# DATA BASES FOR DATA USERS

Edmund M. Murphy, Chief, Development, Regional Statistics Staff Statistics Canada

This paper will concentrate upon the output side, the dissemination side of data base construction using examples drawn from the dissemination programme designed for the data base created from the 1971 Canadian Census of Population and Housing.

One may conceive of three major steps in constructing a data base:

a. collecting or assembling the data
b. processing the data to create the data base itself
c. drawing information from the base and placing it in the user's hands.

A dissemination programme must consider all three phases. The technology of the third phase, the publication phase, has the most apparent effect upon dissemination; however, unless the user is given adequate information about the two preceeding stages, the dissemination programme may become an exercize in user deception rather than user education.

At every step of the collection and processing phases assumptions are made, approximations used, shortcuts taken that can have an important effect upon the quality of the data. The user must be fully aware of the impact of these procedures lest he mislead himself and those who use his analyses.

1. The Collection Stage.

It is a truism that the first step in using a data set should be to read the instruction manual for the enumerators who collected the data. Apart from such utopian and perhaps unfair demands on the user, it is a responsibility of a dissemination programme to give the user a description of the collection procedures as they affect the data.

The presentation of such information should take two forms. First, a general discussion of methods and procedures, and second, a discussion, where necessary, of the impact of collection procedures on specific data sets.

For the 1971 Canadian Census the major methodological innovations were self-enumeration (drop-off, mail-back) and the extensive use of sampling. Preliminary processing procedures that have an impact on the data include acceptance standards for incomplete questionnaires, call back procedures, plus regional office coding of occupation, industry and such open ended questions as ethnic group and language.

Describing the impact of these procedures is no simple task. A general description of the procedures presents no problem, but, aside from estimates of sampling error, it is difficult to go beyond cautionary notes when discussing the effect of specific procedures upon specific data items. What is needed is a quantitative, item specific estimate of the reliability of the data.

2. The Processing Stage

Between the collection stage and the creation of the data bank, the information is subjected to a series of processing steps that have an additional effect upon the data. Again a dissemination programme must include a description of these procedures and an estimate of their impact upon specific data items.

In addition to the regional office processing described with the collection procedures, the major steps in processing the 1971 Canadian Census data were the editing and the imputation programmes and the weighting procedures applied to the sample data.

In general weighting procedures and their effect upon reliability are relatively easy to describe and to quantify. The weighting procedures flow from the sample design, and their impact can be presented as part of the estimates of sampling error.

The 1971 Canadian Census used a ratio ranking procedure for weighting. For private dwellings, enumeration areas were aggregated to weighting areas of roughly 4,000 population. Weights were assigned so that for each weighting area sample estimates and complete counts matched on certain key items common to both the sample and complete count questionnaires.

Presenting to the user a summary of the editing and imputation programmes and their impact upon the reliability of the data is not so simple. Stating the philosophy of these programmes is easy enough: to remove obvious errors and inconsistencies, on the one hand, and to make meaningful estimates of missing data on the other. However, to indicate the operational steps involved is a problem of another order of magnitude. A list of the edit specifications for the 1971 Census is itself an impressively thick volume.

Again the question is one of balance. The user must not be swamped with indigestable masses of information, but he must be given enough information about these procedures to enable him to properly interpret his data and the results of his analyses.

For the editing programme general lines of strategy could be indicated giving the level of priority for each type of data. For example, one could indicate, perhaps in a flow chart, that age is checked against other ages in the household, that marital status is checked against age as is educational attainment, etc.

For the imputation programme one can indicate the general rules. In the 1971 Canadian

Census there were two types of imputation procedures. The first used deterministic rules, a husband's age could be estimated from the age of his wife, for example. The second type was the well known "hot deck" procedure where imputation groups are defined by a set of basic variables and missing data for a record are imputed from the previous complete record falling in the same imputation group.

The user must have a firm idea of the rules of these procedures. Moreover, he must also have a quantitative, item specific estimate of their effect. This latter information is even more difficult to convey than the former. Not only does the impact of these procedures vary from data item to item but from area to area.

These variations by area are especially critical in Census data, which is a major source of data for small areas. The 1971 Census will publish a great deal of data for enumeration areas, which have a population of about 400 to 600. The impact of the editing and imputation programmes can vary significantly from EA to EA, and the user should be warned when he is using an EA that is largely hypothetical. The same problems exist to a lesser degree for larger areas.

3. The 1971 Census Master File and Data Dictionary.

The purpose of these collection and processing steps is to create the data base. For the 1971 Census data this base is called the Census master file; from the user's point of view this file <u>is</u> the Census data. It is from these data as coded, edited, imputed and weighted that all 1971 Census tables must be drawn. To analyse a Census table or to design a special tabulation from the data, the user must understand the data on the file. He must know what data are recorded on the file, what categories are used to record the data, for what subpopulations the data are reported, and what are the strengths and weaknesses of the various data items.

For 1971 the Census is producing a document that contains much of this information — the Census Data Dictionary. In this document the user will find a definition of each data item recorded on the master file, the categories used in coding the responses, which of the last three Censuses contained this data item, differences in definition between these Censuses, the sampling fraction where applicable, the population for which the data are reported, and finally, a remarks section pointing out any problems in interpreting and using the data.

By reading this Dictionary the user can discover for example which ethnic groups are coded on the file and by implication which are grouped as "others"; he can see that the labor force status information is reported only for those 15 and over; that age data is available up to an open ended group 100 and over, etc. He can also discover that no information exists in the Census data base on health.

This Census dictionary is a major start towards providing the user with the information he needs about this data base. The introduction to the dictionary provides the needed overview of the Census collection methodology and processing procedures. It does lack, however, at this stage of development, two of the elements noted above as being essential to a full description of a data base. There is no discussion of the detailed strategy used in the editing and imputation steps, nor is there any indication in the dictionary itself of the reliability of the data for particular items or areas.

4. Root Mean Square Error

On another front, Statistics Canada has also made progress in indicating to the user the reliability of the data. After the evaluation studies of the 1971 Census are complete, the information obtained will be summarized in one measure, the Root Mean Square Error. This combines in one measure both sampling error and non-sampling error. The measure is constructed so that it can be used, in general, like a sampling variance to estimate confidence limits for a figure of a given magnitude.

When this information is added to the Census Data Dictionary as part of a section in each definition on "Reliability", the Census Data Dictionary will come close to being a model for a description of a data base. Even in its present form it is valuable aid to any user in understanding the 1971 Census data base and the tables that are drawn from it.

5. The Dissemination Programme

The first stage in the dissemination programme is to describe to the user what data are available, to give him an understanding of the data base. The next step is to give him some data.

5.1. The Media

Before discussing the various strategies for placing data in a user's hands, a brief discussion of the media of presentation is desirable. There are, in general, three types of media applicable to data transmission — printed publications, optical devices, such as microfilm, and computer readable storage such as magnetic tape. As one reads down this list the media become increasingly more expensive to read, to peruse, and increasingly less expensive to store and manipulate. Printed reports are bulky, difficult to store and the data must be later transferred to machine readable form for manipulation. On the other hand, one does not have to sell his patrimony to buy equipment to read a book.

At the other extreme, data on computer readable media can be introduced directly into a

machine, but one needs a computer to actually see the material and documentation and compatability problems often exist. Optical devices such as film are easy to store and reading devices are at least cheaper than a computer, but again the data must be transferred to another medium for analysis.

The 1971 Census data will be available in all three types of media, the standard printed reports, microfilm rolls and microfiche, and user summary tape.

Since printed materials are very easy to peruse but difficult to store, the evident strategy is to provide summary information in printed form and to use film or tape for transmittal of detailed data. The 1971 Census tabulation programme takes a step in this direction. The preplanned tabulation programme contains some 3,500 tables, about 3.8 million pages of computer printout. A small proportion of these will be published in printed reports. The rest of the tables, except for those produced only for check-out purposes and internal analysis, will be available on microform or as part of the summary tape programme.

5.2. Data to the user

After having described the data base, and insured a range choice of publication medium. The next step is to decide what information is to be transmitted. A problem that often arises is that of confidentiality. For most data bases, it is usually contrary to public law, professional ethics, or common decency to release to the public data that can be related to an identifiable individual unit. One is thus left with three types of publication:

a. Publication of summary data as in tabulations
b. Publication of individual records adjusted so that individuals cannot be identified
c. In-house manipulation of individual records and publication of the results.

For the 1971 Census data, Statistics Canada is considering options b. and c., but the major thrust of the dissemination programme is upon a large and flexible tabulation programme. Such tabulation programmes have two major components, a set of preplanned tables plus systems for permitting the user to design and obtain tables to his own specifications.

5.2.1. The preplanned tabulations

No matter how flexible and efficient the special request tabulation system may be, it is still cheaper and quicker for a user to obtain an "off the shelf" table when he can find one that suits his needs. The twin problems for a preplanned tabulation programme are to design the tables so that a large proportion of user needs will be met and to inform the user as to what tables are available.

The proper design of the programme should involve at least two elements: professional judgement of what would be useful and desirable, plus the equivalent of market research to ascertain the efficiency of previous tabulation programmes and to estimate changing demand for tabulations.

The 1971 Census tabulation programme emphasized the first aspect, professional judgement. The market research consisted largely of an analysis of the special requests generated by the earlier Censuses. Requests for special tabulation are one important source or information about the effectiveness of a preplanned programme, and this type of research will be continued and expanded for the 1971 special requests. In addition, the Census Division is setting up a special unit to conduct market research in the broader sense as an input to the 1976 and 1981 tabulation programmes.

If the preplanned tabulation programme is large, as it is for the 1971 Census, the problem of informing the user of exactly what is available assumes major proportions. The 1971 Census approaches this problem in two ways. The published tables are described in the traditional Census catalogue, and the entire preplanned programme is listed in a new document, the Tabulation Directory.

This directory is the first step toward an automated search system. Each table in the preplanned programme is completely described on an entry in a computer file. This file could be made available on tape to a user who could then devise a programme to search it, or he could list the file or parts of it in a manner that suits him. In addition the Census Division will make available a listing of the file. This listing will be cross sorted for ease of access. For example, all tables containing data on a given variable will be listed in one place cross-referenced by the other variables in the table. A user interested in a table showing age by sex by occupation can turn at once to a given section of the directory and find a list of all tables containing these variables.

5.2.2. Special Request Tabulations

No preplanned tabulation programme could or should meet all requirements for data from a given base. If every need is being satisfied, then the preplanned programme is probably too large, some of the tables are being used by a minute number of people. Thus every dissemination programme must provide for the timely and flexible production of special request, user defined, tabulations.

The basic element of such a system is a description of what tables could be produced from the given data base. The description of the data base discussed above should serve this purpose. For the 1971 Census, the Data Dictionary does give the user the information he needs to design tables that can be produced from the data base.

The next step is to develop a system that will deliver the requested tables quickly and cheaply. Since internal priorities and staff shortages are responsible for many delays in servicing special requests, the system should minimize the necessity for in-house staff intervention.

The need for in-house experts to advise and assist users can never be completely eliminated; however, the system should permit the user who is so inclined to accomplish himself the work necessary to generate a special tabulation. Ideally, there would be direct interaction between the user and the data base; the user types in a request on a terminal in his office and a second later the table appears. (The same kind of interaction is also a goal of the 3rd dissemination alternative mentioned above, in-house analysis of individual records and publication of results.)

There are several problems to be resolved before such direct user interaction is a reality, but the 1971 Census has made a major step in this direction with its STATPAK tabulation system. The TARELA language developed for the STATPAK programme is close enough to ordinary notation that the user can after some study essentially write the control cards for the programme himself. This system is new, and it has not yet been tested on real census data with real users, but it is a very promising step. In theory at least, the user can design the table, write the control cards, and send the request to the Census Division. His request will be checked, the control cards punched, the table run, and the results returned to the user with the obvious gains in overall turnaround time and costs.

A system such as this imposes an obvious burden upon the user. He must take the trouble to decide exactly what he wants, check its availability, write the control cards, and accept responsibility for the results. It is easier to write a note asking for "something by age and sex and occupation for the city of Sorel". However, with an interactive special request system, the user will have ample rewards for his efforts in cheap and timely answering of his request for special tabulations.

Another requirement for a special request programme is flexibility, especially in delineation of areas. The 1971 Census dissemination programme will give the user a great deal of flexibility in areal delination with its Geocoding programme. This is a computer system that codes all Census data to small units — block faces in 14 urban areas, Enumeration Areas for the rest of the country — and permits the user to construct his own areas as aggregations of these units simply by drawing his area on a map.

5.2.3. Confidentiality

Even with summary data, problems of confidentiality of individual records arise. It is often quite easy to pick out individuals in a tabulation and deduce information about them. One Census example is a table showing that the only Doctor in a certain small area has a certain income. In addition to these direct disclosures, there exists the problem of residual disclosure, obtaining information by combining two or more tables which themselves contain no direct disclosures. And finally, many data bases are plagued with problems of dominance, one unit's data dominating the data for a given area.

Confidentiality is a severe problem for any data base and especially one with direct user access. The procedures used to protect against disclosure must work, they must be automatic, and they must not distort the analytical meaning of the data.

For the 1971 Census data, which has problems with direct and residual disclosure but rarely with dominance, a random rounding procedure has been adopted. This procedure essentially shifts the unit of census tables from one to five.

The procedure gives good protection against disclosure; no number smaller than five appears in any table nor can a number smaller than five be derived by manipulating tables. It is automatic; a computer algorithm rounds each figure to one of the two nearest multiples of five using a random number generator and a schedule of probabilities. And it has little impact upon the analytical meaning of the data; only very small numbers are significantly changed, and these are very unreliable in census data in any case.

This, procedure will not be appropriate for, all data bases, but any data base which incorporates direct user interaction or rapid in-house preparation of special requests will need confidentiality procedures with similar properties: adequate safeguard, automatic application and minimal impact upon analytical meaning.

6. Summary

The construction of a data base has one goal: to place in the hands of analysts and policy makers the information they need to do their job. The major tasks of a dissemination programme are to provide the potential user with an adequate description of the data base including the procedures by which it was constructed, to plan a series of standard data packages to meet the needs of the general user, and to provide quick, accurate, and cheap special request data for users with specialized needs.

The 1971 Canadian Census will provide such a data base by means of its expanded summary tape and microfilm publication programme, its Statpack and Geocoding special request systems, and its expanded data documentation.

(Further information on the 1971 Canadian Census Data Dissemination programme may be obtained from:

Census User Enquiry Service, Census Division
Statistics Canada, Ottawa K1A 0T7)