Access to Statistics Canada's Microdata

Donna Dosman and Susan Stobert

Statistics Canada, donna.dosman@canada.ca
Statistics Canada, susan.stobert@canada.ca

1. Introduction

In Canada, providing national statistics is a federal responsibility. As Canada's central statistical office, Statistics Canada is legislated to serve this function for the whole of Canada. Under the Statistics Act, Statistics Canada is required to "collect, compile, analyze, abstract and publish statistical information relating to the commercial, industrial, financial, social, economic and general activities and conditions of the people of Canada."[i]

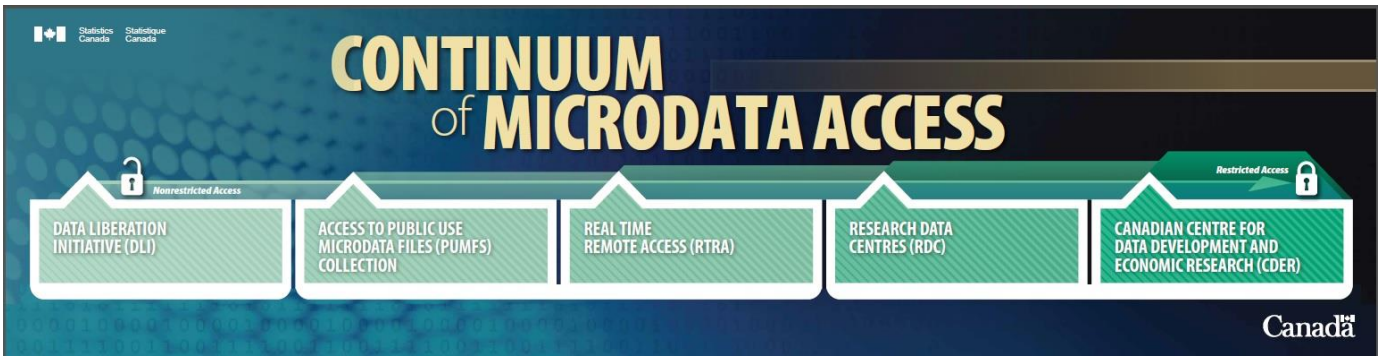
Access to trusted statistical information underpins democratic societies, as it supports evidence-based decision-making in the public and private sectors, and informs debate on public policy issues. One of Statistics Canada's main objectives is to provide statistical information on Canada's economic and social structure to be used to develop and evaluate public policies and programs and improve public and private decision-making for the benefit of all Canadians. In particular, Statistics Canada is committed to ensuring researchers and analysts have access to microdata while preserving the privacy and confidentiality of individuals.[ii]

Statistics Canada's commitment to maintaining the confidentiality of the information obtained from the Canadian public is enshrined in the Statistics Act and the Agency's various policies and practices related to data collection, analysis and dissemination activities as well as the Privacy Act. All information provided to Statistics Canada through surveys, the Census or any other sources is confidential. To this end, disclosure control measures have been developed for each of the access programs examined here.

Statistics Canada provides many avenues to data access along a continuum of access – from data tables on the web site to allowing researchers to work directly with the data files. Statistics Canada currently has four different modes of access for microdata which have varying degrees of flexibility, utilizes a number of methods and procedures to protect the privacy and confidentiality of individuals, serve different user communities and in some cases provide access to different datasets. The diagram below illustrates these four modes of access from the most flexible mode access to microdata being the public use microdata on the left to a more restrictive mode of access which allows direct access to de-identified social microdata [1] in secure centres and on the

[1] "De-identification" is the general term for the process of removing personal information from a record or data set. De-identification protects the privacy of individuals because once de- identified, a data set is considered to no longer contain personal information.

very right of the continuum the most restrictive mode of access to business data which allows indirect access in a secure centre.



2. Access to Statistics Canada's Social Microdata

Three modes of access have been developed for social microdata and these include: Public Use Microdata Files, Real Time Remote Access and the Research Data Centre Program.

2.1. Public Use Microdata Files

2.1.1 Program description

Public Use Microdata Files (PUMFs)[iii] are microdata which have been modified to ensure that no one individual can be identified from a combination of variables on the file. Statistics Canada has been creating PUMFs since the 1970's and these datasets are readily available directly from the division which created the data or through a subscription-based service which provides access to the collection of PUMFs as well as support and training service. Universities are the primary subscribers to the service, the Data Liberation Initiative, with professors and students being the primary users of these types of files for course work, research papers, assignments, exploratory analysis or preparatory work for analysis in the Research Data Centres. The files are also used by small businesses and local governments since they are easily accessible and free of charge. The PUMF collection is updated continually with the size of the collection now consisting of about 450 cycles of data.

Access to the data requires that institutions subscribing to the PUMF services sign an institutional licence agreement which outlines the institutional responsibilities as a distributor of the PUMFs to their staff and students. The key responsibility is to ensure that researchers abide by the terms of licence; there are two clauses relevant to the protection of the confidentiality: do not share the data and do not link or merge the data with other databases in an attempt to re-identify respondents. If any of the terms and conditions are breached, the institution must notify Statistics Canada and the agreement may be terminated.

2.1.2 Protecting confidentiality in PUMFs

The PUMFs are anonymized microdata files of a sample of units, usually from a household survey. The sampling frame provides some measure of protection but the file is further anonymized utilizing a number of strategies depending on the variable. These include:

- Suppression of identifying variables
- Regrouping of some continuous or variables with detailed coding structures into grouped categorical variables
- Data suppression where regrouping is not sufficient and perturbation,
- Creation of special derived variables by combining the responses of two or more variables
- Reduction in the geographic detail
- Examination of the distribution of weights – low weights, geographic information implied by the weights
- Limit design and related information.

Once a project team has created a public use file the output is reviewed by a senior management committee to ensure that adequate steps and measures were taken to protect confidentiality.

2.1.3 Limitations of PUMFs

The primary goal of the PUMFs is to make microdata available, that is liberating access, while ensuring the confidentiality of respondents. Protecting against identification implies decreasing the analytical potential of a microdata file. In particular, low level of geography, small sampling weights and detailed categories of variables are often suppressed or collapsed limiting the usefulness of these files. Achieving the optimum balance of protection and usability is indeed labor intensive. There is, however, specific needs for PUMFs as opposed to other modes of access, for example, in classroom teaching and for international researchers.

2.1.4 Future Developments in PUMF Access

Given the labor intensive process required for the creation of PUMFs and advancements in methods of protection against identification, further research is required to find means of automating some of the protection procedures to decrease burden on microdata owners. In the near future Statistics Canada will also explore how best to streamline the licensing and access procedures to improve the accessibility of these files.

2.2 Real Time Remote Access

The Real Time Remote Access (RTRA) system[iv], introduced in 2010, is an on-line remote access tool allowing users to run SAS programs, in real-time, against microdata sets located in a central and secure location. A full range of descriptive statistics is available through the RTRA such as frequencies, means, medians, percentiles, proportions, ratios, and shares[2].

Researchers using the RTRA system do not gain direct access to the microdata and cannot view the content of microdata files. As such, a security clearance, the swearing of an oath and other legal contracts with Statistics Canada are not necessary.

The RTRA is very flexible as it can be accessed from any computer with an internet connection at any time of the day and from any location; users need a secure username and password and

[2] The share procedure show the relationship between tow variable. For example, this procedure can be used to calculate the share of income to spending by sex.

output is available in an electronic secure vault via the internet. Confidentiality rules are applied automatically to the outputs before they are released to the researcher.

Researchers must be associated with a government department, non-profit organization or an academic institution and each user needs to apply for access. The most frequent users are government researchers. Once their application is approved the researchers have access to any data set in the RTRA collection. The RTRA data collection includes a range of survey and administrative datasets and the collection size as of July 2017 was 238 datasets with updates being added every quarter. The goal is to grow the collection to include a wide range of data, however, two of the key priorities for the RTRA collection is to include data sets not in the PUMF collection or data sets in the PUMF collection but where extensive confidentialization has occurred.

2.2.1 Protecting confidentiality in the RTRA

To protect confidentiality for the microdata a number of strategies have been put in place which include the design of the portal, restrictions on programming and confidentiality vetting. The strategies include:

- Researchers cannot see the data
- Some SAS codes have been restricted to ensure that the researcher cannot list microdata in their output
- The number of submissions per day are limited
- Sensitive variables are removed and there are limited geographical variables on the files
- Additive and controlled rounding is applied to all frequencies. This method of controlled rounding does not affect the accuracy of the data
- The rounding base is set for each dataset using information on the weight distribution, confidentiality vetting rules applied in the Research Data Centres (RDCs) and whether there are other rounding practices used for the dataset
- Controlled rounding is applied independently on each cell, including subtotals and totals
- Only weighted data can be produced
- Minimum threshold suppression rules do not apply to RTRA.

2.2.2 Limitations of the RTRA Program

Many Canadian researchers are more familiar with statistical programs other than SAS, as a consequence the fact that the system only accepts SAS programs is a limitation and barrier of access to the system. Two other limitations of the current version of the RTRA system are: first, the removal of sensitive variables, especially geography; and, second, the limited functionality of the system. It has been noted that the range of descriptive statistics are limited. Specifically, the lack of significance testing and no modeling have been noted by researchers as major limitations.

2.2.3 Future Development of RTRA

In the next few years several of the above noted limitations will be addressed. Specifically, a new graphical interface will be introduced which will eliminate the need to use SAS to run the program which should make the system accessible to more users. Expansion of the functionalities of the RTRA system to include a number of new statistics, including level change,

percent change, significance tests and percent distribution shall improve the users' experience and allow an expanded use of this mode of access.

2.3 Research Data Centres

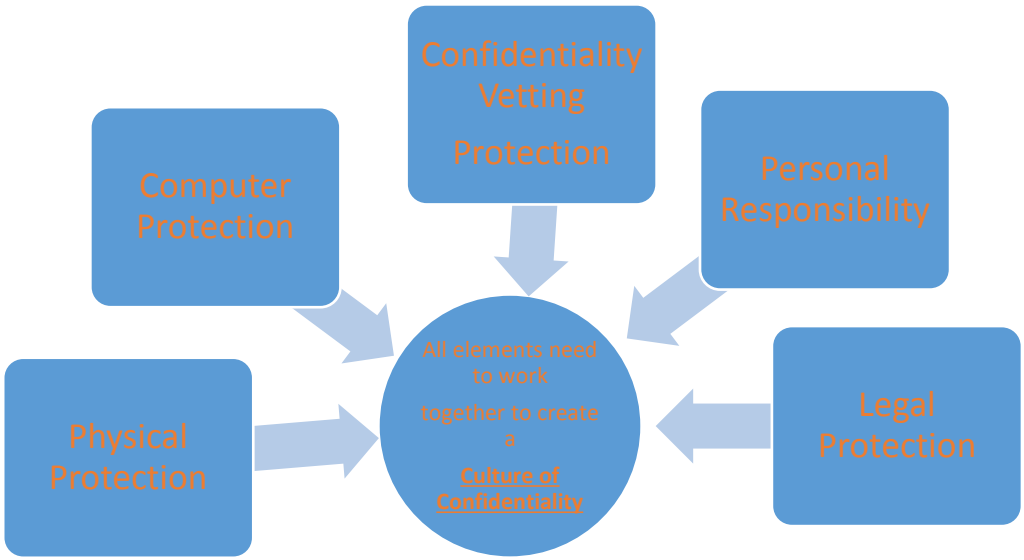
There are 30 academic Research Data Centres (RDCs)[v] across Canada in university settings and 2 Federal RDCs in Ottawa. The Federal RDCs are supported by government agencies while the academic Research Data Centres (RDCs) are part of an initiative supported by Statistics Canada, 3 research granting councils, and a university consortia whose aim is to strengthen Canada's social research capacity and to support the policy research community. RDCs provide researchers with access, in a secure facilities located at either universities or government facilities, to microdata from population and household surveys as well as increasingly administrative data from social and health programs along with integrated data. The RDCs are staffed by Statistics Canada employees and are operated under the provisions of the Statistics Act in accordance with Statistics Canada's policies and standards for security and confidentiality from Treasury Board. RDCs are accessible only to researchers with approved projects who have been sworn in under the Statistics Act as 'deemed employees' of Statistics Canada. As a 'deemed employee' researchers have access to the full analytical file in the centres.

The mandate of the RDC program is to promote and facilitate social science research using Statistics Canada's microdata, while *protecting the confidentiality of data*. Since the academic program was started in 2000 the number of cycles of data has grown to over 400 and has supported over 5400 projects and nearly 5600 researchers.

2.3.1 Protecting confidentiality in the RDC's

Statistics Canada has developed and maintained a multifaceted approach to data confidentiality and security to ensure the RDCs uphold Statistics Canada's commitment to maintaining the confidentiality of personal information. Security and confidentiality in the RDCs is understood holistically as a comprehensive and integrated framework of controls and responsibilities within an established culture of confidentiality.

The RDCs' integrated approach to protecting data confidentiality and security comprises five elements:



Physical protection in each RDC facility includes: security requirements for the physical structure, the locked storage of the server within the RDCs and access to the centre; and, clear procedures and protocols outlining the requirements for the storage of output, controls on printing and shredding, limitations on use of personal electronic devices and requirements for handling visitors to the centre. Statistics Canada employees are trained to maintain and actively monitor the physical protection controls in the RDCs.

Computer protection includes the: prohibition of any external connection other than to the RDC secure network; requirement of encryption of firewalls and encryption between centres; logical hardening of researcher workstations to prevent the download of data; and project based user accounts which limit access to only data for which approval was granted.

Personal responsibility and legal protection of each researcher as a deemed employee of Statistics Canada is established through the swearing or affirming the Oath of Secrecy, security status accreditation, and signing a Microdata Research Contract and the Values and Ethics Code. Statistics Canada paid employees have these same legal responsibilities established through swearing or affirming the Oath of Secrecy, and signing agreement with the obligations and terms of their employment contract and the Values and Ethics Code. Researchers attend a training session where they learn about the procedures and protocols and that they are personally responsible to comply with them.

Data protection is ensured by Statistics Canada's analysts who review all researchers' requests for removal of data outputs from the RDC. These outputs are subject to a risk based assessment of potential disclosure (i.e. vetting rules) based on established rules and procedures.

- Basic principles for vetting rules for survey data:
    - Minimum cell sizes based on sensitivity of the variables and the size of the sample

        - Weighted results are releasable
        - No low – levels of geography are releasable
- Basic principles for vetting rules for administrative data:
    - Minimum cell sizes
    - Depending on data set:
        - Scoring method used to assess risk of disclosure
        - Controlled rounding
- All released output are emailed to the researcher once vetted for confidentiality

All information including statistical results and all notes or documents must pass the vetting process controlled by RDC analysts to ensure any potential risk to confidentiality is substantially mitigated.

A *Culture of confidentiality* is established in the RDCs through the initial training of each researcher during an orientation session and provision of the Researcher's Guide. The training of researchers is maintained and enhanced by Statistics Canada's employees through ongoing discussions and interactions during all stages of the research project life cycle. A researcher may be required to participate in another orientation session if an RDC analyst deems it necessary to repeat this training for the researcher. Statistics Canada's employees are trained and supervised to establish and maintain the culture of confidentiality within the RDCs.

Collectively these elements are mutually re-enforcing. The extent of risk management for the RDCs is enhanced by the integrated nature of the security and confidentiality framework summarized here. Risk management for the RDCs needs to be understood in terms of the whole framework for data protection and confidentiality.

2.3.2 Limitations of the RDC Program

While the introduction of the RDCs improved access to Statistics Canada microdata it is recognized that there are some limitations with the established set of procedures and protocols. The current security protocols require staff to be present at all time during specific operating hours which does not offer flexibility to researchers. Confidentiality vetting rules for some datasets and some analytical methods or results are viewed by researchers as overly complex and restrictive.

2.3.3 Future of the RDC program

New funding was awarded by a granting agency to upgrade the computing capacity and the physical security of the academic RDCs. The physical security equipment include includes cameras in the RDCs and upgrades to the door access controls. These upgrades along with some changes in protocols will allow centres to remain open after business hours without staff. Moreover, lessons learned from the past 16 years on vetting rules will be used to review and simplify the process where possible while still ensuring the protection of confidentiality.

3. Access to Statistics Canada's Business and Economic Microdata

3.1 Centre for Data Development and Economic Research

The Canadian Centre for Data Development and Economic Research (CDER)[vi] is the program at Statistics Canada through which researchers can access Statistics Canada's holding of economic and business microdata to undertake approved research projects in a secure environment. CDER was first established in 2011 in Statistics Canada's headquarters in Ottawa to provide access to researchers from the federal government community. Access was extend in October 2012 to all researchers, including: academics from Canadian and foreign institutions, researchers from think tanks, researchers from federal and provincial government departments, and researchers from federal agencies. CDER is a cost recovery program, users are responsible for covering the entire costs of their projects. The project approval process in CDER mirrors that used by the RDCs, except that the subject-matter experts consulted are different.

In contrast to the RDC program that facilitates access to mainly data on individuals, CDER provides access to data on businesses. It provides access to data collected by Statistics Canada through surveys and administrative sources. The most frequently accessed surveys include: the Survey of Innovation and Business Strategies, the Survey of Financing and Growth of Small and Medium Sized Enterprises, and Annual Survey of Manufacturers. To these and other surveys can be linked data on firm performance that come mainly from administrative data from the Canada Revenue Agency, the federal agency that administers tax laws for the Government of Canada. These administrative data include: corporate income tax, income tax declarations of unincorporated businesses, payroll deductions, and personal income tax files. Survey and administrative data based microdata on imports, exports, foreign direct investment, Canadian direct investment abroad, research and development expenditures, patenting, advanced technology use, pollutant releases, and greenhouse gas emissions are also available.

CDER is also directly engaged in data development. Almost all research projects require the use of linked data. However, linkage is only one aspect of the data development at CDER. Data integration and derivation of variables important for analysis are also key activities. For example, in order to undertake a project on the contribution of immigrant business owners to employment creation in Canada, administrative sources needed to be used to identify the individual business owners and their ownership shares of private Canadian corporations, the identities of the business owners had to be linked to administrative data on immigrants, and the employment creation from the administrative microdata on businesses needed to be reconciled with the aggregate statistics already being published by Statistics Canada.

3.1.1 Protecting the Confidentiality of the Firm-level Data

The nature of the firm-level data differs from that of the individual-level data. The distribution of firm-level variables are often skewed, the data are sparse in certain dimensions (e.g., industry and geography), and certain firms dominate their industries. This introduces challenges in regards to disclosure and residual disclosure. There are also financial incentives involved in trying to identify the particulars of a firm's activities compared to an individual's.

In response to these concerns and concerns expressed by business respondents, additional risk mitigation strategies were adopted at CDER compared to the RDCs. These include: a limit on

tabular output to only what is necessary to motivate the research project; centralized vetting in consultation with subject-matter divisions at Statistics Canada's headquarters; the use of an in-house generalized system, G-CONFID, to perform disclosure analysis that takes into account dominance issues; and the use of an access system where researchers do not see the actual micro data, but are able view at CDER facilities, aggregate/analytical results based on the actual data (e.g., regression results) before they are vetted for confidentiality.

3.1.2 Limitations of the Program

Going forward, the challenges CDER faces are similar to many of the ones faced by the RDC program. Growing size of data files and the associated computational needs necessitate, and the high cost of repurposing administrative data for use in analysis, including the creation of data documentation.

The key challenge for CDER is the extending access to business microdata outside of Statistics Canada's headquarters in Ottawa by leveraging the existing and future infrastructure being considered for the RDCs.

3.1.3 Future Directions

The decision on how to give access to business micro data is based on many considerations. The legislative framework that effectively governs access varies across countries, so too does the attitudes of citizens, both respondents and data users, in regards to privacy and confidentiality. Informed decisions are based on this context, which is evolving over time.

Statistics Canada is currently working on developing synthetic data and limited disclosure databases for use in settings outside of its headquarters in Ottawa. It is working with the experts that developed the Synthetic Longitudinal Database in the United States, and it is working internally to create research databases for small and medium sized enterprises where dominance issues and the likelihood of re-identification are less probable. It is also evaluating platforms that perform disclosure analysis automatically, to speed up the vetting process.

Work continues in partnership with data suppliers and academics to build databases from various sources: surveys; administrative data from government and non-government sources, some of which have already been made public by the data owner and others that have not; and data that have been already been derived from other sources. The combination and manipulation of source data to create researcher databases raises the question of whether the development of variables for analysis creates perturbation into the data, the effect of it on disclosure analysis needs to be fully examined.

4. Access to Statistics Canada's Microdata Moving Forward

Statistics Canada is exploring the most effective means to modernize its digital infrastructure with the aim that the data are exploited to their fullest potential. While Statistics Canada is still collecting survey data there are increased efforts to acquire and utilize big data and administrative in innovate ways. In parallel there is development of an integrated social data

framework to allow researchers to better understand the interrelationships of issues. These new external and integrated datasets will then be added to the RDC data collection.

References

[i] *Statistics Act* http://laws-lois.justice.gc.ca

[ii] *Corporate Business Plan Statistics Canada* 2016/2017 to 2018/2019 http://statcan.gc.ca/eng/about/bp

[iii] *Public Use Microdata Files and the Data Liberation Initiative* http://www.statcan.gc.ca/eng/dli/dli

[iv] *Real Time Remote Access* http://www.statcan.gc.ca/eng/rtra/rtra

[v] *Research Data Centres* http://www.statcan.gc.ca/eng/rdc/index

[vi] *Canadian Centre for Data Development and Economic Research* http://www.statcan.gc.ca/eng/cder/index