Documenting data rescue

The Ontario Data Community Data Rescue Group
and the Data Rescue & Curation Guide

Overview

- The Data Rescue Group
- The rescue project
- Communication with data producers
- Standards & tools for data rescue
- Example of our work on a specific study
- Data Rescue & Curation Guide
- Future directions

Ontario Data Community

- One of a number of interest groups and communities for member institutions of the Ontario Council of University Libraries (OCUL)
  - OCUL celebrated its 50th anniversary this year
  - Under different names, the Data Community has been a part of OCUL since 2004
- Includes representatives from each of the 21 member universities of OCUL
- Guided the development of <odesi> http://odesi.ca, a data portal available to the OCUL community

Data Rescue Group

- A semi-formal group made up of members of the Ontario Data Community
- Formed in response to the discovery that the Government of Canada had released a number of survey datasets on the Canada Open Data Portal in not particularly user-friendly format
- Mandate
  - Monitor & track data needing rescue
  - Undertake data rescue projects, in collaboration with others in Canada (libraries, government)

The Rescue Project

- The Canada Open Data Portal collection:
  - Survey data on Canadians, particularly government data dealing with health and social issues, is in high demand by our users
  - Includes a number of Health Canada surveys on topics that are not well covered in other Canadian public data sources
    - HIV, sexual behaviour, children's health and safety, First Nations populations, etc.
- These government survey files had considerable historical and research significance and had not previously been available to researchers
- We decided to further explore the collection

Initial inventory

- Canadian data librarians are used to dealing with the well-documented and structured government survey files released by Statistics Canada.
- These were... different.
- They fell into one of the following categories:
  - Data available, data dictionary and sometimes other documentation available
  - Data available, questionnaire, tables or other documentation that *might* help us construct a data dictionary available
  - Data available, no documentation
  - Documentation and/or tables, no data
- Some of the survey documentation made reference to earlier survey rounds and related surveys and we listed these as well

Perceptions of Drinking Water Quality in First Nations Communities and General Population

A series of health-related data sets from various quantitative public opinion research studies.

Publisher - Current Organization Name: Health Canada

Licence: Open Government Licence - Canada

Resources

Resource Name	Resource Type	Format	Language	Links
Perceptions of Drinking Water Quality in First Nations Communities and General Population	Dataset	CSV	English	Access

Geographic Information

Ontario - Ottawa

	A	B	C	D	E	F	G	H
1	Pin	TYPE	LANGI	Q30	Q31	Q24	AQ24	Q1
2	AAAMQYF	All Canadi	English	No -> Thar	#NULL!	7	7	5 Very goo
3	AAAQMJC	All Canadi	English	No -> Thar	#NULL!	5	5	2
4	AABAGYY(	All Canadi	French	No -> Thar	#NULL!	4	4	2
5	AAQMJZB	All Canadi	French	No -> Thar	#NULL!	4	4	5 Very goo
6	AARBQQJ'	All Canadi	English	No -> Thar	#NULL!	4	4	2
7	AARZQZS\	All Canadi	French	No -> Thar	#NULL!	3	3	5 Very goo
8	AASJMMC	On-Reser\	French	Yes	Yes	3	3	5 Very goo
9	AAYSMAC	All Canadi	French	No -> Thar	#NULL!	2	2	5 Very goo
10	ABASMRM	All Canadi	English	No -> Thar	#NULL!	2	2	2
11	ABASQAR	All Canadi	English	No -> Thar	#NULL!	2	2	5 Very goo
12	ABBMQQ\	All Canadi	English	No -> Thar	#NULL!	3	3	4
13	ABBRMSS.	All Canadi	English	No -> Thar	#NULL!	2	2	5 Very goo
14	ABJAZGYG	On-Reser\	French	Yes	Yes	4	4	4
15	ABJAZMR\	On-Reser\	French	Yes	Yes	2	2	5 Very goo
16	ABJJMGM	All Canadi	English	No -> Thar	#NULL!	6	6	4

Working with Data Producers

- We reached out to two of the government bodies depositing data in the open data portal
- Both organizations responded positively
- Library & Archives Canada
  - LAC is slowly migrating their collection of archived datasets
  - Conference calls - discussed their project and ours
  - They were interested in learning from us and considering changes to their practices, though they have limited resources to do so

Working with Data Producers: Health Canada

- We reached out to a communications manager and she assigned an employee to work with us
- They are not sure what they have and locating surviving survey files is a slow and uncertain process
  - At this point they are relying on our group to discover evidence of surveys that have been conducted, after which they will search for the data
  - After several searches, found that the initial waves of one major survey series no longer existed
  - In a happier case, a missing survey wave had been mislabeled; identified through case counts from old reports
- They welcome the opportunity to have us provide stewardship for what is available

Standards & tools

OCUL has excellent infrastructure in place through the <odesi> data portal

Our process

Alcohol Consumption Survey, November 1978

Download data and documentation from Government of Canada Open Data

- Data file, codebook, questionnaire, data dictionary, any other files available
- All original files will be published in the archive with the new files created

Related Publications

- May be listed in the codebook/user guide
- Author agency website
- Libraries - academic or public - Vancouver Public Library, Alcohol Research Group, California
- Google search

Our process - Syntax

Create syntax (command code) file

- ASCII text file and codebook or data dictionary
- Run new syntax file against data file
  - Check new frequencies in new data file against documentation (if you have it)
  - If there is no documentation make a note of this in the User Guide
  - Example:

    The ASCII data file and Codebook were downloaded from the Open Data website on August 10, 2015. Using this Codebook, syntax was created and run against the ASCII data file. The SPSS data file created is provided here. Use with caution. There was no documentation to double check the frequencies created by this data file.

Our process - Metadata

Create metadata file (DDI compliant XML file)

- Nesstar Publisher or similar tool
- Different tools are listed on the DDI Alliance webpage
- Use available documentation to describe dataset as much as possible
  - Include question text, notes and universe for each variable
- Related publications - include reference to any publication about the dataset or study
- Notes - issues in creating and/or describing the dataset

Variable groups

- Group variables based on documentation
- If no grouping in documentation, we use a default set of groups we have developed

Our process - Value add

Value Added Enhancements

- Additional cleanup of the data file
- Declare missing values
- Code open-ended questions
- Restructure multiple questions
- Recoding variables

Example:

Variable - Past seven days how many drinks did you have? (V66_DRINK1, V67_DRINK2, V68_DRINK4, V69_DRINK 8, V70_DRINK12)

Original - coded 0 covering both does not drink and had no drinks

Change made:

Where V65_DRINK (Use alcoholic beverages) = 2 (no), the five variables above were recoded to value of '8' - Does not use alcohol

Also, where V65_DRINK = 3 (refused), the five variables above were recoded to value of 9 (refused)

Our process - Completed datasets

Published to <odesi> with the following files:

- User Guide and Data Dictionary
- Questionnaire
- Original codebook and data file (from producer's website)
- Command code file
- Any additional documents related to the publication (publications, documentation, etc.)
- Names need to distinguish between original and new files

<odesi>

DESCRIPTION | TABULATION | ANALYSIS

- Alcohol Consumption in Canada: A National Study, 1978/11
  - Metadata
    - Study Description
    - Data Files Description
    - Other Documentation
      - Codebook and Data Dictionary - OCUL Ontario Data Community Data Rescue Group Version
      - Questionnaire - Library and Archives Canada Version
      - User Guide and Codebook - Library and Archives Canada Version
      - SPSS Syntax/Command file - OCUL Ontario Data Community Data Rescue Group Version [text]
      - Data file - Library and Archives Canada Version
      - Metadata - Library and Archives Canada Version
  - Variable Description
    - Administration
    - Demographic
    - Household
    - Occupation
    - Income
    - Being home
    - Advertisements about drinking
    - Alcohol use
      - Thoughts about alcohol use: First mention
      - Thoughts about alcohol use: Second mention
      - Discussed alcohol use since seeing DoD ads
      - What have you discussed? First mention
      - What have you discussed? Second mention
      - Use alcoholic beverages
      - Past seven days how many days did you have: one or more drinks
      - Past seven days how many days did you have: two or more drinks?
      - Past seven days how many days did you have: four or more drinks?
      - Past seven days how many days did you have: eight or more drinks?
      - Past seven days how many days did you have: twelve or more drinks?
      - Know one or more people with a drinking problem
      - Relationship to person with a drinking problem: A friend?
      - Relationship to person with a drinking problem: A work associate?
      - Relationship to person with a drinking problem: A family member?
      - Relationship to person with a drinking problem: Other person
      - Two worst consequences of heavy drinking: First mention
      - Two worst consequences of heavy drinking: Second mention
      - Anything you can personally do to encourage others to drink moderately
      - What can personally do to encourage moderation: First mention
      - What can personally do to encourage moderation: Second mention
      - Respondent's estimate of the number of drinks a moderate drinker has over a seven day period
    - Weight

Dataset: Alcohol Consumption in Canada: A National Study, 1978/11

Variable V60_ADSDODTHOUGHTA: Thoughts about alcohol use:

LITERAL QUESTION

What were your specific thoughts about alcohol use? A. First mention

Values	Categories	N	
1	can control drinking	20	12.5%
2	reducing consumption	61	38.1%
3	social-psychological consequences	5	3.1%
4	medical problems	8	5.0%
5	do not drink	19	11.9%
6	drinking and driving	16	10.0%
7	alcoholism fears	7	4.4%
8	other dangers	8	5.0%
9	others mention	16	10.0%
10	no second mention	0	0.0%
0	Not applicable: answered no to V56_ADSDOD or V59_ADSDOD	1908	

SUMMARY STATISTICS

Valid cases	160
Missing cases	1908

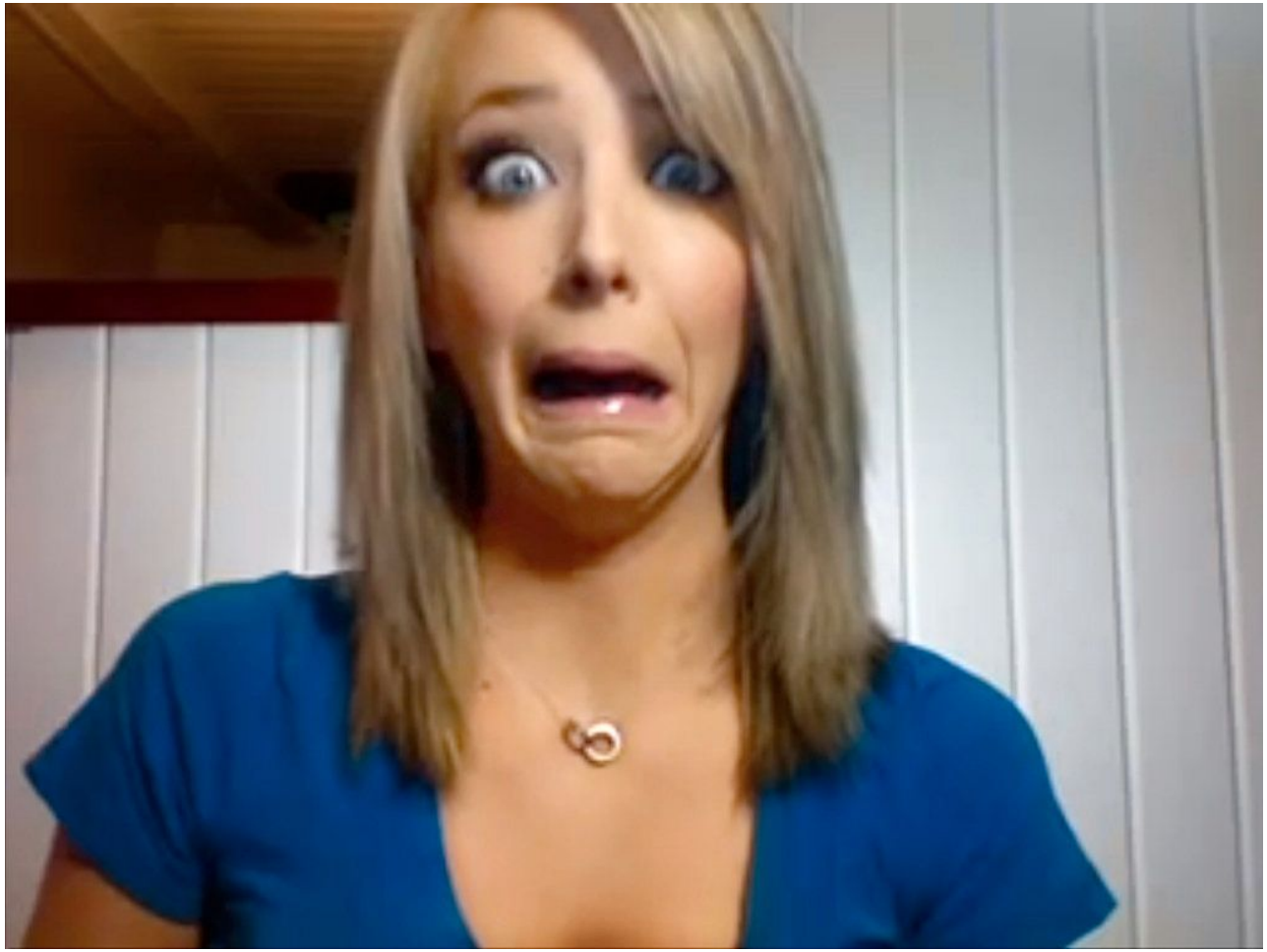
This variable is numeric

Note: In many cases, it is advisable to weight analysis results before reporting them. Correct weighting requires careful conside
select the Weight icon and choose the weight variable to be used. All results need careful interpretation. The data collector

Note: Dans la plupart des cas, il est recommandé de pondérer les résultats d'analyse avant d'en rendre compte. Une pondérati
avant d'appliquer des pondérations. Pour appliquer les pondérations, sélectionner l'icône de poids et choisir la variable de p
la collecte et de la production des données ne peuvent être tenues responsables de l'analyse et de l'interprétation des données.

Roadblocks

- Lack of documentation
  - Weight variable
  - Misidentified skip patterns
  - No frequencies to double check data file created
- Resources of the agency providing the dataset
  - Historical knowledge or people to answer questions
  - Ability/time/interest to provide more than an ascii file and codebook
  - Legacy media - tapes, legacy hardware - time/ability to convert to more useable formats
- Staff time/ability to re-create datasets and properly describe data
  - Finding more documentation
  - Understanding datasets with limited documentation
  - Competing priorities
- Providing datasets with identifying information...



De-identification

- Issue: surveys that contained detailed geographic and other identifying variables along with sensitive information (e.g. HIV status)
- Our contact recognized that this was a problem but had no de-identified version of the survey, or resources for carrying this out
- Undertook to carry out this process ourselves:
  - Dropped census geography, area codes, etc. as well as indirect community identifiers such as community size
  - Regrouped the categories on several variables (education, employment, etc) where the sample size was small
- Checked variables against later survey wave that had been de-identified (apparently by survey firm)
- Verified by cross-tabulating demographics and reviewed separately by a second librarian who made additional changes

Data Rescue & Curation Guide

- Started through the process of documenting the steps used in preparing the sample files
- Expanded into a general guide to data rescue
- Audience
  - Librarians and data curators
- Goals
  - Provide an accessible and hands-on approach to handling data rescue and curation of at-risk data for use in secondary research
  - Provide examples and workflows for addressing common challenges
  - Improve librarians' and data curators' skills in providing access to high quality, well-documented, reusable research data

Data Rescue & Curation Guide

- Topics covered in the guide
  - Background on initiatives working with at-risk datasets
  - Fundamentals: rescuing data files & metadata
    - Case study on retrieving data from different storage media
    - Case study on poorly documented data, reviewing for de-identification
  - Data curation to improve access
    - The OCUL data processing workflow (described above)
    - Links to other documentation on use of DDI & Nesstar that have been created by OCUL for the <odesi> project
  - Further value-added enhancements *(section in progress)*
  - Appendices
    - Expanded data processing workflow - step-by-step
    - Procedures for creating codebooks & data dictionaries
    - Glossary of terms used in the document

Data Rescue & Curation Guide

We’re looking for input & feedback!

http://bit.ly/2pL6xD6

The document is open for commenting

Future directions

- Rescue more Government of Canada files from open data portal and other sources
- Data rescue & curation guide
  - Continue to improve the data rescue guide
  - We are considering creating a briefer version aimed at researchers
- Continue to work with data producers, encouraging them to improve their practices (or at least try to!)
- Monitor the landscape for data rescue projects
- Collaborate with other groups interested in this kind of work

Members of the Data Rescue Project

- 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

questions?