Problems in handling process-produced data

Vries, John de

John de Vries

Problems in Handling Process-Produced Data

I. Introduction

I am assuming that the other contributors to this section on „Preservation, Storage and Assess" will deal more specifically with the technical aspects of „Storage" and „Preservation" of large-scale data sets and, possibly, with the computing aspects of the problems of „Access". I wish to address myself more specifically to a variety of problems in the area of „Access" from the initial stage of acquiring information about Process-Produced Data, through the requirements for adequate documentation, to the final phases of actually analyzing Process-Produced Data sets.

Readers will note that most of the examples and illustrations given throughout this paper refer to demographic data, mainly from Canada. Although these illustrations reflect my personal experiences, biases and ignorance, I am sure that you can find similar examples from your own country and your own academic discipline. I am strongly convinced that, generally, the problems I will discuss are not unique nor are they particularly more serious in one country than in another. At best, one can state that some countries appear to have made somewhat greater progress towards solving the following sequence of problems than have other countries.

I should also note that there are a variety of definitions of „process-produced data" used throughout the social scientific literature. A fairly cursory search through various journals yielded *five* definitions, which are not identical with each other. I will not enter into a detailed discussion of the respective merits of these definitions at this point; it is, however, obvious that sooner or later we will have to come to grips with this proliferation of terms.

In the following „tale of woes" I construct the hypothetical sequence of problems which the researcher interested in process-produced data is likely to encounter during his search for and acquisition of administrative records as a data source.

* I wish to point out that an earlier version of this paper has been presented at the First Open Conference of IASSIST (Canada) in Toronto, May 11–12, 1977. The earlier version was specifically aimed at IASSIST participants and thus dealt in greater detail with the kinds of actions required to overcome some of the problems which one encounters in handling process-produced data. Interested persons are invited to request copies of the earlier paper.

II. Problems of Information and Acquisition

Under this heading, we can classify a large collection of problems which fall into several sub-categories:

A. Information about Existence

To a large degree, information about the existence of particular process-produced data sets is rather hard to obtain. In fact, it is sometimes difficult to find out where such information may be obtained. Admittedly, this situation has been improving in the United States and in Canada, but progress is slow, and information about progress is, again, difficult to obtain. The federal statistical agencies in both countries have recently developed some activity in this area. In the United States, we now have access to catalogues from the National Technical Information Service and from the Machine-Readable Archives Division of the National Archives and Records Service[1]. In Canada, we are running behind to some degree. The Machine-Readable Archives Division of the Public Archives of Canada was created in 1974 and has begun to make selected data sets available to researchers. To my knowledge, no catalogue of available data sets has been produced as yet. In addition to the Public Archives, Statistics Canada (the federal statistical agency) has started a Federal Data Clearinghouse, in its Standards Division. This group received Cabinet authorization in November 1974; to my knowledge, no published reports of its activities have appeared thus far. Among the stated objectives of the Federal Data Clearinghouse are the following products:

a) an information base to enable users of statistical data to locate potentially relevant information sources, related activities and outputs;
b) a Data Index which will contain, indexed by class of information, variables, indicators and time series appearing in major federal publications and machine-readable data files. This index will reflect the degree and nature of availability of the statistical outputs to various classes of users.

It is not at all clear that the Federal Data Clearinghouse will be of great utility to potential users outside the federal government (except, probably, other levels of government).

It would appear that we need, fairly urgently, some overview of available process-produced data in machine-readable form. The North American Action Group on

[1] Traugott, M.W., and Clubb, J.M., Machine-Readable Data Production by the Federal Government, in: American Behavioral Scientist, Vol. 9, No. 4 (1976), pp. 399–401.

Process-Produced Data in the International Association for Social Science Information Service and Technology (IASSIST) has therefore begun to compile an „inventory of inventories", that is a directory of catalogues which list data bases (primarily, though not exclusively, of the process-produced kind). At this stage, we see the final product of this exercise as an annotated bibliography of inventories to be updated periodically. Although we are nowhere near completion of this undertaking, it is already clear that existing inventories (usually produced by governmental departments) vary tremendously with regard to their informational value to outside users. It would seem fair to expect that any inventory of machine-readable data produced by government provide the following essential items of information on each file (that is, in addition to at least a summary of the physical characteristics of the file, such as number of records, record length, medium in which data are carried, etc.):

1. references to existing documentation, methodological discussions, etc.;
2. an indication of the „degree of availability" of a file; we would expect that a threefold classification would be sufficient at this first level of information, that is:
   a. available without restrictions;
   b. potentially available, depending on circumstances;
   c: not available under any conditions (e. g. for reasons of national security, or because of restrictive legislation on confidentiality, etc.);
3. an indication of the cost of acquiring a copy of the data set (obviously only under the condition that it were available);
4. the government department; and the name of the position therein, to contact for further information.

Although these would appear to be rather minimal items of information, I should point out that the government of Ontario recently published an inventory of machine-readable data files produced by its various departments in which *none* of the above items were included!

B. Information About Specific Data Sets

Let us assume that we know that a specific data set exists and is, in fact, available to interested researchers. We then require additional information about the data set, for example the following:

1. The Data-Gathering Process. Since we are dealing with administrative records of some type, we must assume that the agency responsible for collection of the data (even if, in this case, „data collection" is no more than some form of bookkeeping or administration) is striving for complete coverage of the event being recorded (I should point out that the same goal of complete coverage exists in the case of enumerations, for example, with censuses. In this regard, process-produced data differ

from sample surveys, where one does *not* aspire to cover the relevant universe. Even within the selected sample. response rates seldom exceed 80 percent.). In other words, when the administrative records describe *events* (for example, bankruptcies, marriages, charges laid by the police, etc.) we must assume that, in principle, all events of this specific nature, which occurred within a given territory and within a particular time-span should be recorded and, as a result, produce an administrative record which describes some pertinent aspects of each event. Again, enumerations are similar in this respect: if the records describe existing elements of population (legal residents, registered voters, institutions of higher learning, mental hospitals, and so on) we must assume that one of the basic aims of the enumeration is to include all members of the population.

In order to assess the degree of completeness of the administrative data set, we must have some fundamental information about the administrative process which yielded the data in the first place. For example: is reporting the event (or responding to enumerations) required by law, or is it a voluntary act? Within the vast array of possibilities among process-produced data, there are obviously wide variations. For example, in the area of crime statistics, the „charges laid by the police" are frequently an automatic result of the reporting of a crime. That is, the reporting of a crime should be recorded by some police officers; if no record is produced, this could be a result of laziness, inaccuracy, inefficiency or corruption. In other words, the recording of the „event" in this case is required by law, is carried out by a public servant as a part of his duty, and should have a high correlation with the occurrence of the event itself. In contrast, take the registration of marriages. It is clearly the case that „legal" marriages (that is, those which are accompanied by some formal ceremony) are normally registered, when the registration is part of the formal ceremony itself – as in most civil marriages. However, if the formal ceremony is a religious one, taking place in a secular nation-state, recording of the event may or may not take place; such recordings may not have any relevance with regard to the *public* recording of marriages. To make matters even more complicated, common-law unions clearly are not registered by any administrative process. Thus, civil marriages *must* be registered, since such registration is a legal requirement. However, as a measure of the „nuptiality" of a society, such records may be quite inadequate, since in most societies one is not legally required to register the beginning of some period of heterosexual cohabitation.

In cases where registration is required by law, we should know that the sanctions are in the case of non-compliance. (It should be obvious that the sanctions are different in the cases of voter registration and filing an income tax return.) Other questions which fall in this domain are: to what degree are laws regarding compliance in fact enforced? To whom, or which agency, does one report? How much time is permitted to lapse between the occurrence of an event and its registration? Are the events located in *time* (time of occurrence of the event, some measure of duration of the event, some measure of duration of the „event" or both) and *space*?

To make matters even more complicated in this area, let us consider yet another aspect. So far I have assumed that some event occurs and that *someone* is expected

to register the event, that is, to produce some administrative record. I have assumed that the main source of discrepancy would be the non-existence of the administrative record. But there is, of course, a second type of discrepancy; this occurs when an administrative record exists which has *no* counterpart in the universe of events (for example, to file a claim for insurance benefits or some form of public welfare, based on some fictitious event).

It is my opinion that these kinds of information are generally quite difficult to obtain. Often, documentation of this nature exists only in unpublished „working papers" and memoranda, for which bibliographic entries are scarce or non-existent. Moreover, it is not always clear *who* (or which branch of a particular agency) is responsible for the external distribution of such documents. Not uncommonly, documents of this nature are labelled as „classified" and are thus totally inacessible to external users.

I have tried to develop a fairly lengthy list of questions in this specific area, because it is especially in this regard that process-produced data differ from most (if not all) other types of social science data. It is my impression that a large proportion of the researchers using these data is insufficiently critical of the reliability and validity of available data and has failed to ask pertinent questions which could reveal the quality of the data.

2. **Administrative Decisions Regarding Data Manipulation.** Although we are working from a definition of process-produced data which specifies that they are by-products of administrative processes, it is obvious that virtually no administrative agency can handle such records in their original state. As all other machine-readable data (I am obviously, in this section, not terribly concerned with process-produced data which only exist in manuscript form), such records must have gone through processes of recording, transcription, coding, keypunching, reformating, editing and recoding. It is inconceivable that all these procedures are foolproof. Thus, errors slip into the records at various phases of the manipulative phase. As a result, it is likely that, at various points, decisions have to be made regarding any errors or inconsistencies which can be detected. Such documents as coding schemes, ways in which missing values were handled, procedures for editing and cleaning data are all essential for the proper interpretation of a given data set by a secondary user.

C. Governments as Research Resources

The two sections above reflect a broader dimension, which also needs to be discussed. It is quite clear that governments, at various levels (i. e. national, regional, local) have a virtual monopoly over a large share of the existing process-produced data. This is most obviously (and trivially) the case when we deal with records which are the by-products of internal governmental processes (e. g. minutes of Cabinet meetings, federal budgets, and so on), but it also extends to data which are

by-products of administrative processes relating to other sectors of society. In part, of course, this reflects the expense involved in gathering data on a massive scale (for example to record every birth occurring in the country); in part it reflects the fact that governments have legal, political, economic and persuasive powers which private researchers can not bring to bear.

The fact that information about process-produced data is either non-existent or relatively inaccessible may well be taken as an indication that the relations between the social science research community and governments, as data collectors and sources of data, have not yet been worked out satisfactorily – at least in North America. It is my impression that lines of communication are quite imperfect. The research community appears to have relatively little impact on the collection, manipulation or distribution of data by various governments; prospects for improvement of this situation do not seem to be very good[2].

In general, social scientists appear to have had some success in acquiring census data in machine-readable form, both at the aggregated level and at the level of individual records. The United States quite clearly took the lead when it released public use samples of the 1960 Population Census. This initiative was followed in 1970 by an even more generous program of access to individual data. Canada was a somewhat hesitant follower with its set of public use sample tapes from the 1971 census. As the recent article by Judith Rowe[3] indicates, several other countries are prepared to give researcher access to at least samples of individual census records, after appropriate safeguards against disclosure of identification have been taken. It is indicative of the state of affairs with regards to administrative records that no equivalent programs exist to give researchers access to individual records in this area. Three causes can be mentioned for this state of affairs: first, laws regarding access to information are often unduly restrictive, often making access to data impossible for individual researchers; secondly, even when the legislation allows for some discretion on the part of governmental authorities, the importance of the data for research is frequently not recognized; and thirdly, the research community has generally not produced the kinds of efforts regarding administrative records which eventually yielded the various public use samples of population censuses.

It would seem necessary that the social science community increase, where possible, its involvement in the shaping of further legislation on freedom of information, privacy and access to governmental data. Moreover, strategies should be developed to facilitate the flow of data and information within the constraints set by existing legislation. For example, it may be useful for someone to produce an unbiased discussion of the various techniques by which records can be protected against disclosure of identity (such as collapsing of sensitive categories in some variables; elimination of unique or near-unique identifiers, introducing random disturbances into

[2] Op. cit. pp. 402–407.

[3] Rowe, J., Non-American Census Data in Machine-Readable Form, in: Demography, Vol. 14, No. 1 (February 1977).

some proportion of the records and some proportion of the variables; scrambling groups of records, and so on). Finally, there appears to be a great need to educate various „publics" (governmental decision-makers, politicians, voters, the general public) about the need for access to public data.

III. Problems of Documentation

Let us assume that we have aquired a hypothetical process-produced data set, accompanied by the types of documentation outlined so far. We will then still need several items of documentation which may, or may not be included.

A. Specification of the Universe

To use the terminology employed in the description of sample surveys: we need to know how the relevant population is defined, or which members of the population are defined as being part of the data set. (Remember that I am referring to a population of *events*, not necessarily a population of individuals.) To indicate the effects of different universe specifications on the nature of the data produced, let me describe the Canadian data on unemployment. There are three sources of information: the Monthly Labour Force Survey, Statistics from the Unemployment Insurance Commission, and the population census (Note that only the second source is of the process-produced variety). The Monthly Labour Force Survey is „. . . a multi-stage probability sample of the civilian, noninstitutional population of age 14 and over of Canada excluding the Northwest Territories and the Yukon". The census, on the other hand, covers the total Canadian population[4]. Thus, residents of the Northwest Territories and the Yukon are included in the census (but not in the Monthly Labour Force Survey). Moreover, the census includes the institutional population, the Armed Forces, Indians on reserves and Canadians resident abroad. On the other hand, the census questions on employment status are only asked for persons 15 years and over, while the Monthly Labour Force Survey has a lower age limit of 14.

While the former two data sources refer to the *prevalence* of unemployment, statistics from the Unemployment Insurance Commission refer to the *incidence* of employment, among a specific population. The statistics refer to the claims for un-

[4] Dominion Bureau of Statistics, Canadian Labour Force Survey (Methodology), Ottawa 1966, p. 8.

employment insurance benefits (thus, claims are the *events* of which a record is kept). Therefore, if we assume that each claim represents one and only one claimant, the statistics will give us the number of claimants of unemployment insurance benefits. Claimants must, by necessity, belong to the insured population, This population is *not* identical to the total civilian noninstitutional population. It consists „ . . . mainly of the paid worker segment of the labour force. Members of the Armed Forces are in the insured population"[5]. There are, however, exceptions. For example, employed persons with very low earnings, employed persons 70 years and over, and employed persons to whom a retirement pension under the Canada Pension Plan has at any time become payable do *not* form part of the insured population (Statistics Canada, 1973: 81–82). To make life even more complicated, the census definitions of „employed" and „unemployed" are not identical to those used in the Unemployment Insurance Commission's statistics.

I have spent some time delineating these differences in order to illustrate the importance of knowing the *exact* definition of the universe to which the data pertain. It is not always easy to find the exact specifications as they were stated in principle. It is often even harder to find out how membership in the universe was determined in practice. The preceding discussion, by the way, also illustrates one of my earlier points: the difficulty of obtaining this type of information about existing data sets. All the above information about variations in universe specification can be found in an unpublished discussion paper from the Labour Force Survey Section of Statistics Canada. Such papers do not normally appear in the common bibliographic tools (such as library card catalogues, and so on).

B. Estimates of Errors of Coverage

Demographers have traditionally made a distinction between errors of coverage and errors of content. Errors of coverage deal with violations of the assumption that members of the universe are presented by a record, once and only once. That is, the assumption that there is a one-to-one relation between an event (a birth, death, bankruptcy, admission to a mental hospital, and so on) and a record in some administrative data file. Thus, errors of coverage are of two kinds:

a. the exclusion, or omission, of elements whcih should have been included, and
b. the inclusion of elements which should have been excluded.

In the former category, we can distinguish the following subtypes:

- an event occurs which is not recorded (e. g. failure to register a birth, death, etc.);
- an event occurs, is registered but (incorrectly) defined as not belonging to the particular universe;

[5] Statistics Canada, Concepts and Definitions used in the Canadian Labour Force Survey, Ottawa: Labour Force Survey Discussion paper, p. 8.

– an event occurs and is registered, but the record is „lost" during the phases of recording and data manipulation.

In the latter category, possible subtypes are:

– a registration is made for an event which did not really occur:
– a registration is made for an event which did occur, but which did not belong to the particular universe;
– an event, which did occur and belonged to the particular universe, was recorded more than once;
– during the phases of recording and data manipulation, multiple records of a single event were erroneously produced.

For any given data set, we obviously require not only a discussion of the possible sources of error, but also some estimates of the magnitude of the various types of error. Unfortunately, we usually find ourselves in the position where no such estimates exist, or where they are severely biased.

C. Estimates of Errors of Content

The second kind of error is usually called „errors of content". In this case, an event has been recorded, but at least one of its characteristics has been recorded incorrectly.

Of special importance in this respect are those variables which anchor the event in space and time. Let me, again, give a few illustrations from the vital statistics field.

Temporal measures may be in error in several ways. For example, births may be recorded with an indication of the date of birth, or of the date of baptism (as happened with many early birth records). Quite clearly, under conditions of severe infant mortality, the two points in time would lead to drastic differences in estimated birth rates (because infant deaths which occurred prior to the baptism would, of course, not require a baptism. In such case, it is likely that no registration occurred at all). Another problem can occur when an event takes place close to the end of a calendar year (or some other unit of time over which information is summarized). In such cases, the registration of the event will often take place in the following year. This may lead, trivially, to an error in recording the date of occurrence. But another possibility arises. Usually registrations of vital events have a „cut-off date". For example, if an effective period of three months is allowed to pass between the occurrence of an event and its registration, one can „close the books" for calendar year y on April of year y+1. Now suppose we have an event which occurred at some time during year y, but which was not registered before the cut-off date in year y+1. Among the possibilities are: the event could show up in the summaries for year y+1; it could, later, be included in revised statistics for year y; it could be omitted from summary statistics altogether. I should indicate that this last possibility is by

no means just the figment of someone's fertile imagination: during the 1960's, an estimated 10,000 births per annum were indeed, in this fashion, excluded from the vital statistics of the province of Quebec.

Such problems are obviously of great importance when one studies variations of phenomena over time. In a similar fashion, spatial measures are of special importance. Each event which is recorded will have at least on spatial referent. In some cases, such delimiters are of no value (for example, data on attendance in the House of Commons), but those are exceptions. Normally, we will require clear specifications of which measure is recorded. For example, vital statistics require the recording of either the place of usual or legal residence (of the mother, in the case of births) or the place where the event occurred, or both. In the case of births, it is not uncommon for residents of small towns or rural areas to have their babies in nearby cities. If we were using data based on „place of occurrence" instead of the usual residence of the mother, we may find artificially high birth rates for cities with maternity wards in hospitals, and artificially low rates for places without such facilities. I should point out that this type of distortion may indeed have led to some erroneous statements regarding ecological variations in the incidence of mental illness (derived from statistics on admissions to mental hospitals).

In addition to these „anchoring" measures, there are, of course, other variables which may contain errors. We need to know, in fairly great detail, what experience the administrators have had with these data: which types of error are most common, what procedures have been followed to isolate errors, how errors have been corrected, and which methods have been employed to validate the corrections.

D. Codebook Equivalents

We should distinguish two kinds of process-produced data: they either refer to single events, or they are the products of various computational procedures and take the form of aggregated data, or summary statistics of some kind. In the former case, we are essentially dealing with codebooks which differ little from those which accompany sample surveys. We need to identify variables, give the exact wording of the questions, indicate the various permissible response categories, and so on. Ideally, we would also like to have a sample copy of the registration form. Aside from the fact that relatively few data sets are available in this format, and that in most cases nothing like a codebook exists, we are not faced with unique problems.

In the latter case, where there is no longer a one-to-one relationship between the machine-readable record which the researcher gets and the originating event, problems do arise. As an example take the simple case where a data file gives us total populations, at a specified point in time, for areal units in a country. Quite clearly, we will require a statement about the universe covered (Note that this may be a somewhat different universe from the one used to collect the data. For example, many

countries do not provide published data on the armed forces). In addition, one would like to be given summary statistics on such variables, such as minima, maxima, means or medians and standard deviations.

Finally, in the case where the data refer to elements which have spatial counterparts in the real world, a codebook should have sufficient information to locate these elements on maps (or in the real world). One possibility might be to include, with the data, coordinates of sufficient detail for one of the common computer mapping packages, so that users may produce their own maps.

IV. Problems in Data Transformation and Analysis

Although I am now moving beyond a general interpretation of „access", let me briefly enumerate the following issues:

A. „Ragged" File Structure

Often process-produced data have a file structure known as „ragged": logical records do not have a fixed length, but their length varies in relation to the number of elements they contain. For example, data pertaining to households could have formats varying with the number of persons in the household. Most of the standard computing packages do not lend themselbes easily to the handling of such data files.

B. Simultaneous Use of Several Records within a File

Most statistical packages do not allow the simultaneous handling of variables from more than one logical record. Both in the analysis of time-series data, and in an analysis of contiguity, one runs into difficulty with the most common packages.

C. Peculiarities of Aggregated Data

Most statistical packages appear to be the products of the research style of survey researchers. As such, techniques which are more suitable for the analysis of aggregate data tend to be rather poorly developed.

Almost invariably, the analysis of process-produced data relating to *events* requires the use of the corresponding enumeration (to standardize for variations in the exposure to risk, for example). Thus, *two* data sets are often involved. The following problems may occur:

1. **Standardization of stimuli:** in its most fundamental form this problem manifests itself through differences in definitions of the universe. For example, one would have problems if one were to use the Canadian unemployment data from the Unemployment Insurance Commission in conjunction with data from the population census to calculate measures of the incidence of unemployment in Canada.

At a less fundamental level, we may find cases where the universes do coincide, but where characteristics are measured in slightly different fashions. This problem is a common condition when the „linkage" involves data from two points in time. For example, the Canadian census definition of „mother tongue" changed between 1931 and 1941. The earlier criterion required the language still to be spoken, while the later criterion only required the language still to be understood.

2. **Standardization of responses:** even in cases where the stimuli are identical, the available sets of responses may not be exactly identical, either in terms of available categories or in terms of the operational definitions of some of the categories. For example, anyone studying illegitimate fertility in Canada has to come to grips with the fact that, since 1949, Ontario has adopted a different definition than the other provinces: in the former case, births are illegitimate if the mother was never married, while in the other provinces births are illegitimate if the parents are not married to each other.

3. **Standardization of delimiters:** as I already pointed out, variables dealing with space and time are of special importance. In North American society, we are faced with many ways to divide territory: municipalities, counties, planning areas, school districts, electoral districts and, in Canada, the „bilingual district". For reasons which are totally mysterious, the boundaries of such sets of districts usually overlap. Moreover, boundaries often change over time. Therefore, any „linkage" of aggregated data sets is likely to present the user with areal units which are not perfectly identical and which can not easily be mapped onto each other. We clearly need estimation techniques to deal with such situations, since one can usually not gain access to date at such level of areal detail that a reconsiliation of boundaries can be established.

V. Conclusion

I have attempted to provide a list of some problems one is likely to encounter during the various phases of accessing and handling process-produced data. I have tried to indicate that most of these problems are nearly unique to process-produced data; thus, we gain little from the experiences of survey analysts. Given the enormous variety of process-produced data which now is beginning to find its way into the research community, solutions for the array of problems indicated are likely to come from persons working in different scientific disciplines and in different countries. Any efforts to increase and facilitate communication between these highly dispersed researchers should therefore be strongly supported.