

Balancing the need for detail and confidentiality in the Canadian Census

Pierre A. Gauthier,
2006 Census Content Manager, Statistics Canada

Paper presented at the 2002 Population Census Conference in Ulaan Baatar, Mongolia

ABSTRACT: Most modern censuses now accumulate a wealth of data that can be used to inform program and policy-making at the national, regional and even municipal level. However, in order to be of benefit, this wealth of data must be analyzed extensively. The needs of the research community have evolved in recent years towards increasing levels of detail, small area data, and micro-data. As a result, statistical agencies are placed in the position of balancing the need for detailed data with the need for confidentiality-protection. Canada bases its Census data-dissemination program on five general principles: maximize the amount of relevant analysis; protect confidentiality as a highest priority; tailor data products to user groups; produce accurate, accessible, relevant and timely data; and apply disclosure control methods without unduly restricting analytical potential. Statistics Canada improves access to detailed data by providing more census data for small areas, by increasing access to detailed tabulations and by increasing access to micro-data. Canada's Census of Population uses two main methods of disclosure control in its tabular data, area suppression and random rounding. For its public-use micro-data files, it relies on data-reduction techniques. The 2001 Census Data Release approach features the Internet as the primary dissemination vehicle and data products designed to meet the needs of four major user groups.

1. Introduction

(1) Although a census is generally defined as a head count, modern censuses have become much more than that – they are now indispensable and central components of a nation's statistical system, often collecting detailed information on ethno-cultural characteristics, labour market activity, education and income. The collection of this detailed information as part of a census of population makes good sense from many perspectives. The additional costs of collecting detailed information are far less when this is done as part of a census activity than they would be from a survey, and the possibility of collecting data for very small geographical areas and for small sub-groups within the population is one that no other survey offers.

(2) As a result, most modern censuses now accumulate a wealth of data that can be used to inform program and policy-making at the national, regional and even municipal level. However, in order to truly be of benefit to a country, this wealth of data must be analyzed extensively and used for detailed, comprehensive research.

(3) In Canada, the Census of Population is a unique and rich source of socio-economic data, and Statistics Canada has introduced a multitude of ways of making it accessible to the research community. However, the needs of this community have also evolved in recent years. Armed with ever more powerful computers and software packages, researchers are requesting more detail than ever in their data. Increasingly, researchers want direct access to the census database, more complete micro-data files, and complex frequency tabulations with cells numbering in the tens of millions.

(4) Faced with these demands, a statistical agency is placed in a delicate position: balancing the need for detailed data with the need for protecting confidentiality. This paper examines this balancing act from the perspective of the Canadian Census of Population. It looks at some general principles that underlie the Census dissemination program, and some of the ways that Canada has addressed the needs of its many data users while guaranteeing the confidentiality of the data provided by respondents.

2. General Principles

The following are some of the general principles that guide Statistics Canada in designing its census data dissemination program:

a) Census dissemination programs should be designed to maximize the amount of relevant analysis that will be carried out using those data.

(5) The activities of Statistics Canada are founded on the premise that greater accessibility to well-grounded, objective statistical information on socially and economically relevant issues is of benefit to all Canadians. In complex societies, policy-makers and decision-makers need to make their decisions based on facts, and need to evaluate the impacts and effects of their decisions using objective, reliable information. Few would argue that sound statistical information is a prerequisite for social and economic progress. On the other hand, statistical inquiry is intrusive. It places a burden on respondents to take the time to provide the information requested, and it may sometimes be perceived as a violation of their privacy. A natural tension therefore accompanies statistical activities.

(6) Statistics Canada aims to conduct only those statistical inquiries where the benefits to the country and society as a whole are so great that they warrant the burden placed on respondents. We are fortunate in that a majority of Canadians agree with the judgements we have made in this regard, and are willing to comply with our ever-increasing demands from them. But we need to carefully monitor the total burden from all of our surveys, to ensure that we retain this privilege. Nothing would be more harmful to our statistical system than the loss of public confidence and good will.

(7) A key component of statistical inquiry is making the data collected available for analysis. Generally, it is researchers outside of the national statistical organization who turn raw numbers into useable information. Most census programs do include an analytical component, but this is complemented by the type of detailed, in-depth research that is typically conducted in academia, think-tanks, and in the policy departments of federal, regional and municipal governments.

b) The confidentiality of the census data collected should be a priority of the highest order.

(8) A national census is a natural, cost-effective way of gathering social and economic information from a very large sample or even from the entire population of a country. It makes good sense to use the opportunity of a census to collect a variety of useful and relevant data about a population's cultural, educational, family and other characteristics. However, this data collection activity (if it is to be repeated successfully every five or ten years) must be accompanied by a guarantee that the data collected will remain confidential.

(9) Statistics Canada is obligated by law to protect the confidentiality of respondents' information. Disclosure control measures are taken to protect the agency's data in such a way that these confidentiality requirements are not violated. The principles of disclosure control activities are governed, almost entirely, by the Statistics Act. The confidentiality provisions of the Statistics Act are extremely rigorous. The translation of their meaning to specific applications is a difficult but important task. The primary goal is to ensure that no identifiable person's data can be inferred to within a narrow range. Furthermore, it is necessary to protect information whether it concerns something likely to be considered sensitive (such as income) or not (such as basic demographic information). This is a principle which a census (as any statistical program) must adopt if it is to be successful in obtaining detailed socio-economic information from respondents.

c) Data products from the census should be tailored to different user communities.

(10) In Canada, the Census of Population maintains close working relationships with each of its major user groups. The users of census data are often thought of as falling into one of four groups: major users, occasional users, the general public and the media, and researchers. The needs of each of these groups have to be balanced in a census dissemination program.

(11) In the case of major users, the majority of whom are analysts in federal, provincial and municipal governments, the requirements tend to be for highly detailed frequency tabulations, often for small areas or particular sub-groups of the population. Occasional users, for the most part, want easy access to simpler tabulations and analytical pieces. The general public and the media want data presented in the form of a story, and data at the community level. Researchers, by contrast, generally want either public use micro-data files or access to the actual database. Statistics Canada has developed a dissemination program for census data that meets the needs of each of these user groups. This program is described in section 5 of this paper.

d) Census data must be accessible, affordable, relevant and timely.

(12) The cost and burden associated with Census data-collection can only be justified in the presence of a dissemination program that ensures accessibility, affordability, relevance and timeliness. These can mean different things to different user groups, depending on their data needs. The initial release of population counts can occur less than one year after data collection, while users of micro-data files often have to wait up to three years after Census Day before these are available. Statistics Canada's goal is to ensure that the data for public consumption are produced first, followed by planned tabulations, followed by customized tabulations and finally by micro-data files. In all cases, confidentiality protection is ensured, using a variety of means, which are described in section 4 of this paper.

(13) In order to ensure accessibility and relevance of its census data, Statistics Canada consults extensively with its clients on their emerging social and economic data needs. These consultations are designed to gather information from users on their current and longer-term needs, on the issues which they see as emerging, and on the types of products and services which they would find most useful to meet their data needs. The consultations are also designed to highlight the data sources at Statistics Canada that can be used to complement census data, such as longitudinal surveys, cross-sectional surveys, and administrative data.

e) Disclosure-prevention methods must be applied without unnecessarily restricting the potential for meaningful analysis of the data.

(14) There is an obvious tension between the desire to release as much data as possible and the need to protect confidentiality. In the case of both frequency tabulations and micro-data files, various methods of disclosure control are applied to the data before public release. However, these methods need to be applied judiciously, as they typically introduce error into the data, reducing its usefulness. The goal is thus to ensure that confidentiality protection provisions are met while preserving the usefulness of the data outputs to the greatest extent possible. A general principle is that the smallest amount of perturbation or suppression of data should be applied which still meets the overall objective of preventing the relation of any particulars provided to any identifiable individual.

3. Canadian Census approaches to improving access to detailed data

(15) Statistics Canada has responded to the increasing demand for detailed data in a variety of ways. The initiatives introduced over the past thirty or forty years have resulted in increased reliance on the census (as well as our survey and administrative data products) for effective decision-making. In developing our census dissemination program, we have always placed an emphasis on the protection of confidentiality. This section will describe some of the ways in which we have attempted to make more detailed data available, and the following section will describe some of the techniques we have implemented for disclosure control.

a) Provide more census data for small areas

(16) One of the major ways in which census data can be made more useful for detailed analysis is to provide census data for small areas. To enable a national census to produce accurate data for small areas, a geographical infrastructure of boundaries and mapping capacity covering the whole country is a prerequisite. Such an infrastructure requires that each dwelling be associated with a precise geographic location on the ground, where the degree of precision determines the fineness with which small areas can be defined. In Canada, dwellings are geographically coded to the block-face level in urban areas, and to the Census Block in rural areas. A fine level of geographical coding allows great flexibility for the production of tabulations for non-standard, user-defined areas (e.g., health districts), and it allows the creation of smaller standard areas for dissemination.

(17) In Canada, we have recently introduced two new standard geographical areas, the Dissemination Area and the Census Block. The Census Block is the smallest standard geographical area for which census data are released (although they are primarily intended to be used as building blocks for creating user-defined areas). Generally, the only data that are released at the Census Block level are population and dwelling counts. Dissemination Areas are also new in 2001, and are intended to replace the Enumeration Area (the area assigned to one enumerator at the time of data-collection) as the smallest standard geographical area. Dissemination Areas are designed to respect most legal boundaries (e.g. city limits) and to be stable from one census to the other, allowing for inter-censal comparisons. Dissemination Areas are composed of one or more Census Blocks and have a target population of 550 persons (although in very sparsely populated areas, this target cannot be met and the populations are much smaller).

(18) The availability of data for these very small geographical areas, combined with the availability of census and geography data in digital format, with sophisticated search, linkage and GIS tools, and with dissemination over the Internet greatly increase the usefulness of Census data. However, they also provide data users with the ability to manipulate census data in such a way as to increase the risk of disclosure of information for individuals or for small groups of individuals. This may precipitate a need for the introduction of additional disclosure control measures.

b) Increase access to detailed frequency tabulations

(19) Over the past twenty years, Canada's Census of Population has gradually moved away from data dissemination using paper publications to more electronic formats. Data are increasingly being provided over the Internet and made available on diskettes and CD-ROMs. In addition, for selected main users, the possibility exists of having indirect access to the main data-file through off-site terminals co-located at their workplace. These allow specially-trained members of those organizations to prepare their own customized tabulation programs, to have these run on the census base by Statistics Canada, and have the resulting data returned to them electronically (once these data have been screened for confidentiality). This process saves time and money for both the users and Statistics Canada. In addition, the elimination of the paper intermediary permits the production of far more detailed tabulations, often running into the tens of millions of cells.

(20) The 2001 Census will see a major change in that the Internet will be the primary means of data distribution. Statistics Canada's Internet site will feature a Census component where users will be able to download a wealth of information for free. These free data products include cross-tabulations, community profiles, thematic maps, and all relevant meta-data products. In addition, the Internet site is equipped with sophisticated search tools and print capabilities.

(21) Finally, the Canadian census offers a custom tabulation service, where users can define their own tabulations both in terms of geographical areas and the level of detail in the selected characteristics. It is not uncommon for users to request data for a very specific sub-group of the population, and to request a number of different customized geographies, created using block-faces as the building block. These tabulations are run directly against the master file, and are subjected to disclosure-control measures before being turned over to users. Statistics Canada provides this service on a full cost-recovery basis, and demand remains high. Custom tabulations provide the most detailed frequency tabulations, and many users request comprehensive series of tabulations, in order to cover a particular topic exhaustively.

c) Increase the access to micro-data

(22) Despite the usefulness of the geographical and mapping products and the detail of the electronic data and custom tabulations, many researchers, particularly in academia, require access to micro-data to fully explore the relationships between variables. The ability to run multiple regressions, ANOVAs and other statistical tests directly on the micro-data far outstrips the analytical potential of tabular data.

(23) Statistics Canada has been making Public Use Micro-data Files (PUMF) available from the Census of Population since the 1971 Census. There has been increasing pressure from researchers to increase the sample sizes of the PUMFs and even to give them direct access to

the master file. Statistics Canada has explored many models for increasing the access to micro-data.

(24) One successful initiative is known as the Data Liberation Initiative (DLI), wherein a consortium of Canadian Universities no longer need to purchase their Census data file by file. Instead, participating universities pay an annual subscription fee that allows their faculty and students unlimited access to DLI micro-data, databases, and meta-data files. This allows these universities, for the first time, to offer a full range of data services to students. There is evidence that the DLI is making important contributions to Canadian teaching and research.

(25) Another recent initiative is the creation of Research Data Centers, which are secure Statistics Canada facilities located across the country, where micro-data files are housed. These centers are staffed by Statistics Canada employees and they operate under the provisions of the Statistics Act. Access to the centers is limited to researchers who have projects that have met strict criteria and received approval, and who have taken an oath of confidentiality under the Statistics Act. They conduct their research in the centers as “deemed employees”, subject to the same penalties as all Statistics Canada employees. These researchers have direct access to the micro-data only for the purpose of conducting research. Their output files are subjected to disclosure control procedures and are verified by a Statistics Canada employee before they leave the data center. This approach, which has so far been limited to sample survey data, is being considered for the 2001 Census.

(26) A last option for increasing the access to micro-data is remote access. In this case, researchers submit a request for their planned analysis to Statistics Canada. Those who are granted access are provided with documentation and access to a non-confidential dummy database to test their analysis programs. They then submit these programs (usually written in SPSS or SAS) electronically to Statistics Canada. These programs are then run directly against the master micro-data files and the output is subjected to disclosure control procedures by Statistics Canada staff. Once the outputs are deemed to be acceptable for public release, they are sent electronically back to the researchers. This option is not currently available for Census data, but it is under discussion.

4. Disclosure control

(27) Like most national statistical organizations, Statistics Canada has been actively exploring the issue of disclosure control over the past decade, largely due to the increased demand for highly-detailed data and the increased use of new tools and techniques for data-mining, linkage and geographical coding. Several task forces have been created to examine the methods currently in use and to suggest options for improving disclosure control. The conclusion seems to be that disclosure control in this environment poses many challenges, which will only increase as more data are released electronically and data are made accessible over the Internet. Assessing the risk of disclosure is often subjective, and there is no accepted, accurate way of measuring the risk of disclosure, nor of controlling the risk of residual disclosure. The methods of disclosure control are often heuristic, to some extent, relying on common sense and good judgment rather than mathematical certainty.

(28) Other considerations must also be taken into account. There is now increasing interest in the concept of group privacy, where no confidential data about any individual have been released, but enough data have been released for a small and identifiable group of people to

allow the inference of characteristics for members of that group with a high probability. In these cases, members of that group may believe they could suffer a disadvantage by the release of such information.

(29) Although no formal universally acceptable theory for risk of disclosure has been put into practice, a number of methods have evolved over time to deal with the practicalities of ensuring that all reasonable controls are in place. Since the Census of Population has the potential to produce estimates for very small areas and sub-populations, several rules to protect against direct and residual disclosure have been put in place for both tabular data and micro-data. The Census of Population currently uses two major classes of disclosure control techniques for its tabular data, suppression and rounding. The major problems in frequency tabulations surrounds very low frequency cells, where individuals can potentially be identified, and cells with a count of zero (where knowing that no individual in some identifiable sub-population has a particular characteristic may constitute disclosure). Both of these problems are addressed by suppression and rounding.

(30) The Census of Population applies extensive suppression and rounding rules to all data tabular being released, including custom tabulations. These include:

- extensive random rounding procedures to base 5 or base 10, depending on frequencies;
- suppression of areas below a given population threshold;
- suppression rules specific to statistical calculations (means, medians, rates, etc.); and
- informal checking of custom tabulation requests to identify potential for residual disclosure.

For its Public Use Micro-data Files, the Census of Population relies mostly on data reduction techniques to prevent disclosure. The data reduction techniques include sampling, ensuring that the populations for certain identifiable groups are sufficiently large, making the variable categories coarser, and top and bottom coding.

(31) The topic of finding better ways to protect micro-data files is an active research area, and Statistics Canada is looking for ways to improve disclosure control while still allowing for appropriate analysis. One area that has been used successfully in several countries is that of data perturbation, with techniques that include data-swapping and the addition of random “noise”. These techniques, in theory, preserve the integrity of distributions while adding an element of uncertainty into census characteristics.

5. The 2001 Census Releases

(32) As mentioned above, one of the key principles of the Canadian Census dissemination strategy is the provision of data products that are targeted to the needs of our different user groups. For our more sophisticated users, we will offer highly detailed and complex tabulations of data and micro-data files. These products are very costly and time-consuming to produce, and so they tend to be made available only some time after the data have been publicly released.

(33) For the 2001 Census, much thought has gone into what data to make available at the time of the data release, when public and media interest are highest. The data releases offer

an excellent opportunity to generate public discussion around the census results, increasing the visibility and the profile of the Census of Population. It also increases public support for the census, which can be leveraged at the data-collection phase of the next census, through the public communications program.

(34) For the past several censuses, the data have been released in waves over a period of about one year, starting about one year after Census Day. Each release surrounds a grouping of census characteristics, such as families, cultural characteristics or labour force activity. In the past, these tended to focus on the major highlights in the data being released, but without much of a theme or thread tying the data together.

(35) The 2001 Census is trying a different approach to data release. The challenge for the census is to identify one or more story lines around the census data which can be picked up by the media and be of interest to Canadians, but to do so before the actual data are available so that there is sufficient time to prepare the release tables and materials. The solution to this challenge for 2001 centers around the identification of interesting and relevant themes beforehand, based on the analysis of data from previous censuses, data from other sources (e.g., surveys and administrative databases) and from analytical and research papers published over the past few years. Once these themes are identified, the skeleton of the release is put in place, and it awaits the availability of the actual data to populate the tabulations and analytical articles shortly before release.

(36) The 2001 Census dissemination program features the use of the Internet as the primary delivery vehicle for standard data products. It will also feature the availability of more data on the day of release than ever before. This implies that all tabulations, maps, meta-data and analytical pieces must be ready and loaded onto the Internet site by the day of release (Note: Statistics Canada pre-announces all of its data releases and these dates are not subject to change). For 2001, a total of 21 themes have been pre-identified and grouped into 8 major releases. This approach places pressure on staff to have everything ready sooner than ever before, but judging by the huge number of visitors to our Internet site at the time of our release of population counts in March, it is definitely worthwhile.

(37) In addition to the data available on day of release, there will be many tabulations and meta-data products distributed via the Internet in the 2001 Census dissemination program. While many tabulations and community profiles will be made available to all Canadians on our Internet site, the size of these must be kept quite small due to the nature of the existing technology. It is felt, however, that these simpler products adequately meet the needs of the general public and the media. Census Profiles, in particular, have proven popular and useful, presenting summary information about the characteristics of the population of an area, based on a large number of detailed variables. Groups of variables will be released in each of the eight major releases, which when combined will form a complete profile.

(38) For some users, there will be an additional level of access available on the Internet, which will allow them to download more complex tabulations and community profiles, and to have greater flexibility in specifying the geographies for which the data are presented. This level of access will be given to data users in various levels of government, to universities who are members of the Data Liberation Initiative, and to Depository Libraries and our Regional Reference Centers, so that they can respond to requests for more specific information.

(39) For our most sophisticated users, data will be provided through our custom tabulation service and through our Public Use Micro-data File. Various modes of remote access are also being explored. These users can specify their levels of geography with almost unlimited flexibility and can cross-tabulate almost any combinations of characteristics they wish (subject to random-rounding and data-suppression).

6. Conclusion

(40) The technological advances of the past forty years have brought with them a phenomenal increase in the capacity for research and data analysis. For Statistical organizations and census bureaus, this means that more and better use of our data will be made than ever before. It also means more media coverage, more public interest, and more use of our data by decision-makers at every level. It means increased demand for our data, and more requests for small area data, for access to micro-data, and for highly complex tabulations. On the whole, these are very positive developments.

(41) However, this increased demand does come at a price: increased vigilance. Statistical organizations must conduct research into risks of disclosure and explore new ways of guaranteeing that they can protect the confidentiality of the data they have collected. In most countries, respondents are obligated by law to complete their census forms. In return they receive a promise that every effort will be made to safeguard their information and protect against the identification of that information with any individual. Should a country fail to keep that promise, they would surely lose the confidence of the responding public, making it all the more difficult to collect census information in the future.

(42) Statistical organizations can and must continue to provide data to researchers in ways that allow meaningful and relevant analysis, but they must do so in ways that do not jeopardize confidentiality. This is a challenge that will not go away, and that will in all likelihood become exacerbated as data-mining and record-linkage tools become more powerful. Research into methods of estimating risks of disclosure and preventing should be a priority for all national statistical organizations throughout the 21st Century.

REFERENCES

Brackstone, Gordon (2001). Strategies and approaches for small area statistics. Paper prepared by Statistics Canada for the Conference of European Statisticians, Plenary Session, June 2001

Statistics Canada, 2001 Census Preview of Products and Services, Catalogue No. 92-376-XPB

Statistics Canada , Quality Guidelines, 3rd Edition (1998), Catalogue No. 12-539-XIE

Tambay, Jean-Louis, J. Burgess and P. White (2001). Small Area Data and Confidentiality. Paper presented to the Statistics Canada Senior Management Conference, November 2001.