Against the Grain

Volume 29 | Issue 6 Article 12

December 2017

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

Kristi Thmpson
University of Windsor, kthom67@uwo.ca

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.*

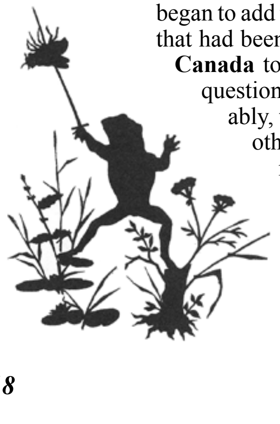
6. **James A. Jacobs** and **James R. Jacobs**, "A Long-Term Goal For Creating A Digital Government-Information Library Infrastructure," *Libraries+ Network*, last modified February 27, 2017, *https://libraries.network/blog/2017/3/7/a-long-term-goal-for-creating-a-digital-government-information-library-infrastructure.*

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

Documentation as Data Rescue: Restoring a Collection ...

from page 33

Canadian data librarians are used to dealing with well-documented and structured government survey files released by **Statistics Canada** through the **Data Liberation Initiative (DLI)**. These user-ready files are published in formats compatible with popular software packages for data analysis such as SPSS. They come with documentation that explains where, when, and how the data were collected, what questions were asked in the original surveys, and what codes were applied. The **Health Canada** data files lacked all this crucial supplementary information, and I found this mystifying in more ways than one: the data themselves were difficult or impossible to understand, and I was also puzzled that they had been released in such incomprehensible condition.

At our next meeting, I raised this issue with the **Ontario Data Community (ODC)**, a provincial network of academic data librarians and other professionals under the aegis of the **Ontario Council of University Libraries**. During our subsequent discussion, I discovered that members of the **ODC** were already working with additional Government of Canada data collections that were not available from **Statistics Canada**, including some vintage surveys held by **Library and Archives Canada** and census files residing in various university collections. After further discussion and investigation, in December 2015 a small group of volunteers from the **ODC** formed the **Ontario Data Rescue Group**. In forming our group, we were joining a tradition of Canadian university data rescue work, including efforts at **Carleton University**[3] and the **University of Alberta**.[4] We decided that as one of our first projects we would focus on the **Open Data Portal** and develop an inventory of at-risk survey files in need of rescue, with the hope of eventually sharing rescued data on the Ontario academic data portal **ODESI**.[5] We were particularly excited to discover survey files on topics that are not well covered in other Canadian public data sources, such as HIV and sexual behavior, adolescent drug use and attitudes, children's health and safety, and First Nations populations.

Unlike many data rescue projects, our group faced a situation in which the data files we were targeting were available through a stable government portal and in no apparent risk of disappearing. They were even available in open, non-proprietary formats such as .csv (comma separated values, a text format used by MS Excel and read by virtually any database software). The issue was not, in fact, a fear that this data would disappear, or that the software to read it would become obsolete, or any of the other usual data loss concerns. The issue was simply one of documentation.

In order to understand why data with inadequate documentation is in need of rescue, it is important to explain structured data files. For a piece of software, a new remote control, or an **IKEA** bookcase, a lack of documentation may make things difficult, but a determined user will often be able to proceed through trial and error. A survey data file is just columns of numbers, so this is not an option. An unlabeled column (or "variable") that contains nothing but the numbers "1" through "7" might represent a respondent's opinion on drug labeling practices, their level of education, a count of their current sex partners, or a measure of vegetable consumption. Without some way of knowing both what type of information is associated with each column and what each code in the column represents, a data file is useless.

Some of these data collections had been released with data dictionaries, which are text files that give a technical description of what each column contains. These files are not exactly user-friendly — a great deal of work is needed to ready them for actual use — but it is at least theoretically possible for a knowledgeable person to make use of them.

In other cases, the data was not accompanied by a data dictionary, but the original survey questionnaire was included. These files are even less useful; while the questionnaire could be used to make educated guesses about what question each column of data corresponds to, the meanings of the numbers in the columns could still be unclear and would probably require additional guesswork. In addition, the final version of a survey data file will often include a number of columns that do not correspond directly to questions in the original survey. Data can be grouped or recategorized, new variables can be created by combining other ones, and other variables can be added to document technical information relating to the original survey such as notes on which respondents were asked which questions. A questionnaire containing seventy questions might accompany a dataset with over one hundred variables. Dealing with a raw data file without a data dictionary is rather like attempting to translate a document in an unknown language without a dictionary.

A few data files were released without any documentation at all, only columns of numbers and a survey title. These surveys were in the most urgent need of rescue. Our only hope for rescuing these files was that many of these surveys appeared to be quite recent — when dates were available they ranged between 2009 and 2014. We hoped that this meant that people involved in the original data collection might still be working at **Health Canada** and would have access to original survey questionnaires and other files

Rescue

We first used the general contact addresses provided on the **Open Data Portal** to request the missing data dictionaries. We were not surprised when these requests failed to produce any results; several of the portal data pages already had comments from members of the public pointing out the uselessness of undocumented data and complaining that similar efforts had been futile. However, additional research through online government document collections turned up **Health Canada** reports relating to the surveys we were looking at. While not as useful as a questionnaire or a data dictionary, this documentation did provide some context and details on the surveys. We also came across references to related surveys and added these to our list of data in need of rescue.

Most importantly, the reports provided a contact email for "questions and concerns" regarding the surveys. Our messages were answered by an initially confused but very helpful employee in the communications and public affairs department of **Health Canada**. After some further correspondence we were put in touch with a health department researcher who agreed to search through old project files and see what was available. We started by requesting material on a 2011 survey, *Knowledge, Perceptions, Awareness and Behaviours Relating to Immunization among First Nations and Inuit*, as this was one of the surveys for which we had no documentation at all.

Our new **Health Canada** contact was happy to respond to our questions and had the technical background to provide useful answers. We soon obtained complete documentation for the immunization survey, as well as for the 2011 *Children's Health and Safety Survey*, another survey from 2009 on drinking water quality, and a major collection of surveys on use of and attitudes toward drugs by young adults. We were particularly gratified to receive data files for some surveys that were already formatted for the statistical software package SPSS. This meant that we could skip the lengthy process of writing command files to read the data and move directly to reviewing the data, checking it against the documentation, and preparing to publish it for research use.

Health Canada does not seem to have a good system in place for keeping track of its older research data. Locating surviving survey files has been a slow and uncertain process, and at this point the agency is relying on our group to discover evidence of surveys that have been conducted, after which our **Health Canada** contact will search for the data. So far the oldest survey we have requested is a historically significant HIV attitudes survey from 2003. Unfortunately, after several searches our contact told us that as far as she could tell no data files for that particular survey seemed to exist. It was too late for rescue. In one happier case, our group managed to locate a survey that staff at **Health Canada** thought lost. Our contact sent us a set of files that contained what appeared to be multiple versions of the third wave of a study on adolescent drug use. After searching through old reports and using technical details, such as the respondent counts, we managed to identify one of the files as a missing fourth wave of the survey.

As of this writing, we also are working with some older data collections, some of which date back to the 1970s. We have not been fortunate enough to locate preformatted files for these surveys, but many of these older files are accompanied by data dictionaries. Our first successful restoration of an older dataset was of the *Alcohol Consumption Survey 1978*. The open data portal included the all-important codebook and data dictionary, and we have been able to locate some of the contextual files that are so valuable to researchers in various library government

continued on page 35

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*

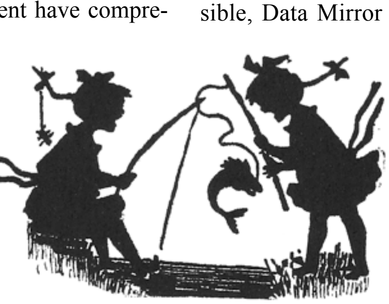
5. **Ontario Council of University Libraries**, ODESI, last modified November 21, 2017. *http://odesi.ca/*

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.