DATA PASS ACQUISITION GUIDELINES

INTRODUCTION

The Data Preservation Alliance for the Social Sciences (Data-PASS) will preserve electronic social science data that has been identified and selected for acquisition and preservation. These data collections will be identified and selected in adherence to our content selection guidelines. Their value and significance will then be evaluated according to our appraisal guidelines. Data collections that are deemed suitable for acquisition will then be subjected to processing for preservation and release.

Before a dataset is made available for distribution, a number of steps are taken to ensure that it will meet the needs of the research community and preservation requirements. Effective data processing ensures that no matter what condition the materials were in when initially acquired, the material that is made available and preserved by the archive will be as accurate, complete, and well documented as possible.

It should be noted that the complexity and diversity in the formats of the material that will be acquired for this project will require some acquisition and processing decisions to be made on a case by case basis. Further, because a variety of potentially time consuming and costly steps are involved in processing research data for dissemination and secondary analysis, the project managers and the processors involved must be given some flexibility in the decision making process. It should also be noted that material that is initially processed at a more basic or routine level, could be re-processed in the future, if demand or a change in its relevancy warrant additional processing to occur. What follows should be considered primary guidelines that will be followed by each of the partners.

IDENTIFICATION AND APPRAISAL:

The partners intend to identify the most significant social science data of the past seventy-five years, based on a wide variety of research criteria. We will systematically identify and select social science data that are classic or destined to be classic, not already available at an existing data archive, and at-risk of loss. (For more on this process, please review our Content Selection Guidelines.) All data collections that are identified for further investigation will be assigned a persistent identifier so that the progress of each data collection can be tracked and monitored.

Once the data is identified as appropriate for preservation, the collection will enter the Appraisal phase. The Operations Committee will review the information on the data, based on our Appraisal Guidelines including, significance of the data to the research community, significance of the source and context of data, uniqueness and usability of the data and so on. (For more on this process, please review our Appraisal Guidelines.) If the committee approves of the data as part of the Data-PASS project, it will be officially assigned to the most suitable organization. Once the assignment is made, it will be up to the partner organization to determine the priority of processing and the level of processing that will occur.

ACQUISITION OF STUDY DATA AND DOCUMENTATION:
The following initial checks should be performed to verify the findings of the Operations Committee appraisal of the data. If discrepancies are found, the Operations Committee should be notified.

- **Review of Appraisal Checklist:** The appraisal checklist, as completed by the Operations Committee will be reviewed. A list of all materials that arrived as part of the data collection will also be reviewed.

- **Verify the contents of the data:** The submitted files will be opened and reviewed to make sure they are not corrupt.

- **Perform cursory confidentiality check:** A cursory confidentiality check will be performed on all non-electronic documentation, removable electronic media and online electronic files to discover data fields that can potentially identify individual respondents, such as name, telephone number and SSN.

- **Submit all Materials for Acquisition:** A copy of all the files will be acquired and preserved. These files should be preserved in the format(s) that they were received.

- **Inventory the Acquisition:** The list of all the acquired materials with their original names and formats will be included in a master database that will include information on all data collections that are processed as part of this partnership.

PROCESSING PLAN:
Once all the electronic and physical materials have been acquired, the data collection will move to the processing phase. Although each partner has different descriptions for the amount of processing that may occur, much of it can be identified as falling into one of three categories: **minimal** processing, **routine** processing, or **intensive** processing. The required level of processing required for each individual data collection will influence the overall processing plan.

Minimal Processing
Minimal processing is usually planned for materials that arrive with much of the processing already completed by the depositor. If suitable for release, these studies are made available to researchers soon after they are acquired. Confidentiality concerns are resolved, hardcopy documentation is converted to electronic form, and electronic format documentation (e.g., Microsoft Word documents) may be converted to a Portable Document Format (PDF). The format of the data may also be converted to multiple formats (e.g. SAS, SPSS, Stata), but little or no further processing is undertaken.

With regard to the Data-PASS project, minimal processing may also occur for data collections that are considered important to the goals of the Data-PASS project, but could not be further processed in an effective manner at the present time.

Routine Processing

Routine processing is planned for materials that arrive in a relatively complete manner and may not justify the cost of more intensive processing. Typically, those reasons would be that a study comprises a single site, is not replicated, or is a cross-sectional study. It is also possible, for the purposes of this project, that the material is not highly relevant to current issues in research, policy, or practice, but may hold nominal current value or potential future value. Studies targeted for routine processing go through additional content authentication processing steps beyond those of minimal processing.

As occurs under minimal processing, checks for confidential information contained within the data are completed and transformations to prevent disclosure are performed. It is possible, however, that the loss of the confidential data could detract from the significance of the data collection. Creating a restricted dataset provides a viable alternative to removing variables that can identify the participants. In this scenario, a public-use dataset that has these variables removed or recoded will be released. The original dataset will be kept as a restricted-use dataset that preserves the original variables. The restricted-use dataset will be released only to clients/users who have agreed in writing to abide by the rules governing the use of these restricted datasets.

As part of routine processing, checks to ensure that the material is complete and accurate are conducted. The documentation is scrutinized to make sure detailed information is available for every variable. Data definition statements that can be used with statistical software packages can often be created to facilitate the use of the data by researchers. In addition to these types of files, files that include both the data and required information to analyze that data can also be made available (e.g. portable or transport files).

Intensive Processing

Intensive Processing is designed for studies that are generally multi-site, multi-state, or replicated; national studies or nationwide data; international studies; longitudinal or cohort, panel, or time series collections; and studies that are highly relevant to current public policy or research concerns. For the purposes of this project, it is also possible that intensive processing would occur for a highly important data collection that is in an endangered format or includes minimal documentation. In addition to all authentication processing steps listed above, further operations may be conducted for intensively processed studies.

Consistency checks may be conducted to ensure that skip patterns in the questioning were followed correctly. Other consistency checks can be conducted to identify unlikely relationships among variables (e.g., children older than parents) or respondents outside the sampling frame (e.g., juveniles included in a study of retirees).

RESOLVING ISSUES OF CONTENT DISCREPANCIES

Throughout the acquisition and processing procedure, incomplete information or other concerns that can affect the usefulness of the data may be uncovered in the received materials. In these occurrences, we will seek to ensure that we archive and distribute the most complete version of a data collection. Clarification will be sought through communication with either the data producer, a representative of the data producer, or someone familiar with the data collection.

BUILDING THE DOCUMENTATION SET:

Documentation about the data is equally important to the data itself. It should include all the information that would be required to gain a complete understanding of the data collection. Some types of documentation include a codebook and data collection instruments (survey questionnaires). A codebook, in turn, can include frequencies and a data record layout. Additional information that could prove useful includes background information, such as the study's history, as well as sampling methodology, weighting procedures, processing details, and related publications.

Because many of the data collections we anticipate receiving may be old and their preservation may have been somewhat neglected to this point, we may receive minimal or incomplete documentation. In these cases, attempts will be made to obtain additional information using diverse resources and means to create meaningful documentation for the data collection. Processing notes containing a description of any anomalies, alerts to the user about known errors, and discrepancies in the data and documentations will also be added to the created codebook.

A study description that consists of Study-and-File-Level metadata will also be created for each data collection. Using the documentation provided, summaries will be written that include descriptions of the study design, sampling procedures, goals of the research, and technical characteristics. The metadata created for each data collection will be in accordance to standards that have been agreed to by the partners. (For more on this process, please review our Metadata Guidelines.)

PACKAGE THE STUDY:

To package the study or to prepare it for preservation and dissemination, quality checks will be performed on the final versions of both the data and documentation. When the final checks have been completed, the data collection will be go through procedures to

make it available to the research community and prepared for long-term preservation. For long-term preservation, two backups of all machine-readable files on Digital Linear Tapes (DLTs) will be created. One tape will be made available for **on-site** storage, the other for **off-site** storage. Additional information on data security will be made available via our Data Security Standards, to be written in early to mid 2006.

EMBARGOED DATA

As organizations committed to providing information to the public, the partners will make many of the data collections available to our constituent communities. However, it may be the case that not every data collection will bring with it dissemination rights; those collections will be acquired for preservation alone. Each partner has experience with such "embargoed" data. We are well suited to limit the amount of material that is unavailable for dissemination or to limit the time that it is unavailable. Additional information on disclosure standards will be made available through our Disclosure Standards, to be written in early to mid 2006.

SUMMARY

Just as there is relative agreement in the way the materials are initially transferred to the partners, there is similar consensus in how the final product is made available to researchers. The end of the archival process for each data collection added to our archives will be long-term preservation, release, and dissemination. When final quality control standards are checked, the study will be released from the processing and quality control stages and permission will be made available for release and dissemination. All materials created since the initial stages of acquisition will also be secured for long-term preservation.

The availability of the data collection will be announced to the research community. Each archive will announce the study in a similar fashion as they do their other studies. However, recognition that this collection is part of the Data-PASS partnership will also be required. Information about these data collections will be available through the Data-PASS website and a shared catalog.

Approved by the Operations Committee on February 21, 2006.