This is a response from the UCLA Social Science Data Archive http://dataarchives.ss.ucla.edu/
Prepared by Libbie Stephenson, Director 310-825-0716 libbie@ucla.edu

The Data Archive has been in operation since 1977 and serves the entire UCLA campus of faculty and students engaged in quantitative research, including archiving original collections of data through surveys, and providing access for the re-use of de-identified (public use) data by faculty and students for research and instruction. Faculty members who are engaged in survey research use the Data Archive as a support to data collection processes and life cycle management, and in making the data publicly accessible. The Data Archive is a member of the Inter-university Consortium for Political and Social Research, an organization of over 700 members, worldwide, engaged in the use and preservation of data used in quantitative research. "ICPSR maintains a **data archive** of more than 500,000 files of research in the social sciences. It hosts 16 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields." UCLA is one of the heaviest users of ICPSR public use data, well above (approximately 3x heavier usage) the median usage by other institutions with the same Carnegie Classification. The re-use of public data at UCLA is vital to the generation of scholarship and social science pedagogy.

Our response to the proposed ANPRM rules is based on experience in working with researchers from many disciplines; sociology, political science, economics, public health, law, geography, history, and quantitative researchers from medicine and nursing. Overall, we agree with most of the UC System-wide responses and below are comments on certain issues where we feel additional clarification may be useful or where we offer some additional or alternative viewpoints.

We also reflect the current viewpoints and practices of the data archive community. We acknowledge that changes in technology have created potential informational risks and new forms of re-usable data, such as data generated by social media and mobile technologies. We agree that informational risks need to be addressed by researchers as they carry out disclosure review assessments to produce de-identified data. We also suggest that consent forms should routinely provide language about the sharing of data and its re-use for research, where appropriate. However, we insist that ethical approaches to survey research, as well as the current and emerging technical processes used to address disclosure risk are more than appropriate and additional regulation is not only unnecessary but would hamper the ability to conduct social science research or to train new scholars. The suggestion that public datasets conform to HIPAA levels of disclosure would make most data that would normally be public use available only with considerable reduction of usable data items or not at all.

Our most significant reaction to the proposed rule changes focus on

- failure to adequately define the meaning of public use data vs restricted use data;
- failure to acknowledge current practices for addressing disclosure risk for social science surveys;
- failure to address federal requirements placed on researchers to share and make available for re-use, data collected with funding provided by federal agencies;

Definition of public use vs restricted use data

The UCLA OPRS has policies governing the creation and use of public use data as follows: http://ohrpp.research.ucla.edu/file/10054/7-2.pdf

> A. **Publicly Available** means that the general public can obtain the data. Data are not considered "publicly available" if access to the data is limited to researchers.
>
> B. **Public Use Data Sets** are data sets prepared by investigators or data suppliers with the intent of making them available for public use. The data available to the public are not individually identified or maintained in a readily identifiable form. Data suppliers may have both (a) publicly available de-identified as well as (b) restricted use data from the same data set. Examples of public use data sets include portions of the U.S. Census data, and the Health and Retirement Survey. Data shared informally among colleagues does not constitute public use data.
>
> C. **Research projects that merge public use data sets** in such a way that individuals may be identified or which are designed to enhance a public use data set with identifiable or potentially identifiable data are not covered by this guidance, and require prior UCLA IRB approval or certification of exemption from UCLA IRB review.

The UCLA policy goes further to state:

> The UCLA IRBs have determined that data sets …[that]…have been stripped of identifiers … are publicly available. As a result, **research using these data does not meet the definition of "human subjects research", and therefore does not require IRB review and approval or certification of exemption from IRB review.**

The reuse of public use is one of the key approaches in social science research and instruction. One study can generate a multitude of research questions and methodological approaches, yielding numerous scholarly publications. Furthermore, many students use data for assignments and graduate level work in virtually every discipline. Quantitative methods are taught in many campus departments and these courses involve the re-use of public data. The data collected by UCLA researchers that is made publicly available can be used to demonstrate the research impact of the university through the production of this scholarship. The proposed rules will put a harness on the work of social scientists and will reduce the ability of the university to share the fruits of scholarship with the public.

Public data and disclosure risk

The ANPRM documents suggest that rather than place data into a restricted use category, all data will simply conform to HIPAA levels of non-release of respondent details. We strongly disagree with the idea that all survey data should conform to HIPAA privacy rules. For the kinds of research conducted in the social sciences this would make most public use survey data no longer useful for scholarship because useful items such as age, gender or race would be almost always removed. When needed, survey data can be made available with certain restrictions already covered by existing IRB policies as well as research and archival practices. We think that more thought needs to be given to achieving a balance between disclosure risk and analytic utility.

We further suggest that the social science research community already has established ethical practices to protect respondents, through removal of direct and indirect identifiers, recoding or masking variables and other statistical techniques. Data archives such as ICPSR (http://www.icpsr.umich.edu) are developing

techniques to address privacy at the computing system level to carry out network scans and PC audits, and are building secure web-based data analysis tools and virtual data enclaves. We believe that the social science community is already addressing ways to reduce or eliminate disclosure risk and the suggested regulations only add a degree of frustration rather than supporting solutions to maintaining respondent privacy and confidentiality.

In addition, the Data Archivist at UCLA works with individual researchers to prepare data for public use (using established practices in removal of identifiers and other technical or statistical techniques), and provides advice and services in describing data management in grant proposals, as well as throughout the research process and eventual release of public use datasets. Rather than impose restrictive rules, we suggest that funding agencies require the addition of a data life-cycle manager, trained in disclosure risk assessment, on all funded survey research projects.

Privacy and data sharing regulations

Funding agencies currently require that researchers share their data and must address data sharing in grant proposals. The NIH policy which began in 2003 states that investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing.

The National Science Foundation states: "Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants."

We insist that the proposed rules would make it difficult to actually share data for re-use in such a way that it meets HIPAA levels of non-disclosure and makes new survey research meaningful. There needs to be better coordination of the rules and the policies of the funding agencies. The goals of data sharing are well stated by the NIH:

> [Data sharing]… is particularly important for unique data that cannot be readily replicated. Data sharing allows scientists to expedite the translation of research results into knowledge, products, and procedures… Sharing data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new datasets when data from multiple sources are combined."

We believe that the rule changes proposed in the ANPRM would pretty much prevent any of these goals from being achieved. We do not think the University, the state or the country can afford to lose these opportunities.

Specific responses to questions:

Question 2. We emphatically agree that continuing review is not needed for survey research, especially when the data collection is finished and the data are simply being analyzed.

Question 14. We suggest that the current terminology of 'exempt' be maintained; we do not see the point of the 'excused' re-classification. However, we note that there have been University of California cases where survey data including all respondent identifiers have been requested under FOIA. Thus far respondents have been protected because the university was able to maintain that the respondents were under IRB authority and therefore information about them could not be released. See the UCLA Civil Rights Project http://civilrightsproject.ucla.edu/

Question 46: We suggest that text be included regarding pre-existing de-identified survey data. We maintain that there should be one consent form for the original collection of the data and that there be simple language in the consent form assuring respondent privacy and confidentiality if the data are shared with other researchers, (per funding agency requirements). For example, Stanford uses the following text: "Subsequent uses of records and data will be subject to standard data use policies which protect the anonymity of individuals and institutions."

Question 54. As described above: We further suggest that the social science research community already has established ethical practices to protect respondents, through removal of direct and indirect identifiers, recoding or masking variables and other statistical techniques. Data archives such as ICPSR are developing systems to address privacy at the computing system level to carry out network scans and PC audit, and are building secure web-based data analysis tools and virtual data enclaves. We believe that the social science community is already addressing ways to reduce or eliminate disclosure risk and the suggested regulations only add a degree of frustration rather than supporting solutions to maintaining respondent privacy and confidentiality.

In addition, the Data Archivist at UCLA works with individual researchers to prepare data for public use (using established practices in removal of identifiers and other technical or statistical techniques), and provides advice and services in describing data management in grant proposals, as well as throughout the research process and eventual release of public use datasets. Rather than impose restrictive rules, we suggest that funding agencies require the addition of a data life-cycle manager on all funded survey research projects.

Question 55. We suggest language be included about the role of data archives which preserve data for the long term. It is certainly the case that data obtained 10-15 years ago may contain identifiers now considered to be linked to privacy disclosures. Most data archives maintain connections with data producers to ensure that any data that may not conform to new definitions of de-identified information are handled appropriately. In one UCLA case data obtained from the California Office of Statewide Health and Development containing suspect geographic identifiers were returned for versions without the level of earlier detail. These are normal existing ethical practices and do not need any regulation.

Question 59. See response to question 54. We suggest that the social science research and archival community is addressing the technical, analytical and ethical aspects of informational risks. We believe it is too soon to try to regulate informational risks in survey research because they are constantly evolving. We also suggest that the ways surveys are conducted are also changing and therefore the methods used to protect respondents are changing. Traditional face-to-face interviewing has been replaced by computer assisted interviews or web surveys. We do not agree that HIPAA levels of non-disclosure can cover every type of survey or those that will be developed in the future with new methodological or technical approaches. See: http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/confidentiality.jsp

Question 61. We suggest that data protection standards guidance is well established in survey research but would like to see these methodologies more prominently discussed and have no objection to publications such as the one produced by NIST.

Question 62. In the social sciences datasets produced by HIPAA covered entities will have so little detail that they do not need data use agreements. However, we do see a possible rationale for data use agreements geared to protect researchers in their re-use of public data. That is, we seek to protect researchers who may be subject to law suits in two cases: first, in the case of the original researcher disagreeing with the outcome or conclusions of the re-use; and second in the case where respondents disagree with the outcome or conclusions of re-use. Right now the focus is all on protecting respondents, but there are no protections to researchers in any of the proposed rules.

Question 63. Currently all data deposit forms (used when data is provided to an archive for long term preservation and public access) and data use agreements explicitly state that trying to re-identify anything in the data is prohibited. For example, ICPSR uses the following: "No attempt will be made to identify any individual person, family, household, business, or organization. If an individual person, family, household, business, or organization is inadvertently identified, or if a technique for doing so is discovered, the identification or discovery will be immediately reported to ICPSR, and the identification or discovery will not be revealed to any other person who is not a signatory to this agreement." There is no reason to expressly prohibit it since this requirement is already in force for survey data.