Top Three Issues in the Long-Term Preservation, Management and Curation of Scientific Data in Digital Form

Henry E. Brady – Professor of Political Science and Public Policy
Director of the Survey Research Center and UC DATA
University of California, Berkeley

There are many problems confronting efforts to preserve and manage scientific data in digital form including making decisions about what to keep, developing plans for discarding data when it is no longer useful, providing adequate meta-data, ensuring long term preservation given frequent changes in media and software, and finding and training staff to do these tasks. I will focus, however, on three problems that are especially pertinent to the social sciences.

1. Linkage of Data Sets – The social sciences are benefiting enormously from the easy availability of large-scale, computerized datasets such as vital statistics, census data, employment records, educational data, welfare and social security records, voting data, medical records, commuting and transportation data, and many other kinds of information. These datasets are even more useful when they can be easily linked together to study events or transitions such as the transition from welfare to work, from illness to a job, from school to citizen, from prison to everyday life, or from home to work. Coding of geographic, contextual, genetic, environmental, and other information can make these data even more valuable for understanding the impact of neighborhoods, institutions, physical distance, individual characteristics, and other factors. Yet these data come in many different forms (different types of databases, different units of observation, and various levels of reliability), and linking them poses significant challenges. If data libraries are to be truly useful for social sciences, they must provide users with the software tools to link these very-large and unwieldy data-sets easily, reproducibly, and reliably.

2. Confidentiality – Although the availability and the linkage of social science data provide tremendous opportunities for answering important social science questions, they also exponentially increase the dangers of disclosing personal information through the possibility of identifying individuals – even though social science researchers are almost always interested in general statements about behavior and almost never interested in individuals. The problem of confidentiality has increased the requirements for Human Subject Reviews, decreased the availability of many kinds of data, and made linkage especially suspect. A number of technical and institutional methods are being developed to deal with these problems, but we are still far from having generally accepted approaches to them. Moreover, although confidentiality has been an especially difficult problem for the social sciences, it is increasingly a problem for the medical, environmental, and even the geo-sciences.

3. Institutional models – One answer to the problems of linkage and confidentiality is to develop better institutional models that provide ways that researchers can have access to data in ways that protect individuals while allowing for extensive data linkage. One example is the Census Research Data Centers which allow researchers to access non-public Census data under rigidly controlled conditions. This model, however, only allows for access on a case-by-case basis, and it does not currently allow for long-term access to data. Institutions are also important for a larger reason: At the moment, we have nothing comparable to the "University Library" which has historically made rational acquisitions through "collection specialists" working with researchers, developed meta-data through classification and indexing, and paid for the development of documentation and the preservation of information. Thus, researchers with data typically do not have any place to go on the University campus, and even if there is an institution concerned with digital social science data, it is typically woefully under-funded and unable to help the researcher archive and preserve data. Some libraries and some computer centers have begun to take up these challenges, but each has other responsibilities and agendas which impede their efforts.