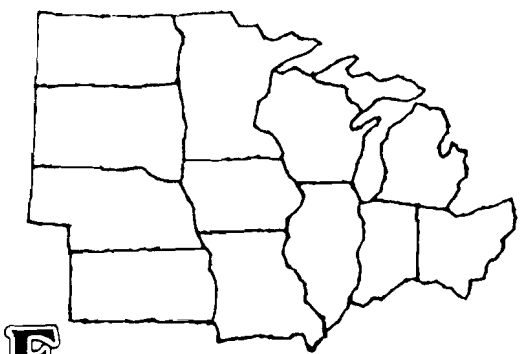
Robbin, A. (1979). Understanding the machine readable numeric record: Archival challenges, with some comments on appraisal guidelines. *Midwestern Archivist, 4*, (Fall 1979), 5-17



THE MIDWESTERN ARCHIVIST

VOLUME IV, NUMBER 1, 1979

THE MIDWESTERN ARCHIVIST

VOLUME IV, NUMBER 1, 1979

CONTENTS

UNDERSTANDING THE MACHINE READABLE NUMERIC RECORD: ARCHIVAL CHALLENGES, WITH SOME COMMENTS ON APPRAISAL GUIDELINES

ALICE ROBBIN

INTRODUCTION

Quantitative social research has been considered necessary to understand the nature of our past and present society and its problems, carry out the research process, and understand the impact of social programs. For almost two centuries, social researchers have utilized governmental statistics to examine a wide range of social phenomena on individuals and social structures and to offer competing analyses of social research and policy. As early as 1828, Melchiorre Gioja, in his *La Filosfia della Statistica*, indicated that the essential scope of social science also embraced the collection of data on a wide range of social phenomena, such as affluence, poverty, knowledge, ignorance, happiness, civilization, and morality, "the set of data relative to a country, which in the daily conduct of affairs, can be useful to each one or to the majority of its citizens, or to its government, which is their agent, delegate, or representative . . . ."[1] In the late nineteenth century, Durkheim relied on published summaries of statistical data to illuminate the causes of suicide. Since the beginning of the twentieth century, technological and intellectual advances have made possible the recording of these statistics in machine readable form, the collection of new types of data, and increased utilization of governmental records.[2]

Recent decades have witnessed the enormous expansion of information recorded by government. Statisticians, economists, and program administrators have increasingly required more detailed and varied data on the status of state and local units of government; on the composition and characteristics of the population; and on the nature, scope, and effects of governmental programs designed to benefit the well-being and general welfare of the citizens of the state. Federal legislation has mandated the collection and analysis of this information as a prerequisite for the allocation of federal resources. Records are used to allocate resources at the state and local levels of government; to plan, audit, and evaluate the distribution of resources; and to ensure adequate planning for human needs and an equitable delivery of benefits to the citizens of the state. These records include information on taxes, public assistance payments, budgetary decisions, pollution, land use, energy consumption, transportation, education, surveys of the population or of particular groups, and of transactions and activities of individuals, businesses, associations, and other organizations. Some of these records have long-term value for historical and other research into social, economic, and political behavior as well as governmental activities and processes. These records offer an unprecedented range and detail for understanding the dimensions of social change.[3]

A major reason for the expansion and use of information and its expected increase in the future is that computer technology has radically altered government's record collection methods and management practices, and made available new tools for the scholarly manipulation of data. Computer technology has greatly increased the potential for information collection, storage, management, and utilization, and it also facilitates the collection and documentation of highly detailed information describing individual characteristics, human transactions, and organizational processes. The computer makes possible the more efficient and effective administration and management of government. It is a powerful tool for implementing programs and policies that require long-term maintenance of records. This tool also assists program administrators to identify human and social problems and to devise and implement solutions in a timely and more efficient manner.[4] Computer technology provides scholars with a variety of tools to manipulate large quantities of information, thus enhancing the dimensions of their investigations. Unlike the traditional published format, raw data in a form suitable for direct input to

and manipulation by modern computing machinery facilitates analysis and reanalysis by permitting its reorganization and summarization at minimum cost and with maximum speed. Modern computers allow the scholar or analyst to order and summarize complex situations, make more understandable the "reality" of a particular subject, establish the relationship between independent events, and ascribe causality or make prediction. Where analysts differ in their needs, the computer allows them to reorganize, summarize, and analyze the same information from a variety of perspectives.[5]

The computerized record offers advantages to the archivist and scholar. Information in this form can be manipulated—sorted, edited, reformatted, counted, tabulated, and cross-tabulated—, selectively retrieved for visual inspection, and utilized for a variety of purposes in ways which were not possible with conventional records. Data gathered for specific purposes may be of additional utility for a variety of other purposes—not all of which were or could have been anticipated when the original data were gathered.[6] The facility with which data manipulation can now be accomplished has given a considerable impetus to the development of new social science theories[7] as well as the analysis of more complicated social phenomena.[8]

In the past, raw information was frequently destroyed because it could not be stored, thereby precluding effective scholarly use of the materials. Today's computer technology allows preservation and compact storage of enormous quantities of highly detailed information.[9] Maintenance of information in conventional paper format often required severe access restrictions to protect anonymity of particular cases, a problem which can be alleviated by utilizing the computer to delete or mask the identity of the individuals involved.[10]

The value of any archival record is enhanced by the existence of other sources, which when used in tandem more completely describe the social, administrative, or economic process. This is especially true of the machine readable record, where source materials from one file can be more easily linked with other files to provide more complete documentation of particular events and transactions, thus augmenting the potential value and increasing the analytic or explanatory potential of the original records.[11]

ARCHIVAL CHALLENGES

The machine readable record poses new problems for the archivist. The categories, volume, and detail of information in machine readable form require new appraisal criteria, techniques, and resources. These unconventional source materials will place an added burden on the archival staff to preserve and make available materials formerly not part of the archives accessions.[12]

As scholarly research increasingly depends on statistical analysis, the archivist must reevaluate traditional appraisal criteria and expand the scope of his accessions. Appraisal is complicated because there is often little information to describe the historical development of these records. Linkage of data items from a variety of other records makes it difficult to identify source documents. Complex logical and physical structures require special computational and software access and retrieval capabilities. Resources are unavailable to identify, accession, and preserve all the records.[13]

Once these records are accessioned the archives may be required to provide new services. Donor agencies may require retrieval and/or linkage of particular records. These activities may occur, for example, if particular data series require special archival protection because records contain confidential information or will be used for program evaluation in the future. The archivist will be required to learn new technical skills and to develop new administrative procedures to facilitate effective use of a new, rich source.[14]

Another problem relates to the technology which has produced these records. Most of the machine readable information will never be converted in its entirety to a form which can be examined without the aid of electronic or mechanical equipment. The archivist must identify available computer facilities because machine readable records are typically generated on a variety of machines. Because computer technology is changing so rapidly, the archivist must acquire advanced information about computer technology and future technological developments. Therefore, the archivist requires specialized technical skills and access to appropriate computational equipment and programs.[15]

Data quality is also a problem because many machine readable data files are poorly constructed and processed. Not properly edited and corrected during the processing stages, files contain many errors. No

standards exist for what constitutes a "clean" data file, that is, what data checking and editing procedures ought to be carried out. Because data files are generally in poor condition, analysis can become a more expensive undertaking than generally assumed; more computer time and staff resources are needed to prepare a data file for analysis than are required to perform the actual analysis.[16]

Data are not always available at the microlevel, but have been summarized and aggregated. When data are aggregated to a higher level than their original state, significant distortions and biases may be introduced into the file. As a result, verification of the original data collection and reports, and analysis not envisioned by the original collectors of the data may be precluded.[17]

The development of idiosyncratic coding schemes for standard data items may create confusion and inconsistencies for all data analysts. Not only are these coding schemes developed on an *ad hoc* basis, they are incomplete and inconsistent with federal guidelines; therefore, they create expensive and time-consuming problems for the person doing comparative analysis or merging data files.[18]

Some of the characteristics which increase the value of machine readable records also create archival problems. Machine readable records are easily reformatted, updated, and otherwise changed, copied, and transmitted. These activities performed on the record make identification and definition of the record copy especially difficult. For example, machine readable files are often copied and diffused through numerous agencies, confusing the issue of which file or version of a file, is the record copy. The ability to update and the reusable nature of the medium lead to the mistaken notion that machine readable records are ephemeral. Unlike conventional records, which have a growing physical presence that demands attention, machine readable records can be painlessly and easily destroyed by deleting a few electronic or magnetic signals. Data processing personnel who typically control access to and retrieval of machine readable records and other individuals who have participated in the development of a data system are for the most part unaware of the potential future value of the information they handle. The archives faces psychological, political, and procedural difficulties in convincing agencies that machine readable records are not ephemeral and must be considered potential archival materials just as conventional records are.[19]

The documentation required to understand the contents of the machine readable records and how to access them is another problem for archivists. Access to machine readable records is possible only through a "map" or set of directions (typically called a data dictionary, codebook, or "user's guide"), information describing how to access the physical structure of these records (sometimes called file documentation), and information on the relationship of the records to associated software or hardware (sometimes called system documentation). Lack of information about the data's physical structure inhibits access. Many data files are transferred from one computing site to another without adequate description of the physical structure of the file and how the data were written. Few data files are received with accompanying technical information on the file structure, recording density, and character configuration. Computer center personnel typically have experience with one computer manufacturer and often only within one computing environment. As a result, there is no need to know about other computers and their internal characteristics. Although standards for writing data in transportable mode have been written, few programmers are familiar with them and few researchers know how to request data written in a transportable mode. The quality and amount of descriptive information about machine readable records vary greatly because there are no standards for what constitutes good documentation. There are also other reasons for the generally poor quality of documentation. Documentation has been perceived as unimportant. Inadequate attention has been given to its preparation. Most data collections have evolved to meet the needs of an individual researcher or program and there has been little thought that the data would be utilized by others not involved in the original data collection. Neither program administrators nor scholars are rewarded for data gathering, so it is not unexpected that they should not be rewarded for documenting their data gathering procedures and file creation and processing. Funds are rarely available to document a data collection. In addition, text documentation is poor because the operations necessary to produce the data file are not well understood; partly this is due to poor conceptualization of the research problem and a faulty research methodology. Most text documentation is created after-the-fact, that is, after the data collection and processing, or completion of the project, when most of the staff has disbanded and memories have faded. In the case of administrative records, documentation is often

nonexistent. When documentation is insufficient or nonexistent, it becomes difficult to maintain control over the records. This lack of control has made it difficult to determine the provenance of the machine readable record and directly affects its appraisal, analysis, and classification.

Even if documentation is adequate, most machine readable records are stored on magnetic tape for long-term preservation and this medium is short-lived when compared to conventional storage media such as paper and microform. Although the quality of this storage medium has improved in recent years, magnetic tape is still a fragile medium of storage. Maintaining an archival file requires regular and systematic inspection, cleaning, and copying. This poses an archival problem unique to this type of record: a wholly new set of costs associated with preservation and maintenance.[21]

An official data gathering agency collects personal or confidential information for its routine administrative functions and for measurement of other governmental activities and functions. Sometimes this information is obtained through either implied, actual, or believed threat of retaliation, with no assurances that the information will not be used against the individual who provided it, but most often information is obtained by "good will" and voluntary cooperation, with assurances that it will not be released to third parties in any identifiable form. As official data gathering agencies fulfill their research, statistical, and administrative functions, the data they gather constitute a valuable resource for the social scientist's research and statistical activities. As funds for primary research become more difficult to obtain and greater demands are placed upon government to become more accountable for the programs it institutes, the research community is becoming a more active consumer of both research and administratively produced information of a confidential nature and is beginning to rely increasingly upon the archives to preserve information.[22]

The archives has always played an important role as custodian of sensitive information, providing both security for collections of sensitive information and legitimate access to scholars and individuals. Archivists attempt to balance the personal requirements of the individual subject's or organization's right to privacy and society's need for knowledge.[23] But while personal privacy and freedom of information have been cherished in our society, these concepts have created an ambiguous role for the archivist. A recent examination of

confidentiality and the role of the archives elaborates on the problem.

> The keeper of the historical record has been placed squarely in the middle of a growing, ambiguous situation. The traditional passivity of the archivist no longer affords him protection. The archivist, in fact, has become a central element in the conflict. He has to become simultaneously an advocate and a protector of both sides of a complex and sensitive issue, while foregoing the sheltered and far more placid role of dutiful purveyor of the record.[24]

State laws as well as a growing number of federal laws define and limit access to information of a confidential nature, but do not define an archival role, set of responsibilities, or administrative guidelines.[25] The undefined status of the archives' responsibilities has resulted, in many cases, in a reluctance to deal with these records.[26]

APPRAISAL GUIDELINES

The value of all records series is enhanced when "there are enough examples over time to permit investigations of sufficient scope for drawing valid conclusions."[27] Archivists must consider the file's potential for linkage with existing record series in the collection or located at other archival institutions, since this condition is even more applicable to machine readable records. The machine readable record should be evaluated not only in terms of its record content, but its relationship to other machine readable record series, paper documents, and other information which may be translated into machine readable form. Machine readable records which record information over a period of time and record this information in a standardized compatible format (i.e., definitions of informational components do not change) enhance the archival value of the record series and enhance the "validity of research based on archival sources."[28] While informational records of this type should be appraised for retention, the unique record series may also have important research values and should not be rejected on the basis of its nonregularity.[29]

Documentation is the necessary component for evaluating machine readable records. Documentation provides a conceptual framework for the collection process, communication and coordination of processing staff, historical reference, general instruction for communication between specialist and nonspecialist, and is a report on successful

output. Documentation should be comprehensive and include or anticipate all the normal questions a user might have about the data file. Documentation must identify sources and contents, creators of the record series, its statutory or functional origins, processing history, subject matter contents, and reports on and uses of the data.[30]

Maynard Brichford, archivist of the University of Illinois-Urbana, and author of the Society of American Archivists Basic Manual Series volume on *Archives and Manuscripts: Appraisal and Accessioning*, notes that "credibility" of the record "is a basic research value."[31] This notion can be transferred to the machine readable record and used as an important appraisal criterion for data quality. Data should accurately represent the original information when it was collected, reflect the subject contents described by the documentation, and be carefully edited, processed, or prepared.

Data are most useful when they are at the lowest level of aggregation. The archivist must place a high priority on retaining machine readable records which record the primary, original data, or those data which are not summarized or aggregated, because these records offer greater value for future analysis and linkage to other records. These data allow the scholar to examine interrelationships and changes over time among variables, and they enhance the potential of records which are not yet available to the public. While confidentiality rules may preclude access to some of these data, it is imperative that they be preserved in their original, microlevel form.[32]

Were the standard for machine readable records reserved only for carefully prepared files, few machine readable records would be appraised for retention. And, indeed, an evaluation of poor data quality should not mean rejection of the records. Machine readable records are evidence of the policies and practices of an agency; it is therefore important to preserve the information upon which an agency bases its decisions and justifies its activities. As Dollar notes, "the manner in which data is [sic] collected and used can reveal a great deal about the agency's perception of its mission." He cites the example of the data collection and analyses of the Immigration and Naturalization Service, which suggest that the Service dramatizes and even exaggerates the problem in order to "justify both the activities of the agency itself and future control over immigration into the United States."[33]

Technical considerations are a critical aspect of appraising the machine readable record. Much more than other records, the machine

readable record is tightly bound to its storage medium and the equipment and programs required for access and retrieval. How records are stored and accessed depends on the computer on which they were processed and the software used to write them. Whether the data are independent from the software and computer will depend on how they were initially prepared and used. Information must be supplied on the physical characteristics of the file, whether the data are available in batch or interactive mode, and the logical composition of each record type. Proper computational equipment, software, and personnel (specialized data and computer expertise) must be available to the scholar and archivist, if they are to access and retrieve machine readable records.

The current computer storage and retrieval technology emphasizes development of data base management systems which create machine readable files in internal character configuration and structure that create difficult appraisal and preservation conditions for the archivist. Machine readable records, which need no associated software for processing and analysis offer the greatest possibility for increased future use by scholars. Appraisal policy must emphasize preservation of machine readable records which are software and computer independent. If the records are not independent, the appraisal staff must decide whether the records warrant their conversion to independence and whether resources can be allocated to convert the data.

The archivist appraising machine readable records must consider the costs of processing, preservation, storage, and use. These costs are higher than for other records, particularly because the records cannot exist without their associated documentation, computational equipment, and software. The medium of storage is fragile and requires regular and periodic examination (thus increased computation, software, and capital costs). Staff are required to have technical expertise beyond the training of a traditional archivist to process, describe, preserve, and disseminate these records.[34]

In addition to costs incurred in normal archival accessioning, the archivist must recognize the future cost of the machine readable record to the scholar. Those considerations include: Will the user have access to support personnel and resources necessary to access and retrieve the data? Will resources be available to process the data because they have not been properly prepared? Will resources be available to document the data? Will library and programming (consulting and software

development), and updating capabilities be available to assist in understanding and analyzing the records? Any appraisal policy for machine readable records must weigh the archives' and users' costs of processing, preservation, storage, dissemination, and available resources and expertise against the future usefulness of the records.

Administrative rules and statutes have begun to restrict access to valuable information resources because these records contain identifiable and confidential data. The restrictions which limit or prohibit access, or destroy the quality of machine readable records reduce the archival value of a collection.

The value of administrative statistical records depends on their accessibility and thus restrictions should be established only to protect institutional security and individual privacy. The situation demands creative solutions if the individual's personal privacy is to be balanced against society's need for knowledge. The archives' policy should be to maintain the flow of information without compromising the integrity of the information, while protecting details which "identify individuals on the basis of unique characteristics or as members of an identifiable group."[35] Various administrative strategies and statistical procedures can be applied to reduce confidentiality-related problems, so that considerations of privacy and confidentiality are "independent of research values" and not the "primary justification for either the destruction or retention of records"[36] and so that there are guarantees to the security and controlled release and use of this information.[37]

CONCLUDING REMARKS

This article presents an introduction to some of the problems and challenges archivists will face with records in machine readable form. It offers a guide to understanding what criteria should be applied to judge the value of these records. These criteria do not differ markedly from those applied to other records. What is different however are the techniques and strategies for appraising these records.

Lest anyone believe that these remarks represent a definitive statement on appraisal guidelines for machine readable records, it is worth repeating Schellenberg's wise remarks on the subject of appraisal: "These standards serve as guidelines to steer the unwary through the treacherous shoals of appraisal work. They are often little more than general principles. They can never be very precise. They should always

be applied with judgment and good sense."[38] In the Forward to the Schellenberg book, *Modern Archives*, H. L. White says that "archival establishments are in no sense cemetaries of old and forgotten records . . . ."[39] Nowhere is this more true than with machine readable records which record the processes and policies of government. Current machine readable records present an unprecedented opportunity and an exciting challenge for applying modern techniques and technology to the preservation of records for future use.

FOOTNOTES

1. Guido Martinotti, "Data Processing, Government, and the Public: Reflections on the Italian Case," *International Social Science Journal* 30 (1978) 147.
2. The design of the Hollerith punch card created a medium on which to document large amounts of information. With the computer, vastly higher data processing speeds and higher computing rates could be attained and therefore few limits applied to the amount of information which could be handled. With the development of the survey method social science had an alternative to census and other government records as valid sources of quantitative data on human behavior. John Lansing and James Morgan (*Economic Survey Methods*. An Arbor, Mich.: Survey Research Center for the Institute for Social Research, The University of Michigan, 1971, p.1) note that although secondary analysis has had a long tradition, it has not been until recent decades that "survey research has become a scientific tool, . . . able to produce quantified, reproducible information . . . used to test hypotheses or to provide unbiased measurement of quantities or relationships."

   But it has been the computer technology and software developed for file management and statistical analysis that have permitted the efficient and rapid analysis of data. For an extended description of these technological and intellectual advances, see Y. Lucci and S. Rokkan, *A Library Center for Survey Research Data* (New York: Columbia University School of Library Science, 1957). Stein Rokkan in "Data Services in Western Europe: Reflections on Variations in the Conditions of Academic Institution-Building," *American Behavioral Scientist* 19 (1976) 443-454, "National Primary Socioeconomic Data Structures, III: Norway," *International Social Science Journal* 30 (1978) 621-652, and Carolyn L. Geda, "Social Science Data Archives,"*American Archivist* 42 (April, 1979) 158-166.
3. Joseph W. Duncan, "The Demand for Regional and Local-Area Statistics: Issues Concerning the National Response," *Statistical Reporter* 78 (January, 1978) 97-100. Examples of programs requiring the collection and reporting of statistical information on particular populations include the State and Local Fiscal Assistance Act of 1978, part of the Crime Control and Safe Street Act of 1968 (as amended in 1971 and 1973), and the Education Amendments (1974) to the Elementary and Secondary Education Act of 1965. These laws in one way or another delegate responsibilities for data collection and analysis to state and local units of government.

   A number of rich data sources have been created as a result of federally mandated policies. These include large scale data gathering for research and policy analysis by the Office of Economic Opportunity (OEO) for the *Surveys of Economic Opportunity, 1966 and 1977* [machine readable data file], to examine the characteristics of the poor in the United States and the impact of OEO's action programs on those poor people, and, the U.S. Department of Labor's funded surveys, *The National Longitudinal Surveys of Labor Market Experience, 1966-* [machine readable data file], analyze sources of variation in labor market behavior for four age cohorts.
4. The Wisconsin Department of Health and Social Services illustrates the transition toward automation. In increasing numbers since 1977, county social service agencies have had terminals which provide on-line access to a central computer. Application information is in machine readable form instead of hard copy, and the data are updated to reflect changes in a client's situation.
5. David Nasatir, *Data Archives for the Social Sciences: Purposes, Operations and Problems* (Paris: Librarie de L'UNESCO, 1973), Chapter One.
6. The U.S. Bureau of the Census collects data on population characteristics for apportionment, production of goods and services, employment, distribution of

income, homeownership, and frequency of moving. These and other descriptive statistics are useful not only for measuring inequality or for looking at differences between parts of the population, they are also used in analyzing the impact of taxes or other governmental policies on different groups in society. When these data are repeatedly collected over time, they produce information on trends; provide excellent descriptions of the social and economic composition of the nation; and can be used to develop explanations of social, political, and economic processes.

7. Herbert Hyman argues that secondary analysis benefits science in many ways, "all stemming from one fundamental feature of the method. It expands the types and number of observations to cover more adequately a wider array of social conditions, measurement procedures, and variables than can usually be studied by primary surveys. Thus it produces a more comprehensive and definitive empirical study of the problems the investigator has formulated." See Herbert H. Hyman, *Secondary Analysis of Sample Surveys: Principles, Procedures and Potentialities* (New York: John Wiley and Sons, Inc., 1972), p. 11.
8. A good example is the following: In the early 1970 s, David Featherman and Robert Hauser's "Design for a Replicate Study of Social Mobility in the United States," (in Kenneth C. Land and Seymour Spilerman, eds., *Social Indicator Models* (New York: Russell Sage Foundation, 1975), pp. 219-251) replicated and extended Peter Blau and Otis Dudley Duncan's *Occupational Changes in a Generation* [machine readable data file], a survey of the extent and sources of social mobility in the United States. As Featherman and Hauser note, "the . . . value of our research lies in its potential contributions—substantive and methodological—to the construction of a time-series of indicators (both descriptive and analytic) of the distribution of social and economic opportunity in the United States . . . (1) in the assessment of widespread beliefs about equality of opportunity and factors affecting it . . . (2) in locating and defining the problems of specific population subgroups . . . (3) in providing an overall model of the process of social and economic achievement which can serve as a frame of reference for discussions about specific aspects of that process . . . (4) in providing a set of current trend estimates on major features of the process of social achievement . . . and (5) in improving the measurement of processes of social and economic achievement."
9. For example, the State Historical Society of Wisconsin has preserved corporate income tax returns, but until the availability of the computerized record, there was no possibility of maintaining the individual tax returns of all Wisconsin citizens.
10. Not only can the computer protect individual identities, by selectively deleting information, but it can also summarize information which in its raw state could compromise personal privacy. Special techniques can also determine the probability of inadvertent disclosure of individually identifiable information. Whereas archivists traditionally have been wary of accepting confidential information, they are now more inclined to accession this type of material, since they have a technology that assists them in their role of records custodians.
11. One example of data linkage is the *Wisconsin Assets and Income Project.* Here tax returns for a sample of Wisconsin taxpayers, containing summary data available from the archives of the State of Wisconsin for the period of 1946 to 1964, were linked with social security benefit payment and earnings records, supplemental data on ages and deaths from the Wisconsin Motor Vehicle Department (driver's license records) and from the Wisconsin Board of Health (death records), and personal interviews and reports of certain members of the interview sample to obtain reports on their portfolio holdings and recent portfolio changes for a household assets diary. See Martin H. David, William Gates, and Roger Miller, *Linkage and Retrieval of Microeconomic Data* (Lexington, Mass.: Lexington Books, 1974).

12. In the creation of machine readable records, raw data files which are input in the process of record development may be as valuable as the master files. Fishbein notes that "most of the literature about statistical records recommends, with occasional exceptions, the destruction of raw data sources and intermediate tabulations." He also notes that there are no "generally applicable criteria" for the preservation of records of science and technology, but because the interest in the history of science and the effect of technological change on society has increased, archivists must protect the research sources for these areas of history. (Meyer H. Fishbein, "A Viewpoint on Appraisal of National Records," *The American Archivist* 33 (1970) 147.)

    Two other examples of categories of records which might and have not been considered of archival value are described by Fishbein and Mitchell. Fishbein cites a study by Robert W. Fogel on the role of railroads in the American economy at the turn of the century. Fogel used the series of Interstate Commerce Commission railroad tariffs to determine freight costs. Fishbein notes that "by our earlier standards such records were, at best, of dubious archival interest" (p. 182). Mitchell cites the destruction of prison records of the Ohio Penitentiary. Their value for sociological research, "to investigate the reasons for self-inflicted wounds among inmates, the reasons for recidivism, and the effects of the work-release program," could have justified their preservation, but they were destroyed on the basis that the information had no value for historical research. See Thorton W. Mitchell, New Viewpoints on Establishing Permanent Values of State Archives," *The American Archivist* 33 (1970) 166.

    Charles Dollar, Director of the Machine Readable Archives Division of the National Archives and Records Service, has stated that over the last 25 years the National Archives has preserved only 3 percent of all federal print or textual records. Were the Machine Readable Archives Division to observe this "rule of thumb" for computer readable records, the National Archives "would face the prospect of preserving between 240,000 and 270,000 reels of computer tape as of 1977." (Charles M. Dollar, "Problems and Procedures for Preservation and Dissemination of Computer Readable Process-Produced Data," paper prepared for the QUANTUM-SSHA Conference on Quantification and Methods in Social Science Research: Possibilities and Problems with the Use of Historical and Process-Produced Data, Cologne, West Germany, August 9-12, 1977, p. 9.) Fishbein demonstrates optimism with his statement that the computer has made it possible to deal with voluminous records and suggests that archivists must reject "the idea that the volume of information is a deterrent to research." (Fishbein, p. 183). Nevertheless, recent public records *are* of a voluminous nature and are expected to increase exponentially in the coming decades. As a result, decisions about which records to retain become far more difficult than they were in the past.

    Whereas once the paper record constrained the amount of information that could be retained, the computer imposes few constraints on the number of pieces of information which can be stored. For example, the *National Longitudinal Surveys of Labor Market Experience* [machine readable data file] contain thousands of items of information for each of four age groups (cohorts) surveyed. Each cohort has been interviewed on an almost yearly basis since 1966-67. The size of each logical record (information for one individual) now approaches 20,000 characters of information, some several thousand separate items, and increases by about 2-3,000 characters yearly. Documentation describing these items now exceeds 2,000 pages per cohort. To understand the informational items and the relationships to each other and across years and to retrieve selected pieces of information with appropriate computational equipment and software require resources available only

to highly sophisticated and well trained analysts working in experienced computational settings.

13. Without proper documentation data items from several sources may appear to be from the same initial documents. It is useful when linking data from different sources to maintain a certain degree of redundancy of data items, since this facilitates validation of data origin. An analyst might link certain items from death records, other items from health records; several variables containing the same information would be retained as a check on the accuracy of the record linkage. Without proper documentation, another analyst (or even the individual who linked the data files) might not be able to identify and rectify inconsistencies or original errors. Many data files can only be used with special purpose statistical and data management software which imbeds the records in a specially designed internal structure. This structure constrains the use of these records to a computing environment which provides the same software and data base management systems. There are greater costs associated with machine readable records than with other records. Expenditures include more specialized archives staff, a greater capital investment in materials, equipment, and support services.
14. The National Archives Machine Readable Archives Division, for example, provides data dissemination services, tailored to meet individual user needs, rather than merely providing a full set of data; in this way, the Division takes on a new function, typically reserved for an information service.
15. Computer technology has progressed with extraordinary rapidity, to the extent that various computers used less than 15 years ago are no longer available. This means that data files designed to be used with particular computers may not be usable on contemporary machines. In fact, this is a problem faced by the U.S. Bureau of Census, where data files created in the early 1960s, cannot be transported to any other installation because only the Bureau has the equipment to read these data.
16. The creation of machine readable data files is not regarded as a legitimate administrative or scholarly activity because the file only serves the end of providing information or providing a tool for publication. For example, social scientists are not rewarded for producing clean data. There are few professional incentives for making data available to another analyst. Federal agencies, which provide the bulk of funding support for data collection, do not require that data be available as soon as primary analyses are completed or within a specified time period, thus providing incentives for good data preparation (which will more easily occur in an environment where it is known that the data will be evaluated at the time a report is made public). Professional journals do not require that a data file upon which an analysis is based be submitted at the same time the article is submitted for publication. The General Accounting Office is only now beginning to consider guidelines for data quality evaluation. (See Alice Robbin, "The Data Archive Perspective on Machine Readable Data for Secondary Analysis: Technical Standards for Text Documentation, Quality Data, and Improving Access to Information," paper prepared for the Alternative Designs Conference, National Institute of Education, Washington, D.C., October 30-31, 1978, pp. 21-22.)
17. Hedrick, Boruch, and Ross describe a variety of problems they encountered with inappropriate aggregation, problems which reduced the utility of evaluative data for answering specific research questions. Their examples range from data files where an individual's total achievement test scores and total inventory scores are retained, but not the individual's response to each item within a test or inventory; to RAND analyses which included data aggregated at the classroom level which led to conclusions that the Alum Rock Voucher Program had either no effect or negative effects on student achievement scores. (Terry E. Hedrick, Robert F. Boruch, and Jerry Ross, "On Ensuring Availability of Evaluative Data for Secondary Analysis," *Policy Sciences* 9 (1978) 259-280.

18. These include such items as geographic location, occupational and work history, educational institutions, electoral districts, political affiliation, industry, payroll periods for employment reports, governmental agencies, nation-states, and biographical information on political elites. The federal government, through the Office of Federal Statistical Policy and Standards sets recommended guidelines for coding standard data items. For a further discussion of this issue, see A. R. Eckler and Thomas J. Mills, "Planning and Coordination of the Federal Statistics System," *Statistical Reporter* 78 (January, 1978) 97-108.
19. The rapidly growing computer storage and retrieval technology includes the development of data base management systems which allow a user to access different files from different locations, bring these files together, select elements of these files, and create entirely new files on a temporary basis. Dollar comments on the effect of this new technology: "Problems of provenance could become insoluble, since there would be no so-called audit trail to reveal the sources of data." (Charles M. Dollar, "Appraising Machine Readable Records," paper prepared for delivery at the Annual Meeting of the Society of American Archivists, Salt Lake City, Utah, September 30-October 4, 1977, p. 11.) Steve Johnson notes that "the universe of potentially accessionable machine readable data is so large and so little subject to intellectual control by librarians, archivists, or subject specialists, that the importance of a given record, relative to other files, is difficult to determine." (Steve Johnson, "File Documentation as a Key to the Appraisal of Machine Readable Data Sets," student paper prepared for F. Gerald Ham, Madison, Wisconsin, 14 September 1977.)
20. See Alice Robbin, "Characterizing Text Documentation as a Minimal Information Management System," *SIGSOC Bulletin*, 4/5 (1974/75) 56-68; Alice Robbin, "Systemic, Structural and Policy Problems of Access and Retrieval of Numeric Machine Readable Data, Some Modest Solutions To," *Proceedings of the Association for Population/Family Planning Libraries and Information Centers-International Eleventh Annual Conference*, (1978) 125-146; and Robbin, "Data Archive Perspective."
21. An example of the fragility of this medium is worth recounting. More than a year ago, the Data and Program Library Service bought 150 magnetic tapes which were guaranteed to last 20 years. These tapes were used to write data files which are intended for long-term retention. Within a year, these tapes had deteriorated. Obviously, there was poor quality control when the tapes were manufactured, but this could also have been exacerbated by tape drive problems.

    Poor humidity and temperature conditions hasten the deterioration of magnetic tape. Also, if the tape is not turned every 2-3 months and not rewound and used every several years, it can stretch and cause data destruction. Careless handling by computer operators and poorly balanced computer tape drives can harm the magnetic tape and its contents. However, recent technological developments in new storage media, such as the "miracle chip", suggest that within two decades, institutions will be able to retain enormous amounts of data on a very small device, which will also ensure more stable storage conditions. For a good discussion of the future of storage devices for machine readable data, see Carl Hammer, "A Brief Forecast of Computation Information and Management," *The International Journal of Management Processes and Systems* 1 (1977) 3-10.
22. See *Personal Privacy in an Information Society*, The Report of the Privacy Protection Study Commission, (Washington, D.C.: U.S. Government Printing Office, July 1977) for a full discussion of the principals and requirements of the Privacy

Act of 1974 and legislative recommendations made by the Commission to protect the privacy of individuals.

23. Alice Robbin, "Ethical Standards and Data Archives," in Robert F. Boruch (ed.), *Secondary Analysis* (San Francisco: Joseey-Bass, Inc., 1978), 7-18.
24. Laura A. Guy, "Personal Information Privacy and the Archive: A Dilemma for the 1970 s, . . . and Beyond," student paper prepared for F. Gerald Ham, Madison, Wisconsin, December 1978.
25. For a description of the various state statutes, see Charles Knerr, *Confidentiality of Research and Statistical Data: A Compendium of State Legislation* (Washington, D.C.: National Criminal Justice Information and Statistical Service, Law Enforcement Assistance Administration, U.S. Department of Justice, 1978). Increasingly, the federal government is applying strict criteria to what information may be collected, what information shall be released, and who shall have access to the information. See Robert F. Boruch and Joseph S. Cecil, "Privacy Legislation, Its Character, and Its Accommodation by the Social Research," paper presented at the Alternative Designs Conference, sponsored by the National Institute of Education, Washington, D.C., October 30-31, 1978, for a wide ranging discussion of the effect of federal statutes on social research. See also *Privacy and Security of Criminal History Information: Summary of State Plans (Washington, D.C.: National Criminal Justice Information and Statistics Service, Law Enforcement Assistance Administration, U.S. Department of Justice, 1978) for a description of the development of legislation at the state level to deal with the issues of privacy and security* of criminal history information. *Privacy and Security of Criminal History Information: Compendium of State Legislation* (Washington, D.C.: National Criminal Justice Information and Statistics Service, Law Enforcement Assistance Administration, U.S. Department of Justice, January 1978) is an important monograph for all state archivists.

    While the U.S. federal policy on access to microlevel data is far less restrictive than the policies of most West European nations, the federal government and research community are taking a conservative view and are erring, on the side of increased restrictions on access. See Paul J. Miller, ed., *Proceedings of the CESSDA/IFDO International Conference on Emerging Data Preparation and the Social Sciences' Need for Access to Data* (Cologne, West Germany: Zentralarchiv fur Empirische Sozialforschung, 1979).
26. Robert J. Anderson's student paper "Public Welfare Case Records: Archival Perceptions and Archival Practices," prepared for F. Gerald Ham, Madison, Wisconsin, 18 December 1978, describes the refusal of many state archivists to deal with public welfare case records because of their confidentiality problem. Confidentiality is deemed the most significant barrier to accessioning by archivists who participated in his survey.
27. Maynard J. Brichford, *Archives and Manuscripts: Appraisal and Accessioning* (Chicago: Society of American Archivists, 1977), p. 8.
28. The *Graduated Work Incentive Experiment in New Jersey, 1968-1972* [machine readable data file] is an excellent example of a machine readable record series worthy of retention because it documents the U.S. government's decision to fund a major research project with important policy implications. Data were collected with instruments designed to yield reproducible information which could be tested against data collection by other agencies (U.S. Bureau of the Census, *Current Population Surveys* [machine readable data files]; and University of Michigan Institute for Social Research, *Panel Study of Income Dynamics* [machine readable data file]). The records of the project, correspondence, original paper records (now microfilmed) provide a history of the personnel, relationships between data subjects, project staff, and government, and outcomes of federal agency program support for policy purposes. See D. Kershaw and J. Fair, *The New Jersey Income-Maintenance Experiment, Volume I: Operations, Surveys and Administration* (New York: Academic Press, 1976) for a very readable discussion of this project.

29. For example, there are a number of important one-time surveys which have been made over the last 20 years, which although not replicated, provide us with important insights into the health of the nation. A number of surveys within the *Current Population Surveys* [machine readable data file] series contain questions which were asked at only one point in time. This does not mean that these questions will not be asked again in the future, as we see from an Institute for Social Research (University of Michigan, Ann Arbor) study made first in 1957 and just recently replicated *(How Americans View Their Mental Health* [machine readable data file]).
30. Robbin, "Characterizing Text Documentation." See also Paul T. Zeissert's position paper "Some Views on Good Documentation," in National Bureau of Economic Research Conference, *Standards for Documentation of Large Social Science Statistical Data Bases* (New York, April 18-20, 1974) 2.
31. Brichford, *Archives and Manuscripts*, p. 8.
32. Harold Watts, "Microdata: Lessons from the SEO and the Graduated Work Incentive Experiment," *Annuals of Economic and Social Measurement* (1972) 184.
33. Dollar, "Appraising Machine Readable Records," 12.
34. Ben DeWhitt, "Archival Uses of Computers in the United States and Canada," *American Archivist* 42 (April, 1979) 152-157.
35. *Personal Privacy in an Information Society*, p. 586.
36. Brichford, *Archives and Manuscripts*, p. 9.
37. A variety of technical procedures have been devised to prevent disclosure, but not inhibit or constrain secondary analysis. (See Robert F. Boruch, "Educational Research and the Confidentiality of Data: A Case Study," *Sociology of Education* 44 (1971) 59-95; Boruch, "Strategies for Eliciting and Merging Confidential Social Research Data," in P. Nejelski, ed., *Research in Conflict with Law and Ethics* (Cambridge, Mass.: Ballinger, 1976); and Donald T. Campbell *et al.*, "Confidentiality-Preserving Modes of Access to File and Interfile Exchange for Useful Statistical Analysis," *Evaluation Quarterly, A Journal of Applied Social Research* 1 (May, 1977) 269-300.) The recommendations of the Privacy Protection Study Commission on the release of and access to microlevel data for statistical research represent a sensible approach to the problems of data disclosure; but few members of federal or state agencies appear familiar with the Commission's recommendations. *Ad hoc* decisions on release and access by members of federal agencies will continue to be a problem until enough expertise exists to make rational decisions on access to microlevel data (Robbin, "The Data Archive Perspective," October 1978, p. 23).
38. T. R. Schellenberg, *Modern Archives: Principles and Techniques* (Chicago: The University of Chicago Press, 1956), p. 133.
39. *I bid.*, p. viii.