THE SELECTION, APPRAISAL, AND RETENTION OF DIGITAL SOCIAL SCIENCE DATA

M Gutmann[1]*, K Schürer[2], D Donakowski[3] and Hilary Beedham[4]

*[*1]Inter-university Consortium for Political and Social Research, University of Michigan, Ann Arbor, MI 48104*
Email: gutmann@.umich.edu
[2]UK Data Archive and Economic and Social Data Service, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ
Email: schurer@essex.ac.uk
[3]Inter-university Consortium for Political and Social Research, University of Michigan, Ann Arbor, MI 48104
Email: dwdonako@icpsr.umich.edu
[4]UK Data Archive, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ
Email: beedh@essex.ac.uk

ABSTRACT

The number of data collections produced in the social sciences prohibits the archiving of every scientific study. It is therefore necessary to make decisions regarding what can be preserved and why it should be preserved. This paper reviews the processes used by two data archives, one from the United States and one from the United Kingdom, to illustrate how data are selected for archiving, how they are appraised, and what steps are required to retain the usefulness of the data for future use. It also presents new initiatives that seek to encourage an increase in the long-term preservation of digital resources.

Keywords: Social Science, Data access, Data sharing, User needs, Confidentiality

1 INTRODUCTION

Starting with the 1890 US Census and the 1911 census in England and Wales, data used for social, economic, and political research were the first materials to be converted to digital format for analysis by computer technology (Anderson, 1988; Higgs, 2004). The first tabulating machines were invented to record and archive the data collected on Hollerith cards. Using technological advances such as this, social scientists were able to invent a reliable, scientifically validated means for ascertaining public opinion that has enabled citizens to have a voice about the nation's affairs (Converse, 1987). For three-quarters of a century, public opinion polls, social surveys, and other kinds of structured interviews have tracked socio-economic data, as well as people's values, attitudes, knowledge, and behavior – measuring and recording our cultural and social heritage. Preserving these data, both for posterity and for secondary analysis, is the job of social science data archives.

The number of data collections produced and the variations in their quality and in their potential for future secondary analysis, however, do not allow for the long-term retention of all the social science research data. It is therefore necessary to make decisions regarding what can be preserved and why it should be preserved. This paper reviews the processes used by two data archives, one from the United States, the Inter-university Consortium for Political and Social Research (ICPSR), and one from the United Kingdom, the UK Data Archive (UKDA), to illustrate how data are selected to be archived, how those data are then appraised, and what steps are required to retain the usefulness of the data for future use. The problems and limitations that occur as part of the process are also cited, as well as new initiatives designed to address those issues. But first, it is important to understand what types of data are collected and by whom they are collected.

1.1 What constitutes digital data in the Social Sciences

Historically, most of the research conducted in the social sciences has produced two types of data: categorical or closed-ended survey responses and qualitative or open-ended survey responses. A large proportion of the information used by social scientists comes from responses to questions that are closed-ended, either asking for specific factual information ("How old are you?") or electing categorized responses to questions of opinion ("Which of the presidential candidates do you think you will vote for in the upcoming election?") The common characteristics of these surveys is that their information can be readily converted to numerical values and made digital. Some of these data, including the US census and the UK census, have their origins in administrative records that require universal coverage, Other data collections are produced by government-sponsored, university-sponsored, or other research organizations, using sampling techniques that allow relatively small sample surveys to represent a larger population, whether local, regional, or national.

Not all research questions, however, are suited to categorical responses. Often the exact words that a survey respondent used to discuss his or her life, attitudes, or experiences are important. These qualitative survey responses require that the full text of the interview be made available for study and analysis and analytic software is now available to do this. While many of these qualitative studies are the product of randomized samples of respondents, other researchers may be more directed in their interviewing strategy, choosing key actors or selecting representative lives to chronicle.

There are also increasing amounts of 'non-survey' type data. These data come in the form of images or video. Sound files and mixed media are also increasingly collected as part of social science research. The ongoing Project on Human Development in Chicago Neighborhoods (PHDCN) is taking an integrated approach by combining many of these types of data including surveys, systematic social observations via videotapes, and official records from administrative sources (PHDCN, 2004).

1.2 Who produces the data

The materials that encompass social science data have diverse origins. The largest producer of social science content in the United States and the United Kingdom is the government. Much of this content comes from surveys conducted by government agencies, while other activities are conducted under contract with other organizations. At the opposite end of the spectrum are private survey and marketing firms, which have been instituted since the 1930s to gauge public opinion and to evaluate the demand for products (Converse, 1987). In the US, new organizations were formed after World War II to meet the demand for research. For example, the RAND Corporation was spun off from Douglas Aircraft in 1948, and the Survey Research Center at the University of Michigan grew out of United States Department of Agriculture (USDA) survey efforts also in 1948. Other organizations established in that time period, include the National Opinion Research Center (NORC), founded in 1941, and the Research Triangle Park, which was set up in 1958. In the middle of the spectrum are academic researchers, who design research projects supported with local funds, corporate or foundation support, or government grants. Some researchers manage their own data collection, but most large-scale projects are done under contract by local or national-level survey organizations.

It is important to note that not all archived data were originally produced for research purposes. Administrative databases, such as police and public school records, have been collected for very different purposes. They can nonetheless yield valuable and timely information for researchers to further advance their investigations, much as the content of government archives of survey data has allowed traditional research for many years.

1.3 Who owns the data

The government maintains ownership of the data it produces, and generally maintains ownership of research that it contracts. Ownership of data that are collected with government resources depends on the type of funding that was made available. Much of the social science research conducted in the US with the assistance of the government is funded with grants. In these cases, as when the data are privately funded, ownership is maintained by the organization conducting the research or the researchers conducting the study.

In the UK the situation is similar insofar as ownership of data is not passed from the originator to the archive and distribution is undertaken under license. Data collected by or on behalf of all government departments are always “Crown Copyright” (i.e., owned by the government). Where a government department contracts an external organization to collect data for government purposes, ownership does not pass to the organization that collects the data and in some instances, the organizations are bound by contract to offer the data to the UKDA for preservation.

Many government data producers are bound by law to preserve information and, under certain controls, to make their data available to researchers for further analysis. In the US, for example, Federal data that are determined to be of continuing value are archived at the National Archives and Records Administration (NARA). This requirement extends to some of the data produced by Federal contractors, but it does not always include data from Federal grantees.

Private businesses and university-based researchers have assumed until recently that the data they generated were their property and that they had limited obligations either to share their data with others, or to ensure its preservation. In both the US and the UK, pressure on academic researchers to share their data has increased, especially in situations where their research has been funded by government agencies and/or private foundations. Despite the notion by some that survey and other social science data are private property, an international movement to archive, preserve, and share as much as possible emerged when digital data began to appear in volume.

Depositing digital social science information in an archive usually does not change its ownership status. An archive may house the data and make them available to others in the research community, but ownership is usually maintained by the individual or organization that produced the data, which grants a limited license to redistribute the data to the archive The major exception among social science archives in the United States is the Murray Research Center, which generally requires depositors to transfer copyright at the time of deposit.

1.4 Who archives the data

There are many data archives around the world. In the United States, for example, the Odum Institute at the University of North Carolina (founded in 1924) and the Roper Center for Public Opinion Research (founded in 1947) predate the creation of modern computers and the widespread production and preservation of digital data.

The data archiving movement began in earnest in social science departments in Europe and the United States in the 1960s. The Central Archive for Empirical Social Research was created in 1960 as an institute of the University of Cologne's Faculty of Economics and Social Sciences. ICPSR was founded as a consortium of political science data producers in 1962. The UKDA (formerly known as the Social Science Research Council Data Archive and the Economic and Social Research Council Data Archive) at the University of Essex was founded in 1967. Other social science archives in the United States, like the

Henry A. Murray Center at Radcliffe College, were founded in the 1970s as the use and preservation needs for social science data broadened.

The first data archives initially collected survey data of specific interest to quantitative researchers in the social sciences. As interests expanded and additional archives were created, sources such as aggregate/macrodata, census, and administrative data were added. Although non-confidential social and economic data collected by the government are generally public and freely available, archives began to include these data in an effort to make them more accessible. The goal of most major social science data archives has been to maintain and provide access to as much data as possible for research and instruction, and to ensure that data are preserved against technological obsolescence and physical damage. Each archive, however, operates in its own way to achieve these goals.

2 DATA SELECTION AND APPRAISAL

Not all social science research data are archived. Some data do not have long-term value or have a low level of potential future interest. Data could have value or interest, but do not include enough supporting information to allow for secondary analysis. Still other data collections may have value, but do not fall under the scope of the archiving organization. And, of course, economic factors also play a role in the amount of data that can be selected and preserved. Each archive sets its own standards, but many of these standards can be traced back to basic guiding principles.

2.1 Identifying data

Archives routinely identify content for acquisition through a wide variety of means, including longstanding agreements for the deposit of data (including serial data collections), unsolicited donations of data from researchers and survey organizations, requests from the research community, press releases and published reports announcing results of a study, papers presented at professional meetings and scholarly conferences, and consolidated lists of studies published by researchers and research organizations**.** ICPSR, as membership-based institution, also has a commitment to acquire data that are recommended by its governing council or its members, either through direct contact or a periodic survey of interests. Equally, the UKDA has a special mandate to archive data generated as a result of Economic and Social Research Council (ESRC) funding and also acquires data from a large number of UK government departments, especially the Office for National Statistics (ONS).

Selected data collections can be brought into an archive to accomplish several different purposes, among them to fill substantive gaps in the holdings, to appeal to new constituencies, to round out existing subject area concentrations, to support new research techniques, and to rescue and preserve data that are in danger of being lost. At both ICPSR and the UKDA, when a data set or a collection is identified as having potential suitability for acquisition, contact is made with its primary collector (e.g. lead researcher, survey organization, funding agency) to gather detailed information about the collection (type of data, sampling information, sample size, etc.) and to inquire about its availability for archiving and any conditions that could restrict its distribution.

A significant amount of data are not sought out in one of the ways described above but rather are deposited by researchers who are eager to preserve their data and to ensure that they are made available to others. Some of the data that are received in this manner are submitted to meet granting agency requirements that data be deposited in a public archive.

2.2 Determining suitability for preservation

Once a data collection is identified for potential preservation, a determination of its viability must be made. Appraisal standards vary from archive to archive, but NARA's latest guidelines (2003), based mainly on the traditional intellectual appraisal of the value in records as well as physical considerations such as usability and cost, provide a good example of the types of issues that are raised. The appraisal guidelines list several questions, which are to be considered together, not in isolation, including:

- How significant are the records for research?
- How significant is the source and context of the records?
- Is the information unique?
- How useable are the records?
- Do the records document decisions that set precedents?
- Are the records related to other permanent records?
- What is the time frame covered by the information?
- What are the cost considerations for permanent maintenance of the records?

Even if the answers to these questions suggest that data preservation is in the interest of the social science community, other issues may influence the retention decision. One of the primary considerations in the selection process involves the extent to which the data will advance knowledge. Thus, data must demonstrate importance to the social science community as determined by substantive value, enduring archival value, and uniqueness. Keeping in mind that it is not always evident in the present what data we might need in the future, studies can be scrutinized from a long-range perspective in an attempt to predict whether or not the data will remain valuable. Such important datasets may be acquired even if they fall short in terms of other criteria.

The documentation and supporting information that are included in the data collection are also important in the decision-making process. Data should have comprehensive technical documentation that provides ample information on sampling procedures, weighting, recoding rules, and data collection procedures. Data collections that include such information are strongly preferred in the archiving decision, as they allow users to assess the quality and analytical reliability of the data.

Along with these criteria, other issues could impede the retention and preservation of data even after a decision to preserve it and they will be discussed in the next section. Ultimately, however, the data that are provided for preservation should be in a format that allows for dissemination and preservation, and allows the data to be migrated to other formats. By making these appraisals early in the process, less negative appraisals of final data occur and more time can be spent on preparing approved data for retention and preservation.

Each archive has a procedure for determining whether data that are presented for preservation meet the archive’s appraisal standards. At ICPSR, appraisal decisions are made by the director of each of six individual sub-archives, in consultation with the Director of Acquisitions and the Director of the Collection Development Unit. In the UKDA, a more formal process involves the presentation of the appraisal case to an acquisitions committee, the Acquisitions Review Committee (ARC), which meets every two weeks. Each new study is reviewed and a decision is taken whether or not to accept it based on specified criteria including the re-use value, sample size, and insoluble copyright, legal, or ethical issues.

3 DATA RETENTION AND PRESERVATION

It is important to note that there is no clear break between data selection and appraisal on the one hand, and retention and preservation on the other. In fact, the appraisal process bridges the two aspects and continues until the data are ready for final preservation and/or redistribution to the research community.

The retention and preservation phase involves more than simply receiving or making copies of data. Staff members examine each dataset from the viewpoint of a potential user and then apply both social science knowledge and technical expertise to add value to and enhance the usefulness of the dataset for secondary analysis. It should be noted that while the original materials that make up the data collection are preserved as they were originally received, additional processing is often undertaken to make sure the materials will meet the needs of later users. 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 complete, accurate, and well documented as possible.

Just as differences exist between archives in the standards regarding data selection, so too are there differences in the standards regarding retention and preservation. Because of the diversity of both the format and the complexity of the material that can be acquired by data archives, some acquisition and processing decisions have 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 processors at ICPSR have the flexibility to decide how extensively to process any given collection. At the UKDA, the Acquisitions Review Committee decides the level of processing when the study is first reviewed. This will only be changed if unexpected problems arise that cannot be resolved and any such decision would be referred back to the committee. As with the selection and appraisal phase, there are nonetheless basic principles that most archives follow.

Beyond the question of retention and transfer, other issues of security and preservation arise. The most severe challenges to preservation of digital materials occur months or years after initial intake of such research materials. As such, the archiving effort requires a sustained integration between science and technology. Supported file formats can become inaccessible with changes or updates to the software, and a lack of backward compatibility of the newer software. To prevent material in the archival holdings from becoming unusable or inaccessible, both the media and the storage format of all collections need to be reviewed on a continual basis. In addition, the storage format of any collections deemed to be at-risk either because the formats are no longer supported in current software packages, or because they do not easily lend themselves to conversion by individual researchers, should be converted.

Just as it is important to protect against material becoming obsolete, data security is also a main priority for data archives. High standards for the security of acquired project materials are required. These standards protect against natural disasters, fire, errors, and vandalism. To guard against accidental destruction of any of the acquired materials, archive staff create security copies of all project materials. The current standards of both the UKDA and ICPSR call for multiple copies to be made, each stored in a different physical location, including one copy that is taken off site to a secure location. If archival processing produces further "versions" of core data files, copies of those are made and secured. Security copies are stored on media scientifically tested for its archival preservation qualities, such as high-density Digital Linear Tape (DLT) removable devices.

Most social science data archives rarely review their holdings with a view to de-accessioning material that is not of long-term value. As preservation concerns become more complex, requiring continuous processing to keep archival content useable, and as the size of the archives grow, this may become more of an issue. Although not a formal process, the UKDA keeps its collection under constant review and, on occasion, a study will be de-archived.

4 PROCESSING DATA FOR PRESERVATION AND FUTURE USE.

A key aspect of ensuring that social science data will be useful in the future is reviewing and processing the data. The first step is content authentication. This task includes verifying the receipt of data and documentation files, completing an inventory of the contents of the submission, and ensuring that the correct number of files has been received and are readable. The collection is also evaluated for completeness, verifying that the material is correctly identified in the documentation. Additional checks involving cursory confidentiality reviews assess the material's suitability for public release, which will be discussed in more detail in an upcoming subsection. Finally, working copies of the data and documentation are made and the original files are prepared for archival storage and preservation.

The amount of processing required will be based on the results of this review and the complexity of the material. Although many archives have 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, basic processing, or intensive processing.

4.1 Levels of Processing

At both the UKDA and ICPSR minimal processing is designed for materials that arrive with much of the processing already completed by the depositor, as well as materials that an archive wishes to preserve but for which it currently lacks the resources for additional processing. If suitable for release, these studies are made available to researchers soon after they are acquired and in the format in which they arrived. Confidentiality concerns are resolved, hardcopy documentation is converted to electronic form, and electronic format documentation (e.g., Microsoft Word documents) are converted to a Portable Document Format (PDF), but no further processing is undertaken.

Basic, or routine, processing is designed for materials that arrive in a relatively complete manner and which, for a variety of reasons, may not justify the cost of more intensive processing. Basic processing may also be conducted on material that may presently hold nominal value, but is not highly relevant to current social science research topics or may hold potential future value. Studies targeted for basic processing go through additional content authentication processing steps.

As occurs under minimal processing, checks for confidential information contained within the data are completed and transformations to prevent respondent disclosure, or re-identification, are performed. If it is determined that the loss of the confidential data could detract from the significance of the dataset, a restricted dataset may be created that includes the confidential data. Creating such a dataset provides a viable alternative to removing variables that significantly increase the risk of respondent disclosure. In such cases, a public-use dataset without the confidential variables is released. The original dataset is kept as a restricted-use dataset that preserves the original variables. The restricted-use dataset is released only to approved researchers who have agreed in writing to abide by the rules governing the use of these restricted datasets. Although many of the processes are similar at the UKDA, the management of restricted datasets is somewhat different. A sensitive dataset would be anonymized in much the same way as at ICPSR but would only be made available for secondary use, on a case by case basis and with the explicit agreement of the depositor.

The processor also checks the data to ensure that the material is complete and accurate. He or she verifies the accuracy of the submission by using a statistical software package, such as SAS or SPSS, to compare the number of cases, the number of variables, and other file-specific information to that provided in the documentation. Steps are also taken to create and/or revise the technical documentation. Using this documentation, staff write summaries that include descriptions of the study design, sampling procedures, goals of the research, and technical characteristics. The conversions to the documentation that occur as part of minimal processing also occur in both basic and intensive processing.

Intensive processing is designed for studies that are generally multi-site, multi-state, or national studies; international studies; longitudinal or cohort, panel, or time series collections; and studies that are highly relevant to current public policy or research concerns. In addition to all of the authentication processing steps listed above, additional operations can be conducted for intensively processed studies.

Files that can be used with statistical software packages are often created or enhanced to facilitate the use of the data by researchers. Depending on the material received, this can be a trivial or lengthy process. These “setup” files provide column locations and formats for each variable, descriptive variable and value labels, and missing value declarations. These files can be modified by the researcher to select only variables of interest or for use with other statistical packages. In addition to these types of files, portable or transport files that include both the data and information contained in the setup files can be made available.

These files can be directly read into statistical analysis packages without further documentation of the data to the software.

Because confidentiality of survey respondents is so important, additional checks and analyses can also be conducted that are designed to minimize the risk that the identities of respondents will be revealed to someone using the data (O'Rourke, 2003). While the careful attention to protecting respondents has long been a part of data archiving, new and more sophisticated disclosure risk analysis and limitation procedures are now available in the world of data archiving. We discuss this more fully later.

Throughout the acquisition and processing procedure, incomplete information or other concerns that can affect the usefulness of the material may be uncovered in the received materials. It is common for data to contain undocumented or "wild" codes, usually the result of errors of key entry but sometimes the result of a valid code being undocumented. There are also instances in which the documentation and the data do not match, or the documentation is incomplete. In these occurrences, we seek to ensure that we archive and distribute the most complete version of a dataset.

4.2 Confidentiality

One of the biggest concerns in the preservation of social science data is confidentiality. Most social science data are gathered from human respondents and reflect their thoughts and actions, so the challenge of protecting their identity is a top priority. This priority is addressed in the examination of all material for both direct and indirect identifiers that would allow a data user to identify a specific individual.

Direct identifiers include such things as names, addresses, phone numbers, or any other public identification. They are usually identified in the initial checks that occur when material is first acquired. Direct identifiers can be completely removed from the dataset, including any reference to the fact that it was originally part of the collected material. In most cases, however, the direct identifier is blanked (all occurrences within the dataset are deleted) or recoded (all occurrences are recoded to a predetermined identifier, such as "0" or "9").

Indirect identifiers are more complicated. They are variables that, when used in combination with other data, can identify an individual. Some examples of 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 (such as birth, arrest, graduation), detailed income, and offices or posts held by the respondent.

Indirect identifiers are a potential problem when there are few cases and some of the variables have low frequency counts on a value within the data collection. It can also be a problem if these variables can be linked to other resources, which could make the research participants subject to being identified (de Wolf, 2003). For example, running cross-tabulations may show that there is only one individual in the study born on a certain date and arrested on a certain date. That information could be used to trace back to public files and identify the person.

If it is determined that a variable might act as an indirect identifier (and thus could be used to compromise the confidentiality of a research subject), the variable information must be altered. In some cases, modifying a dataset to remove identifiers can compromise the research value of the data. In special circumstances, users can get access to restricted-use datasets after signing a contract and agreeing to certain conditions. To facilitate research activities that cannot be effectively performed without access to data items that have been removed, masked, or collapsed in the public-use version of a data collection, there are measures in place at ICPSR that will (upon written request) provide a restricted-use dataset to an individual researcher for a specific time period. Provision of such a dataset requires a fully-executed agreement. It should be noted that this service is not available at some archives, including the UKDA, nor is it available for every dataset from ICPSR.

The issue of confidentiality is of great importance when reviewing all the materials associated with a collection. The issue can also arise when direct identifiers are included in the supporting technical documentation. When this occurs, such information is deleted from the documentation that is ultimately made available, and therefore, directly effects the creation of the metadata for the material.

4.3 Metadata

Another factor that greatly determines a data collections long-term viability and therefore, its preservation value, is the quality and richness of its metadata. Metadata, quite simply, are data about data. The metadata requirements for digital social science content are immense, and can therefore require extensive editing during the processing phase. In marked contrast to the digital representation of a book, for example, social science data are seldom easily understood on their own. Put simply, a social science survey rendered in digital format is just a list of numbers, often nothing more. In order for a future content user to make sense of those numbers, extensive metadata must be prepared. These metadata tell us the structure of the information, the underlying questions that were asked, the ways those questions were interpreted, and any changes that were made to the data, either to preserve them (preservation metadata) or to ensure that no information is reproduced and disseminated that will allow the identities of individuals to be revealed if they were promised confidentiality.

Four types of metadata can be created, preserved, and distributed:

- Study-level metadata, also known as abstracts, study descriptions, or metadata records. This is the highest level of metadata, describing the study or collection as a whole, and is primarily intended for resource discovery purposes. These metadata outline the purpose of the study, the major conceptual categories studied, the characteristics of the sample, measures, etc.
- File-level metadata. These describe the properties of individual files in a data collection.
- Variable-level metadata. This type of metadata describes individual measures or groups of measures. These are detailed in technical documentation such as codebooks and data definition statements and are essential to effective and accurate interpretation and use of the numeric and character data.
- Administrative and structural metadata. These are critical to ongoing maintenance and preservation of the electronic data collections. They must be complete enough to permit future archivists to discern how the files were produced and how they might be migrated or emulated in an evolving technological environment. Management of this type of metadata can vary from organization to organization.

The major social science archives are converging on an XML-based metadata standard called the Data Documentation Initiative (DDI). The DDI is “an endeavor to provide a straightforward means for social and behavioral scientists to record clearly and then to communicate to others all the salient characteristics of the empirical data for which they are responsible” (DDI, 2004).

If a digital record does not have sufficient metadata to adequately explain its content and relevance, or these metadata cannot be reproduced, it will usually remain unavailable for secondary analysis. However, if the required metadata can be created during the retention phase, efforts will be made do so, particularly if the data collection is considered to be one of great importance or has the potential to be of great importance to the research community.

Because of the importance of metadata, properly documenting data for long-term preservation and access must become part of the daily practice of researchers. Promoting these changes to established practice requires collaboration among scientists, data managers, educators, and archive customers.

4.4 Final processing steps

The end of the archival process for each study added to a collection is long-term archiving, release, and dissemination. When final quality control standards are met, the study is released from the processing and quality control stages and permission is given for its distribution. All materials created since the initial stages of acquisition are then secured for long-term preservation.

5 LIMITATIONS

Despite more than a half-century of aggressive archival efforts, not all historically valuable digital social science research content has been preserved. Some of the reasons for this have been discussed in the two previous sections. However, there are other reasons for the failure to preserve studies that do not directly apply to either selection or retention. These topics will now be discussed.

5.1 Sporadic Archiving

The archiving of digital data within the social science community is sporadic. There are a variety of reasons for the erratic attention that is paid to preservation. A general problem is the low priority given to data management and preservation. Some individual researchers have been reluctant to deposit their data in archives, either because they want to avoid sharing their data with potential competitors or because they lack the time or expertise to prepare the metadata required for effective sharing. New research projects tend to get more attention than projects that attempt to preserve and reuse existing data, even though the payoff from optimal utilization of existing data may be greater. Institutional data producers may have been under contractual obligations with those who paid for data collection to protect proprietary information. And sadly, some data just “fall through the cracks,” as occurs when data producers do not know where or how to archive their data, or when data collections are forgotten about once the funding has ended and the publications have been written.

Recently in the US, major Federal supporters of social science research, the National Institutes of Health (NIH) and the National Science Foundation (NSF), have initiated policies that encourage and in some cases require grantees to share their data. The impact of these requirements has yet to be seen, but they are not yet strong enough to ensure the preservation of digital social science content. Institutional repositories themselves have not yet been proven to increase the amount of research results made freely available, as well. This may be largely due, however, to a lack of awareness and incentive to deposit.

In the UK, ESRC, and several government departments including the Office for National Statistics, as well as a number of charitable bodies that fund research have all written clauses into their contracts with data collectors that require them to offer the resulting material to the UKDA. In the case of the ESRC, future funding can be withheld from researchers who fail to offer their data, and this has served as an incentive for data collectors to offer their material for preservation.

5.2 Funding issues

Funding issues are by no means unique to the social sciences community. Incentives to promote the depositing of data to the archives have been limited. These incentives must be identified, expanded, and promoted to encourage digital content creators to deposit their research with archives for future access and reuse. The NIH rules, as mentioned above, only apply to their largest grants, those with direct costs that exceeding $500,000 in a single year. Most social science research costs less than that, and therefore does not fall under the obligatory data sharing requirement. In fact, the data sharing rules may create more problems than they solve, because they can lead to a proliferation of Web sites for self-dissemination by researchers, with ineffective long-term preservation. In the UK, ESRC requires any funded data to be offered for deposit with the UKDA, regardless of the size of the award. As in the US, there equally has been an increased tendency toward self-dissemination via Web sites in the UK, which place long-term access and preservation into question.

6 NEW INITIATIVES

The current challenge that data archives face is to find and preserve the most significant digital content not yet archived, and to encourage the owners of this content to make their data available to the archiving community. The long-term preservation of these resources is difficult, if not impossible, for individual institutions to resolve on their own due to the complexity and scale of the task. This challenge is being joined by efforts both in the US and the UK to not only increase the amount of digital data that are being preserved, but also to reach universal agreement on procedures and standards in archiving digital data. These efforts are also encouraging a more formalized network of institutions that will collaborate and create an international network of specialist organizations that can offer advice and archival services to the international research community. This vision is already being realized by the application of established data exchange networks such as that between ICPSR and the UKDA, by the adoption of agreed upon standards such as the DDI, and the work of the Council of European Social Science Data Archives (CESSDA) which encourages cooperation with other international organizations sharing similar objectives. This vision is also being advanced through new initiatives such as those introduced below.

6.1 National Digital Information Infrastructure and Preservation

Recognizing the challenge of preserving digital data, the United States Congress instructed the Library of Congress to lead a collaborative project, called the National Digital Information Infrastructure and Preservation Partnership Program (NDIIPP). The program announcement called for the acquisition of "holdings of historical and cultural materials or information from around the globe that document key social and political developments necessary to understand contemporary events of high importance to national legislators and policy-makers" (Library of Congress, 2003). In response to the announcement, ICPSR has partnered with five other institutions. They are the Roper Center for Public Opinion Research at the University of Connecticut, the Howard W. Odum Institute at the University of North Carolina-Chapel Hill, the Henry A. Murray Research Center at Harvard's Radcliffe Institute, and the Harvard-MIT Data Center, as well as the electronic records custodial division of NARA, to create a project that will identify, appraise, acquire, catalog, and preserve social science research data that is in jeopardy of being lost.

This project will result in the acquisition and preservation of "at-risk" digital social science. It will provide an opportunity for the partners to develop common policies for the deposit and dissemination of data. It will also establish standards for metadata and software/hardware that will ensure that the enormous quantities of data being collected will be maintained in compatible and accessible media. From the different sets of procedures currently used by the partners, as well as a review of procedures used by other organizations, the most efficient and effective that each has to offer will be implemented. In so doing, a "best practices" standard will be put into operation and disseminated throughout the social science research community.

6.2 Digital Curation Centre (DCC)

The Digital Curation Centre (http://www.dcc.ac.uk/) is a UK-based consortium comprised of four partner institutions: the University of Edinburgh (lead partner) and the University of Glasgow, which together host the NeSC, (National e-Science Centre); UKOLN, at the University of Bath; and the Council for the Central Laboratory of the Research Councils. In the interest of securing consensus, the DCC proposed the term "curation" to cover the active management and appraisal of data over the corresponding lifecycle of scholarly and scientific interest. As part of its mission, the DCC undertakes a research program to address the wider issues of data curation and offers outreach and practical services to aid data curators. Launched in November 2004, the DCC recognizes that data have importance as the evidential base for scholarly conclusions, and for the validation of those conclusions. It therefore supports a collaborative network of data organizations in the UK and encourages continued improvements to the quality of data curation and digital preservation.

6.3 E-Science

The overall mission of the E-Science program is to provide a strategic focal point for the identification, development, and delivery of an integrated national research and training program aimed at promoting a step change in the quality and range of methodological skills and techniques used by the UK social science community. ESRC has invested in e-science by establishing this national centre for e-social science, the NCeSS (http://www.ncess.org/), which has a distributed structure comprising: a co-ordinating hub based at the University of Manchester and supported by the UKDA; and a set of research-based nodes distributed across the UK.

The hub is the central resource for e-social science issues and activities in the UK, integrating them with ESRC research methods initiatives and the existing e-science core program. It will provide a one-stop shop for awareness raising, expertise, training, technical infrastructure, data resources, computer facilities, and user support for e-social science research.

The hub will be supported by a number of specially funded nodes. The task of the nodes is to develop grid technologies and apply them to substantive social science research problems. To achieve this, each node will pursue a designated part of a wide-ranging research agenda. Four nodes have already been commissioned; further nodes will be commissioned in 2005. The existing nodes include work on Collaboration for Quantitative E-Social Science Statistics; Modelling and Simulation for E-Social Science; Understanding New Forms of Digital Record for E-Social Science and the Mixed Media Grid. It is hoped that one outcome of the completed work of the nodes will be the greater interoperability between disperse data collections.

7 CONCLUSION

At the core of the archival mission of social science data archives are their preservation procedures, as well as their data selection, or acquisition, procedures. The goal of the archives is to ensure that the digital materials in the holdings (including data and documentation files for both the original materials and any public use files that may be produced) remain accessible, complete, uncorrupted, and usable over time. However, rapid technological change will always threaten the viability of digital materials produced in previous years and under obsolete technological conditions. Innovations in technologies, strategies, and resources have not kept pace with the rapid growth in digital material (National Science Foundation and Library of Congress, 2003). Initiatives like the NDIIPP and the Digital Curation Centre, however, bring us closer to that goal.

Future advances in our ability to preserve an optimal level of digital social science research will depend, in large part, on collaborative efforts such as those presented above. New collaborative efforts across disciplines can enhance the ability to preserve important information in all disciplines, by allowing researches in these fields to learn from each others experiences, and introducing innovative perspectives and ideas. Within disciplines, collaboration must also occur to a greater extent, not only among the researchers themselves, but also with archivists and other information specialists, who can assist in the creation of data collections that can stand the test of time.

8 ACKNOWLEDGEMENTS

The authors would like to thank the Electronic Resource Preservation and Access Network (ERPANET) and the Committee on Data for Science and Technology (CODATA) conference organizers for providing the opportunity to discuss these issues at the Workshop on the Selection, Appraisal, and Retention of Digital Scientific Data, held at the Biblioteca Nacional in Lisbon, Portugal in December 2003.

9 REFERENCES

Anderson, M.J. (1988) *The American Census*. New Haven, CT: Yale University Press.

Converse, J.M. (1987) *Survey Research in the United States: Roots and Emergence 1890-1960*. Berkeley, CA: University of California Press.

Data Documentation Initiative (2004) Homepage of the Data Documentation Initiative. Available at: http://www.icpsr.umich.edu/DDI/codebook/index.html.

De Wolf, V. A. (2003) Issues in Accessing and Sharing Confidential Survey and Social Science Data. *Data Science Journal 2,* 266-74.

Higgs, E. (2003) *The Information State in England:The Central Collection of Information on Citizens Since 1500*. Basingstoke: Palgrave Macmillan.

Library of Congress (2003) Program Announcement to Support Building a Network of Partners. Retrieved June 21, 2004 from the Library of Congress site: http://www.digitalpreservation.gov/index.php?nav=4.

National Archives and Records Administration (2003) Appraisal Policy of the National Archives and Records Administration. Retrieved on June 19, 2004 from the National Archives and Records Administration site: http://www.archives.gov/records_management/initiatives/appraisal.html.

National Science Foundation and Library of Congress (2003) It's About Time: Research Challenges in Digital Archiving and Long-term Preservation. *A Final Report from the Workshop of Research Challenges in Digital Archiving and Long-term Preservation, April 12-13, 2002.* Retrieved on June 21, 2004 from the National Digital Information Infrastructure and Preservation Program site: http://www.digitalpreservation.gov/repor/NSF_LC_Final_Report.pdf.

O'Rourke, J.M. (2003) Disclosure Analysis at ICPSR. *ICPSR Bulletin, Fall 2003,*2-9. http://www.icpsr.umich.edu/org/publications/bulletin/fall03.pdf .

Project on Human Development in Chicago Neighborhoods (n.d.) Retrieved on June 19, 2004 from the PHDCN site: http://www.hms.harvard.edu/chase/projects/chicago/res_pubs/data_archives/index.html.