Against the Grain

Volume 29 | Issue 6 Article 13

December 2017

Data Mirror-Complementing Data Producers

John Chodacki
University of California Curation Center, john.chodacki@ucop.edu

Documentation as Data Rescue: Restoring a Collection ...
from page 34

document collections. We used the documentation to write a syntax file to read the data into a statistical software package, which we then used to check that the data matched the technical information in the data dictionary and that everything present in the data file was accounted for in the documentation. During this process, we needed to backtrack several times as we discovered inconsistencies between the data file and the documentation. We also performed some customizations to make the data easier to use and interpret before loading it into a data portal and saving an archival copy to a secure academic cloud. The syntax used to make changes to the data is retained with the documentation to help keep the process as transparent as possible for our data users. While working on this survey we kept notes on the steps that were taking to help streamline the process. These notes have been incorporated into the *Data Rescue and Curation Guide for Data Rescuers*, a how-to manual being developed by the group.

Lessons

One lesson from the experiences of the **Ontario Data Rescue Group** is that librarians without any technical or statistical background can still make valuable contributions to data rescue projects. Much of our work has involved searching for reports in government document collections and collating information on the different research projects from which our data rescue targets were derived. Data rescue does not always mean heroically saving files from deletion by malevolent custodians. Sometimes it means the library detective work of searching through archives of neglected government documents, cross-checking details to track changes in content over time, or trawling departmental contact lists in hope of reaching that one person who knows where a file originated.

Data rescue is a time-sensitive endeavor. Data collections that have been separated from the data creators, making it difficult to track down lost contextual information, are particularly at risk. Even data being preserved and shared with the best of intentions may be in need of rescue and curation. The point of curating data is to make sure that it will be available for use both now and into the future, because data without adequate accompanying documentation cannot be used.

The Ontario Data Rescue Group consists of:

Alexandra Cooper, **Queen's University**
Jane Fry, **Carleton University**
Walter Giesbrecht, **York University**
Vince Gray, **University of Western Ontario**
Vivek Jadon, **McMaster University**
Amber Leahey, **Scholars Portal**
Susan Mowers, **University of Ottawa**
Kristi Thompson, **University of Windsor**
Leanne Trimble, **University of Toronto**

Endnotes

1. **Government of Canada**, "Canada's Action Plan on Open Government 2012-2014," last modified Sept. 22, 2016. *http://open.canada.ca/en/canadas-action-plan-open-government*
2. Ibid, "Directive on Open Government," last modified October 9, 2014. *http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=28108*
3. **Jane Fry**, "Data Rescue in Canada, A Case Study," presented at the International Association for Social Science Information Services & Technology (IASSIST) conference (Ithaca, New York, June 2, 2010). *http://www.iassistdata.org/downloads/2010/2010_c3_fry.pdf*
4. **Alberta Research Council**, "The Alberta Hail Project Meteorological and Barge-Humphries Radar Archive," (UAL Dataverse, 2016). *http://dx.doi.org/10.7939/DVN/10672*
5. **Ontario Council of University Libraries**, ODESI, last modified November 21, 2017. *http://odesi.ca/*

Data Mirror: Complementing Data Producers

by **John Chodacki** (Director, University of California Curation Center) <john.chodacki@ucop.edu>

Data Mirror is a collaborative project between the **University of California Curation Center (UC3)** and **Code for Science & Society (CSS)**, a non-profit organization committed to improving access to data for the public good. We are interested in preserving federal data because we know that the research produced, collected, or funded by the federal government are an integral part of the rich tapestry of the nation's cultural and scholarly record, and are critical resources for advancing scholarship, public policy, and governmental transparency and accountability. However, we in the library and preservation community often forget that the data producers within the federal government have comprehensive preservation strategies and workflows of their own. Although we are focused on helping solve problems, many times we unnecessarily create duplicative or parallel solutions that cut the federal research groups out of the conversation and can cause additional issues down the road. The Data Mirror project (*datamirror.org*) is working to exemplify a different possible path forward.

Data Mirror is a complete, and routinely updated, copy of the main federal government research data portal, *data.gov*. Hosted by the **UC3** at the **California Digital Library (CDL)**, Data Mirror points back to the "datasets of record" on federal agency websites for routine access. Why? Because those are the copies that are cared for and handled by the data producers themselves, and therefore, those copies should be referenced and used by researchers. However, should these access paths become interrupted or inaccessible, Data Mirror also includes pointers to **CDL**-managed copies, as well as additional registered replicas hosted by other institutions. In this model, *data.gov* and the mandates that it works under remain the center of the workflow. Basically, Data Mirror works as a back-up of the existing systems and offers redundancy to the *data.gov* metadata catalog and preservation services to its underlying datasets. Providing alternative search and retrieval opportunities helps to ensure that these important data remain available for study and use in perpetuity while keeping existing Federal workflows intact. Without building entirely new systems or processes, government research groups can continue to rely upon their existing workflows.

We have worked directly with the team at *data.gov* to ensure we are respecting their existing workflows. With the support of the wider library and preservation community, we would like to enhance the Data Mirror portal to include the ability for our communities to propose enriched metadata or the addition of new datasets through the portal, which would be communicated back to the agencies and *data.gov*. It is that round-tripping of federal data preservation (through existing channels!) that would truly build long-term collaboration between those producing government data and those focusing on the preservation of government data.