Introduction

The Data Preservation Alliance for the Social Sciences (Data-PASS) was formed to take shared responsibility for the long-term accessibility to social science datasets that are of value to current and future researchers and policy-makers. Preservation of subject confidentiality is a core value for partners in the Preservation Alliance for the Social Sciences (Data-PASS). Members of the partnership use a variety of methods for protecting the confidentiality of the study participants described in the many social science research studies within each organization's holdings. This document outlines our policies for the confidentiality of materials acquired for the Data-PASS project.[1]

To maintain accessibility and support dissemination, the quality and integrity of the information within and about a data collection must be controlled throughout the various stages in its life-cycle. Based on current security procedures already in place at each organization, these standards protect against the destruction and loss of the data, whether through natural disasters, fire, vandalism and/or error.

Data security is usually understood to involve availability (e.g. through redundancy and management of the computing environment), integrity (e.g., through backups and verification methods), and controlling access (by authenticating users, and authorizing actions on the data). In the archival context, we include data migration within "security", since we use migration to ensure the availability or the intellectual content of the data we maintain, as well as to maintain its integrity.

Data confidentiality is closely related both to security (which includes physical systems security), and to digital rights management (which provides a framework for authorization and access control). This document is one of three that together, address all of these issues. This current document discusses the practices that we have identified to maintain data confidentiality.

Confidentiality Issues in Research Data[2]

In general, dissemination of any research data that describes human subjects is a potential subject of confidentiality concerns. There are a number of recognized broad ethical principles governing research: respect for persons; 'beneficence' (a positive obligation to protect subjects from harm and to benefit them), and justice. Application of these principles to the conduct of research leads to the requirements of informed consent, systematic assessment of the benefits and risks to the subject (including elimination of

[1] These standards are intended to reflect best practices among academic data archives. Government agencies, may be governed by regulation. In particular, NARA, a current Data-PASS partner, is governed by: *Code of Federal Regulations*, 36 CFR Part 1234, Electronic Records Management, Subpart C --Standards for the Creation, Use, Preservation and Disposition of Electronic Records Access to all federal records accessioned into the National Archives is governed by the federal Freedom of Information Act, as amended [5 U.S.C. 552, as amended]. See also, 36 CFR Part 1256.

[2] Some of the following section is excerpted from Altman & Franklin, *Managing Social Science Research Data*, Chapman & Hall/CRC, (Forthcoming) 2008.

any risks not required to achieve the research goals and subject benefits), and fair procedure for subject selection.[3]

These principles carry over to the dissemination of data collected as part of research. There exists an ethical obligation by the researcher and the archive to avoid disclosure of potentially harmful information without the informed consent of the subject, and to minimize the potential harm (whether economic, legal, or psychological) from information disclosures where these are authorized and necessary for research purposes.

In addition to these general ethical obligations, there is an evolving body of law that governs the disclosure of personal information and which creates specific obligations for researchers and archives. It is impossible to discuss all of these laws, since they are complex, and differ in scope, source, and coverage. As of the time we write this, the laws having the greatest impact on disclosure of research data for the most part stem from United States code: the Health Insurance Portability and Accountability Act of 1996 (which governs the disclosure of health-related information); Family Educational Rights and Privacy Act (FERPA, which governs the disclosure of educational information); The Privacy Act of 1974 and The Freedom of Information Act (controlling disclosure of information collected by the federal government); the Code of Federal Regulations on Protection of Human Subjects (45 CFR 46; and 21 CFR 56, controlling disclosure of information obtained through research at institutions that accept U.S. government funding) and the Graham-Leach-Bliley Act (protecting financial information). Other important laws regarding the treatment of personal data include the California Financial Information Privacy Act, The Video Privacy Protection Act, and Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 (protecting the personal information of European citizens and residents); and various national common and codified laws addressing invasion of privacy and defamation.

The specific requirements of these laws are complex and varied. Nevertheless, it is possible to identify some general requirements (and non-requirements) regarding data dissemination.

First, it is important to note where confidentiality laws apply:

- Confidentiality laws do not restrict the dissemination of information that does not describe human subjects.
- And, confidentiality laws do not restrict information that cannot be *directly or indirectly linked* to individual subject.

Release of data that can be *directly or indirectly linked* to an individual may be restricted by applicable federal, state, or local laws. (We discuss how we determine when data can be directly or indirectly linked in the next section.) Generally speaking, however, these laws do *not* restrict the release of such personally-identifiable data for which any of the following hold:

[3] See Belmont Report: Ethical Principles and Guidelines for the Protections of Human Subjects of Research, 1979. Available from: < http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm>

- All the information in the data has been previously released, or its release would not constitute a potential unwarranted invasion of personal privacy; thus, for example, information in data about identified living persons who are public figures, and where the data relates to their public roles or is not particularly sensitive (as below), is generally not restricted; or,
- A sufficient length of time has passed since the collection of the information so that the data can be considered "historic;" or,
- All identified subjects have given explicit informed consent allowing the public release of the information in the dataset; or,
- All the information was collected with an explicit statement concerning the public nature of the data, such as information collected for governmental regulatory purposes; or,
- For federal records (data created by a U.S. federal government agency or under a federal contract), all identified subjects are deceased and no federal statute explicitly restricts the release of the data

The following information is likely to be deemed particularly sensitive (if identified, dissemination of this information warrants increased scrutiny, although is not necessarily prohibited:

- Information gathered from minors, the mentally impaired, or prisoners
- Information describing financial records, educational records, health records, or video rental records

Data-PASS Partner Practices

The Data-PASS partners have identified and employ a number of practices to manage the disclosure of information that is restricted due to confidentiality laws and concerns. (These techniques need not be applied to data that is not restricted by confidentiality laws, such as data previously released, and so forth, as described above.)

These practices fall into three categories:

- *De-identification*, which is the deterministic removal or modification of directly and indirectly identifying variables.
- *Statistical Disclosure Control*, the use of statistical techniques such as permutation, introduction of random noise, cell-suppression, and observation censoring.
- *Usage restriction*, which can include restricting use of the data to authenticated and authorized users, requiring that use of the data be restricted in duration or location, auditing use of the data, or imposing other specific terms of use.

These practices should be, of course, accompanied by appropriate *documentation* of the specific de-identification, disclosure control and usage restriction policies and procedures of the archives, and by sufficient *record-keeping* to verify that these policies were honored.

These techniques should be used only when necessary to protect subject privacy, since de-identification and disclosure limitation can reduce the usefulness of the data. For example, identifiers are crucial for linking data; data that can be linked to other studies provide additional secondary benefits: it allows for additional, and external, accuracy checks; it can obviate the need for later redundant data collection; it can enhance the value of small data collections; and it can be used to add contextual information to other studies. Disclosure limitation can distort the relationships between variables, lose information through aggregation, sometimes in unpredicted ways, and can mask rare events or important outliers.[4] For these reasons, archives sometimes apply a hybrid approach of creating a de-identified public-use data set, and maintaining an identified version under much more restricted use conditions for linking and more detailed analyis.

De-identification

De-identification is the most common method used by the Data-PASS partners when disseminating confidential data that is restricted by confidentiality laws (as described above). It is important to note, in particular, that the deposit agreement used by our partners requires, in effect, that the depositor de-identify any confidential data prior to depositing it. Depending on the partner which accepts the deposit, and on the collection into which it is accepted, the data *may* be subject to additional review by the archive staff.

De-identification involves the removal of any information that could be used directly or indirectly to link the data to individual subjects. For strictly quantitative data, de-identification typically consists of removal of all identifiers explicitly enumerated in major regulations. Currently, HIPAA provides the most complete list which provides safe harbor if these identifiers are removed, and the archive has no other reasonable basis to believe that the data is identified:

- names of subjects (and relatives, employers and household members)
- any geographic subdivisions smaller than a state
- any dates smaller than a year
- phone #'s
- fax #'s
- social security numbers
- e-mail addresses
- medical record numbers
- health plan numbers
- any other account numbers
- certificate/license numbers
- vehicle ids; device ids
- URL's
- IP addresses
- biometric identifiers (fingerprints, retina, voice print, DNA, etc.).

[4] Commission on Behavioral and Social Sciences and Education,2000. *Improving Access to and Confidentiality of Research Data*, National Academy Press.

For most quantitative data, this results in data that can be connected to individuals only in unusual circumstances, and thus generally satisfies confidentiality concerns.

Removal of all of these identifiers is not strictly necessary to prevent linking of the information in the dataset to individual subjects. For a particular dataset, especially one that has few other demographic variables, an archive may analyze the data (e.g. by cross-tabulation) and determine that it is not possible to identify individuals, even where geographic subdivisions smaller than a state, dates smaller than a year, and other non-individual identifying information is included. Such determinations do merit increased scrutiny of the dataset since they increase the potential risk of harm to the identified subject and/or issues of liability for the archive.

On the other hand, removing the identifiers above also does not provide an absolute guarantee that an individual subject can never be identified. So, for data that was gathered from extraordinarily sensitive research subjects, or contains an extraordinary number of demographic variables (increasing the likelihood of indirect identification), or where there is another reasonable basis to believe that the remaining information could be used to identify a person, archives may choose to subject the data to additional confidentiality review.

Furthermore, qualitative data, containing free-form textual responses, video or audio data is more difficult to de-identify while retaining research value. Names may be replaced with pseudonyms or identifiers, so that the responses of an individual can be linked within a dataset but not externally. Qualitative data must be reviewed as a whole to ensure that the content of answers to separate questions, when taken together, do not contain sufficient information (such as a list of all the education institutions attended) to identify a subject. Full-face images and voices must be masked to prevent identification. Since these techniques are more complex, all qualitative data is reviewed by archival staff before its release.

Statistical Disclosure Control

Statistical disclosure control is a supplement to de-identification that has been used for several decades – primarily by government agencies releasing aggregated data. Statistical disclosure control can involve randomly permuting responses; introducing random noise to continuously measured observations or aggregated statistics; suppressing aggregate results that are based on too few observations; and censoring extreme observations (e.g. recoding all ages above 80 to “80 and above”). The general goal of statistical disclosure control is to modify individual observations (or summary statistics from table cells) to prevent identification, while preserving the overall statistical properties of the data.

Statistical disclosure control further reduces the likelihood that an individual can be identified based on data disclosure. And, in some special circumstances, it can provide a provable guarantee that the disclosed data does not allow identification. However, it is

often computationally intensive, and may restrict the types of analysis that can be performed on the data, and/or affect the correctness of subsequent analyses. There are many different techniques being researched, none of which is clearly recognized as a best practice. Thus it is used most frequently by large statistical agencies producing large datasets[5] – and indirectly by archives re-disseminating the already modified datasets produced by these agencies. However, in unusual circumstances, where the archive has a reasonable basis to believe that de-identification is insufficient, it is used by some of the partner archives. [6]

Usage Restriction

Restrictions on usage can be used as a substitute or complement to the modifications of the data described above. Restrictions can take several forms.

First, each partner archive has a set of general terms of use that govern access to any data obtained from it. These terms should include requirements that the user of data not attempt to violate the confidentiality of subjects described in the data, and should report to the archive any identifying information that they discover in data obtained through the archive. Second, each dataset can be associated with additional terms of use, described in its documentation and metadata.

Depending on the sensitivity of the data, as determined by the individual archive, these terms of use may be enforced by a number of increasingly restrictive methods (corresponding to increasing sensitivity of the data):

1) The website for the data archive should provide prominent links to the terms of use under which data may be obtained.
2) The user may be required to agree to a click-through terms of use in order to obtain the data. (The Data-PASS catalog contains mechanisms to administer these terms, based on the usage restriction metadata for each study.)
3) The user may be required to obtain a login account, provide basic identifying information that is subject to some level of automated check (such as an IP-address check, or having the user reply to an e-mail sent to a registered address).
4) The user's registration information may be subject to further, manual verification of identity, institutional affiliation, and organizational authority.
5) A specific application may be required for particular data resources. Such applications usually include a statement of the intended use of the data, and should require the user to represent that they will honor confidentiality and other restrictions. Approval for use of the data may be limited to a particular duration or subset of the data, and may include reporting requirements.

[5] In particular the Census uses many of these techniques to protect the confidentiality of data collected under USC Title 13. See the Census Bureau Standard on Disclosure Control (and related checklist) at: http://www.census.gov/quality/S14-0_v1.2_Disclosure.htm

[6] For an extensive set of citations to disclosure control techniques see: http://www.icpsr.umich.edu/HSP/citations/index.html

6) Use of the data may be restricted to a particular computer system or, network. (E.g., users may be required to use the data only on a physically secure system that has been disconnected from all networks. Alternatively, users may be required to use data only through a secure login to a data server hosted at the archive.)
7) All operations on the data by the user may be audited, with the audit logs subject to review by the archive. (This implies that all data use is restricted to a computer system hosted by the archive, and that only the results of analyses are retained by the researcher. If copies of the data were disseminated outside of the archive's administrative control, auditing cannot be ensured.)

Balancing Confidentiality and Access

Most of the data handled by the Data-PASS partners is not sensitive (or not deposited in a sensitive form), and is intended by the researcher to be widely used. Archives have a variety of tools available to minimize the potential for harm to subjects stemming from the release of data.

Authored by: Micah Altman, Copeland Young
Revised and approved by Operations Committee on 04/03/2007