WP. 19
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (iii): Research data centres and virtual labs

DEVELOPMENT OF A REAL TIME REMOTE ACCESS INFRASTRUCTURE AT STATISTICS CANADA

Invited Paper

Prepared by Michelle Simard, Statistics Canada

Development of a Real Time Remote Access Infrastructure at Statistics Canada

Michelle Simard*

* Statistics Canada, Household Survey Method Division, R.-H. Coats 16-F, Tunney's Pasture, Ottawa, Canada,K1A 0T6, michelle.simard@statcan,gc,ca

Abstract: Like many national statistical organizations (NSOs), Statistics Canada is facing increasing national and international demands from researchers for access to detailed microdata. Statistics Canada has recently initiated a review on how to improve access while at the same time protecting the confidentiality of respondent data. One of the options being considered is to develop a real time remote access (RTRA) infrastructure. RTRA is essentially an on-line remote access facility allowing users to run, more or less in real time, data analyses on microdata or lightly masked microdata sets kept in a central and secure location.

1 Introduction[1]

There is great interest in the use of detailed microdata by all sectors of Canadian society. This reflects government policies and programs that are focusing increasingly on specific segments of the population, such as the elderly, recent immigrants and the business sector. This greater interest extends to academia and to the international research community. Advances in technology have significantly opened up avenues, for both statistical agencies to produce and disseminate detailed data and for researchers to mine and analyse them.

To meet the increasing need for evidence-based policy development, Statistics Canada (StatCan), in partnership with other federal departments, developed in the 1990s a number of high profile longitudinal surveys. The Survey of Labour and Income Dynamics (SLID), the National Population Health Survey (NPHS) and the National Longitudinal Survey of Children and Youth (NLSCY) are a few surveys that were established to provide data in such areas as labour market and income, health and child development. These surveys have yielded large amounts of both longitudinal and cross-sectional data, well beyond the capacity of StatCan analysts to exploit fully. Consequently, academic researchers were commissioned by policy departments to analyze the data but, by the very nature of longitudinal analyses they required access to confidential individual records.

[1] Statistics Canada (2007) provides an exhaustive description of the issues of Microdata Access at StatCan. Sections 1 and 2 contain some excerpts of the sections of the report.

This situation presents the dilemma facing StatCan: how can the researchers access more data while still protecting respondents' confidentiality? The *Statistics Act* protects the confidentiality of the data StatCan collects. Any loss of public trust can be very detrimental to response rates and to the overall the quality of the country's statistical programs.

This paper discusses the current thinking and discussions on implementing a new mode of access. Section 2 provides background information on StatCan's current access situation, section 3 describes the issues surrounding the development of the real time remote access (RTRA) and section 4 discusses the confidentiality aspects. Finally, section 5 presents the characteristics of the prototype being built as well as the future testing and plans.

2. Background

Canada's *Statistics Act* does not allow confidential data to be provided to anyone who is not a Statistics Canada employee (or "deemed" employee), except under the provisions of Sections 11, 12 or 17(2) of the *Act*. Sections 11 and 12 cover the sharing of information with provincial statistical agencies and other organizations, respectively. Section 17(2) allows certain types of release of identifiable individual data by the Chief Statistician.

Given this framework, Canadian microdata can be accessed directly in only three ways. First, public-use microdata files (PUMFs) containing records that have been masked sufficiently to ensure there is negligible risk of identification and disclosure, are released under a contractual agreement. The agreement requires the user to use the data for statistical purposes only and not to attempt to re-identify records on the file. The second means of access to microdata is through research data centres (RDCs), which provide researchers with access, in a highly secure university setting, to microdata from population and household surveys. They are operated under the provisions of the *Statistics Act* in accordance with all the confidentiality rules and are accessible only to researchers with approved projects who have been sworn in under the *Statistics Act* as "deemed" employees. Thirdly, a microdata file can be released to a specified organization under the authority of Section 12 of the *Statistics Act*. Such files contain only the records for those respondents who have agreed to share their responses with that organization.

In addition, a number of surveys offer facilities for remote access, whereby researchers submit analytic requests to the appropriate survey area, which then runs the request on the unmasked data set(s) in StatCan's secure environment, checks the outputs for potential disclosure and returns the results to the researcher. This service relies upon the manual intervention of a technical officer to confirm that the outputs

returned are properly vetted for confidentiality. Depending upon the availability of the technician, his/her other workload and the complexity/volume of the outputs, this process may take anywhere from two to five 5 working days.

In January 2007, a working group was established to review international developments in improving access to microdata for research purposes and to propose what, if any, new approaches should be pursued. The scope of the review was limited to household survey datasets, with an emphasis on data from longitudinal surveys. Particular consideration was to be given to facilitating access for analysts outside Canada for the purposes of cross-national research. Three areas in particular were investigated: synthetic data files, RTRA and data sharing. More details can be found in Statistics Canada (2007).

This document focuses on the work done for RTRA. This approach builds upon the already successful RDC infrastructure and remote access services but with the added benefit of being able to return the results of the queries back to researchers in a much more timely fashion and without having to go to one of the RDC locations.

3. The RTRA plans at Statistics Canada

The objective is to deliver, in three to five years, a full package RTRA program with remote job submission for modelling and intricate/delicate programming as well as a tabulation tool for metadata driven descriptive statistics. The program would be in a user friendly format. There are three major phases in the development of the RTRA infrastructure. The already completed first phase was to gather StatCan business requirements allowing the Agency to gain a deep understanding of the different components such as the security, legal and functionality requirements. The second phase, currently underway, is to convert the security requirements into concrete tools - basically it is to build a prototype. The third phase will be to test the overall tools put in place to measure the level of security in a gradual approach and prudently expanding the program. The first and second phases are presented in the next section while the third phase is discussed in section 5.

3.1 The first phase: Defining business requirements

The first phase was to gather business requirements so that StatCan can build a successful remote access facility. This process basically involved learning from the past and present experiences of similar NSOs and establishing an overall direction for the project. The working group carefully examined different RTRA facilities, among them; Statistics Netherlands, the Australia Bureau of Statistics (ABS), the Office of National Statistics (ONS) and the *Institut de la Statistique du Québec (ISQ).* A long-term plan was developed based on this investigation. The first step

involved determining the requirements, i.e. the basic features, procedures and governance of the program. The key elements that were identified are as follows:

- Identifying the scope: descriptive statistics and modelling for survey and administrative data.
- Defining the approach: gradual implementation with the construction of a prototype using SAS. The program will be available to pre-identified partners, i.e. federal departments first; but will be expanded after careful evaluation of the risks.
- Developing the process model: determine how to access the system. This includes the Informatics, contractual and legal aspects. Diagram 1 presents the process model being developed for the prototype; from the request for access by a pre-identified researcher requesting access up to his/ her reception of output.
- Developing the governance model: determine how to fit and manage this new access as part of a suite of services i.e., direct access to the files, indirect access to the files, remote access and RDC.
- Ensuring proper organizational implementation: privacy impact assessment and risk and threat assessment, communication/marketing plan for the launch of the system. Establish partnerships for the development of the RTRA and determine infrastructure costs to accommodate the RTRA system.

3.2 The second phase: Determining the security framework

The second phase, currently underway is to convert the requirements related to security into concrete measures. There are four "security" control points, for which policy choices must be made: (1) security of the data sets housed; (2) security of the data in transit; (3) validation of registered users; and (4) confidentiality rules for output. Point 4 will be the sole topic of section 4. The various security components will be viewed as a package so appropriate trade-offs can be made. For example, the greater the masking of the data, the less stringent user authentication needs to be, and the less restrictive the rules governing output.

The model currently favoured is similar to that of the ABS Remote Access Data Laboratory (RADL). After signing a legal undertaking, users in Australia receive a username and password that they can use to link to a server through the Internet. They can submit a job using software packages (SAS, Stata and SPSS in the case of ABS). The software available is modified to prevent the use of particular commands and to comply with rules regarding the nature and size of outputs. Requests that do not comply with the rules can be submitted with a note and they will be kept in "quarantine" until an employee vets the output for confidentiality before releasing it. All jobs are kept for auditing purposes for the RADL. Thus, the system relies on several layers of protection: lightly masked data, signed legal undertaking, user training, user authentication, output restrictions and audits.

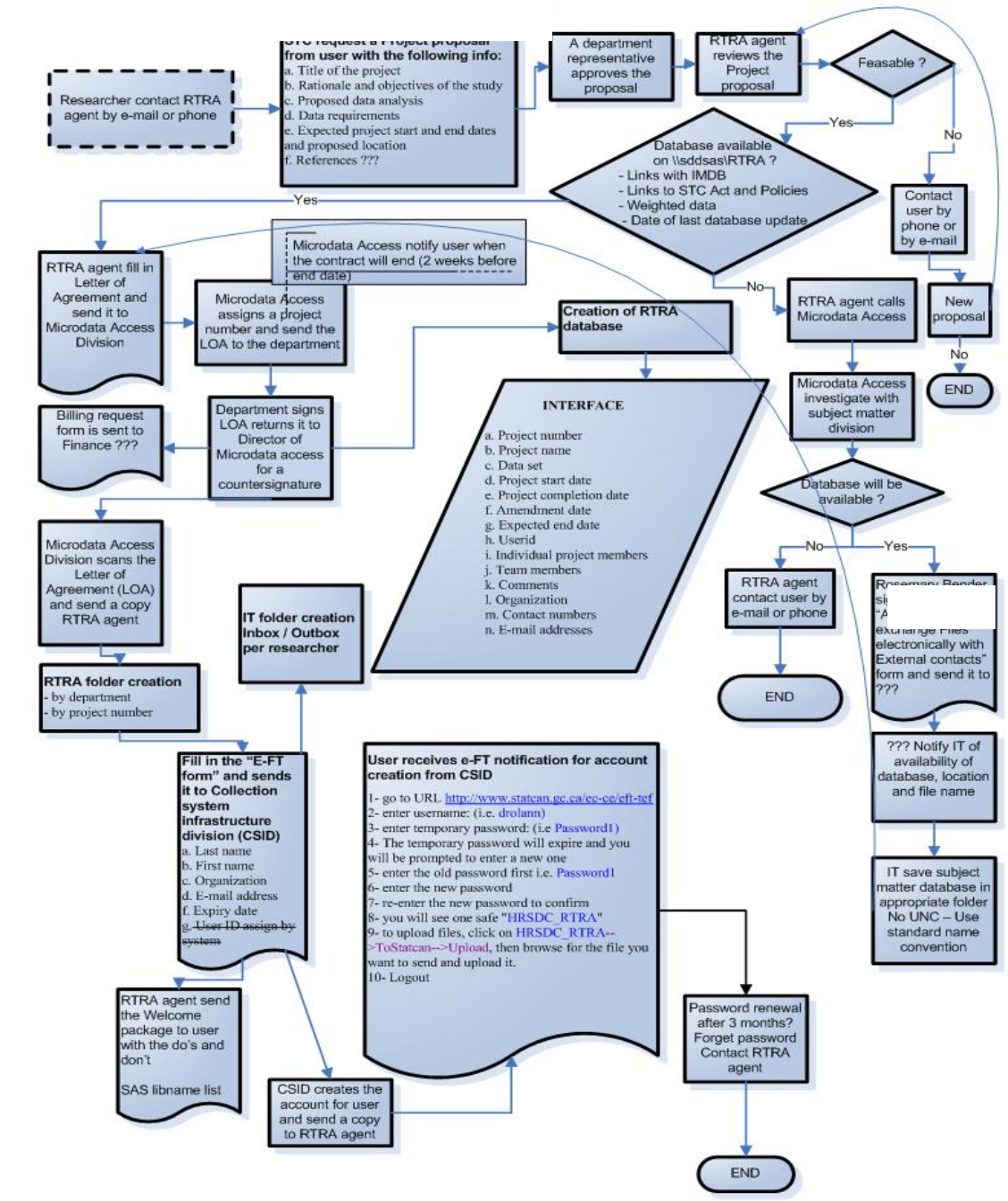

Diagram 1: Statistics Canada's overall access process

At StatCan, the prototype will be built upon the existing e-file transfer (EFT) system; a highly secure infrastructure developed for the Agency's two networks. It will be

used as a base platform to navigate the air gap between networks. Diagram 2 presents the prototype built based on the EFT. Similarly to RADL, a password-type approach has been implemented. SAS has been identified as the unique software package so far. The legal aspects are currently being developed and are carefully designed with the RDC in mind. They include the contractual arrangements, scope of the contract, the use of the data, the multi-request and linkage issues and the penalties aspects.

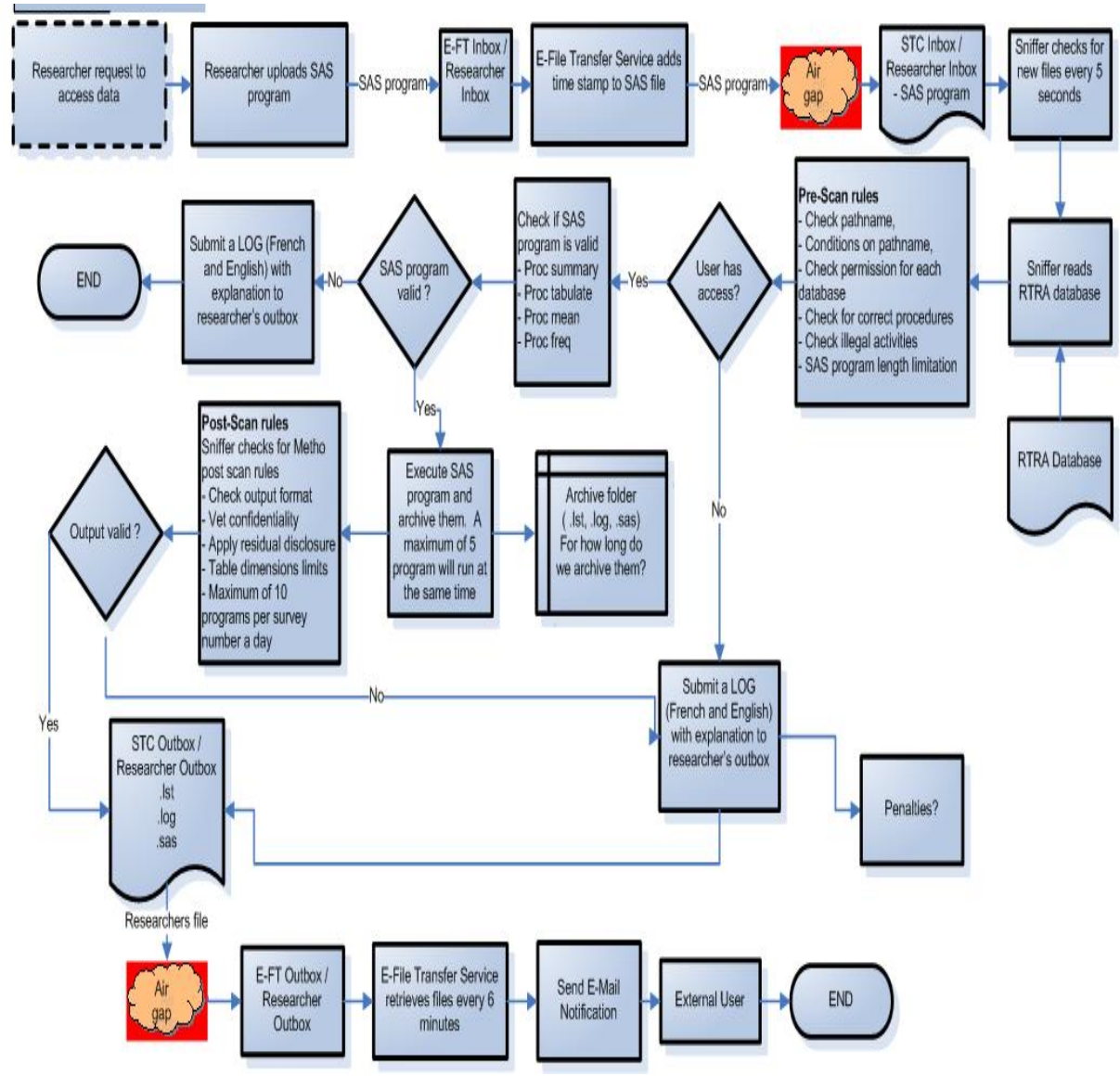

Diagram 2: RTRA prototype

4. The confidentiality aspects

There is no absolute criterion for defining confidential data. The boundary between confidential and non-confidential can be interpreted as the threshold between negligible and non-negligible risk. StatCan has a risk management policy to safeguard the confidentiality of the microdata. This includes a governance structure, security measures (physical and electronic security) and use of deemed employees. This will be the basic model used in the beginning. Over time, as there is an expected increase in the scope of access and as the expertise is developed, the risk management will be adapted.

In implementing RTRA, it will be necessary to develop rules for disclosure control. In looking at the literature and into the different NSOs, four different aspects are usually considered. The four are: Slightly masked microdata files; Automatic disclosure rules for tabular outputs; Pre-Scan or control rules for the inputs (manual and automatic) and; Post-Scan or control rules for the outputs (manual and automatic). Again, the strategy that StatCan will favour will be a trade-off of the four potential methodologies. The decision process in choosing the proposed methodologies involves managing risk while taking into account the other levels of security.

To mask or not to mask

For StatCan RTRA, although users would have to register, and there would possibly be some audit controls, there would be no requirement for physical security checking and swearing in of users. Furthermore, the databases could not be held within the StatCan secure network if they were to be accessible externally. Finally, automated checks provide very limited means for tracking multiple and overlapping requests, and therefore result in a non-negligible risk of secondary disclosure. For these reasons, it would be prudent if the microdata files were to be lightly masked, e.g. removal of detailed geography, some grouping of response categories, top-coding etc. but very far from the degree necessary for a PUMF. A possible strategy would be to modify slightly the geography and/or to use the Skinner-Elliot measures (Skinner and Elliot, 2002) to identify the probability of being unique and modify or mask only high risk records. The decision about whether to mask or not is still being evaluated.

Tabular output

Since the scope of the RTRA is descriptive statistics and modelling for survey and administrative data, disclosure rules need to be developed for tabular outputs. For this purpose, some options are currently being evaluated: these include suppression strategies using package such as Tau-Argus (Hundepool *et al*, 2009) and controlled tabular adjustment (Cox, 2004) as well as random and controlled rounding methods.

The criteria for evaluation are the robustness and risk associated with the method, the richness of the released information, capacity to evaluate the risk associated with multiple requests, the capacity to automate/program in the RTRA environment, the promptness of outputs produced, the capacity to handle large datasets and outputs and the ease of harmonization among all survey programs and with RDC rules. Controlled rounding as described in Boudreau, Filep and Liu, (2004) is the current favourite for frequency data. The use of a minimum unweighted count as a rule for manual verification is also being considered.

Pre and Post-Scan rules

As mentioned, the rules should be very similar to those used at RDC even though there are basic differences between the two venues. Tambay 2007 provides a good overview of the issues and potential methods to consider for these types of rules. Apart from that, little progress has been made on these two fronts.

5. The prototype, the tests, the future

In April 2009, a first test to evaluate only the logistic was completed using the EFT. This was completed with a deemed employee working in another Federal Department, a PUMF and a SAS program. It was evaluated as being very successful, since there were no manual interventions in the process. Everything was completed automatically with no hiccups. The second test scheduled for the spring of 2010 will offer tabulation procedures and the use of a limited number of cross-sectional survey data files. PROC Means, PROC Tabulate, PROC Summary and PROC Freq, in this stage of development will be the only allowed procedures. The decisions about whether or not to use slightly masked data and which disclosure method to use for tabular outputs for the test have not yet been finalised.

In subsequent years, the plan is to enhance features and establish standard vetting procedures. Some of the challenges will be: adjusting the services to client feedback and requirements; taking into account the new wide area network infrastructure under development for the RDC: adding more cross sectional survey sources to the one that are already accessible; developing longitudinal data vetting procedures and administrative data vetting procedures; and expanding the process model for academics and the private sector.

Conclusion

Statistics Canada is just at the beginning of the process of developing an application for real time remote access. Starting with a basic model, the modalities will be expanded as the risks of disclosure are better understood and managed. Over time, this application will become part of a continuum of access to Statistics Canada microdata which ranges from public use microdata files to the confidential microdata in the RDC. RTRA falling within this range will provide academic and policy researchers with an effective and timely access to microdata.

Acknowledgment

The author would like to thank Rosemary Bender, Johane Dufour and Jack Gambino for their comments and a particular thank you to Jean-Louis Tambay for his expertise and sound advices on the topic of confidentiality.

References

Boudreau, J.-R., Filep, K. and Liu, L.(2004). *Iterative rounding for large frequency tables. Proceedings of the ASA Joint Statistical Meeting. Toronto, 2004.*

Cox, L. H (2004). *Resolving confidentiality and data quality issues for tabular data. Proceedings of the ASA Joint Statistical Meeting. Toronto, 2004.*

Hundepool, A, Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Nordholt, E., Seri, G. De Wolf, P.-P., (2009). *Handbook on Statistical Disclosure Control. January 2009.*

Skinner, C.J, Elliot, M.J. (2002). *A measure of disclosure risks for microdata.* *Journal of the Royal Statistical Society*, *Series B, 64,* 855-867.

Statistics Canada (2007). *Increasing Access to Statistics Canada's Microdata. First Report of the Working Group. Version 2.1. Internal document, March 2007.*

Tambay J.-L. (2007). *Types of Disclosure Rules for Real-Time Remote Access (RTRA). Internal document, October 2007.*