April 4, 2017

Title: Stronger together: the case for cross-sector collaboration in identifying and preserving at-risk data

Authors: Matthew S. Mayernik, Robert R. Downs, Ruth Duerr, Sophie Hou, Natalie Meyers, Nancy Ritchey, Andrea Thomer, Lynn Yarmey

Introduction

In the past few months, a range of grassroots initiatives have gained significant momentum to duplicate US government agency data. These initiatives are inspired by recent reports that scientific data and documentation have been removed from government websites, and by concerns over US budget proposals that slash scientific budgets [1]. National media outlets have reported on numerous "data rescue," "data refuge," and "guerrilla archiving" events that have taken place around the US and in Canada during the past few months [2]. Many of these events have focused on creating copies of Earth science data generated and held by US federal agencies. These activities have attracted hundreds of volunteers who have spent considerable time and energy working on duplicating federal data.

Early connections have been made between the rescue volunteers and the federally-funded data community; these conversations have highlighted some of the different perspectives and opportunities regarding agency data. The two goals of this document are to provide the perspective of Earth science data centers holding US federal agency data on this issue, and second, to provide guidance for groups who are organizing or taking part in data rescue events. This paper is not a how-to document, and does not take a position on the political aspects of these efforts. Given the extent of the US government data holdings in the Earth sciences and other domains, it is inevitable that any grassroots data rescue will have to make strategic choices about how to invest their efforts. This document is intended to describe considerations for data rescue activities in relation to the day-to-day work of existing federal and federally-funded Earth science data archiving organizations.

The authors use the ‘data rescue’ terminology throughout this text to connect with the stated goals of the grassroots ‘data rescue’ communities, though we do wish to push back on the assumption that the data being targeted by these efforts are necessarily in need of ‘rescue.’ As we discuss below, many of these data are, in fact, well managed and safe, though sometimes in ways that are less-than-obvious to someone new to the domain. We look forward to working with these communities to develop a shared sense of risk for federal data.

Our context

As data managers, data curators, etc. (whom we will refer to in general as “data professionals”), we work with researchers, agency personnel, and the cross-agency data community to understand

requirements, identify critical metadata, standardize practices, and disseminate our work. We have strong communities-of-practice that emphasize accountability, sustainable solutions, efficiencies of scale, and the development of shared solutions to data challenges. Like librarians and archivists, we know our collections well and we are proud of our work, knowledge, and communities [3]. We recognize the value of data preservation, description, and long-term reuse, and in many cases have spent our careers improving our infrastructure, relationships, and workflows to better support the research community, agencies, and broader public.

We see our data as being intertwined with metadata, domain and technical requirements provenance, and users. With environmental data in particular, simply capturing a web site rarely captures the data itself as each URL may point to a different part of the data package. Data may be held as files on an ftp site, as records within a database of some form, and may even be held on off-line storage media such as tape. Capturing data files or contents of a database in isolation will rarely be sufficient to enable subsequent use of the data since often the contextual information needed to understand the format and meaning of that data will be contained in a series of metadata records, web pages, or other forms of documentation. The exact nature of these parts and pieces that comprise a complete data set will vary from center to center and sometimes from data set to data set. All of these pieces and their connections have been optimised over time to meet defined needs based on the best knowledge and resources available at the time.

We see data ‘risk’ as involving technical, metadata, policy, and resource considerations. In recent public conversations, “at risk” has been used to imply that data may be deleted or become inaccessible to the public now or in the future. From a public perspective, this would be visible as broken links [4] or the removal of a particular portal. However, from a data center perspective, these seemingly ‘lost’ data may well still be well-preserved and even accessible through professionally-managed federal infrastructure as data management systems are usually detached and insulated from changes visible on the web. As of this writing, it is unclear what currently open and publicly available US federal government data are actually “at risk”. In the US, laws and budgets have been proposed which would phase out the EPA and substantially reduce the budgets of other environmental agencies [5], but the implications of these proposals on the data stewarded by federal agencies are unknown. It is certainly true that having multiple copies of a data set held by multiple organizations is central to successful data archiving, but the legal precedent for actually deleting data is not clear. Regardless, data may be characterized as being “at risk” for many reasons, with risk factors including obsolete technology or data formats, lack of metadata, lack of expertise to interpret the data, and lack of funding to maintain data [6]. It is important to be clear about what risk factors are being used to motivate data rescue events. Being unclear about this can lead to confusion or misinformation being spread. Wired.com, for example, published two news articles within a week of each other that respectively characterized

NASA data as being a) in need of saving and b) not actually at much risk of being lost in their current homes [7].

Data center background

Numerous federal data centers, staffed with data professionals, infrastructure specialists, and often researchers themselves, offer usable, trustworthy data along with data preservation services. Data centers, archives, and digital repositories provide valuable services to support the long-term value and use of data by a particular domain or community [8]. Most data centers holding US agency data have some form of preservation plan that at least involves distributing multiple copies of the data in different geographic locations. Some of these centers, such as the NOAA National Centers for Environmental Information (NCEI) also have federal legal mandates for archiving particular data. US Federal data centers are also in many cases part of national or international data networks. At least nine US agency data centers, and a number of other federally-funded data centers, are members of the World Data System (WDS), an international federation that promotes and supports trustworthy data services [9]. Becoming a WDS member involves undergoing a certification process to validate the data center's procedures for effectively stewarding data over time. As another example, through the Federal Big Data program, NASA, NOAA and other agencies have started making copies of their most popular very large data sets available on a variety of cloud providers.

In addition, the National Archives and Records Administration (NARA) manages the archives from many federal agencies, including each presidential administration's websites, and the data and documents hosted on those sites [10]. Each White House's website also includes subsites from different government agencies and committees that are likely relevant to scientists (e.g. the Office of Science & Technology Policy (OSTP). Unfortunately the archiving process results in many broken links, and the appearance that documents and data are disappearing with each new administration.

It must be made clear, however, that data generated by federal agencies are not uniformly managed, and not all federal data resources are housed in formal data centers. There are vast quantities of data held by researchers in the federal government, academia, and industry, that have not been deposited into any repository. There are many reasons for this, which we will not discuss here as there is substantial literature on the topic [11]. What can happen in these cases is that as the researcher approaches retirement they start thinking about their legacy and typically drop off boxes of materials at their library, archive, or favorite data repository. Data centers may be woefully unprepared to do anything with this largess especially if the researcher is not available to answer questions.

In other cases, a repository may have been around since before the digital age. Many data centers maintain a legacy library and/or archive full of data in analog form, e.g. as maps, prints, or books, and consequently not fully available to the community for use [12]. These legacy collections of out-dated media (7-track and 9-track tapes, paper tape, floppy disks, etc.) need to be migrated to modern media and data formats, since the technologies for reading those old media are obsolete and often the media itself is degrading. Moreover, often these data predated the ASCII or unicode eras and may need considerable bit-level manipulation in order to be translated into something useable by today's technologies. These are data at considerable risk of being unusable by anybody unfamiliar with the original data collection effort.

Government data centers are typically happy to work with anybody interested in accessing or using their data. Some details on how to engage with data centers are described below.

Recommendations

This section outlines recommendations on how emerging grassroots data rescue initiatives can productively partner and collaborate with current data center services.

- Confirm the current risk level of data sets in line for rescue. For instance, though a data set may have appeared to vanish from a previously reliable access point, it may have a) been moved to another location; b) be duplicated in other locations that are less popular, less widely known, or require different methods of access (for instance, through an API rather than a web browser). Do a bit of research on the data set, its creators, and managers to confirm that it really is "at risk" like it might first appear.
- Interact with the data center
    - Contact the data center before rescuing their data at large scale for preservation or access purposes. Data center personnel will be able to give pointers for what data under their purview may be "at risk."
    - If you already have a lists of data sets or resources that somebody has declared may be "at risk," contact the relevant agencies for help in reviewing the list, getting connected to appropriate data center contacts, and facilitating documentation of data rescue activities.
    - Data center personnel may be able to guide you to the best mechanism for accessing the data (and associated metadata) from their facility. Hacking/scraping web pages may not be the most efficient way to download data and capture associated metadata, for example.
    - Let the data center know about your plans before asking dozens of volunteers to hit their web sites with large numbers of downloads. Contacting the center first will allow them to potentially provision additional web server space at a specific times.

    - Log-ins – Data systems may require log-in for data access in many legitimate cases, e.g. protection of sensitive or legally-restricted data, to gather usage metrics, to communicate data updates to the appropriate user communities, etc. Not all conditions for access are ill-intentioned. Sometimes sensitive data need to be protected (for instance, data that reveal the location of endangered animals, rare specimens, or rare artifacts need to be hidden from potential poachers), and data containing information about people needs to be carefully managed to mediate or prevent the sharing of personal data.
    - In some cases, technically bypassing log-ins or firewalls can be illegal. Data center staff can tell you why something requires a log-in, and how to access such data in compliance with appropriate policies.
    - If you are having difficulty contacting center staff, send a message to the data center's Help desk or User Services email address, which may be accessible from the FAQ, or Support link on their website. In the absence of a timely response from the data center's Help desk or User Services staff a written request should be sent to the data center's director.
- Gather all associated metadata and keep them with the data. Data center personnel can help you identify all of these parts and pieces and how to access them.
    - Metadata might consist of information structured in a web page, an XML file, a database, or other mechanism, and might include documents, images, or maps.
    - Gather, maintain and use all persistent identifiers (PIDs) associated with data. Many government data repositories assign Digital Object Identifiers (DOIs) or other kinds of identifiers to data sets following cross-agency standards [13]. PIDs enable persistent location, identification, and citation of particular data, which is critical to tracing their provenance and usage.
    - Provenance / chain of custody – All data must be traceable back to their original sources, and must have a demonstrable chain of custody including validation mechanisms such as checksums.
- Syncing efforts
    - Plan for maintenance and versioning. Many federal data sets change over time, with new data being added, or values being changed as errors are identified and fixed. Creating snapshots of data may exacerbate problems related to authoritative versioning and communication of changes.
    - If you allow users to access the rescued data
        - Link the rescued data back to the original source, using PIDs if possible.
        - Provide usage/download metrics back to the original source. Diverting traffic from the original data center to another data location actively hurts the data center, as they rely on usage metrics to understand community

needs, determine priorities, and demonstrate the value of their services to the scientific communities, the general public, and their funders.
        - Identify who is going to provide human services for these data, e.g. answer questions, provide help to users in understanding what data sets actually represent, and help people interpret data correctly.
    - Security - Government web sites have to meet legal requirements for information security, e.g. keeping their files safe from hackers [14]. Adopting this requirement for rescue efforts will help ensure that users know they are getting real uncorrupted files.
- Contribute expertise and effort in rescuing legacy data
    - Multiple international organizations and collaborative working groups have been working on the rescue of legacy data for decades [15]. As one example, the Data Rescue Interest Group within the Research Data Alliance (RDA- Data Rescue IG) currently is working on guidelines for the rescue of legacy data [16]. These initiatives would benefit tremendously from additional attention, effort, and resources.
    - If volunteers are participating in data rescue events as concerned citizens, there are many opportunities to contribute to these ongoing efforts to rescue legacy data. A very valuable contribution would be to hold events where people participate in citizen science-based data rescue efforts, e.g. http://weatherwizards.org/, https://www.oldweather.org/. Contributions of funds and technical expertise could also have significant impact on these efforts.

Conclusion - Working Together

Data management, curation, and preservation efforts are chronically under-resourced and overlooked, and we all care about data safety, accuracy, use, and preservation. The trustworthiness of data is critically intertwined with the factors described above, e.g. metadata, provenance, transparency, security, and community [17]. If those factors are not taken into account, rescued data will be of no use regardless of how many times they are duplicated. There are many data professionals and other stakeholders in the data management community collaborating formally and informally to provide stewardship and to identify at-risk data, curate at-risk data, and mitigate the chances for data to become "at risk." The grassroots data rescue efforts like DataRefuge and others have brought together an energetic, diverse community of passionate citizens and professionals with valuable skills and expertise. Initial connections between DataRefuge and broader communities such as the Research Data Alliance and ESIP have shown value, and point out important gaps and opportunities moving forward. The more we can work together to preserve the data that matter to us all, the more effective and sustainable the our work will be.

References

[1] Varinsky, Dana. (2017). Scientists across the US are scrambling to save government research in 'Data Rescue' events. Business Insider, Feb. 11, 2017. http://www.businessinsider.com/data-rescue-government-data-preservation-efforts-2017-2
Science News Staff. (2017). A grim budget day for U.S. science: analysis and reaction to Trump's plan. Science, Mar. 16, 2017. https://doi.org/10.1126/science.aal0923

[2] See for example: Dennis, Brady. Scientists are frantically copying U.S. climate data, fearing it might vanish under Trump. The Washington Post, Dec. 13, 2016. https://www.washingtonpost.com/news/energy-environment/wp/2016/12/13/scientists-are-frantically-copying-u-s-climate-data-fearing-it-might-vanish-under-trump/
Temple, James. Climate data preservation efforts mount as Trump takes office. MIT Technology Review, Jan. 20, 2017. https://www.technologyreview.com/s/603402/climate-data-preservation-efforts-mount-as-trump-takes-office/
Khan, Amina. Fearing climate change databases may be threatened in Trump era, UCLA scientists work to protect them. Los Angeles Times, Jan. 21, 2017. http://www.latimes.com/science/sciencenow/la-sci-sn-climate-change-data-20170121-story.html
Harmon, Amy. Activists rush to save government science data — If they can find it. New York Times, March 6, 2017. https://www.nytimes.com/2017/03/06/science/donald-trump-data-rescue-science.html

[3] Yarmey, K. and Yarmey, L. (2013). All in the Family: A Dinner Table Conversation about Libraries, Archives, Data, and Science. Archive Journal, Issue 3. http://www.archivejournal.net/issue/3/archives-remixed/all-in-the-family-a-dinner-table-conversation-about-libraries-archives-data-and-science/

[4] Herrmann, Victoria. (2017). I am an Arctic researcher. Donald Trump is deleting my citations. The Guardian, Mar. 28, 2017. https://www.theguardian.com/commentisfree/2017/mar/28/arctic-researcher-donald-trump-deleting-my-citations

[5] US H.R.861 - To terminate the Environmental Protection Agency. Introduced Feb. 3, 2017. https://www.congress.gov/bill/115th-congress/house-bill/861/all-actions

[6] Anderson, William L., Faundeen, John L., Greenberg, Jane, & Taylor, Fraser. (2011). Metadata for data rescue and data at risk. In Conference on Ensuring Long-Term Preservation in Adding Value to Scientific and Technical Data. http://hdl.handle.net/2152/20056

Downs, Robert R. & Chen, Robert S. (2017). Curation of scientific data at risk of loss: Data rescue and dissemination. In Johnston, Lisa (Ed). Curating Research Data. Volume One, Practical Strategies for Your Digital Repository. Association of College and Research Libraries. http://dx.doi.org/10.7916/D8W09BMQ
Griffin, R.E. (2015). When are old data new data? GeoResJ, 6: 92–97. http://dx.doi.org/10.1016/j.grj.2015.02.004
Ryan, H. (2014). Occam's razor and file format endangerment factors. Proceedings of the 11th International Conference on Digital Preservation (iPres), October 6-10, 2014: Melbourne, Australia (pp. 179-188). https://www.nla.gov.au/sites/default/files/ipres2014-proceedings-version_1.pdf
Thompson, C.A., Robertson, W. D., & Greenberg, J. (2014). Where have all the scientific data gone? LIS perspective on the data-at-risk predicament. College & Research Libraries, 75(6), 842-861. https://doi.org/10.5860/crl.75.6.842

[7] Molteni, Megan. Diehard coders just rescued NASA's Earth science data. Wired, Feb. 13, 2017. https://www.wired.com/2017/02/diehard-coders-just-saved-nasas-earth-science-data/
Molteni, Megan. Old-guard archivists keep federal data safer than you think. Wired, Feb. 19, 2017. https://www.wired.com/2017/02/army-old-guard-archivers-federal-data-safer-think/

[8] See e.g. Ramapriyan, H.K., Pfister, R. and Weinstein, B. (2010). An overview of the EOS data distribution systems. In Land Remote Sensing and Global Environmental Change (pp. 183-202). Springer New York. http://doi.org/10.1007/978-1-4419-6749-7_9

[9] https://www.icsu-wds.org/community/membership/regular-members

[10] https://www.archives.gov/presidential-libraries/archived-websites

[11] Douglass, K., Allard, S., Tenopir, C., Wu, L., Frame, M. (2013). Managing scientific data as public assets: Data sharing practices and policies among full-time government employees. Journal of the Association for Information Science and Technology, 65(2): 251–262. https://doi.org/10.1002/asi.22988
Tenopir, C., et al. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS One, 10(8): e0134826. https://doi.org/10.1371/journal.pone.0134826

[12] US Geological Survey. 2016 Data at Risk Project. https://www.fort.usgs.gov/ldi/2016-data-at-risk-project
National Oceanic and Atmospheric Administration. Climate Database Modernization Program.

https://www.ncdc.noaa.gov/climate-information/research-programs/climate-database-modernization-program

[13] Earth Science Information Partners (ESIP). 2011. Interagency Data Stewardship/Citations/provider guidelines. http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines

[14] https://www.dhs.gov/fisma

[15] E.g. the International Environmental Data Rescue Organization (IEDRO, http://iedro.org/). See also: Tan, L. S., S. Burton, R. Crouthamel, A. van Engelen, R. Hutchinson, L. Nicodemus, T. C. Peterson, F. Rahimzadeh. (2004). Guidelines on Climate Data Rescue. WMO/TD No. 1210. Ed. by P. Llansó and H. Kontongomde. Geneva, Switzerland: World Meteorological Organization. http://www.wmo.int/pages/prog/wcp/wcdmp/documents/WCDMP-55.pdf.

[16] Guidelines to the Rescue of Data At Risk, https://www.rd-alliance.org/guidelines-rescue-data-risk
See also the Interest Group's home page: Research Data Alliance Data Rescue Interest Group. https://www.rd-alliance.org/groups/data-rescue.html

[17] Yakel, E., Faniel, I., Kriesberg, A., & Yoon, A. (2013). Trust in Digital Repositories. International Journal of Digital Curation, 8(1). https://doi.org/10.2218/ijdc.v8i1.251
Yoon, A. (2017). Data reusers' trust development. Journal of the Association for Information Science and Technology, 68(4): 946-956. https://doi.org/10.1002/asi.23730