This PDF is a selection from an out-of-print volume from the National
Bureau of Economic Research

Volume Title: The Role Of The Computer In Economic And Social Research

Volume Author/Editor: Nancy D. Ruggles

Volume Publisher:

Volume URL: http://www.nber.org/books/rugg74-1

Publication Date: 1974

Chapter Title:

Chapter Author: I. P. Fellegi, S. A. Goldberg

Chapter URL: http://www.nber.org/chapters/c6612

Chapter pages in book: (p. 1 - 18)

THE COMPUTER AND GOVERNMENT STATISTICS

I. P. FELLEGI AND S. A. GOLDBERG*

Statistics Canada

INTRODUCTION

Technological barriers to the effective realization of many of the overly enthusiastic pronouncements of the sixties in regard to the utilization of the computer are being removed, while the cost of hardware is actually declining. Yet one can sense in the literature on computer utilization a tendency toward greater caution. This tendency is presumably the result of the failure of many automation projects to realize fully their promised goals, and to complete their implementation within anything like their scheduled times, and is undoubtedly a healthy one. This paper is an attempt to present some views in regard to automation in the light of our recent experience in Statistics Canada. It draws upon and updates an earlier paper presented at the London meetings of the International Statistical Institute, which dealt with the theme of the present session; namely, government statistics and the computer.[1]

We shall concentrate here on problems associated with automation in a statistical office. In so doing, our intention is to promote the cause of automation, rather than to inhibit it. By identifying problems and facing them squarely, we should be able to cope with them more effectively.

Three classes of problems are discussed in this paper: data problems; problems related to the implementation of projects for automation; and problems arising from the overriding need to maintain confidentiality. The discussion of these problems is preceded by some reflections on the goals of the statistical office and of automation. We end the paper with some remarks on issues confronting management.

SOME REFLECTIONS RELATED TO GOALS OF A STATISTICAL OFFICE AND OF AUTOMATION

Clearly the goals of automation must derive from the goals of the statistical office, which in turn must derive from the goals of society. It is appropriate then to start with the question, Why should a statistical office aim to automate its operations?[2] This should provide perspective to the subsequent discussion.

The statistical office, like other organizations, must strive to become more efficient; that is, to increase its output per dollar spent. If analysis indicates that

---

the increase in timeliness, quality, or volume of useful output exceeds the cost of the conversion to automated procedures, this is a good and sufficient reason for conversion.

However, output per dollar is not the only relevant factor to consider in deciding whether or not to automate. Thus, we may find that the satisfaction, without automation, of a sharply rising demand for statistical information would involve substantial increases in manpower, rendering the management of the statistical office more difficult and unwieldy. In such a case, even elementary considerations of efficiency and effectiveness might well make automation desirable.

However, considerations of effectiveness become much more prominent when we consider the basic goals of society, which should be reflected in the goals and programs of the statistical office. The goals of society are seldom defined unambiguously. Some "traditional" goals relating to economic growth can, it is true, be quantified and monitored, though even this is far from simple. However, the more recently formulated social goals have so far been articulated in most societies only in broad, vague terms, such as the elimination of poverty, abatement of pollution, a more egalitarian society in terms of income and asset distribution, better control of crime and socially maladjusted behavior, less congestion in urban areas, participatory democracy, reduction of social tensions, and a better quality of life. Precise definitions of these notions and measurable concepts are still to be worked out and a great deal of experimentation will be involved, by both users and producers of statistics—not only to devise appropriate measurements but to evolve appropriate analyses of the data. Thus, the statistical agency finds itself facing a multiplicity of goals not fully defined in advance, and the traditional cycle of statistical operations—collection, processing, and publication—is becoming insufficiently flexible. What is required, in addition, is the storage of the most detailed data cells and a capacity for retrieval of an almost infinite variety of unanticipated cross-classifications and aggregations, consistent with confidentiality requirements.

There is also a growing requirement to be able to interrelate data collected in a variety of surveys or through other sources. In fact, policies and programs are frequently interrelated so that the formulation of any single one must take into account its effect on the others. The dangers of formulating policies in isolation and on the basis of partial knowledge are being recognized increasingly, as reflected in the following quotation from the Report of the Special Commission on the Social Sciences of the National Science Board of the United States.

> Partial policies, based upon some narrow band within the whole spectrum of relevant knowledge, may cause more indirect harm than direct good, and will always require costly modification, improvisation and patching, when unexpected and unwanted results appear. We need only mention insecticides, irrigation, mass transport, and urban housing to bring to mind a host of striking examples of the pitfalls of narrow approaches to *broad* problems.[3]

Broad approaches to the analysis of problems require a great deal of detailed data and highly flexible retrieval capability, facilitating their joint and interrelated

[3] *Knowledge Into Action: Improving the Nation's Use of the Social Sciences*, National Science Foundation, 1969, pp. 51–52.

use. The more widespread utilization of economic and social models based on a large number of equations, and the development of simulation models designed to study the behavior of persons, businesses, and governments—as well as society as a whole—under alternative assumptions, clearly require massive data storage and quick and flexible retrieval capability in addition to an unprecedented level of data cleanliness (editing).

Finally, we quote from a recent paper on "The Elusive Management Information System."

> Those who have delved into the Management Information System problem from a management viewpoint soon discover that the classical systems methodology for identifying output requirements simply doesn't work in a Management Information System environment. Conventional requirements analysis, as a prelude to systems design, is based on the assumption that user needs can be adequately defined in advance, whereas experience has shown that management's need for information is largely *ad hoc* in nature and cannot be predicted even by the most thoughtful and articulate of managers.[4]

What Head has said about management information systems applies equally well to statistical information systems. Increasingly, we must treat the data we have collected as a capital which can be drawn upon repeatedly to satisfy a variety of user needs. Implicit in this approach is the requirement of fast, flexible, and inexpensive access to such data.

While the preceding section suggests the goals that the statistical office should try to reach through automation, the attainment of these goals is enormously difficult, so that they are probably best regarded for the moment as merely indicating the *direction* in which the statistical office should strive to go. Here, we can no more than hint at the prerequisites for achieving the goals, and at the immensely complex obstacles which lie in the way, before turning our attention to the more *immediate* problems which must be coped with in order to carry forward successfully a practical program of automation.

First, as already suggested, the statistical office must develop facilities for storing the data collected in highly disaggregated form (to be referred to as "microdata sets"), and for retrieving quickly a large variety of unanticipated, as well as anticipated, tabulations. Clearly, if only aggregates are stored or retained, this limits the flexibility for subsequent reaggregation and cross-classification. The storage of microdata has major implications for processing: the microdata must be particularly well edited, adjusted for nonresponse, weighted in the case of sampling, and so on—and these operations must be carried out in relation to individual records. Thus, statistical standards related to concepts, classification, edits, and imputations must be much more rigorous than formerly when the statistical end products were, by comparison, much fewer in number, more highly aggregated, and largely preconceived.

Second, statistical standards related to collection (coverage, fieldwork, and so forth) must be tighter than they have been previously in order to ensure that the large variety of detailed tabulations and cross-classifications will have predictable (and acceptable) margins of error. Without elaborating on this point, what

[4] R. V. Head. *Datamation*, Sept. 1, 1970.

is involved is an increasing responsibility to follow up the nonrespondents (probably necessitating a greater role for the field function) and greater integration of follow-up procedures for different surveys, which in turn requires careful scheduling and planning. Put differently, our thesis is that the integrity of the data must increasingly be assured during its collection and processing at the level of the individual observations, since we can no longer design our surveys with a view to minimizing the error of a few predetermined aggregates only.

Third, the more data is accumulated in the statistical office, the more important it is for the statistical office to carry out analyses of the data in order to identify the relationships existing between the data derived from different surveys and other sources, with a view to providing users "guided tours" of the character, limitations, and potentials of the data. Without such "guided tours," the abundance of data can be just as confusing as the lack of it. Such "guided tours," and the analyses on which they are based, can have major beneficial feedbacks into the statistical system: the more data is available, the more it is important to standardize concepts, to study and weed out inconsistencies, to structure the data base in a coherent manner and, in general, to carry forward what is known as the function of statistical integration. In sum, the statistical office should strive to provide an analytical information service, rather than just a data-collection facility—a very challenging task.

The last problem we shall point out is that of privacy and confidentiality. Previous practices of publishing only predetermined aggregates lent themselves more easily to confidentiality checks than does the retrieval flexibility mentioned above. The latter carries with it dangers of inadvertently disclosing confidential information, which must be overcome. Moreover, the very process of storing individual returns in machine-readable form and building up facilities to interrelate them creates fears in the public mind in regard to privacy. These fears must not be disregarded on the grounds that they are unfounded, since statistical offices merely require individual observations as raw material to produce aggregations which remove identifiable events.

Later we shall discuss in more detail some of the points raised above. Now we turn to the first of our immediate practical problems in regard to automation; namely, data problems.

DATA PROBLEMS

As indicated above, the storage of individual data (microdata sets) and the facility for retrieving unanticipated tabulations necessarily put an increased strain on data reliability. The whole theory of survey sampling and census-taking evolves around the minimization of the error of a particular statistic for a given overall cost. While some attempts have already been made to generalize this theory to multipurpose surveys, this generalization still refers only to a set of *anticipated* needs. It is difficult to conceive how any theory could be developed to guide us in the face of completely unanticipated retrieval requirements. At the very least, however, we have to strive toward flexibility in systems and operations and experienced staff to adapt the allocation of effort in the light of emerging needs.

At the heart of most automated survey-processing systems is a subsystem to accomplish the editing of individual returns, their correction, and the imputation for nonrespondents. We believe that the proper utilization of the edit and imputation subsystem in an automated survey is vital in overcoming the data problems referred to above. However, this subsystem is typically the most complex of the automated survey-processing system and great care must be taken that its implementation does not get out of control (in terms of time, resources, testing the results, and so on). We shall return to this matter later on. Here we want to emphasize the role of automatic editing and imputation with respect to the quality of the stored data.

It appears to us that automatic editing and imputation must have a dual role in the processing system. The first role is the most visible one; namely, the identification of inconsistencies in the reported data and their elimination, and the identification of nonrespondents with the creation of some imputed data for them. This must, however, be an iterative process. The computer is not only capable of carrying out the necessary edits and imputations. It is also capable of providing print-outs indicating in summary form the impact of the changes in the data due to computer processing. Whenever this impact (in terms of both gross and net changes) is small, there is no point in attempting to overrule the computer through subsequent manual corrections. Where the impact is large, however, this can be interpreted as a danger signal. The input data in such cases can be examined and external manual corrections can be made (thereby overruling the original computer corrections). This strategy facilitates the utilization of human intervention in areas where it is most beneficial. At the same time, by reducing the scale of the human intervention required, it facilitates a more careful and considered intervention.

The most important intervention probably relates to nonrespondents. It is unrealistic to assume a 100 percent response rate even though we aim to approximate it. Computer editing and imputation and the resulting summary measures can focus the follow-up effort in the areas where it can have the most important impact.

In this connection, it is important to emphasize that the storage of microdata and the availability of flexible retrieval systems should have an impact on what are considered to be important, or unimportant, cases of nonresponse. Traditionally, when we were aiming to produce only a few major aggregates from a survey, we identified as the most important nonrespondents the large units; or, in the case of repetitive surveys, units which had a volatile pattern of reporting from survey to survey, thus making imputation on the basis of historical data difficult. This strategy needs to be reassessed in the era of microdata-set storage. Some of the unanticipated retrieval requirements may well relate to small survey units for which we have traditionally tended to allow a higher nonresponse rate. At the very least, in addition to following up the "important" nonrespondents, we should also follow up a sample of the remainder in order to evaluate the impact of imputations on all types of nonrespondents.

Increased flexibility of retrieval may expose data weaknesses not only within individual surveys, but also between surveys. One of the objectives of automation of individual surveys is the creation of a data base permitting easy retrieval of

information from the files resulting from the particular survey. In a sense, there-
fore, we are in the process of developing a multitude of small data banks whose
subject-matter scopes overlap. As more such data banks are developed, they can
create incompatibility problems of immense complexity. Integration tools are
urgently needed to overcome these problems. One important integration tool is
a central register, containing the universe of units included in individual surveys.
Such a register can facilitate an unambiguous definition of the scope of different
surveys; it can provide a high-quality frame for the selection of samples; it can
provide a tool for the consistent classification of identical units in different
surveys; and it can facilitate the comparison of microdata collected in different
surveys. Statistics Canada is in the process of establishing a central register covering
all business (and other, e.g., institutional) reporting units, taking into account
the many complexities in the identification of respondents in different business
surveys.

At least as important as a uniform frame for related surveys is the problem
of uniform concepts. What is needed with increasing urgency is a data-element
dictionary identifying the statistical concepts underlying each data element (e.g.,
"hours worked," "total retail sales"), which can serve as a standard convention
enabling the unique referencing of identical data elements with identical terms and
different data with different terms. Such a data-element dictionary would also
serve as the foundation for an overall system of file descriptions of all machine-
readable files. Clearly, such an overall file-description system is a necessary part
of the strategy of flexible retrieval and other statistical manipulations: we must
be able to reference identical data by the same terms for the convenience of users.

One of the most important data problems facing the statistical office is that
of drawing users' attention to the degree of reliability of the statistics released.
This was difficult enough in the era of preconceived tabulations. It will be ex-
ponentially more difficult in the period of flexible retrievals. We must develop
general models incorporating the contribution to the total mean squared error of
sampling error and reporting errors, as well as conceptual, classification, and
processing errors. While it might not be possible to estimate the mean squared
error of each individual retrieved aggregate, this should be estimated for a sufficient
variety of such aggregates to enable the development of general models of the
mean squared error which could provide guidance to users of the order of magnitude
of the errors associated with the data.

Statistics Canada has developed a very flexible retrieval system which is
being implemented in connection with the 1971 Population Census. We are in
the process of developing generalized tables which will be provided to any user
who obtains a retrieval of census data, and which will show, for each of several
broad classes of estimates, the size of the mean squared error as a function of the
size of the estimate. Similarly, we have developed such general tables for the
sampling error of Labour Force Survey data. However, these efforts must be
considered as only the very beginning of a program to develop tools which will
enable users to identify and appraise errors.

An important part of controlling data problems and of developing the
necessary error models referred to above is the conduct of a rigorous program of
evaluation of surveys. Traditionally, we are used to computing sampling errors,

and the formulas used in the past for these computations incorporate the impact of the random reporting and processing errors. Such formulas, however, do not measure the bias which may arise from any of a number of possible sources. Starting with the 1961 Census, we began to conduct, as part of our quinquennial population censuses, extensive evaluation programs aimed at measuring the different components of the mean square error. These evaluation studies had far-reaching impact on the methodology employed in the 1971 Census. Several of our other surveys incorporate control features which provide as a by-product some evaluation information. However, we have a very long distance to travel before we can claim to understand with the necessary clarity the sources of errors in surveys and the complex forces at work in generating them.

PROJECT IMPLEMENTATION PROBLEMS

While data problems are probably the most fundamental of those underlying the automation of statistical processes, the most immediate and frustrating problems relate to project implementation.

Several of our surveys now utilize machine-readable files to accomplish the mailing out of questionnaires, the checking in of respondents, the identification of nonrespondents and the corresponding follow-up actions, the editing and correction of the returns and imputations for nonrespondents, tabulations, and variance tabulations. As indicated above, we are in the process of automating a comprehensive central register of business units. We have automated the data processing of the 1971 Census, including the editing, correction, sample weighting, and production of predetermined tabulations, as well as the creation of a general retrieval system to permit the retrieval of any cross-tabulation for any area based on the census data. We also have established a general time-series data bank capable of storing tens of thousands of time series and retrieving any number of them for print-out or subsequent manipulation. However, with a few notable exceptions, our accomplishments to date have been accompanied by considerable frustrations and delays. We have made a conscious effort to learn from our own experience but there are still many lessons ahead of us.

In the London paper referred to above, we articulated five important lessons we learned from our previous automation experience. Although, in the light of our subsequent experience, these lessons need to be supplemented, they are, as far as they go, as valid today as they were two years ago. It might be worth reproducing these five points here in summary form.

1. We emphasized the importance of developing an overall design for the automation project, identifying its individual components and specifying in detail at least some of them before programming begins. We emphasized that the development of such an overall design may go through several iterations and that a number of areas of functional responsibility are necessarily involved in it: subject matter, survey operations, survey methodology, computer systems analysis, computer operations, and so on.

2. The ultimate system implied by the overall design should be split up into fairly self-contained modules made up of subsystems and program packages, with minimum interaction between them, to make it possible to lift out and replace

any one as necessary. The modules should be small enough for the implementation to be carried out in a reasonably short time. Whenever possible, the systems and subsystems should be capable of being broken down into programs simple enough for an *average* programmer to implement.

3. Specifications for the corresponding computer systems have to be developed in collaboration between experts in survey methodology, computer systems, and subject matter.

4. While the importance of working in an interdisciplinary milieu was emphasized, it was also felt that it is individuals, rather than committees, who produce the most efficient work. Therefore, it was stressed that individual responsibilities must be assigned within a working team, and that a project manager must be designated, whose particular responsibility it is to ensure adherence to objectives and budgets, to monitor and coordinate the implementation, and to ensure communication between areas of functional responsibility involved in the project.

5. Finally, the vital importance of effective communication between members of a project team was emphasized.

Even though we were trying to live by the prescriptions that we have just outlined, we nevertheless encountered some difficult problems of implementation during the last two years. One of these is that the stress on the necessity for an overall design and the need for complete and unambiguous specifications resulted in some unduly long gestation periods. Having emphasized that computer systems, once implemented, are difficult to change, having emphasized the need for complete and unambiguous specifications, we tended to inculcate a sense of finality in those who had the task of writing such specifications. Yet, in spite of this, we did not succeed in eliminating the need for changes in specifications after programming began, or even after some of the subsystems were developed and tested. This was due to the fact that the ultimate systems we were aiming at could not be thought through in complete detail in advance of their implementation. The full impact of some complex subsystems like editing and data correction could not be anticipated in advance. Moreover, if one took sufficient time to think out and specify in advance all the requirements of the system, the time consumed would likely be so long that the environment would change in the meantime, and the requirements of the system also. So, even where detailed specifications were prepared and were successfully implemented, it turned out that the specifications resulted in some unanticipated impact on the data and changes to the systems had to be made. Given the complexity of the ultimate systems we were aiming at, changes in specifications resulted in considerable delays and they often had unanticipated effects on other parts of the system (due to the interaction of the various subsystems).

While modular programming helps when changes to programs have to be made, it appears that we have to extend the concept of modularity from programming to implementation. In fact, what we may need is a phasing of implementation. By phased implementation, we mean one of two possible alternatives: either the implementation of a subset only of the overall system, with additional subsystems added to it gradually as required; or the initial implementation of a skeletal version of the whole system, with provisions for its subsequent expansion and

refinement. The latter is the one we now favor. We want to emphasize that we are not advocating compromising the ultimate objectives. Rather, we advocate, where feasible, their attainment through several generations of gradually more complex systems.

The process of phased (or gradual) implementation just outlined has several advantages. First of all, it reduces the complexity of design testing and implementation, since *at any point in time* one would be aiming at a more modest incremental goal. The achievement of the more modest goals provides a convenient checkpoint for everyone participating. The particular features of the new systems can be tested thoroughly in a real application (rather than just using test data). Shorter implementation periods also give us a better sense of achievement.

This point is more important than it may at first appear. Personnel responsible for the operation of the survey are typically faced with an overwhelming task at the time of automation. They are required to learn all the complexities of operating a major new system, and they are also required to monitor and test the data produced by the new system and to compare them with the results of the old one. (This is necessary both as a system test and as part of the effort to monitor and control historical continuity of series.) Since this kind of testing of a new system cannot be accomplished using test data alone, the personnel in question are often faced with the parallel running of the old system and the new. They seldom have sufficient skilled resources to do this, since the publication of statistics (usually based on the old system) assumes priority, while the testing of the new system and the analysis of the data produced by it must take second place. The consequences of this, in terms of a long period of double work load, inadequate testing, acceptance on the basis of insufficient evidence, and so on, are obvious enough. By contrast, in a phased implementation of a mail survey, for example, the mail out and check in can be automated and implemented after a relatively short period of testing. The implementation of automatic mail out might immediately save some resources, which could then be available for the testing of subsequent subsystems. The next module to be implemented might, for example, be the tabulation module (leaving editing and data collection a manual operation for the time being). Finally, a simple version of editing and imputation might be implemented, which could gradually be refined to more sophisticated versions.

Thus, the first major advantage of phased implementation is that it reduces the impact of automation on the operating personnel (or, rather, it phases this impact over time in a more realistic fashion). It minimizes the problem of running a dual operation and minimizes the management problem which arises when one has to accept or reject the end result of what might be several years of development. A second advantage of phased implementation is psychological. The successful implementation of each new generation of the automated system and its acceptance for operational use provides a psychological boost, not only to members of the developmental team, but also to the receiving division: it provides them with tangible benefits.

The third advantage of phased implementation is that the implementation of each phase provides management with a checkpoint at which a decision can be made as to whether additional features of the intended overall system should be proceeded with eventually, proceeded with immediately, or abandoned altogether.

It is very difficult to make a rational judgment concerning the end result of a major developmental project which, after the investment of many man-years, runs into serious problems: the temptation is inevitably to try to fix up such systems and salvage as much as possible. In the case of phased implementation, an earlier version of the same system is always available as a backup to meet current operational needs, and problems with newer versions of the system can be resolved without major crises.

There may seem to be an inconsistency between the concept of phased implementation presented here and our earlier views, presented in our paper in London:

> The computer is more than a hardware equivalent of a host of clerks. The application of computer processing is a fundamental parameter which must be taken into account in the design of surveys with the aim of rendering efficient the survey as a whole, given its objectives, not just its parts. In addition, it is our view that in order to derive the full benefits of computer processing, one should in general actually plan to go all the way: to comprehensive automation of all phases of the survey.[5]

The danger of piecemeal implementation is that it might tend to lock us into the framework of what exists. The danger is real. We must emphasize that modular implementation is no substitute for an overall system; rather, it should be looked upon as a way of accomplishing a comprehensive change with as little pain as possible. The *outlines* of the overall system embodying the final automated survey must not be shortchanged. The modules that are selected for implementation must be consistent with the overall design. *Moreover, they have to be conceived in a general enough fashion to be capable of expansion and change.* To the extent, however, that these modules are required to be compatible with the existing survey (or with some relatively small modifications of it), this may undoubtedly be the source of some conflicts. As additional modules are implemented, previous modules may have to be changed somewhat. We believe that so long as these changes are anticipated, their impact can be minimized. Even so, it is probable that if one were to compare the cost of implementing the ultimate system at once with the cost of implementing it in a gradual fashion, phased implementation might well be the more expensive of the alternatives. However, such a simplistic cost calculation would be based on the assumption that the systems are successfully implemented on schedule. This may not be the case. Thus, one should look at the incremental cost of phased implementation as insurance money. It may well be worth *planning* to spend, say, 50 percent more on the implementation of the ultimate system in order to ensure that we do not have to pay an *unplanned* 300 percent more.

Fortunately, important tools are being developed to assist us with the problem of modular implementation. A series of generalized programs have been developed in Statistics Canada (and elsewhere) capable of coping with a particular phase of processing of a variety of surveys. For example, we have a generalized program to produce address labels for mail-out purposes; another generalized program for the comparison of two files, for example in preparation for historical editing;

---

[5] Fellegi and Goldberg, *op. cit.*, Book 1, p. 158.

another generalized program for editing and data correction (embodying a variety, though certainly not all, of commonly used editing and correction techniques); and retrieval programs capable of producing tabulations (including weighted tabulations) from a number of files.

We believe that these generalized programs and their extensions will play an increasing role in the automation of our surveys. They have several major advantages over special custom-made systems. First, because they are generalized and widely used, they soon become reasonably well debugged. Secondly, because they are widely used, they must be well documented. Thirdly, their application reduces the implementation time in that it avoids the time-consuming process of program design, coding, testing, and debugging. Fourthly, and most important of all, they facilitate experimentation and relatively easy subsequent changes.

This last point requires elaboration. When specifications are prepared for a complex custom-made program, the consideration of alternatives is of necessity a theoretical exercise. Because of the complexity of what is to be implemented, assessing and testing its impact would require very large volumes of test data and almost impossible amounts of manual calculations. Thus, very often one finds out the full impact of such programs on the data only after they have been implemented at a considerable cost and with considerable effort. The most fundamental advantage of generalized programs is that because of their ease of implementation, they facilitate the testing of alternatives at a reasonable cost.

While generalized programs may not have precisely all of the features required for the implementation of the ultimate system, they may have an exceedingly important role to play in the scheme of modular implementation: they provide an easy means of implementing skeletal versions of the ultimate system, which may subsequently be replaced, if necessary, with custom-made programs at a later date. In fact, as successive versions of generalized programs are developed, they may be capable of coping with very complex and sophisticated processing requirements as well.

One final point on project implementation: It is essential that the anticipated benefits of automated surveys, as well as their cost and time of implementation, be estimated in advance, and that these estimates be checked out after implementation in the form of post-implementation audits. This is an important feedback of the automation process, enabling all concerned to improve gradually their ability to estimate both costs and benefits, as well as enabling the management of the statistical office to derive general lessons regarding policy that can be applied to future projects.

CONFIDENTIALITY PROBLEMS

It is a salient feature of statistical information that it always relates to a well-defined population, rather than to a particular respondent. However, if a population of interest is sufficiently narrowly defined, it may contain only one respondent. If this respondent can be identified on the basis of the statistics, their release would violate statistical confidentiality. Undoubtedly, this type of potential disclosure is not overly difficult to detect, although a manual process of checking each retrieved tabulation prior to its release may be expensive and

time consuming. Much more difficult is the problem of so-called residual disclosure. Residual disclosure occurs when a number of tabulations are released, none of which violates confidentiality in itself, but which together enable a user to deduce information about a single identifiable respondent. The simplest example of residual disclosure occurs when one statistical aggregate is based on all but one of the respondents involved in another published aggregate and, consequently, the difference between the two aggregates discloses information about that one respondent. Unlike direct disclosure, residual disclosure is notoriously difficult to detect, even in the case of preconceived tabulations. Obviously, a policy of flexible retrievals exponentially increases the problems of checking for residual disclosure, since each new retrieval must be checked against all the previous retrievals.

In a recent article, one of the authors has developed a precise mathematical theory dealing with tests of residual disclosure. In all but the simplest situations, the practical implementation of this theory would involve prohibitive amounts of calculation. How, therefore, can the overriding objective of protecting the confidentiality of statistical returns be reconciled with the need to utilize the data that has been collected as extensively as possible?

In connection with the general retrieval system mentioned earlier, which is being implemented on the 1971 Census data base, we have developed a general approach to the solution of the confidentiality problem. As far as direct disclosure is concerned, the retrieval system automatically checks each tabulation cell to ensure that it satisfies the predetermined requirements of statistical confidentiality. (This test is generally based on the requirement that each aggregate must relate to more than a predetermined minimum number of respondents.)

In order to avoid the more complex problem of residual disclosure, each *ad hoc* tabulation will be subjected to a small amount of random disturbance. This will involve a random reallocation of a small proportion of selected respondents within the table. The particular strategy of reallocation is so designed as to minimize the impact of this random disturbance on the accuracy of the data. This random disturbance will prevent users from manipulating the retrieved statistics in order to derive other statistics which were not tabulated. In the case of the simple example above, if let us say, someone took the difference between two published statistics and observed that the difference appeared to relate to a single respondent, he could not be sure whether this was a real event, or rather the result of a random disturbance on one or the other of the tabulated aggregates.

While, undoubtedly, the strategy of random disturbances will increase somewhat the mean square error of the statistics, this may well be a necessary price for making the data available at all in the detail requested. The procedure can be automated and, thus, it can be fast and inexpensive. Even apart from this consideration, however, the strategy of random disturbances appears to be a "necessary evil," since a rigorous checking for residual disclosure is impossibly complex even with our modern computers; it is simply unmanageable using manual methods. Statisticians are well used to making compromises between the often conflicting requirements of cost, reliability, and timeliness. It now appears that a new dimension of the compromise must involve confidentiality and reliability.

As we have seen, the requirement for retrieval flexibility sharpens the need for rigorous procedures to safeguard against statistical disclosure. Another methodological and technological development which also reemphasizes the fundamental importance of safeguarding confidentiality relates to record linkage. The association of information with particular respondents was traditionally of no substantive interest to the statistical office; identification has traditionally been only a means of control during the collection of data. More recently, with the advent of computer technology and some relevant statistical methodology, the identification of respondents has assumed an added importance to statistical offices: they may wish to compare information collected in different surveys, either for editing purposes or to enrich the information content of one or the other of the surveys through record linkage.

While the interest of the statistical office in record linkage is entirely statistical, the process of linkage aggravates the confidentiality problem. Linkage usually involves the identification of respondents through a unique number, or through descriptive information—name, address, and so forth. It appears that society is not only concerned about the ends but also about the means as far as privacy is concerned. The statistical office must emphasize at every opportunity that the data it collects are not available to other government agencies for any purpose whatsoever; that the data are made available as *statistics* in the form of tabulated aggregates. This assurance, must, however, be followed up by a consistent policy of ensuring the security of data both physically (in the sense that the schedules collected are secure and accessible only to employees of the statistical office) and in its publication policy. The development of precise rules for disclosure testing and their implementation is, therefore, of paramount importance.

Given the concern of society regarding record linkage, we believe that this is not the right time to engage in it on a massive scale. At any rate, even from a substantive statistical point of view, record linkage may have serious limitations. First, it is hardly of benefit when the records involved are obtained from two independent sample surveys. After all, the probability that the same person is involved in both surveys is approximately equal to the product of the sampling fraction in the respective surveys: typically a rather small quantity. Hence, for maximum benefit, at least one of the surveys involved in record linkage should be a census or an administrative file with no sampling, or little sampling, in its coverage.

In the case of administrative files other problems may be present: the scope of their coverage is usually determined by the administrative requirements for which they were set up, rather than by any statistically meaningful definition. Also, the concept underlying the information recorded in administrative files might well be different, or it might be differently coded, differently edited, and so on. The problem of scope of coverage may diminish as the major administrative files involved in our modern state become more and more universal. The problems of concepts and care of processing, however, can only be reduced through close collaboration between the statistical and administrative agencies involved.

Finally, the dual problem of error of record linkage should also be noted: some records which should be linked will not be linked (in a broad sense, this is not dissimilar to the problem of nonresponse in surveys) and some records which

should not be linked might be linked (in the same broad sense, this phenomenon is somewhat similar to the problem of response errors). Assuming the absence of a unique identifier, these two problems can only be controlled if there is adequate information recorded in the two files for carrying out the record linkage with an acceptable standard of precision, and if the resulting linked file, as well as the residual unlinked files, are subjected to careful editing and, if necessary, follow-up.

While most of the technological and methodological problems have been solved, record linkage is seen to present real conceptual problems and problems related to public concern. An alternative method of bringing together the information content of two files has recently been proposed. This pseudolinkage method consists essentially of allocating information from one file to another on the basis of some common core of information and a set of assumptions. For example, if both surveys contain age-sex and labor-force information, and one file contains, in addition, some expenditure data, while the other contains some income data, then the distribution of expenditure within an age, sex, labor force cross-classification cell can be estimated from one survey, and this same distribution of expenditure can then be imputed randomly to the other survey file within the same age, sex, labor-force cell. Clearly, any such activity would be based on the assumption that the distribution of the substantive characteristic which is to be imputed from one file to another is entirely determined by the characteristics which are common to the two files. This is equivalent to assuming that whatever correlation exists between the characteristics that are not common to the files is entirely explained by the common information which is the basis of the imputation. In the case of the example above, the imputation would be based on the assumption that whatever correlation exists between income and expenditure is entirely explained by age, sex, and labor-force status (or, put differently, that within an age, sex, labor-force-status cell, there is independence between income and expenditure). In this example, such an assumption would clearly not be reasonable. The trouble is that even in situations where the assumption does not appear to be blatantly unreasonable, one can never be sure of its validity without testing. In important applications, one should not assume independence in the sense explained above, but rather should take a special survey to test for it. The hypothesis is, of course, capable of being tested if a survey is appropriately designed for it.

In conclusion, we assume a very cautious posture with respect to record linkage. At present, we are applying it, for example, in situations where two lists have to be made free of duplication in order to arrive at a single comprehensive mailing list. We have not applied it as yet for the actual bringing together of substantive data from different surveys. We have some plans for applying the pseudolinkage outlined above in relation to data collected in two recent surveys on incomes and expenditures, and the 1971 Census. The validity of the pseudolinkage techniques, however, must be tested again and again, probably separately for each application. We believe that we have a long way to go before we can claim to have created adequate data bases from each of our surveys individually. We would attach a much higher priority to developing uniform statistical standards between surveys, uniform concepts where applicable, standard codes, and flexible retrieval systems. At any rate, it appears that these would be prerequisites for record linkage as well: it is hardly worthwhile to bring together data between

surveys if deficiencies in the concepts involved, and in the standards of collection, coding, and so forth, would render the linked information of dubious validity.

CONCLUDING REMARKS: SOME ISSUES CONFRONTING MANAGEMENT

In spite of a somewhat cautious tone throughout this paper, a logical analysis of what has been said will lead to the conclusion that the process of automation is, in spite of the frustrations involved, desirable and essential. The fundamental reasoning that leads to this conclusion starts with the fact that the extensive utilization of data already collected (together with the collection of additional series as required) is the only conceivable way of satisfying the variety of information requirements of society. As indicated above, this information need cannot be fully predicted in advance and, therefore, the posture adopted by the statistical office should be one of readiness. Thus, there is a compelling need for storing microdata and creating automated, flexible retrieval systems.

Starting from this premise, and considering the strain that such extensive utilization of data puts on data reliability, the cleanliness of the stored microdata becomes an essential consideration. Thus, we are inexorably led to the necessity of automated editing and correction at the microdata level. We are also led to other considerations of statistical standards, such as the necessity for complete and up-to-date frames for surveys and censuses, which in the case of economic surveys, renders the requirement for a central register of business units almost unavoidable. An important source for the updating of such a register is the surveys which are currently using it. Accordingly, we are led to the desirability and necessity of using the central register of business units either in the form of a central mailing list or as a coordinating device with which individual survey lists are kept fully in parallel (at least in the respective subject-matter fields in which the particular surveys operate).

We have more than a decade of experience to indicate that the mere existence of a central register does not ensure that the mailing lists of individual business surveys conform to it or feed it in a systematic fashion with respect to the updating of information. In the heat of current operations, the emphasis is on completing the survey—and often the liaison with the central register falls by the wayside. This problem can only be remedied by automating the information flows between the central register and the user surveys. As a consequence, we are led to the necessity both of creating a machine-readable central register and of creating machine-readable mailing lists for individual surveys, the two to be connected by an automatic process of information flows. Starting from the nature of information utilization, we are thus led inexorably to the full process of automating individual surveys and creating coordinating devices. One of the challenges involved in the management of statistical offices is the recognition of this fundamental unavoidability of automation.

Another challenge, which, of course, is at the heart of most management problems, is how far and how fast we should go. We have no magic solutions to offer. Individual decisions will have to be made according to individual circumstances. However, a few important considerations basically related to costs and benefits are sufficiently general to be applicable to almost any situation.

Short-term costs are very real and highly visible (although not necessarily accurately predicted at the beginning of development). For example, the total cost of developing the generalized retrieval system we mentioned earlier, which will be implemented in connection with the 1971 Census, was estimated to be in the neighborhood of two and a half million dollars. About a third of this figure was expended in systems analysis, programming, and computer testing. The cost of the time-series data bank, also mentioned earlier, is estimated at about $750,000, of which somewhat more than half a million dollars involved systems analysis; programming; and management, professional, technical, and clerical staff.

However, once such systems are developed, the cost of effective data dissemination through them is relatively small in comparison to the costs involved in the collection of statistics. We are firmly convinced that the benefits of these two systems will be far in excess of the costs. The precise estimation of these benefits in advance is, however, exceedingly difficult. Given the long developmental time required for such systems, any advance estimation of their benefits would necessarily involve forecasting the information requirements of users some years in advance—a very knotty problem. In the case of the census information system, one can speculate, however, that a relatively few significant uses of the system would repay its entire development cost. A single bridge that will not be built at the wrong place, the more precise delineation of one or two urban renewal areas, a single instance of a more precise identification of poverty areas which might receive economic aid from the government, a few small-scale household surveys that need not be taken because the data will be available from the census through the system—any one of these applications might repay the development cost of the system.

The estimation of benefits to users is rendered particularly difficult because users are a heterogeneous group, both with respect to their requirements and to their own capability of utilizing large amounts of information. It is relatively safe to forecast that user requirements will increase in respect to the amount of data required, the type of disaggregations, the kinds of manipulations, the format and medium of the retrieved data, and the desired retrieval response time. However, users have a wide variety of requirements and varying levels of computer sophistication. For example, a few users may genuinely require, and be able to take advantage of, access to speeds approaching "real time," while other users might be satisfied with overnight, or even longer, turnaround. If the statistical office wants to keep in step with the users having the most advanced requirements, then it must be prepared to take greater risks in terms of development costs. Given the rapid changes in computer hardware and the software industry, forecasting the requirements of advanced users several years in the future might appear to be a visionary process. Yet, again given the long development time required by automation, we believe that only a general policy commitment to attempt to *keep ahead* of the information requirements of users will succeed in ensuring that the statistical information service does not *fall seriously behind* the demands put on it.

Although it is not overly difficult to argue about the ultimate benefits of automation in relation to its cost, this fact by itself does not necessarily guarantee a green light for automation. There are too many other activities that the statistical office can undertake which have impressive cost-benefit ratios. Thus, cost-benefit

considerations, while they might reaffirm the ultimate need for automation, do not necessarily provide generally applicable guidelines with respect to the question of how fast we should proceed with it in comparison with other activities—for example, the alternative of undertaking several new surveys. Here again, we believe that the development and utilization of generalized programs should have a high priority because of the broad facilitating nature of such programs. The automation in a one-by-one fashion of individual surveys in a large statistical office like Statistics Canada, having hundreds of surveys under way, is a gigantic task. Moreover, if one takes into account that custom-made programs which are parts of complex systems are difficult to change—while at the same time, due to alterations in technology, in methodology, and in user needs, various changes to these systems are unavoidable—one can foresee a situation where the updating of earlier automated systems could keep one-half to two-thirds of a large programming staff occupied. Modular programming can diminish this ratio. Nevertheless, we believe that only through extensive development and utilization of generalized programs will this ratio be significantly improved in favor of new developments as against changes to existing systems.

In our paper delivered in London, we dealt extensively with the problems of management involved in the reorientation of personnel to the era of automation and its implications: the initial fear of the computer, the fear of loss of control, the difficulties of infusing a genuine spirit of interdisciplinary cooperation, the role of training to alleviate some of these problems, and the need for documentation and systematic follow-up of the progress of the various projects. We shall not repeat this material here, although we believe it is still relevant. However, we do want to add two additional points. First, it is difficult enough to overcome the initial fears when they relate to the unknown. It is particularly difficult to overcome such fears successfully when they are based on actual unpleasant experience with automation. It is, therefore, essential to try to achieve a few notable successes at the beginning. This points, once again, to a strategy of modular implementation of automated systems. No amount of abstract indoctrination can take the place of the impact of success, and this applies not only to the management of subject-matter divisions but also to all the staff involved in operating surveys, including technical and clerical personnel. The second point concerns the difficulty of incorporating personnel involved in the operation of current surveys in interdisciplinary development teams, due to the lack of spare resources which can be devoted to new developments. Again, phased implementation and generalized systems might help (since they reduce the scope of dual operations). Ultimately, however, it may be necessary to have a special group of personnel which could be seconded to operating divisions for the duration of developmental projects and their implementation.