

Data Recovery Projects

Presentation to ACCOLEDS

Vince Gray

December 1, 2016

1. Canadian Aggregated Census Project
2. Discussions between Libraries and Archives Canada and OCUL's Data Rescue Group
3. Statistics Canada public use longitudinal microdata files
4. CIHI's Discharge Abstract Database

Canadian Aggregated Census Project

Does anyone else remember the National Archives' Machine Readable Archives Division? They sponsored a pilot project called CULDAT – the Canadian Union List of Machine Readable Data Files – it was one of the first projects I worked on in 1986. In some ways, I see the Census Aggregate File project as a successor to CULDAT. If you're interested in the early days of data in Canada, I can talk a bit more about that project.

What it is: part 1

- A bilingual inventory and finding tool
 - Born-digital Census of Canada aggregated data products (e.g., USTs, BSTs, Profiles, Nation series, etc. - from 1961 onwards)
 - Who has what?
 - What documentation is available?
 - What format(s) are the files in?
 - What is unavailable?

What it is: part 2

- Make the available files retrievable by any user as open data (without IP restriction)
 - Delivered by OCUL [& other participants???
 - Utilise Dataverse and/or <odesi> and/or ???

Finding tool

- English:
 - <http://mdc.lib.uwo.ca/census/pubsearch.htm>
- French:
 - <http://mdc.lib.uwo.ca/census/pubsearchf.htm>
- Can switch language of search pages or results display

Editing records

- Can add holdings information or add notes
- Bilingual web-based form for data entry
 - <http://mdc.lib.uwo.ca/census/search-edit.htm>
 - Password: **Ed1t0r**

Adding records (new files)

- Same form and password allows editing or adding records
 - <http://mdc.lib.uwo.ca/census/search-edit.htm>, password **Ed1t0r**
- Dropdown lists are provided for controlled vocabulary fields
- *Hint:* for controlled vocabulary fields, paste the English into French fields or vice versa

Current Status of Project

- 2,935 files identified (1961 to 2011 inclusive)
 - Includes all census cycles and the 2011 National Household Survey
 - A slight increase over the 395 of last year ...
- ***Should consideration be given to adding geographic files?***

Need more information?

- Contact:

- Vince Gray (vince@uwo.ca) or
- Alex Cooper (acooper@queens.ca)

Scariest two words I know: LIVE DEMO

[Search](#)

[Edit](#)

Discussions between
Libraries and Archives Canada
and OCUL's Data Rescue Group

Context

Good news:

- LAC releases data through Canada's Open Data Portal (<http://open.canada.ca/en/open-data>)

Bad news:

- Much of the data is poorly documented

Project Findings

- Members of the OCUL Data Rescue Group worked up a file (1978 Alcohol Consumption) to present to LAC
 - Weight variable was “unreliable”
 - Had no frequency counts for variables
 - Values were used for more than one purpose (e.g., 0 = zero OR not applicable)
 - Values were inappropriately coded (e.g., 1-9=1-9, 0=10)
 - Value labels were incorrect
- [<odesi>](#) metadata

Discussions

1. Could LAC use DDI to code files?
2. Could LAC improve the metadata and data (e.g., release comma or tab-delimited ASCII rather than column-delimited)?
3. Would LAC be able to put enhanced files up on opendata site if cleaned up by other groups?

Machine Readable Archives Division also printed catalogues of what they had – we’ve offered them back to LAC if they can’t find print copies. LAC still seems to have many (maybe even most) of these data files.

Outcomes

1. LAC unlikely to adopt DDI
 - LAC has insufficient manpower to dedicate to data (SPSS work is done by a contractor!)
2. Improve data & metadata?
 - Possible, since many files are in SPSS format
 - Little additional work to extract comma-delimited files rather than column-delimited
 - Should be able to use Display Dictionary command to supplement documentation (save as ASCII/HTML)
3. LAC can't load anything but what they have in storage – but might be able to point elsewhere

Machine Readable Archives Division also printed catalogues of what they had – we've offered them back to LAC if they can't find print copies. LAC still seems to have many (maybe even most) of these data files.

Outcomes

4. LAC is interested in engaging with the Canadian data community
 - Realizes could do more with data, and that there are groups out there which could help
 - OCUL is trying to establish a “permanent” contact in LAC with whom to talk about data

More information:

- Contact the head of the OCUL Data Rescue Group, Leanne Trimble ltrimble@utoronto.ca

Machine Readable Archives Division also printed catalogues of what they had – we’ve offered them back to LAC if they can’t find print copies. LAC still seems to have many (maybe even most) of these data files.

Statistics Canada public use
longitudinal microdata files

Well-known Problem

Statistics Canada does not distribute longitudinal data file in PUMF form due to risk of breaching anonymity

- Survey of Labour and Income Dynamics has only cross-sectional PUMFs; nothing from LSIC, YITS, etc., all longitudinal data collection cancelled

BUT

- Before there was SLID or the Canadian Income Survey there was the Labour Market Activity Survey (LMAS)

Now, I'm going back in time to before a number of you were born, I'm afraid ...

<http://equinox.uwo.ca/drl-only/catalogue/scanned-cd-covers/011248001.pdf>

INSTALLATION	INSTALLATION
A. Hardware Requirements Before you install or run LMSSDA, make sure you have the following: 1. An IBM compatible PC, running MS-DOS, with disk memory which is also running DOS/386, version 3.00. 2. A CD-ROM drive connected to the PC. 3. Monitor CD-ROM extensions installed on the PC. 4. A variable disc connected to the PC. You do not need a hard disc, but 20 megabytes or larger of free space is recommended as your random working space. B. Accessing the LMAS System To access the LMAS CD-ROM system do the following: 1. Place the LMAS CD-ROM disk for the year and file type desired in your CD-ROM drive. Year and file type are indicated on each individual compact disk. 2. At the DOS prompt change the default drive to the CD-ROM location. 3. Identify the CD-ROM disk letter, e.g. D: 4. Type LMAS and press <Enter>.	A. Matériel requis Avant d'installer ou de mettre en marche le logiciel, assurez-vous d'avoir le matériel suivant : 1. Un ordinateur personnel compatible IBM, fonctionnant sous l'implémentation MS-DOS, d'une mémoire de disque, le plus d'un plate d'extension MS-DOS. 2. Un lecteur de disque CD-ROM relié à l'ordinateur personnel. 3. Le module d'extension de Microsoft pour disque CD-ROM installé dans l'ordinateur. 4. Un disque variable relié à l'ordinateur personnel. Vous n'avez pas besoin d'un disque dur, mais il est recommandé d'avoir un espace libre de 20 mégaoctets ou plus comme zone de travail pour l'installation des données. B. Accès au système de l'ÉIA Pour obtenir accès au disque CD-ROM de l'ÉIA, procédez comme suit : 1. Placez le disque CD-ROM contenant les données de l'ÉIA pour l'année et le genre de fichier désirés dans votre lecteur de disque CD-ROM. L'année et le genre de fichier sont indiqués sur chaque disque compact. 2. Au message DOS, changez l'unité exploitée pour désigner l'ordinateur vers le lecteur de disque CD-ROM. 3. Identifiez la lettre désignant le lecteur de disque CD-ROM, p. ex. «D». 4. Tapez LMAS et appuyez sur <Entrée>.

Catalogue MDP-3853b

Labour Market Activity Survey
1988, 1989, 1990
Individuals and Jobs - B

Enquête sur l'activité
1988, 1989, 1990
Individus et emplois - B




 Statistics Canada
 011248001
 

What was LMAS?

- Successor to Annual Work Patterns Survey (1978-81, 1983-85)
- The objectives of the Labour Market Activity Survey were:
 - to measure the frequency and number of job changes in the Canadian labour market over a one to three year period;
 - to provide information on the characteristics of jobs held during this period;
 - to develop socio-economic/demographic profiles on groups of people who would benefit from Employment and Immigration Canada (EIC) programs; and
 - to identify participants of specific EIC programs.

Topics include: periods of employment and unemployment; job search activities; unionization; pension plan coverage; wage rates; work schedules; and financial compensation during absences from work. Data is presented in two file universes: individuals and jobs, for the years 1986 to 1990. Two longitudinal files 1986-1987 and 1988-1989-1990 are also provided.

This is copied from the sleeve above ...

Software specifications

- Used Folio Views software (developed for U.S. National Longitudinal Survey data releases?)
- Software could extract raw ASCII data, write SPSS code to read files, and create codebook
- Limited in size of any single extraction to ~500 variables
- Does not work with Window XP or newer operating systems

Fortunately, I have refused to

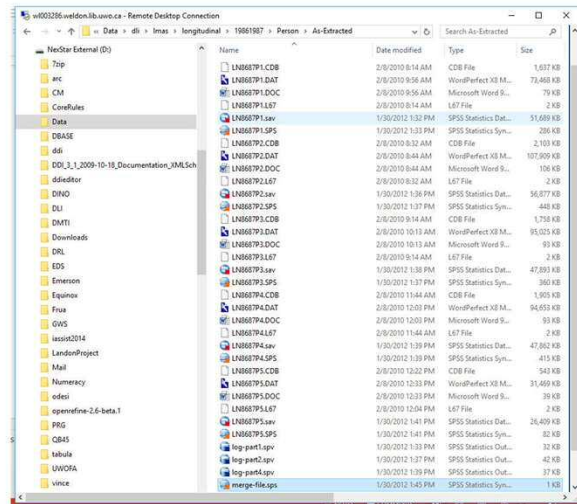
LMAS data release

<http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3853&lang=en&db=imdb&adm=8&dis=2>

- Released on three cds in 1996
 - Disk A:
 - 1986 and 1987 job and person cross-sections
 - 1986-1987 job and person longitudinal
 - Disk B:
 - 1988, 1989 and 1990 job and person cross-sections
 - Disk C:
 - 1988-1990 job and person longitudinal

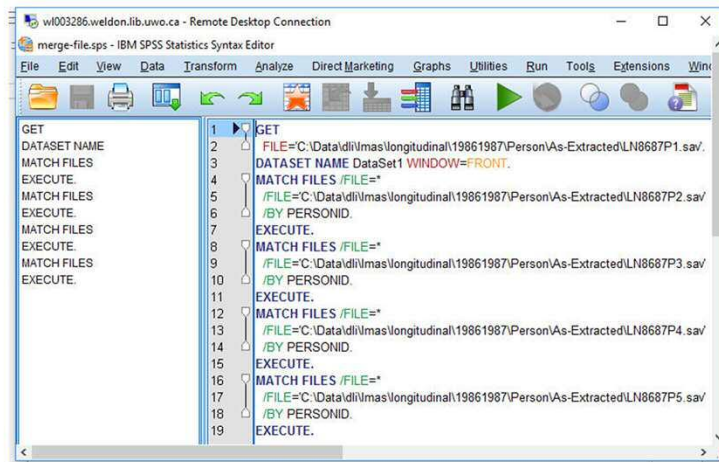
I think that I remember the original software correctly

Multiple file extractions



Each file required multiple separate extractions (five for 86-87 job file), then merging those extractions together in SPSS

Merge files with SPSS: (e.g., 1,635 variables in 1988-90 job file)



The screenshot shows the IBM SPSS Statistics Syntax Editor window. The title bar indicates a remote desktop connection to 'w1003286.weldon.lib.uwo.ca'. The menu bar includes File, Edit, View, Data, Transform, Analyze, Direct Marketing, Graphs, Utilities, Run, Tools, Extensions, and Window. The toolbar contains icons for opening files, saving, printing, and running syntax. The main text area displays a syntax script for merging five data files. The script is as follows:

```
1 GET  
2 FILE=C:\Data\dl\lmas\longitudinal\19861987\Person\As-Extracted\LN8687P1.sav/.  
3 DATASET NAME DataSet1 WINDOW=FRONT.  
4 MATCH FILES /FILE=*  
5 /FILE=C:\Data\dl\lmas\longitudinal\19861987\Person\As-Extracted\LN8687P2.sav/  
6 /BY PERSONID.  
7 EXECUTE.  
8 MATCH FILES /FILE=*  
9 /FILE=C:\Data\dl\lmas\longitudinal\19861987\Person\As-Extracted\LN8687P3.sav/  
10 /BY PERSONID.  
11 EXECUTE.  
12 MATCH FILES /FILE=*  
13 /FILE=C:\Data\dl\lmas\longitudinal\19861987\Person\As-Extracted\LN8687P4.sav/  
14 /BY PERSONID.  
15 EXECUTE.  
16 MATCH FILES /FILE=*  
17 /FILE=C:\Data\dl\lmas\longitudinal\19861987\Person\As-Extracted\LN8687P5.sav/  
18 /BY PERSONID.  
19 EXECUTE.
```

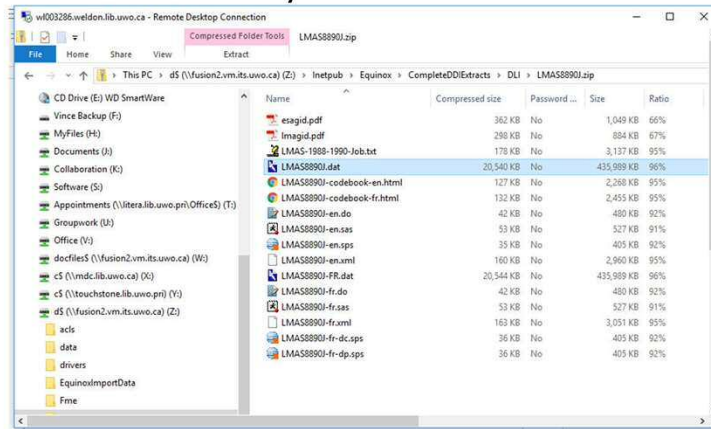
[illegible]

Uses of files?

- Thirty year old data isn't likely to excite current researchers
- **But** it is training material for learning how to do some forms of longitudinal analysis
 - Good sample size
 - Up to 156 repetitions of weekly measures (e.g., labour market participation, employment by job)
 - Up to 15 jobs recorded over 3 years, plus cycle of moving between jobs

Availability

- Some (all?) of these files should be available on the University of Alberta's DLI Mirror Site



I wasn't able to check the mirror site from home on Sunday, but the Equinox files were to be transferred to the mirror

Missing LMAS files?

- Let me know if there are missing LMAS files
 - I'll do what I can to get them to you
- I'm strongly considering applying for a sabbatical to recover data files from old media
 - If there are other files you know of that need recovering, let me know ...

CIHI's Discharge Abstract Database

So, I'm broadening the definition of "rescue" to include a current file that we get regularly (but that is sent in an appalling fashion)

DAD Problems

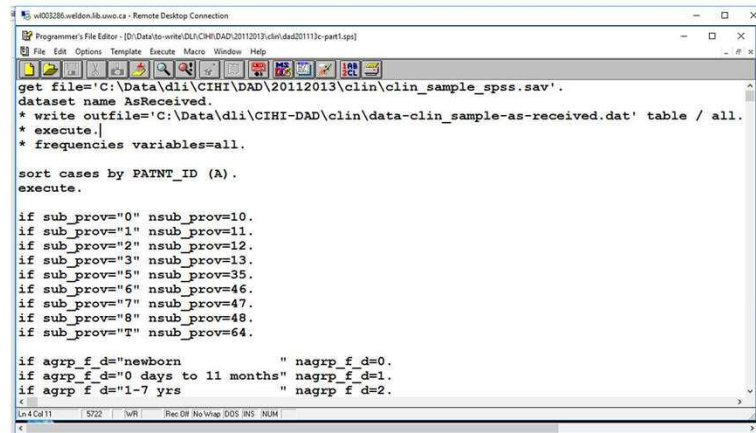
- Quebec and BC are not included
 - We can't do *anything* about that except lobby 😞
- The incident file (particularly) is released *apparently untreated* by CIHI – released as it is stored and used at CIHI
 - For storage purposes(?), punctuation is stripped from ICD-10 and CCI codes
 - Strings are used instead of codes for what should be ordinal variables
 - No summary variables are provided

Lobbying is probably best done at the province level – it's not a CIHI decision to not release the data

Having talked about the problems with the DAD file with CIHI, they seem to have no interest in improving the situation

Three part solution:

1) reformat variables with SPSS



```
get file='C:\Data\dli\CIHI\DAD\20112013\clin\clin_sample_spss.sav'.
dataset name AsReceived.
* write outfile='C:\Data\dli\CIHI-DAD\clin\data-clin_sample-as-received.dat' table / all.
* execute.
* frequencies variables=all.

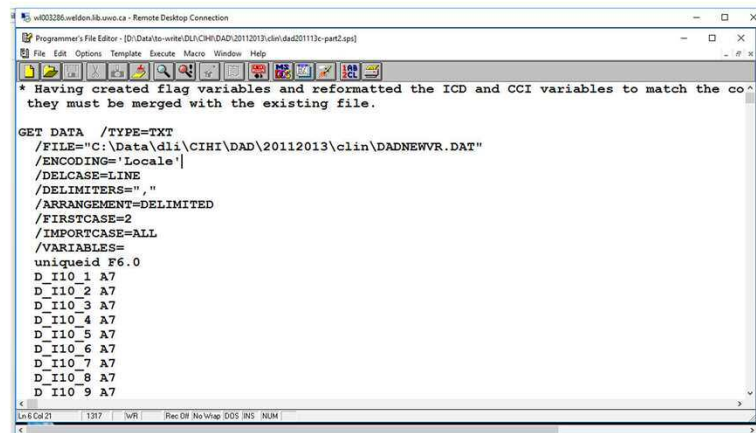
sort cases by PATNT_ID (A).
execute.

if sub_prov="0" nsub_prov=10.
if sub_prov="1" nsub_prov=11.
if sub_prov="2" nsub_prov=12.
if sub_prov="3" nsub_prov=13.
if sub_prov="5" nsub_prov=35.
if sub_prov="6" nsub_prov=46.
if sub_prov="7" nsub_prov=47.
if sub_prov="8" nsub_prov=48.
if sub_prov="T" nsub_prov=64.

if agrp_f_d="newborn" ngrp_f_d=0.
if agrp_f_d="0 days to 11 months" ngrp_f_d=1.
if agrp_f_d="1-7 yrs" ngrp_f_d=2.
```

Three part solution:

2) programming to construct summary vars



The screenshot shows a SAS Programmer's File Editor window titled "Programmer's File Editor - [D:\Data\to-write\DLI\CHH\HAD\20112013\clin\data201113c-part2.sps]". The window contains the following SAS code:

```
* Having created flag variables and reformatted the ICD and CCI variables to match the co  
they must be merged with the existing file.  
  
GET DATA /TYPE=TEXT  
/FILE="C:\Data\dl1\CIHI\HAD\20112013\clin\HADNEWVR.DAT"  
/ENCODING='Locale'  
/DELCASE=LINE  
/DELIMITERS=" "  
/ARRANGEMENT=DELIMITED  
/FIRSTCASE=2  
/IMPORTCASE=ALL  
/VARIABLES=  
uniqueid F6.0  
D_I10_1 A7  
D_I10_2 A7  
D_I10_3 A7  
D_I10_4 A7  
D_I10_5 A7  
D_I10_6 A7  
D_I10_7 A7  
D_I10_8 A7  
D_I10_9 A7
```

The status bar at the bottom indicates "Ln 6 Col 21", "1337", "tWR", "Rec Off No Whap ODS INS NUM".

Three part solution:
3) Merge new & old data with SPSS

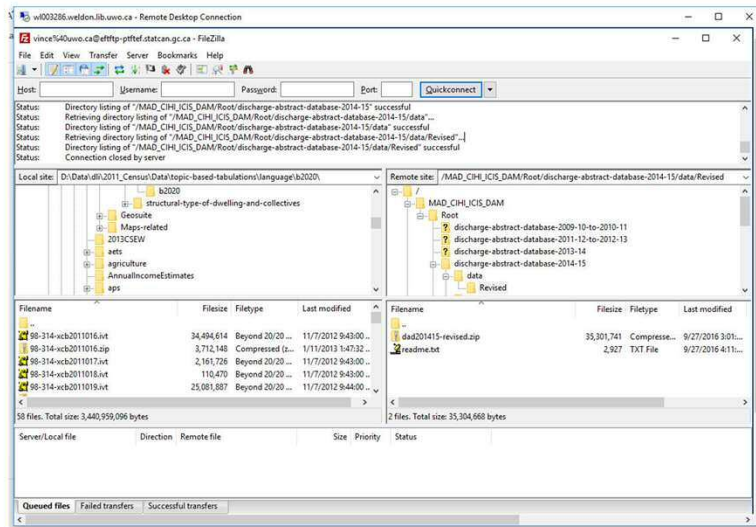
[illegible]

File uses

- Estimate severity of problem (i.e., number of people affected by disease, retrieving particular treatment, etc.)
 - Not perfect, because doesn't cover entire population
 - But, as a preliminary request (e.g., estimate for a grant application), it may be good enough

The Associate Dean for Research at our medical school was very interested in these data, because researchers are always being asked to show the significance of what they are proposing – this would allow them to include SOMETHING

Availability



On DLI EFT site

Ongoing availability issues

- My retirement is looming / approaching
 - I don't anticipate working past 2018
 - If CIHI releases more data, it likely will also need treatment to make it useable
 - Someone needs to step up to carry on reformatting the CIHI files, because CIHI is unlikely to change their practices – **volunteers?**