# Progressing from Safe Data to Safe Output to Safe Metadata: Managing the risk of disclosure

Charles Humphrey
University of Alberta

Canada's national statistical agency is confronted with increasing demands to generate new knowledge about Canadian society from its many surveys. Pressures from within the Federal public service have recently been building in conjunction with a push for evidence-based policy making, initially driven by a crisis in Canada's health care system in the 1990's. The demand for policy-relevant knowledge from Statistic Canada's surveys became a major driver within the agency as it emerged from the 1990's. During the same period, Statistics Canada introduced a series of longitudinal surveys capturing life course events of Canadians in the areas of employment and income, health, childhood and youth experiences, transitions into the workplace by young adults and the experiences of new immigrants to Canada. The cumulative cost of these new longitudinal surveys raised the eyebrows of Treasury Board officials who wanted to know the public return on these expenditures. What new knowledge has been gained from these surveys that has contributed to improving the lives of Canadians?

With a limited research capacity in the national statistical agency and with a Federal public service that had many of its policy analyst positions eliminated during the downsizing of the public labour force in the 1990's, the challenge became one of finding ways of engaging the wider Canadian research sector to produce new knowledge with the agency's data resources. The paradox confronting Statistics Canada was how to maintain its ethos of data confidentiality, prescribed by legislation and embodied in its organisational culture, while increasing access to the detailed microdata sought by researchers outside the agency.

How best to ensure a strong return on public investment in data is not a new puzzle for the Chief Statistician. However, the combination of recent factors has heightened the stakes. There has never been as much detailed data about Canadian society as exists today. However, the mechanisms to provide access to these data must minimise risks defined by legal constraints and societal expectations about confidentiality. To address risk levels, current assessments tend to focus on the threat of disclosure while minimal attention is given to the risks of impeded knowledge creation. What are the costs imposed by our current mechanisms for data access on the generation of new knowledge? This is a question that has not received a lot of attention.

## A Tale of Two Microdata Dissemination Mechanisms

Over the past thirty years, research access to Statistics Canada microdata has been available through two main dissemination mechanisms. The first of these, *anonymised microdata*, arrived on the scene in 1975 with the release of public use files from the 1971 Census of Canada. Since this initial offering, over eighty of the agency's social surveys have incorporated microdata in their line of products, with files ranging in price between $1,000 and $2,500 per survey today. Many of these social surveys are part of an on-going series, contributing to the ever-increasing number of public use files. In 1996, an academic subscription service, known as the Data Liberation Initiative (DLI), was introduced to buffer the direct expense of these data from researchers and students. While the demand for public use microdata had been fairly constant over time, DLI contributed to a significant rise in the popularity of these files. Students enrolled since 1996 in a Canadian post-secondary institution with a DLI membership have only known a world in which Statistics Canada public use microdata have been available without cost, although certain licence conditions still apply. This convenient access through DLI has heightened researchers' expectations for greater access to microdata.

The second mechanism of access was introduced in late 2000 with the opening of the first of nine Research Data Centres (RDC's) located across the country on selected campuses. By the end of 2001, all nine Centres were operational and in the following few years, four other RDC's and six branches have been added. These facilities are highly secure Statistics Canada offices housing the confidential master files from social surveys. While each university technically owns the facility in which the RDC is located, Statistics Canada, as the tenant, staffs and runs the operations in these Centres, which are only open when a Statistics Canada employee is present. Researchers are allowed entrance to one of these data enclaves only after becoming a 'deemed employee' of Statistics Canada, signing a contract and having an approved project that specifies in advance the confidential data that is needed. Researchers cannot bring laptops or portable storage devices into the Centre and the only output they can remove must pass disclosure analysis.

The twenty-five year gap between the introduction of public use files and the opening of Research Data Centres can be explained by a combination of factors. Certain barriers had to be overcome before the second mechanism of access became acceptable. To begin, major shifts in computing technology occurred in this quarter century, moving from mainframe to personal computers to network computing. Each shift introduced greater computing power at a more affordable price to a wider range of individual researchers. By 1995, twenty years after Statistics Canada's first public use file, access to computing power had all but disappeared as an obstacle to processing microdata. A new impediment, however, had been imposed in the mid-1980's by Statistics Canada's pricing policy on data products. For example, the cost of a bundle of similar products from the 1981 and 1986 Censuses rose from $10,000 to over $200,000, respectively. This barrier eventually was addressed in 1996 through DLI, which established a flat subscription fee to post-secondary institutions for all standard data products, which includes public use microdata.

Almost concurrent with the introduction of DLI and a wider flow of public use microdata, the decision was made within Statistics Canada not to produce pubic use files for its wave of new longitudinal surveys. Some cross-sectional public use files were produced for a few of these longitudinal surveys but the production of such files had ended by 2000 for all except the Survey of Labour and Income Dynamics. While researchers outside of Statistics Canada never had access to the longitudinal data, they also now lost access to even these modified cross-sectional public use files. By the end of the 1990's, the push for evidence-base policy making and the drive by Treasury Board for a benefits analysis of the longitudinal surveys paved the way for Research Data Centres, only the second access mechanism to Statistics Canada microdata but the first for confidential data.[1]

## Approaches to Managing Risk

Two basic approaches to managing disclosure risk underlie these two mechanisms of access. With public use microdata, the survey division minimises this risk by making a safe data file: the cases remain real but the detail in a subset of variables is altered or removed. The steps typically taken include reducing the level of geographic coverage, capping values, reducing detailed coding to a smaller number of general categories and removing sensitive variables or cases. The manager of a survey makes these transformations while endeavouring to preserve as much substantive content with research value as possible. Once a file has been constructed, the manager meets with the agency's Data Release Committee and presents her or his public use version. The Committee, which consists of a mix of senior Statistics Canada officials and methodologists, then assess the level of risk represented in this file. If the Committee members feel the risk is too great, the manager is sent back to make further modifications. If they are satisfied, the release of the public use file is scheduled for announcement in **The Daily**, the official mouthpiece of Statistics Canada

Public use files are part of the production stream of a survey. Therefore, the survey manager must include the cost of creating this product in the survey's budget. If the Data Release Committee rejects the proposed public use file, the manager may find herself or himself in a squeeze, depending on how much budget remains for a second go at a public use file. I have had survey managers

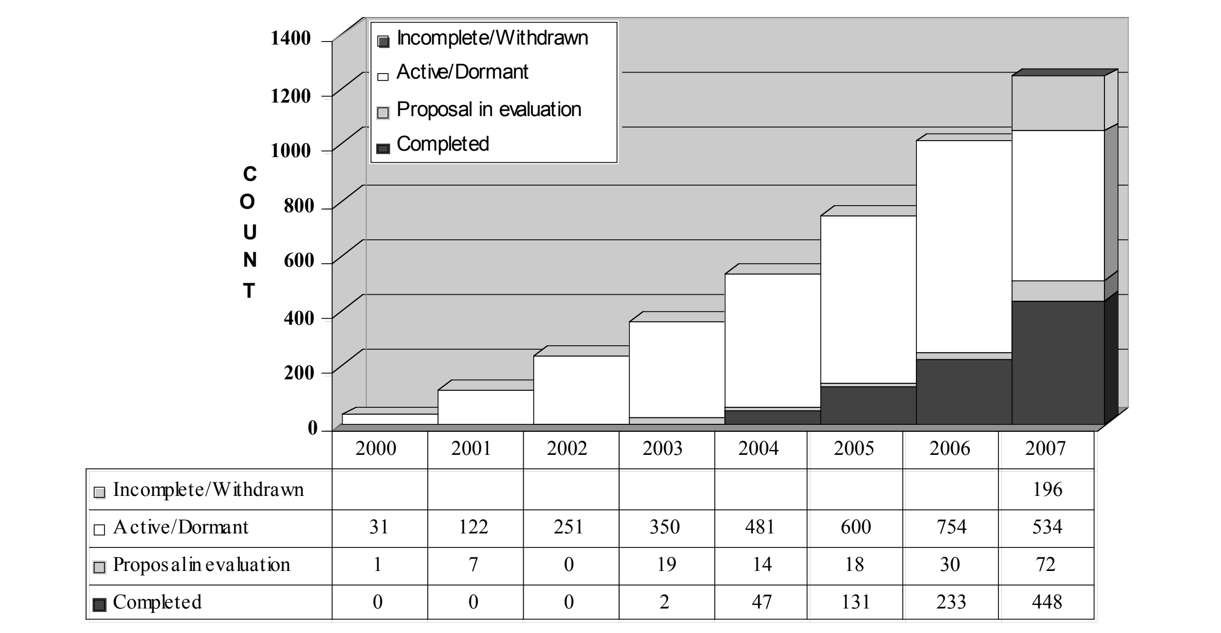
[1] A couple author divisions did introduce a service for their longitudinal surveys called "remote access." These really were "remote job submission" services where researchers were asked to submit files via email containing analysis runs in the syntax of one or two of the major statistical systems. The division would run the code as received and then vet the output for disclosure analysis. Results clearing disclosure analysis were encrypted and sent to the researcher via return email. No uniform "remote job submission" service developed within the agency and instead the offerings varied from survey to survey. For example, some divisions charged fees for this service but the rates tended to differ between surveys. One division even provided the service for free for a period. There was no wide uptake of this service by researchers.

say that the experience of going before the Data Release Committee is unpleasant and one to avoid. Consequently, public use files are produced with great anticipation of the response from the Data Release Committee. These tend to be one-shot efforts and while a revised version of a public use file may be released to correct errors subsequently discovered in the data, I have never known a second, competing version of a public use file to be created. For one thing, a second release would allow for comparisons leading to disclosure issues. The pressure to get approval on the first try means that the perceived priorities of the Data Release Committee receive the highest attention. What researchers may want in the data ends up a lower priority.

Anonymised microdata, as much as we treasure them, are created at the tail end of a survey's production stream with a focus on satisfying the Data Release Committee's concerns about confidentiality. From the data producer's perspective, this is how it should be. From the researcher's point of view, she or he is left to respond to a safe public use file, which may not satisfy her or his analysis requirements.

The second approach to managing disclosure risk involves controlling the use of confidential data within secure environments and assessing all output that is removed from these environments for possible disclosure. The overhead to implement and operate a data enclave environment is substantial. The physical environment must meet the host agency's security standards and the rules of operation must be rigorous enough to satisfy the Chief Statistician, who is ultimately responsible for ensuring the confidentiality of survey responses. As an academic director of one of the original nine Research Data Centres and who has been involved in the development of this programme from the beginning, I can attest to the adherence to the strictest practices established at the very beginning both in constructing and operating these environments. In building the facilities, blue prints, materials and construction methods had to be approved by Statistics Canada's security experts. Local area networks were not allowed external connectivity to either the campus network or the wider Internet. The only computers with access to the data are the ones connected to the internal LAN within the enclave.

Access to the data is determined on a project-by-project basis. Researchers must submit a project proposal that is peer reviewed and an assessment is made whether the use of confidential data is really required to conduct the research. Furthermore, researchers must undergo a security clearance to become a 'deemed employee' of Statistics Canada. Accounts are created for approved projects but the only data accessible under an account are the files that were approved for each project. A contract is signed with every project and a mandatory orientation session introduces the researcher to the rules of working with confidential data. Electronic entrances to the facility track traffic flow in and out of the RDC; however, researchers with approved projects can only be in the Centre while a Statistics Canada employee is present. No laptops or external



| | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|---|---|---|
| ▨ Incomplete/Withdrawn | | | | | | | | 196 |
| □ Active/Dormant | 31 | 122 | 251 | 350 | 481 | 600 | 754 | 534 |
| ▨ Proposal in evaluation | 1 | 7 | 0 | 19 | 14 | 18 | 30 | 72 |
| ■ Completed | 0 | 0 | 0 | 2 | 47 | 131 | 233 | 448 |

Figure 1
Cumulative Growth of Projects Up to April 13, 2007
Source: Gustave Goldmann, RDC Manager's Report, April 2007

storage devices can be brought into the enclave and the only output that can be removed has be to vetted for disclosure risk and approved by a Statistics Canada Analyst.

Access to the confidential data of the cross-sectional and longitudinal surveys identified in approved projects is the reward for complying with these strict rules. This method manages risk by controlling the environment in which the research is conducted. Steps are taken to ensure that the users of the data are 'safe' and that only 'safe' output is removed from the Centres

The number of projects within Canada's Research Data Centres (see Figure 1) indicates the willingness of researchers to work with the rules stipulated by Statistics Canada. Because Centres are located in most large cities across the country, many researchers have local access, although the researchers in Saskatchewan, Prince Edward Island and Newfoundland are without an RDC in their province. Some issues exist around the hours of access for researchers. Centers tend to be open only during the business hours of 9:00 to 5:00, Monday through Friday, preventing researchers from working evenings and weekends. Recently, some sites have begun offering extended hours to allow access outside normal business hours.

While improvements are being made to accommodate researchers across Canada, researchers from other countries interested in analysing Statistics Canada confidential data remain at a disadvantage. This issue surfaced in a conference co-sponsored by Statistics Canada in January 2006 that addressed the future of longitudinal social and health surveys. Approximately seventy worldwide experts were invited to discuss the value of longitudinal surveys. Early in the conference, a researcher presented results from a set of similar longitudinal surveys on a graph consisting of several countries, none of which were Canada. The researcher paused and asked rhetorically of his own graph, "Where's Canada?" The answer was obvious: Canada was not included in the analysis because the researcher could not obtain access to Canadian longitudinal data.

Finding ways to provide access to researchers outside of Canada and to those who don't have convenient access to a Centre within Canada presents an opportunity to discover yet a new approach to assessing risk assessment. Is there something between public use files and data enclaves?

### A Third Method: Metadata-driven Risk Assessment

Both anonymised microdata and data enclaves operate within an all-or-nothing approach to providing data access. With public use files, the survey division prepares an entire file that must pass the scrutiny of the Data Release Committee. Approval for release is contingent upon minimising the risk for the complete set of variables and cases. The researcher's needs, as defined by the variables and cases that she or he requires, are not a factor in determining public use access. Similarly, a researcher with an accepted project in an RDC receives access to the entire confidential data file regardless of her or his need for the complete file. For admission to an RDC, a key question is whether the researcher's proposal is possible given the data being requested. This evaluation is not done with the intention of protecting the data but rather is done to determine whether the data will support the proposed research.

Taking into account information representing the data requirements of the researcher presents a third approach to assessing disclosure risk. This method would begin by identifying the data requirements of the researcher and then would assess the risk of providing only the data requested from the confidential file. Data extraction tools are already available for public use data allowing researchers to select the specific subsets of variables and cases needed for their research.[2] While confidentiality is not an issue with public use files, tools nevertheless exists that allow researchers to choose data on the basis of their data needs.

[2] In Canada, Sherlock, LANDRU and IDLS are three popular data extraction services on the Web that allow researchers to create subsets of public use data disseminated through DLI. Currently, the DLI unit in Statistics Canada and a consortium of Ontario universities are developing their own NESSTAR servers to offer similar services.

If this idea is not particularly new, why has it not been tried earlier with confidential data? For one reason, the two current approaches flow from the production cycle of a survey. The system is geared to produce a public use microdata product that is disclosure safe and yet rich enough to satisfy a wide range of research questions. The research community may be consulted about the information loss in making a public use file but individual research needs fall outside the formula in creating the final anonymised version.[3] Another reason is that survey units do not have on-going dissemination services around which to build requests for individual subsets from their microdata. Once the production of a survey is complete, the unit begins work on the next survey. Until Research Data Centres were established, no continuous dissemination service existed for confidential data in Statistics Canada.

Some survey units have developed specific data extraction tools for use with their own microdata. For example, SLIDRET is software that helps specify a unit of analysis and a corresponding set of variables from the longitudinal Survey of Labour and Income Dynamics (SLID). SLIDRET is available to researchers in RDC's but only those with projects approved to use SLID. A version of SLIDRET without any data is available for researchers to identify variable subsets outside RDC's. This is useful to a researcher with an RDC SLID project to prepare her or him for generating a SLID extraction in the RDC. However, no service currently exists that allows a researcher to request a subset for use outside an RDC and to have the data assessed for confidentiality risk. Other survey units have created tools to help researchers identify variable subsets to test statistical software syntax prior to running the code with the actual microdata.

There seems to be no coordination in the development of these tools across survey units. Each unit tailors its extraction utility to their data. From the perspective of someone working in Statistics Canada, this may not appear to be a problem since each survey is fairly autonomous. For the researcher outside the agency who needs to work with several surveys, the variety of software requires additional learning, although the tools may provide similar functions. Each tool, while helpful for the survey to which it belongs, operates within its own stovepipe. This scattered development of tools may be a manifestation of not having a common dissemination service for microdata extractions. On the other hand, they also pose an impediment to achieving a common service at this point.

To introduce this new third approach to risk-managed microdata dissemination, two criteria must be met. First, the service will need to be established outside the survey production cycle. An extraction service involving some aspect of disclosure assessment will need to be housed in an environment supportive of

[3] For example, the production of the public use microdata files for the 1991 Census of Canada was preceded by a consultation of the academic community to discover the variables researchers least wanted to see modified by the anonymisation process. In recent years, some survey managers seeking similar advice have informally consulted the DLI External Advisory Committee.

these tasks. It must have the capacity to respond to individual requests and to be responsive to the priorities of individual researchers. Secondly, the system supporting requests for individual data subsets must be uniform across all surveys. Introducing a new survey into the system should be straightforward and based on a common, underlying information structure or metadata model.

The Research Data Centre Network (RDCN) could very well meet both of these criteria. The RDCN recently received an award of $4 million CDN dollars from the Canada Foundation for Innovation (CFI) to enhance its infrastructure. This investment is seen as a catalyst in making a generational change in the Network and its operations. From the beginning, each Research Data Centre has operated independently with its own local area network. Encrypted data files are shipped via courier to each RDC where a Statistics Canada Analyst decrypts and stores the data on the local file server. This type of computing network has been required by Statistics Canada as a security measure. With no Internet connectivity, no attack is possible on RDC files.

One of the CFI-funded projects will establish a lightpath articulated private network on Canada's national optical research network, forming in a national intranet for the thirteen Centres and six branches spread across the country. Statistics Canada's Communications Services Section, which is responsible for the agency's computing security, approved the use of lightpath technology in July 2007 after being convinced that this technology meets their security requirements.[4] With point-to-point connections between Statistics Canada's Federal Research Data Centre in Ottawa and each of the academic Centres, this high-speed network will transform the RDCN into a truly functional intranet. The coordination of services across the full Network will now be possible and make the sharing and distribution of tasks among Statistics Canada Analysts easier. With better management of the workloads across Centres, other applications and services may be possible within the RDCN. One new offering could be an extraction service for vetted confidential data requests, which addresses the above criterion for a dissemination service located outside the survey production cycle.

The second criterion is dependent on developing interoperable, standardised metadata across all surveys in the RDCN. The common, underlying information upon which a variety of applications can be built is standards-based metadata. Three projects within the CFI award are dedicated to producing a combination of DDI-compliant metadata and a set of software tools to exploit the data through mining its metadata. A conversion project will generate metadata for the complete data collection now in the RDCN and provide for the addition of new surveys over the next three years. The long-term sustainability of producing DDI

[4] One senior member of the Statistics Canada unit approving the use of lightpath technology said that he wished all Statistics Canada's external network connections were on this technology. For additional security, data encryption and decryption boxes will be placed between each Centre's connection to the national optical research network and the Federal RDC end-point.

metadata is now being addressed by several of the survey divisions providing the RDCN with microdata. For example, the Health Statistics Division, which is responsible for the National Population Health Survey and the Canadian Community Health Survey, is currently investigating ways of generating DDI metadata as part of their survey production cycle. They are also discovering how this metadata can be repurposed to drive other tasks within their production stream. For example, DDI metadata can generate a variety of reports, including drafts of questionnaires, data dictionaries or reports on data quality. It can also be used to share information with elements in the Integrated Metadata Database (IMDB), Statistics Canada's implementation of an ISO 11179 metadata registry.

In addition to the value metadata can add to the survey production cycle, it can also serve as standard input for a variety of discovery tools. The NESSTAR server is an example of a data extraction system driven by underlying DDI metadata. In the example of an extraction service for assessing disclosure risk, the metadata would be open to the research community outside an RDC and would thus be independent of the confidential data. The NESSTAR Data Publisher, for example, currently allows publishing the metadata on a NESSTAR server without the microdata. Two enhancements would have to be made to work with existing tools, however. First, the metadata would need to incorporate an element indicating risk level at the variable level. Secondly, a tool to assess the combination of risk measures for a subset of variables would need to be developed and integrated into the workflow of the RDCN.

## A User-driven Continuum of Risk Management

This approach would establish a service that determines initially the microdata a user wants to analyse, would assess the risk level for this request and would direct the user to the appropriate dissemination mechanism. This vision of managing disclosure risk is driven by the researcher's data needs. Using DDI metadata to explore all of the variables in the confidential and public use microdata versions of a survey, a researcher would browse and identify a set of variables and cases that she or he feels is needed in a research project.

This subset would be captured in DDI metadata and transmitted to a server where the risk level for the request would be assessed. Using the pre-assigned risk scores of individual variables, an overall score for disclosure risk would be determined and compared to thresholds developed from histories of disclosure analyses and from simulations.

The threshold within which the total risk score falls would be used to identify the dissemination mechanism for a project. If the variables fall within the public use threshold, the process would be expedited by sending the researcher a data dictionary in DDI metadata that would produce a subset directly from a public use data extractor. If the request requires data from a confidential microdata file but the level of disclosure risk is low for the entire subset, a data dictionary with the

DDI metadata would be filed with the RDCN for processing. The extract would be subsequently generated and the contractual terms for use sent with the data to the researcher.[5] Finally, if the risk score falls in the range requiring an RDC project, an application for a project would be generated from the metadata and sent to the researcher for submission approval, which initiates the application to work in an RDC. Once the approval is received, the researcher will arrive at an RDC with the extraction already generated from the initially requested subset.

This places the researcher in the driver's seat to identify the data she or he needs, while the agency responsible for safeguarding the data retains control over the actual dissemination mechanism. Fewer barriers exist at the start-up stage, which should increase overall data usage. For example, identifying the data needed in a research project and making a data request both would be simplified. The re-use of information streamlines the various stages leading to approval and determination of the appropriate access.

## Future Work

For the above model to become a reality, research is needed on both developing risk level measures for individual variables and ascertaining thresholds for overall scores. The assignment of risk levels to individual variables may be fairly straightforward. Some of this knowledge likely can be learned from current assessment strategies used by survey managers in preparing public use files for the Data Release Committee. For example, a variable containing postal codes would create a very high risk-score whereas a variable capturing whether a respondent can touch her or his toes may be assigned a low risk-score. The more difficult work will be in understanding risk given the combination of variables. Even with this, evidence is being created by every disclosure analysis currently conducted in an RDC. We just need to capture this evidence in a form that can be analysed.

One approach for building this evidence is under development. One of the CFI-funded projects will create the tools to generate metadata for the analysis files created by researchers in RDC's. The full life cycle record of an RDC project will ideally be captured in this metadata, including a summary of the disclosure analysis. As more disclosure analysis experiences are captured in metadata, this information can also be used to assess threshold levels of risk. The growing volume of metadata from projects will form a database of disclosure risk histories that can be processed.[6]

[5] The data would be encrypted but the key for decryption would likely be withheld until the contract and conditions of use were finalised. Once the paperwork is completed, the key would be sent to the researcher.

[6] Residual disclosure resulting from independent tables released from the same data but different projects or possibly the same project is a valid concern. With metadata describing the analytic files used by researchers and the results released through disclosure analysis, we may be getting valuable evidence that allows us to determine the actual risk of residual disclosure.

Several of the pieces required to move toward a user-driven continuum are in place now or are under development.  The few remaining parts can be built on the foundation already established.  Through metadata we can improve access to detailed microdata while safeguarding the confidentiality of respondents.