Distr.
GENERAL

WP.8
16 May 2011

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2011)
(Luxembourg, 23-25 May 2011)

Topic (i): Architectures, models and standards

Statistics Canada's Real Time Remote Access Solution

Supporting Paper

Prepared by Karen Doherty and David Price, Statistics Canada, Canada

I. Introduction

1. Access to, and analysis of, Statistics Canada's data are fundamental to the fulfilment of Statistics Canada's mandate. By providing researchers with access to its statistical outputs, the Agency's data continue to make an immeasurable and vital contribution to public policy debates in Canada that benefit many of the nation's citizens and businesses.

2. Statistics Canada has traditionally provided access to its statistical data through various means:

   - aggregate data posted on the Agency's Web site;
   - public use microdata files (PUMFs); and
   - special and customized tabulations of aggregate data.

3. The creation of over 20 Research Data Centres has permitted access to confidential microdata files to researchers across the country. Although these Research Data Centres (RDCs) are located in universities, they operate as satellite Statistics Canada offices and provide researchers with controlled and secured access to Statistics Canada microdata for purposes of statistical research.

4. However, like many other national statistical organizations, Statistics Canada is facing increasing demands from the research community for even greater access to detailed microdata. This greater interest in the analytical use of microdata by researchers reflects its importance in the development and analysis of government policies and programs, many of which focus on one hand on specific demographic populations such as the elderly and recent immigrants, and on the other hand on the business sector.

5. This interest in the use of statistical information extends not only to academia and policy makers within the country but also to the international research community for cross national studies by the World Bank, World Health Organization, etc.

6. Ongoing advances in informatics technology have significantly opened up avenues for statistical agencies to produce and disseminate detailed data and for researchers to mine and analyse.

7. This increasing demand for access, coupled with operational procedures and requirements, have caused researchers to express frustration with the roadblocks that impede their ability to conduct the research they wish to do. One of the time consuming conditions is a requirement to submit a formal detailed research proposal for approval before access is granted to microdata. Some researchers are frustrated that they have travel to the nearest Research Data Centre and adjust their work schedule to the RDC's hours of operation.

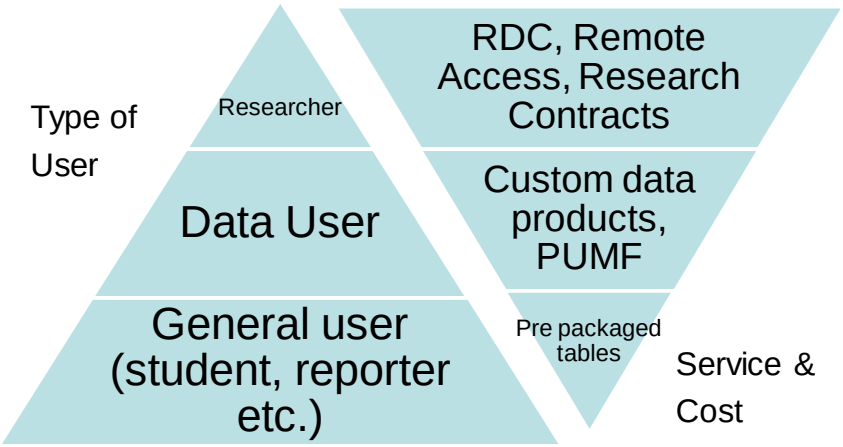
8. In recent years, Statistics Canada has invested much time and effort in examining ways to meet the demands of researchers while ensuring that the legislative requirement under the *Statistics Act* to protect the confidentiality of identifiable respondent data is met.

II. The Real Time Remote Access System

A. The Business Solution

9. To address the concerns of researchers Statistics Canada is developing a Real Time Remote Access (RTRA) system. This system is essentially an on-line remote access facility that would allow researchers to run – more or less in real time – data analyses on microdata or lightly masked microdata sets kept in a central and secure location under the control and care of Statistics Canada.

10. The RTRA system enhances the options available to researchers within the Agency's overall data access strategy.



11. Once fully developed the RTRA system is expected to address many of the impediments The system provides researchers with access to the data they need at any time, on any day, directly from their own office.

12. Finally, it should be noted that rather than replacing current methods of access to confidential microdata such as by the Research Data Centres, this new system would be complementary to the suite of options that are currently offered to researchers.

B. Development of a Working System

13. The first phase, completed in 2009, included the identification of business requirements providing a clear understanding of the different components of any potential system such as security, legal and functionality requirements.

14. The second phase consisted of the development of a pilot version of the system which was made available to a limited number of researchers in the spring of 2010. This initial model had a limited audience (researchers employed by other federal government departments) and placed certain restrictions on both the types of requests that could be submitted and the level of detail of the statistical outputs.

15. The datasets available to researchers at this point in the development of RTRA include:

- General Social Survey – Cycle 19 and 20 (GSS)
- National Graduates Survey (NGS)
- Aboriginal Children's Survey (ACS)
- Aboriginal Peoples Survey (APS)
- Participation and Activity Limitation Survey (PALS)
- Canadian Community Health Survey (CCHS)
- Survey of Labour and Income Dynamics (SLID) and Labour Force Survey (LFS) to be added this year

16. The third phase will be a test of the complete functionality of the system that would be expanded incrementally so as to permit the necessary ongoing evaluation of the system in terms of the required security measures and the identification and mitigation of risks. New datasets will be added and new functionality will be developed.

C. The First Phase: Defining the Business Requirement

17. This phase involved learning from the experiences of other national statistical agencies in order to establish an overall direction for the project. A working group examined the Real Time Remote Access solutions offered by Statistics Netherlands, the Australia Bureau of Statistics, the Office of National Statistics in the UK and the *Institut de la Statistique du Québec.* This phase involved determining the key requirements of a Statistics Canada model, i.e. the basic features, procedures and governance of the program. These were identified as follows:

- Identifying the scope: descriptive statistics and modelling for survey and administrative data.
- Defining an approach, i.e., a gradual implementation with the construction of a prototype using the program SAS that will be tested in conjunction with limited partners (other federal government departments); expansion would follow balanced with careful evaluation of risks.
- Developing the process model, that is to say determining access to the system. This step would involve examination of the IT solution as well as contractual and legal aspects.
- Developing the governance model for the system.
- Ensuring proper implementation of the system by conducting a privacy impact assessment, required threat and risk assessments and creating communication/marketing plans for the launch of the system. This would also involve establishing the necessary partnerships for the development and management of the RTRA system.

D. The Second Phase: Determining the Security Framework

18. The model developed for this phase is similar to the one used by the Australian Bureau of Statistics. The prototype was built upon the existing e-File Transfer (e-FT) system used by Statistics Canada. The e-FT is a highly secure program developed for data transfers between the Agency's networks (private and public) and outside data users and respondents. It provides a mechanism for researchers to submit requests and pick up results and to transfer this information across the air gap that exists between the public and private networks operated by Statistics Canada.

19. Another important aspect of this second phase is to address the security requirements by establishing concrete measures. There are four identified "security" control points:

- the security of how the datasets are housed;
- the security of the datasets in transit;

- the validation of registered users; and
- the confidentiality rules for output.

20. The various security aspects will be viewed as a whole in order to strike the right balance of risk versus security. For example, the system of higher levels of masking on the data would result in less stringent user authentication and/or less restrictive rules that govern the output.

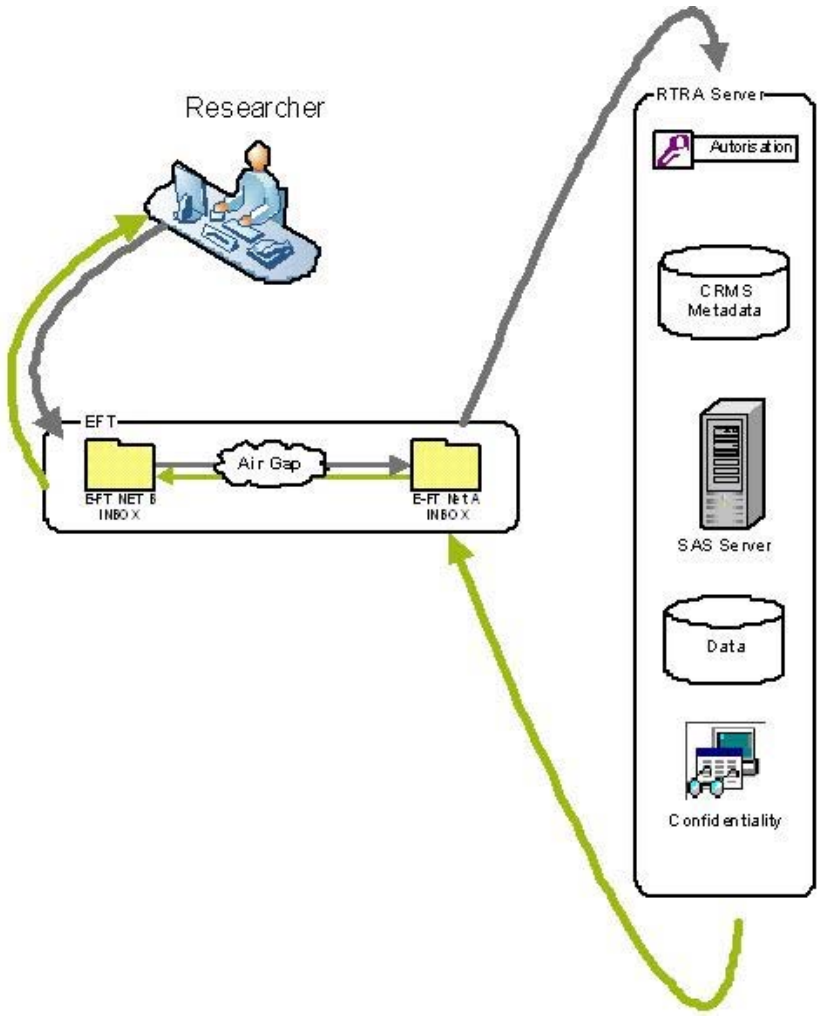
E. How RTRA Works

21. Researchers are issued a username and password that they use to link to the Statistics Canada e-FT external server via the Internet.

22. Once the system is in full production mode, researchers will be required to sign a contract that outlines rules and responsibilities as well as penalties and disciplinary actions if they are found to have breached the rules.

23. Each researcher takes a brief RTRA training session to familiarise themselves with the system and understand the implications of accessing the data remotely.

24. The RTRA system allows a researcher to submit a SAS program to a SAS server that has been modified to prevent the use of particular commands and to comply with rules regarding the nature and size of the statistical outputs. The request passes through Statistics Canada's IT security firewalls before being delivered to the secure internal Statistics Canada server. The datasets used for the pilot consist of confidential microdata that have been lightly masked to remove outliers that could easily identify an individual, and all tabular outputs are automatically weighted and vetted for confidentiality. Following the vetting for disclosure of confidential identifiable information the tables are sent back to the researcher in the specified format.



25. Because the system runs a pre-scan operation, requests that do not comply with the guidelines and parameters of a submission will not be run by the RTRA system. To monitor such incidents, a log is generated by the system indicating how the program did not comply with the guidelines. The log is sent back to the researcher so that the researcher can adjust their submission.

26. All submission files and resulting log files are kept indeterminately for auditing purposes.

27. The pre-scan of requests does the following:

- limits access to data files;
- ensures that the programming guidelines have been followed;
- uses automated SAS processes to control output.

28. The post-scan of outputs does the following:

- applies a controlled rounding algorithm to output tables (vetting against disclosing confidential information);
- limits each submission to 10 tables;
- limits each researcher to 10 successful program submissions per day;
- supports two formats for output (.sas7bdat) and HTML.

F. The Methodological Challenge of Ensuring Confidentiality of Microdata

29. From a methodological point of view, there is no absolute criterion for defining confidential data. However, the boundary between confidential and non-confidential data can be interpreted as the

threshold between negligible and non-negligible risk. Therefore, in terms of disclosure control, Statistics Canada applies numerous risk management practices to safeguard the confidentiality of microdata.

30. In the development of the RTRA system, it has been necessary to develop specific rules for disclosure control. Based on literature and the practices of other statistical organizations, there are usually four key aspects:

- 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).

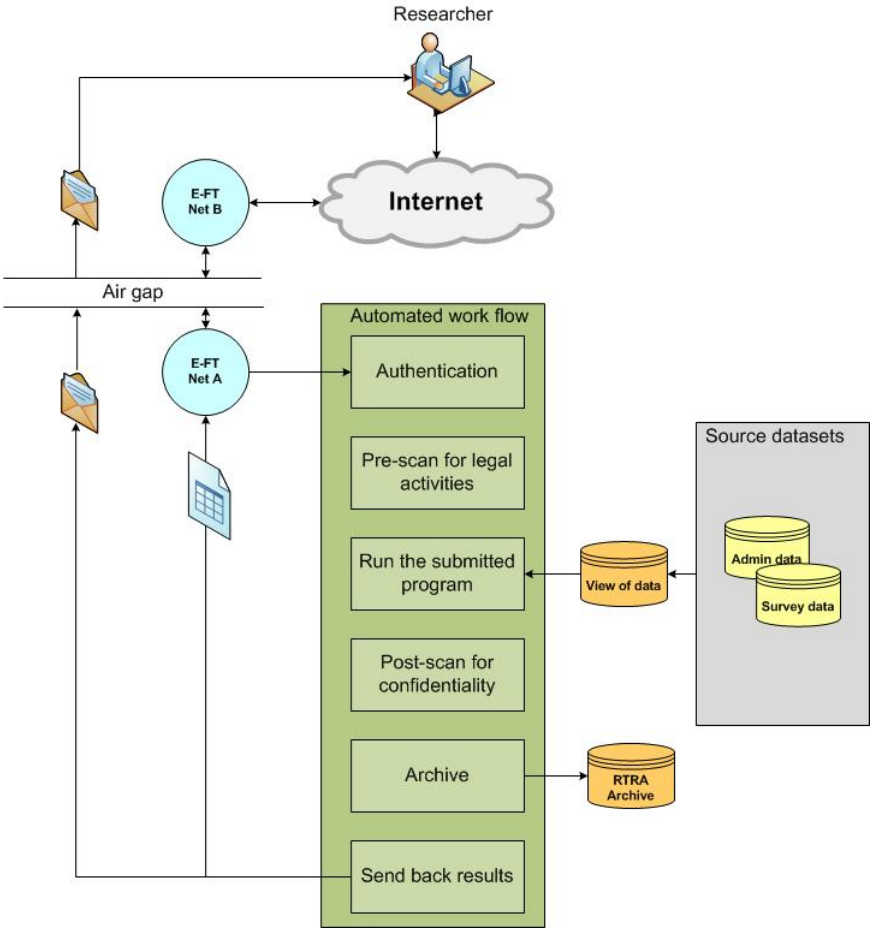
31. The strategy that Statistics Canada is adopting involves trade-offs of the four potential methodologies. Any decision in choosing the proposed methodologies will involve managing risk and take into account other levels of security.

III. The architecture of RTRA

A. The Design

32. At this point RTRA is limited to tabular outputs only and to a particular set of household survey data files. The current version is also limited to requests delivered in SAS format.

33. The following diagram shows the components of the RTRA system.

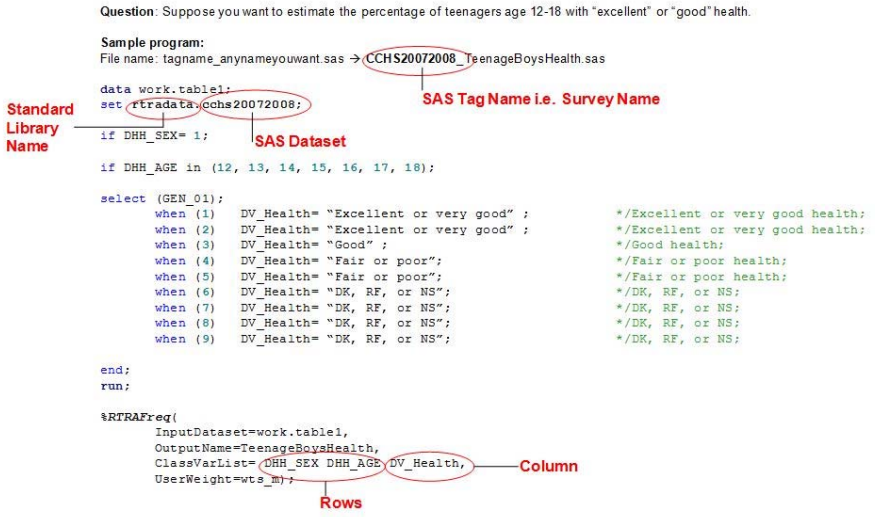


34. The system uses the following technologies:

- File transfer: e-File Transfer service (COTS implementation)
- Individual components of the work flow: SAS
- Authentication of the user rTRA credentials: SAS and Statistic Canada's Customer Relations Management System
- Archive: file folders
- SAS data views
- Automated work flow: currently uses SAS Sniffer but we are looking at the feasibility of migrating to the StatScan Enterprise Bus solution once it is operational or to a workflow manager product.
- Post-scan: currently uses a Statistic Canada in-house controlled rounding tool (RNDII.exe) developed in C++ but are looking into the feasibility of adopting Statistic Canada's Generalized Tabulation System (G-Tab) once it is in production or working in partnership with other parties to develop or adopt another tool.

B. The User Interface

35. The user creates a request, currently in SAS, as in the example below.



36. The user then logs onto the RTRA system which is accessed from the StatCan website.

Login

After Login

37. The user then submits the request.

38. The resulting data will be delivered to the external FTP server via Statistic Canada's e-FT system.

IV. The Future Direction and Long-Term Development

39. In subsequent years, the plan is to enhance features and establish standard vetting procedures. Already a number of challenges have been identified that will need to be addressed:

- adjusting services based on client feedback and meeting requirements;
- being in synch with the new wide-area network infrastructure used by the Research Data Centres;
- making more cross-sectional surveys available to researchers;
- developing vetting procedures for longitudinal survey data and administrative data; and
- developing the system so that it could tap into a potentially wider audience of academics and the private sector.

40. Work to implement the following new functionality is planned for 2011:

- Quality indicators for frequency tabulations are being introduced. S.E.'s, C.V.'s and C.I.'s will be incorporated into RTRA by June 2011.
- Means, medians, percentiles, ratios and proportions will be available by August 2011.
- Support for SPSS and other programming languages is being investigated.
- Census information will be introduced by November 2011. The constraint is that the Census output must be equivalent to what is produced today from the Census Tabulation System.
- Work with the G-Tab (new Generalized Tabulation System under development at Statistic Canada) development team will begin in 2011 to provide input on how to automate confidentiality by types of output. The hope is that G-Tab will eventually replace the confidentiality and quality indicator modules in RTRA

V. Conclusion

41. After a slow start, the RTRA is now starting to gain traction among Government of Canada researchers. As the system evolves and new functionality is added, Statistic Canada believes that this tool will become a key component of the toolset available to researchers, whether they are policy researchers in government departments and agencies at the federal, provincial or municipal level, academic researchers in Canadian universities or in fact, any other researcher that agrees to the RTRA terms and conditions of use.