DLI + RDC Partnership

Vivek Jadon

May 22, 2019

Outline

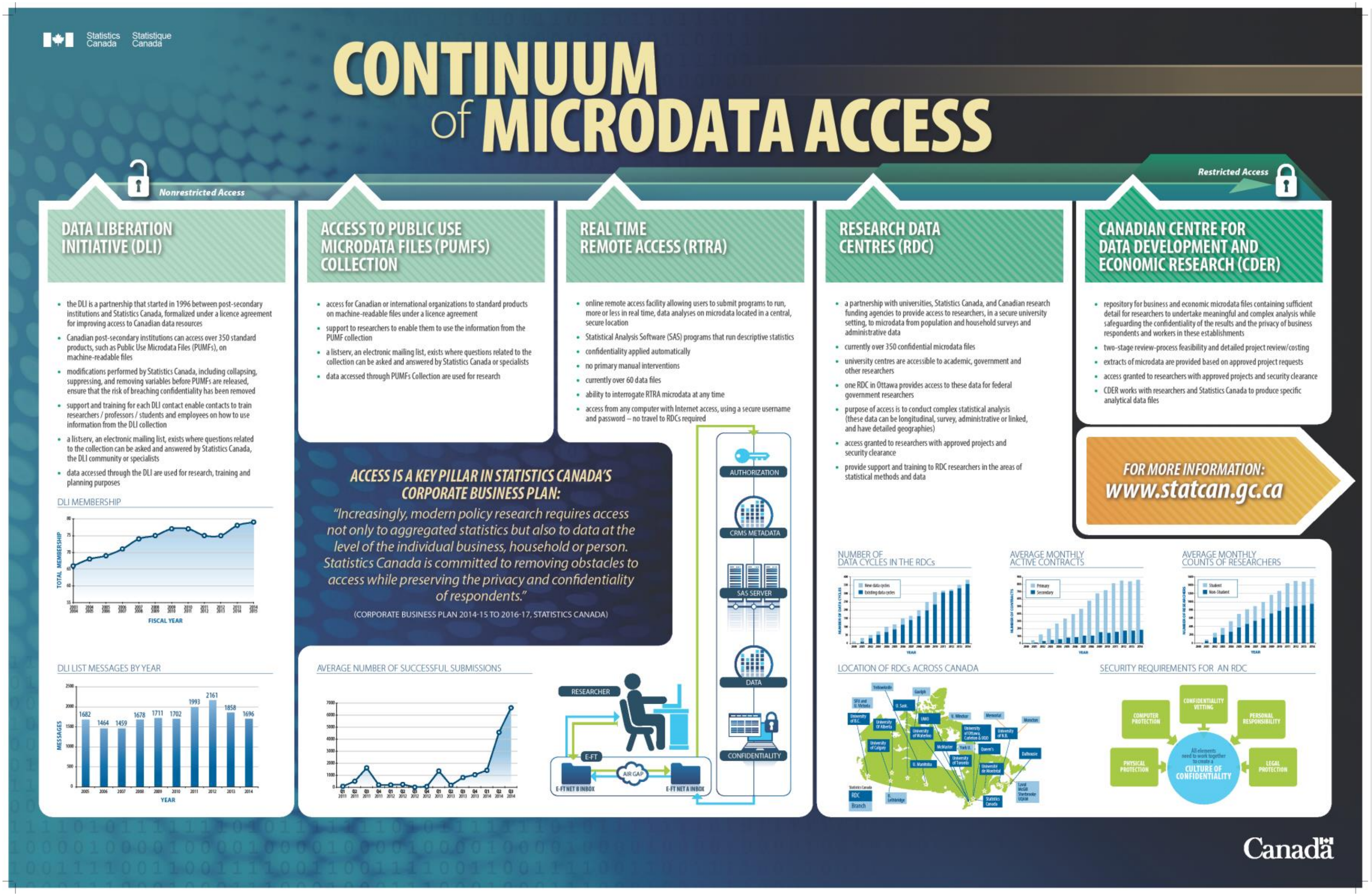
- Continuum of “Microdata” Access at Statistics Canada
- DLI Program and Collection
- DLI and RDC Programs: Common goals
- DLI and RDC programs: Common Differences
- Partnership between DLI and RDC program and next steps

Continuum of Access

- Statistics Canada provides access to its statistical information through a variety of services and initiatives that function as dissemination channels.

- There are three characteristics that make up this continuum:
  - **Cost:** which runs from free to expensive
  - **Restrictions or conditions:** which run from open or no restrictions to very restricted
  - **Type of Information:** which runs from statistics to data.

Continuum of “Microdata” Data Access



Statistics Canada Statistique Canada
Nonrestricted Access
Restricted Access
DLI MEMBERSHIP
TOTAL MEMBERSHIP
FISCAL YEAR
DLI LIST MESSAGES BY YEAR
MESSAGES
1682
1464
1459
1678
1711
1702
1993
2161
1858
1696
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
YEAR
AVERAGE NUMBER OF SUCCESSFUL SUBMISSIONS
RESEARCHER
E-FT
AIR GAP
E-FT NET B INBOX
E-FT NET A INBOX
AUTHORIZATION
CRMS METADATA
SAS SERVER
DATA
CONFIDENTIALITY
NUMBER OF DATA CYCLES IN THE RDCs
New data cycles
Existing data cycles
AVERAGE MONTHLY ACTIVE CONTRACTS
Primary
Secondary
AVERAGE MONTHLY COUNTS OF RESEARCHERS
Student
Non-Student
YEAR
LOCATION OF RDCs ACROSS CANADA
SECURITY REQUIREMENTS FOR AN RDC
COMPUTER PROTECTION
CONFIDENTIALITY VETTING
PERSONAL RESPONSIBILITY
PHYSICAL PROTECTION
LEGAL PROTECTION
All elements need to work together to create a CULTURE OF CONFIDENTIALITY
Canada

CONTINUUM of MICRODATA ACCESS

DATA LIBERATION INITIATIVE (DLI)

- the DLI is a partnership that started in 1996 between post-secondary institutions and Statistics Canada, formalized under a licence agreement for improving access to Canadian data resources
- Canadian post-secondary institutions can access over 350 standard products, such as Public Use Microdata Files (PUMFs), on machine-readable files
- modifications performed by Statistics Canada, including collapsing, suppressing, and removing variables before PUMFs are released, ensure that the risk of breaching confidentiality has been removed
- support and training for each DLI contact enable contacts to train researchers / professors / students and employees on how to use information from the DLI collection
- a listserv, an electronic mailing list, exists where questions related to the collection can be asked and answered by Statistics Canada, the DLI community or specialists
- data accessed through the DLI are used for research, training and planning purposes

ACCESS TO PUBLIC USE MICRODATA FILES (PUMFS) COLLECTION

- access for Canadian or international organizations to standard products on machine-readable files under a licence agreement
- support to researchers to enable them to use the information from the PUMF collection
- a listserv, an electronic mailing list, exists where questions related to the collection can be asked and answered by Statistics Canada or specialists
- data accessed through PUMFs Collection are used for research

REAL TIME REMOTE ACCESS (RTRA)

- online remote access facility allowing users to submit programs to run, more or less in real time, data analyses on microdata located in a central, secure location
- Statistical Analysis Software (SAS) programs that run descriptive statistics
- confidentiality applied automatically
- no primary manual interventions
- currently over 60 data files
- ability to interrogate RTRA microdata at any time
- access from any computer with Internet access, using a secure username and password – no travel to RDCs required

RESEARCH DATA CENTRES (RDC)

- a partnership with universities, Statistics Canada, and Canadian research funding agencies to provide access to researchers, in a secure university setting, to microdata from population and household surveys and administrative data
- currently over 350 confidential microdata files
- university centres are accessible to academic, government and other researchers
- one RDC in Ottawa provides access to these data for federal government researchers
- purpose of access is to conduct complex statistical analysis (these data can be longitudinal, survey, administrative or linked, and have detailed geographies)
- access granted to researchers with approved projects and security clearance
- provide support and training to RDC researchers in the areas of statistical methods and data

CANADIAN CENTRE FOR DATA DEVELOPMENT AND ECONOMIC RESEARCH (CDER)

- repository for business and economic microdata files containing sufficient detail for researchers to undertake meaningful and complex analysis while safeguarding the confidentiality of the results and the privacy of business respondents and workers in these establishments
- two-stage review-process feasibility and detailed project review/costing
- extracts of microdata are provided based on approved project requests
- access granted to researchers with approved projects and security clearance
- CDER works with researchers and Statistics Canada to produce specific analytical data files

ACCESS IS A KEY PILLAR IN STATISTICS CANADA'S CORPORATE BUSINESS PLAN:

> *“Increasingly, modern policy research requires access not only to aggregated statistics but also to data at the level of the individual business, household or person. Statistics Canada is committed to removing obstacles to access while preserving the privacy and confidentiality of respondents.”*

(CORPORATE BUSINESS PLAN 2014-15 TO 2016-17, STATISTICS CANADA)

FOR MORE INFORMATION:
www.statcan.gc.ca

Data Liberation Initiative (DLI) Program

- DLI provides access to Stat Can’s Standard products, Databases, 350 Public use Microdata sets and Geographic information files.
  - Main focus of DLI collection is on Socio-Economic data: Health, Education/Literacy, Labor Market, Income, Travel, Justice, Census of Population etc.
  - Databases such as the Small Area Business and Labour Database, Inter-Corporate Ownership, Financial Performance Indicators, Trade data etc.
  - An enhanced line of Census products
  - Aggregated data on subject such as Justice and Education
  - All standard geographic files and databases

Data Liberation Initiative (DLI) Program

- Metadata from DLI surveys are marked-up in DDI/XML format for discoverability at the variable level from Stat Can **Nesstar site** and **Odesi Data Portal**.

- DLI members have support through a very active listserv.

- Currently 77 subscribing institutions -- McMaster University Library is part of Stat Can DLI program.

DLI Collection: PUMF

Public Use Microdata File (PUMF)

- Each Public Use Microdata File is based on a corresponding master data file. The modifications performed by Statistics Canada before the PUMF is released ensure that the risk of breaching confidentiality has been removed. Since the results of any analysis performed do not have to be scrutinized before they are released, the file is considered "Public".

- Modifications made to the Master files for conversion to PUMFs may include: **collapsing of variables** (e.g., age groups instead of individual years of age); **collapsing variables into one variable** (e.g., multiple language questions collapsed into one language variable for analysis); **suppressing variables** (although the variable is part of the master file, it will not show up in the public file); and **removing outliers** (removing cases that are extremes - often used with income).

- By using these techniques to anonymise the files, combining variables will not result in the user identifying a respondent from any given survey.
- **Benefits**
  - Free
  - Very few restrictions on access & use of the data
  - No approval process to access the data
- **Limitations**
  - Content is limited (screened and grouped for confidentiality)
  - Not all surveys have a PUMF
  - PUMFs are cross-sectional, i.e., represent data collected at one point in time

DLI Collection: Master file

Master file

- contains the full sample of respondents, not a sub-set of them
- includes additional categories in variables; more detailed information
- allows research on lower levels of geography
- provides discrete values for certain variables, such as age or body weight, instead of categories (as found in PUMFs)
- may offer other concepts that are not available in PUMFs

Moreover, master files contain derived variables and bootstrap weights

- RDC access is useful when PUMF does not exist or provide adequate level of details for quality research or when longitudinal data linkage is required

DLI Collection: Synthetic Files

- **Synthetic Files**
  - These microdata do not contain actual “real” cases but contain “**pseudo-cases**” that provide aggregate results close to the “real” cases
  - These files have been prepared to create analysis runs with the master file without possibly disclosing or identifying any of the cases
  - The results are NOT to be reported, but are strictly to be used to prepare analysis of master files
  - Usually associated with longitudinal files, e.g. NLSCY, NPHS etc.

DLI and RDC Programs:

Common Goals: Access

- A goal of DLI is to create affordable and equitable access to “standard data products” for post-secondary institutions.

- A goal of the RDC program is to provide access to “confidential data” for approved research projects using procedures allowed under the conditions of the Statistics Act.

Common Goals: Knowledge Creation

DLI

- A goal of DLI is to facilitate the creation of new knowledge by providing access to STC standard data products for research.
- While useful in evaluating DLI, knowledge products are not mandatory for the continued operation of DLI, although outcomes have shown to be important in maintaining participation by author divisions in DLI.

RDC

- A goal of the RDC program is to support the creation of new evidence or knowledge relevant to policymaking by providing access to STC confidential data for approved research projects.
- A research outcome is required for every approved project, which is stipulated in the contract signed between STC and a project PI.

Common Goals: Training

DLI

- A goal of DLI is to support the training of students in quantitative reasoning through the use of real Canadian data, which is seen as an important step in building a data culture in Canada.

RDC

- A goal of the RDC program is to support the training of students in quantitative methods developed for the analysis of longitudinal surveys.

DLI and RDC Programs:

Access Differences

DLI

- Access is to "standard data products", which have been created for public dissemination.

RDC

- Access is to confidential data, which are protected under the Statistics Act and are only available to STC employees or "deemed employees" who have been given approval to use the data. These data products have not been created for dissemination.

DLI and RDC Programs:

Access Differences

DLI

- Access is determined by a paid institutional membership and a license that defines approved users and uses of these data products.

RDC

- Access is determined by a peer-approval process for projects, a security clearance prior to establishing “deemed employee” status, and a contract.

DLI and RDC Partnership

- **Building relations between RDC Analysts and DLI Contacts**
  - RDC analysts and DLI contacts consult with each other about making proper referrals – e.g. RDC analysts refer all PUMF questions to Library Data Service (LDS)
  - Invite and involve RDC Analysts in DLI and RDC training.
    - RDC Analysts have participated in DLI workshop
  - Make joint presentations and offer seminars to promote collection and services on campus
  - Provide links to RDC website from LDS webpages and vice versa

DLI and RDC Partnership

Next Steps

- Continue to provide referrals where appropriate
- Continue promotion of collection and services as a continuum – Not discrete silos
- Joint publicity of events DLI training, RDC forum etc.
- Collaborate on training and share training resources where appropriate.
- Educate each other on any new developments taking place in our areas.

Contact Information

Data Liberation Initiative (DLI)

- Mills Memorial Library, Room L104/C
- 905 525-9140 Ext. 23848
- vivek@mcmaster.ca
- Hours of Service:

  9:30 am – 1:00 pm

  2:00 pm – 5:00 pm
- library.mcmaster.ca/data/