes. First, we were all excited by a new challenge, something that would drive each of us and help us grow personally. Second, we all interviewed our potential employers as much as they interviewed us. We had questions that needed answers, at least perfunctorily, and we had benchmarks that we were looking for. Though I would generally encourage self-confidence and making that humbling leap into leadership, it is important to ask yourself questions. Is this the right time? Is this the right position? Is this the right institution? An interview is not a one-way street, and becoming a leader, not just a manager, director, or dean, is not either. Being a leader is about lifting up your team, not yourself, so it is important that you know what you are getting yourself into from the start.

Conclusion

This is part one in a three-part series. In part two, Powers, Garnar, and Fife will address integrating themselves into new organizations and teams. They will focus on the essential nature of humility for new leaders, asking questions, identifying stakeholders, and accepting that you can still lead, while admitting that you do not know everything.

("Strategic open data preservation," continues from page 484)

We wanted to make sure the open data community members we worked with were a part of this larger community, called the Libraries+ Network, because there was so much these individual communities could learn from each other. The details are laid out in the recently released report[9] from the workshop.

To quote Danielle Robinson, "Usually academia does not hack! We form subcommittees." In the open data community, one sees a lot of hacking, which for this purpose we are using the *Oxford English Dictionary* definition of "providing a quick or inelegant solution to a particular problem."

Hacking has its benefits. These events were successful because event hosts could work quickly to put them together without having to wait for central coordination and committees to agree upon standards. It also has its drawbacks: a greater possibility of duplication of effort or needing to revisit work that was done to "clean it up." Neither of these workflows is right or wrong, but each has valuable components that can help the other. Our next steps are to take the lessons and viewpoints learned from these events and the Libraries+ Network and think bigger about what we can accomplish together.

Notes

1. Libraries Plus Network, "Libraries Plus Network," *Libraries+ Network,* accessed August 13, 2017, https://libraries.network/.

2. University of Pennsylvania Program in the Environmental Humanities, "DataRescue Events," *PPEH Lab,* accessed August 13, 2017, www.ppehlab.org/datarescue-events/ and www.ppehlab.org/datarefuge/.

3. Big Ten Academic Alliance, "Big Ten Academic Alliance Geoportal," accessed August 13, 2017, https://geo.btaa.org/.

4. Mozilla, "Mozilla Network," *Mozilla Network,* accessed August 13, 2017, https://network.mofoprod.net/.

5. Mozilla, "The Internet Health Report v.0.1," *The Internet Health Report,* January 2017, https://internethealthreport.org/v01/.

6. Mozilla, "Mozilla Science Lab," *Mozilla Science,* accessed August 13, 2017, https://science.mozilla.org.

7. Mozilla, "Mozilla Science Fellowships," *Mozilla Science,* accessed August 13, 2017, https://science.mozilla.org/programs/fellowships/overview.

8. Danielle Robinson, *Data-Rescue-PDX: Volunteer Guide, and Other Materials for DATA RESCUE PDX, 2017,* https://github.com/daniellecrobinson/Data-Rescue-PDX.

9. The Grove, "Libraries+ Network Meeting Report," May 8, 2017, https://libraries.network/s/may-meeting-report.pdf.