A STRATEGY FOR ARCHIVING GOVERNMENT DATA TO MEET THE NEEDS OF THE RESEARCH COMMUNITY

Dr. Jake Knoppers
Senior Advisor (Information Management)
Public Archives of Canada

BACKGROUND *

Government Administration and Survey Data as a Research Resource

By far the greatest collector and user of information is the government. The largest portion of the data collected is for administrative purposes, e.g. collection of taxes; distribution of socio-economic benefits; regulation of industry, trade and commerce, etc. As a result of these activities, the government builds and maintains enormous stores of information or data banks. This information is collected continuously albeit with some periodicity, e.g. annual tax returns, monthly filings.

In addition to the information which is collected pursuant to Acts of Parliament and initiatives of the government, the various departments also carry out innumerable surveys, or commission surveys and opinion polls to be carried out by third parties.

Altogether, Statistics Canada has identified 2,352 major data banks so far, including both manual and machine readable. (1) Some data relates to individuals while other data refers to larger aggregates such as firms, households and national accounts. The number of subjects in each data bank ranges from as few as 200 to as high as 21,500,000.

As a matter of fact, there is a growing recognition that the various departments of government may have gone too far in their myriad data collection activities. The government has responded to these concerns through the "reduction of paperburden" initiative. This finds expression in such management tools as forms control, data collection approval mechanisms, and data locator systems.

Perhaps the desire to reduce paperburden, while meeting the needs of the research community, can be dealt with more effectively if one institutes a new strategy for archiving these government data in a rational and consistent manner.

The Needs of the Research Community

Because of the nature of the research scientist's work, his concern for access to government records is both highly selective and strongly motivated. Neither the research scientist nor an informed public can rely solely on the summary tables produced by government departments and statistical agencies. Nor do such tables of aggregated statistics take into consideration the concerns of all scholars. It must be possible for the public and the research community to obtain the raw micro-data, conduct independent analyses, and draw their own conclusions.

Since the major portion of the data collected by government for administrative purposes involves privacy considerations, ways must be found to balance the freedom to do research against the individual's (including legal persons') right to privacy. This principle is stated quite succinctly in a resolution of the Social Science Federation of Canada, passed at their May 17, 1979 meeting. The resolution reads:

> "There are socially significant fields of research for which access to personal records is indispensable. There is, therefore, a need to use personal data held by government agencies, for statistical and research purposes, in order to promote scientific understanding of important contemporary problems. This use of government data is not incompatible with the need to protect the privacy of individuals. Therefore, any federal or provincial laws for the protection of personal privacy or for access to government documents should make a clear distinction between administrative or regulatory uses of personal information, which directly affect a person, and statistical or research uses, which do not, and should explicitly recognize the legitimacy of using personal data for statistical or research purposes. Accordingly, provisions in these laws should set out the right of researchers to obtain access to personal data under specified conditions, and should specify these conditions, the most important being a written undertaking not to reveal data on specific individuals without their express consent.

* The opinions expressed in this paper are those of the author, and are not intended to reflect the position of the Public Archives of Canada, for which he prepared it as a consultant, nor that of any other Canadian government department or agency.

> Such laws should also provide, in case such access is refused, a right of appeal to an independent authority, such as an ombudsman, or a court, or preferably both." (2)

Finally, the data needs of the research community and those of the government as well would be met if it were possible to develop a longitudinal, rational, and integrated master sampling frame of government administrative records. The creation of such a mechanism and entity would also lead to a high probability of reduction of paperburden and extraneous survey data collection.

The Role and Responsibility of the Public Archives of Canada

The broad mandate of the Public Archives of Canada is to collect public (i.e., government) records, documents and other historical material of every description which reflect Canadian society from its beginnings to the present day. The Public Archives has the overall control responsibility for the "life-cycle" of all records, once their creation has been approved. The components of the "life-cycle" of records include:

> Identification, registration and classification; care and custody; controlled circulation; the establishment of retention and disposal schedules; and final disposal through destruction or permanent retention.
>
> Records include all diferent kinds of physical storage media for information, e.g., paper, electronic, film, etc.

It is the responsibility of the Dominion Archivist to identify and appraise all government records as to their historical or research value, granting permission for the destruction of those records which are of no archival value and thus not eligible for permanent retention. The gathering and preserving of archival records is done with the purpose of making them available to the government, researcher and the public.

The role and responsibility of the Public Archives to the research community is that the desired micro-data found in government administrative records and surveys is preserved for use. An Advisory Council on Public Records in which the social sciences are represented provides a vehicle whereby the research community can make its wishes known. (3)

THE POLITICAL AND LEGAL ENVIRONMENT

Present Legislation and Public Attitudes

One of the rights which has been recognized in law in many countries is the right to information privacy, i.e., that individuals have the right to be informed about the storage of personal data about themselves and to control the collection, use, and dissemination of such information. This definition is grounded on the conviction that, in the end, information about individuals belongs to the individuals themselves — indeed, such information is the essence of individuality. While recognizing that an individual divulges information about himself for a purpose — in exchange for a good, service or benefit, or as required under law — this approach holds that the information is nonetheless his, and that he retains rights with respect to it.

In general, with respect to the operations of government, privacy legislation is designed to enable individuals to control how, when, and to what extent information about them is communicated to others—especially where such communications are related to the decision-making or administrative processes which affect them personally. Consequently, privacy legislation normally contains provisions designed to control national or federal data banks, with controls on collection, storage, dissemination, retention and corrections of personal information. In many countries, the exercise of the rights under privacy is assisted through requirements to list or report on all personal data collection activities. The emphasis in many countries is on the surveillance of *computerized* rather than manually recorded information. Much of the pressure for greater openness in government has been relieved by legislation on privacy, even though this aspect has been treated separately from disclosure of government information of the more general kind.

In Canada, privacy legislation finds its expression in Part IV of the Canadian Human Rights Act. It requires that after March 1, 1980 the government shall inform data sources who knowingly provide information to a government institution, during the course of collection, the purpose and use to which the information will be put. (4) This rider applies specifically to administrative uses of information for decision-making purposes which impact directly on the individual concerned.

Apart from privacy concerns, many other Acts of Parliament set stringent conditions on access to their records. Some Acts state specifically which officers or which other Departments can have access to their data. Most of these access restrictions for administrative purposes are presently also used to deny access for research purposes and in some instances even to deny access to archivists when the latter wish to appraise the historical or research value of records.

Required Legal Environment

Before outlining some of the basic legal principles which Parliament should embrace through legislation, the research community would do well to take note of a comment in the decision of the U.S. Supreme Court in the "Kissinger Case." (5) The Justices quoting the Senate Report to the Federal Records Act of 1950 highlighted that:

> "It is well to emphasize that records come into existence, or should do so, not in order to fill filing cabinets or occupy floor space, or even to satisfy archival needs of this and future generations, [one would assume that this means the needs of the research community] but first of all to serve the administrative and executive purposes of the organization that creates them. There is a danger of this simple, self-evident fact being lost for lack of emphasis..." (6)

The statement applies universally to administrative records of all organizations, including those in the public sector. Changes in the legal environment which would assist in meeting the needs of the research community must at the least be compatible with administrative requirements. Fortunately, a legal environment which stresses efficient and cost-effective management of recorded information will also, as we shall see, encourage the making available of administrative data for research purposes.

The legal environment required to allow for access to government administrative and survey data should embody the following concepts:

- that no government record may be destroyed or altered in any form without proper authority;
- that the government be able to identify, inventory, and describe all their information holdings;
- that the government be able to identify and know the information contents of all its data collection activities;
- that institutions of government be permitted to disclose government records under their control to any scholar or research institution for research purposes; such records to be in both identifiable and anonymized microdata and in original and complete form without legal barrier or statutory exemptions;
- that the government accepts the principle that the security sensitivity of classified records declines with the passage of time and that declassification schemes are established and implemented;
- that for all government records a "life-cycle" is established and adhered to (i.e., records retention and disposal schedules);
- that the research community apart from their rights as individuals under privacy and as individuals or groups under freedom of information be granted the right to advise the government in its decision-making process pertaining to the disposal (i.e. destruction or permanent retention) of government records; and
- that the Dominion Archivist can effectively declare as archival any government record (or a copy), to be kept permanently for historical or research purposes.

On the researcher's side a comparable legal environment must also be created. Such an environment would be based on the principles of ethics in research which *inter alia* would also include the right of privacy and protection of the individual where personal information is involved. Researchers might want to consult government records that are not normally accessible to the public. One would expect that when access to such information is granted for legitimate research purposes, it would be granted only:

- if the research subject has consented to the intended use;

or

- if the vested interests of the research subject are not harmed or involved because of the type of information, general public knowledge of the information, or the type of data processing involved; or
- if the researcher agrees in writing not to reveal, publish or otherwise disclose information which would make it possible to identify any individual person, business, or organization except "X" years after the birth of an individual, or when the individual is dead or "X" years after death, or "X" years after the taking of a census or survey, or "X" years after the receipt of information relating to a business, institution or organization outside of government.

The reason for the *"X" years* is that these are policy decisions for the government to make, hopefully in consultation with the research community. However, the written undertaking or contract between the researcher and the government not to reveal information in any form that could reasonably be expected to identify the research subject, i.e., guaranteeing anonymity, would cover over 90% of the research needs.

Should an individual researcher not agree with such conditions, one would expect him to raise the question of public access (as distinguished from access for research purposes under specified conditions) either through the path open to him under freedom of information legislation or through the consultative channels between the research community and the government.

The above is an outline of the basic political-legal environment for government records which would ensure that both administrative and research needs can be met. The reason for the statements made in this section will become apparent in the model solution.

A MODEL SOLUTION

Basic Approach

The public at large benefits from legitimate uses of administrative data for research and analysis by government, businesses, non-profit organizations, academics, etc. Such data is used to analyze the effectiveness of the delivery of existing socio-economic programs; to study the causes of disease, poverty, crime, or migration; and to discern trends in society which may be of interest to the nation as a whole, to particular interest groups, or to the individual researcher. Yet the public is also concerned about the increasing burden of providing the required information ("paperburden") and about the real or imagined possibility of the misuse of the data provided by them.

Further, at the moment, the legislation affecting the use of administrative data for research purposes is far from clear and on the whole presents a negative approach. Instead of dealing

with all these laws and regulations on a case-by-case basis, the model solution proposed here presents a global approach which nevertheless should be applicable on the detailed level. The success of the approach taken here depends, however, on the establishment *a priori* of the following overall requirements for government records.

- *THE GOVERNMENT MUST ESTABLISH THE INFORMATION IN ITS POSSESSION (PHYSICALLY LOCATED IN ONE OF ITS INSTITUTIONS) OVER WHICH IT HAS OWNERSHIP OF THE CROWN IN RIGHT OF CANADA AND THE CONDITION OF SUCH OWNERSHIP*. This will also be a requirement for information falling under FOI, Privacy and Archives legislation, and is of particular relevance to information created as part of joint ventures of the government, e.g., federal-provincial, as well as for information created or collected as a result of the contracting out of work using public funds.

- *NO GOVERNMENT RECORD MAY BE DESTROYED OR ALTERED IN ANY FORM WITHOUT THE CONSENT OF THE DOMINION ARCHIVIST OR HIS DESIGNATE, AND IMPROPER DESTRUCTION MUST CALL FOR AUTOMATIC PENALTIES*. Through the establishment of authorized records retention and disposal schedules, the Dominion Archivist ensures efficient and cost-effective management of government records, destroying those records of no further use or value and selecting for permanent retention those records having historical or research value.

- *THE DOMINION ARCHIVIST MUST HAVE THE RIGHT TO DECLARE ANY GOVERNMENT RECORD OR A COPY THEREOF AN ARCHIVAL RECORD AND WHERE A COPY IS CONCERNED TAKE EARLY POSSESSION WHERE EFFICIENCY OR THE NATURE OF THE RECORD DICTATES*. The greater proportion of administrative data are created as a result of ongoing programs. Such data ae generically known as case files, i.e., files containing records related to specific repeatable actions, events, persons, organizations, products, objects, etc. and which are usually filed and retrieved by name, number or any other systematic identifier. For the sake of administrative efficiency and timely delivery of programs the government has in recent years resorted to technologies which help it reduce the "paper-shuffling." Information received from individuals in paper form is microfilmed, coded and maintained in higher density storage media. A long letter from a program participant is reduced to a single computer change-order slip; old addresses and records of those no longer participating are, in due course, deleted automatically. Normally when a file such as that of a Canada Pension Plan beneficiary finally reaches the archives, all that might be found inside the file jacket are a few computer change orders, the current address and the current benefits. Information on that individual's fifty or sixty years of interaction with government will not be there. Since modern storage technologies lend themselves quite readily to copies, the historical and research interests would be met if procedures were instituted so that at least a complete microdata sample of such administrative data would be deposited in a continuous fashion with the Public Archives.

In return for authorizing the alteration of records from one storage medium to another, e.g., paper to microfilm or machine readable form, the Dominion Archivist would require a sample to be forwarded to the archives. Sampling schemes adopted for such series of case files or data banks should:

- be applicable to the paper, micrographic and machine readable components of a data bank;

- be consistent with the technology used for processing the information of that data bank;

- be cost-effective and administratively implementable, given the human and financial resources at hand;

- be statistically acceptable;

- allow for longitudinal studies, i.e., be periodic and consistent;

- allow for comparative studies, i.e., be integrated with other data banks so as to create as high a probability as possible in capturing information on the same data subject from different data banks.

The last point touches on the problem of record linkage. Record linkage is the process whereby data from different record systems are linked on a case-by-case basis, on grounds that any of the single record systems is incomplete either with respect to data or required coverage or both. Record linkage is used to "reconstitute" large portions of (past) populations or to consider "statistically" any large population or sample in a multivariate context.

For most large data banks the basic file series is organized by unique identifiers such as the Social Insurance Number, corporate taxation number, or like unique systematic numerical or symbolic identifiers. Derivative file series, e.g., frauds and persecutions, appeals, are organized either as part of a main file series with colour coding or as a separate file series, often in alphabetical sequence. Other case files whose existence is the result of voluntary rather than mandatory participation are often organized alphabetically within some subject, geographic region, or industrial or product coding schemes.

The approach of the archivist has always been that of sampling. It is estimated that on the whole the Public Archives presently selects for permanent retention less than 5% of all the paper records. Case files can be selected for permanent retention on the basis of any of the following criteria:

- a case file may be important for the issues involved, due to the issue itself or the context in which it occurred;

- a case file may be regarded as important for its influence in the development of principles, precedents, or standards of judgement in such matters as the definition of the jurisdiction, operations, and mandate of the department concerned;

- a case file may be regarded as important for its contribution to the development of methods and procedures;

- a case file may be regarded as important for its documentary/illustrative value;

- a case file may be regarded as important for its research value, which even if minimal for one case file would be high enough to warrant permanent retention if a sufficient number of case files were selected.

Basically, the archivist uses two general kinds of sampling techniques, "judgement sampling" and "probability sampling." In both cases the archivist selects for permanent retention a collection of records on only a few members of the universe. A more common term for judgement sampling is selective retention, while probability sampling or just "sampling" is a technical term for a procedure whereby one selects a number or "sample" of items from a defined "population" of items in such a way that every item in the population has a known chance of being selected. For example, if one decided to sample by terminal digit 5 of the Social Insurance Number one would know that the probability of any individual being in the sample would be 1 out of 10 or 10%. Should one instead decide to sample first letter of the surname, for example by the letter "R", the probability would be 1 in 20 or 5%. The exact details of sample scheme for administrative and survey data will be presented at a later date. (7)

If one can now assume that the Dominion Archivist has acquired administative data banks *in toto* or in sample form, these are some of the basic parameters that might be applied to the release of microdata under controlled conditions:

- there must be a legitimate and important research purpose to be served by the process;

- the researcher must sign a written contract specifying the degree of detail below which information taken from the microdata may not be disclosed;

- at time of the conclusion of the research project or no later than some specified date, the researcher, if granted temporary possession of microdata, must either return such data to the Public Archives or submit an affidavit as to its destruction;

- there must be a prohibition against dissemination of the microdata to a third party without the written authorization of the Dominion Archivist;

- the researcher must submit a copy of the publication containing the data derived from the microdata to the Public Archives and in some cases prior to publication;

- significant and mandatory sanctions or penalties for improper disclosure of microdata would have to be founded in law;

- the researcher must have available an ombudsman mechanism to deal with conflicts relating to the terms of a research contract before it is signed;

- where the researcher is allowed to take the microdata to his own facilities for processing and analysis, the facilities should provide a level of security comensurate with that required for the microdata

A Sample Application—The Federal Government

In terms of actual application, we can consider the following (hypothetical) example for a socio-economic program (SEP). The program is about to shift from a totally paper mode to a multi-storage media approach, i.e., paper, microform and EDP

The program consists of five different levels of file series, namely:

Title	Organization	N of Participants
Benefits & Claims	By terminal SIN digit	22,000,000
Appeals	ditto, colour coded	1,100,000
Frauds & Persecutions	alphabetic	220,000
Judicial Decisions	alphabetic	44,000
Medical Files	By terminal SIN data	2,000,000

The administrators of the program propose to destroy all paper records for file series 1,2 and 5 upon receipt, microfilming the "relevant" portions instead. They further propose to destroy all files for file series 1 and 2 two years after last action, and file series 3 and 5 five years after last action, while maintaining only summaries of 4, the judicial decisions.

After some deliberation, the Dominion Archivist approves the records retention and disposal schedule subject to the following limitations:

For file series 1 the programme administrators are to deposit at the Public Archives,

- a .01% sample of all paper records consisting of those participants whose SIN number ends in 5555;
- a .1% sample of all records microfilmed consisting of those participants whose SIN number ends in 555;
- a 10% sample of all EDP records consisting of those participants whose SIN number ends in 5.

For file series 2, the samples for (a) would be 1%, e.g. all red colour coded file jackets where the SIN number ends in 55. For (b) and (c), no separate samples would be taken as it is assumed that for (c) there would be a code for "appeals" in the EDP record.

For file series 3, a 5% sample would be taken of all those whose first letter of the surname starts with an "R." (Note: 10% of the "R's" would also be found in the level 1, EDP samples. This would ensure high linkage probability.)

For file series 4, the sample would consist of two components,

- those case files for which the judicial decision was of special significance because of person involved, nature of case, precedents, etc.;
- a 10% sample consisting of the letter "R" (4.94%) and the letters "A" (2.92%) and "N" (1.69%). While the letter "B" for example would give a 10% sample directly such a sample would not be "linkable" to file series 3. One should therefore approach alphabetic sampling in terms of "common building blocks."

In our hypothetical example, the department agrees with this sampling plan, knowing that the Public Archives will abide by the access and disclosure provisions of the legislation pertaining to this data. Some of the data can be made readily available in anonymized form, e.g., the EDP portion, and the remainder will become available at a later date. This assures a respect for the totality of the "fonds."

Now this is a single case. If the same approach were applied to cover all the data banks, a national archival sample would be created. The matter of a national archival sampling strategy for administrative data would include advice to the Dominion Archivist from such agencies as Statistics Canada, the national research councils, and the research community. In order to maximize the probability of record linkage for longitudinal and comparative studies, a national archival sampling strategy would have to be applied uniformly to all administrative data. The immediate benefit would be a reduction in survey costs and resulting paperburden.

It may well happen that a government institution "A" wishes to use the data collected by another department "B" (administrative or survey data) for research purposes. However, the legal questions involved may take some time to resolve. Department "A" makes its case to the Public Archives, which in turn requests Department "B" to deposit a copy of the data in question, e.g., a machine readable data set and documentation, in the Public Archives, which will hold the magnetic tape until the legal differences are resolved. Such a mechanism might be of particular interest to government departments, in that it provides for a single "neutral" depository of administrative data, thus ensuring that longitudinal series can be created without a sudden hiatus caused by some legal problems of direct transfer from "A" to "B."

The Public Archives would be the neutral repository of these samples, which would be subject to the transfer conditions and prevailing legislation. When demand from the research community warrants, "public use files" would be prepared. In other instances, a research file would be created from the master file. The Public Archives may even decide to split the master file into two separate components, removing all unique identifiers to a XREF file and substituting a control number instead.

The research community would peruse the federal government inventory of data banks (required under FOI and privacy legislation) and the national archival sample to determine whether the administrative data being archived (and the variables contained therein) met their research needs. If not, a researcher or a research project could make its wishes known to the Dominion Archivist for the permanent retention of certain administrative data, even though legal and financial problems pertaining to access for research purposes may take some time to resolve.

The ombudsman mechanism could be fulfilled through the creation of a special committee of the already existing advisory council on public records consisting of representatives of the learned societies and like representative bodies.

References:

1) **Federal Inventory Annual Report, March 1980,** prepared by Federal Inventory Data Base Group, Statistics Canada. For further information contact Mr. Dave Sally, Federal Data Bank Manager, Statistics Canada, Ottawa, Canada, K1A 0T6.

2) As cited in J. Knoppers, "A Freedom of Information Act and the Future of Social Science Research", *Social Sciences in Canada,* 7 (December 1979) 4:9-10.

3) At present the Canadian Historical Association and the Canadian Political Science Association represent the research community on the Advisory Council on Public Records. The Public Archives is giving serious consideration to widening the diversity of representation of the research community.

4) The full text of Part IV of the Canadian Human Rights Act can be found as an appendix to the *1980 Index of Federal Information Banks* which is available for reference in every Post Office and Canada Manpower Office across Canada.

5) *Kissinger vs Reporters Committee for Freedom of the Press et al.*, Supreme Court of the United States, March 3, 1980, No. 78-1088 and 78-1217. The case involved summary or verbatim transcripts of Kissinger's conversation notes and telephone notes which he "unlawfully removed" from the State Department and at a later date deposited at the Library of Congress with severe access restriction. In short, the Supreme Court ruled that the plaintiffs had "no standing" and that only the National Archives and Records Service and/or the State Department could pursue Kissinger for return of the notes.

6) *Senate Report to the Federal Records Act of 1950*, S. Rep. No. 2140, 81st Congress, 2nd Session, at 4(1950).

7) The distribution of SIN numbers by first letter of surnames for R is 4.94%. The author is currently undertaking a project to establish an overall mixed numeric-alphabetic sampling scheme for case files falling under privacy legislation. This includes a study on the distribution of first letters of surnames of language, ethnic and provincial groupings.

THE CONDUCT OF USER SURVEYS

Dennis D. McDonald, Ph.D.
King Research, Inc.
Washington, D.C.

I'd like to start by making a few statements in general about user surveys, and then to discuss their goals, their different varieties, and finally some methodological pointers.

First, there's no such thing as an "end user."

Second, information isn't a commodity like a bar of soap which can be bought, sold, and priced "over the counter."

Third, don't believe that users and potential users can tell you pointblank what they really want and need in the way of information services.

Finally, the more you know about a user before you conduct a user survey, the better off you'll be.

GOALS

User survey goals can be classified into the following five categories:

- **Input prior to system development.**

  You may want to find out the needs of a potential user group before you invest a lot of dollars in the development of systems or services. This kind of user survey is difficult to conduct since what people *say* they may need and what they actually end up using may be two entirely different things.

- **Information about your competition.**

  You may be in the process of developing a system or service, and you may want to assure yourself that related information products and services, i.e., your "competition," are not satisfying the same needs. This kind of survey is essentially a form of intelligence gathering, and conducting it will force you to consider how your product or service will supplement what is already available.

- **Identity of current users.**

  You may have a system in operation, and you may want to find out not only how and why people are using your system, but whether they are satisfied with what your system is supplying. This is what most people consider to be a "user survey."

- **Potential users.**

  You may have a system which is operating and satisfying the needs of a certain core group of users. But you want to expand the system's use. You must then seriously think about potential user populations and how to ask