COMPUTERS, PRIVACY, AND RESEARCH ACCESS TO CONFIDENTIAL INFORMATION

MARGARET L. HEDSTROM

The advent of computer technology and the heightened public concern with personal privacy in recent decades coincide with several other developments that make the issue of computers and privacy an important yet complex one for archivists to address. Changes in government policies, especially the expansion of state supported benefits, mandated the collection of more personal information and expanded the quantity of documentation available on a wider range of citizens. At the same time, the focus of historical research has broadened to include an interest in the composition, attitudes and behavior of non-elites (precisely the same population on which government agencies compile and maintain extensive documentation). Other academic fields, such as sociology and public policy, rely extensively on theories of behavioralism and on quantitative methodology which necessitate access to data on the characteristics and behavior of the general population.[1]

Despite an increased research interest in the social, economic, and demographic characteristics of the population and the greatly improved practices among archives for identifying and making such information available, research efforts are often hindered by federal and state legislation which restricts access to a substantial portion of the historical record. Frequently, such legislation has not taken the interests of researchers into account, due in part to the lack of influence by the scholarly community over the shaping of privacy legislation. While archivists have expressed concern about privacy and the legislation passed to protect it, they have been reluctant to become involved in determining how access to

personal information will be regulated. They have viewed their role as ambiguous and felt trapped in a compromised position, recognizing the need both to protect individual rights to privacy and to make important records available to researchers. By viewing their role as ambiguous, archivists have been reluctant to intervene on the side of either the individual citizen, who seeks increased protection of personal privacy, or the researcher who desires access to restricted information.

This article attempts to define a conceptual approach to the issue of privacy which will transcend the ambiguity of the archivists' role and provide the basis for a more active role in shaping privacy legislation and making restricted information available for research in a form that does not threaten anyone's privacy. Following a review of the impact of computer technology on record keeping, and a presentation of the conceptual approach for dealing with the privacy issue, several practical methods for handling confidential information in machine readable form will be presented.

At the outset, some basic terminology must be defined. *Machine readable records*, also referred to as computerized or automated records, are records created with the use of a computer which require access to a computer to transform the information into a human readable form. Microforms, audio recordings, and motion pictures are not included in this definition. *Data files* are the machine readable equivalents of records series or files of paper documents. *Micro-level data* refers to data that have not been aggregated or summarized, where each observation pertains to one case, transaction, or event. *Data subjects* are individuals and organizations on whom records are maintained. Finally, while this discussion focuses on personal information about individuals, most of the comments apply equally to other types of confidential information, such as records of the financial activities of businesses.

The application of computer technology to record keeping has had a fundamental impact on the maintenance, storage, accessibility, and retrievability of personal information. The increased capacity of government agencies, private companies, and research facilities to maintain and manage personal information has raised concerns about the issues of access to information and

privacy. While the public shares an awareness that computerized records pose new threats to privacy,[2] the nature and extent of the problem are not well understood. Public apprehension, which has focused on the notion of omniscient computerized data banks that could assemble detailed dossiers on the characteristics and activities of large segments of the population, may be misdirected due to legal, political, fiscal, and technical constraints on the centralization of computerized information.[3] Meanwhile, many of the implications of computerized records for privacy go unnoticed by laypersons who lack the interest or knowledge to evaluate the effect of new technology. The public appears most concerned with the technical capabilities of computers to invade privacy. Other fundamental issues receive less attention: the lack of control over the maintenance and dissemination of computerized records, the unequal distribution of power between government agencies and businesses and their constituencies and clients to access and use the information, and the barriers to external evaluation of the quality and use of the information.

On the surface it is apparent that the application of computer technology can facilitate invasions of personal privacy by making information more widely accessible at lower cost and greater speed. The ability to transmit information over telephone lines and to transfer data from one computing facility to another allows branches, offices, and bureaus of one or several organizations to share information more easily and often instantaneously. Through the use of telecommunications, information in machine readable form can be disseminated to decentralized locations, and decentralized banks of data can be transferred to a central location. Retrieval of information is simplified as well, because the computer is able to scan records and select in seconds those that meet certain criteria. Computers can also be used to link records on one individual or entity from disparate sources by matching elements of information common to all sets of records, such as name, address, and social security number.

It is frequently argued that computers not only have improved access to personal information, but also have increased the amount of personal information collected. However, the extent to which computer technology has increased the amount and types of personal information collected by public and private agencies is

debatable, and a direct link between computer applications and the increased collection and maintenance of personal information is difficult to demonstrate. The U.S. Privacy Protection Study Commission concluded that "the broad availability and low cost of computer technology provide both the *impetus* and the *means* to perform new record keeping functions."[4] Other studies of computerized records in the United States,[5] Canada,[6] and Wisconsin[7] suggest that computers have altered the form in which information resides more than they have altered the content of the records. Further evidence indicates that the most sensitive types of personal information, such as narrative evaluations of students, employees, or clients, are not likely to be found in machine readable form.[8]

The potential threats to personal privacy posed by the increased access and retrieval capabilities of automated record systems can be distilled into three legitimate areas of concern: 1) the use of computers to scan hundreds of thousands of records in seconds and locate the records on any individual; 2) the ability to selectively retrieve records of all individuals who share a set of attributes; and 3) the ability to construct detailed dossiers on individuals by combining records from a variety of sources. While all of these activities are possible with manual records, the introduction of computer technology makes them much more feasible from both time and cost perspectives. However, it is easier to discuss the theoretical possibilities for potential invasions of privacy than to evaluate actual effects. Since empirical evidence is pitifully lacking, no one really knows how much personal information is maintained in machine readable form, who has access to it, or how it is used. Furthermore, many discussions of the threat of computers to personal privacy fail to recognize that the threat is not inherent in computer technology. Rather, it stems from the misuse of the technology and from inadequate or unenforceable restraints on records linkage, centralization of information, and dossier building.

In contrast to the potential invasions of privacy made possible by technological advances, both the lack of knowledge about personal information in automated record systems and the absence of systematic control over its dissemination and use by public and private agencies pose very real threats to personal privacy which

seldom are acknowledged. While computer technology has greatly increased user agencies' capabilities to access and retrieve information, it has simultaneously erected barriers to access to public records by individuals outside these agencies. For example, following a survey of machine readable records in major data producing agencies in Wisconsin, the Wisconsin Survey of Machine Readable Public Records found that systematic control over machine readable records is completely lacking. Agencies have not compiled inventories of machine readable files, and extensive research is required to identify what confidential information exists in machine readable form and to ascertain who has access to it.

Access to public records and an informed public evaluation of the impact of the computerization of records on personal privacy are hindered further by the visual obscurity of machine readable records and by the veil of mystery that surrounds the technical operations of computer centers. Even if one has the technical skills necessary to use machine readable data files, essential technical information required to gain access to the records is often available only from systems analysts and computer programmers who work intimately with the particular automated systems involved. Finally, policies and practices regarding access to confidential information often are arbitrary and vary among agencies. In some cases, access is denied to information that is neither statutorily nor administratively defined as confidential, merely because the records are in machine readable form. In other cases, hard copy source documents may be defined explicitly as confidential by statute, yet access to the same information in machine readable form may not be restricted.

The lack of control over machine readable records presents a formidable barrier to individuals who may desire access to records maintained about them in order to assess the accuracy and timeliness of the information. The difficulty of locating and identifying data files that might be applicable to a research project also makes research by parties outside of the agencies difficult, if not impossible. Finally, the arbitrary application of restrictions on access to public records by government agencies, under often spurious arguments of protection of personal privacy, has given the agencies an undue amount of discretion in determining who

will and who will not be allowed to examine public records.[9] Such actions have begun to erode the effect of the Freedom of Information Act and of state open records laws, as well as to create a climate in which government secrecy can flourish.

Given the lack of control over machine readable records, the obstacles to monitoring their maintenance and use, and the power of government agencies to affect individuals through their use of personal information, machine readable records can pose the greatest threat to privacy when they are maintained in an administrative setting. By contrast, when confidential records are placed in an archives and made available for research, issues of a more practical and procedural nature must be confronted. Computers can be used to make restricted information available to researchers in such a way that individual identities are masked or deleted. In addition, the risk of misuse of personal information by the research community is exceedingly small. As Robert Boruch and Joseph Cecil point out, "we know of no significant instance in which a legitimate researcher, when given access to proprietary records, has exploited the records at a cost to the individual on whom the record is kept."[10]

The legal and ethical questions of providing access to confidential information would be simplified if a clear distinction were made between administrative and research uses of the records. It is important for archivists and researchers to develop the concept of such a distinction for both manual and automated records, because much of the recent privacy legislation has been aimed toward regulating administrative uses of personal information in all formats, with little or no consideration for its research applications. The lack of provisions allowing access to confidential information for research purposes has resulted in regulations that make some legitimate research projects impossible. It has also led to the costly duplication of effort by researchers who independently collect data that already exist in administrative records, and to frustrating experiences for those who attempt to gain access to restricted records. A corollary to this problem is that as increasing amounts of survey research data are collected on human subjects, it is crucial that personal information, offered to researchers with assurances of confidentiality, does not filter back into the administrative arena

where it could be used as the basis for a decision affecting the individual. Finally, as empirical evidence plays an increasingly important role in the design and implementation of social and economic policies, it is important that third parties gain access to the data in order to evaluate the validity and accuracy of the evidence and assumptions upon which such policies are based.

The Privacy Protection Study Commission urged that a distinction, or "functional separation," be made between administrative and research records. The Commission defined the distinction in an organizational sense, as a well defined separation of the research and statistical components of an agency from the decision- and policy-making components.[11] While such a distinction is crucial, it should not be based on the organizational setting in which the records are created, but on the way in which they are used. Many small agencies do not have separate research divisions, and many large agencies are increasing their use of internally generated administrative records for research, due to fiscal constraints on the collection of data and to reluctance by the public to respond to survey questionnaires.

A distinction can be made between administrative and research records, based on how use of the information in the records affects the individuals on whom the records are maintained. Generally, administrative records containing personal information are used to make policy decisions on a case-by-case basis which have a direct bearing on the individual whose records are involved. Personal information is used administratively to determine if an individual meets the qualifications for a social service benefit, a driver's license, or a mortgage; if one has paid all taxes owed the state; or if one's education and training qualifies him for a license to practice a particular trade. Research or statistical use of personal information, on the other hand, is not intended to affect directly any individual whose records are included in the research project's database. Rather, the researcher is interested in reporting anonymous information based on aggregate statistics which describe a particular group or allow for generalizable inferences from a sample to a larger population.

In order to resolve many of the legal and ethical questions associated with confidential information, it is essential to develop different regulations for its administrative and research uses. Such

regulations would allow access to restricted administrative records for research use, as long as the identities of individuals are not publicly divulged. However, administrative use of confidential information that was collected independently in pursuit of a research project would be prohibited. Until such time as this distinction is clearly developed and implemented, a number of procedural measures can be employed to make restricted records available for research.

Machine readable records offer the opportunity to circumvent some of the restrictions on access to confidential information, and they provide archivists with a range of choices in determining the conditions under which confidential information will be made available. Archivists can place administrative restrictions on the user, use technical procedures to create disclosure-free public use versions of the data files, or use a combination of the two approaches. Each data file must be evaluated separately and the most desirable approach selected with careful consideration of the characteristics of data in the file and the extent of the restrictions that will be required. In addition, the archivist must consider the potential risk of disclosure, the potential harm to a data subject if disclosure occurred inadvertently, the cost to the archives of any procedure employed, and the effect of any alterations of the data on its potential research uses.

The most practical and widely used procedure for providing access to confidential information, in both manual and machine readable form, is to develop contractual or administrative requirements for the user before access is granted. Such requirements can range from a simple written statement by the user, in which he or she agrees not to disclose the identities of any individuals, to elaborate contractual agreements between the user and the archives or originating agency. A contractual agreement might spell out the conditions under which the archives will allow access; the right of the archives or the agency to review any written or publicly presented papers, reports, or publications; and the willingness of the researcher to accept well-defined penalties for violation of the agreement. Such an agreement might also contain a clause absolving the archives of responsibility for unauthorized disclosure if it has made a good faith effort to enforce the agreement. To my knowledge, the legality of such agreements has

not been tested in the courts, and it may be burdensome for the archives to administer these agreements due to the time required to evaluate the researcher's findings prior to dissemination. However, this approach has the advantage of placing primary responsibility for the proper handling of confidential information on the researcher, where it belongs, and not on historical agencies whose mission is to provide access to information for research. This safeguard can be coupled with some of the technical procedures described below to further reduce the legal liability of the archives in the rare event of misuse of confidential information by researchers.

A number of technical measures have been developed to provide access to confidential information in machine readable form by masking the identity of individual data subjects or by slightly altering the micro-level data to prevent disclosure of individually identifiable information. These procedures fall under the general rubric of creating disclosure-free or public use versions of data files. Following these procedures, the archives would accession and retain a complete version of the data file with all personal identifiers intact, but it would alter the data that it releases to researchers in such a way that the identity of individuals would be protected.

Since the distinction between research and administrative uses of records rests on the premise that researchers have no intrinsic interest in the personal identity of data subjects, the easiest and most obvious solution is to create public use files with all personal identifiers removed. Computer technology makes it possible to delete all names, addresses, social security numbers and other forms of personal identifiers from a data file without altering other information in the file. However, removing personal identifiers may not always be advisable or adequate. Archivists must consider two additional problems: that personal identifiers are essential for some types of research; and that in some cases, even the removal of personal identifiers might not adequately prevent identification of some data subjects.

In most studies, personal identifiers serve merely as an accounting device and they can be replaced with a set of case numbers unique to the research project, without reducing the value of the records. However, personal identifiers are necessary in

some research projects to allow for evaluation or reanalysis of the data, auditing of the results by a third party, recontacting of the original data subjects for follow-up studies, or linkage to other records on the same individuals. The availability of personal identifiers is essential for longitudinal studies where characteristics of a sample are examined at several points in time. Longitudinal studies are used to investigate issues such as the effects of different long-term employment patterns, or innovative educational programs on learning patterns.

Personal identifiers are essential to research projects that require linkage of two or more data sets on the same individuals. For example, to assess the impact of exposure to a toxic chemical on health, it would be crucial to link records of the incidents of exposure with subsequent medical records. Likewise, a researcher might want to augment the information available in a restricted data file with records that are retrievable by name from a different source. Personal identifiers are required if a second group of researchers wants to evaluate the quality of a set of original data or to reanalyze the data with the addition of more variables. While personal identifiers are not essential to the vast majority of research projects, it is impossible for archivists to anticipate all of the potential applications for a data set. Consequently, personal identifiers should never be permanently removed from a file, since such actions could jeopardize future uses of the data.

A compromise solution to this problem is to create a "link file."[12] According to this procedure, the computer is instructed to delete all personal identifiers and to assign a unique identification number to each case in a public use version of the file. The computer creates a separate file that contains the personal identifiers and the unique identification number for each case. Using this procedure, the researcher can return to the archives for assistance in linking data from one file with additional records, without ever ascertaining the identity of individuals in the restricted data file.

In addition to research needs for personal identifiers, the problem of deductive or statistical disclosure occurs when deletion of personal identifiers is not sufficient to prevent identification of data subjects. In some data sets, cases are so unique that they can be identified on the basis of their attributes alone. For example, a data

set on the characteristics of the population of a small town, including occupations, would permit identification of the doctor, the dentist, the hairdresser, the butcher, etc. Statistical disclosure can also occur when statistical tables are built in such a way that individual cases can be identified through manipulation of values in the tables. The likelihood of statistical disclosure increases when other sources of information are available on the group under study. For example, if a city directory listing the occupations of its residents were available in conjunction with a data set containing occupational information, some individuals could be identified on the basis of their occupation.

The risk of statistical disclosure is exceedingly small, yet it may require additional measures beyond the removal of personal identifiers to prevent or minimize the risk. Some of the methods employed by the U.S. Bureau of the Census for releasing demographic data include deleting geographic information for any area with a population of less than 250,000; deleting information in instances where less than five cases share the same attributes; and deleting outlying values, such as incomes above a certain level.[13] The release of samples, rather than data on an entire population, also reduces the risk of statistical disclosure by making it impossible to determine which cases are included in the sample.

Several methods can be employed to alter the values of data in a file in order to mask the identity of individual cases.[14] Some data files contain only one variable which is considered sensitive, while the remainder of the information is innocuous, yet valuable for research. The computer can be instructed to omit that particular item without disturbing the remainder of the information in the file. Computers can also be used to adjust the exactness of the data, by rounding the values for each case and fitting them into a range of values. For example, exact income figures can be rounded and released as income categories. Statisticians have developed sophisticated methods of introducing "noise" or random errors into data files in such a way that the exact values for some cases are altered, but the values of aggregate statistics are not affected. A strategy called microaggregation can also be employed, by which a few cases are grouped into small clusters and aggregate or average values for the groups are released. Any of these methods could reduce the utility of the data for some types of statistical analysis.

They should be applied only after careful consideration of the characteristics of each data set and after consultation with the researcher.

A number of practical methods have been described that can be used to accommodate restrictions on access to machine readable records. While computer technology can be used to provide access to confidential information without disclosing the identity of individuals, it should not be interpreted as a panacea. Many types of confidential information are not, and will not be, available in machine readable form; and evidence suggests that the most sensitive information is not computerized. The lack of systematic control over machine readable records, both by the originating agencies and by archives, makes access to these records difficult for many researchers. In addition, the procedures used to create public use versions of restricted access data files often are expensive to design and implement; they require new technical skills that are not common among archivists; and they may significantly reduce the utility of a data file for research.

The legal and ethical questions regarding access to confidential information are essentially the same for both manual and automated systems. Archivists must address these questions and resolve them in such a way that restrictions on access to confidential information used for research are applied only when disclosure could potentially harm a data subject and only to the extent necessary to minimize the potential of harm. The development of different procedures for administrative and research uses of confidential information could go a long way toward reducing the tension between respecting and protecting rights to personal privacy and making important public and private records available for research. In the interim, the techniques discussed here, and many others, can be used to circumvent some restrictions. Because techniques developed to provide access to confidential information in machine readable form are expensive and often unnecessary due to the integrity of researchers in handling proprietary information responsibly, restrictions should be reduced as much as legal and ethical considerations permit. Thus, while a number of practical strategies have been developed to circumvent restrictions and to make machine readable records available in some altered form, the

judicious application of these techniques is dependent upon further clarification of administrative procedures and basic ethical and legal issues.

FOOTNOTES

1. The author would like to thank Max Evans, F. Gerald Ham, Christine B. Harrington, and Alice Robbin for their review and comments on this article. Its development was also supported in part by a grant from the National Historical Publications and Records Commission to the State Historical Society of Wisconsin and the University of Wisconsin-Madison.

2. The evidence from survey research shows a general public concern about the impact of computers on privacy. For example, a recent poll by a Stanford University researcher revealed that nearly two-thirds of the respondents felt that computers have affected personal privacy, and more than three-fourths of that group thought that the effects were negative. "Computer Security," *Mosiac*, 9, #4, (1978), p. 8. A comprehensive discussion of public attitudes toward privacy can be found in Harris, Louis and Associates, *The Dimensions of Privacy: A National Opinion Research Survey of Attitudes about Privacy* (Stevens Point, Wisconsin: Sentry Insurance, 1979)

3. Alan Westin et al., *Databanks in a Free Society* (New York: Quadrangle Books, 1972), pp. 238 ff.

4. U.S. Privacy Protection Study Commission, *Technology and Privacy* (Washington, D.C.: Government Printing Office, 1977), p. 1.

5. Westin et al., *Databanks in a Free Society.*

6. Canada, Department of Communications and Department of Justice, *Privacy and Computers* (Ottawa: Information Canada, 1972).

7. This survey was part of a Pilot Program to Accession Machine Readable Public Records of Wisconsin State Agencies: F. Gerald Ham and Martin David, co-principal investigators; Max Evans and Alice Robbin, project co-directors; NHPRC Grant #80-8 to the State Historical Society of Wisconsin and the University of Wisconsin-Madison. The survey found no significant increase in the amount of personal information collected and maintained by state agencies as a result of automation of their records systems. In fact, most machine readable records contain extracts of much more elaborate documentation maintained in paper files.

8. This generalization may become less accurate in the future as further reductions of storage costs and technological advances (especially in the area of word processing technology) make textual information more subject to computerization.
9. For example, the proposed Intelligence Reform Act of 1980 would exempt most files of the Central Intelligence Agency from the Freedom of Information Act. See: George Lardner, Jr., "Moynihan Unleashes the C.I.A.," *The Nation*, Vol. 230, No. 6 (Feb. 16, 1980): 161, 176–78.
10. Robert Boruch and Joseph Cecil, "On Solutions to Some Privacy Problems Engendered by Federal Regulation and Social Custom," *Federal Regulations Ethical Issues and Social Research*, Murray L. Wax and Joan Cassell, eds. (American Association for the Advancement of Science, 1979), p. 179.
11. Privacy Protection Study Commission, *Personal Privacy in an Information Society* (Washington, D.C.: Government Printing Office, 1977), p. 572.
12. There are several elaborate variations on the link file strategy. The link file strategy is explained thoroughly in Robert Boruch, "Maintaining Confidentiality of Data in Educational Research: A Systematic Analysis," *American Psychologist* 26 (1971): 424–426; Robert Boruch and Joe S. Cecil, *Assuring the Confidentiality of Social Research Data* (University of Pennsylvania Press, 1979), pp. 108–112.
13. Paul T. Zeisset, "Census Bureau Confidentiality Practices and Their Implications for Archivists," unpublished paper presented at the Conference on Archival Management of Machine Readable Records, Ann Arbor, Michigan, Feb. 7–10, 1979.
14. These methods and others are discussed in U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, *Report on Statistical Disclosure and Disclosure-Avoidance Techniques* (Washington, D.C.: Government Printing Office, 1978).