Archival Data Resources

MICHAEL W. TRAUGOTT
University of Michigan

The development of data archives containing computer-readable resources relevant to the study of aging reflects the growing awareness that social scientific models can be appropriately applied to solving problems in this area and that greater advantage must be taken of substantial investments already made in underutilized original data collections. Relatively immediate research returns and policy analyses are possible from the extended analysis of such resources conducted in response to pressing national needs. And such efforts can be achieved without the lag and expense associated with launching new studies each time a new question is raised about problems of the aged or the process of aging. At the same time, the use of increasingly complex research designs associated with ever-larger data bases is straining the computational resources available to the average researcher, in both technical and financial terms. With computer-readable data resources becoming available in ever-increasing amounts, the challenges for archives are to develop means for simplifying access to increasingly obstreperous data sets on a routine basis and to train a new cohort of analysts in their use.

The availability of computer-readable resources through a social science data archive serves both a scientific and an administrative function based upon the concept of secondary or extended analysis. The underlying notion of secondary analysis is that no principal investigator ever exhausts the full analytic potential of a data set. In one sense, this is the case simply

---

AUTHOR'S NOTE: Prepared for delivery at the Conference on Demographic and Health Information for Aging Research: Resources and Needs, National Institute on Aging, National Institutes of Health, Bethesda, Maryland, June 25-27, 1979.

RESEARCH ON AGING, Vol. 3 No. 4, December 1981 487-501

because any single researcher, or group of researchers, has a limited perspective on the data with regard to a model based upon a specific set of operationalizations. As the process of theory development and testing occurs or, more likely in the case of progress in the social sciences, as additional disciplinary perspectives are brought to bear in the field, data collected for one purpose come to be viewed quite differently in conceptual and operational terms by other researchers.

From a scientific perspective, the significance of archival data rests upon the principle of replication, the notion that confidence in the validity of a relationship between two or more variables is increased by the number of times and the variety of settings in which it can be reproduced. As an elemental fact, the original relationship should be reproducible by secondary analysts in the original data base. But more importantly, data archives provide a central repository for multiple collections of data in which the same relationship may be tested in a variety of research contexts.

Replication has much greater social significance than as a simple scientific tenet. Where matters of public policy and planning which affect the lives and well-being of individual citizens are concerned, programmatic alternatives must be subjected to public scrutiny and reasonable systematic reanalysis before they are implemented. The process that we call "policy evaluation" really represents nothing more than successive replications of basic relationships over time in which we are testing the hypothesis that desired goals have in fact been achieved.

From an administrative point of view, archival resources represent a cost-effective means of increasing research opportunities and, implicitly, results. These cost reductions are available in time as well as money. The greatest single cost involved in the conduct of research is for data collection and processing. This cost is associated primarily with the complexity of the research design and the number of units of analysis or observations required. Beyond the initial data collection effort, considerable resources must be devoted to transcribing information to computer-readable form and then checking the resulting data matrix for consistency and other technical errors. Then inevitable prob-

lems that are found must be resolved and corrected before analysis begins. In addition to financial resources required for research and support staff, field costs, and the like, this process often requires substantial periods of time, even years in the case of extensive longitudinal studies. The alternative, when appropriate archival resources are available, is construction of a research design followed by a search through documentation for meaningful operationalizations of important theoretical concepts. Although there is often an iterative process involved in modifying the original design in accord with available data resources, this process consumes relatively little time. And when the selection of archival data has finally been made, copies of the files can usually be delivered to the analyst in a matter of days or weeks at a cost effectively equal to the cost of duplicating a magnetic tape. If archival resources are sufficiently rich and diverse, they provide an unusual and effective opportunity for meaningful research by skillful analysts and policy makers. It is certainly the case that a wide variety of important data sets relevant to the study of aging are currently available from a number of sources.

Data Resources in the Field of Aging

A variety of computer-readable data resources for the study of aging may already be obtained from several different sources. These organizational sources include federal agencies, principally the Department of Health, Education, and Welfare, but with the significant inclusion of the Bureau of the Census in the Department of Commerce; various academic data organizations; and a small number of profit and nonprofit organizations which serve primarily as intermediary processing and dissemination mechanisms for data originally produced by the federal government.

The growth of nonfederal data archives can be directly attributed to the difficulties social scientists have encountered in acquiring and analyzing data resources from federal agencies. Perhaps the most frustrating problems confronting researchers who seek to employ federally collected data are simply obtaining

information about the availability of relevant data files and identifying appropriate modes of access to them. Then the researcher must often carry out extensive processing to reorganize, subset, and otherwise restructure data files before analytical work can begin.

The costs of acquiring federal data files vary widely by agency and in relation to their size. These costs are often superficially low because of the extended processing that must be carried out to make the files easily usable or because the researcher must often "overpurchase" inappropriately organized data files containing substantially more data than are actually needed for a given analysis. Because of potential obstacles such as these, social science data archives have assumed an intermediary role where dissemination of federally collected data is concerned. Acting as a kind of "buyers' cooperative," these organizations assume initial costs of purchase of data where necessary, capitalize the reprocessing costs for data and documentation through a central facility, and then provide redissemination services, including subsetting capabilities, by which their constituents obtain exactly the data in which they are interested. A number of federal agencies now commission data archives to carry out these intermediary functions and provide financial support for these and related activities, such as methodological training for researchers interested in using such resources. As an example of such an activity, the Administration on Aging has supported the development of a National Archive of Computerized Data on Aging (NACDA) by a two-year grant to the Institute of Gerontology and the Inter-University Consortium for Political and Social Research at the University of Michigan.

The data resources being developed by NACDA as part of the Consortium archive are representative of the major areas in which such materials are available for the study of aging and the aged. These resources may be broadly characterized as aggregate demographic data, information from administrative records, and individual-level data from surveys. In each of these three categories, data are available from federal and nonfederal sources which have been underutilized in the past and which present significant opportunities for increasing our understand-

ing of aging through careful and comprehensive secondary analysis.

Aggregate demographic data. The United States Bureau of the Census is, of course, the largest and most important data collection agency in the nation. The tabulations from their decennial efforts provide us with the basic geopolitical statistics about the elderly population to be serviced, the size and characteristics of the full range of age cohorts in the population, and basic employment and occupation data as well. Increasingly, census population counts are being associated with other variables, such as geographical boundary information, which makes them amendable to computerized cartographic analysis. Development of these capabilities is certain to improve the utility of these materials for area planning agencies. Census data are of equal importance to researchers and policy makers as they measure the size and composition of service populations and form the basis for allocation of public expenditures.

The decision of the Bureau of the Census to release almost its entire 1970 data product in computer-readable form was a great benefit to all social scientists. This action also presented a major challenge to those who sought access to the data, as over 2000 reels of magnetic tape were included in the release of these materials in the original format. The Bureau has already made a substantial investment in planning for the release of computer-readable products from the 1980 census, which will be even more complete and extensive than the 1970 census data. And the census will be conducted on a quinquennial rather than a decennial basis beginning in 1985, with the clear implication that large quantities of current statistics will be available on a more frequent basis from now on.

These future developments demand that more planning be devoted now to ensure effective utilization of these resources. Although this topic will be dealt with in greater detail below, attention must be given to ascertaining which particular elements of the computer-readable products of 1980 and subsequent censuses are of greatest interest to researchers and policy makers interested in aging, what the most likely technical formats are in

which such data will be required, and what acquisition and dissemination mechanism should be established to satisfy these needs on a cost-effective basis. Given assumptions that some common needs exist and that the volume of data and cost of acquisition are so great that large numbers of individuals will not be able to afford to purchase their own copies, the task is to determine how we can maximize access to these valuable resources.

Information from administrative records. Some of the most policy-relevant and significant research findings in the various subfields of gerontology have been derived from computer-readable data files consisting of administrative records appropriately converted for public use. In particular, various data files made available by the Social Security Administration and the National Center for Health Statistics are of research value because they consist of large numbers of cases containing information in great detail. These data files allow analyses of rare populations or samples of individuals with particular characteristics to be undertaken with confidence in the reliability of the findings.

For example, computer-readable data are available annually from the National Center for Health Statistics (NCHS) in the Mortality Detail File, which consists of information on all deaths registered on individual death certificates in the United States since mid-1972. These data files consist of almost 2,000,000 such records for each calendar year. Containing basic information on the age, race, and sex of each deceased person, as well as cause of death, these data files provide the primary opportunity for relatively detailed epidemiological studies at the national level. The National Center for Health Statistics also collects and maintains data on institutions and services, as well as on individuals. This type of information is provided in data files such as the Master Facilities Inventory, which is intended to be a comprehensive list of the facilities in the United States that provide medical, nursing, personal, or custodial care on an inpatient basis to groups of unrelated persons. One data file contains information on hospitals, including classification by facil-

ities, services, staff composition, and staff size. The Nursing Home and Other Facilities file describes nursing, personal, or domiciliary care homes by such features as type of ownership, number of beds, number of residents, and services provided.

Data from the records of the Social Security Administration (SSA) are useful primarily for research questions related to earnings, pensions, and disabilities. Data from this agency are often combined with information collected by other government offices or bureaus, or are used by the agency as the basis for drawing samples of individuals for the purposes of conducting further surveys, some of which are described below. One of the largest of the public use files created from SSA records is a data base containing longitudinal earnings information from 1937 to the very recent past. Some of these data have been augmented with information from the Internal Revenue Service for selected years, while in other cases this information has been combined with individual-level data collected as part of the Census Bureau's Current Population Surveys.

It is obvious from the sensitive nature of certain of the items contained in these data files that extreme care must be taken in preserving the confidentiality of individual administrative records. By and large these data were originally assembled on individuals for bureaucratic purposes having no relationship to research interests. Their initial compilation in computer-readable form was for internal programmatic analyses by governmental employees only. Their availability in public use form, after extensive purging of information which might serve to identify individual case records, is a very recent phenomenon of the 1970s; but it is a movement for which nongovernmental researchers and analysts are extremely grateful.

Individual-level data. For more than 25 years survey research methods have been used to collect data about individual-level attitudes and behavior on a systematic basis. To the extent that any sample drawn appropriately of standard size to be representative of the United States population as a whole, there is usually some number of respondents available for analysis in the upper age cohort. However, there are also many surveys

available that are either based entirely upon samples of the middle aged and elderly or that contain special strata and subsamples to permit analysis of this group of people.

When talking about individual-level data collected by the survey method, it is useful to distinguish between data files resulting from the projects conducted by federal agencies and those collected by individual principal investigators (usually based in colleges or universities). The reasons for this are twofold. The sample sizes in the federal surveys tend to be very much larger and often result from more complex designs. The information collected in these efforts tend to be more or less "objective," consisting of reports of behavior, expenditures, and the like, particularly as they relate to available federal programs. The data collected by academic-based researchers tend to range much more broadly in content, often containing more "subjective" assessments of the individual's life cycle and status, satisfaction, and interpersonal relations. The sample sizes are usually relatively small in comparison to federal data collection efforts. These distinctions are somewhat sharply drawn here for emphasis to indicate the general directions in which a secondary analyst searching for data resources relevant to a particular research question might turn.

In the case of federal survey data collections appropriate to the study of aging, the major sources of information are, again, the Bureau of the Census, the National Center for Health Statistics, and the Social Security Administration. The Bureau is responsible for release of the Annual Demographic Files derived from the March Supplement to the Current Population Survey. These data contain the basic elements of income, age, race, sex, household size and composition, education, and occupation used to construct the official annual intercensal estimates of the characteristics of the U.S. population. The data files for each year contain approximately 200,000 data records on individuals as well as families, and they are available beginning with 1968. In 1976, the Bureau conducted a special Survey of Income and Education for the Department of Health, Education, and Welfare consisting of a sample of approximately 158,500 households. In addition to demographic data, information was obtained on

school enrollment, disability, health insurance, bilingualism, food stamp recipiency, assets, and housing costs. These data are reported by household, family, and individual in a complex hierarchical structure. The significance of this collection is that it was designed for the estimation of state-level statistics, and these data are the major current source for this type of information.

The Longitudinal Retirement History Survey, begun by the Social Security Administration in 1969, represents the most substantial effort of this agency in this type of data collection. Taking off from a sample of administrative records, this panel study concentrates on a cohort beginning with individuals at pre- and postretirement years and follows them through into retirement. The emphasis in the study is on expenses, income, and retirement benefits; and extensive analyses of provisions of the Social Security system as well as private pensions have been carried out using these data. Information for the first three waves of the survey are currently available through 1973, and it is anticipated that data from a fourth wave will be available in the near future. Other data files released by the Social Security Administration include earnings data for individual respondents in the Longitudinal Retirement History Survey, a 1972 Survey of the Status of the Elderly, a 1970 Survey of Newly Entitled Beneficiaries, and a 1973-1974 Survey of Low Income Aged and Disabled.

The major survey conducted by the National Center for Health Statistics which has particular relevance for researchers in aging is the Health Interview Survey. This survey, for which the 1975 data constitute the last public use file available, represents a continuous sampling and interviewing of the civilian, non-institutionalized population of the United States in order to collect information on the social, demographic, and economic aspects of illness, disability, and medical services. Data are available on the amount and distribution of self-reported illness and the effects of disability, as well as on the utilization of medical care facilities. For many social scientists, these data represent the major alternative to longitudinal epidemiological studies.

In the area of special surveys of the elderly conducted by individual principal investigators not associated with federal service agencies, the National Archive of Computerized Data on Aging has begun to assemble data sets from a series of separate national cross sections covering a 20-year period. At the beginning of the sequence are 2 surveys of the elderly conducted by Ethel Shanas through the National Opinion Research Center in 1957 and 1962. Then, in 1968, Kermit Schooler directed the National Senior Citizens Survey, which also included a reinterview with slightly more than 500 of the original respondents in 1971. In 1974, Louis Harris and Associates conducted a survey for the National Council on Aging dealing with the myths and reality of aging. While these five studies are obviously not exact replications of each other, they provide the basis for extensive analysis of changing American attitudes and opinions toward aging and the aged, information not generally available in the federal surveys. Even though these data are cross-sectional and not longitudinal, they contain important data on residential environments, social relationships, moral values, and other subjective indicators of the quality of life of elderly Americans.

There are three additional survey projects of substantial dimensions that are of particular relevance to social gerontologists, although the data collection efforts were originally designed for other purposes. The first of these is Herbert Parnes' National Longitudinal Surveys of Labor Market Experience, designed originally to analyze the sources of variation in the labor market behavior and experience of four age-sex cohorts of the U.S. population through interviews conducted at least biennially. Initiated in 1966, the cohort of "mature men" who were then 45 to 59 years of age now includes many individuals who have passed into retirement. The cohort of "mature women," who were 30 to 44 years of age 13 years ago, remain in their preretirement years even today. But interviewing continues, and these respondents are still being tracked. Each cohort entered the survey through an initial sample of 5000 individuals, consisting of 1500 blacks and 3500 whites drawn to provide a weighted national cross section. The major topics covered in the surveys

include labor market experience, the development of socio-economic and human capital, and environmental variables.

A second major survey project is James Morgan's Panel Study of Income Dynamics, consisting of annual interviews with almost 5000 American families since 1968. Originally designed for a study of the determinants of family income and its changes, the sample is composed of approximately 3000 families in a representative national cross section and about 2000 low-income families. Data are available at both the family and individual levels, and the decomposition and recomposition of the original families through regular life cycle events results in the continuous addition and deletion of families and individuals in the data base. The information obtained in the survey falls generally under the headings of the economic status, behavior, and attitudes of the respondents and their families; and it includes such specific topics as employment, income, housing, disability, time use, and family background and composition. Because of the relatively large size of the sample and the longitudinal design, this survey provides unparalleled analytical opportunities to assess the economic behavior and circumstances of the aged in a representative national context.

Finally, an extended series of data files dealing with access to medical care are about to become publicly available. Ronald Anderson at the Center for Health Administration Studies at the University of Chicago has recently completed the fifth in a series of national surveys on this topic which spans more than 20 years. Surveys were conducted by the National Opinion Research Center in 1953, 1958, 1963, and 1976. The topical coverage of the most recent survey, which will become available first, includes information on usual sources of care, perceptions of health and recent episodes of illness, a summary of disability and physician visits, the utilization of specific medical services, information on insurance coverage, and standard background data. Interviews were obtained from 5554 individuals in 3896 families, which included 619 rural southern black families and 531 southwestern Spanish heritage families. These data represent the major nonagency source of public information on medical

care access; and the sample size is sufficiently large to support analysis of subsets of older respondents.

The information presented above illustrates the range of public use data files currently available from social science data archives for researchers and analysts interested in the study of aging. There are substantial data resources available for secondary analysis in a variety of gerontological subfields. These data are available relatively quickly and inexpensively, especially in comparison to the costs of original data collection. But increasingly these data files present problems for researchers with limited technical support facilities and/or computer funds because of their size and complexity. This situation raises a set of issues that must soon be faced if the pace of research is to match the need for increased knowledge and understanding.

Research support needed in the study of aging. Experience and information accumulated during the initial project period devoted to development of the National Archive of Computerized Data on Aging suggests two areas to which particular attention should be devoted to support further research in the field of aging. These are the development of additional technical resources to support analysts who want to use archival data resources of increasing size and complexity, and the need for better and more available training for individuals interested in utilizing such resources.

From our daily interactions with social scientists interested in a wide variety of topics in addition to the study of aging, it is clear that secondary or extended analysis has become a common research form. The perception that data represent a resource that is to be shared, particularly when they were collected with federal funds, is also a widely shared value. As a result we now see a substantial increase in the number and variety of data resources being made available for such analysis. These are all positive accomplishments for which all social scientists must be grateful. Analytical problems arise now not from the scarcity of relevant data resources, but from manipulating them due to their very size and complexity. The potential exists for research opportunities to be limited to individuals at only the technically

best endowed institutions unless action is taken to reduce sharply the costs of access to these resources.

In recognition of the fact that social phenomena are complex and that social scientists interested in the study of aging are concerned with a significant subset of the total population, the research designs for studies in this area are being developed with corresponding complexity as well. Data are being collected simultaneously for multiple levels of analysis, with all of the data being incorporated in hierarchical structures in the same data file. It is increasingly common, for example, to find data for a family unit, individual members of the family, and health or retirement-related data for each individual all sorted within the same data file. For substantive reasons, this is an appropriate format because the analysis of individual-level effects or treatments should take place within the appropriate social and environmental context. And for technical reasons it is also appropriate to store data in this nested structure because it is efficient and effective in reducing data processing costs. But the fact remains that general-purpose analytical software designed to operate on a variety of data structures of this general form is not yet widely available.

While other data management software is available that can be used to construct the more common rectangular data records utilized by most general-purpose software systems from these hierarchical structures, the very size of many of these data files results in prohibitively expensive retrieval costs. We're talking now about sample sizes of 40,000 households and almost 100,000 individuals in the Current Population Survey data, for example, as opposed to the standard 1500 to 2000 respondents in more routine public opinion surveys; and the information for each unit of analysis is contained in several hundred variables. In the major longitudinal studies, such as those of Morgan and Parnes, which permit more sophisticated analyses because of the very nature of their designs, the number of variables runs to the thousands. We are approaching or may have already exceeded our ability to make these data easily used and widely available on a routine basis.

The implication of these circumstances, if we want to continue to provide meaningful access to archival data resources of increasing value, seems to be a rethinking of current dissemination procedures. It is now customary for archives to make copies of their data files on magnetic tapes, which are then mailed or otherwise transported to the interested analyst. This system may have to be reviewed in light of the need to associate specific large and complex data files with appropriate hardware configurations on which resides the necessary software to analyze these data with a variety of techniques in a cost-effective way. Much of the technical infrastructure required to support such a system of access is already in place. This includes national computer networks that link major computing installations across the country, providing access to remote users for the cost of regular charges on the host system plus the additional cost of a small charge for connect time and the convenience of a local telephone call. What is still required is a system for associating relevant data with appropriate hardware and software and for obtaining sustaining support for staff who would provide consultation to individual researchers desiring access to such resources. Given that one or more of these national research centers could be established on a basis similar to that described above, researchers could then request support for analysis in terms appropriate to that facility's software and data resources, rather than being limited to those of his or her own institution.

As a corollary to this need, additional training opportunities must be provided for researchers interested in the study of aging. At one level, basis training in social science research methods must be made available to young scholars and retooling opportunities extended to these already working in the field. This will result in a general increase in the quality of research being conducted. Furthermore, the availability of large and complex data bases more often incorporating longitudinal elements in their design requires that advanced methodological offerings be available as well. And should a movement develop whereby data and other technical resources become consolidated in central locations, training in modes of access through computer networks incorporating distributed processing techniques would also be necessary.

Researchers interested in the study of aging should be optimistic about prospects for the future because they stand on the threshold of a period in which unprecedented technical resources for extended analysis are being put at their disposal. More data of higher quality are available now than at any time in the past, and the prospects for significantly incrementing this resource base are considerable. We must devote appropriate energies to planning that will ensure that researchers and policy makers can adequately utilize these materials in expanding our knowledge about aging and the aged and in designing appropriate social programs to meet their needs for services.