Department: **UAMS Institutional Review Board**
Policy Number: **13.2**
Section: **Confidentiality**
Effective Date: **July 31, 2002**
Revision Date: **June 10, 2004**

SUBJECT: Confidentiality in Archived Data- Social Science

The Principles

Social scientists have a deep and genuine commitment to preserve the anonymity of the subjects whom they study in the course of their research. Most often applied to individuals who consent to be interviewed in surveys, this commitment extends also to groups, organizations, and entities whose information is recorded in administrative and other kinds of records.

The social sciences broadly defined (as well as a number of professional associations) have promulgated codes of ethics that require social scientists to ensure the confidentiality of data collected for research purposes. (See, for example, the "Ethical Guidelines for Statistical Practice" of the American Statistical Association, which stresses the appropriate treatment of data to protect respondent confidentiality.) Both the rights of respondents and their continued willingness to voluntarily provide answers to scientific inquiries underlie this professional ethic. That ethic applies to all participants in the research enterprise, from data collectors to archivists to secondary analysts who use such data in their research.

Sets of regulations also bind all of us in the research enterprise to measures intended to protect research subjects as well as data obtained from such subjects. These regulations range from federal and local statutes to rules instituted by universities and colleges.

The Practice of Protecting Confidentiality

Two kinds of variables often found in social science datasets present problems that could endanger the confidentiality of research subjects. Most familiar are the **direct identifiers** that may have been collected in the process of survey administration. 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. Data collectors should remove all such identifiers when preparing public use datasets.

Another category of variable can often be problematic as well: these are the **indirect identifiers** that might be used in conjunction with respondent attributes and publicly available information to identify individual respondents. This category is harder to deal with, since it includes items that are often quite useful for statistical analysis (indeed, that is probably why such information was collected in the first place), and is dependent on the content of the data collection and the nature of the research subjects included in the dataset. 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.

How to Handle 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 dataset. Typical kinds of treatment commonly used are:

- Removal--Eliminating the variable from the dataset entirely.
- Bracketing--Combining the categories of a variable.
- Top-coding--Restricting the upper range of a variable.
- Collapsing and/or combining variables--Merging the concepts embodied in two or more variables by creating a new summary variable.

Excerpted from "Guide to Social Science Data Preparation and Archiving", Inter-University Consortium for Political and Social Research