

The Less Obvious Functions of Archiving Survey Research Data

WARREN E. MILLER
University of Michigan

The archiving of data produced by survey research differs in kind from the archiving of other basic research materials only in some very limited senses. The generic meaning of archive is of course attached to repositories of public records. The classic definition emphasizes the storing of historical documents and refers to collections of material originally produced by "natural processes" quite unrelated to the scholarly mission to which archived materials may be subsequently devoted. Archives of survey data, on the other hand, are collections of material that come into being only to serve someone's original research needs. Archived survey data have in common with the materials of a more traditional archive the status of primary evidence. Archives of survey research data thus differ fundamentally from reference libraries insofar as the library is a collection of derivative, secondary source materials. Of course, where the library is a research library its holdings may consist predominantly of original source material and if the researcher's objective has to do with a topic such as the sociology of knowledge, even the books of a public library containing the results of research may become original sources of evidence.

AMERICAN BEHAVIORAL SCIENTIST, Vol. 19 No. 4, March/April 1976
© 1976 Sage Publications, Inc.

Insofar as archives of survey research data are like other research libraries, their impact on the scholarly community is, in principle, unique only to the extent that a somewhat different community of scholars is served by the archive. Thus, what can be said about data archives and survey research doubtless has many parallels for other communities of scholars served by other repositories of original research materials. Nevertheless, so much that has been written about the new survey research archives has emphasized their technology or their particular content that it seems useful to emphasize some of the less obvious functions that are served by their existence.

The ultimate impact of the development of archives of data produced by survey research is difficult to predict. The growth of archives has been rapid and has been compressed into a period of little more than a decade. The great variety of archives now in existence confounds most attempts to generalize about the contribution of the archival movement. Nevertheless, there are perspectives that seem reasonably well-founded and that speak to at least two sets of likely consequences of the continued evolution of this new resource for the scholarly community.

I. DATA ARCHIVES AND THE SCIENTIFIC COMMUNITY

One of the major consequences of the existence of archives of survey research data concerns their impact on relationships which individual scholars enjoy with each other. Whatever subsequent developments and technology may do to the organization, content, and style of archival activity, it seems certain that members of the research community will be related to each other in a manner that would not be possible in the absence of the archival institution.

One of the most obvious ways in which relationships among scholars are affected relates to the degree to which the archive facilitates collegial behavior. The somewhat romantic concept

of the invisible college of scholars is being replaced by an equally exciting but probably much more real and visible collectivity of scholars who are knowingly joined together intellectually by virtue of shared access to archived bodies of data central to their common intellectual endeavors. This is crucially important to the social science community because of the nature of the phenomena which the social scientist seeks to understand. Although the analogy should not be pushed too far, the existence of the archive gives the community of social scientists the opportunity for common experience that natural scientists can find within their laboratories. The entire gamut of basic experimental observations for chemistry or physics can be replicated in hundreds of laboratories equipped with standardized materials and instrumentation. In an earlier period, the social scientists' equivalent was found in the ubiquitous class of undergraduates who became the subjects of experiments, or in the multitude of communities that served as hosts for systematic observation. Unfortunately, both undergraduate classes and local communities are very limited sites from which to explore many generalizations that bear replication, whether for purposes of basic research or of science education. Today common access to bodies of data that are substantially less limited in their scientific utility means a standardization and a commonality of research experience among widely separated scholars. As bodies of data relevant to particular research topics grow, each of the members of a far-flung constituency can keep pace and maintain contact and intellectual rapport with each other through their common access to the empirical base of their work.

The evolution of the data archive has been accompanied by the evolution of analytic techniques and perspectives relevant to the data. Thanks to the rather phenomenal homogeneity of the world of computers, the spread of analytic perspectives has come close to keeping pace with the flow of data in their diffusion across academic institutions. From the earlier days of the first statistical package to the current scene in which the

OSIRIS or, more recently, the SPSS systems are so widely distributed, access to computers and computer software designed to execute complex data analyses has more than kept pace with the availability of archived data. For substantial bodies of research, the days of the lone scholar with a unique data set and a limited analytic capacity are a thing of the past. They have been superseded by a time in which hundreds of scholars have common access to the same bodies of data that can be manipulated by all with a rich variety of alternative methods and techniques.

The facilities of the data archive have also vastly accelerated the pace of that phase of the scientific enterprise in which the private insight of the scientist becomes a public commodity. The tensions and the complementarities between the highly private and personal act of discovery and the public critique, replication, and reconfirmation are in no fundamental way altered by the existence of the archive. However, those parts of the scientific enterprise which are social now develop at a vastly accelerated pace. The presentation of alternative interpretations, the discovery of flaws in theory or logic, or the reaffirmation through replication, can all occur with a swiftness and an ease previously denied because of the barriers separating the private from the public aspects of science.

This is not to say that all barriers have crumbled immediately. The utility of the archive is often still limited by the inclination of the individual scholar to retain his or her monopoly on information. When the survey data archive movement began, one of the largest obstacles to the acquisition of important bodies of data rested on the well-established norm that a researcher's primary data base was properly regarded as personal property against which others had no claim. In a very real sense, information was power and often the scientist who had managed to accumulate a store of information derived substantial status from the sheer fact of the accumulation. Of course, the transformation of the data collection process from a totally private affair of the individual scholar to the organized institutional setting funded by public monies has also played a

large role in the norm. As a body of data is generated through the existence of a public institution supported by public funds, it becomes reasonable to presume that the data must ultimately pass into the public domain, accessible to colleagues and professional competitors. That principle acquires real meaning, however, with the existence of an archival capacity that can implement the ideal and move the data from the scholar's personal research files on to computer tapes that can be copied and distributed to as many scores of interested scholars as may exist.

Within an amazingly short time, new generations of scholars have indeed come to presume a new norm of access to data which at times threatens all but the most pure of intellectual motives that might lead a principal investigator to launch a data collection. At the extreme, the argument articulated by the most aggressive is that all data should pass immediately into the public domain and be made available to interested scholars wherever they may be. The argument is made that the principal investigator has an intrinsic advantage over all others because of the prior thought which shaped the data collection and the readiness with which the principal investigator can, therefore, move to exploit the data in accord with the stated research objectives. With this line of reasoning, even the most aggressively competitive of the external colleagues starts with some disadvantage in the race for prior publication, and that disadvantage should be deemed large enough to support the principle of immediate release of data.

A somewhat more experienced view takes account of those instances in which the principal investigator's timetable is knocked askew with unexpected delays in the execution of even the best formulated research plans. If, as is often the case, the analysis does not proceed with perfect predictability, the complications that result may lead to unanticipated and totally unavoidable delays for the principal investigator. An even more pressing constraint on principal investigators stems from inadequacy of funding for social science research. Virtually no

research project of recent years that has produced a multi-purpose collection of data has been funded at a level that would permit the principal investigator's immediate realization of even the principal objectives of the research.

Continued amicability among data producers and data users will depend on better facilitation of the primary research efforts of the former and a more sophisticated understanding of their problems by the latter. In any event, the age of the survey archive distributing data of common interest to many scholars is an age in which the line between the private the the public aspects of science has been blurred.

The ramifications that follow from the ability to implement a policy of open access to survey data are in some senses totally revolutionary. The older work ways created and perpetuated a status hierarchy among academic institutions and therefore among scholars based in those institutions. In part, the hierarchy was "simply" a matter of prestige in which the higher the institution's status the greater its ability to attract the most competent of scholars. Usually, but not always, it followed that prestige institutions provided commensurately better facilities for their prestigious faculty members. As a consequence, the first job for the new Ph.D. might do more to shape the subsequent career than any combination of factors related to native talent, excellence of training or dedication to research. With the advent of the archive, the most crucial institutional barriers that militate against excellence of performance by scholars at lesser institutions have been greatly reduced. Inequities, such as unequal teaching loads, of course remain, but minimal local facilities now permit an unprecedented range of scholars to compete on closer to equal terms. Whether there will be a real change in the overall quality of research, or increased equality in the distribution of professional rewards remains to be seen, but certainly opportunities for careers that include a substantial research component have been increased for a large number of social scientists.

II. DATA ARCHIVES AND THE QUALITY OF RESEARCH

A second way in which survey archives have altered the nature of the scientific enterprise has less to do with relationships among the members of the scientific community and more to do with the quality of their work. The state of the art throughout much of the social sciences is still one in which elegant and persuasive argument plays a large role in establishing that which is accepted as "true." However, with access to the prime evidence underlying the exposition, the importance of the evidence has most certainly increased. With the basic data open to public inspection there is, first of all, a new critical regard for the quality of the data. Irregularities that originate in one or another stage of the execution of the data collection are hard to disguise. Slipshod manipulation of the data is rapidly exposed as students, or skeptical colleagues, attempt replication. In a very real sense the presentation of survey data to public examination evokes an inverse of Gresham's law: "good research drives out the bad."

Over the past few years a new application of standards of scientific performance has developed. The technical phases of survey research are no longer irrelevant mysteries but are the subject of detailed scrutiny. The design and execution of the sample is more properly appraised. The adequacy of the information-eliciting techniques can be evaluated and they are open to substantive criticism. Standards for data processing have been raised as the understanding of options available to research scholars has broadened.

All of this has given new point to the truism that asserts the intrinsic unity of method and substance. The sophisticated scholar of today examines the methods and techniques that may have been responsible for an anomalous finding before constructing a new substantive theory to accommodate the finding. It is also true that a new awareness of the impact of method on substance may have extended the interest in method at the expense of preoccupations with substance. In the long

run, this would seem to be a necessary part of the scientific endeavor.

Along with renewed respect for the standards appropriate to the research enterprise has come a new appreciation of the importance of documentation. With the reproducibility of results as an elementary foundation for assessment, clear exposition of all the operations of the research process becomes essential. The growth of the survey archives has thus been responsible for the development of new occupations and indeed new professions within the domain of social research. The data archivist who may be a specialist in documentation has become an established feature of social science departments or divisions in many academic institutions. Depending on the scale of local activity, the infrastructure for research will not only include specialists in research design and research methods but experts in facilitation of the use of computers and specialists in providing access to complex structures of data. These developments are unique largely in their spreading incidence. Centralized archives such as that of the Inter-University Consortium for Political and Social Research (formerly the Inter-University Consortium for Political Research) have an extended experience in developing the highly differentiated staff functions that are needed by a general purpose service-oriented archive. Today, however, other local archives, engaged more in retailing than in wholesaling data, are also developing similar organizational structures in responding to the demands from their immediate research community and they are thereby adding to the positions that ultimately mean new career lines within the social science enterprise.

The growth of survey archives carries with it some inherent problems, indeed dangers, that will doubtless receive more attention in the future than they have thus far. The largest danger possibly lies in the extent to which the availability of archived data determines lines of scientific inquiry and thwarts the growth of other theoretically important courses of study. Given the limited funding available for archival work, highest

priority for the processing of studies has perhaps been too easily justified by the criterion of the generality of interest in the individual data set. Special purpose studies based on more or less idiosyncratic research interests central to the work of only a handful of scholars are not likely to be given high priority for archival processing. The large multipurpose data collections that *are* given precedence may of course produce serendipitous results for the off-beat scholar, but the production of happy accidents should not be equated with foregoing the uniqueness of highly specialized lines of inquiry.

The availability of data may also thwart the improvement of data. In recent years it has become fashionable to speak of social indicators. At times the phrase properly refers to the social implications of the aggregation of individual measures. Often, however, the indicator is but a readily available surrogate for the optimal operationalization of a theoretically important concept. The expert in secondary analysis becomes adept at making do with that which some other principal investigator has made available. "Making do" often has many virtues, but it also implies the acceptance of shortcomings whose importance and consequences necessarily remain untested. In short, the custom of relying on archival material may dull the critical sense and stifle the innovative inclination to the detriment of at least some phases of the intellectual enterprise.

III. CONCLUSIONS

The latent dangers associated with enthusiasm for the proliferation and expansion of survey archives will be real dangers only when the marginal utilities of archiving another data set clearly exceed the utility of launching the least rewarding new data collection. With an eye to maximizing research opportunities for the largest number of scholars, that day will be found in the very distant future. The broader contributions of the archive to the changing and the improving of the work ways of research scholars, and of the relationships

among the members of the research community, constitute contributions that should continue to be enhanced as a complement to the continued need for more original research by uniquely innovative scholars.