## INTRODUCTION

The Data Preservation Alliance for the Social Sciences (DataPASS) will initiate its project to preserve electronic social science data by identifying and selecting data that ought to be acquired and preserved. The partners intend to identify the most significant social science data of the past seventy-five years. The target data range from historic fertility and family longitudinal studies to contemporary studies of individuals and public opinion poll data. The target data will also include historic and contemporary data funded by grants from the National Institutes of Health and the National Science Foundation, as well as government data, political process data, private public policy research data, and data collected by individual researchers in all social science fields. In addition, the partners will also identify and select social science data based on a wide variety of research criteria, such as macro- and micro-data, data collected for research purposes, data produced as a by-product of running or administering a public or private program, methodologically innovative social science research data, and respondent-level survey data. In conducting this identification and selection process, the partnership will work toward supporting social science research among academics, students, and public and private sector researchers through the preservation and dissemination of major social science data. The content identification and selection guidelines developed by the partnership and outlined in this document will provide the general guidance used by the partnership to identify and acquire social science data broadly defined.

A major consideration will be the extent to which data are considered at risk, i.e., those data for which longevity and access are threatened, and where delay in preservation may increase the risk of loss of the data. The partners will continuously examine the technological and social factors that influence the speed of data extinction. If data are available at an alternative site and if there is confidence that availability will continue over time, the risk of loss is diminished. Data may be given higher priority for being acquired when the information is not available from another reputable, accessible data archive. The partners will continually consider the issue of risk because of its central role in the activities of the project.

## APPRAISAL GUIDELINES

The total amount of data identified as appropriate for preservation may be more extensive than the combined capacity of the partners to acquire and preserve. Therefore, the partners will apply standard selection, or appraisal, criteria. These criteria will incorporate elements of accepted archival practice to identify the most important content to preserve and an evaluation of the risk of losing the content should acquisition not take place.

Appraisal is the process of determining the value of data and whether they are temporary or permanent. Appraisal is not a rote exercise. It requires informed judgments,

knowledge of and sensitivity to researchers' interests, recognition of resource considerations, and a willingness to acknowledge and understand comments and suggestions from diverse perspectives.

Archivists appraise records for their evidential and informational values. Evidential value refers to information about the creator, which may be an individual, an organization or a governmental entity. Informational value refers to information about persons, places and things outside the creator. Social science data are generally appraised for their informational value. For data about persons, social science data will more likely have permanent value if they describe a socially significant group of people and contains a wide range of data about them. Examples of socially significant groups would be all residents of the United States, all members of the U.S. military, and all children in the public elementary schools in the United States. Permanent social science data should normally include basic demographic and other contextual information about each individual in addition to the topic of the data collection. For data about places, social science data will more likely have permanent value if they describe places in a clearly defined and socially significant area or if they describe a clearly defined and socially significant type of place, and if they contain basic demographic and other contextual information about the people in those places in addition to the topic of the data collection. Examples of significant areas or types include Europe, the United States, or cities in the United States with over 500,000 people. Permanent social science data that normally describe things have information about socially significant institutions or organizations, such as businesses, educational institutions, fraternal or religious organizations, or government agencies. Examples would include all Federally chartered banks, all colleges and universities in the United States, or all police departments. In addition to the topic unique to the data collection, permanent social science data about institutions and organizations will normally contain economic and/or demographic information.

The guidelines below provide a consistent framework for appraisal decision making. Applying the guidelines to specific cases will not be a mechanical process akin to adding up points or checking boxes. However, using these guidelines will make decision making easier and will result in appraisal judgments that are consistent and that can be readily explained and defended. In developing appraisal recommendations, the partners will address the following questions. The questions should be considered together, rather than in isolation.

**1. How significant are the data for research?**

The future research potential of data is the most difficult variable to determine. What is of relatively low research use today may become of great research use in the future. Perhaps even more important and difficult to predict are the issues and topics that will be considered of significance in the future. Nevertheless, it is important to consider this question in making appraisal decisions. It is necessary to consider the kinds and extent of current research use and to try to make inferences about anticipated use by social scientists, other researchers, and by the government. In making this judgment, factors to

consider include the substantive value of the collected information; time frame of the information; uniqueness of the collected information; relationship to other holdings in the archival collections and to the collections of other repositories; relationship to previous studies; the scope of the data (local/regional/national, depth and breadth of information); the influence of these data and the investigator in the social sciences; the data collection methodology; and ability to use the collected information for secondary studies. Privacy and confidentiality issues do not override the long term value of the data.

**2. How significant is the source and context of the data, particularly in regard to scientific progress and society?**

The significance of the functions and activities performed by the creator of the data and the social science context within which the data are created are important considerations for the appraiser. Thus, data must have demonstrated importance to the social science community as determined by: (1) substantive value and its influence on social science knowledge, (2) enduring archival value, and (3) uniqueness. It will also be important to place a high value on data that permit policy analysis and research addressing broad public policy issues including human society and human interaction, the environment, the role of networks and communication, and political and economic theory and practice. The appraiser must relate the source and context of the data to the strategic framework and objectives found in these guidelines.

**3. Is the information unique?**

Appraisal must be conducted in context with other data. The appraiser must determine whether the data under consideration are the only or are the most complete source for significant information. Data that contain information not available in other sources (including data in your repository as well as files accumulated by other repositories) are more likely to warrant permanent retention than records containing data that are duplicated in other sources. However, a repository may decide to retain data that contain information available elsewhere in the case of data that are more complete or more easily accessible than the alternative source or a more closely related to other information in that repository. Even if data are unique, however, they may not warrant continued preservation depending on the other appraisal criteria.

**4. How usable are the data?**

Consider these three issues:

> A. How does the way the data were gathered, organized, presented, or analyzed affect their usability? For example, does the scope of the data cover a national population sample or a representative subsample of the population of the U.S., a state, region, or foreign country? Do the data offer enough depth and breadth of information to support a wide range of research methodologies? In reference to polling data, the polling methodology should also meet the basic disclosure criteria as stated in the American Association for Public Opinion Research Code for Professional Ethics and Practices.

B. How do technical considerations affect the usability of the data? For example, some electronic records may pose such technological challenges that extraordinary measures may be required to recover the information, while other records containing similar documentation (either electronic records or records in another format) may be usable with much less effort. Appraisers will also want to give strong preference to data collections that have comprehensive technical documentation providing ample information on the data structure and code tables, sampling procedures, weighting, recoding rules, and data collection procedures to allow users to assess the quality and analytical reliability of the data. Furthermore, if data lack essential components, such as data layouts and code tables, it probably does not warrant being acquired. It will also be important to give a higher priority to data in a readily useable digital form and accessible to researchers at a variety of computing and technological settings.

C. How does the physical condition of the preservation media affect the usability of the data? For example, some media may have deteriorated to the point that the data they contain are not readable.

**5. What is the timeframe covered by the information?**

"Timeframe" may refer to the date span of the entire body of data or the length of time that individual records or file units typically cover.

A. The longer the date span for which there are extant files, the more valuable the data are likely to be for research. In addition, longitudinal studies, which show the effects of social change or historical events over time, can be particularly valuable, although consider carefully the potentially high attrition rate for these studies.

B. Some data are made up of individual documents or files whose content covers many years. This attribute includes records where the documents in individual files are accumulated over a relatively short period but contain information pertaining to events covering a long period of time.

**6. Are the data related to other data in the archives?**

Data that add significantly to the meaning or value of other data already appraised as archival are more likely to warrant retention than data lacking such a relationship. Examples would be data that fill substantive gaps, that round out existing subject area concentrations, or that are new version of or additions to data collections in the holdings. Data that are chronological continuations of data already held by the archives are likely to warrant permanent retention, particularly if the older segments of the data are subject to high reference use. Consideration may also be given to methodological issues affecting data linkages, such as the extent to which units of analysis, measurement error, and

sampling or other collection methodology permit useful linkages to other holdings.

**7. What are the cost considerations for long-term maintenance of the data?**

This consideration should play a significant role only in marginal cases. In such cases, an appraisal should balance the anticipated research potential of the data with the resource implications of retaining them permanently. If data carry significant costs for acquisition, processing, archiving and distribution, the value of the data must clearly outweigh the costs. Other things being equal, data with low long-term cost implications are more likely to warrant permanent retention than those data with high long-term costs.

**8. What is the volume of data?**

Data that are clearly permanent based on the other appraisal guidelines listed above should be designated for permanent retention regardless of the size/volume of the data. The size/volume of a collection should be a factor in the decision making only when the permanent value is marginal. In those cases, the costs of preservation/maintenance may outweigh the value of the data. Other things being equal, data that are compact are more likely to be appraised as permanent than those that are voluminous.

Approved by the Operations Committee on February 9, 2005.