How Safe is Safe Enough when we Preserve Social Science Data?

Myron P. Gutmann[1], Nancy Y. McGovern[1], Bryan Beecher[1], T.E, Raghunathan[2]

[1]Inter-university Consortium for Political and Social Research, Institute for Social Research, University of Michigan

[2] Department of Biostatistics, University of Michigan and Survey Research Center, Institute for Social Research, University of Michigan

Email address of corresponding author: gutmann@umich.edu

Abstract. Those responsible for digital preservation are aware of a tension between the need to expend resources on preservation and the scarcity of those resources. Ideal preservation would save many copies forever, but this has a large potential cost. We need to be certain that we are preserving the right number of replicas. The paper raises issues that derive from a core attribute of most social science data, which is that social science data is often created by drawing random samples from a population and studying the behavior or attributes of the sample. The sampled character of these data has implications for preservation. While it is less than desirable to lose cases from a sample, even after some loss the sample still has validity and can be used for future research. From this the paper argues that replication for preservation purposes may require thinking at the level of cases or variables and not entire data files. There may be varying numbers of replicas within a data file, depending on the attributes of the overall sample, and the attributes of cases and variables. The situation is also more complex because of the need to protect to protect confidentiality of data.

Introduction

Researchers and practitioners in the area of digital preservation are increasingly aware of a tension between the simultaneous need to expend resources on preservation, and the scarcity of those resources. Ideally, we would maintain many tens of copies of the digital objects that we are charged with preserving, and we would ensure that every bit of information in those objects is preserved for an indefinite period of time. Even with very substantial resources, however, preserving large bodies of digital data is very difficult, as David Rosenthal (Rosenthal and Reich 2007) demonstrated in a recent presentation. At the same time, our resources for digital preservation are always limited. In the world of research, the tension takes the form of the trade-off between the cost of acquiring new data, the cost of preserving all the data already acquired, and the cost of staffing the research enterprise that analyzes both older and newer data. In effect, resources we expend on preserving data inevitably come at the expense of collecting new data and analyzing data. There is no easy resolution. For social science data, where data sets have not historically been very large (in comparison in those assembled for genomics, space science, or earth science, for example) this is an issue as we consider building networks of replicated storage and move to a situation of ever-larger digital objects in the form of video recordings of behavior and social interactions. Given the cost of replication schemes for petabyte-scale archives or networks of archives, the question of just how much data should be preserved, how many copies should be made, and how those copies should be judged is of great importance. How safe is safe enough?

This paper begins to set out the boundaries of the questions and answers needed to understand how much and how well preservation should be done for social science data, the area in which we have experience. In effect, the paper raises issues that derive from a core attribute of most social science data, and the implications of that attribute for digital preservation: that attribute is that a large proportion of social science data are created by drawing random samples from a population and studying the behavior or social (or economic) attributes of the sample in order to draw conclusions about the whole population. Even where the data are not explicitly a sample, or not really random, we will argue that the sampled character of social and behavioral science data has implications for preservation. Those implications come from the realization that while it is less than desirable to lose cases from a sample, even after some loss (from partially failed preservation or from other failed attributes of the social science enterprise, for example invalid responses, failure to answer individual questions, refusal to participate, or loss to follow-up), the sample still has validity and can be used for future research. We argue that only when we draw these conclusions can we make effective decisions about preservation strategies for social science data, and that these conclusions also may have relevance for other data preservation activities.

Social Science Data

The ideas described in this paper build on our experience preserving and disseminating quantitative digital social science data. We begin by describing some of the norms and practices of social science data collection, research, and preservation, and then turn to the relationship between preservation and analysis. For the purposes of this discussion we restrict ourselves to social science data that are quantitative in nature -- as opposed to qualitative -- and that are designed to be analyzed using statistical methods. While qualitative data are an important part of social science research, they require a separate discussion. We also restrict ourselves to micro-data here, where the data reflect the responses to questions made by individuals, families, or institutions, rather than aggregated summaries of individual data that can be represented in tables. One set of examples for understanding the distinction between microdata and aggregate tabular data is through the data produced as part of the United States Census of Population (http://www.census.gov/main/www/cen2000.html), where we can contrast data about individuals in the Public Use Microdata Samples [PUMS] (http://www.census.gov/Press-Release/www/2003/PUMS.html) and the data tabulated to create tables about geographic areas in the Summary Files (see for example Summary File 1 at http://www.census.gov/Press-Release/www/2001/sumfile1.html).

A significant proportion of quantitative social science research makes use of data about individuals, families, or institutions (e.g., businesses, schools, hospitals), that are collected by asking those respondents questions about their attributes, behaviors, and attitudes. The Census, which we have already mentioned, asks about demographic attributes (age, sex, marital status, number of children, migration history), employment, place of residence, characteristics of housing, among other things. Social surveys (broadly defined to include surveys of social structure, attitudes, behavior, economic characteristics, and many other things) ask many kinds of questions, most of which produce either categorical or continuous responses (and sometimes qualitative responses, which we are not concerned with here). The data files that are produced from these collection activities are then subjected to various kinds of quantitative analysis.

Three attributes of social science data files and their analysis are worth noting here. The first is that the research is usually done on samples designed to be representative of the population, rather than the whole universe itself. Administrative data such as the Census, of course, are designed to cover the entire universe, but a large proportion of research done using Census data are done with samples (Anderson 1988). This is so for two reasons. First, the Census only asks the entire population a small number of questions (the "short form"), while they ask a much larger number of questions (the "long form") of a sample population. Second, the data that the census makes available for public use in the PUMS are a subsample of the entire long form population. More generally, most social surveys are conducted on a sample drawn from a national population, because it would be impossibly expensive to ask questions of the entire population and because 75 years of research (Converse 1987) has shown that it is possible to draw significant conclusions about the whole population from research on a sample of that population. It is also the case that each sample case may not represent the same proportion of the original population as all others, which means that each case may or may not have the same "weight." In a simple example, the Census PUMS files represent one percent or five percent of the total U.S. population, giving each of their cases a weight of 100 or 20, respectively. But other sample designs might include different weights for different cases, with some representing 5 individuals in the underlying population, some representing 20, and still others representing 100 (Kish 1965).

Second, these data are often analyzed using inferential statistics. What this means is that analysts use methods that ask whether a statistical relationship is true for the sample population, and if it is whether it is strong enough ("significant") for a larger population, given that it is true for the sample. This analytic strategy has at its starting point the fact that the data are a sample, and that inferences are to be drawn. What matters is that the sample is constructed properly, that cases are weighted in an appropriate way, and that the statistical method is appropriate for the data, the issue under analysis, and the hypotheses to be tested. That the sample does not include every potential respondent in the universe is taken for granted. Moreover, methods for dealing with missing data are well developed (Raghunathan 2004). The broad conclusion to draw is that it is usually possible to draw valuable conclusions from quantitative social science data even if those data are sampled or otherwise incomplete.

We can visualize quantitative social science data as operating at four levels. At the highest level is the population universe. Beneath that is the sample population of individuals who have been chosen as survey respondents through some selection process, and then agreed to provide data. Within that sample are individuals, each of whom has a sample weight within the population. Finally, we have a series of individual data items for each respondent. The data used for any given social science research project are typically a subset of those individual data items for all or a subset of individual respondents, weighted to reflect the entire population. The success of the analysis is partly a function of the completeness of that body of data, but much more a product of the success of data collection, the representativeness of the sample, the appropriateness of the methods used for analysis, and the inferences that can be drawn. Put another way, having all the data items about all the individuals in the population is much less important than having a well-collected sample and appropriate analysis.

The third attribute of social science data worth noting is that an increasing fraction of data preserved in an archive has attributes that require protection of confidentiality. The requirement to preserve confidentiality may place limits on the extent to which copies of the data can be located away from the archive that has accepted responsibility for the data.

These attributes of quantitative digital social science data guide our understanding of their preservation. Virtually all useful data are samples, which means that they are by definition incomplete. Second, they are analyzed using statistical methods that take into account their nature as samples and include procedures for dealing with missing data. These methods are designed to ensure that researchers draw valid inferences from the data, which is what makes them useful. Third, the need to protect confidentiality may put limits on the number of copies that can be located away from the archive that has accepted responsibility for preserving the data. In the next section we will expand upon these ideas to argue for a preservation strategy that operates at a level of granularity appropriate for the data and research -- the variable, so that if some variables are lost or missing the rest may still be useful.

Preservation Standards for Social Science Data

Quantitative social science researchers and organizations have a history of activities that promote the preservation and sharing for secondary use of research data (Clubb et al. 1985; Scheuch 2003). Beginning with the creation of the Roper Center for Public Opinion Research in 1946 (opened to the public in 1957), archives for the preservation and sharing of social science data have been established widely and broadly used. ICPSR, where we are involved in preservation activities, was founded in 1962 and is one of these long-lived archives.

For much of the 60-year history of social science data archiving, the dominant preservation strategy was to maintain copies of data files on magnetic tape, replicated locally. At ICPSR, that is exactly what we did: until we began to deliver data over the internet in the mid-1990s, we maintained our preservation copies of data holdings on multiple copies of magnetic tape, with one copy local for ease of copying and sending to users, and one copy off-site. When we moved to on-line data delivery we continued the tradition of dual copies on tape in addition to a copy on disk for distribution, with one copy on-site and one off-site for security. No standard approach to preservation systems and preservation metadata accompanied these ad hoc tape replication systems. Every archive had its own strategy for documenting where and how data were preserved.

Only recently have digital repositories begun to develop standard ways to understand the preservation process and preservation requirements. The publication of the Open Archival Information Systems (OAIS) Reference Model in 2001 opened the door to systematic approaches, a process that coincided, as McGovern shows (McGovern 2007), with the construction of the first systematic digital preservation policies for archives and other digital collections. The process of reconciling earlier practices with the benefits of the OAIS approach has not always been easy, but archives such as ICPSR have made progress (Vardigan and Whiteman 2007).

If compliance with the OAIS is the first step on the road to improving data preservation, a second important step is finding new ways to think about preservation. New research on preservation has emphasized the variety of ways that digital content is at risk. Baker et al (2006) enumerate a long list of threats, starting with disaster, human error, component faults, media faults, and obsolescence, and concluding with attack, organizational faults, and economic faults. The risks are great, and despite their variety (especially of human and technical causes) they can be summarized as loss of a large-scale storage system, loss of an entire media device (a whole disk drive or storage tape), loss of an individual file that resides on a media device, and loss of some number of bits within a file (Baker et al. 2006). In addition to loss of bits, there are well-known examples of “bit-rot,” where bits are readable on a disk but the information recorded on them apparently changes, so that a Black Female because a White Male, a transformation not revealed without effective audit procedures.

In response to those threats, there are well-described strategies for ensuring digital preservation, with the most common emphasizing replication (multiple copies) to prevent media loss, migration or emulation to insure against obsolescence, and audit systems to ensure that losses are noticed when they occur and corrected before they prevent recovery. For this discussion we will concentrate on the role of replication, with some comments about audit processes.

What research about preservation-related replication shows is that the number of copies required to ensure preservation is larger than the long-time requirement of two or three copies (Rosenthal et al. 2005; Baker, Shah et al. 2006). One reason for this is that more and more digital archives have shifted their core preservation and replication processes to disks from tapes, which has required them to rethink their operations. How many copies are necessary when data are held on disk rather than tape? And where should those copies be located?

Replication requires multiple copies, with several of them located away from the primary archival location. Our strategy at ICPSR, where we have moved our preservation platform from tape to disk, is to keep one copy on high-density tape and multiple copies on disk, synchronized with the San Diego Supercomputer Center's Storage Resource Broker (Baru et al. 1998) and Chronopolis systems. We are also in the process of working with other social science data archives in the United States to build a private LOCKSS network, using Stanford's LOCKSS software (http://www.lockss.org/lockss/Private_LOCKSS_Networks), and following the model of the MetaArchive for Southern Culture (http://www.metaarchive.org/). The LOCKSS system advocates a minimum of six (preferably seven) replicas, which is significantly more than past preservation strategies. It also ensures through its polling system (Reich and Rosenthal 2001; Reich and Rosenthal 2004) that attention is paid to audit of preserved digital content, which is vitally important for knowing when digital content has been lost and how it may be restored from other replicas.

Beyond replication, our strategy for maintaining sufficient redundancy relies on a mechanism for encoding data that allows a high level of granularity, so that the record describing a single case's responses is separately recorded and is capable of extraction from the record. One way to visualize this is to compare the attributes of an individual in a social science data file with a digital photograph of that individual. In the data file we can separately determine age, sex, marital status, employment, educational attainment, and so on. Even if one or several of those attributes is lost because of a partial preservation failure, the rest are available. In the case of the photograph, the image is a whole, and if part of the digital representation is lost, it may not be possible to recognize the image at all.

We attempt to assure this variable-level preservation in two ways. First, where possible, we organize our canonical preservation copies of data in non-proprietary formats. This means, for example, that data are stored as rows and columns with ASCII encoding and a plain text description of the structure of the data file, and not in one of the proprietary formats used by the major statistical analysis programs (e.g. SAS, STATA, SPSS). This is not always possible, of course, because some data are structured in ways that make a non-proprietary format extremely difficult (for example spreadsheets and relational databases). Second, we do not compress our preservation copies, because use of compression algorithms reduces the possibility of recovering part of a data collection where bit-level failure takes place and partial recovery is possible.

Despite these well-established and successful approaches to preservation of quantitative social science data, much remains to be done. We're eager to develop the field while using our experience and holdings as a test-bed for new approaches. In this paper we want to use the points we have made about social science data and preservation to raise new questions. To be specific, we ask five questions, which derive from our earlier discussion of the attributes of social science data and how they are created. The point of the questions is to emphasize that we need to find new ways to think about replication that are more sensitive to the scientific characteristics of the data being preserved, more attuned to what can be accomplished through audit processes, and hopefully less expensive.

1. What is the right level of granularity for replication of digital objects? Is it the data file, the respondent case, or each variable for each case? If one of the objects smaller than the data file is the right level of granularity, is it appropriate to have different numbers of replicas for different objects?

2. Is it necessary to enlarge the number of replicas because there are some objects that are more important than others for analysis of a data set, for example because they have very large weights or because certain variables have outlier values or have high theoretical importance?

3. Is it possible to reduce the number of replicas because we know that the data are a sample, and effective analytic inference is possible in a smaller number of cases or variables?

4. If it is appropriate to have different numbers of replicas for different objects, may some of those replicas be on the same device, or must all replicas always be on diverse devices?

5. How does confidential data shape a replication strategy, both in terms of numbers of replicas and the need to control the location of replicas?

In the remainder of our paper we attempt to answer these questions, and in so doing set out a research agenda for better understanding preservation of quantitative social science data.

New Ways of Thinking about Digital Preservation for Social Science Data

Not all social science data are equal, and therefore not all data require the same preservation strategy. We want to find the most efficient way to ensure the long-term preservation of those data, while ensuring their usefulness for analysis and protecting the confidentiality of respondents. The five questions we enumerated in the previous section allow us to create a multidimensional matrix to identify the ways that we should think about replication.

1. <u>Level of granularity</u>. Given the nature and structure of social science data, the ideal level of granularity for preservation is probably the case. While there are variables that may be more significant than others, the critical area for replication in sampled data is the individual respondent. Having said that, any preservation strategy should be capable of identifying objects as small as a variable within a case, in order to ensure that the largest possible number of objects can be recovered from a partially lost replica. We recognize that audit practices will need to be refined to manage a process where the granularity is smaller than what is currently considered a replicable unit, which now is primarily a file consisting of data about a number of cases.

2. <u>Are there objects for which we need more replicas?</u> We argue that more replicas would be needed for some cases in a sampled data collection where weights differ between cases. We can imagine a scale where cases with large case weights would be replicated

more than those with smaller case weights, perhaps with a linear relationship (so a case with a weight of 20 would be replicated four times as much as a case with a weight of 5).

3. Can we reduce the number of replicas for sampled data? If the potential unit of loss is the case (and not the entire study), and if replication and audit can operate at the level of the case or an even smaller object, then sampled data require fewer replicas than for data that includes the complete universe. This is so because all sampled data assume analysis that draws effective inferences, which can compensate for sample size, sample weights, and missing data.

4. Must all replicas always be on diverse devices? This question is important because if we visualize replicas at a scale smaller than a complete study (made up of many cases each with many variables), then one can suggest that within-case (where we have multiple replicas of a single variable for each case) and within-study (where we have multiple replicas of a single case within a study) replication could occur on the same device. This would ensure against failure at the bit-loss level, but not necessarily at the device level.

5. How does confidential data shape a replication strategy? It is essential that the number of copies and number of users of confidential data be carefully controlled, which would suggest that fewer replicas should be maintained. It is also much more difficult to have a public replication scheme (of the type used in both public and private LOCKSS networks) with confidential data, because of the potential that one of the external replication servers might be compromised. The solution is to keep the number of replicas the same as they would be for public use data, but to keep those replicas in carefully controlled settings, probably on diverse devices within the physical premises of the responsible archive or archives that are connected by secure communications.

We can summarize our findings in the following way: social science data preservation lends itself to a strategy that makes use of replication but that increases the granularity at which replication takes place. This increased granularity allows us to create a strategy where there are more replicas for some variables than others and more for some cases than others. On the other hand, the fact that data are sampled allows there to be fewer replicas than might otherwise be required, because effective inference is possible with missing cases and missing variables for some cases. Such a strategy would also make use of a diverse replication strategy in which multiple copies of some objects are on the same device, and replication across devices and locations also takes place. Finally, confidential data must have the same number of replicas at the same level of granularity as public data, but replicas of confidential data must be more carefully controlled, either all residing within the responsible archive or distributed among archives with similar responsibilities using secure communications.

What do we gain?

In this paper we have laid out a new approach to digital preservation that takes into account the ways that quantitative social science data are constructed and used by researchers, and managed by repositories. What we have offered is a new way to think about replication that increases the granularity to the case or variable level, takes into account the fact that data are sampled, and protects confidential data.

This new way of thinking about preservation is not without costs. We need to create new tools to manage preservation at the granularity that we suggest is necessary. These tools will create preservation metadata for the within-case and within-study replication that is necessary, so that we can explain to data users what is happening, extract analytic files from preservation copies, and audit preservation replicas at the new level of granularity. We also need to study the usefulness of this change in the way that we think about preservation replicas. We envision undertaking simulation studies based on known data about the likelihood of various kinds of preservation failures, given our ideas, to see whether these ideas would provide greater assurance of long-term preservation. We also envision simulation studies of the potential to draw useful analytic inferences from data with various levels of replication granularity and various levels of replication failure.

These potential costs offer a strong potential for benefits, which is how we choose to conclude. Social science data have their own characteristics, and the more we understand them as we design preservation systems, the more successful our preservation systems are likely to be. In estimating the number of replicas required for preservation, the approach we suggest is likely to get us much closer to the right number than any other we know.

References

Anderson, M. J. (1988). *The American census: a social history*. New Haven: Yale University Press.

Baker, M., et al. (2006). A fresh look at the reliability of long-term digital storage. *Proceedings of the 2006 EuroSys conference*. Leuven, Belgium**:** 221-234.

Baru, C., Moore, R., Rajasekar, A., Wan, M. (1998). The SDSC storage resource broker. *Proceedings of the 1998 Conference of the Centre for Advanced Studies on Collaborative research*. Toronto, Ontario, Canada, IBM Press.

Clubb, J., Austin, E., Geda, C., Traugott, M. (1985). Sharing Research Data in the Social Sciences. *Sharing Research Data*. S. Fienberg and M. Straf. Washington, D.C., National Academies Press**:** 39-88.

Converse, J. M. (1987). *Survey research in the United States: roots and emergence*. Berkeley: University of California Press.

Kish, L. (1965). *Survey sampling*. New York: John Wiley and Sons, Inc.

McGovern, N. Y. (2007) "A Digital Decade: Where Have We Been and Where Are We Going in Digital Preservation." *RLG Diginews* **11**, http://www.rlg.org/en/page.php?Page_ID=21033#article3

Raghunathan, T. E. (2004). "What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data." *Annual Review of Public Health* **25**(1): 99-117.

Reich, V., Rosenthal, D. S. H. (2001) "LOCKSS: A Permanent Web Publishing and Access System." *D-Lib Magazine* **7**, 10.1045/dlib.magazine: 10.1045/dlib.magazine http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/june01/reich/06reich.html

Reich, V., Rosenthal, D. S. H. (2004). "Preserving today's scientific record for tomorrow." *BMJ* **328**(7431): 61-62.

Rosenthal, D. S. H., Reich, V. (2007) "Can We Afford to Preserve Large Databases?" *Coalition for Networked Information, Spring 2007*, http://www.cni.org/tfms/2007a.spring/abstracts/PB-can-rosenthal.html

Rosenthal, D. S. H., et al. (2005). "Requirements for Digital Preservation Systems: A Bottom-Up Approach." *D-Lib Magazine* **11**(11).

Scheuch, E. K. (2003). "History and visions in the development of data services for the social sciences." *International Social Science Journal* **55**(177): 385-399.

Vardigan, M., Whiteman, C. (2007) "ICPSR meets OAIS: applying the OAIS reference model to the social science archive context." *Archival Science*, 10.1007/s10502-006-9037-z: 10.1007/s10502-006-9037-z