Preserving Quantitative Research-Elicited Data for Longitudinal Analysis. New Developments in Archiving Survey Data in the U.S.

Treffeisen, Jürgen (2004): Archivische Überlieferungsbildung bei konventionellen Unterlagen im deutschsprachigen Raum. Eine Auswahlbibliographie. In: Historical Social Research 29 (4). 227-265.

Trugenberger, Volker (1992): Archivalien-Erschließung mit EDV in der staatlichen Archivverwaltung Baden-Württemberg: das Beispiel Reichskammergerichtsakten. In: Historical Social Research 17 (3). 136-141.

Tuchman, Gaye (1994): Historical Social Science. Methodologies, Methods and Meanings. In: Denzin, Norman K./Lincoln, Yvonna S. (Eds.) (1994): Handbook of Qualitative Research. Thousand Oaks/London/New Delhi: Sage. 306-323.

Van den Nieuwenhof, Patrick (2003): Archivilization of Sciences Archives. New Techniques Making Science Archives Understandable. In: Historical Social Research 28 (4). 242-255.

Vinovskis, Maris A. (1980): Problems and Opportunities in the Use of Individual and Aggregate Level Census Data. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 53-70.

Volkens, Andrea/Bara, Judith/Budge, Ian (2009): Data Quality and Content Analysis. The Case of the Comparative Manfesto Project. In: Historical Social Research 34 (1). 234-252.

Waldow, Florian (2001): The Suggestive Power of Numbers. Some Remarks on the Problem of the Accuracy of Quantitative Indicators in Comparative Historical Research. In: Historical Social Research 26 (4). 125-140.

Weber, Max (1906-1922): Die Wirtschaftsethik der Weltreligionen. Reprinted in: Gesammelte Aufsätze zur Religionssoziologie. 3 Volumes. UTB: Stuttgart.

Werner, Thomas (1991): Transforming Machine Readable Sources. In: HSR 16 (4). 62-73.

Wettengel, Michael (1995): Archivierung maschinenlesbarer Datenbestände im Bundesarchiv. In: Historical Social Research 20. (4). 123-126.

Winchester, Ian (1980): Priorities for Record Linkage. A Theoretical and Practical Checklist. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 414-430.

Witzel, Andreas/Medjedović, Irena/Kretzner, Susanne (Eds.) (2008): Secondary Analysis of Qualitative Data. Historical Social Research 33 (3). 7-214.

Preserving Quantitative Research-Elicited Data for Longitudinal Analysis. New Developments in Archiving Survey Data in the U.S.

Mark Abrahamson, Kenneth Bollen, Myron P. Gutmann, Gary King & Amy Pienta *

Abstract: *»Maßnahmen zur Langzeitarchivierung von Umfragedaten in den USA«*. Social science data collected in the United States, both historically and at present, have often not been placed in any public archive – even when the data collection was supported by government grants. The availability of the data for future use is, therefore, in jeopardy. Enforcing archiving norms may be the only way to increase data preservation and availability in the future.

Keywords: Longitudinal Analysis, Survey Data, Archiving, Secondary Analysis, Data Access, Data Preservation.

1. Introduction

We believe that important research in the social sciences in the 21st century will have either of the following features:

1) it will analyze long-term social processes, covering much longer time frames than we have typically examined in the past, or
2) it will be comparative, examining the same question or issue across a sample of nations.

And the very best research will do both of these. If our crystal ball is right, it will mean a much more important future role for social science data archives. Individual researchers are not likely to be able to assemble the required data sets on their own. If the major archives cannot provide the requisite data, both

* Address all communications to: Mark Abrahamson, The Roper Center for Public Opinion Research University of Connecticut, 369 Fairfield Way, Unit 2164, Storrs, CT 06269-2164, USA; e-mail: m.abrahamson@cox.net. Kenneth Bollen, The Odum Institute for Research in Social Sciences, CB 3355 Manning Hall, University of North Carolina, Chapel Hill, NC 27599–3355, USA; e-mail: bollen@unc.edu. Myron P. Gutmann, Inter-university Consortium for Political and Social Research, University of Michigan, PO Box 1248, Ann Arbor, MI 48106, USA; e-mail: gutmann@umich.edu. Gary King, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge St., K350, Cambridge, MA 02138, USA; e-mail: King@harvard.edu. Amy Pienta, Inter-university Consortium for Political and Social Research, University of Michigan, PO Box 1248, Ann Arbor, MI 48106, USA; e-mail: apienta@umich.edu.

the long-term and the comparative studies that need to be conducted will be frustrated, and social science will not move forward.

If research questions concerning modernization, globalization, gender change, democratization and other key change processes are to be examined in a sophisticated way, with empirical data, it obviously presupposes the availability of data covering long time spans: fifty maybe a hundred years. Maybe more. Where will these data come from? In this paper, we are going to examine the question of such data availability, focusing primarily upon the U.S. with a few comments about other parts of the world. First, we will discuss why the U.S., despite having one of the longest traditions of survey research, has not archived important social science data collections, particularly those that reside in private research organizations. Second, we will show what the actual state of archiving is in the U.S with respect to federally funded social science data collections.

2. The Problem of Missing Social Science Data Collections

The middle of the 20th century was a very important period in empirical research in the social sciences. Studies conducted in every institutional realm provided both:

1) perspectives which greatly influenced the development of social theories and
2) data which provided baselines for ensuing studies of trends in social mobility, attitudes toward health care, religious and political participation, and so on.

In the United States, these studies were conducted in academic departments, in university institutes and centres, and private research firms associated with universities. At the time, very little thought was given to preservation of the data, even when the data were collected with government funding, hence ought to be publicly available. Some of the data collected during these years have vanished, and it is doubtful they will ever be found. Some are on tapes and IBM cards that are in faculty offices, institute storage rooms and warehouses. Their future availability is in jeopardy, and this is going to mean serious limitations for social scientists wishing to do historical research or use these data as base lines to examine long-term trends (Platt 2007).

An interesting example, and an experience that proved to be one of the catalysts for the Data Preservation Alliance for the Social Sciences (Data-PASS) project to be discussed at greater length below, involved Tom Smith's frustration after the Al-Quaida attacks on New York's World Trade Towers in September, 2001. Dr. Smith is a senior investigator at NORC (the National Opinion Research Center at the University of Chicago), a founding director of the General Social Survey, and a long-serving member of the Roper board. He wanted to compare public reactions to the 9/11 attack with those that immediately followed the assassination of President John F. Kennedy in 1963. How did the emotional mood of the nation compare at the two times? Surveys covering the post 9/11 attack were readily available and he knew NORC had collected data after the 1963 assassination. Data to permit the comparison, he initially thought, should not be difficult to obtain.

However, it took over four months for him to locate the 1963 data in NORC's warehouse where they had been stored with limited finding aids. Getting his hands on the data did not end the problem, though, because the data were on IBM punch cards. Smith had to drive boxes of these cards over 800 miles (to New York City) to find a card reader before he could do the analysis. Smith's tenacity was remarkable, but it is easy to understand why, after the experience, he wrote a paper that praised the existence of data archives (Smith and Forstrom, 2001)!

3. The Major Social Science Data Archives in the U.S.

It is the above situation that underscores the importance of data archiving. In the United States, there is a long history of preserving social science data. The Howard W. Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill (http://www.irss.unc.edu/odum) was founded in 1924 is the oldest university-based social science research institute in the U.S., and probably the oldest archive of digital social science data.

In the United States, several other archives have become major centers for preserving social science data over the last half century given the absence of a national data center as is commonly found throughout Western Europe. In addition to Odum, there is the Roper Center of Public Opinion Research at the University of Connecticut (http://www.ropercenter.uconn.edu/) holds the largest archive of public opinion survey data. ICPSR, the Inter-university Consortium for Political and Social Research is the largest non-governmental social science data archive in the world (http://www.icpsr.umich.edu). The Electronic and Special Media Records Service Division, National Archives and Records Administration (NARA) preserves and provides access to permanent electronic records created by or received by the U.S. Federal Government (http://aad.archives.gov/aad/). And finally, The Henry A. Murray Research Center at the Institute for Quantitative Social Science, Harvard University (http://www.murray.harvard.edu/) holds important data collections related to human development, especially longitudinal and qualitative data, and data that illuminate women's lives.

These university-based, independently formed archives (with the exception of NARA) are responsible for trying to preserve the important social science data that have been produced in the U.S. For the most part, these archives have been successful, but there are several types of data collections that their combined collection strategies have missed over the years. This situation stands

somewhat in contrast to the social science data archives in Western Europe that tend to be centralized and funded by the government. For example, in Germany there has been the Zentralarchiv für empirische Sozialforschung (ZA) that has archived and disseminated social science data since the 1970s (www.gesis.org/en).

Historically, the European archives and U.S. archives developed in parallel, with some interaction, but they remain different in their approach to archiving (Scheuch, 2003). The U.S. archives work independently without government support trying to identify and preserve important collections whereas European archives often have the specific mandate to collect the data products of a particular funding body. An example of this is the UKDA (http://www.data-archive.ac.uk/) that archives the social science data collections funded by the Economic and Social Research Council.

4. The Data-PASS Solution to the Problem

To acquire and preserve classic "at-risk" data is an objective of the Library of Congress – which includes the research arm of the U.S. Congress, and the largest library in the world. Three years ago, the Library sought to preserve data and artifacts in many realms; from props that had been used in making motion pictures to data from sociological research. In the social sciences, the Library funded the Data-PASS project, to create a partnership among the major social science archives in the U.S. The objective was to create a long-term and sustainable partnership to preserve digital materials of original social science research. Included among the partners are: ICPSR at the University of Michigan, IQSS at Harvard University, the Odum Institute, the National Archives (NARA) and the Roper Center at the University of Connecticut. For the past three years, members of this group have been cooperating to try to rescue "at risk" content in the social sciences, and to build a foundation to help ensure the preservation of content being created now and in the future.

The initial plan was to identify and then acquire some at-risk social science materials. The further we have gone, the more potential at risk content we have discovered. It is much more than we ever thought existed. We have completed three years of initial funding, and recently began working on an 18 month extension of the original award.

When the studies we were seeking were held by government agencies, we – the archive partners – were typically able to acquire them with little difficulty, complete the archiving process, and make it available to the Library of Congress and to potential users. We have preserved a lot of data in this way. Roper's main contribution here has been to archive studies previously conducted in many parts of the world by the USIA. The USIA part of the story has been very successful – we have made steady progress, and will in the near future have completed archiving these studies. On the other hand, Roper's (and

Odum's) other major objective has been to acquire and archive studies being held by private firms, but we have not been nearly as successful in this endeavor.

5. The Challenge of Private Research Organization Data

In the U.S., *private research organizations* played an important role in data gathering during the middle of the 20th century. Most notably included were: the Research Triangle Institute, the National Opinion Research Center (NORC), Westat and ABT Associates. These firms were involved in a significant portion of governmental and scholarly research in many areas of social science research. Some of the mid-century's most prominent researchers were affiliated with the private firms. However, the U.S. archives had been focused largely on preserving government data and data resulting from university-based research activities. And so, the contract data collections of organizations like NORC, Westat, and RTI were less likely to be archived as a result of collection practices of the U.S. archives. The extent to which this same situation exists in Europe and other parts of the world is not known, but is likely not a problem exclusive to the U.S. As a part of Data-PASS, Odum agreed to work with RTI, and Roper to work with NORC to preserve data from those respective organizations.

Roper and Odum each began to work with the private firms with whom they had been partnered, RTI and NORC, respectively, with a great deal of optimism. Both archives had a long history of working with data producers to acquire and preserve data, and the officials of the private research organizations fully endorsed data preservation as an objective. NORC investigators had conducted a number of classic studies of great historical interest during the 1950s and 1960s. Included were analyses of occupational ratings and occupational prestige, attitudes toward health and illness and the training of health delivery specialists, and so on; many of you are probably familiar with at least some of them. Roper developed a high priority list of 25 studies that were especially important as a starting point. Once those data were archived, the plan was to develop a second high priority list, including the most important of the remaining studies. NORC officials seemed very receptive to the idea of data preservation, so there was ample reason for optimism. Odum had the same reaction after their initial contacts with RTI.

However, several problems arose in both the Roper and Odum partnerships that have prevented a lot of progress. The most vexing problems have been economic. If the data sets could be considered assets, the private firms wondered, what is their value? It they are put into the public domain, what is lost to the firm? This is a very difficult question to answer. In most instances the data sets had simply existed, unused, in storage, but suddenly someone wanted them. Did that imply that they had value? Given the business climate in which

these firms now must operate, these were reasonable questions for them to raise. Thus, although the Data-PASS partnership thought it was eliminating the economic barrier to archiving these older data (by offering to cover the costs of archiving the data), the for-profit survey firms still associated archiving the data publically with potentially lost opportunity costs.

Furthermore, the business model followed by the private research organizations requires that the time of all officials be billed. When Roper first approached NORC or Odum approached RTI, it was not even clear who were the relevant officials. No one in either organization had preservation of old data in their job description. We each went through several vice presidents trying to find the right one. Helping the organizations to identify who could take responsibility for the task took longer than one might expect. And once library and warehouse people whose hours could be billed were identified, we discovered that the costs of locating and recovering the at-risk data were greater than originally budgeted in the Data-PASS project.

In the U.S. archives, similar to the experience in European archives, the data creator typically bears little or no cost for data archiving activities which include preparing the data and documentation for long term use and preservation. However, the data creator is typically responsible for locating, organizing and providing the basic set of data and documentation upon which an archive builds and enhances the files it preserves. In this case, the U.S. archives secured funding from the Library of Congress to offset the extraordinarily large costs of locating and recovering files that had been long ago abandoned. However, the resources available were inadequate for the large costs associated with the staff time at the private research firms we interacted with.

When data were collected in the middle of the last century, they were often placed haphazardly in warehouses, or left in the back of institutes, with little thought given to preservation or future use. There were no clear maps or finding aids that would lead to the data sets we wanted to preserve. Our strategy was simply to pay the labor costs for people actually to locate the data where possible.

One the largest challenges facing data archives today is the protection of the confidentiality of the study participants themselves – minimizing the possibility of reidentification through the archived data. Although this is a common reason that archives have a difficult time getting data owners to agree to archiving, this and other more common issues offered little interference in our attempts to work with private research organizations.

Three years later we have learned a great deal about the problems and prospects of partnerships between public data archives and private research organizations (see also Crabtree, Maynard, and Timms-Ferrara, 2007). And we are still optimistic that we will overcome the obstacles and archive the at-risk data. However, if one asks to see tangible results of our efforts and expenditures with the private research organizations to date, we have little to show at this time.

6. Federally-Funded Research Data in Archives (and not)

Next, we review our work preserving more contemporary studies; this work is being carried out under the leadership of one of the Data-PASS partners, ICPSR at the University of Michigan. Amy Pienta, Myron Gutmann and their colleagues wondered how much publicly funded social science research is being publicly archived. This includes studies funded by the National Science Foundation (NSF) and the National Institutes of Health (NIH). These are the two agencies of the U.S. government that fund the most social and behavioral science research.

Over the last two decades, both of these government agencies implicitly and explicitly (most recently in written policy statements) began to indicate in grant awards that they expect investigators to make their data available to the research community, in a timely way. This ordinarily means placing the data in an archive, after the principal investigator's initial use of the data. Specifically, they state that all proposals must contain detailed plans for how and where data will be stored. Professional associations – such as the American Sociological Association (ASA) and the American Psychological Association (APA) – also state in their ethics and/or research guidelines that data should be made available to others in a timely manner.

Apart from the fact that research data collected with public funds ought to be in the public domain, scientific paradigms are built on the assumption of replication which requires others' access to original data. In addition, science is, in part, defined by its cumulative nature, and when data are readily available it makes it much easier for investigators to build on previous work. Further, some of the most useful data sets are very large longitudinal studies. Because they are so large, no single researcher or research team is able to fully utilize the data. It is inefficient not to permit many investigators to bring fresh perspectives to the data set, in order to fully mine it. So, given all these reasons why investigators receiving public research funding, in particular, ought to be putting their data into public archives, to what degree has that been happening in the U.S.? Throughout much of Western Europe (Germany, Switzerland, the UK, and so on), archiving of data acquired with government funds is mandated and specifically funded. They manage to include a very high percentage of publicly funded studies. Furthermore, these Western European archives make substantial effort to include privately funded studies as well. How well does archiving in the U.S. compare?

Despite norms and expectations for data sharing, researchers give many reasons why their data are not publically archived. Researchers cite reasons such as the cost (time and financial) of creating basic documentation and data files, concerns about getting "scooped" by other researchers interested in the same topics, the need to protect the confidentiality of the respondents, and feeling responsible for errors in the data and documentation. And, given that most of

the data sharing expectations in the U.S. have been either implicit and/or not enforced, scientists have largely not been held accountable for archiving their data. Thus, ICPSR wanted to enumerate how much federally-funded social science data in the U.S. have not been publicly archived.

The funded research portfolios of both NSF and NIH were examined for grants that might have involved primary data collection in the social and behavioral sciences. The objective was to exclude non-research awards: training grants, funds for conferences and workshops, and so on. Most of the examined awards were made between the mid-1970s and 2006. There were a total of nearly 11.000 awards in the pool and, and screening all of them turned out to be a very tedious task, and one in which there was often incomplete information. The research team concluded that they almost certainly made some errors in screening.

From the records of granting agencies and other professional associations, the research team then searched for principal investigator's e-mail addresses for all of the studies that appeared to involve primary data collection and they sent questionnaires to the principal investigators at those addresses. They asked principal investigators whether they had actually collected data (sometimes they planned to, but failed) and, if so, whether their data were subsequently given to any public archive.

The results indicated that fewer than 350 of the awards that produced data had been archived – that was 20%, despite the explicit expectation that the data be archived. The investigators in almost one-half of the non-archived studies claimed that they still access to a copy of the data. This includes over 800 studies that are "at risk" of being lost, but could yet be archived. The principal investigator's reported that the non-archived data were stored in diverse media. Most were on hard disk or magnetic tape, but few survived in paper only collections. In a few cases studies in this category had been made available on personal or departmental web sites. The Data-PASS partners share a long-term plan (hope?) to see the surviving data placed in the archives of one of the Data-PASS partners before they are lost.

We have seen that nearly one fifth the publicly funded studies that generated primary data were archived, and roughly one-half were not, but still could be – what about the remaining studies? They are, in effect, lost. Investigators said they might have been left in their former offices; they were lost when they moved; their co-investigators might have them (though they did not); and so on. They said everything except their dogs ate them. These data are simply gone, despite the explicit archival expectations of funding agencies.

Part of the problem is the culture of social science in the U.S. which claims to value making data publicly available in a timely manner, but apparently does not value it very strongly. Part of the problem lies with the funding agencies that do not attempt to employ any follow-up sanctions when investigators fail to archive their data. In fact, the agencies typically do not follow up at all!

Expectations without punishments has produced the kind of situation that the classical French sociologist, Emile Durkheim, described when he contended that if there were no punishment, there was no norm!

We are going to have to do better in the future. One way to proceed is working to change the culture of social sciences so that routine archiving becomes more valued. To be realistic, however, we may also have to consider sanctions in order to make archiving a norm, in Durkheim's sense. Perhaps that means investigators who receive public funds, but do not archive their data in a timely way are not eligible for public funding in the future. There are many difficult questions that would have to be resolved before any such policy could be put into place; but pertinent discussions must begin because every day we wait, more data is at risk of being lost.

References

Crabtree, Jonathon / Maynard, Marc / Timms-Ferrara, Lois (2007): "Developing Partnerships in the Social Sciences" Presented at the e-Social Science Conference, Ann Arbor, MI.

Pienta, Amy M. / Gutmann, Myron/ Hoelter, Lynette/ Lyle, Jared/ Donakowski, Darrell (2008): "The LEADS Database at ICPSR" Presented at the American Sociological Association meetings, Boston, MA.

Platt, Jennifer (2007): "Some Issues in Comparative, Macro and International Work in the History of Sociology" In: American Sociologist 38, 4: 352-363.

Scheuch, Erwin K. (2003): "History and Visions in the Development of Data Services for the Social Sciences" In: International Social Science Journal 55, 17: 385-399.

Smith, Tom / Forstrom Michael (2001): In Praise of Data Archives. In: IASSIST Quarterly 25 (Winter): 12-14.