Machine-Readable Data Production by the Federal Government

Access to and Utility for Social Research

MICHAEL W. TRAUGOTT
JEROME M. CLUBB
University of Michigan

Governments have long been a major source of information for the study of human affairs. As historians know, the records of rulers, their courts, and their officials are among the prime sources for the study of some of the earliest periods of human history. The growth of national bureaucracies, as Stein Rokkan notes elsewhere in this issue, brought with it not only standardization and centralization of administrative record-keeping but also a vast and continuing expansion of the nature and volume of records kept. The tax records, the vital statistics, the records of governmental expenditures and economic transactions, and the like collected and maintained by early bureaucracies have come to be basic sources of quantitative information for later investigators. To these administrative records can be added the results of periodic and more or less systematic assessments in the form of national censuses and other enumerations of the extent and characteristics of national populations, of the nature of economic activities, and of national resources.

Expansion of governmental activities provided a further stimulus to information-gathering and record-keeping. Assumption of regulatory powers meant expansion of the volume of information collected bearing upon business and commercial

AMERICAN BEHAVIORAL SCIENTIST, Vol. 19 No. 4, March/April 1976

activities that were in many nations formerly considered matters of purely private concern; adoption of more complex taxation systems meant systematic and continuing collection of more detailed and more extensive information on the income, possessions, occupations, and expenditures of individuals and organizations; acceptance by national governments of a measure of responsibility for the well-being of citizens meant collection of extensive and detailed social statistics bearing, again, on matters that were formerly considered as of exclusively private concerns; and multiplication and intensification of the contacts between nations has made national governments not only repositories of data relevant to their own populations and national activities but to the populations and activities of other nations as well. Nor has the information gathering of national governments been limited to such routine activities. Concern for the impact and effectiveness of particular national programs and policies and for assessment of problems has frequently led to the commissioning of special data collection and research efforts which also constitute for the social scientist a basic source of systematic information.

The so-called "computer revolution" has, of course, proven to be a further and massive stimulus to governmental information-gathering and record-keeping. Computer technology has greatly increased the potential for information storage, management, and utilization. Indeed, information utilization of the magnitude now characteristic of national governments would be unimaginable without the computer. At the same time, and as is suggested subsequently in this essay, the potentials of computer technology have been by no means fully realized. As a consequence the computer revolution has also worked to complicate greatly information management and utilization within governments themselves and has worked as well to erect significant barriers to effective use of governmentally produced information by social scientists.

The federal government of the United States is by no means an exception to these trends. If anything, indeed, these trends are particularly marked in the case of the United States. Perhaps no better testimony to the magnitude and character of

the information-gathering activities of the federal government can be found than the often expressed, and often exaggerated, concerns for the hazards to the privacy and well-being of individuals, groups, and organizations which these activities present. From another perspective, however, the federal government can be seen as constituting, potentially, a massive, rich, and incredibly diverse archive of machine-readable data of immense value for the purpose of social inquiry. As will be suggested, these potentialities have not been fully realized nor are the prospects entirely heartening that they will be realized in the foreseeable future.

It is obviously impossible in a brief essay to even catalogue, much less describe, the rich machine-readable social science data resources produced by, and ostensibly available from, the federal government. It is only possible in such an essay to consider some of the major categories of federally produced data resources, to mention some of the means of access to those resources, to call attention to some of the difficulties confronted in their use, and to note in passing several developments that look toward more effective access by social scientists to these resources.

I. CATEGORIES OF FEDERAL DATA

The decennial censuses constitute the most extensive data collection efforts of the federal government. The decennial enumeration of the population for the purpose of apportionment of the House of Representatives was mandated in Section 2, Article 1 of the Constitution. A review of the development of the forms used to record information from the first census in 1790 to those used in the last census in 1970 illustrates the shift in emphasis, content, and technique that has occurred as a function of the expansion of the role and concerns of the federal government, the growth in the complexity of society and the availability of faster tabulation procedures (U.S. Bureau of the Census, 1973c). And in the intervening years, what is now known as the Bureau of the Census has become the single

largest collector and disseminator of national statistics, with a continuous data program that includes monthly surveys and special quinquennial censuses in addition to the censuses of population and housing.

The federal government also commissions independent and special data collection efforts, most of which are directed toward the development of information for planning and evaluation purposes. A prime example of this type of activity is the "Coleman Report" (Coleman, 1966), a project supported by the Department of Health, Education and Welfare. In addition, numerous data-gathering projects are conducted in support of national commissions. The breadth of these efforts ranges from historical data on domestic civil violence to contemporary public attitudes toward gambling. Although originally collected to serve specific governmental purposes, such data frequently have the same value for secondary and extended analysis as do data orginally collected by academic and private researchers.

It is in the area of administrative record-keeping, however, that the computer—along with growth in the role of government and the increasing complexity of society—has had its greatest effect. In the same manner that the analytic power of the computer has enabled researchers to perform statistical calculations with speed and accuracy previously unavailable, it has become possible to store and manage quantities of administrative records which would represent an impossible human clerical task. These "process data," as they are sometimes called, have potential value for a wide variety of research purposes. They provide a means, for example, to examine the functioning of government itself—to investigate the flow of resources and personnel between agencies, the relative priorities of the federal government and its various agencies, the administrative procedures employed, the movement of resources between government and society, and the interventions of the federal government in societal and economic matters. In the aggregate these data resources frequently can be used to assess the depth and breadth of social services and to measure

the penetration of social programs in terms of their actual and expected effects. In some instances, disaggregated data are available that allow an intensive look at some social phenomenon which may or may not be directly related to the purpose for which the administrative records were originally obtained. An excellent example of such data, discussed in greater detail below, are the Continuous Work History data files available from the Social Security Administration. Constructed from administrative records supplied by employees and employers, the merger of such data on an individual basis provides a unique source of information on occupational mobility.

Through increased availability of data of these types, the federal government has become a leading source of social science data. Not all of the information which the federal government collects is publicly available. And special precautions must often be taken to protect the anonymity of the individuals from whom the information was obtained. Nevertheless, the federal government is the source of the most extensive and varied data collections presently available to social scientists.

II. THE BUREAU OF THE CENSUS

Without question, the largest and most centralized source of governmentally produced and publicly available machine-readable information is the Bureau of the Census of the Department of Commerce. Thus it is useful to consider briefly some of the major data products of the Bureau, both to call attention to the activities of that agency and to illustrate the magnitude of the information resources of the federal government. Here again, however, it is impossible to do more than touch upon some of the Bureau's activities and products.

The scope of the Bureau's activities can be described along many dimensions. The official figure quoted for the cost of the 1970 decennial censuses of population and housing was $220 million in total, of which $30 million was for data processing

expenses; the annual budget of the Bureau, including fees collected for work performed for other governmental agencies as well as its own appropriations, now exceeds $130 million (U.S. Bureau of the Census, 1975). The Bureau conducts monthly interviews in almost 50,000 households in the course of its Current Population Survey program. Other recurring Census programs include the quinquennial Census of Governments, the Census of Manufacturers, and the program of the Economic Censuses. Because of this high level of data collection activity and the planning and methodological research that is required for such efforts, the Bureau of the Census also serves as a primary source of advanced methodological information in the areas of sampling and interviewing procedures.[1]

Although selected data files had been made available in machine-readable form prior to the 1970 Census of Population and Housing, these were primarily internal working files which directly supported various Bureau publications such as the series of County and City Data Books. The planning for the 1970 Census, however, specifically included provisions for the dissemination of machine-readable data products which contained much more detailed information than would ultimately be contained in published reports. The justification for such an effort was acknowledgment of the fact that most of the data collected through the Census would be processed and tabulated only once and presented in a relatively standard series of published reports. It was recognized that an enormous potential existed for extended analysis and use by researchers, business firms, local governments and the like if the basic data could be made generally available in a usable machine-readable form. As a result of this effort, over 2,000 reels of magnetic tape containing data from the 1970 Census were made publicly available. More importantly, perhaps, a precedent was established for increasing the flow of data to interested researchers and analysts; and the official catalog of the Bureau is now divided into two sections, one for publications and the other for data files and special tabulations (U.S. Bureau of the Census, 1946 to the present).

The vast bulk of the data released in conjunction with the 1970 Census was aggregated tabulations for various geographical areas. There were six so-called "counts" of information released in this form, involving substantive content ranging from simple age, race, and sex information, based upon a 100% enumeration of the national population, to more complex data on income, ethnicity, and occupation obtained from a 5% sample of the population. The smallest geographical units for which data were tabulated were city blocks, and the largest were states.

In addition to data aggregated by geographical area, a large volume of individual-level data from the 1970 Censuses of Population and Housing was also made available in the form of Public Use Samples (U.S. Bureau of the Census, 1972). The sampling fractions of individual records used to construct these data files ranged from one-in-one hundred at the local level to one-in-one thousand or one-in-ten thousand at the national level. While excluding information which might reveal the identity of specific individuals, the various Public Use Samples contain responses from both the basic and sample questions used in the multiple enumeration forms employed in the 1970 Censuses. In effect, the Public Use Sample files constitute giant surveys of the populations of groups of counties, states, or the nation. The responses can be tabulated or analyzed in any form the researcher desires, with the only practical restriction being the cost of manipulating such large quantities of data (National Data Use and Access Laboratories, 1973).

One of the supplementary services which the Bureau of the Census provides are "special tabulations" of its own data bases. Sometimes this work involves reaggregation of basic records to alternate geographical areas; in other cases it involves recategorization of the data values. Special tabulations are performed by the Bureau because certain data needs can only be met by returning to the original individual-level data files, to which only Bureau personnel have access. In the case of the 1970 Census, the Bureau has performed several special tabulations and made them publicly available. One of these is the so-called Fifth Count File C tabulations, originally performed for the

R. H. Donnelly Co. Using the same tabular format for variable categorization as the Fifth Count data, basic data were reaggregated to the tract-level. The result was extensive information, recorded on over 50 reels of magnetic tape, not previously available for that level of aggregation.

The Bureau of the Census also performs similar services for other governmental agencies. The Bureau has been the official source of population and income totals and estimates for the Office of Revenue Sharing in the Department of the Treasury. The initial planning and disbursement activity of the Office of Revenue Sharing was based upon information from the 1970 Census of Population and minor adjustments made to those figures to account for boundary changes, annexations, and consolidations. Recently, the Population Division of the Bureau has issued new estimates of population and annual income for over 38,000 local governmental units as of October, 1974. While the availability of this information fills a valuable administrative need for the Office of Revenue Sharing as well as a financial one for local governments, the incidental release of these data to the research community provides an important analytic and methodological resource (U.S. Bureau of the Census, 1974).

The quinquennial Censuses of Governments and the Annual Surveys of Governments conducted by the Bureau are further sources of machine-readable data of value to social scientists. Each of these activities is conducted in two parts, with separate information reported for governmental employment and finances. The 1972 Census of Governments was the first to be released entirely in machine-readable form. In combination, the Employment and Finance Files amount to over 9,000,000 card-image equivalents of data for approximately 79,000 governments. The annual Survey of Governments provides information for the fifty states and for a sample of approximately 16,000 local governmental units. The sample of local governmental units is stratified by type of government and magnitude of expenditures, and cities of 25,000 population are included with certainty. The survey thus provides relatively complete information for larger municipalities. The survey is

conducted in October of each year, and the data from 1971 to the present, again including both Finance and Employment files, are available in machine-readable form.

In addition to these files of aggregate data released in conjunction with the various censuses conducted by the Bureau, there are also machine-readable data products which support special publications. Foremost among these are the data files for the County and City Data Book series (U.S. Bureau of the Census, 1953, 1957, 1962, 1967a, 1973b). The data in these volumes, along with the machine-readable data files on which they are based, are constructed from selected elements of the Censuses of Population, Housing, Manufacturing, and Governments and include as well minor components taken from other federal data sources. These files provide data for states, counties, cities, and Standard Metropolitan Statistical Areas (SMSAs). Machine-readable files and published reports are also available which provide data aggregated to the level of Congressional Districts (U.S. Bureau of the Census 1961, 1963, 1973a), a straightforward reflection of the Bureau's mandate to enumerate the population for purposes of congressional redistricting.

Increasingly data from the Current Population Surveys are being made available to interested users. At present only two of the monthly surveys, involving approximately 50,000 households each, are available routinely. The basic data files from the Annual Demographic File: March Supplement are available dating back to 1968. This file contains information on approximately 200,000 persons for each survey, including detailed data on age, race, sex, ethnic origin, income, and occupation. The second survey is the biennial Voter Participation File: November Supplement, which contains information on voter registration and participation, including reasons for nonparticipation, as well as standard demographic information, for the population eligible by age to vote. These data files contain information for approximately 95,000 such individuals in each survey. Other monthly data files from the Current Population Survey dating back to 1959 are available on a special

tabulation basis. It is anticipated, however, that more of these surveys will eventually be made available on a routine basis.

Supplementing its independent data collection activities, the Bureau of the Census is also the largest collector of data for other government agencies. One of the projects of this nature in which the Bureau is presently involved is the National Crime Survey Panel conducted for the Law Enforcement Assistance Administration (LEAA) of the Department of Justice. Interviewing is carried out on a monthly basis in approximately 5,000 households and 1,250 businesses which constitute a representative national sample (U.S. Department of Justice, 1974b). In addition, local area data are provided by supplemental sample surveys in many of the nation's largest cities; this procedure will result in triennial data for each of thirty cities and annual data for the five largest cities (New York, Chicago, Los Angeles, Philadelphia, and Detroit; U.S. Department of Justice, 1974a). Including the pilot surveys and design as well as the panel activity which began in January of 1973, this project represents a very sizable data collection effort. Although few activities are as large as this one, the Bureau is engaged in the conduct of a wide range of such data collection efforts, including the National Health Survey for the National Center for Health Statistics, surveys on disabled veterans for the Veterans Administration, and surveys on recreation, fishing, and hunting for the Bureau of Sport Fisheries and Wildlife of the Department of the Interior. Many of these activities also yield information of potential importance to social scientists.

III. OTHER FEDERAL AGENCIES

A second source of data from the federal establishment is the administrative records that result from the routine conduct of agency business (National Technical Information Service, 1974). The advent of electronic computers and their use to manage records has unquestionably saved many agencies from being buried under the weight of their own paperwork. In the process, however, the use of computers has also had the

important side benefit of generating data files, usually containing only a small sample of such records, which are an invaluable source of information for social scientists. The primary sources of such information include the Social Security Administration, the Internal Revenue Service and the various national centers for social and demographic statistics.

The Internal Revenue Service has made data available relating to the distribution of individual income based upon tax returns aggregated to the level of five-digit ZIP code areas in the United States. Although a reasonable amount of detail is provided from the categorized returns, no identification of individuals is provided. The Office of Revenue Sharing provides machine-readable financial data from the Planned and Actual Use Reports submitted by the local governments which are the beneficiaries of the program. In conjunction with information from the Census of Governments Finance File, these data constitute an important resource for the analysis of public policy-making at the local level.

Some of the very largest available data files are produced by the Social Security Administration from its own administrative records and from studies it has commissioned of the records of state agencies. The annual Continuous Work History Sample, available back to 1957, contains data for 1% of all of the social security numbers for which wage and salary information was reported in the year (National Technical Information Service, 1975: 50-55). Each annual file, stored on ten reels of magnetic tape, contains information on both employees and their employers, derived from Form SS-5, Employees Application for Social Security Number and Form SS-4, Employers Application (for ID No.). These data provide a unique opportunity to analyze the economic and occupational structure of a cross-section of the entire American labor force. In a more specialized vein, the Social Security Administration also makes available analogous samples of records of annual benefit payments, Medicare and Medicaid payments, and disability applications and awards.

Particularly where administrative records such as those mentioned above are concerned, the preservation of the

anonymity of individuals, business firms, and other organizations is a primary concern. It goes almost without saying that names and other identification information are removed from the publicly available data files. In addition, significant elements of the individuals' geographic location are also removed, and often such locational information is limited to only state or county. The Bureau of the Census has also adopted strict rules and practices to avoid disclosure of confidential information and to prevent invasion of individual privacy. In preparing the Public Use Samples of individual records, for example, the Bureau has adopted the practice of identifying no geographical locales of less than 250,000 population. In both its published report series and geographical machine-readable data files, the Bureau also suppresses all cross-tabulation entries for which cell frequencies are five or fewer persons or firms in order to safeguard the privacy of individual respondents.

A further generic category of machine-readable data available from federal sources is composed of information collected through the research activities of various national committees and commissions. In the main such data files reflect special-purpose, one-time data-gathering efforts usually addressed to well-defined issues. In topical coverage these files tend to be quite varied and range, for example, from a survey of public attitudes on drug use and abuse for the Commission on Marijuana and Drug Abuse to a study entitled "The American Public Looks at Violence" conducted for the National Commission on the Causes and Prevention of Violence (U.S. National Archives and Records Service, 1975). Frequently data from these sources, as is also often the case where other categories of federal data are concerned, have been only partially exploited by the agencies responsible for their creation. In some instances that exploitation amounts to little more than generation of a few summary statistics or frequency distributions for reporting purposes. Thus the data from these studies often constitute a fertile field for secondary and extended analysis. Unfortunately, preservation of these data has usually been dependent upon the interest and concern of the various commission and committee

staffs, and a good deal of the data collected for these purposes have been lost.

IV. ACQUISITION OF FEDERAL DATA

Most of the federal agencies that produce data resources of interest to social scientists also include sections or divisions devoted to dissemination of these resources. In general, these services vary in size and effectiveness. The largest data dissemination service in the federal establishment is the Data Users Service Division of the Bureau of the Census.[2] Two sections within the division are of particular importance to social scientists interested in the use of Bureau data. The Data Access and Use Laboratory is the primary source of information bearing upon Bureau data and publications. The Users' Service Staff is responsible for the actual dissemination of Bureau data, and procedures for ordering data are routinized and prices standardized.

Several Bureau publications are available and provide current information on the availability of data and services as well as examples of specific applications of data. The quarterly *Bureau of the Census Catalog* (1946–) was mentioned elsewhere. The catalog provides basic descriptions of available data, size of files, technical format, and ordering information. A second periodical, *Data User News* is a brief monthly newsletter containing information on new Bureau products, services, and programs which replaces an earlier publication entitled *Small-Area Data Notes.* The series *Data Access Descriptions* (1967–) appears on a variable schedule (four to six issues per year) and serves as an introduction to means of acquiring Census Bureau data.[3] The reports of the Census Use Study are special publications which appear episodically and relate to specific applications of census data in local, state, and federal agencies (U.S. Bureau of the Census, 1970). Although the work of the Census Use Study has been successively centered in New Haven, Los Angeles, and Indianapolis, its activities are purposefully designed to have

general utility for public policy makers. The reports which are issued by the project illustrate the use of census data in this context.

A second major supplier of federal data is the National Technical Information Service (NTIS), a central source for the public sale of government-sponsored research, development and engineering reports in addition to federally generated machine-readable data files. In fact, the distribution of machine-readable data is really a minor function of the NTIS operation. The agency is obligated by Title 15 of the U.S. Code to recover its costs from sales to users, and its self-sustaining nature is reflected in the price ($60.00) of the current NTIS catalog (National Technical Information Service, 1974). Data from the Bureau of the Census constitute the bulk of the listings in the catalog, although more than 500 data files and data bases produced by 60 federal agencies are listed. One relection of the unusual status of NTIS is a quoted price for supplying Bureau of the Census data that in some cases exceeds the price at which the same data could be obtained from the Bureau itself.[4]

The current NTIS directory also lists materials other than quantitative data. The directory provides a listing of computer software available from federal agencies, and the listed software includes computerized models of economic growth, simulations of natural and man-made processes, management systems for administrative records, as well as statistical routines. The charge for some of the listed software is greater than the simple cost of reproduction, but still considerably less than original development costs or those that would be incurred in producing comparable software from scratch. The NTIS catalog is currently the only centralized source of information on computer software available from governmental agencies.

Computer-readable files of federal data are also disseminated by the Machine-Readable Archives Division of the National Archives and Records Service.[5] The division is charged with responsibility for acquisition and preservation of those machine-readable federal data files that are deemed appropriate for long-term retention by the National Archives and for making

these data available to interested scholars, to the general public, and other users. The operation attained division status within the National Archives in August, 1974 and has only recently published its first catalog of data holdings (U.S. National Archives and Records Service, 1975). The catalog lists slightly fewer that 100 datasets recorded on more than 1200 reels of magnetic tape. The bulk of the present holdings of the division are from two sources—commissions established by both the legislative and executive branches and administrative records from federal agencies. The data sources include the National Commission on Population Growth and the American Future, the President's Commission on Campus Unrest, and the National Commission on the Causes and Prevention of Violence. Although it is a relatively new operation, standards for supplying data have been established. The activities of the division of the National Archives have been hampered in the past by the tendency in some agencies to erase or write over data files in order to reuse magnetic tapes. It can be expected that the holdings of the division will increase as the preservation ethic pervades those areas of federal agencies in which machine-readable data files are used. Thus the division is likely to become a major source of machine-readable federal data.

The National Technical Information Service and the Machine-Readable Archives Divison of the National Archives serve in effect as intermediaries between data-producing agencies of the federal government and data users outside the federal establishment. In addition to these governmental operations, a number of private organizations—of both profit-making and not-for-profit form—provide similar services. Three not-for-profit organizations that are oriented toward national constituencies can be mentioned here. The first of these, the National Data Use and Access Laboratory (DUALabs), was originally formed to facilitate access to and utilization of the data files from the 1970 census.[6] Most recently DUALabs has expanded its activities to include the processing of other federal data files and has entered into contractual relationships with various federal agencies to prepare their machine-readable data

files for dissemination. The Inter-University Consortium for Political and Social Research (ICPSR), a membership organization with more than 220 academic affiliates, also disseminates a number of the federal data files mentioned above.[7] In general, the ICPSR has selected federal data files that have attracted interdisciplinary interest within the academic community and has devoted attention to improvement of the technical format and documentation of these files in order to facilitate their use by secondary analysts. Thus ICPSR holdings of federal data are much more limited than those of DUALabs. A third center of such activity is the Oak Ridge National Laboratory where the Urban Research Section has expanded its activities to include dissemination of information from a large socioeconomic data base much of which is drawn from federal sources.[8] Two of the largest commercial operations which disseminate federal data—the National Planning Data Corporation and WESTAT Research, Inc.—can also be mentioned.[9] These are but a few of the organizations that perform an intermediary function where dissemination of federal data is concerned. The Bureau of the Census lists almost 200 organizations that perform such services on a local or regional basis (U.S. Bureau of the Census, 1973c).

For those unfamiliar with the problems confronted in the acquisition and use of federal data, it may seem incongruous that intermediary organizations such as these even exist, particularly in view of what may seem to be a large federal establishment devoted to data dissemination activities. In fact, these organizations perform an invaluable service in facilitating access to what often borders on an unusable product, in technical and economic terms. It is important to recall that most of the data made available by federal agencies are materials originally collected and processed for internal purposes only. With major exceptions, these data are neither collected, managed nor documented with external users in mind.

Typically the data are processed originally with an eye to convenient preparation of internal reports in some standard format that has been used by the agency for a number of years.

These procedures were audited and approved at some time in the past, and there is reluctance to alter them. As a result, data are often processed that include special characters or multiple punches, for example, because approved software is available to generate required cross-tabulations and other summaries. Because of standardization of procedures and because the same departments, perhaps even the same individuals, in the agency have been continuously responsible for performing these operations, only minimal documentation is required for current files. All of these factors work to the advantage of the agency in the performance of its required tasks but are detrimental to the extended use of such resources by secondary analysts. What is rational for production purposes is not necessarily rational for archival purposes or for the purposes of secondary analysis.

Additional problems are peculiar to the recent machine-readable data products of the Bureau of the Census. Normally the Bureau presents information on a state-by-state basis, and the organization of the data files for the 1970 Census reflects the fact. Moreover, the internal data-processing operations of the Bureau involve the use of an older, low-density (seven-track, 556 b.p.i.) tape system. Although tapes for dissemination can be written at higher densities, the data are not reconsolidated as they are copied. As a result, a given data file for Wyoming may occupy only a small fraction of a magnetic tape written in the Bureau internal format; it will require even less space on a reel written at a higher density. Analogous data for New York may span three reels of tape at the lower density; exact copies of the file will occupy only parts of three reels at the higher density. As a consequence the user must frequently acquire more data recorded on more reels of tape, with attendant higher costs, than would be required if consolidation had been carried out.

In terms of technical format of data records, the Bureau of the Census has made a concerted effort to adopt a "lowest common denominator" approach to accommodate the requirements of diverse computational environments. Thus it has adopted the use of 12- and 16-digit data fields even though the actual values recorded in these fields, aside from those for units

at the highest level of aggregation, are usually much smaller. Unfortunately, data fields of this width cannot be employed with many computer software systems and the user must carry out costly reformatting to convert data to a usable form.

The case of the 1970 Census well illustrates the employment of practices that seriously complicate and increase the costs of data acquisition and use. As supplied by the Bureau the basic aggregate data from the census were organized in a series of "summary counts." The earlier counts include a relatively small number of variables and are simple in structure. Some of the later counts include much larger numbers of variables and are highly complex in structure. Within each count the data are organized in state files each of which may include data at several levels of aggregation. Thus data for several types of units (states, counties, SMSAs and component areas) may be included in the same state file. The user whose interests are limited to data for a single state or a limited number of states and to data variables included in one of the earlier counts is reasonably well served by the Bureau's organization of the data. On the other hand, this organization is much less suited to the needs of users whose research requires data variables from two or more counts or whose interests involve data at a single level of aggregation (whether SMSA, county, tract, or other unit) for the entire nation or for a large number of states. In such cases the user must acquire a substantially larger body of data than his research actually requires with the consequence of significantly higher data acquisition costs. Moreover, the researcher then faces the cumbersome and costly task of subsetting the data to extract the specific units and variables required. These problems are, of course, greatly increased for the researcher whose research interests dictate data from the large and complex files characteristic of the later summary counts. Indeed, manipulation and subsetting of these files require software that is not available at many computer installations. It is fair to say that for some researchers, problems of this sort constitute an insurmountable obstacle to use of 1970 census data.

Taken in total, acquisition and use of federally produced data can involve all too often a process that can perhaps be best

described as "triple costing." Initially, federal taxpayers' funds are used to support original data collections and processing. Because of procedures such as those described above, an intermediate organization then carries out additional data-processing and performs dissemination services with either direct governmental support or, in some instances in the case of private organizations, with partial support through federal grants and contracts or grants from private foundations. Finally, the individual researcher or research group must purchase required data and then face the additional processing that is required to convert the data to usable form. And frequently funds are sought, and obtained, from federal research funding agencies or private foundations to support data purchase and processing. The need for improved management strategies and more effective use of public funds is self-evident.

V. CONCLUSIONS

The preceding pages have touched upon only a fraction of the data production and dissemination activities of the federal government. The massive data collection efforts of the Department of Health, Education, and Welfare were barely noted and the work of such agencies as the Department of Defense, the Bureau of Labor Statistics, and numerous others passed unmentioned. It is fair to say, however, that many volumes would be required to describe adequately the information-gathering activities of the various agencies of the federal government.

It was suggested elsewhere that the agencies of the federal government could be seen, in terms of potentialities, as constituting a vast data archive containing data relevant to the investigation of the society, economy, politics, and government of the United States in virtually all of their dimensions. It is probably not an exaggeration to say that these resources could provide a basis for major breakthroughs in the scientific understanding of human affairs. But social scientists have only

limited and ineffective access to these resources, and the record to date can most accurately be described as one of lost opportunities rather than realized potentials.

Explanations for the shortcomings of federal data dissemination mechanisms and policies are not difficult to find: the development of information technology has progressed more rapidly than human capacity to use it; funds have perennially been too scarce to allow creation of optimal facilities; neither the agencies of government nor social scientists themselves have fully recognized the rich research potential of federal data resources; and social scientists have not been active enough in making their needs and interests known or in seeking to employ these resources in their work. And encouraging signs of progress can be observed. Formation of the National Technical Information Service and the Machine-Readable Archives Division of the National Archives are two such signs. Comparison of the management and dissemination of the machine-readable data files that resulted from the 1970 census and subsequent enumerations with the management and dissemination of equivalent files from the 1960 census provide clear evidence of noteworthy—indeed, spectacular—progress. The Bureau of the Census has devoted substantial time and energy to planning for the 1980 census. It has sought the views of social scientists and taken cognizance of their needs and interests; and in its planning efforts the Bureau has directed attention to the need to develop means to the more effective dissemination and use of the 1980 data. Continuing technological development promises to ease the process of solving current problems, and a variety of other indications of progress can also be observed.

Yet achievement of the scientific potentialities of federal information resources will depend on a variety of factors—correct assessment of the directions of technological development, adequate diagnosis of pressing societal problems, effective identification of the most promising directions of social scientific research, meaningful support from the community of social scientists, and solutions to problems of confidentiality and privacy including more realistic appraisal of those problems. Above all, perhaps, adequate financial support will be required

both for scientific research itself and for the development and maintenance of necessary supporting facilities. The potentialities and promises seem clear, but at this writing their achievement remains in doubt.

NOTES

1. The reader should consult the full series of Working Papers and Technical Papers, examples of which are listed in the references.

2. The primary contact is Michael G. Garland, Chief, Data User Services Division, Bureau of the Census, Washington, D.C. 20233.

3. The monthly publication *Data User News* is available for the annual subscription price of $4.00 from the Bureau of the Census. The series entitled *Data Access Descriptions* contains four to six issues per year, numbered consecutively by date of issue. The usual price for an issue is $.50 or $1.00.

4. Inquiries about the catalog or available services should be directed to Program Manager, NTIS, 5285 Port Royal Road, Springfield, Va. 22151. The standard charge for a copy of data on a magnetic tape from NTIS is $97.50 for 1960 census data, which exceeds the current charge of $80.00 by the Bureau of the Census.

5. Requests for additional information about the data holdings and servicing policies of the Machine-Readable Archives Division of the National Archives should be directed to Dr. Charles M. Dollar, Chief.

6. Inquiries should be addressed to John Beresford, President, Data Use and Access Laboratories, 1601 N. Kent Street, Arlington, Va. 22209.

7. Information about the services and data holdings of the Inter-University Consortium for Political and Social Research can be obtained from Dr. Jerome M. Clubb, Executive Director, ICPSR, Box 1248, Ann Arbor, Michigan 48106.

8. Inquiries should be addressed to Andrew Sobel, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830.

9. Requests for additional information should be directed to Data Services Division, WESTAT Research Inc. 11600 Nebel Street, Rockville, Md. 20852 and/or Peter Francese, President, National Planning Data Corporation, P.O. Box 610, 20 Terrace Hill, Ithaca, New York 14850.

REFERENCES

COLEMAN, J. S. (1966) Equality of Educational Opportunity. Washington, D.C.: Government Printing Office.

National Data Use and Access Laboratories (1973) 1970 1/100, 1/1000, and 1/10,000 Public Use Sample State Files. Pricing memorandum of July 1973. Rosslyn, Virginia.

National Technical Information Service (1974) Directory of Computerized Data Files and Selected Software Available from Federal Agencies (March). Washington D.C.: Government Printing Office.

U.S. Bureau of the Census (1946) Bureau of the Census Catalog. Washington, D.C.: Government Printing Office.

--- (1974) Census Tract Papers. Series GE-40, No. 10. Statistical Methodology of Revenue Sharing and Related Estimate Studies. Washington, D.C.: Government Printing Office.

--- (1970) Census Use Study: General Description. No. 1 (March). Washington, D.C.: Government Printing Office.

--- (1961) Congressional District Data Book (Districts of the 87th Congress). A Statistical Abstract Supplement. Washington, D.C.: Government Printing Office.

--- (1963) Congressional District Data Book (Districts of the 88th Congress). A Statistical Abstract Supplement. Washington, D.C.: Government Printing Office.

--- (1973a) Congressional District Data Book, 93rd Congress. A Statistical Abstract Supplement. Washington, D.C.: Government Printing Office.

--- (1953) County and City Data Book, 1952. A Statistical Abstract Supplement. Washington, D.C.: Government Printing Office.

--- (1957) County and City Data Book, 1956. A Statistical Abstract Supplement. Washington, D.C.: Government Printing Office.

--- (1962) County and City Data Book, 1962. A Statistical Abstract Supplement. Washington, D.C.: Government Printing Office.

--- (1967a) County and City Data Book, 1967. A Statistical Abstract Supplement. Washington, D.C.: Government Printing Office.

--- (1973b) County and City Data Book, 1972. A Statistical Abstract Supplement. Washington, D.C.: Government Printing Office.

--- (1967b) Data Access Descriptions. Washington, D.C.: Government Printing Office.

--- (1973c) Population and Housing Inquiries in the U.S. Decennial Censuses, 1790-1970, Working Paper No. 39. Washington, D.C.: Government Printing Office.

--- (1972) Public Use Samples of Basic Records from the 1970 Census: Description and Technical Documentation. Washington, D.C.: Government Printing Office.

--- (1975) The Census Bureau, A Numerator and Denominator for Measuring Social Change, Technical Paper No. 37. Washington, D.C.: Government Printing Office.

--- (1973d) Summary Tape Processing Centers: Address List (June). Washington, D.C.: Government Printing Office.

U.S. Department of Justice, Law-Enforcement Assistance Administration (1974a) Crime in the Nation's Five Largest Cities: Advance Report (April). Washington, D.C.: Government Printing Office.

--- (1974b) Criminal Victimizaion in the United States, January-June 1973: A National Crime Panel Survey Report, Vol. 1 (November). Washington, D.C.: Government Printing Office.

U.S. National Archives and Records Service (1975). Catalog of Machine-Readable Records in the National Archives of the United States. Washington, D.C.: National Archives Trust Fund Board et al.