Protecting Confidentiality in Archival Data Resources

by Christopher S. Dunn & Erik W. Austin *

Data sharing is a disputed norm in scientific affairs (Fienberg et al. 1985; Weil and Hollander 1990; Fienberg 1994; Mishkin 1995). On the one hand, principal investigators argue that they and their research teams are the most competent analysts of originally collected data and best able to safeguard the data against release of confidential information. They know the details and nuances of the sampling procedures, instrumentation, data reduction, and missing data. They have an investment in the original research that should be repaid by first rights of publication. They also argue that for certain kinds of complex studies, for example, observational research, organizational research, longitudinal research, clinical research, and research involving geo-coded data or administrative records linked to survey data, they are the only or principal safeguard against violations of confidentiality of the data. On the other hand, researchers argue that publicly supported data collections should be available to the public, or at least to competent researchers. Data sets can be purged or cleaned of identifying information. Competent researchers can do responsible secondary analyses of the data while simultaneously upholding the normative requirements for protection of confidentiality. The investment of public funds in data supercedes ownership rights at least with respect to access to the data, as also do the norms of science as an activity open to and dependent upon the scrutiny and review of other scientists.

Since 1962, ICPSR has been responsible for many of the technical and normative developments in social science data sharing. As an archive that acquires data from many principal investigators, ICPSR has had to develop and implement procedures that assure original investigators that the distribution of their data will not compromise the protection of confidentiality. As an archive that distributes data to a wide variety of users, ICPSR has had to develop and implement these same procedures to substantially reduce or eliminate the opportunity for secondary users to compromise confidentiality even if they wanted to. Over the past 36 years, ICPSR has had to respond to new technical challenges in protecting the confidentiality of data, while simultaneously charting a course that satisfies both proponents and opponents of data sharing, both data producers and data users.

In this paper, we briefly review the origins of ethical requirements and regulations for the protection of confidentiality of research data and ways that confidentiality can be violated. That discussion sets the stage for a description of the nature and development of ICPSR practices to assure confidentiality of research data. These practices have had to take account of both technical developments in the capacity to store, distribute and analyze data and normative developments in the biomedical and social sciences about data sharing. Finally, we describe some trends in research that pose yet new problems for protecting confidentiality of research data and some new approaches to protecting confidentiality.

Ethics and Regulations

Biomedical sources.

Surprisingly, a review of the foundational documents that raised the consciousness about, and led to Federal regulation of, the protection of human subjects in research revealed very little attention to or concern with the privacy of research data and the protection of confidentiality. The Nuremberg Code (OPRR 1993c) addressed informed consent, social benefits of research, avoidance of suffering and injury, risks to subjects not greater than the importance of the problem, and preparations and facilities for protection of subjects against injury, disability and death. But it did not address issues of confidentiality and privacy. Beecher's seminal publications (1966a,1966b) focused primarily on safeguarding the physical health of research subjects, the absence of voluntary participation, and the need for informed consent. The Belmont Report's (OPRR 1993a) discussion of three basic ethical principles (respect for persons, beneficence, and justice) did not mention safeguarding privacy of research subjects or protecting the confidentiality of data obtained from them. The closest it came was in describing the principle of beneficence as making efforts to secure the well being of persons through minimizing possible harms. Of the basic documents, only the Helsinki Declaration mentioned privacy: "Every precaution should be taken to respect the privacy of the subject and to minimize the impact of the study on the … subject." (OPRR 1993b) But it did not extend this discussion of principles to its practical implication for protecting confidentiality. Finally, in the current Federal

regulations governing human subjects protection, confidentiality is mentioned only as an element of content of an informed consent statement. "…in seeking informed consent, the following information shall be provided to each subject: … (5) a statement describing the extent, if any, to which confidentiality of records identifying the subject will be maintained;" [45 CFR 46.116(a)(5)].

It seems likely that these foundational documents largely ignored privacy and consent issues because of their biomedical research origins, their primary concern with protection of the physical health and well being of the subjects, and the (false) assumption that physicians would be the primary personnel conducting biomedical research with people. Under these conditions, research information from or about human subjects is equated with information obtained under the privacy and confidentiality of the physician-patient privilege in a clinical relationship. So apparently little if anything was said about privacy and confidentiality in the early biomedical discussions.

Early social science data collection organizations. A sharp contrast is presented in the early history of the social sciences. Eckler, a former Director of the Census Bureau, reported that for the first five censuses (1790-1830), copies of returns were publicly posted for corrections or additions of missing information and were deposited with local courts (1972:164). He also reported that the sixth census (1840) was the first to instruct assistant marshals (i.e., field enumerators) that they were to "consider all communications made to him in the performance of his duty, relative to the business of the people, as strictly confidential" (Eckler 1972:165). Eckler speculated that this phrase was introduced into the instructions either to deter or curtail the private use of an increased amount of economic information collected in the 1840 Census, or to improve the reliability of reports to enumerators.

Protecting the confidentiality of data was an important concern for two of the early leaders of social statistics, Francis A. Walker, Superintendent of the 1870 and 1880 Censuses, and Carroll D. Wright, first Commissioner of Labor beginning in 1885 and later Director of the Census. Up until 1902, the Census was a temporary organization brought into existence each decade by legislation and terminated soon after issuing its reports. Congresses during the 19th century were indisposed to creating new, permanent Federal agencies. Walker was appointed Superintendent of the ninth Census (1870), the plans for which had become embroiled in larger political issues of apportionment of House of Representative seats and black suffrage (Anderson, 1988:76-81). The 40th Congress set aside plans for a more scientific census proposed by then Rep. (later President) James A. Garfield and the 1870 Census proceeded under the 1850 Census legislation. In the ensuing decade, Walker suggested a number of scientific and operational reforms for the census and a quinquennial census in 1875 (which never came to pass) (Wright 1900:58). The 1870 Census was the last census that used judicial marshals appointed by the Senate to supervise data collection in the states.

The tenth Census in 1880 used "supervisors of census" appointed by the President and who numbered more than twice as many as the judicial marshals, thereby providing more direct supervision of the actual work of enumeration (Wright 1900:59) and centralized planning and control (Anderson 1988:99). Each enumerator had to make daily reports and submit signed copies of original data schedules. Most importantly (for our present concern), the enabling legislation for the 1880 Census provided elementary forms of protection of confidentiality of the data. First, the oath of office signed by enumerators required that they "will not disclose any information contained in the schedules, lists or statements obtained by me to any person or persons, except to my superior officers." (Wright 1900:937:Section 7 of the Act to provide for taking the tenth and subsequent censuses). Second, Section 12 of the enabling legislation made it a crime to violate the confidentiality of responses:

> "That any supervisor or enumerator, who, having taken and subscribed the oath required by this act, … shall, without the authority of the Superintendent, communicate to any person not authorized to receive the same, any statistics of property or business included in his return, shall be deemed guilty of a misdemeanor, and upon conviction shall forfeit a sum not exceeding five hundred dollars." (Wright 1900:938)

In the eleventh Census (1890) the language about "any statistics of property or business" was changed to "any information gained by him in the performance of his duties." (Wright 1900:946)

As Walker's reforms proceeded (including appointments based on merit rather than patronage), the size of the 1880 Census organization grew but it ran out of appropriated funds in 1881. Walker resigned in 1881, moving to the presidency of M.I.T. After criticism and buffeting by Congress and the popular press, control of the Census remnants and reporting finally passed to Wright in 1885 and was finally completed in 1888 just before the need for legislation for the eleventh Census (1890).

The policy language about confidentiality that had undergone modest changes from 1840 through 1890 applied only to data collectors. Other Census employees, in particular, tabulation clerks, and increasingly in 1880 and 1890, professional staff, were not similarly enjoined. Thus, in the law providing for the twelfth Census (1900), Eckler reported that "confidential treatment of the census records was, for the first time, required of all employees, and penalties for violation were applicable to everyone

(1972:165). Similarly, in the law that provided for the 1910 Census, Eckler reported that "the possibility of disclosure through published reports" was addressed by instructions in the industrial censuses that indicated that publication was to be made in such a way as "not to reveal the report of any establishment" (1972:165). This same provision was not extended to the population and agriculture censuses until 1930, presumably because of the lower risk of identifying people than companies. For the 1920 Census, data sharing with other government officials was strictly limited "by the provision that in no case should the information thus furnished be used to the detriment of the person to whom it relates" (Eckler 1972:165).

Things were more informal in the Department of Labor. Plewes (1985:222) reported that Carroll Wright operationalized the standards for protecting confidentiality of data on a personal basis. He sent telegrams to businessmen, pledging his word "as a government officer that names of your plants and of city and state in which located shall be concealed (Plewes 1985:222). Plewes suggested that obtaining cooperation for data collection about sensitive topics like working hours and conditions, child labor, and wage practices was the motivating force behind these personal persuasions. These practices eventually became associated with such higher objectives as "integrity, impartiality and independence" (Plewes 1985:222). Plewes also noted that (as of the date of his remarks, March 1985), the Bureau of Labor Statistics was one of only two Federal statistical agencies whose policies of protecting confidentiality have existed without the protection of an agency wide confidentiality statute.

The expansion of the Federal government has been accompanied by the expansion of its information collection role and activities. As more and more kinds of data have been collected, issues surrounding the confidentiality of and access to government statistics have also increased. In many instances, agency practices have been formalized into statutory protections of confidentiality of statistical data and prevention of compulsory disclosure. For example, Title 13 of the United States Code governs the activities of the U.S. Census Bureau. In section 9, requirements for the confidentiality of census data are spelled out.

> (a) Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, or local government census liaison, may, …,
>
> (1) use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or
>
> (2) make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or
>
> (3) permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports. (13 USC 9)

Where individual reports are allowed to be shared with government officials, those records are "immune from legal process, and shall not, without the consent of the individual or establishment concerned, be admitted as evidence or used for any purpose in any action, suit, or other judicial or administrative proceeding" (13 USC 9(a)(3)).

Microdata from U.S. Department of Justice supported research also has confidential status and is prohibited from uses in the legal process other than statistical research:

> …, no officer or employee of the Federal Government, and no recipient of assistance under the provisions of this chapter shall use or reveal any research or statistical information furnished under this chapter by any person and identifiable to any specific private person for any purpose other than the purpose for which it was obtained in accordance with this chapter. Such information and copies thereof shall be immune from legal process, and shall not, without the consent of the person furnishing such information, be admitted as evidence or used for any purpose in any action, suit, or other judicial, legislative, or administrative proceedings. (42 USC 3789g)

Professional association ethical guidelines.
A third source of confidentiality restrictions is the ethical guidelines of professional associations. The post Civil War decades of the 19th century and first two of the 20th century brought immense technological development, world changing scientific discoveries in physics and chemistry, major demographic changes in American society, and the development of professions and professional organizations. The American Statistical Association and the American Economics Association were front runners in the movement to lobby for a permanent Census Bureau. These associations were made up of persons who had prior direct experience with the censuses or whose graduate students worked with the Census or with the Department of Labor. Thus, it is not surprising that ethical guidelines or codes of professional social science organizations eventually reflected confidentiality policies. The same people who were leaders in the associations were also leaders in the emerging disciplines and professions of the social sciences and social statistics in which confidentiality policies were first introduced.

In general, professional associations are concerned with promoting the professionalism (and status) of their work. Some essential aspects of professionalism are the ability to control or discipline members at the fringes of respectable practice and the provision of members with resources against outside disciplinary or malpractice actions. Associations have developed codes of ethics that educate

members about allowable practices or ethically suspect practices, and that guide behavior in gray areas. Many address the protection of confidentiality of sources or data.

The American Sociological Association requires sociologists to "take reasonable steps to ensure that records, data, or information are preserved in a confidential manner," and that when confidential records, data or information are transferred to other persons or organizations, "they obtain assurances that the recipients … will employ measures to protect confidentiality at least equal to those originally pledged." (ASA 1997, Section 11.08).

The American Political Science Association addresses the potential conflict between civic and legal obligations to cooperate with governmental organizations and the "professional duty not to divulge the identity of confidential sources of information or data developed in the course of research." (APSA 1998:Section 6) They are also required to observe Federal and university rules and regulations for the protection of human subjects, including protection of confidentiality of data. (APSA 1998: Section 34)

The American Statistical Association, founded in 1839, has recently released a new draft publication, Ethical Guidelines for Statistical Practice, for comments. The section on ethical responsibilities to research subjects includes the following item: "Protect the privacy and confidentiality of research subjects and the data they provide."

(American Statistical Association 1998: Section II.D.3 at http://www.amstat.org/about/ethics.html)

The American Association of Public Opinion Research also has a Code of Professional Ethics and Practices policy pledging confidentiality. "Unless the respondent waives confidentiality for specified uses, we shall hold as privileged and confidential all information that might identify a respondent with his or her responses." (AAPOR 1998: Section II.D.2 at http://www.aapor.org/ethics/principl.shtml)

Summary.
The present emphasis on the biomedical roots of modern human subjects protection regulations and their original implementation in the Department of Health and Human Services obscures some important origins of the protection of confidentiality of records and data. Foundational documents of ethical principles of biomedical research are largely silent on issues of privacy and confidentiality. In contrast, mid 19th century US Census legislation required enumerators to keep information they collected in the course of the Census confidential, principally as an instrumental means to promote subject cooperation and truthful response. The practice of maintaining confidentiality of Census data was extended to all Census employees in 1900. Gradually, professional associations adopted policies for the protection of confidentiality. These policies are based not on instrumental values like improving the cooperation of respondents and accuracy of the data but on ethical principles like safeguarding the privacy of individuals and minimizing potential harm to subjects through disclosure of sensitive information to third parties.

US Census and Bureau of Labor Statistics confidentiality practices initiated in the late 19th and early 20th centuries anticipated two of the four major possibilities for failure to maintain confidentiality. The early statements about treating information as confidential in the Census enabling legislation from 1840-1890 and their extension to all Census employees in 1900 recognized that individuals with legitimate access to microdata could also behave illegitimately by selling or transferring data to third parties. Industrial census guidelines in 1910 and population and agriculture census guidelines in 1930 recognized that individual or microlevel identities could be deduced from macrolevel tabular data with small cell sizes, thereby reflecting the first concerns about statistical disclosure. In the next section we describe four main categories of failure to maintain confidentiality as a preface to describing activities undertaken to protect confidentiality of archival data.

Ways That Confidentiality Can Be Violated
There are four major ways that confidentiality can be violated, resulting in the release or deduction of individual identities and/or identifying characteristics: accidental release; malicious release; compulsory release; and statistical disclosure.

Accidental release may be due to sloppy data management procedures, ignorance or errors on the part of staff, or failure to follow standard procedures.

Malicious release may be due to theft or unauthorized transfer of data by disgruntled staff or by staff or others seeking financial gain, or through breaches of computer systems security.

Compulsory release may occur as the result of legal action or court order.

Statistical disclosure results from logical use or analysis of data to identify cases or events that are infrequent or rare, or unique patterns of characteristics which when associated with data from other sources, lead to subject identification.

The value of these categories is not merely descriptive. They also direct attention toward objects or mechanisms for maintaining confidentiality. The idea of accidental release

suggests that confidentiality is preserved by:

- educating staff about the need for confidentiality protection procedures;
- training and monitoring staff in the application of those procedures;
- performing quality control checks on data files that are developed for restricted use or public release; and
- maintaining adequate security for confidential information.

The central feature of malicious release is the idea that information (and hence data) has value and that there are people, whatever their motives, who may attempt to translate that value into cash or otherwise use the information inappropriately. Disgruntled staff, for example, may satisfy a symbolic urge for retaliation or retribution by unauthorized transfer or release of information. Regardless of whether the motive is instrumental or symbolic, the inappropriate, illegal behavior can be counteracted by deterrence and punishment. These dynamics suggest that organizations should have and use policies that prohibit the unauthorized use, transfer, or release of data. In the case of public release, even though there is no restriction on who can access the available data, there ought to be use restrictions consistent with the research and educational purposes of the organization.

The matter of compulsory release is too complicated and uncertain to be dealt with in an encapsulated discussion here. It is sufficient to note that the ethics of research are not the only requirements that researchers face and that the legal protection accorded the confidentiality of research data is not absolute or uniform across states or in different legal matters. Researchers have been ordered to release confidential data. Some have complied, others have refused and been penalized, still others have had initial orders overturned or modified on appeal. Again, the focus with this type of release seems to be a strong organizational policy against compulsory release that has as its basis the necessity of confidentiality in social research. Where possible, such policies should be backed up by regulatory or statutory nondisclosure protections, such as the DHHS certificates of confidentiality or the US Department of Justice statutes (42 USC 3789g) and regulations (28 CFR 22) prohibiting evidentiary or other non-research uses of justice research data.

The topic of statistical disclosure is also too complex to be dealt with in an encapsulated discussion. But fortunately, there is more information available on this topic than on the others. Statistical disclosure has been the focus of both professional and academic attention. There are a variety of established methods for preventing disclosure (Cox et al. 1985; OMB 1994). There is also developmental work in progress for devising and testing new methods (e.g., Duncan undated; Dutta Chowdhury et al. undated). Some of these have been discussed at this meeting. But the central feature of this way that confidentiality is preserved is its technical focus on the data themselves.

In general then, there are four approaches on which to focus attention for protecting the confidentiality of research data:

- education and training of persons who work with data;
- data management techniques and statistical procedures that can be applied to data;
- organizational policies that mandate confidentiality and data security; and
- government regulations and laws that protect the confidentiality of research data.

The next section of this paper focuses attention on the first two of these approaches at ICPSR.

Practices At ICPSR To Assure Protection of Confidentiality

Data modifications.

These sections borrow heavily from the ICPSR **Guide To Social Science Data Preparation And Archiving.** *(Material taken from Second printing 1997:16-17 is* italicized*.) Two kinds of variables often found in social science data sets present problems that could endanger the confidentiality of research subjects. Most familiar are the* ***direct identifiers*** *that may have been obtained in the process of data collection. These include items such as names, addresses (including ZIP codes), telephone numbers (including exchanges), Social Security numbers, and other linkable identification numbers such as driver license numbers, certification numbers, etc s. Data collectors should remove all such identifiers when preparing public use data sets.* If data sets are received with such variables, ICPSR will remove them as part of the lowest level of study processing. Increasingly, consideration is being given to returning to investigators data sets that are received with direct identifiers. This is because ICPSR practice is to preserve originally submitted data that could become the focus of legal action should it be known that ICPSR maintains a copy of such a data set.

*Another category of variables can often become problematic d*epending on the content of the data collection and the nature of the research subjects included in the data

set. *These are indirect identifiers that might be used (in combination or in conjunction with publicly-available information) to identify individual respondents. This category is harder to deal with, since it includes items that are often the focus of or useful for statistical analysis. That is probably why such information was collected in the first place. Some examples of these indirect identifiers are detailed geography (e.g., state, county, or Census tract of residence), organizations to which the respondent belongs, educational institution from which the respondent graduated (and year of graduation), exact occupations held, place where the respondent grew up, exact dates of events, detailed income, and offices or posts held by the respondent. Such indicators should be reviewed by the principal investigator/data collector and a judgment made about the effect of retaining such items upon the confidentiality of the research subjects before depositing the data in a public archive.*

Sometimes, variables usually considered to be indirect identifiers can become direct identifiers depending upon features of the research design. Job title or occupational role can directly identify a respondent when there is only one such position in an organization, one such organization in a town (or department in an organization), and the town (or organization) is identified, as well as the date of the data collection. For example, if the police chief, Presbyterian minister, high school principal, or any other unique figure in a community or organization identifies their job title or occupational role, and the community or organization is also identified, and the date of the data collection is known, then it is easy to find out exactly who that person was at that time.

Handling indirect identifiers.
If, in the judgment of the principal investigator, a variable might act as an indirect identifier (and thus could be used to compromise the confidentiality of a research subject), the investigator should "treat" that variable when preparing a public use data set. Modifications commonly used are:

- *removal—eliminating the variable from the data set entirely;*
- *bracketing—combining the categories of a variable;*
- *top-coding—grouping the upper range of a variable to eliminate outliers;*
- *collapsing and/or combining variables—merging the concepts embodied in two or more variables by creating a new summary variable*.

The following example is taken from the ICPSR *Guide To Social Science Data Preparation And Archiving* (1997:17).
An example from a national survey of physicians (containing many details of each doctor's practice patterns, background, and personal characteristics) may help to illustrate each of these categories of treatment of variables to protect confidentiality. Variables identifying the school from which the medical degree was obtained and the year graduated should probably be removed entirely, due to the ubiquity of publicly available rosters of college and university graduates. The state of residence of the physician could be bracketed into a new "Region" variable (substituting more general geographic categories such as "East," "South," "Midwest," and "West.") The upper end of the range of the "physician's income" variable could be top-coded (e.g., "$150,000 or more") to avoid identifying the most highly paid individuals. Finally, a series of variables documenting the responding physician's certification in several medical specialties could be collapsed to a summary indicator (with new categories such as "Surgery," "Pediatrics," "Internal Medicine," "Two or more specialties," etc.).

ICPSR staff consult with principal investigators to help them design or modify a public use data set that maintains (to the maximum degree possible) the confidentiality of respondents. The staff will additionally perform an independent confidentiality review of data sets submitted to the archive and will work with the investigators to resolve any remaining problems of confidentiality. The goal of this cooperative approach is to ensure that all reasonable steps have been taken to protect the confidentiality of research respondents whose information is contained in ICPSR's public use data sets.

Research Trends that Pose Problems for Confidentiality
Some types of studies include variables that pose unusually difficult or problematic threats to confidentiality but are also difficult to modify because of their central importance to the study. One such study is the multi level study having hierarchical files with linkage variables between files. Another type is the study that has exact event dates and birth dates. A third type is the study with geo-coded information. A fourth type is the qualitative narrative interview study. A fifth type, the longitudinal panel study, is not especially problematic when ready for archiving, but the need to maintain linkage and locator identifiers from one round to the next makes the study vulnerable to threats to confidentiality during its operational phases.

Multi-level studies, where data is collected about places, organizations, households, persons and events, simultaneously, is especially difficult to handle with the usual means of modifying variables. Often, information in the multiple levels of files will make it easy to identify individual subjects, but the linkage variables between files are essential to maintain the multi-level value of the study. Where identification risks are high because the multiple levels of information make it easy to narrow the focus on individuals, ICPSR will consider making the study a

restricted use data set.

Studies with many precisely dated events and birth dates also pose risks to confidentiality, especially if the event information also might have been publicized in the media or recorded in publicly available administrative records (e.g., court dockets). Exact dates in the study information and event characteristics can be matched against media or administrative record data allowing subjects to be easily identified. Nevertheless, the exact date information is often useful for various forms of time dependent analyses like survival analysis or event history analysis. Removing exact dates reduces the value of the information. Once again, the solution may be creating a restricted data set rather than removing information.

Studies with geo-coded information are also problematic. Depending upon the nature of other information in the study and the degree of area resolution, geo-coded studies may make it easy to identify subjects, especially when public information is available. For example, it would be inappropriate, unethical, and potentially dangerous to release a data set with the address locations of rape victims. Again, resolving these kinds of problems caused by multiple levels of information is not a simple process of modifying indirect identifiers because of the nature of the study.

Qualitative narrative interviews are another type of problematic study. The level of detail provided through in-depth interviews is extensive and often contains many references to people, places, events, associations, organizations, family relationships, persons not liked at work, and so forth. Someone with intimate knowledge of these patterns of information may be able to easily identify the individuals involved. The very richness of the detailed information is simultaneously the value of the study and the threat to confidentiality. Original investigators are loathe to restrict the richness of the narratives, yet are unwilling to release such detailed information because of the ease of identifying individuals involved in the scenes.

Providing Access To Original Indirect Identifiers

It is rarely the case that variables removed or modified to maintain confidentiality are without value for research purposes. Archives and other data providers, therefore, frequently field requests for some form of access to original data values. Three of these forms of access that have been utilized will be discussed here: customized data analysis performed by the archive/data provider; private use data sets; and front-end software.

The first method of providing access to restricted indirect identifiers retains the data in secure form but permits researchers to design analyses that use those data. **Customized data analysis** (often performed at cost to the researcher) affords the opportunity of obtaining analytic results from restricted variables. Typically, researchers will be asked to provide detailed analytic instructions—usually in the form of software commands—and the requested analyses are performed at the archive, with analytic output sent to the requesting party. At ICPSR and elsewhere, the output is examined by staff to ensure that the analysis results will not endanger the confidentiality of respondents. Delivery of a **private-use data** set allows original data values to be provided to a researcher, with the requestor explicitly assuming responsibility for maintaining confidentiality of those data. Most organizations that provide private-use data sets require a transaction form, replete with both researcher and official signatures certifying that such data will be securely held, to be used only by the requesting party in ways that protect respondent confidentiality. A third mechanism bundles an entire data set in an analytic software package which prevents examination of discrete values/cases while allowing statistical access to all variables. This **front-end software** alternative usually prevents extracting or downloading of original values on some or all variables. (The National Center for Education Statistics' Data Analysis System [DAS] is one example of such a software-based method of protecting the confidentiality of research subjects. Other such front-ends are actively being explored, including at ICPSR.)

Each of the mechanisms described above has advantages as well as drawbacks. None are completely satisfactory to both the research community and the repository/holder of original data. Tightest control of original data values is an attraction of the customized data analysis option, but is the least popular with active researchers. It is typically costly (in terms of both time and money), and frequently thwarts the iterative analytic style most common in the social sciences. Private-use data sets permit the most researcher control of the analytic process, at the expense of certainty of protection of respondent confidentiality. Enforcement of private-use data set provisions agreed to by requestors is difficult to effect, and sanctions against violators of promised assurances would inevitably involve a litigious voyage on mostly-uncharted waters. Possibly the most secure yet flexible alternative is the front-end software option. Yet from the archive's standpoint, this is probably the most expensive of the three alternatives; putting data into one of these packages is so time-consuming that it can practicably be utilized on very few data collections. Furthermore, it is doubtful that front-end software is wholly impervious to hacking by a skilled and determined violator. Finally, the learning of "yet another" software package and its guaranteed limitations raises the bar over which interested researchers must jump to access needed research data.

Other Alternatives for Protecting Confidentiality

Yet other mechanisms have been proposed or are being experimented with in the quest for the "ideal" way of

protecting respondent confidentiality. Brief mention will be made of three "positive" alternatives, before we close this section on a draconian note. **Licensing** a researcher to use a data set containing indirect identifiers is a variant on the private-use data set arrangement described above. Like it, a licensed use is agreed to after completion of a transaction form. Unlike private-use data set agreements, however, most licenses impose an up-front fee in the form of a security bond as surety for maintaining confidentiality. The fee has been known to range from a few hundred to many thousands of dollars. Several license mechanisms also require the researcher and her/his institution to assume all legal liability in any instance of breaching confidentiality. Needless to say, the popularity of this form of "access-with-assurance" is quite low in the research community (not to mention in the college/university legal offices).

A second alternative method is being discussed in more detail elsewhere at this conference, and so will be briefly alluded to here. This is the ***"perturbing"*** *o*f original data values to break the certain bond between any given data value and the (possibly identifiable) individual who may have provided the initial information. Since the essence of this technique is the altering of original data values, it remains suspect in the minds of several generations of social scientists. These individuals find it difficult to overcome one legacy of their training—getting error out of research data collections—which clashes with the practice of *introducing error* into a data set (however noble the purpose underlying that introduction).

Perhaps more promising is the concept of ***secure data analysis laboratories.*** *I*n such facilities, original data would be available for data analysis in a controlled setting, precluding such things as making copies of original data, investigating single cases, or transmitting the data offsite. Scholars would apply to visit the site to do data analysis in the laboratory under secure conditions. An experiment using this form of access can be found at Carnegie Mellon University, for its Violence Research Consortium Project supported by the National Science Foundation and the National *I*nstitute of Justice. Data from the National Crime Victimization Survey, which have long been distributed without geographic sector information, are available with geographic information at Carnegie Mellon to the violence consortium members This mechanism represents, for social scientists, a departure from a long-term trend of facilitating the export of research data from an archive or producer site directly to the institution (or desktop!!) of the interested scholar. It should be noted parenthetically that many research materials utilized by both historians and social scientists are available **only** by visiting the site where the research materials are housed. Included among such facilities are traditional archives and other repositories, including some fine social science collections like those of the Henry Murray Center at Radcliffe College.

Undoubtedly more costly for the individual researcher (and perhaps for the archive as well), this mode of access to confidential data may become more common with heightened concern for preserving confidentiality.

The search for suitable mechanisms for protecting confidential microdata promises to become a high-stakes venture. At risk is the Big Kahuna of post-WWII social scientific research practice—**readily available, empirical microdata.** Some in the statistical and social science communities, as well as in government, are beginning to worry about the release of **any** microdata, with a few even predicting its demise.

Conclusion

The very progress of social science research methodology has made it more difficult to safeguard the confidentiality of the research data. Removing direct identifiers is a foundational requirement for public use data sets but that is essentially a trivial task. More difficult tasks involve investigating which variables could be used as indirect identifiers and modifying them without significantly reducing the value of the data collection. Careful attention must be paid to interactions among the context of the study, the nature of the sample, and the characteristics of respondents to prevent ordinarily unrevealing information from becoming the pointer to an individual. But many studies today involve complex research designs with multiple levels of data collection, file linkage variables that are crucial to the statistical analysis, sources of information that are intrinsically locational in nature, or detailed descriptions of events or situations that can be cross-referenced in publicly available sources like the media or administrative records. Maintaining complete archival files for these kinds of studies may involve other procedures than simply eliminating or modifying variables. Procedures used in the past or under development include:

- conducting contracted analyses;
- creating private use data sets;
- developing front end software to limit access to data records;
- licensing data use;
- introducing noise (known statistical error) into data records;
- developing data laboratories in which the data can not be removed from the site.

References

American Association of Public Opinion Research. Code of Professional Ethics and Practices. AAPOR 1998 at http://www.aapor.org/ethics/principl.shtml

American Political Science Association. A Guide to Professional Ethics in Political Science. *(Second Edition)* Washington, DC: American Political Science Association, 1998.

American Sociological Association. Code of Ethics. Washington, DC: American Sociological Association, 1997. At http://www.asanet.org/

American Statistical Association. Ethical Guidelines for Statistical Practice. American Statistical Association 1998 at http://www.amstat.org/about/ethics.html

Anderson, Margo J. The American Census: A Social History. *New Haven and London: Yale University Press, 1988*

Beecher, Henry K., M.D. "Some Guiding Principles for Clinical Investigation," JAMA *195:135, March 1966.*

Beecher, Henry K., M.D. *"Ethics and Clinical Research"*, New England Journal of Medicine *274:1354, June 1966.*

Cox, Lawrence, Bruce Johnson, Sarah-Kathryn McDonald, Dawn Nelson, and Violeta Vazquez. *"Confidentiality at the Census Bureau,"* Proceedings *of the First Annual Research Conference, March 20-23, 1985, pp 199-218. Washington, DC: U. S. Department of Commerce. Bureau of the Census. 1985.*

Duncan, George, Ramayya Krishnan, Rema Padman, Phyllis Reuther, and Stephen Roehrig. "Cell Suppression to Limit Content Based Disclosure." Undated reprint from H. J. Heinz III School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA 15213

Dutta Chowdhury, Sumit, George T. Duncan, Ramayya Krishnan, Stephen F. Roehrig, and Sumitra Mukherjee. "Disclosure Detection in Multivariate Categorical Databases: An Optimization Approach." Undated reprint from H. J. Heinz III School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA 15213

Eckler, A. Ross. The Bureau of the Census. New York: Praeger Publishers, 1972.

Fienberg, Stephen E. *"Sharing Statistical Data in the Biomedical and Health Sciences: Ethical, Institutional, Legal, and Professional Dimensions,"* Annual Review of Public Health *15:1-18, 1994.*

Fienberg, Stephen E., Margaret E. Martin, and Miron L. Straf, editors. Sharing Research Data. Washington, DC: National Academy Press, 1985.

Inter-university Consortium for Political and Social Research. Guide to Social Science Data Preparation and Archiving. *(Second printing)* Ann Arbor, MI: ICPSR, (September), 1997.

Mishkin, Barbara. "Urgently Needed: Policies on Access to Data by Erstwhile Collaborators," Science *270:927-928, (November 10), 1995.*

Office of Management and Budget Report on Statistical Disclosure Limitation Methodology. Washington, DC: U. S. Office of Management and Budget. Statistical Policy Office. Federal Committee on Statistical Methodology, (May), 1994.

Office of Protection from Research Risks. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research *in* Protecting Human Research Subjects: Institutional Review Board Guidebook. *Appendix 6 (A6:7-14)* Washington, DC: U.S. Department of Health and Human Services. Public Health Service. National Institutes of Health, 1993a.

Office of Protection from Research Risks. Declaration of Helsinki *in* Protecting Human Research Subjects: Institutional Review Board Guidebook. *Appendix 6 (A6:3-6).* Washington, DC: U.S. Department of Health and Human Services. Public Health Service. National Institutes of Health, 1993b.

Office of Protection from Research Risks. The Nuremberg Code *in* Protecting Human Research Subjects: Institutional Review Board Guidebook. *Appendix 6 (A6:1-2).* Washington, DC: U.S. Department of Health and Human Services. Public Health Service. National Institutes of Health, 1993c.

Plewes, Thomas. *"Confidentiality: Principles and Practice,"* Proceedings *of the First Annual Research Conference, March 20-23, 1985, pp 219-226. Washington, DC: U. S. Department of Commerce. Bureau of the Census. 1985.*

Weil, Vivian, and Rachelle Hollander. "Sharing Scientific Data II: Normative Issues," IRB: A Review of Human Subjects Research 12(2):7-8, (March/April), 1990

Wright, Carroll D. The History and Growth of the United States Census. Washington, DC: U. S. Government Printing Office, 1900.

* Presented at the Annual Meeting of the International Association for Social Science Information Service & Technology (IASSIST), New Haven, CT, May 20, 1998.

Christopher S. Dunn Director, National Archive of Criminal Justice Data ICPSR and Erik W. Austin Director, Archival Development ICPSR.