

GOALS FOR STATISTICAL USES OF ADMINISTRATIVE RECORDS: THE NEXT TEN YEARS

Thomas B. Jabine, National Academy of Sciences, and Fritz Scheuren, Internal Revenue Service

The purpose of this paper and of today's panel session is to broaden the discussion of issues surrounding statistical uses of administrative records. The paper introduces six specific goals for expanded activities over the next ten years, discusses recent developments relevant to these goals, and describes several important elements of a strategy for achieving them.

During the past several decades, administrative records have become a major element in the statistical programs of Federal and State governments in the United States. The three principal uses are: for program statistics; for general and multipurpose data systems based solely on administrative records (for example, the Statistics of Income program and the Continuous Work History Sample); and for enhancement of information collected directly for statistical purposes in censuses and surveys. Important uses in the last category include, but are not limited to: development of sampling frames, supplementation of data collected directly, and evaluation of census and survey data (some other uses are discussed under Goal 4).

The use of administrative records in statistical programs is likely to increase further, given the ever-increasing costs of direct data collection and the general consensus on the need for reduction of the overall burden on the public to report information for both administrative and statistical purposes. The shift to greater use of administrative records, while it can benefit the statistical system in many ways, is not free of problems. These problems relate primarily to the quality and relevance of data available in administrative record systems, to the technical and legal difficulties in gaining access to administrative records for statistical uses, and to public perceptions of the process of integration of statistical and administrative records and the question of how such perceptions may affect the ability of statistical agencies to collect data directly in censuses and surveys. Frequent changes in the character of administrative records systems pose an added difficulty.

In view of the present environment and trends, we feel strongly that statistical agencies and the system as a whole should adopt an activist strategy in developing new statistical uses of administrative records. This does not mean that we favor rushing into new applications without adequate testing and consideration of their broad implications. What we do argue for is system-wide planning to make the best possible use of administrative record systems and to seek some degree of control over features of those systems that affect their suitability for statistical applications.

The next section of this paper discusses six specific goals for statistical uses of administrative records. These goals are taken from the final section of the just-completed two-volume handbook, Statistical Uses of Administrative Records: Recent Research and Present Prospects (Internal

Revenue Service, 1984a, Volume II, Section VIII). The following section of the paper identifies some general themes that are relevant to some or all of the specific goals, and suggests some essential elements of a broad activist strategy for statistical uses of administrative records. The final section summarizes current progress toward meeting the six goals and tries to answer the question: Where do we go from here?

SIX SPECIFIC GOALS: RECENT DEVELOPMENTS AND PROPOSALS

In this section, we present and discuss six goals for expanded statistical uses of administrative records. Figure 1 lists the six goals as they first appeared in the recently published two-volume handbook (Internal Revenue Service, 1984a, Volume II, Section VIII).

Progress in meeting these goals will depend primarily on the major general-purpose and multipurpose statistical agencies and on OMB's Office of Statistical Policy. Little progress is possible unless these agencies can agree on these or some other common goals and work together to achieve them. Such agreement is, however, only the first requirement. Collaboration must also be sought from: the Federal agencies that are the custodians of the administrative record systems to be used; the State agencies responsible for important administrative systems, such as unemployment insurance; and, where legislation is needed, from the Congress. In addition, all major initiatives must be considered from the point of view of how they will be perceived by the public, both as providers of individual data and as users of statistics. Public acceptance and cooperation by the custodians of administrative records are more likely to be secured if there is agreement on goals within the Federal statistical system.

We now proceed to discuss each of the six goals separately. Where necessary, we clarify the nature of the goal; then we summarize relevant recent developments; and then, for some of the goals, we suggest new initiatives. The length of the individual goal discussions varies considerably: this reflects, in part, variation in the extent to which each of them is being actively pursued at this time. Following the discussion of individual goals, we discuss briefly their general character and interrelationships.

Discussion of Individual Goals

GOAL 1. Explore fully and develop the uses of major administrative record systems in the conduct and evaluation of the decennial population censuses and for current population estimates.

The Census Bureau has taken several steps to explore possible uses of administrative records

Figure 1. Six Goals for Expanded Statistical Uses of Administrative Records

| | |
|---------|---|
| GOAL 1: | Explore fully and develop the uses of major administrative record systems in the conduct and evaluation of the decennial population censuses and for current population estimates. |
| GOAL 2: | Starting with the Continuous Work History Sample (CWHHS) as a basis, develop a Linked Administrative Statistical Sample (LASS), using administrative records from several agency sources to construct a longitudinal person-based system that will serve a broad spectrum of users. |
| GOAL 3. | Give high priority to statistical applications of administrative records that will enhance our ability to monitor and analyze long-term environmental health effects. |
| GOAL 4. | Make greater use of administrative records in all phases of household surveys. |
| GOAL 5. | Develop and activate a "business directory" for use as a sample frame for economic censuses and surveys and a source of geographic and industry codes for common use by all eligible Federal and State statistical units. |
| GOAL 6. | Continue efforts to develop more consistent and compatible procedures for defining and identifying reporting units in both administrative and statistical data collections. |

Source: Internal Revenue Service, 1984a, Volume II, Section VIII

in the 1990 census. An internal working group, the Subcommittee on an Administrative Records Census, Committee on the Use of Administrative Records in the 1990 Census, reviewed the proposal by Alvey and Scheuren (1982) for an administrative record census and reported its findings concerning the operational and technical problems that might be involved in implementing the proposal (Bureau of the Census, 1984). The working group concluded that there would be serious problems related to coverage, the accuracy of geocoding, and obtaining basic data such as age, race and sex for all persons. While these problems might be overcome by the use of supplementary data gathering and the development of new methodologies, the working group considered it unlikely that an administrative records census procedure could be validated and put in place for the 1990 census. However, it considered that some elements of the proposed administrative records census might be usable in a standard census and recommended a large-scale test of an administrative record census procedure in conjunction with the 1990 census.

Late in 1983, at the request of the Census Bureau, the Committee on National Statistics established a Panel on Decennial Census Methodology to advise it on several aspects of planning for the 1990 census, including possible uses of administrative records. The Panel was requested to prepare recommendations by mid-1984 on procedures to be tried during the pretesting phase. The Panel's interim report (National Research Council, 1984) does not include definitive recommendations on all uses of administrative records. However, it does recommend use of administrative records, as well as reinterviews, in content evaluation studies, and some investigation of the use of tax assessor's records to improve or evaluate the quality of housing data collected in the census. With respect to a census proposal to test reverse record check and "megalist" techniques for evaluating

coverage of hard-to-count groups, the Panel felt that more clearly defined goals and procedures were needed. However, it encouraged continuation of nonfield test research on multilist or composite list methods.

For the first 1990 census pretests to be conducted in 1985, the Census Bureau has ruled out uses of IRS data which would require the use of specially designed individual income tax forms in areas selected for pretesting. This decision undoubtedly stems from reluctance by census staff to do anything that would suggest direct linkages between tax collection and census-taking activities, especially given the recent public resistance encountered by some Western European countries to censuses involving close linkages with administrative registers (Butz, 1983). However, the Census Bureau remains open to proposals for testing other uses of administrative records in the census.

The use of IRS records for evaluation of census coverage of persons in the working age population was tested by the Census Bureau in a match of a sample of primary and secondary taxpayers listed on individual income tax returns filed in 1980 (for tax year 1979) against 1980 census records (Childers and Hogan, 1984). The study demonstrated that matching problems resulting from the use of post office boxes, rural routes and business addresses on tax returns can be overcome with proper followup procedures. The researchers concluded that the IRS frame has great potential value for census coverage evaluation.

During the next few years, changes in forms and processing systems used by IRS could have important implications for future uses of IRS records in connection with censuses and current estimates of population and income. Some examples of changes that may be considered are: inclusion of taxpayer telephone numbers on returns, coding taxpayer occupation on all returns and adding it to the master file, obtaining dependents' SSNs

and dates of birth on individual returns, integration of returns and information documents, and establishing a computerized linkage between EINs for parent and subsidiary corporations. Changes like these would also create significant opportunities for enhancing the content of SOI samples and other statistical data systems which make use of tax information.

A final comment concerning goal 1 is that even if it does not prove feasible to conduct a direct test of an administrative records census procedure as part of the 1990 census, it may still be worthwhile to construct, at least for some part of the country, the kind of matched administrative file proposed by Alvey and Scheuren (1982) and to compare counts from that file, by whatever level of geographic and other detail is available, with corresponding 1990 census counts. Such a test would have some utility for census evaluation (and might provide a sampling frame for a direct-match against the census) and would also produce valuable information about the eventual feasibility of an administrative records census.

GOAL 2: Starting with the Continuous Work History Sample (CWHS) as a base, develop a Linked Administrative Statistical Sample (LASS), using administrative records from several agency sources to construct a longitudinal person-based system that will serve a broad spectrum of users.

The LASS proposal envisions the enhancement of the present content of the CWHS with data on income, occupation, morbidity, mortality, and other topics, taken primarily from other Social Security files, tax returns and vital records and linked to the CWHS files on the basis of Social Security numbers (Internal Revenue Service, 1984a, Volume I, selection 3.02).

A prerequisite for meeting this goal is to find a formula that will remove some of the present restrictions on the release of CWHS microdata files to users other than SSA and IRS. As a first step, these two agencies have been working with the Bureau of Economic Analysis (BEA) to find an acceptable procedure for resuming release of CWHS files to the latter agency. Prior to passage of the Tax Reform Act of 1976, the BEA, in addition to being a major user of CWHS microdata files itself, played an important role in evaluating the quality of CWHS data and in creating reformatted files that were of interest and could be released to many other users.

As a result of the current negotiations, a tentative proposal for releasing post-1978 CWHS files to BEA has been developed, and its basic structure appears to be acceptable. An early formal agreement among the three agencies is hoped for. When BEA receives the new files, it plans to evaluate them, especially with respect to the quality of industry and geographic codes. Their findings will have a significant bearing on the prospects for further efforts to implement the LASS proposal.

Meanwhile, there have been several technical developments that will contribute to the development of LASS once acceptable solutions have been found to the access questions. The Statistics of Income Division of IRS has designed its annual

sample of individual tax returns to have a partial overlap with the CWHS. Procedures have been developed for computer-assisted coding of occupation from tax returns, and the possibility of using the National Death Index to add mortality and cause-of-death information to the occupation-coded 1979 SOI file is now being explored (Sailer et al., 1984). The BLS is now incorporating employer identification numbers in its Unemployment Insurance Address File: this will make it technically possible to obtain more up-to-date industry codes for single-unit employers in the CWHS.

GOAL 3: Give high priority to statistical applications of administrative records that will enhance our ability to monitor and analyze long-term environmental health effects.

Options for using administrative records in this area have been identified and evaluated by numerous persons, agencies and committees in recent years (for one example, see Committee to Coordinate Environmental and Related Programs, 1983). What is needed now is a firm long-range commitment to improve existing systems such as the National Death Index and to develop one or more linked data systems such as LASS.

Recent developments related to LASS have been reviewed under Goal 2. Other important initiatives are: developmental work on coding of occupation on death certificates (Crouse et al., 1983) and matching Census Bureau samples to the National Death Index (Rogot et al., 1983). Both of these projects involve joint work by the Bureau of the Census and the National Center for Health Statistics.

Statistics Canada has made important advances in the use of administrative records for epidemiological studies of occupational morbidity and mortality. Smith (1983) reports on administrative record sources and the database management and linkage systems used to exploit them for this purpose. The Canadian achievements offer a good target for the U. S. statistical system to aim at.

GOAL 4: Make greater use of administrative records in all phases of household surveys.

An important application of administrative records in household surveys, exemplified by the 1973 Exact Match Study, Internal Revenue Service, 1984a, Volume I 4.04), is the use of exact matching techniques to link data from administrative records to data collected directly from households and individuals in surveys. Goal 4 envisions, in addition to expanding the use of linked administrative data, the use of administrative records at several stages of survey development and execution, including: definition and stratification of sampling units, development of frames, sample selection, imputation of missing data, estimation and evaluation.

A new working group of the Administrative Records Subcommittee has been organized to review and evaluate techniques for integration of survey and administrative records in both demographic and economic surveys (further details on this and other working groups appear in the next section).

In the initial stage, members of the group are making presentations describing ways in which they are currently using administrative records in surveys. It is expected that this sharing of experience will lead to wider application of promising techniques and proposed solutions to problems that inhibit effective use of administrative records in surveys.

A special problem that occurs in connection with the linkage of survey and administrative record data is known as the "reidentification problem." This may be explained as follows: suppose that statistical agency A has conducted a household survey and has then enhanced the survey records for sample persons or households by a direct match which permits the addition of administrative data for these same units obtained from agency B. The linkage has been carried out in such a way that agency B has had no access to the individual survey records. Now it is desired to release a public use microdata file containing the linked survey and administrative data, with the usual precautions of removing all individual identifiers from the records and suppressing geographic and other detail that might make it possible for users to identify specific individuals.

Agency B, which supplied the administrative data in the first place, may have a valid interest in using the public-use microdata file for various kinds of program-oriented analyses. The reidentification problem lies in the possibility that Agency B could identify individual records in the public-use file by linking them to its own administrative files, using the common data items, and thereby gain access to the survey data for these identified individuals.

Solutions are needed to the reidentification problem, because the custodians of administrative record systems are reluctant to incur the costs of making data available for such survey enhancements unless they themselves (or at least their research and statistical components) can use the resulting files for their own statistical analyses. One avenue for possible solutions is to take advantage of existing legislation, such as the Privacy Act of 1974, or to seek new legislation. By declaring such linked microdata files to be Privacy Act record systems, it might be possible to eliminate the possibility of disclosures for nonstatistical purposes. If this were not regarded as sufficient protection, it might be necessary to develop new legislation providing for the designation of such files as protected files, i.e., files which could be released generally, or possibly only to other Federal agencies, for statistical use, but for which any attempt to identify individual records would be legally barred, with severe penalties for violations.

Another avenue for solutions is technical, e.g., by using various forms of error inoculation or "data swapping" procedures (Internal Revenue Service 1984a, Volume II, selection 6.09). A promising recent addition to this family of procedures is "blurring", a procedure designed for use when it is necessary to protect records with a small number of continuous variables from disclosure (Oh and Scheuren, 1984). For each variable the records are sorted into groups of a specified size and group averages are substituted for individual values. Blurring has significant advantages over traditional grouping procedures in

which records are grouped on all variables simultaneously. The Statistics of Income Division of IRS, with the help of a contractor, is undertaking further research on this technique.

GOAL 5: Develop and activate a "business directory" for use as a sample frame for economic censuses and surveys and a source of geographic and industry codes for common use by all eligible Federal and State statistical units.*

Efforts to convert the Census Bureau's SSEL into a common business directory, for use by all qualified Federal and State statistical agencies, have failed. The latest setback was the Administration's decision, in November 1983, to withdraw support from the proposal for broad confidentiality legislation--the so-called "enclave bill"--which included provisions for release of the SSEL to other Federal and State agencies for statistical purposes.

Before the Census Bureau and other agencies pursue attempts to obtain separate legislation covering only the SSEL, it may now be necessary to pause and make a determined effort to understand why the SSEL, which has appeared to many of us to be based on a sound concept and to have the potential of bringing substantial benefits to the U. S. statistical system, has failed to gain the necessary acceptance. Are there other approaches which would have a better chance of succeeding?

Following are some of our present thoughts on this subject. They are intended to provoke further discussion of the issues, not to provide definitive answers. First, the SSEL was planned and developed almost exclusively as a sampling frame which could be made available to other agencies as needed for their own surveys. At this time there does not seem to be a large demand from other agencies for this type of use. The agencies that do major economic surveys, such as BLS and the Statistical Reporting Service of the USDA, have developed their own frames. For some agencies, the SSEL does not meet their specific requirements: it cannot, for example, generate a list of all manufacturers who produce a specific product.

Other potential uses of the SSEL, which could be just as important, have not been provided for in its present structure. It has already been noted that the Census Bureau uses the SSEL as an aid in coding industry in the census of population. Could a modified SSEL be used similarly to provide better quality industry and place-of-work codes for wage earners included in the CWS and better quality industry codes for units included in the SOI samples of sole proprietorships, partnerships and corporations? The Census Bureau is planning a pilot study of the feasibility of enhancing SIPP data by adding economic census data on the characteristics of the establishments in

*To the extent that this goal proposes changes in tax law or the policies of this Internal Revenue Service, it represents only the view of the principal author.

which sample persons work, with the linkage being made via the SSEL (Haber et al., 1984). Along these same lines, could the SSEL be better designed to facilitate other linkages of this kind, such as linkages of census establishment data with SOI financial and tax data for taxpayer units? How could the SSEL interact with the BLS ES-202 system to improve the quality and lower the cost of industry coding in both systems?

The Treasury Department has strongly and consistently opposed the release to other agencies of SSEL information coming directly from IRS sources, including both tax returns and Form SS-4, Application for an Employer Identification Number. There are two possible ways of structuring a common business directory to satisfy Treasury's concerns. The first would be to release only information for units for which the Census Bureau (or another cooperating agency) has collected information directly. This would cover a very substantial part of the business universe, including all establishments that respond directly to the economic and agricultural censuses. A second possibility would be to ask permission from business taxpayers and applicants for employer identification numbers to transfer a limited number of items from their returns or applications to a common business directory, to be used only for specified purposes. A combination of these approaches might effectively cover the business universe.

It is difficult, except in a very few agencies like the Census Bureau, to disentangle and compartmentalize the statistical and non-statistical uses of directory information. Is the idea of a separate statistical directory unrealistic: does it limit too severely the number of agencies and programs that can use the directory effectively and the reductions in cost and reporting burden that can be realized?

A constructive dialogue on these and related issues is urgently needed to develop new proposals for reducing the high costs, financial and other, of maintaining fully independent business data systems.

GOAL 6: Continue efforts to develop more consistent and compatible procedures for defining and identifying reporting units in both administrative and statistical data collections.

This goal relates primarily, although not exclusively, to business or economic reporting units. The comparability of economic data from different statistical and administrative data systems is seriously hampered by differently defined reporting units in these systems. Some lack of uniformity is inevitable because of differing data requirements; however, much could be done to achieve greater standardization and to develop identifiers that would facilitate linkages between systems.

Some progress is being made in this area by the Establishment Reporting Unit Match Study (ERUMS) Working Group of the Administrative Records Subcommittee (one of several working groups--for further details see the next section). Detailed plans have been prepared for a sample matching study of BLS and SSA reporting unit records in one State, and the sample selection and

matching operations will soon begin. Members of the ERUMS Working Group have already become better informed about the differences between the two systems whose records are being matched. By the time the study is completed, the nature of these differences will be much clearer, and it should be possible to develop practical recommendations for achieving greater comparability between the two systems.

Even more than the other goals, achievement of this one requires close collaboration between statistical users and the custodians of administrative record systems. The primary uses of these systems are non-statistical, and changes in them to enhance their value for statistical uses are acceptable only if it can be shown that their primary uses will not be impeded in any way.

The burden on persons or businesses that report under the administrative systems is also an important consideration. In some areas, reporters are unnecessarily burdened by administrative system requirements that overlap but are not fully compatible with respect to identifiers and definition of reporting units, and that are imposed by agencies that share only limited information with each other. An example of this is the administrative systems under which employers are required to report on employment and wages, both in the aggregate and for individual employees, for the purposes of Federal and State income tax withholding, social security, unemployment insurance, and workmen's compensation (see Rich, 1983). The Small Business Administration is concerned about the burden of these requirements on small businesses, and has proposed to OMB that a Federal-State commission on tax coordination be established, as had been previously recommended by the Commission on Federal Paperwork (1977), to develop proposals for better integration of these reporting systems. Changes in these systems could have significant effects--both favorable and unfavorable--on statistical applications, so the statistical users of these systems would be well advised to seek a voice in the process and to follow developments closely.

General Nature and Interrelationship of the Six Goals

The 10-year period chosen as a time referent for these goals is arbitrary. It is unlikely that we will be able to say, at any particular time, that goal 1, or goal 3, or goal 6 has been fully achieved. Progress is likely to be incremental. Some increments could be quite large, for example, the passage of legislation permitting joint use of a common business directory by qualified statistical agencies. Other increments may seem to be comparatively small, for example, the recent addition of social security number to Form SS-4, application for an Employer Identification Number, which opened up new possibilities for linking establishment data for sole proprietorships with data on personal attributes of the proprietors. The goals should be thought of as broad objectives that can guide and focus work on statistical uses of administrative records, not as precise targets to be met by particular dates. Within this broad framework, of course, agencies will need to define more specific objectives and develop schedules for

their implementation.

The six goals are interrelated in several ways. For example, a major use of the LASS system proposed in Goal 2 would be to monitor environmental health effects. Thus, LASS would be one of the vehicles for meeting Goal 3. Goals 5 and 6 are closely related. Increased standardization of business reporting unit definitions, identifiers and classifiers would make it easier to construct a common business directory, using inputs from different record systems. The availability and use of a common business directory would, in turn, promote still greater comparability between systems.

At first glance, there may appear to be a clean break between goals relating to data systems in which the units are persons or households (Goals 1 to 4) and those that apply to business or economic data systems (Goals 5 and 6). However, it would be a mistake to think of these two classes of data systems as being totally independent. The SSEL, for example, while it was initially designed as a sampling frame for economic censuses and surveys, is also used by the Census Bureau to aid in the classification of employed persons by industry in the decennial census of population. A common business directory, provided the technical and legal problems could be solved, might eventually be used to assign more accurate industry codes to persons with earnings in the CWS.

Goal 4, the expanded use of administrative records in household surveys, could apply equally well to economic surveys. Household surveys were initially chosen for emphasis because the use of administrative records in connection with economic surveys is generally farther advanced. The frames used in economic surveys come primarily from administrative records, and other uses, such as for imputation and estimation, are frequent. However, we now feel that it would be appropriate to include both kinds of survey under this goal, and the new working group mentioned earlier under Goal 4 will, indeed, look at uses of administrative records in both demographic and economic surveys.

ELEMENTS OF AN ACTIVIST STRATEGY

The activist strategy we have proposed requires interagency coordination and system-wide strategic planning at a higher level than now exists. Besides these two basic elements, for which the need is obvious, there are several other kinds of activities that can help us to exploit the opportunities that exist for better use of administrative records in developing, enhancing and evaluating statistical data systems and to address the difficulties inherent in the process.

Interagency Collaboration at the Technical Level

Recognizing the importance and inevitability of increased use of administrative records, the OMB's Federal Committee on Statistical Methodology has given considerable attention to this subject since 1978, when the Subcommittee on Statistical Uses of Administrative Records (SUAR) was established. The SUAR Subcommittee compiled information on 38 major administrative systems. It prepared a report (Office of Federal Statistical Policy and Standards, 1980) which reviewed several

important uses of these systems for statistical purposes and presented some broad recommendations, addressed both to the statistical agencies and to the program agencies which maintain administrative record systems, for steps that should be taken to facilitate ongoing statistical applications of administrative records and to encourage the development of new ones.

After completing its report, the SUAR Subcommittee, in accordance with the normal practice of the Federal Committee on Statistical Methodology, was dissolved. However, it was evident that many aspects of the subject deserved further, continuing attention. A successor subcommittee, the Administrative Records Subcommittee, was formed. The meetings of the Administrative Records Subcommittee make available to the agencies represented on it a forum for exchange of information on current developments in administrative record systems and their statistical implications. In addition, several ad hoc working groups have been formed to study specific topics and to prepare findings and recommendations. Topics taken up by the working groups include: industry coding, the definition of economic reporting units, a detailed review of selected administrative record systems containing data for persons, planning and conducting a matching study of economic reporting units from different agency systems, and the integration of administrative record and statistical data in household and business surveys. Figure 2 gives further particulars of the objectives, current status, and outputs of these working groups.

In view of the recognized importance of maintaining interagency contacts and undertaking collaborative studies on various aspects of the statistical uses of administrative records, the Administrative Records Subcommittee is now considered by its parent body, the Federal Committee on Statistical Methodology, to be a continuing subcommittee.

Starting at about the same time as the establishment of the SUAR Subcommittee, there has been a parallel and related undertaking to organize, for each annual meeting of the American Statistical Association, several invited and contributed paper sessions devoted to current developments and research on statistical uses of administrative records. In order to make them conveniently available, papers from each of these sessions, which started with the 1979 annual meeting, have been published in special "pre-print" volumes (U.S. Department of Health, Education and Welfare, 1979, 1980; U. S. Treasury Department, 1981, 1982, 1983). To put the whole subject in broad perspective and to cover different aspects of the topic in a structured format, the two-volume handbook described in the introduction to this paper was compiled and published (Internal Revenue Service, 1984a).

These kinds of technical cross-fertilization activities are valuable and should continue and perhaps intensify. The participation of Statistics Canada in this process has been important. While the U.S. statistical system has made important technical contributions to government statistical agencies throughout the world, the process need not be one-sided. In addition to learning about "catastrophes" associated with statistical uses of administrative records, as exemplified by Butz' (1983) report on the West German census, we

Figure 2. Working Groups of the Administrative Records Subcommittee

| Name | Objectives | Year Started | Report or Status |
|--|--|--------------|--|
| Industry Coding | To review major establishment industry coding systems and recommend steps to improve | 1981 | OMB Statistical Policy Working Paper 11. <u>A Review of Industry Coding Systems</u> , March 1984 |
| Establishment Reporting | To explore the possibilities for implementing true establishment reporting in major administrative record systems | 1981 | <u>Establishment Reporting in Major Administrative Record Systems</u> , unpublished report of Working group, December 1982 |
| Matching | To document selected Federal administrative files that may be used for data linkages for statistical purposes | 1982 | Drafting final report |
| Establishment Reporting Unit Match Study (ERUMS) | To undertake a pilot matching study to compare characteristics of records in BLS employer wage reporting system and SSA establishment reporting system | 1983 | Study in progress |
| Integration of Administrative and Survey Data | To explore the potential for greater use of administrative records in the design and conduct of demographic and economic surveys | 1984 | Getting underway |

need to seek out and learn from the successes, such as Statistics Canada's database and linkage system for epidemiological studies (Smith, 1983), Danmark Statistik's transition to a "full register-statistical system" (Thygesen, 1983), and France's COLIBRI II system used for on-line coding of occupation, industry and place of work in their 1982 population census (Lery and Stephany, 1984).

Monitoring Changes in Administrative Systems

The content, accessibility and quality characteristics of most administrative record systems change frequently. Major changes in content may occur as the result of legislation which places new data requirements on a system or eliminates existing ones. Accessibility may be affected by a switch to new computer or data-base management systems or by new policies covering disclosure of data. The quality of information can be affected by many factors: changes in report forms, new compliance procedures, new data edits, and so on.

It is easy to find examples of such changes and their effects on statistical data systems based on administrative records. An example of a major change was the shift from quarterly to annual wage reporting in the social security system in 1978. This change required substantial revisions in the content and structure of the Continuous Work History Sample (CWHS) data files (Internal Revenue Service, 1984a, Volume I, selection 3.01). In addition, delays in developing and implementing new procedures for posting annual wage

data to individual earnings records caused corresponding delays in the availability of annual CWHS files for several years following the changeover.

The plans for shifting to annual reporting were known to statistical users of social security earnings data well in advance, and they had time to develop plans for adapting to it. Frequently, however, significant changes become evident to statistical users only after they occur, sometimes leading to embarrassing consequences. Examples are well-known to experienced users: one is referred to in the next section of this paper.

Because statistical agencies have only limited control over changes in administrative data systems, those that depend primarily on direct collection of data are understandably reluctant to place more reliance on administrative records: they become less and less masters of their fates. Nevertheless, pressures to reduce costs and respondent burden mean that greater reliance on administrative records is an option that must be taken seriously.

Changes in administrative record systems represent both threat and opportunity for statistical users. Whichever is the case, it is essential for users to do their utmost to monitor changes: to learn about them before they occur and whenever possible to ensure that they are made in ways that minimize disruptions to statistical applications. To date, provisions for monitoring have generally been less than systematic, especially on the part of users not in the agency that has custody of the administrative records.

Over the past few years, the activities of the Administrative Records Subcommittee have helped

to make information on administrative record systems more widely available; however, there is still need for a much more systematic effort to obtain information on prospective and actual changes in these systems and to disseminate that information to users throughout the statistical system. A proposal for an Administrative Record Monitoring System (ARMS) was presented to the Administrative Records Subcommittee (Internal Revenue Service, 1984a, Volume I, selection 1.06). The basic ideas are straightforward; the need is for modest resources to operate such a system. The potential benefits--early identification of new opportunities for more effective use of administrative records and avoidance of costly failures to adapt to changes--are substantial. Furthermore, systematic monitoring is essential if statistical agencies are to have a more active role in determining the character and uses of administrative record systems.

Monitoring Trends in Interrelationships Among Administrative Record Systems

Effective statistical use of administrative records, whether an active or passive strategy is adopted, requires more than careful monitoring of changes in individual administrative records systems; it also requires attention to broad trends affecting several systems and the way they interrelate. We give one example of such a trend: the trend toward integration of Federal and State administrative record systems.

Because of the division of authority between Federal, State, and local governments in the United States, individuals, businesses, and non-profit organizations are subjected to various reporting requirements that overlap with respect to the specific items of information required by different levels of government. This condition leads inevitably to the question: can the reporting burden associated with particular functions be reduced by methods such as: (1) use of a single report form from which the data will be processed and shared by the various governmental agencies needing them, (2) use of a standardized form, copies of which can be submitted to all agencies needing the data, or (3) standardizing the reporting unit definitions and identifiers and the definitions of common data items used at all levels?

There has been some recent progress along these lines. No States have yet taken advantage of the "piggy-backing" procedure which allows the Treasury Department to enter into agreements with States to collect individual income taxes and pay over to the States the amounts collected. However, many States have structured their income tax laws to minimize the additional work required of taxpayers to file their State tax returns once they have completed their Federal returns.

Most non-profit organizations are required to file information returns with Federal, State and, in some cases, local tax agencies in order to give evidence of their continuing tax-exempt status. Recently the reporting burden on these organizations has been reduced as the result of a cooperative effort between the IRS and State tax agencies. In nearly all States, tax-exempt organizations filing the Federal information return, Form 990, are now permitted to file a copy of this

return to the State taxing authority to meet its filing requirements. In preparation for this arrangement, certain data items were added to the Form 990 so that it would meet the basic needs of all States. As a result, some additional information is available for inclusion in the statistical data bases for tax-exempt organizations that are developed from the Federal returns on Form 990.

Although the non-profit organization initiative required careful, detailed work by the parties to it, it was relatively straightforward. An important factor was that no collection of taxes was directly involved. Other proposals for integration, such as those that might surface as a result of SBA's recent suggestion to OMB that a Federal-State commission on tax coordination be established (see goal 6 discussion), involve much larger reporting systems that are used to collect taxes or to pay cash benefits and will require much more time, effort and balancing of diverse objectives in order to succeed. The potential benefits for the public, the program agencies and the statistical users will also be greater.

There will inevitably be further efforts to reduce reporting burden. It is important for statistical users of administrative records to follow these developments and, if possible, to participate in the design of better integrated reporting systems. The system changes can have both positive and negative effects on statistical uses; greater participation by statistical users at the planning stage can help to minimize the latter.

Designing Data Systems to Facilitate Linkages with Other Systems

The potential value of many statistical data systems can be enhanced by including identifiers that will permit linkages for statistical purposes with other data systems. For household surveys, the collection of social security numbers (SSNs) of sample persons is an obvious possibility. The value of survey-collected SSNs is affected by the completeness and accuracy with which they are reported. Based on lessons learned in the 1973 Exact Match Study (Internal Revenue Service, 1984a, Volume I, selection 4.04), special techniques were used to obtain SSNs for persons included in a 1979 panel survey conducted for the Income Survey Development Program. As a result, valid SSNs were obtained for more than 95 percent of sample persons aged 14 or older. (Kasprzyk, 1983).

Enhancement of a survey base file with administrative record data can be longitudinal as well as cross sectional. In the Social Security Administration's Retirement History Survey, earnings and benefit data were linked to the records for sample persons over a period of about ten years during which the same persons were being interviewed biennially. There are other possibilities for longitudinal tracking; for example, annual occupation entries on individual income tax returns might be used to track occupational changes for a labor force or health survey base population.

In the business sector, a common business directory would be more useful if it were planned as a vehicle for various kinds of data linkages, rather than simply as a frame for economic censuses and surveys. A recent development of

some interest is the decision by the BLS to require cooperating State agencies to provide an Employer Identification Number (EIN) for each reporting unit included in the file transmitted annually to the BLS for updating of its Unemployment Insurance Address File. Leaving legal and policy issues aside for the moment, it can be noted that this step would make it technically feasible for BLS to provide updated industry codes for single-unit employers to the SSA or the Census Bureau. For the latter agency, this would be especially useful for units not covered directly in economic censuses and surveys.

A special design consideration arises if both of the systems to be linked are based on samples. Direct linkages are, of course, only possible to the extent that the samples in the two systems overlap. Starting with the tax year 1979, linkages between the Statistics of Income (SOI) sample of individual taxpayers and the Continuous Work History Sample (CWSH) were made possible by incorporating in the SOI sample a subset of returns for which the SSN digital selection pattern coincided with that used to select a part of the CWSH. This was done to provide a vehicle for developmental work on the LASS system (see earlier discussion of Goal 2).

Fostering a Two-way Relationship

Any relationship between two parties works better if it has some give and take on both sides. How can this principle be applied to the relationship between custodians of administrative record systems and statistical users of these systems?

It is essential that the flow of information about identifiable individuals (persons, businesses, etc.) be in one direction only--from administrative systems to statistical agencies. If this rule were ever violated, the consequences for the U.S. statistical system could be catastrophic--we may refer again to the lessons of the West German census (Butz, 1983).

However, there are some things that statistical agencies can do to help custodians of administrative record systems. The latter, like the statistical agencies, care about the completeness and accuracy of information in their systems. In some respects, accuracy of administrative data is even more important because the data are used to determine eligibility, benefits, liabilities and penalties for specific individuals. It should be relatively easy for the statistical agencies that process administrative records and, frequently, link them with census or survey data to produce aggregate data bearing on the coverage and accuracy of the administrative records. Sharing aggregate data of this kind would not violate the basic rule concerning the one-way flow of individually identifiable data, provided adequate steps were taken to prevent statistical disclosure.

Another possibility would be for the statistical agencies to make microdata files containing linked administrative and statistical data, with identifiers removed, available to statistical units of the agencies where the administrative records originated. As discussed earlier in connection with Goal 4, the "reidentification problem" makes statistical agencies reluctant to do this. However, it would be a forthcoming gesture for

the statistical agencies to participate actively in efforts to find legal or technical solutions to this problem.

To the extent that linked microdata files cannot be released, the statistical agencies might make a special effort to produce, from these files, tabulations that would be of value to the administrative record custodians for program and policy analysis.

WHERE DO WE GO FROM HERE?

Our review of recent developments related to the 6 goals presents a mixed picture of reasonable progress in some areas and minimal or no progress in others. Possibilities for increased use of administrative records in the decennial census and current estimates programs (Goal 1) are being systematically explored by the Census Bureau with the help and encouragement of other agencies and advisory groups. Efforts to develop a Linked Administrative Statistical Sample (Goal 2) are stalled until legal and technical access issues can be resolved. The quality of recent data in the CWSH files (which would be the core element of the LASS system) is largely unknown and needs to be determined. There has been some progress in making other elements of the system--occupation from tax returns and mortality from the National Death Index--more accessible.

New developments in the use of administrative records to monitor environmental health effects (Goal 3) include coding of occupation on death certificates and matching CPS files to the National Death Index. However, overall progress in this area is discouragingly slow and we are a long way from achieving the effective data mechanisms which have been operating in Canada and other countries for some time now.

Progress in the integration of survey and administrative data (Goal 4) is being made in various agencies and programs--the plan to link SIPP and economic census data is a particularly interesting development. The new interagency working group on integration of administrative and survey data may prove to be a catalyst for further innovative applications of administrative records in surveys.

The idea of a business directory available to Federal and State agencies for statistical purposes (Goal 5) seems not much closer to fulfillment than when first conceived about 50 years ago. Lack of a directory has hampered efforts to standardize the definitions of economic reporting units and their classifiers (Goal 6); however, the work of the ERUMS Working Group should make a useful contribution.

Why has there been so little progress in developing significant new uses of administrative records? Some might argue that the present mix of administrative and statistical sources is about right for the U. S. statistical system at this time. We do not subscribe to this view, and we attribute lack of progress to the following factors:

- o Reluctance of statistical managers to place reliance on data sources over which they do not have direct control. This attitude is understandable. In social terms, one might

liken it to the desire of some people to return to subsistence farming as a way of life, rather than continuing to be part of a complex society in which nearly every aspect of existence requires reliance on the good faith and performance of others. There is no question that reliance on administrative records has brought severe headaches to some statistical program managers. The difficulties associated with the use of IRS records in the 1982 economic censuses are a good example of the kinds of problems that occur. The IRS (1984b) has prepared a detailed account of what happened and has offered suggestions for avoiding recurrences. Our own view is that circumstances make increased reliance on administrative records inevitable, and that what is necessary is to develop strategies to make it work.

- o Concerns about the coverage, relevance and accuracy of administrative record data. The coverage, content and accuracy of potentially usable administrative records systems seldom provide an exact fit to the data requirements of statistical users of the records. There are ways to compensate for lack of fit: use of multiple frames, edits, imputation, etc. In the longer run, the ideal solution is to influence the design of administrative systems so that they are better suited to meet both program and statistical needs. Careful evaluation of the suitability of administrative records for proposed statistical uses is always essential, but lack of a precise fit should not lead to automatic rejection if the potential benefits are substantial.
- o Difficulties in gaining access to administrative records for statistical uses. Legislation, regulations and policies adopted since the mid-1970s in response to increasing public concerns about privacy and confidentiality have unquestionably made it more difficult to gain access to records in administrative systems. Even where access is possible, making the necessary arrangements is often frustratingly difficult and slow.
- o Limited resources for research and development. Innovative statistical applications of administrative records require the allocation of scarce resources for evaluation of suitability of the records and for the solution of systems design, data access and data linkage problems. Competing needs for these resources tend to get higher priorities.
- o The 1984 syndrome. The coming of 1984 triggered many articles and discussions about whether the Orwellian society, with its omniscient "big brother", was already at hand or about to arrive. People were more than normally sensitized to the negative aspects associated with "data banks" containing large amounts of information about individuals. More than one recent proposal to link data from different sources for statistical purposes has been met with the response that, "This is not the year to do it." Without in any way denying that public attitudes and perceptions are important, we hope that starting in 1985 it will be possible to evaluate such proposals objectively on their merits.
- o The West German Census syndrome. This is

closely allied to the 1984 syndrome. Butz (1983) has argued very persuasively that the United States statistical agencies should pay close attention to the circumstances that caused postponement of the West German census scheduled for 1983 and to similar, although somewhat less intense opposition to censuses encountered recently in other Western European countries. We agree that this phenomenon cannot safely be ignored. However, it should not be used as a justification for suspending all consideration of new initiatives to reach the goals presented in this paper.

- o Lack of overall strategic planning. Many individual statistical agencies are doing an excellent job of long-range strategic planning. However, they do it in the context of their own functions, programs and interests, and not as part of an overall statistical system designed to meet the information needs of government and the public in the most efficient way possible. As a result, we believe there is a bias that favors direct data collection programs under the control of a single agency in preference to those making greater use of administrative records, but requiring sharing of data and close cooperation by two or more agencies. This situation explains the relative scarcity of resources for research and development work on new uses of administrative records and it may also account for the failure of efforts to obtain legislation needed to implement a shared business directory and other projects that require exchanges of identifiable data.

Faced with all of these obstacles, it is tempting to opt for the status quo, i.e., to continue with the present mix of administrative records and direct data collection, on the grounds that it is working reasonably well. We believe this would be a poor strategy. Continuing pressure on both dollar and burden budgets is inevitable: better use of administrative records is essential if we are to adapt to these pressures without seriously weakening our ability to meet the nation's needs for statistical information.

Two strategies are now open to us. One strategy is to postpone further work on new statistical applications of administrative records until pressures on resources or new data needs make it essential, i.e., to react passively to changes in the external environment. The other strategy is to actively prepare for change and to seek some degree of control over the broad structure of the statistical system and its external operating environment. Some important elements of an activist strategy are:

- o Elevating the interagency discussion of goals for statistical uses of administrative records from the technical to the policy level.
- o Agreeing on priority goals and developing detailed plans and schedules for reaching them.
- o Doing a better job of monitoring actual and prospective changes in administrative record systems and, beyond this, seeking an active role in influencing the design of major systems to facilitate the extraction of relevant, timely and accurate statistics from them.

The activist strategy, which is clearly the one preferred by the authors, will be feasible only if there is general agreement among the major statistical agencies on the need for the Federal statistical system to become a reality as opposed to the present loosely-connected group of agencies which constitutes a system in name only.

ACKNOWLEDGMENTS

The authors would like to thank Maria Gonzalez, Chair, Federal Committee on Statistical Methodology and Daniel Garnick, who chaired the Subcommittee on Statistical Uses of Administrative Records, for their support of and participation in interagency technical activities described in this paper. Thanks are also due to Donny Rothwell for significant contributions and Anne Sprague for typing assistance through several drafts.

NOTES AND REFERENCES

- Alvey, W. and Scheuren, F. (1982) Background for an Administrative Records Census. Pp. 47-65 in Statistics of Income and Related Administrative Record Research. Washington, D.C.: U.S. Department of the Treasury, Internal Revenue Service.
- Bureau of the Census (1983) Feasibility of an Administrative Records Census in 1990. Unpublished report of the Subcommittee on an Administrative Records Census, Committee on the Use of Administrative Records in the 1990 Census. Washington, D.C.: U.S. Department of Commerce.
- Butz, W. (1983) Public Protest and the Postponement of the West German Census. Unpublished report, Bureau of the Census. Washington, D.C.: U.S. Department of Commerce.
- Childers, D. and Hogan, H. (1984) Matching IRS Records to Census Records: Some Problems and Results. Paper presented at the Annual Meetings of the American Statistical Association, Philadelphia.
- Commission on Federal Paperwork (1977) Taxation: A Report of the Commission of Federal Paperwork. Washington, D.C.
- Committee to Coordinate Environmental and Related Programs (1983) On Occupational Cancer Estimation: Report of the Occupational Cancer Risk Subcommittee. Washington, D.C.: U.S. Department of Health and Human Services, November 4.
- Crouse, W., Schuster, L., Rosenberg, H., Kametani, D. and Sestito, J. (1983) Using the Census Bureau's Occupation and Industry Coding System for Coding Death Certificates. Pp. 53-58 in Statistics of Income and Related Administrative Records Research: 1983. Washington, D.C.: U.S. Department of the Treasury, Internal Revenue Service.
- Haber, S., Ryscavage, P., Sater, D., and Valdiserà, V. (1984) Matching Economic Data to the Survey of Income and Program Participation: A Pilot Study. Presented at the Annual Meetings of the American Statistical Association, Philadelphia.
- Internal Revenue Service
(1981) Statistics of Income and Related Administrative Record Research. Washington, D.C.: U.S. Department of the Treasury.
- (1982) Statistics of Income and Related Administrative Record Research: 1982. Washington, D.C.: U.S. Department of the Treasury.
- (1983) Statistics of Income and Related Administrative Record Research: 1983. Washington, D.C.: U.S. Department of the Treasury.
- (1984a) Statistical Uses of Administrative Records: Recent Research and Present Prospects (2 vols.). Washington, D.C.: U.S. Department of the Treasury.
- The selections from the report that were referred to in this paper are:
- Volume I:
- 1.06 Jabine, T. Proposal for an Administrative Record Monitoring System.
- 3.01 Buckler, W. and Smith, C. The Continuous Work History Sample (CWHS): Description and Contents.
- 3.02 Kilss, B., Scheuren, F. and Buckler, W. Goals and Plans for a Linked Administrative Statistical Sample.
- 4.04 Kilss, B. and Scheuren, F. The 1973 CPS-IRS-SSA Exact Match Study.
- Volume II:
- 6.09 Spruill, N. Measures of Confidentiality.
- (1984b) IRS Support for the 1982 Economic Census: Some Lessons for the Future. Unpublished report. Washington, D.C.: U.S. Department of the Treasury.
- Kasprzyk, D. (1983) Social Security Number Reporting, the Use of Administrative Records, and the Multiple Frame Design in the Income Survey Development Program. Pp. 123-141 in Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program (ISDP). Washington, D.C.: Social Science Research Council.
- Lery, A. and Stephany, A. (1984) COLIBRI II: Saisie-chiffrement en Ligne des Bulletins du Recensement de la Population de 1982. Courier des Statistiques, No. 30: 21-28. Paris: INSEE.
- National Research Council (1984) Planning the 1990 Census: Priorities for Research and Testing. Interim Report of the Panel on Decennial Census Methodology, Committee on National Statistics. Washington, D.C.: National Academy Press.
- Office of Federal Statistical Policy and Standards (1980) Report on Statistical Uses of Administrative Records, Statistical Policy Working Paper 6. Washington, D.C.: U.S. Department of Commerce.
- Office of Management and Budget
(1982) Establishment Reporting in Major Administrative Record Systems. Unpublished report of the Establishment Reporting Work Group, Administrative Records Subcommittee, Federal Committee on Statistical Methodology. Washington, D.C.: Office of Management and Budget.
- (1984) A Review of Industry Coding Systems, Statistical Policy Working Paper 11. Washing-

- ton, D.C.: Office of Management and Budget.
- Oh, H. and Scheuren, F. (1984) Statistical Disclosure Avoidance. Notes for presentation to Washington Statistical Society, May 22.
- Rich, S. (1983) "15 Forms to Hire a Domestic?" The Washington Post, November 13, 1983.
- Rogot, E., Feinleib, M., Ockay, K., Schwartz, S., Bilgrad, R. and Patterson, J. (1983) On the Feasibility of Linking Census Samples to the National Death Index for Epidemiologic Studies. American Journal of Public Health, November.
- Sailer, P., Crabbe, P., and Kilss, B. (1984) Taxpayer Data Classified by Occupation: Results from the 1979 Statistics of Income Sample. Presented at the Annual Meetings of the American Statistical Association, Philadelphia.
- Smith, Martha E. (1982) Development of a National Record Linkage Program in Canada. Pp. 41-46 in Statistics of Income and Related Administrative Record Research: 1982. Washington, D.C.: U.S. Department of the Treasury, Internal Revenue Service.
- Social Security Administration (1979) Statistical Uses of Administrative Records with Emphasis on Mortality and Disability Research. Washington, D.C.: U.S. Department of Health, Education and Welfare. (1980) Economic and Demographic Statistics. Washington, D.C.: U.S. Department of Health, Education and Welfare.
- Thygesen, L. (1983) Methodological Problems Connected with a Socio-Demographic Statistical System Based on Administrative Records. Pp. 227-242 in Bulletin of the International Statistical Institute, Vol. 50, Book 1, Madrid.