Sharing Research Data

SHARING RESEARCH DATA

Stephen E. Fienberg, Margaret E. Martin,
and Miron L. Straf, Editors

Committee on National Statistics
Commission on Behavioral and Social Sciences and Education
National Research Council

NATIONAL ACADEMY PRESS

Washington, D.C. 1985

National Academy Press • 2101 Constitution Avenue, NW • Washington, DC 20418

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Research Council was established by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and of advising the federal government. The Council operates in accordance with general policies determined by the Academy under the authority of its congressional charter of 1863, which established the Academy as a private, nonprofit, self-governing membership corporation. The Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in the conduct of their services to the government, the public, and the scientific and engineering communities. It is administered jointly by both Academies and the Institute of Medicine. The National Academy of Engineering and the Institute of Medicine were established in 1964 and 1970, respectively, under the charter of the National Academy of Sciences.

Library of Congress Cataloging in Publication Data

Main entry under title:

Sharing research data.

"Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council."

Bibliography: p.

1. Communication in the social sciences—Addresses, essays, lectures. 2. Intellectual cooperation—Addresses, essays, lectures. 3. Social sciences—Research—Addresses, essays, lectures. I. Fienberg, Stephen E. II. Martin, Margaret E. III. Straf, Miron L. IV. National Research Council. (U.S.). Committee on National Statistics. V. National Research Council (U.S.) Commission on Behavioral and Social Sciences and Education.

H61.8.S53 1985 300'.72 84–27275
ISBN 0–309–03499-X

Printed in the United States of America

Contents

PART I:
REPORT OF THE COMMITTEE ON NATIONAL STATISTICS

PART II:
SOME PERSPECTIVES—COMMISSIONED PAPERS

Sharing Research Data in the Social Sciences

Jerome M. Clubb, Erik W. Austin,
Carolyn L. Geda, *and* Michael W. Traugott

During the past two decades an extensive literature has appeared exploring issues related to access to basic computer-readable data for empirical social science research. In the main, the authors of this literature emphasize the scientific, public policy, and pedagogical values and advantages of data sharing, and they often advocate a policy of open access to data in maximally usable form. Obstacles to data sharing are discussed, specific categories of data are noted as exceptions to the general sharing rule, arguments against complete open access to research data are sometimes offered, and the precise nature of obligations to share data are debated, but few if any of the authors cate-

Jerome M. Clubb, Erik W. Austin, Carolyn L. Geda, and Michael W. Traugott are at the Inter-university Consortium for Political and Social Research, Center for Political Studies, Institute for Social Research, University of Michigan.

An earlier draft of this paper was discussed at length by Stephen Fienberg, Clifford Hildreth, Margaret Martin, Miron Straf, Joe Cecil, and Terry Hedrick. Although we were unable to meet all of their many comments and suggestions, this paper has benefitted greatly from their efforts. We alone, however, are responsible for its shortcomings.

gorically oppose data sharing or some form of open access.

These same two decades have been marked by movement among social scientists toward implementation of the general principle of open access to basic research data. Institutional mechanisms have appeared to facilitate access to data, and various agencies that fund research in the social sciences have stressed that the resultant data collections should be made available to other researchers. One consequence of these developments is that abundant, if somewhat unsystematic, concrete evidence of the value of open access to basic research data is now available.

At the same time, however, discussion and disagreement continue, and acceptance and implementation of the general principle of data sharing are far from complete. Social scientists are still often refused access to data, or if access is granted, copies of data are sometimes received in technically unusable form. In some cases data are shared, but only after prolonged delay. In other cases data are shared only within relatively limited networks of researchers, often within a single discipline or subdiscipline. Access to data by people outside such networks is either difficult or precluded. Difficulties in gaining access to data are not simply the product of unwillingness of researchers and research groups to share, but also result because mechanisms to provide information about the availability of data, and particularly mechanisms that operate across disciplinary boundaries, are not yet well developed. It is only in very recent years, for example, that concerted efforts to develop bibliographic control over computer-readable data collections have begun, and there is as yet no centralized reference service for computer-readable social science data.

Failure to move more rapidly toward acceptance and implementation of the principle of open access to basic data is sometimes asserted to be a reflection of the supposed transitional nature of the social sciences—from essentially literary values, with their emphasis upon private and unique individual creativity, to the scientific values of public and cooperative pursuit of cumulative knowledge. In our view such an explanation is neither particularly useful nor accurate. If it were accurate, other areas of inquiry would also have to be seen as transitional in nature, since difficulties and disagreements concerning access to data and to data collection facilities are also encountered in other sciences. In our reading much more obvious and, in some respects, more useful explanations are also available. First, there are serious concrete technical obstacles to effective data sharing, although at least some of them could be readily overcome. Second, there are reasonable arguments against a generalized norm of data sharing and against complete open access to research data, arguments that reflect conflicting values and goals as well as the reward structure characteristic of science. These issues constitute the most serious obstacles to data sharing.

In this paper we examine the issues confronted in sharing basic social science data. The initial section summarizes scientific and other values and advantages gained through open access to data. The second section provides an indication of the magnitude of data sharing that now occurs. The third section considers technical obstacles to generalized access to basic data in usable form and suggests means by which some of these obstacles might be overcome. The fourth section considers further arguments against data sharing and the conflicting values, goals, and obligations that seem often to underlie disagreement and discussions of data sharing; for these, solutions that go significantly beyond continued exhortation are less easily identified. The fifth section considers modes and facilities for data sharing, and the sixth section briefly considers practices of data sharing in several other areas of inquiry. We offer conclusions and recommendations in the final section.

This paper has a number of limitations that should be made explicit. Data-sharing practices vary rather widely in the social sciences, and it is unlikely that the full range of this variation has been adequately taken into account. While data-sharing practices in several rather specific areas of the natural and biomedical sciences are examined, this examination is somewhat unsystematic and far less than complete. To explore in anything approaching comprehensive fashion questions of data sharing and access to data collection facilities in the many and diverse areas of the other sciences would be a major research undertaking in its own right. Thus we are able to offer here only a few highly tentative generalizations.

There are a very large number of organizations and facilities in the academic, government, and private sectors that function in some way to share and provide access to computer-readable data relevant to social science research. Our discussion of these facilities is most complete for academically based organizations; it is significantly less complete in the case of organizations in the public and private sectors. Our discussion of data-sharing practices and facilities is also heavily based on the United States; practices, facilities, and experiences in other nations are less to computer-readable data collected and processed more or less specifically to serve the goals of social science research and the purposes of monitoring social processes. We distinguish between computer-readable *data* for research and computer-readable *information* of the sort found in data bases containing bibliographic citations and abstracts of published textual material. The latter are shared through many mechanisms and are outside the scope of this paper. There are similar questions regarding access to other categories of research source material, such as oral histories, and it is likely that somewhat similar principles and imperatives would apply to these other categories of source material as apply to computer-readable data for social science research. The personal papers of statesmen, political, government, and other public figures constitute primary source materials for the research of historians and other social scientists as well as of scholars of literature and the arts, and access to such materials is often restricted and is at best uneven. However, the issues confronted in dealing with such materials are complex, controversial, and widely debated, and we have been forced to rule them outside the scope of the present paper.

The operational records of government agencies and other organizations are also not considered in this paper. These records constitute research resources of very considerable value for investigation of social processes, and they are also of central importance for purposes of policy and performance evaluation and public accountability. Such records, moreover, are increasingly maintained in computer-readable form so that transactions and activities are documented in greater detail than formerly, and the records can also be manipulated for analytic purposes. However, these records fall within the purview of governmental, business, and other organizational archives that are today largely ill-equipped to manage them in their computer-readable form or to make them available for scientific use. A recent collection of essays (Geda et al., 1980) provides a useful summary of the issues and problems presented by these materials and calls attention to the risk of loss of major research opportunities. These issues and problems are not reviewed in the present paper.

VALUES AND ADVANTAGES OF DATA SHARING

Beginning in the early 1960s, numerous books and articles have appeared that discussed the values and advantages to be gained through open access to basic social scientific data and that explore means for providing this access. Much of the early literature emphasized the impact of change in the technology of social science research. It was recognized that the social sciences were undergoing the introduction of complex technologies analogous in some ways to the costly instrumentation of the natural sciences. The consequences of this new technology were seen as providing abundant research opportunities, but these opportunities were also seen as accompanied by need for change in work practices and uneven access among social scientists to research resources and as interposing new obstacles to effective research.

The advent of computer technology and its application to social science research meant that researchers had the capacity to manipulate large data collections and to use complex methods of analysis in ways that previously had been virtually precluded. At the same time, however, researchers faced high costs for data collection and for processing data to computer-readable form, uneven access to computational facilities and capabilities among social scientists, and the possibility and value of multiple uses of data collections. Hence the early literature emphasized need for mechanisms that would facilitate generalized access to data and to computational capabilities required for their use.

It also became increasingly clear that standard publishing mechanisms offered few effective solutions to the problems of access to research data: the size of research data collections, and the attendant high costs of publishing basic data, precluded this option. Furthermore, publication of scientific research data that already exist in computer-readable form was seen to add an unnecessary and expensive loop to the process of data sharing: to be used effectively in research applications, such published data must be reconverted to computer-readable form by each and every analyst who wishes to use them in research. Finally, in more recent years numerous observers have noted that the publishing of research results falls far short of satisfying goals represented by the term "data sharing." Few if any professional journals or monographs permit or encourage the depth of exposition of research data and methods that underlie reported research findings; it is therefore rarely the case that published research reports satisfy a reader seeking to evaluate the basic data and techniques used in a research investigation.

Increased use of sample surveys as a primary mode of data collection constituted a further impetus to data sharing. By the 1960s, numerous collections of sample survey data existed, some of them dating to the mid–1930s, and the survey method of data collection had attained highly sophisticated form. It was clear, however, that mounting a large-scale sample survey was beyond the financial reach of most social scientists and, consequently, many researchers were increasingly disadvantaged. Again, the possibility of multiple research applications and the cumulative values of data from well-designed sample surveys was stressed.

To realize new research opportunities and to capitalize on new technology required creation of new data facilities. These facilities were viewed, in some cases, as functioning analogously to the laboratories and the research installations of the physical sciences. They would provide mechanisms to implement the obligations of original data collectors to share their data with other researchers. They would devise and implement standards for data collection and processing, contribute to the development of general-purpose computational capabilities, and provide training in new approaches to social science research.

Some of these same themes continue to underlie much of the literature since the 1960s. (A partial list of the earlier and subsequent literature is provided in the references and bibliography section.) Like the earlier literature, subsequent contributions to this general discussion explore a variety of more specific advantages and values of generalized access to basic computer-readable social scientific data. In view of this large body of literature, we need only briefly summarize those values and advantages here.

Replication and Verification

Improved capacity to verify and replicate reported research findings is among the most commonly discussed advantage of generalized access to data. Obviously, use of computers and computer-readable data and increased use of large bodies of data that are costly to collect increase the complexity of verification and replication as compared with more traditional data sources and research methods. The costs of a major survey are large, and repetition of the survey for purposes of replication and verification of an original effort is usually precluded. Thus replication and verification can often be accomplished only through access to the data from the original survey. In addition, many of the phenomena studied by social scientists are in some senses nonrecurring. National elections are, of course, repetitive, but the specific contexts and characteristics of elections vary. As a consequence, findings based on data collected for one election often cannot be verified and replicated with data collected for a subsequent election. Hence, the values of verification and replication can often be served by access to the original data.

The need for simple verification of research findings is frequently minimized since fraudulent research reports are thought to be rare. The risks of datacollection or analysis errors are greater, and erroneous findings due to such errors are probably more common. However, there are also occasional reports of fraudulent research, some of them with continuing and even dire consequences. For these reasons the opportunity for verification using original data is often seen as a vital element of the research process and as dictating generalized access to data.

Access to basic data is often seen as facilitating three somewhat different forms of replication of reported findings. One of these might be described as "exact" replication. In this case the same data and methods are used to determine whether the same results are obtained. The second form replicates and tests reported findings using the same data but different analytic methods or assumptions. Both of these are obviously forms of verification and are sometimes seen as particularly important when data and research bear directly on current social policy concerns. The third form of replication looks toward testing the generality of reported findings. In this case data from different contexts—national or temporal, for example—are used to discover the conditions under which particular relations do or do not apply and, hence, to generalize research findings.

Methodological Improvement

Further values served by open access to basic data are improvement of measurement and data collection methods. In this view, the obligation to share data with other researchers subjects data and data collection methods methodological improvement is encouraged. In somewhat similar fashion, the availability of extended collections of data is seen as holding benefits for the design of new data collection efforts: in opportunities for exploratory research to determine in differing contexts the adequacy of question wordings, unobtrusive scales, and indicators, leading to improved measures and measurement validation.

Secondary Analysis

The value of data collections for extended, or secondary, analysis is, of course, frequently discussed. The research potential of a welldesigned data collection is rarely exhausted by the original data collector, and data collections usually have value beyond those for which they were originally designed. Thus data collections generally have multiple research applications. Moreover, the availability of extended collections of data provide a basis for realization of further values: in the possibilities of combining data, derived measures, or analytic results from diverse collections in order to address new research questions and in the comparative and longitudinal perspectives provided by the availability of data collected at different times and in different places. Realization of the latter values, it should be noted, not only dictates that data be shared, but also that data be preserved and remain accessible for extended periods of time.

Further values of data sharing for research are economic in nature and follow from opportunities for secondary analysis. Generalized use of data is believed to reduce research costs. The ready availability of data means that researchers often do not need to collect data de novo but can pursue research interests and goals by drawing on existing data. In this way, duplication of data collection efforts and investments are reduced, and the research value of investments in data collection are more fully realized. Opportunities to carry out meaningful research are, in effect, democratized, and more social scientists are able to conduct research and contribute to the development of knowledge.[1]

Generalized access to basic research data in readily usable form is also seen as serving a variety of additional values, including pedagogical ones. Original data are now frequently used in both substantive and methodological instruction at the graduate and undergraduate levels as well as, occasionally, at the secondary school level. Probably the best-known and most widely used examples of instructional applications of this sort are the SETUPS (Supplementary Empirical Teaching Units for Political Science) series developed collaboratively by the American Political Science Association and the Inter-university Consortium for Political and Social Research.

Twenty-one of these units have been prepared and more are now being developed or are planned. Each unit includes a brief monograph that poses a substantive or methodological problem or set of problems and a specially tailored data file to address that problem. By using original data in this fashion, students are able to more directly experience the research process and come to better understand the empirical bases and the contingent nature of research findings. In a more general sense, instructional use of empirical data improves social scientific and numeric literacy and enhances students' critical capacity to evaluate the results of applications of social science methods, whether reported in scholarly publications or in the mass media.

Ready access to data is also seen as holding values for public policy purposes. The availability of data facilitates and encourages use of empirical data in policy formation and evaluation and so improves policy. Ready access to data also means, in this view, a capacity to more rapidly address policy questions.

Numerous illustrations of the values summarized above could be cited. Three somewhat diverse illustrations are touched upon here. One example is provided through research by James S. Coleman and his colleagues (1966) on the equality of educational opportunity. The second is taken from a quite different area of inquiry: research into the economic history of the antebellum South and the economics of slavery, carried out by Robert W. Fogel and Stanley L. Engerman and reported in *Time on the Cross* (1974). In both cases, the reported research engendered widespread debate and controversy, sometimes acrimonious, among both scholars involved in the areas of inquiry and others. However, because the original data on which the research was based were generally available, scholarly debate could often be conducted on empirical rather than purely speculative grounds.[2] The underlying data could be explored and evaluated and the findings empirically tested and contested. The consequence in both cases was that, despite controversy, debate was of a higher order and more effectively conducted; weaknesses of original data collection and research were better identified, and new and potentially rewarding areas for further research found.

A third illustration is of a still different order and is provided by the American National Election Studies, which are directed by Warren E. Miller. These surveys have been conducted by the Survey Research Center and the Center for Political Studies of the Institute for Social Research (located at the University of Michigan) for each national election since 1952. Data from the surveys provide an incomparable resource for cross-sectional and longitudinal investigation of the formation and durability of political attitudes and of American political processes. In more recent years, moreover, similar studies—stimulated in part by these studies—have been conducted in many other nations, including Australia, Austria, Canada, Denmark, Finland, France, Israel, Italy, Japan, the Netherlands, Norway, Spain, Sweden, the

United Kingdom, and West Germany. In some of these nations, their series now span well over two decades. The various studies show marked similarity in theoretical foci, in the structure of questions and measures, and in other design characteristics. Thus, taken collectively, the data from these surveys constitute a powerful resource for both longitudinal inquiry and cross-national comparison, and they also exemplify the advantages, for purposes of designing new data collection efforts, of general availability of data collections.

Distinctions and Reservations

While the values summarized above are recognized and stressed, discussions of data sharing also draw distinctions, both explicitly and implicitly, between different categories of data in terms of the importance of sharing and the obligations of researchers to provide access. Data collections that threaten privacy or place individuals or organizations "at risk" are usually seen as requiring special treatment, although such concerns were less frequently expressed in the earlier literature than they are now, and distinctions are also made in the case of proprietary data collected for the purposes of private enterprise. Issues of privacy and confidentiality and questions of proprietary data are discussed in a subsequent section; here we are concerned with distinctions that center on such issues as the presumed intrinsic importance of data collections, the purposes they were designed to serve, and the relative ease with which particular categories of data collections can be replicated.

Distinctions are often drawn between large-scale data collections, particularly sample survey data collected at public expense, and smaller bodies of data collected at personal expense. There is widespread agreement that the former category of data should be shared and made generally available in a timely fashion, although there is less agreement as to what constitutes "timely." Sharing smaller data collections, particularly those created at individual expense, is often seen as less important, and obligations to provide access to such data are considered less pressing. These distinctions seem to be based on the presumed lesser value of smaller data collections for the purposes of secondary analysis, the sources of financial support for data collection, and the greater ease and lower cost at which smaller data collections can be duplicated. A similar distinction is sometimes also made for data collected from published or other public record sources. The presumption seems to be that because the original data can be found in published or otherwise publicly available sources, they can also be collected and processed by the secondary user; consequently, sharing is less obligatory or useful.

Further and more specific distinctions are also sometimes made in terms of the purposes data collections are intended to serve and their potential for affecting government, public affairs, and human life. Hedrick et al. (1978)

suggest, for example, the importance of general and immediate access to data collected for purposes of formulating and evaluating public policy. And their views might be extended to include other categories of data for applied social science research. Such data are designed to provide a basis for social program and policy decisions, and their potential for directly affecting people's lives is great. Thus in this view there is greater need for rapid evaluation of data and for replication of analytic findings than in the case of data designed to serve the purposes of more basic social science research.

Distinctions such as these may be useful and even necessary in pragmatic terms. Obviously, it would not be realistic to envision sharing and open access to all data collected by social scientists. However, distinctions of this sort may be difficult to implement in practice, and they may appear in conflict with the values and advantages summarized above. It is, after all, difficult to anticipate the potential secondary research applications of data collections whatever their size, focus, or content. Even data from the most limited case study, for example, can sometimes be combined with other data to provide a basis for more extended explorations. The view that data collected from public sources and processed to computer-readable form can be readily duplicated is at best only partly correct. Such data collection efforts usually involve large investments of time and energy, and to duplicate them is obviously wasteful. Of greater importance, data collections of this sort often draw on multiple sources, some of which may not be easily accessible, and often use complex derived measures and aggregations. Given the imperfections of the mechanics of citation, it is frequently impossible to completely identify precise sources and methods and to reconstruct derived measures and indexes. Hence duplication of such data collections and replication and verification of reported findings are often difficult if not impossible.

The recent controversy centering upon research reported by Martin S. Feldstein that shows social security as a disincentive to saving is a case in point (Feldstein, 1974, 1980; Leimer and Lesnoy, 1980). In this instance, the original sources from which the data were obtained were not as easily identified or available to others as was apparently assumed, and complex derived indexes could not be readily reconstructed. Because the data were not shared, the process of replicating and verifying the reported findings was slowed, a programming error that marred the original analysis was not more promptly discovered, and effective debate and evaluation of the findings were delayed.

It is likely that few people would contest the importance of early and general access to data explicitly designed to provide a basis for policy formation or evaluation or for social action. However, to argue that access to data for more basic research is of lesser importance presents difficulties. It is worth noting that Isaac Ehrlich's research on the deterrent effects of capital punishment, one of the controversial recent examples of contestable research with immediate policy consequences (Ehrlich, 1975; Bowers and Pierce, 1975; Passell and Taylor, 1975) was apparently not commissioned to provide a basis for policy decisions. The capacity to predict that particular research will or will not have policy consequences is far from perfect, and it is plausible tc argue that most research has the potential for policy consequences.

It may well be that for practical reasons distinctions such as discussed in this section must be made. However, the values and advantages of general and timely access to data appear commanding, and the rule should be, it would seem, to err on the side of these values and advantages rather than to move prematurely to distinctions.

INCIDENCE OF DATA SHARING

The importance and value of data sharing in the social sciences can be illustrated in a number of concrete, albeit somewhat unsystematic, ways. As will be noted at several points below, nothing approaching comprehensive information is available documenting either the incidence of data sharing or the multiple use of data collections. Several illustrations indicate, however, that very considerable sharing occurs and that data sharing is one of the vital underpinnings of research and instruction in the social sciences. The illustrations below also suggest that significant progress has been made toward realization of the values summarized in the preceding section.

Social Science Data Archives

Data sharing occurs in a variety of ways, including informal sharing among individual scholars and research groups as well as through organizations that function as data repositories and dissemination services. Indeed, one indication of the importance of data sharing is the development in the United States and other nations during the past two decades of numerous organizations that serve as mechanisms to provide general access to the basic data of social science research. These facilities include national—indeed, international—"social science data archives" in the academic sector, various private organizations that provide access to data, as well as organizations that maintain and disseminate data collected by government agencies. In addition, numerous local facilities maintain data collections, usually obtained from national data organizations, for use by a particular university community, government agency, or private firm. (A selected list of data organizations appears as the appendix to this paper.) The existence of these facilities and the resources invested in them suggests, of course, the value and importance of data sharing and multiple use of data collections.

The Inter-university Consortium for Political and Social Research (ICPSR) serves, among other functions, as a social science data archive. It is based on institutional memberships: some 270 colleges and universities in the United States and more than a dozen other nations are currently members. In return for an annual membership fee, individuals at member institutions have access to ICPSR data holdings and related services. (Access to data and services is also available, at a charge, to individuals located at nonmember institutions in the government, private, and academic sectors.) At present, ICPSR data holdings include more than 12,000 data files. A primary source of ICPSR data holdings is individual researchers and research groups who deposit data that they have collected in the course of their own research. Data are also obtained from government and private agencies, and the ICPSR staff collects and processes data, usually from public record sources. The size of ICPSR data holdings is a concrete indication of the willingness of researchers to share data.

The data holdings include virtually all forms of social science data and span much of the spectrum of social science research. They range from relatively small cross-sectional surveys through large, extended, continuing surveys. In the latter category are the series of American National Election Studies (referred to above); the Panel Study of Family Income Dynamics carried out each year since 1968 under the direction of James N. Morgan; the National Longitudinal Surveys of Labor Market Experience conducted by Herbert S. Parnes; and the General Social Survey conducted by the National Opinion Research Center under the direction of James A. Davis and others. Also included in this category are the series of surveys conducted since 1971 in the nations of the European Economic Community under the auspices of the Commission of the European Economic Community.

Extensive collections of public record data are also included in the archive. These include comprehensive voting records for the United States Congress from the Continental Congresses to the present and voting returns at the county level for elections to the offices of president, governor, and United States senator and representative from 1789 to the present. ICPSR also holds extensive data from the United States censuses from 1790 to the present, including unpublished data from the censuses of 1960 and 1970 (comprehensive data from the 1980 census are now being added) as well as data from the Current Population Surveys and various other data collection activities of the Bureau of the Census. The archive also includes data from censuses of various other nations, voting records from the United Nations, and data collected by the United Nations and other international agencies.

In substantive terms, the ICPSR data bear upon the society, politics, and economy of the United States and a variety of other nations in both contemporary and historical perspective. Extensive data are also included that bear upon the operations of the international political system and economy, the formal and informal interactions between nations, and domestic and international violence. Included as well are data collections pertinent to education, crime and deviance, criminal justice, public health, aging, and developmental processes more generally. The data holdings, in short, are a shared resource that is relevant to the study of social, economic, and political processes in virtually all their dimensions.

Dissemination and use of these resources is at least suggestive of growth in both the incidence and importance of data sharing. The volume of data supplied by ICPSR for research and instructional applications has steadily grown through the years. In fiscal 1983, for example, some 307 colleges, universities, and other organizations were supplied data amounting in total to over 138 billion characters of information. By comparison, in fiscal 1976 only 8 billion characters were supplied.

There is no solid information as to the nature of the actual use of the ICPSR data; figures given in the above and following paragraphs reflect institutional distribution of data by ICPSR. Data are supplied to a college or university and maintained by a local data facility for faculty, staff, and student use. In some cases data are supplied to one university for redissemination to other colleges or universities in the vicinity. Multiple uses of the same data are the rule, but few statistics on the number of discrete uses of a particular body of data supplied have ever been assembled. It is known that for the years from 1975 through 1980, more than 500 books, articles, dissertations, and conference papers were reported to the ICPSR staff as based entirely or in part on data obtained from ICPSR, and there is reason to believe that these constitute only a portion of the papers and publications that used these data. Several samplings of professional journals and programs for the meetings of professional associations indicate that no more than half of the publications and papers based upon ICPSR data are reported to the staff. We cannot comment on the importance of these publications and papers as contributions to social science research, but we note that the magnitude of data supplied and the number of publications suggest rather extensive interest in data sharing and also indicate a measure of realization of the values of data sharing.

Data Collections as National Resources

A further indication of the incidence of data sharing is of a different order. In recent years research funding agencies have supported several major data collection efforts that are explicitly designed to serve the research interests of extended communities of scholars rather than those of individual researchers or research groups. These data collections, in other words, are explicitly designed to serve the research interests of extended communities of scholars rather than those of individual researchers or research groups. These data collections, in other words, are explicitly intended to be shared. Four examples are noted here. The multiwave Panel Study of Income Dynamics and the American National Election Studies began as specific research projects (the former in 1968 and the latter in 1952) and were subsequently continued to provide data to be immediately available to all interested researchers. The General Social Survey began in 1972 as a general-purpose scholarly resource. A fourth example is provided by the two *World Handbooks of Political and Social Indicators* (1964 and 1972), which also involved collection of extended data for general scholarly use.

Here again, partial information on the use of these data collections can be provided. To date more than 200 copies of the data collection provided by the Panel Study of Income Dynamics have been supplied by ICPSR to academic institutions and other organizations, and additional copies of the data have been supplied directly to researchers by the project staff. Over the past 18 years the data files produced by the American Election Studies have been used by tens of thousands of researchers and their students throughout the world. Copies of the machine-readable data files from one of the most recent surveys in this series, the 1978 American National Election Study, have been supplied by ICPSR to more than 100 academic and other institutions. More than 1,000 publications and other research contributions based on this series of studies have been reported (Center for Political Studies, 1980), and here again there is every indication that the actual incidence of publications and papers based entirely or in part on these data has been significantly underreported. Information about the use of the third and fourth data collections noted above is more limited. ICPSR, however, has furnished well over 1,000 copies of specific files from the General Social Survey series to various institutions, and the Roper Center for Public Opinion Research, which also distributes the data, has supplied additional copies. Jodice et al. (1980) report some 300 research applications employing data from the two *World Handbooks of Political and Social Indicators*.

As noted in the preceding section, shared data are used not only for research but also for teaching. As in the case of research use, only limited indications are available as to the actual incidence of instructional applications of shared data. Data for the SETUPS teaching units (described above) are maintained and disseminated by ICPSR, as are data for a number of other teaching packages. To date more than 1,150 of these instructional data files have been supplied by ICPSR for use at well over 350 colleges and universities. Here again, these figures undoubtedly seriously understate actual use. The data in question were supplied to institutions to be maintained for continuing use, and it is at least highly likely that these data were used in more than one class. No record is available of these multiple uses, nor is there a record of the instructors who have used shared data to fashion their own packages for instructional applications.

Again these illustrations are intended only as indications of the incidence of data sharing and of its value and importance for research and teaching. Nothing approaching complete information is available, and it is certain that these illustrations provide only a very partial indication of the incidence of data sharing and of multiple applications of shared data collections. Taken in total they strongly suggest, however, that data sharing has become an important mechanism to support research and teaching in the social sciences.

TECHNICAL OBSTACLES TO DATA SHARING

While data sharing in the social sciences appears widespread, there are also important obstacles that often slow the sharing process or completely prevent it. For the purposes of the present discussion these obstacles can be grouped into two categories. The first includes essentially technical problems, most of which, at least in principle, can be solved. The second category relates to what might be described as conflicting values and obligations and to the reward structure of the social sciences and, for that matter, of the sciences more generally. In this area, solutions are less easy to identify.

Stated in general terms, technical obstacles to sharing computer-readable data in the social sciences reduce to matters of machine and software-system incompatibilities, data-file structures, and standards and procedures for recoding, processing, and documenting data. In earlier stages of the development of computer technology, essentially technical factors sometimes constituted virtually insurmountable barriers to transferring data from one computer installation to another. At the present stage of technology, however, difficulties encountered in transferring data from one installation to another are largely due to the practices of original data collectors and processors rather than to technical factors.

Machine Incompatibilities

Earlier, for example, computational equipment was characterized by considerable variation in terms of conventions used for internal representation of information. Variations existed not only between equipment produced by different manufacturers, but even between machines produced by the same manufacturer. Today, however, very significant standardization has occurred. Variations still exist, but they can be overcome by what might be termed a lowest-common-denominator approach. That is to say, data recorded in character mode can be more consistently transferred from one machine to another than data recorded in binary mode. Common conventions for internal storage and representation of character-mode data (either ASCII or EBCDIC) have been more widely accepted than for binary-mode data. Similarly, data organized in card-image or rectangular logical record format, whether recorded on magnetic tape or other media, can be more readily transferred between installations than data organized in other forms. The only major exceptions to these generalizations involve recently developed microcomputers and the nonstandard data storage devices (floppy and hard disks, cassettes, etc.) they use. Acceptance of common conventions is less general across this equipment than in the case of larger computational devices.

Incompatibilities of Software Systems

Technical requirements and characteristics of data management and analysis computer program systems also sometimes complicate date sharing. Data organized for analysis using the Statistical Package for the Social Sciences (SPSS), for example, cannot be analyzed using the Statistical Analysis System (SAS) without reformatting and reorganization. Here again, the character-mode, card-image, or logical record approach referred to above constitutes a common denominator. Data records in these forms can be organized and restructured ("filebuilt," to use the jargon) to meet the requirements of these systems or any other available general-purpose computer software system. To do so, however, requires rather elaborate and time-consuming effort. Some of these systems include capabilities that allow data prepared for another system to be "read" and somewhat routinely converted to the required form and structure. Conversion capabilities of this sort could probably be added to all such systems.

Many of the problems encountered in converting data prepared according to the conventions of one software system for use by another revolve around the database dictionaries rather than the data records themselves. Database dictionaries contain technical and substantive information about the data file and each of the data elements in it. By prerecording this kind of descriptive information in computer-readable form in a database dictionary, the actual retrieval and analysis of data is greatly simplified. Indeed, the development of database dictionaries, begun in the late 1960s, stands as an important innovation in facilitating ready access to and use of large and complicated data collections. Yet most database dictionaries in use in the social sciences are tied specifically to certain software packages like SPSS, OSIRIS, or SAS; their conversion for use by other packages is usually not straightforward. Thus, researchers attempting to use data prepared by others must often forgo direct use of information contained in the "foreign" database dictionary or, alternatively, they must reenter the information into a computer-readable form compatible with locally available software. As mentioned above, conversion capabilities could be added, or are being added, that would allow computer installations to accept database dictionaries prepared for other systems. These additions would surmount a significant barrier to effective data sharing.

Difficulties are also encountered in transferring large and complexly structured data files for use at other installations. The first issue is a matter of limitations of machine capacity at recipient installations and can usually be overcome by provision of custom subsets of larger files tailored to specific needs. The second is a matter of availability of appropriate computer program capabilities. Increasingly, social scientists have begun to use complex structures to organize data, such as hierarchical and, to a lesser degree, network structures. While these file structures are appropriate for the data and facilitate data management and research applications, computer programs to work with such structured data are not available at many installations. Data structured in these fashions can usually be converted to more standard rectangular ("flat") form, but to do so requires appropriate software, and the result of a "flattening" operation is a data file that is substantially larger than the original structured file. At present, however, this difficulty remains relatively confined, since files with complex structures are not yet widely used. It is also a difficulty that can be overcome through further development of general-purpose computer programs.

Data Preparation and Documentation

Further obstacles to data sharing result from matters of data preparation and documentation. Data received from original collectors often have undocumented codes, inconsistencies, and other errors; coding conventions and formats that are not acceptable on other systems; and inadequate documentation. The result in such cases is data that can be used only with difficulty or not at all. Problems of this sort are sometimes said to be the product of absence of standards for data preparation and documentation. In fact, however, basic standards for preparation and documentation are rather widely accepted and followed (they are stated systematically in Geda (1979) and Roistacher et al. (1980); the problems arise because the original data collectors and processors are not aware of the existence of the standards or they are simply not followed.

This situation seems to result from several considerations that, on the surface at least, appear fully understandable. Data collectors sometimes prefer to continue to follow data preparation and documentation practices with which they are familiar even though those practices may be at odds with the ones followed by others and with accepted standards. Investment in converting to new practices is seen as unnecessary. Accomplishment of research goals is often not seen as requiring fully "cleaned" and well-documented data.

The requirements of research, in other words, may be different than those of data sharing, and data are collected primarily to achieve particular research goals, not to serve the purposes of data sharing and secondary analysis. Considerations of funding are sometimes at issue. Available financial resources are seen as inadequate to support both data collection and analysis as well as elaborate data preparation and documentation. In this situation, the latter work is given lower priority.

Views such as these are in need of reconsideration, and not solely because of data sharing. It is likely that application of basic standards of data preparation from the beginning of data collection, through data processing, and throughout a project would result in reduced rather than increased project costs. A more readily usable file would be created, and time-consuming interruptions of analysis to correct errors would be avoided. Costly backtracking to recover needed but unrecorded information would similarly be reduced or eliminated, and, certainly, the purposes of data sharing would be better served.

A distinction should be made here between technical and substantive documentation. By substantive documentation we mean such matters as descriptions and explanations of sampling procedures and of the original design of the data collection and of deviations from it; of the assumptions that underlie particular questions, combinations of questions, and derived measures; of the degree to which instruments were pretested and the results of those pretests; and so on. As noted above, basic standards for technical documentation have been established and are in use in the preparation of many research data collections, but practices regarding substantive documentation are less consistent and probably generally less adequate than in the case of technical documentation.

Yet the substantive aspects of documentation are fully as important as technical ones in facilitating effective secondary use of data collections. Data may be in perfect technical order and readily usable in these terms, but if the substantive documentation is inadequate, the data are subject to inadvertent misuse with the result of misleading or erroneous findings. The inadequacies of substantive documentation are apparently widespread and extend to the literature reporting research findings.

Data Access

We argue here that technical obstacles to data sharing are largely related to the practices of original data collectors and processors rather than to the peculiarities of computers and data processing equipment. We have referred, however, to data sharing that involves actual transferral of copies of data collections, whether directly from one researcher or installation to another or through an intermediary data archive or other organization. For some of the purposes of secondary analysis, the process of transferring data is not fully adequate and may indeed present a barrier to data sharing.

Secondary analysis often requires that researchers combine data from diverse data collections to create a new data collection designed for new research goals. The ready availability of data collections means that researchers can carry out exploratory analyses to design new data collection efforts, to assess the efficacy of particular measures and questions, and to perform preliminary tests of hypotheses. But to achieve these benefits under present modes of data sharing, a researcher must acquire data collections and install them on local equipment, a process that often involves time delays and considerable investment in data manipulation. The consequence is likely to be that researchers sometimes forgo the benefits of available data.[3] Difficulties such as these could be reduced through remote access to data collections. Remote access to on-line data collections is now fully feasible in technical terms, but under present conditions is unnecessarily cumbersome and costly and is, as a consequence, only used in limited ways by academic researchers.

CONFLICTING VALUES AND OBLIGATIONS

Before turning to the issues of conflicting values and obligations, it may be useful to briefly consider several related matters. One of these concerns individual creativity. The design and execution of a data collection effort is a creative activity that sometimes involves innovative techniques. Why should secondary analysts be allowed to benefit from the creative work of original data collectors to which they themselves did not contribute, and why should original data collectors be expected to reveal their innovative techniques to others who are potential competitors? A further question concerns the alleged temptations presented by data sharing: since secondary analyses that replicate and confirm reported findings are difficult to publish, secondary analysts, or so this view holds, are tempted to be unfairly critical of the original work. The latter allegation is, of course, related to another allegation that is sometimes made: that original data collectors sometimes refuse to share data out of concern that that their reported findings may be refuted and inadequate methods revealed.

There are several responses to these views. The notion of private individual creativity, at least as phrased above, contradicts the concept of open pursuit of replicable and testable knowledge, particularly in the case of costly data collections that cannot be readily duplicated. Development of innovative techniques, moreover, is a contribution for which professional reward and recognition is often given. Furthermore, critical examination and evaluation of data collection and analysis procedures are necessary elements of the research process and should be listed as benefits of data sharing, not liabilities. Unfair criticism is obviously undesirable, but there are other mechanisms available to discourage such practices that do not involve secrecy. Reports of replications that confirm original results are probably too frequently rejected for publication: greater receptivity on the part of editors and reviewers to such studies, particularly those that involve innovative replications, would be a step toward removing obstacles to data sharing.

Rewards for Data Sharing

These issues are obviously related to the reward structure of the social sciences. What might be termed the reward dilemma is easily stated. In social science research, as in the sciences more generally, rewards come from original research contributions, not from contributing data for use by others. Sharing data may be desirable, it may contribute to the development of knowledge, and it may facilitate the research of others, but it has no place on the curriculum vita. In fact, data sharing may hurt: premature release of data may allow another to publish it first, and any sharing deprives the original investigator, and perhaps students and colleagues, of long-term opportunities to mine data collections.

These are real values that cannot be easily set aside, and they are at odds with the individual and collective values summarized in a preceding section. But the dilemma is obviously overstated, and its various components are not of equal weight. There are rewards for sharing data. Contribution of valuable data for use by others is recognized, albeit often only informally, and one component of the stature of some senior scholars is probably the quality, value, and innovative nature of data that they have collected and shared. However, rewards for sharing data could be strengthened. A minimal step would be to improve citation practices. Journal editors might take greater care to ensure that the sources of data that provide the bases for submitted manuscripts are fully and accurately cited. Although the suggestion may appear trivial, some sort of public recognition of data contributed for secondary use, perhaps in the form of journal or newsletter notes, might be valuable. It is also worth noting that sharing data is beneficial to all. To the degree that a norm of data sharing is followed, original data collectors also have access to the data collected by others.

Concerns for prior publication by others as a consequence of prematurely shared data can also be exaggerated. The concerns often seem to neglect the advantages primary investigators have over secondary analysts. Primary investigators design instruments, measurement procedures, and data collection strategies, and they do so to address well-formulated research questions. Thus, the possibility that secondary analysts, even with immediate access to data, will be able to scoop primary investigators in any significant way seems limited.

There are also steps that could be taken that would further reduce such possibilities. A useful small step might be taken by foundations and other research funding agencies. In some cases funding is sufficient to support data collection but insufficient to support analysis, so that reports of primary findings as well as data sharing are delayed. In these situations, more adequate research support would speed both processes.[4] It is also sometimes argued that funding is adequate to support data collection and analysis but insufficient to support the documentation, cleaning, and processing of data to forms adequate for use by secondary analysts. As suggested above, however, adherence to basic standards of data preparation from the beginning of data collection would probably reduce rather than increase costs and would produce data collections adequate for secondary analysis.

Mechanisms to protect the prior rights of primary investigators, even though data are shared, have been suggested. One of these is to accord to primary investigators for some specified period after release of data a right to review manuscripts by secondary analysts and to request delay of publication. Such a mechanism—and others of a similar nature—may have disagreeable implications and may also admit to abuse, but it has been used and may merit consideration.

Suggestions such as these obviously do not reconcile the dilemma, but the dilemma is still overstated. The scientific value of data sharing appears commanding, and it is probably the case that many, perhaps most, academic data collectors are agreed that sharing data is desirable, with specific categories of data noted as exceptions (see below). There is probably also substantial agreement, in principle, that data should be shared after a specified period, perhaps 1–2 years to allow time for completion of initial analyses and publication. Steps are needed to institutionalize such a norm while recognizing legitimate exceptions, and suggestions to this end are made at the conclusion of this paper.

Such a norm, however, should not be categorical. In the case of several categories of data, a norm of more immediate release would be desirable. There is no obvious reason, for example, why data relevant to social science research that are collected by government agencies and that do not pose hazards to confidentiality or national interest should not be made available immediately. Similarly, it would seem that data collections commissioned to address public policy issues should be subject to early release, and this norm should also extend to data that, though not commissioned for public policy purposes, bear directly on policy issues. And finally, for data that are of immediate value to large numbers of researchers and that relate to critical research issues, a norm of early release would appear desirable, however, with appropriate steps to accord recognition to original data collectors and to ensure that they obtain the benefits of initial publication.

Misuse of Scientific Data

Another area of value conflict involves the possible misuse of scientific data. There are at least two aspects to this issue. One involves the concern that other researchers will misapply data and arrive at erroneous findings, perhaps through use of inappropriate methods or by failing to recognize limiting characteristics of data. A related concern is that secondary analysts will waste their time pursuing avenues of inquiry that the primary investigator has already found to be fruitless. While misapplications of data and wasted effort are obviously undesirable, refusal of access to data on these grounds may sometimes seem to imply omniscience on the part of a primary investigator. The peer review system, moreover, remains the primary safeguard against publication of erroneous findings. Whatever the shortcomings of peer review—and they are surely many—it appears preferable to denial of access to data on the basis of the prior judgments of original data collectors.

The second concern is that data will be used for unscientific purposes, perhaps for profit making or to serve ends that the original data collector considers inappropriate or antisocial (such as deliberately casting particular groups in an unfavorable light). In some instances, such concerns are taken as arguments against all data sharing; in others they are taken as reasons to limit data sharing to established and recognized scholars or to academic researchers. It is easy to sympathize with some of these concerns. Except in the case of data that bridge privacy or place individuals or organizations at risk (discussed below), however, these concerns do not seem to justify complete refusal to share data. To argue, moreover, that use of data should be confined to established or academic researchers only and that use for government or commercial purposes should be precluded raises complex questions, particularly for data collected at public expense. From some points of view at least, the right of an original data collector whose work was supported by public funds to make such a decision would be highly questionable. Similarly, to allow only particular private groups access to data while refusing access to other groups would also present questions of propriety and would involve judgments and distinctions that some researchers would be unwilling to make.

Proprietary Interests

A further set of conflicting values concerns proprietary data. Commercial concerns sometimes collect data that have potential value for social science research. Since these data are collected for profit-making purposes, provision of general access would be competitively disadvantageous.

One example is data collected by the A. C. Nielsen Company on television viewing habits, which includes data on characteristics of households and of small areas; data collected by commercial polling firms constitute a more obvious example. Still other firms collect data that both provide a basis for a profit-making service and are sold, sometimes at high prices, for a profit.

(The Dun & Bradstreet small-area data are an example.) It is unlikely that social scientists can achieve open and general access to such data. But if a data-sharing norm was more fully institutionalized within the social sciences, such firms might be encouraged to provide at least limited access to their data, perhaps in the form of "public-use tapes," for social science research. (Some of the approaches discussed below to provide access to confidential data might also afford a means to allow social science researchers access to proprietary data of this sort.)

A second category of data that is sometimes treated as proprietary is that collected by private firms for purposes of policy or performance evaluation under contract from government agencies. In some cases, the data are retained by the firms as a basis for further work on their own. In this case, however, there is no obvious reason to exempt such publicly funded data from the general norms of data sharing suggested above, and the contracts commissioning such data collection efforts provide a convenient means to ensure data sharing.

Proprietary issues also arise in another way. Some organizations, individuals, and groups of individuals resist being the subjects of research—out of concern for privacy or fear of embarrassment or damage—and are willing to cooperate with researchers only under restrictive conditions. In some instances these restrictions include explicit or tacit understanding that data collected by the researcher will not be made available to others. Even in the absence of such understandings, researchers sometimes fear that release of data will effectively "dry up the source" and result in future refusals to cooperate. Hence, data sharing is understandably resisted.[5] Here again, approaches that might be used to provide at least limited access to data that threaten confidentiality might also be used to provide access to data of this sort.

Confidentiality and Privacy

Among the most frequently discussed and controversial issues about data sharing are those that relate to matters of confidentiality. Some categories of data allow identification of specific individuals or organizations. As a consequence, such data abridge privacy and place individuals and organizations at risk of damage or, at least, embarrassment. Issues of confidentiality and privacy raise complex legal questions that we are not qualified to discuss (see

Cecil and Griffith, in this volume). Here we can only attempt to better define the magnitude of the problems presented by this kind of data and note various means to allow shared use of data without abridging confidentiality or privacy.

Most social science research does not require identification of specific individuals or organizations. For that research, problems of confidentiality would be solved if the simple practice of removing names and substituting numeric identification codes was uniformly followed. Similarly, confidentiality would be further preserved if occasional data values that reflect rare attributes and, hence, allow identification of specific individuals or organizations were consistently removed from data collections.[6] For most data and most research purposes, uniform adherence to these simple practices would preserve confidentiality and privacy.

It is often noted, however, that in some cases combinations of variables can be used to identify specific individuals or organizations through a process of "triangulation." It is also sometimes possible to combine data from different sources in a triangulation process. (The combination of automobile registration information with small-area data from the U.S. census is sometimes given as an illustration of this possibility.) Three means to avoid such possibilities have been suggested and implemented: to introduce limited random error into data; to group data; and to combine variables to create composite variables that do not allow identification of specific individuals.

Obviously, all of these approaches involve some reduction of the research value of data. A fourth approach, removing offensive variables entirely, is even more strenuous in this respect. But before undertaking or advising these rather heroic steps, it might be legitimate to ask why, under what circumstances, at what costs, and at what risks to whom would the laborious process of triangulation be undertaken. Whatever the answer, however, most social science research does not require data that allow identification of individuals, and whenever necessary, means are available to prevent it.

There are categories of research that require use of data with identifiable individuals or organizations. Investigations of elite groups or other small or special populations with rare traits and studies of particular organizations or sets of organizations are examples. In such research, the means noted above cannot be used to protect confidentiality. Even for this research, however, approaches have been suggested and used to allow at least limited sharing of data. One approach involves a form of licensing or "swearing in" as a condition for access to data with the possibility of legal sanctions and penalties for breaches of confidentiality. Another approach involves provision of custom data reductions and analyses: for example, some organizations maintain confidential data collections and provide, to user specifications, subsets of data, summary measures, or analytic results that do not allow identification of individuals. Both of these approaches might also provide a means to allow access to proprietary data. Obviously, using either of these approaches, a researcher is effectively subjected to a measure of surveillance, and some restraints are imposed on the kinds of research and analyses that can be carried out. Even so, they do permit at least limited access to otherwise inaccessible data.

MODES AND FACILITIES FOR DATA SHARING

There are two primary modes for sharing and providing access to social science data. The first of these is simple sharing in informal and somewhat ad hoc fashion among researchers. Individual researchers and research organizations simply request and receive copies of data from other researchers and organizations. In some cases, data so obtained are then supplied to still other individuals. The second mode involves use of intermediary facilities that function as data repositories and dissemination services. In some instances, the facilities are a part of research organizations or data collection agencies; in others they are more or less independent organizations.

Informal Data Sharing

Data sharing in substantial but unknown volume occurs through the first mode, and informal sharing in this manner is often seen as involving significant advantages. One advantage is economic: the original data collector bears the costs of maintaining and supplying data or charges those who request data the minimal costs of copying tapes and duplicating documentation.[7] There are no overhead costs for maintaining an intermediary installation. Other advantages of this mode are the intimate familiarity data collectors have with their own data and their consequent ability to advise and assist secondary analysts. Intermediary agencies are believed to lack this familiarity or conversance with data. Still a third advantage of the direct, informal mode is the absence of bureaucratic obstacles that intermediary facilities are sometimes seen as interposing between original data collectors and secondary users.

Some of the disadvantages of this mode to data sharing are related to its advantages. Since the original data collectors bear the costs of maintaining data collections, they suffer at least the distractions involved in honoring requests for data. If requests for data are numerous, those distractions may become intolerable and, for that reason, the data may become unavailable or may not be preserved for extended periods. Thus the cumulative value of data collections is reduced.

This informal data sharing approach probably occurs most commonly within networks of researchers working in the same areas. Researchers in other areas are less likely to know of the existence of relevant data, and their requests for access may be less readily honored. Hence this mode is less likely to facilitate interdisciplinary use of data or to allow realization of the combinatorial opportunities provided by data sharing. Technical difficulties—in terms, for example, of nonstandard formats and inadequate documentation—are also likely to be more frequent in informal data sharing, and safeguards for data quality are probably less effective.

Sharing Through Intermediary Facilities

The second approach to data sharing, through intermediary facilities, requires somewhat more extended discussion. As noted above, there are numerous such facilities in the academic, government, and private sectors in the United States and other countries. These include nationally oriented social science data archives in the academic community, which function in more or less general-purpose fashion in that they are oriented toward several or all social science disciplines. A number of agencies of the federal government also have data centers that maintain, manage, and disseminate data produced by those agencies. Finally, there are numerous local facilities that provide access to data—often obtained from national data organizations—and provide other data services for a particular university community, government agency, or private firm. Thus it is possible to speak of an extended, if somewhat inchoate, network of data facilities that extends from the level of local installations and clienteles to the national and international levels. (The appendix is a partial list of these facilities.)

At this point we are primarily concerned with the nationally oriented data archives in the academic sector, which seem to be the primary organizational mechanism used for sharing data for social science research. The ICPSR, one of these archives, was discussed above. A second is the Roper Center for Public Opinion Research, located at Yale University and the University of Connecticut. The Roper Center differs from ICPSR in that it is primarily, although not exclusively, oriented toward sample survey data collected by commercial firms and agencies (ICPSR data holdings largely originate from the academic and government sectors). The extended data holdings of the Roper Center are highly diverse in substantive terms, they cover many nations, and they have the advantage of considerable temporal reach: some of the data are from surveys conducted as early as the 1930s. Data archives in other nations include the Zentralarchiv für empirische Sozialforschung, at the University of Cologne; the Danish Data Archives, at the University of Odense; the Social Science Research Council Survey Archive, at Essex University in Great Britain; the Belgian Archives for the Social Sciences, at Louvain la Neuve

University; and the Steinmetz Archives in the Netherlands. There are, in addition, a number of private-sector organizations that provide access to social science data produced primarily by the federal government. Chief among them are DUALabs, Inc., of Arlington, Virginia, and Data Resources, Inc., of Lexington, Massachusetts, among others.[8]

The academically based organizations listed above differ in substantive orientation and in terms of the forms of data they hold. Their basic function, however, is the same: to maintain data resources and make them available for research and instructional applications. The primary source of data is researchers who have collected them in the course of their work, but data are also obtained from government and private sources, and data are sometimes collected by the archives themselves. On a selective basis, the archives process data to eliminate or document errors and inconsistencies, convert them to standard format to facilitate dissemination, and prepare documentation. In most cases data can be supplied, usually on magnetic tape, to researchers in technical forms compatible with requirements of local computational facilities.[9]

The financial bases of the academically based organizations are highly diverse and in some instances resemble patchwork quilts. In some cases support is derived from a combination of member fees or other subventions from participating colleges and universities, fees for services, and subsidies from the universities at which they are located. Grants and awards from government and private research funding agencies are also received, usually to support special projects or for development of facilities. Support for the operations of some of the European archives is provided by national governments or research-supporting agencies. In some cases private-sector data organizations are for-profit operations, while others are not for profit. Government data facilities are, of course, supported by government; the fees assessed nongovernment users for access to data and services range from minimal to very costly. In general, variations in support base have obvious implications for the levels and kinds of services that these organizations provide and the fees (if any) for obtaining data and related services.

From the standpoint of secondary analysis, these data organizations, particularly those in the academic sector, have a number of advantages. Their holdings tend to be substantively diverse and include data of varied forms, and they cover many disciplines. Thus they encourage and facilitate interdisciplinary use of data, and their data dissemination activities are not confined to limited networks of scholars. They are located at universities, staffed and directed by trained social scientists, and they usually draw upon advisory panels and committees composed of active social scientists. Consequently, they are well integrated into the research community. They also relieve original data collectors of the burdens of maintaining and supplying data to others, and they contribute to the development and implementation of more uniform practices of data preparation and documentation. Because they preserve data indefinitely at a central location, the cumulative and combinatorial research value of data collection efforts can be better realized.

Intermediary facilities also have disadvantages, some of which were alluded to above: the overhead expenses required to maintain them; their distance from the original data-collection process; and their intermediary nature itself, sometimes interpreted as posing barriers between original data collectors and others with whom data might be shared. But at this point the advantages for data sharing of intermediary facilities seem to greatly outweigh their disadvantages.

PRACTICES OUTSIDE SOCIAL SCIENCES

A somewhat superficial review of data-sharing practices and access to research resources in other sciences suggests a range of diversity at least as broad as that found in the social sciences. It suggests as well the presence of problems, issues, and disagreements that appear similar to those encountered in the social sciences. But before turning to these matters, the limitations of the comments that follow must be made clear. A comprehensive examination of data-sharing practices in the other sciences would be a monumental task indeed. Such an examination would require both review of a very large and complex literature and systematic interviews with scientists to determine the ways and degrees to which actual practices diverge from stated principles and conceptions of appropriate behavior. It would also require a degree of conversance with the substance, methods, and technologies—and, indeed, the lore and gossip—of diverse areas of inquiry that we lack.

The discussion here is based on a significantly more limited effort. It is primarily concerned with three rather specific areas within the natural and biomedical sciences. It is based on relatively shallow soundings of relevant literature and on more or less extended and systematic discussions with colleagues active in research in these areas. Therefore, the discussion is not well informed in technical terms, but is impressionistic and tentative. However, even this limited effort indicates great diversity, and it provides at least some idea of issues confronted in data sharing in the natural and biomedical sciences.

The principle of data sharing and the collegial norm of contributing data to central resource bases are apparently well established in at least some areas of the natural and biomedical sciences. Particularly when expensive instrumentation is involved or when maintenance of large colonies of experimental subjects is required, scientists—or perhaps more accurately, their laboratories—are seemingly accustomed to the use of computer technology to share data and to administrative arrangements that facilitate exchange of data.

In some cases individual researchers contribute observational data collected with one type of instrumentation in anticipation of receiving analogous data derived from alternative data collection techniques. They actively engage in a two-way flow of data, often with explicit agreements about levels of measurement, units of measurement, and technical formats for supplying data. Not everyone is fortunate enough, of course, to be located at an institution that is technically well endowed, and many scientists simply avail themselves of data from central repositories in their research activity. They are able to perform analyses based on materials that are provided on magnetic tape or to which direct, on-line access is possible for essentially the costs of computer time for data copying and analysis. In these cases, there is only a one-way flow of data from the resource base to the scientist.

The range of data resources and the conditions under which they are available are highly varied, but at least two facilities—one on the sun and one in medicine—appear markedly similar to the social science data archives described above. For physicists and astronomers interested in data on the sun, there are a variety of data collections available from the World Data Center A for Solar-Terrestrial Physics in Boulder, Colorado. This is one of the world data centers established in conjunction with the 1957 International Geophysical Year in order to archive and provide data related to solar and interplanetary phenomena.[10]

Solar-geophysical data contributed by more than 60 institutions located around the world are archived at the Boulder facility. All of these laboratories or observatories have substantial investments in the land-based or satellite instrumentation that is used to collect the data, and it is the accepted norm for the data that they collect to be deposited at the Boulder center. Even the U.S. Air Force prepares a special public-use tape, from its own otherwise classified satellite data, for deposit there. The basic data series available from the Boulder center include information on sunspots, solar radio emissions, coronal holes and flares, solar wind, cosmic rays, and the like; the detailed data series contain hundreds of variables. While some of the series extend back to 1957, most were initiated during the mid–1960s or later.

Data are available from Boulder in three forms—on tape, in printed reports, and by telegram. With continuous data input, the various series are frequently updated. A researcher can obtain computer-readable data on tape in three dimensions: selected variables for selected times at selected locations on the sun's surface or in space. Data are also published by the center in monthly reports, which contain selected variables in a standardized format. These data are published with only a 2-month delay and constitute an extremely timely data source by most scientific standards.

Since many astronomical events are relatively short-lived, the center also operates, for a fee, a rapid notification system. Through this service researchers can be notified by telegram of the occurrence of a major solar-terrestial event, such as a flare of a certain size or larger. In this way, a researcher interested in geomagnetic storms on the sun, for example, can be notified immediately when such an event starts in order to begin independent observation and data collection. After analysis by the researcher, it would be expected that the data would also be deposited with the center.

In more general terms, it appears to be the accepted norm that individual scientists and research groups deposit relevant data produced by independent observation with the various centers. Among astronomers, such data are expected to be deposited after initial analysis and publication was completed, usually 1–2 years. An astronomer who observed a rare event, such as a supernova, would be expected to immediately report its occurrence to the Smithsonian Center for Short-Lived Phenomena so that other scientists could be notified and begin independent observations. We have no information as to actual adherence to these standards or of any sanctions for noncompliance.

The Laboratory Animal Data Bank (LADB) is a second example of data-sharing facilities of this sort. LADB is a computer-based, on-line information resource developed by the National Library of Medicine (see National Library of Medicine, 1980). Its purpose is to provide biomedical researchers with information obtained from laboratory animals on hematology, clinical chemistry, pathology, environment, husbandry, and growth and development. The system was originally developed to meet the needs of the Department of Health and Human Services' Committee to Coordinate Environmental Programs, including the National Cancer Institute, the National Center for Toxicological Research, the Environmental Protection Agency, and the Interagency Regulatory Liaison Group. But the data base is now available to any researcher, for a fee, for on-line or off-line access.

Approximately 50 laboratories routinely contribute data to LADB about each of their experimental animals, the conditions under which they are maintained, and details about their aging and death. The data base now contains information from over 500 animal groups composed of 30,000 individual animals, representing 65 strains or species of animals. There are now more than 1 million observations in the data bank, and data are continually being added.

An individual scientist might use this data base to establish parameters for normalcy in terms of various physiological and biological measurements or to evaluate spontaneous pathological changes in the animals. The information can be obtained in the form of marginal distributions for selected variables, cross-tabulations or correlations, or complete listings of the data for selected subject animals. And, as noted above, researchers can gain access to the data through the contractor that provides computer services for LADB, through the National Library of Medicine, or through direct access to the data base.

Again, these facilities appear markedly similar in function and goals to the social science data archives described above, and they seem to further highlight the advantages of intermediary facilities as mechanisms for data sharing.[11] In at least some other areas of other sciences, however, the norm of data sharing is apparently less well established and less frequently followed.

In some areas of the biomedical sciences, data-sharing practices apparently take quite a different form from those that are relatively widespread in the social sciences. In general, data sharing means publication of research results in journal articles and the like. Very little sharing of the data on which research reports are based seems to occur, and data sharing is not widely advocated as a desirable or necessary practice. While nearly all biomedical researchers would agree in principle to make basic data available to other researchers, the practice is seemingly rarely followed.

There appear to be three main reasons for the lack of data sharing in these areas: a proprietary attitude toward data and research; the form of the data that might be shared; and the relative ease with which data can be collected and research can be replicated. Proprietary issues seem to be the most important elements in the nonsharing equation: researchers place such a high premium on being the first to publish a particular finding and are in such competition with each other to do so that most would be unwilling to make basic research data available to other potentially competitive researchers. This unwillingness to share basic data persists even beyond the publications of findings, since sharing the basic data that underlie a particular investigation would reveal research techniques and methods that the original researcher was continuing to use in ongoing investigation. The apparent concern is that such revelations would not be in the self-interest of the original investigator.

The second obstacle to data sharing in these areas is the form of the data to be shared. The data in question are frequently records of observations, test results, and the like, transcribed in idiosyncratic fashion in typed and handwritten notes and stored in ponderous notebooks and folios. Not only is the technology for sharing such information (i.e., photocopying of some sort) expensive and cumbersome, but the organization of the material often presents serious difficulties of interpretation for other researchers. When sharing of these materials occurs, it is accomplished by one researcher traveling to the research site of another to examine research notebooks, charts, and the like, and by interviewing the original researcher and his or her technicians. This is obviously a time-consuming process, and few researchers have the luxury of traveling or hosting such an exchange of basic data. If a piece of research is called into question, such an examination can be undertaken, but it is not part of the normal routine because of its cost and cumbersome nature.

The third reason for the lack of widespread sharing of basic research data in these areas is the relative ease with which new data can be collected and research thereby replicated. This issue has two related elements: the desire of researchers to be in control of the design, conduct, and conditions of data collection and the relative availability of funding and facilities for data collection. Much of the necessary data can be collected in other contexts with relative ease through the use of clinical and laboratory procedures and facilities to which these researchers have reasonably ready access. In addition, funding is plentiful (in a relative sense) and thus the incentive to reuse data is not strong.

Data sharing does occur in a number of specific areas of biomedical research, and its value is recognized. One example of sharing is the National Cooperative Crohn's Disease Project. Because of the rarity of cases to study, over 15 sites were jointly funded to pool data on the disease and trade that data back and forth among researchers.[12] An indication of concern for sharing is provided by a major journal, *The Journal of Clinical Investigation*, which has undertaken to require explicit discussions of methods, data used, and experimental procedures in manuscripts as a condition for publication. This requirement, however, has apparently led some biomedical researchers to turn to other publishing avenues (which exist in abundance) rather than comply.

These examples seem to illustrate rather divergent practices of data sharing in the other sciences. They also suggest both similarities and differences between the social and other sciences. In numerous scientific areas there appears to be widespread interest in the development of data centers to collect, maintain, and provide access to basic data, and a number of such centers seem similar, on superficial examination, in many essential functions to the data archives and facilities of the social sciences. There are concerns about the establishment and application of adequate standards for collecting, encoding, recording, and documenting data, for data quality, for data evaluation, and for the need for scientifically trained personnel to manage data centers and facilities—all of which are very similar to the data-sharing literature of the social sciences. Indeed, the concluding paragraph of one survey of the data needs of science and technology might with only modest terminological change and a few omissions appear in a discussion of data needs and sharing in the social sciences (Lide, 1981:1349):

> We cannot take for granted that the data generated by the research establishment will automatically flow smoothly to those who need it. Changes in attitude are required by the scientific community, industry, and the federal government. The scientific community must place a higher priority on organizing the data it produces and presenting these data in a form suitable for technological applications. Private industry should put more resources into developing data bases to support long-term industrial needs. The federal government must recognize that its commitment to

> supporting basic research for the long-range benefit of the country also implies a commitment to make the results available in a form that maximizes their utility.

There are also differences. In discussions of data centers and facilities in the other sciences, heaviest emphasis seems to be placed on what might be termed base-line or reference data. These are data based on repeated measurements and are apparently intended to provide the typical or "best" values for particular phenomena or classes of entities or subjects. Discussions are frequently concerned with data about phenomena or subjects that have or can be assumed to have invariant characteristics and that can be measured repeatedly in diverse contexts. These are data collections to which a scientist might refer in attempting, for example, to identify a particular chemical compound or against which experimental or observational results might be compared to determine the degree to which the characteristics of a particular experimental population or set of observations depart from the norm. A report, "Study on the Problems of Accessibility and Dissemination of Data for Science and Technology" (1975), puts it as follows:

> Data with which we are concerned . . . may be regarded as the "crystallized" presentation of the essence of scientific knowledge in the most accurate form. Data, as usually understood in physics and chemistry, are numerical data representing the magnitudes of various quantities. . . . If we further include basic qualitative data such as the chemical structure of molecules, decay schemes of unstable nuclides, sequences of genes on chromosomes, etc., it may not be unrealistic to say that data constitute the reliable essence of scientific knowledge.

In the social sciences emphasis is placed on sharing data to allow their use for secondary analysis—in other words, for new research applications. In the other sciences it appears that heavier, although not exclusive, emphasis is placed on amassing data collections to serve as base-line data against which researchers can compare data that they have collected through their own experiments and observations. Individual researchers may deposit their data with a data center, but it is often to serve these base-line functions rather than to serve purposes of secondary analysis in the social science sense of the word. However, multiple research applications of data collections, in a fashion analogous to secondary analysis in the social sciences, does occur, most commonly in research areas in which costly and rare instrumentation is used for data collection. In these areas researchers cannot hope to consistently satisfy data needs through independent data collection. In areas in which data collection is easier and independent data collection is more consistently feasible, data sharing appears less common, and replication of reported research findings often occurs through new and independent data collection efforts.

CONCLUSIONS AND RECOMMENDATIONS

It is likely that something approaching consensus exists, at least in many areas of the social sciences, to the effect that data should be shared and available to all researchers. Consensus is strongest about large data collections assembled at public expense and less strong about smaller bodies of data collected at individual expense. The proposition that primary investigators should, at a minimum, have first rights of analysis and publication is generally accepted. There is probably less agreement as to mechanisms for data sharing. In some disciplines the expectation seems to be that data will be shared through intermediary facilities; in others, sharing occurs, if at all, primarily through relatively small networks of researchers working in the same area, although it is probable that recognition of the value and advantages of data-sharing organizations is becoming more widespread.

But even with this degree of acceptance of the principle of data sharing, a general norm of data sharing cannot be established and implemented by fiat. Changes in the attitudes of social scientists are required. While there is abundant evidence that the required change is taking place, the primary means to further change, particularly in the case of individual data collections, is moral suasion and demonstration of the value and scientific importance of sharing. There are also, however, more specific steps that could be taken to encourage and facilitate sharing.

One such step might be endorsement by professional associations and other prestigious social science organizations of the obligation to share data. Endorsements of this sort might, moreover, include well-reasoned statements of the value of sharing, discussions of data-sharing mechanisms and procedures, and illustrative examples of research and instructional gains made possible by data sharing.

Modest steps could also be taken to increase the incentives to share data. Citation practices could be improved to provide better recognition of original data collectors. Secondary analysts could be expected to provide complete citations of the data collections used and to acknowledge the original data collectors; journal editors might require such citations as a condition of publication. Secondary analysts might also give greater attention to noting innovations, matters of quality, and design advantages of data collections they use. Modest recognition could also be accorded to original data collectors through newsletter and journal notes when data collections are deposited with a data-sharing organization or otherwise made available for secondary use. In more general terms, some reassessment of the bases for professional rewards is probably in order. Design and execution of a major data collection effort is intellectually demanding, a creative accomplishment, and, when the product is shared with other researchers, a contribution to the development of scientific knowledge that should be better recognized and rewarded than it now is.

At least some of the existing disincentives to data sharing could be reduced if not eliminated. It is apparently true that support for research projects is sometimes sufficient for data collection but not completion of analysis and reporting findings,[13] and, as a consequence, data sharing is slowed or avoided entirely. To overcome this difficulty more adequate funding would be desirable to guarantee researchers the opportunity to reap the first fruits of their data collection. Research funding should also be adequate to support the costs involved in preparing and documenting data for use by others.

Technical obstacles to data sharing could be reduced. As noted at several points above, basic standards for data preparation and documentation are available. If these were routinely followed, data could be more readily shared, and it is likely that project costs would not be increased. Standards for documenting study and sample design and for complex derived and composite measures and indexes and the like are less well developed and adhered to. It should be recognized that documentation of this sort is of central importance to secondary analysis and a primary safeguard against erroneous or mistaken use of data. A small step toward improvement of this form of documentation could be taken by journal editors and peer reviewers. Requiring adequate documentation as a condition of publication would contribute to development of basic standards.

Depositing data with data-sharing organizations would probably be preferable to exclusive reliance on informal data sharing, although depositing data with an organization does not preclude simultaneous informal sharing by the data collector. The advantages of data-sharing organizations are several: they remove the burdens of supplying data from original data collectors; they maintain data collections and so the cumulative and combinatorial values of data are more likely to be realized; and they cross disciplinary boundaries so that interdisciplinary use of data is facilitated.

Data that threaten the privacy of individuals and the rights of organizations pose special problems. To reduce these problems the practice of removing names and other variables that would allow individuals to be identified should be consistently followed. This practice should be extended to include variables that allow individuals to be identified through a triangulation process. In the case of some data collections, however, individuals and organizations are intrinsically identifiable and to allow normal access to these data would abridge confidentiality. Various means can be used to allow limited access to such data for purposes of replication and secondary analysis. These include swearing in and licensing researchers to prevent misuse and provision of custom subsets and analyses that do not abridge confidentiality. Some of these same expedients might be employed in the case of proprietary data.

Questions are often raised as to what data ought to be shared, and distinctions are sometimes drawn between different categories of data in terms of the importance of sharing. Rather than begin with distinctions, it would be desirable to begin with the principle that all data ought to be shared with the reservation of special and limited forms of access for data that threaten privacy and confidentiality. Certainly data collected by government agencies, to the extent that questions of confidentiality and national interest are not present, should be readily and promptly available for research applications. The same rule should be followed for data collections commissioned for purposes of public policy and for performance evaluations. For these categories of data, it can be questioned whether delay of release to allow initial analysis and first rights of publication would be justifiable.

Data collected by individual researchers and research groups should also be made available to others in timely fashion. Some delay of release of data—a period of 1–2 years is often mentioned—to allow researchers to carry out analysis and publication is justifiable. One step toward institutionalizing such a practice would be for journal editors to require as a condition of publication that data be available to others. In the case of data collections supported by government funding agencies, stronger action is possible. Item 754 of the National Science Foundation's *Grant Policy Manual* "Rights in Data Banks and Software," is a significant step toward stating a general norm of data sharing. The item is worth quoting in full (National Science Foundation, 1983:vii–16):

> Unless otherwise provided in the grant letter, data banks and software produced with the assistance of NSF grants, having utility to others in addition to the grantee shall be made available to users, at no cost to the grantee, by publication or, on request, by duplication or loan for reproduction by others. The investigator who produced the data or software shall have the first right of publication. Grantees will be allowed a reasonable amount of time to make necessary corrections or additions to finite data banks that are incomplete or contain errors, ambiguities or distortions. Privileged or confidential information will be released only in a form that protects the rights of privacy of the individuals involved. Any dispute over the release or use of data or software will be referred to the Foundation for resolution. Any out of pocket costs incurred by the grantee in providing information to third parties may be charged to the third party.

The NSF statement has been in force, with some modifications, for over a decade and, along with numerous other Foundation actions, has done much to encourage and facilitate data sharing. The statement is strong, and it would be useful if, at a minimum, other research funding agencies would take a similar position. Even so, the statement falls short of the ideal. In the first place, it provides no guidelines as to the time of release. A primary investigator could delay release of data for half a decade, not an uncommon occurrence at present, and still be in technical compliance with the NSF policy. There is no statement as to the means by which data should be made available: willingness to supply copies if asked would be enough. There are also no provisions for special categories of data, except in the case of confidentiality, and the wording might suggest that the researcher need not provide any form of access to such data, although that is probably not the intent. Policy-relevant data and data of major concern to the research interests of large communities of scholars are not mentioned, and no reference is made to the technical form in which data would be released. There is also no indication of expectation that data would be conserved for any extended period to allow realization of the cumulative value of data collections.

Obviously we cannot expect a policy that specifies precise procedures for all occasions. On the other hand, we might imagine a policy that asked researchers to include in proposals a dissemination plan indicating the time of release of data, the means by which the data would be made available and preserved for long-term use, the technical form in which data would be released, the supporting documentation that would accompany the data, what forms of access to confidential or other sensitive data would be provided, and an assessment of the policy relevance and broad research value of the data. Peer reviewers could then judge the adequacy of the dissemination plan and suggest modifications.[14] Immediate release of data might be urged in cases of policy or broad research relevance. For agencies that commission data collections for policy evaluation, performance assessment, and the like, the requirement of immediate availability of data might be the norm. It may be worth noting in this respect that agencies that support development cf materials of broad scholarly utility—such as reference works, compilations, teaching aids, and the like—usually require that statements of plans for dissemination be included in proposals.

The utility of guidelines such as these in contributing to improvement of data-sharing practices would, of course, depend on the capacity of peer reviewers and agency officials to evaluate plans for dissemination. Here we have little to suggest. It is, after all, a matter of informing and convincing social scientists and agency officials of the values of data sharing, of the availability and utility of technical standards, of the need for long-term preservation of data, of difficulties encountered in transferring data and of means to overcome them, of the advantages (and disadvantages) of data-sharing facilities, and of the advantages (and disadvantages) of informal data sharing. In our experience substantial progress in each of these respects has been made in recent years, and we expect that progress will continue. We have suggested throughout this paper various steps that would speed progress.

NOTES

1. Some of the values of data sharing summarized here might be contested on the grounds that they rest on the notion that the development of scientific knowledge is a cumulative process; an alternative view might be that the development of scientific knowledge occurs through periodic and in some degree unpredictable quantum jumps involving basically new breakthroughs and departures. Even if this is the case, however, it would seem to follow that since breakthroughs and new departures are unpredictable, the opportunity for more social scientists to carry out meaningful research would increase their likelihood.

2. It is worth noting that the data used by Fogel and Engerman were made available to other scholars before their own analysis was completed and well before actual publication of *Time on the Cross*.

3. Lide (1981) calls attention to similar needs in the other sciences.

4. This difficulty is also suggested in a report to the Canada Council on survey research (Canada Council, 1976).

5. We can only ponder whether, at least in rare instances, willingness to cooperate with particular researchers but not others reflects an assumption on the part of subjects that the research findings will be favorable or at least not unfavorable.

6. We do not discuss the practices followed by researchers to provide security for information that links identification numbers used in data collections to actual names and addresses.

7. In this paper, open access to data does not mean free access. Provision of access to data usually involves a cost to the provider, and it is legitimate to transfer that cost to recipients of the data. Organizations that provide access to data also face the costs of sustaining themselves. Hence charges over and above actual costs of providing data are sometimes necessary.

8. Most of the academically based data archives are linked through the International Federation of Data Organizations (IFDO) based at the Universities of Cologne and Milan, and less directly through standing committees of the International Association for Social Science Council of UNESCO. Many of them are also members of the Inter-university Consortium for Political and Social Research. They are also linked through the International Association for Social Science Information Service and Technology—an international organization of individuals active in data organizations.

9. A number of the social science data archives mentioned above and in the appendix perform related functions beyond those of storing, processing, and disseminating machine-readable data. A few of them provide training in the use of data and related software (the ICPSR summer training program in the theory and technology of social research is an example, and the Social Science Research Council Survey Research Archive at the University of Essex also has a program); many will perform custom data analysis upon request; and a few have developed computer software and can provide software as well as assistance in the selection and use of computational facilities for social science research. Catalogues and lists of data holdings are available on request.

10. Other centers are located in Tokyo and Zurich. The centers operate under principles established by the International Council of Scientific Unions, as does the Centre de Donnees Stellaires in Strasbourg, France, which provides similar services. The world centers and their activities are described in International Council of Scientific Unions (1979).

11. Facilities for data sharing in the sciences are highly diverse. Museums often perform data-sharing functions through developing, maintaining, and providing access to specimens contained in their or known collections. One example is the automated herpetology collection of the Museum of Zoology at the University of Michigan: systematic information on over 300,000 specimens of amphibians and reptiles has been encoded and stored in computer-readable form by the museum. The information is accessible to interested researchers through the use of the TAXIR interactive information storage and retrieval system, which uses the main University of Michigan computer system, and can be interrogated remotely by scholars located at sites throughout the

country; for further information, see Van Devender (1978).

12. An entire issue of *Gastroenterology* in October 1979 was devoted to the National Cooperative Crohn's Disease Project and contains numerous other articles describing the project and its findings.

13. The prevalence of this difficulty is unknown; examples could be cited, however, and it is a frequent complaint of data collectors.

14. These suggestions follow recommendations made to the Canada Council by a special consultative group on survey research (Canada Council, 1976). The "Guide for Applicants" of the Social Science and Humanities Research Council of Canada (1979), formerly the Canada Council, includes provisions similar to those of the National Science Foundation.

APPENDIX
SELECTED LISTING OF DATA-SHARING FACILITIES

Behavioral Sciences Laboratory, University of Cincinnati, Cincinnati, Ohio 45221

Belgian Archives for the Social Sciences, Place Montesquieu, 1 Boite 18, B–1348 Louvain-la-Neuve, Belgium

Bureau of Labor Statistics, Division of Planning and Financial Management, U.S. Department of Labor, 441 G Street, N.W., Washington, D.C. 20212

Latin American Population Data Bank, United Nations Latin American Demographic Center (CELADE), Casilla 91, Santiago, Chile

Center for Quantitative Studies in Social Science, 117 Savery Hall, DK–45, University of Washington, Seattle, Washington 98195

Center for Social Analysis, State University of New York, Binghamton, New York 13901

Center for Social Sciences, Columbia University, 420 W. 118th Street, New York, New York 10027

Danish Data Archives, Odense University, Niels Bohrs Alle 25, KD–5230 Odense M, Denmark

Data and Program Library Service, 4451 Social Science Building, University of Wisconsin, Madison, Wisconsin 53706

Data Archives Library, Institute for Social Science Research, 1101 Gayley Center, 405 Hilgard Avenue, University of California, Los Angeles, California 90024

Data Bank, Institute for Behavioral Research, York University, 4700 Keele Street, Downsview, Ontario, Canada

Data Library, 6356 Agricultural Road, Room 206, University Campus, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5

Data Library, Survey Research Center, University of California, Berkeley, California 94720

Data Resources, Inc., 29 Hartwell Avenue, Lexington, Massachusetts 02173

Data User Service Division, Bureau of the Census, U.S. Department of Commerce, Washington, D.C. 20233

Drug Abuse Epidemiology Data Center, Institute of Behavioral Research, Texas Christian University, Fort Worth, Texas 76129

DUALabs, Inc., 1601 N. Kent Street, Suite 900, Arlington, Virginia 22209

European Consortium for Political Research, Data Information Service, Fantoftvegen 38, N–5036 Fantoft-Bergen, Norway

Inter-university Consortium for Political and Social Research, P.O. Box 1248, Ann Arbor, Michigan 41806

Leisure Studies Data Bank, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

Louis Harris Data Center, Manning Hall 026A, Institute for Research in Social Science, University of North Carolina, Chapel Hill, North Carolina 27514

Machine-Readable Archives, Public Archives of Canada, 395 Wellington Street, Ottawa, Ontario, Canada K1A 0N3

Machine-Readable Archives Division, (NNR), National Archives and Records Service, Washington, D.C. 20408

National Center for Education Statistics, Data Systems Branch, 205 Presidential Building, 400 Maryland Avenue, S.W., Washington, D.C. 20202

National Center for Health Statistics, Scientific and Technical Information Branch, Room 1–57 Center Building, 3700 East-West Highway, Hyattsville, Maryland 20782

National Center for Social Statistics, Office of Information Systems, Washington, D.C. 20201

National Opinion Research Center, University of Chicago, 6030 South Ellis Avenue, Chicago, Illinois 60637

National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Road, Springfield, Virginia 22151

Northwestern University Information Center, Vogelback Computing Center, Northwestern University, Evanston, Illinois 60201

Norwegian Social Science Data Services, Universiteet i Bergen, Christiesgate 15–19, N–5014 Bergen-University, Norway

Oklahoma Data Archive, Center for the Application of the Social Sciences, Oklahoma State University, Stillwater, Oklahoma 74074

Polimetrics Laboratory, Department of Political Science, Ohio State University, Columbus, Ohio 43210

Political Science Data Archive, Department of Political Science, Michigan State University, East Lansing, Michigan 48823

Political Science Laboratory and Data Archive, Department of Political Science, 248 Woodburn Hall, Indiana University, Bloomington, Indiana 47401

Project Impress, Dartmouth College, Hanover, New Hampshire 03755

Project TALENT Data Bank, American Institutes for Research, P.O. Box 1113, Palo Alto, California 94302

Public Opinion Survey Unit, University of Missouri, Columbia, Missouri 65201

Roper Public Opinion Research Center, Box U–164R, University of Connecticut, Storrs, Connecticut 06268

Social Data Exchange Association, 229 Waterman Street, Providence, Rhode Island 02906

Social Science Computer Research Institute, 621 Mervis Hall, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Social Science Data Archive, Laboratory for Political Research, 321A Schaeffer Hall, University of Iowa, Iowa City, Iowa 52240

Social Science Data Archive, Survey Research Laboratory, 414 David Kinley Hall, Urbana, Illinois 61810

Social Science Data Archive, Box 596, University of Notre Dame, Notre Dame, Indiana 46556

Social Science Data Archives, Department of Sociology and Anthropology, Carleton University, Colonel By Drive, Ottawa, Ontario, Canada K1S 5B6

Social Science Data Center, University of Connecticut, Storrs, Connecticut 06268

Social Science Data Center, University of Pennsylvania, 353 McNeil Building, CR, 3718 Locust Walk, Philadelphia, Pennsylvania 19104

Social Science Data Library, Manning Hall 026A, University of North Carolina, Chapel Hill, North Carolina 27514

Social Science User Service, Princeton University Computer Center, 87 Prospect Avenue, Princeton, New Jersey 08540

Social Security Administration, Office of Research and Statistics, Room 1120,, Universal North Building, 1875 Connecticut Avenue, N.W., Washington, D.C. 20009

SSRC Survey Archive, University of Essex, Wivenhoe Park, Colchester, Essex, England
State Data Program, 2538 Channing Way, University of California, Berkeley, California 94720
State Government Data Base, Council of State Governments, Iron Works Pike, Lexington, Kentucky 40578
Statistics Canada, 1006-General Purpose Building, Ottawa, Ontario, Canada K1A 0T6
Steinmetzarchief, Herengracht 410–412, 1017 BX Amsterdam, The Netherlands
The United Nations Statistical Office, The United Nations, New York, New York 10017
Zentralarchiv für empirische Sozialforschung, Universitaet zu Koeln, Bachemer Strasse 40, D–5000 Koeln 41, West Germany

REFERENCES AND SELECTED BIBLIOGRAPHY

Bancroft, T.A.
1972 The statistical community and the protection of privacy. *The American Statistician* 26(4):13–16.

Banks, A.S.
1973 Problems in the Use of Archival Data. Prepared for the Panel on Research Problems in Comparative Analysis, Annual Meeting of the International Studies Association, May 14–17, New York.

Benson, L.
1968 The empirical and statistical basis for comparative analysis of historical change. In Stein Rokkan, ed., *Comparative Studies Across Cultures and Nations*. Paris: Mouton.

Bick, W., and Muller, P.J.
1980 The nature of process-produced data—towards a social-scientific source criticism. In Jerome M. Clubb and Erwin K. Scheuch, eds., *Historical Social Research: The Use of Historical and Process-Produced Data*. Stuttgart, Germany: Klett-Cotta.

Bisco, R., ed.
1970 *Data Bases, Computers, and the Social Sciences*. New York: Wiley-Interscience.

Bogue, A.G.
1976 The historian and social science data archives in the United States. *American Behavioral Scientist* 19:419–442.

Bond, K.
1978 Confidentiality and the protection of human subjects in social science research. *The American Sociologist* 13(3):144.

Boruch, R.F.
1972 Strategies for eliciting and merging confidential social research data. *Policy Sciences* 3(3):275–297.

Boruch, R.F., and Reis, J.
1978 An illustrative project on secondary analysis. Pp. 88–111 in R.F. Boruch and P.M. Wortman, eds., *New Directions for Program Evaluation*. San Francisco: Jossey-Bass.

Boruch, R.F., and Wortman, P.M.
1978 An illustrative project on secondary analysis. *New Directions for Program Evaluation* 4:89–110.

Bowers, W.J., and Pierce, G.L.
1975 The illusion of deterrence in Isaac Ehrlich's research on capital punishment. *Yale Law Journal* 85:185–208.

Bowman, R.T.
1970 The idea of a federal statistical data center—its purpose and structure. Pp. 63–69 in Ralph L. Bisco, ed., *Data Bases, Computers, and the Social Sciences*. New York: Wiley Interscience.

Bryant, F.B., and Wortman, P.M.
1978 Secondary analysis: the case for data archives. *American Psychologist* April:381–387.

Byrum, J.P., and Rowe, J.
1972 An integrated user-oriented system for the documentation and control of machine-readable data files. *Library Resources and Technical Services* 16:3.

Campbell, A.
1970 Some questions about the New Jerusalem. In Ralph Bisco, ed., *Data Bases, Computers, and the Social Sciences.* New York: Wiley Interscience.

Campbell, D.T.
1968 A cooperative multinational opinion sample exchange. *Journal of Social Issues* 24(2):245–256.

Canada Council
1976 *Survey Research: Report of the Consultative Group on Survey Research.* Ottawa, Canada: The Canada Council.

Carmichael, N., and Parke, R.
1974 Information services for social indicators research. *Special Libraries* May-June:209–215.

Carroll, J.D.
1973 Confidentiality of social science research sources and data: the Popkin case. *PS—American Political Science Association* Summer:268.

CELADE
1975 Banco de datos de CELADE. *Data Banks and Archives for Social Science Research on Latin America* 6:114–118.

Center for Political Studies, The University of Michigan
1983 American National Election Studies, 1972–1980: Bibliography of Data Use March 64.

Centre D'Etudes Sociologiques
1969 France: note on the creation of a department for secondary analysis. *Social Science Information* 8:147–148.

Chandler, W.M., and Hartjens, P.G.
1969 Secondary analysis in the social sciences: report on an international conference. *Social Science Information* 8:37–47.

Chattopadhyay, M.
1974 *Some Distinctive Features of Data Bank Movement in Social Sciences.* Calcutta: Indian Statistical Institute.

Clubb, J.M.
1975 Sources for political inquiry II: quantitative data. *Handbook of Political Science* 7:43–78.

Clubb, J.M., and Scheuch, E.K., eds.
1980 *Historical Social Research: The Use of Historical and Process-Produced Data.* Stuttgart, Germany: Klett-Cotta.

Clubb, J.M., and Traugott, M.
1979 Data resources for community studies: the United States. In E. Summers and A. Selvick, eds., *Non-Metropolitan Economic Growth and Community Change.* New York: D. C. Heath.

Coleman, J.S., et al.
1966 *Equality of Educational Opportunity.* 2 Vols. Office of Education, U.S. Department of Health, Education, and Welfare. Washington, D.C.: U.S. Government Printing Office.

Committee on Data for Science and Technology
1975 Study on the problems of accessibility and dissemination of data for science and technology. *CODATA Bulletin* (October). Paris: UNESCO-UNISIST.

Converse, P.E.

1964 A network of data archives for the behavioral sciences. *Public Opinion Quarterly* 28(Summer):273–286.

1966 The availability and quality of sample survey data in archives within the United States. In R. L. Merritt and S. Rokkan, eds., *Comparing Nations: The Use of Quantitative Data in Cross-National Research*. New Haven: Yale University Press.

De Grolier, E.

1966 Short note on information retrieval systems applicable to archive data. Pp. 196–202 in S. Rokkan, ed., *Data Archives for the Social Sciences*. Paris: Mouton.

Dennis, J.

1971 The relation of social science data archives to libraries and wider information networks. Pp. 117–120 in J. Becker, ed., *Proceedings of the Conference on Inter-library Communications and Information Networks*. Chicago: American Library Association.

Derivry, D.

1972 A data-bank of French electoral statistics from 1945–1971. *Social Science Information* 11:309–316.

Deutsch, K.W.

1966 The theoretical basis of data programs. In R.L. Merritt and S. Rokkan, eds., *Comparing Nations: The Use of Quantitative Data in Cross-National Research*. New Haven: Yale University Press.

1970 The impact of complex data bases on the social sciences. In R. Bisco, ed., *Data Bases, Computers, and the Social Sciences*. New York: Wiley Interscience.

Deutsch, K.W., Lasswell, H.D., Merritt, R.L., Russett, B.M.

1966 The Yale political data program. In R.L. Merritt and S. Rokkan, eds., *Comparing Nations: The Use of Quantitative Data in Cross-National Research*. New Haven: Yale University Press.

Dodd, S.A.

1977 Cataloging machine-readable data files—a first step? *Drexel Library Quarterly* 13:1.

Dollar, C.M.

1980 Problems and procedures for preservation and dissemination of computer-readable data. In J.M. Clubb and E.K. Scheuch, eds., *Historical Social Research: The Use of Historical and Process-Produced Data*. Stuttgart, Germany: Klett-Cotta.

Dunn, E.S.

1974 *Social Information Processing and Statistical Systems—Change and Reform*. New York: John Wiley.

Edsall, J.T.

1981 Two aspects of scientific responsibility. *Science* 212(4490):11–14.

Ehrlich, I.

1975 The deterrent effect of capital punishment: a question of life and death. *American Economic Review* 65(3):397–417.

European Political Data Newsletter

n.d. European Consortium for Political Research, Data Information Service, Gamle Kalvedalsveien 12, N–5000 Bergen, Norway.

European Science Foundation

1980 Statement Concerning the Protection of Privacy and the Use of Personal Data for Research. Strasbourg, France.

Feldstein, M.S.

1974 Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* 82(5):905–926.

1980 *Social Security, Induced Retirement, and Aggregate Capital Accumulation: A Correction and Updating*. Working Paper No. 579. Washington, D.C.: National Bureau of Economic Research.

Feige, E.L., and Watts, H.W.

1970 Protection of privacy through micro-aggregation. Pp. 261–272 in R.L. Bisco, ed., *Data Bases, Computers, and the Social Sciences*. New York: Wiley Interscience.

Fogel, R.W., and Engerman, S.L.

1974 *Time on the Cross: The Economics of American Negro Slavery*. Boston: Little, Brown.

1974 *Time on the Cross: Evidence and Methods—A Supplement*. Boston: Little, Brown.

Garcia-Bouza, J.

1969a Latin America: a progress report on archival development. *Social Science Information* 8:153–158.

1969b The future development of social science data archives in Latin America. In M. Dogan and S. Rokkan, eds., *Quantitative Ecological Analysis in the Social Sciences*. Cambridge, Mass.: MIT Press.

Geda, C.L.

1979 *Data Preparation Manual*. Ann Arbor: Institute for Social Research.

Geda, C.L., Austin, E.W., and Blouin, Frances X., Jr., eds.

1980 Archivists and machine-readable records. In *Proceedings of the Conference on Archival Management of Machine-Readable Records*, February 7–10, 1979. Chicago: Society of American Archivists.

Gerhan, D., and Walker, L.

1975 A subject approach to social science data archives. *Research Quarterly* Winter:132–149.

Glasser, W.A.

1969 Note on the work of the Council of Social Science Data Archives, 1965–1968. *Social Science Information* 8:159–176.

Glenn, N.D.

1973 The social scientific data archives: the problem of underutilization. *American Sociologist* 8:42–45.

Gordis, L., and Gold, E.

1980 Privacy, confidentiality and the use of medical records in research. *Science* 207(11):153.

Hartenstein, W., and Liepelt, K.

1969 Archives for ecological research in West Germany. In M. Dogan and S. Rokkan, eds., *Qualitative Ecological Analysis in the Social Sciences*. Cambridge, Mass.: MIT Press.

Hastings, P.K.

1964a International survey library association of the Roper Public Opinion Research Center. *Public Opinion Quarterly* 28(Summer):332–333.

1964b The Roper Public Opinion Research Center. *International Social Science* Journal 16:90–97.

1966 Inventory of American production of survey data in 1963. Pp. 83–92 in S. Rokkan, ed., *Data Archives for the Social Sciences*. Paris: Mouton.

1975 Problems of data acquisition in Latin America: the Roper Public Opinion Research Center. *Data Banks and Archives for Social Science Research on Latin America* 6:70–103.

Hedrick, T.H., Boruch, R.F., and Ross, J.

1978 On ensuring the availability of evaluative data for secondary analysis. *Policy Sciences* 9:259–280.

Hofferbert, R.I.

1972 Data Archiving and Confidentiality in the International Comparative Study on the Organization of Research Units. Unpublished paper prepared for UNESCO Science Policy Division. Center for Social Analysis, State University of New York, Binghamton.

1976 Social science archives and confidentiality. *American Behavioral Scientist* 19:467–488.

Hofferbert, R.I., and Clubb, J.M., eds.

1976 Social science data archives: applications and potential. *American Behavioral Scientist* 19(4)(entire issue).

Hopkins, T.K., and Wallerstein, I.

1971 A proposal for a data bank of African materials. *Social Science Information* 10:135–147.

Hyman, H.H.

1972 *Secondary Analysis of Sample Surveys: Principles, Procedures and Potentialities*. New York: Wiley.

International Association for Social Science Information Service and Technology

n.d. IASSIST Newsletter. University of California, Los Angeles.

International Council of Scientific Unions, Panel on World Data Centres

1979 *Fourth Consolidated Guide to International Data Exchange Through World Data*. Washington, D.C.: Secretariat of the ICSU Panel on World Data Centres.

Jodice, D.H., Taylor, C.L., and Deutsch, K.W.

1980 *Cumulation in Social Science Data Archiving: A Study of the Impact of the 2 World Handbooks of Political and Social Indicators*. Königstein/Ts., Germany: Anton Hain.

Klingemann, H.D.

1967 Research and development of library-style retrieval systems for survey data archives. *Social Science Information* 6:119–135.

Lefcowitz, M.J., and O'Shea, R.

1963 A proposal to establish a national archives for social science survey data. *The American Behavioral Scientist* 6:27–31.

Leimer, D.R., and Lesnoy, S.D.

eries Evidence Using Alternative Social Security Wealth Variables. Paper presented at the 1980 meeting of the American Economic Association, Denver, Colorado, September.

Lide, D.R., Jr.

1981 Critical data for critical needs. *Science* 212:1343–1349.

Lipset, S.M.

1963 Approaches toward reducing the costs of comparative survey research. *Social Science Information* 2:33–38.

Lucci, Y., and Rokkan, S.

1957 *A Library Center for Survey Research Data*. School of Library Service. New York: Columbia University.

Madge, J.

1967 Great Britain: establishment of a social survey data bank. *Social Science Information* 6:185.

Martinotti, G.

1968 A note on the new institute of sociology in Milan. *Social Science Information* 7:31–35.

1969 Domains and universes: problems in concerted use of multiple data files for social science inquiries. In M. Dogan and S. Rokkan, eds., *Quantitative Ecological Analysis in the Social Sciences*. Cambridge, Mass.: MIT Press.

Mason, K.O., Winsborough, H.H., Mason, W.M., and Poole, W.K.

1973 Some methodological issues in cohort analysis of archive data. *American Sociological Review* 38:242–258.

Mendelssohn, R.C.

1967 The systems for integrated storage retrieval and reduction of economic data of the Bureau of Labor Statistics. *Social Science Information* 6:197–205.

Merritt, R.L.

1967 European public opinion and American policy: the USIA surveys. *Social Science Information* 6:143–160.

Merritt, R.L., and Lane, R.E.

1966 The training functions of a data library. Pp. 136–144 in S. Rokkan, ed., *Data Archives for the Social Sciences*. Paris: Mouton.

Merritt, R.L., and Rokkan, S., eds.

1966 *Comparing Nations: The Use of Quantitative Data in Cross-National Research*. New Haven: Yale University Press.

Miller, A.R.

1971 *The Assault on Privacy: Computers, Data Banks and Dossiers*. Ann Arbor: University of Michigan Press.

Miller, W.E.

1966 Inter-university Consortium for Political Research: current data holdings. Pp. 95–102 in S. Rokkan, ed., *Data Archives for the Social Sciences*. Paris: Mouton

1967 Promises and problems in the use of computers: the case of research in political history. In E. Bowles, ed., *Computers in Humanistic Research*. Englewood Cliffs, N.J.: Prentice-Hall.

1969 The development of archives for social science data. In M. Dogan and S. Rokkan, eds., *Quantitative Ecological Analysis in the Social Sciences*. Cambridge, Mass.: MIT Press.

1976 The less obvious functions of archiving survey research data. *American Behavioral Scientist* 19:409–418.

Miller, W.E., and Converse, P.E.

1964 The Inter-university Consortium for Political Research. *International Social Science Journal* 16:70–76.

Minister of Supply and Services

1976 Survey Research: Report of the Consultative Group on Survey Research. Ottawa, Ont.

1979 Guide for Applicants: Research Grants Program. Ottawa, Ont.

Mitchell, R.E.

1965 Survey materials collected in the developing countries: sampling, measurement, and interviewing obstacles to intra- and inter-national comparisons. *International Social Science Journal* 17:665–685.

1966 A social science data archive for Asia, Africa and Latin America. Pp. 103–121 in S. Rokkan, ed., *Data Archives for the Social Sciences*. Paris: Mouton.

1967 Abstracts, data archives, and other information services in the social sciences. Pp. 304–314 in D. Sills, ed., *International Encylopedia of the Social Sciences*. New York: MacMillan.

Nasatir, D.

1967 Social science data libraries. *The American Sociologist* 2:207–212.

1973 Data Archives for the Social Sciences: Purposes, Operations and Problems. Reports and Papers in the Social Sciences. UNESCO.

National Library of Medicine

1980 Laboratory Animal Data Bank. Fact sheet. Bethesda, Maryland.

National Research Council

1975 *An Assessment of the Impact of World Data Centers of Geophysics*. Washington, D.C.: National Academy of Sciences.

1976 *Geophysical Data Centers: Impact of Data-Intensive Programs*. Washington, D.C.: National Academy of Sciences.

1978 *National Needs for Critically Evaluated Physical and Chemical Data*. Washington, D.C.: National Academy of Sciences.

National Science Foundation

1983 *Grant Policy Manual*. Revised. NSF 77–47. Washington, D.C.: National Science Foundation.

Nelkin, D.

1981 Intellectual Property: The Control of Scientific Information. Unpublished manuscript, Cornell University.

Nesvold, B.A.

1976 Instructional application of data archive resources. *American Behavioral Scientist* 19:455–466.

Overhage, C.F.J., and Harman, R.J., eds.

1965 *INTREX: Report of a Planning Conference on Information Transfer Experiments*. Cambridge, Mass.: MIT Press.

Passell, P., and Taylor, J.B.

1975 The Deterrent Effect of Capital Punishment: Another View. Discussion paper 74–7509. Department of Economics, Columbia University.

Pool, I.S.

1965 Data archivist libraries. Pp. 175–181 in C.F.J. Overhage and R.J. Harman, eds., *INTREX: Report of a Planning Conference on Information Transfer Experiments*. Cambridge, Mass.: MIT Press.

Potter, A.M.

1967 Great Britain: Social Science Research Council data bank. *Social Science Information* 6:77–80.

1968 British social science research data bank. *Information Retrieval & Library Automation* 4:5–6.

Raben, J., and Marks, G., eds.

1980 *Data Bases in the Humanities and Social Sciences*. Amsterdam: North Holland.

Raven-Hansen, P.

1981 Quid Pro Quo for Public Dough. Paper presented to the New York Academy of Sciences. National Law Center, George Washington University.

Relyea, H.

1980 Freedom of information, privacy and official secrecy: the evolution of federal government information policy concepts. *Social Indicators Research* 7:137–156.

Roistacher, R.C.

1980 *A Style Manual for Machine-Readable Data and Their Documentation*. Report No. SD-T-3, NCJ–62766. Washington, D.C.: U. S. Government Printing Office.

Rokkan, S.

1962 The development of cross-national comparative research: a review of current problems and possibilities. *Social Science Information* 1:21–38.

1964 Organization: archives for secondary analysis of sample survey data: an early inquiry into the prospects for western Europe. *International Social Science Journal* 16:49–62.

1965 Second conference on data archives in the social sciences, Paris, 28–30 September 1964. *Social Science Information* 4:67–84.

Rokkan, S., ed.

1966a *Data Archives for the Social Sciences*. Paris: Mouton.

Rokkan, S.

1966b International efforts to develop networks of data archives. Pp. 11–32 in S. Rokkan, ed., *Data Archives for the Social Sciences*. Paris: Mouton.

1973 Data exchanges in Europe: the role of the European consortium. *European Journal of Political Research* 1:95–101.

1976 Data services in western Europe. *American Behavioral Scientist* 19:443–454.

Rokkan, S., and Aarebrot, F.

1969 The Norwegian archive of historical and ecological data: progress report, August 1968. *Social Science Information* 8:77–84.

Rokkan, S., Deutsch, K., and Merritt, R.

1963 International conference on the use of quantitative political, social and cultural data in cross-national comparisons, Yale University, 10–20 September 1963. *Social Science Information* 2:89–108.

Rokkan, S., and Scheuch, E.K.

1963 Conference on data archives in the social sciences. *Social Science Information* 2:109–114.

Rokkan, S., and Valen, H.

1966 Archives for statistical studies of within-nation differences. Pp. 122–127 in S. Rokkan, ed., *Data Archives for the Social Sciences*. Paris: Mouton.

Rose, R.

1974 The dynamics of data archives. *Social Science Information* 13:91–107.

Rowe, J.S.

1975 The use and misuse of government produced statistical data files. *RQ* 14:(3):201–203.

1978 Government documents in machine-readable form: microdata for studies of labor force participation. *Government Publications Review* 5(3):379–382.

1979 Publicly available machine-readable data files. *Population Index* 45(4):567–575.

Rozsa, G., and Foldi, T.

1980 International co-operation and trends in social science information transfer. Librarianship and Archive Administration. *UNESCO Journal of Information Science* 2(4):234–239.

Ruggles, R., and Ruggles, N.

1967 Data files for a generalized economic information system. *Social Science Information* 6:187–196.

Russett, B.M.

1966 The Yale political data program: experience and prospects. In R.L. Merritt and S. Rokkan, eds., *Quantitative Data in Cross-National Research*. New Haven: Yale University Press.

Russett, B.M., Alker, H.R., Deutsch, K., and Laswell, H.D.

1964 *World Handbook of Political and Social Indicators*. New Haven: Yale University Press.

Sasfy, J.H., and Siegel, L.

1981 The Impact of Privacy and Confidentiality Laws in Research and Statistical Activity. MITRE Corp. paper no. 81W00073.

Schellenberg, T.R.

1965 *The Management of Archives*. New York: Columbia University Press.

Scheuch, E.K., and Stone, P.J.

1964 The general inquirer approach to an international retrieval system for survey archives. *American Behavioral Scientist* 7:23–28.

1966 Retrieval systems for data archives: the general inquirer. In R.L. Merritt and S. Rokkan, eds., *Comparing Nations: The Use of Quantitative Data in Cross-National Research*. New Haven: Yale University Press.

Scheuch, E.K., Sonte, P.J., Alymer, R.C., Jr., and Friend, A.

1967 Experiments in retrieval from survey research questionnaires by man and machine. *Social Science Information* 6:137–167.

Schoenfeldt, L.F.

1967 The Project TALENT data bank. *Social Science Information* 6:161–173.

1970 Data archives as resources for research, instruction, and policy planning. *The American Psychologist* 25:609–616.

Smith, K.W., and Rowe, J.S.

1979 Using secondary analysis for quasi-emperimental research. *Social Science Information* 18(3):451–472.

Sobal, J.

1981 Teaching with secondary data. *Teaching Sociology* 8(2):149–170.

Sodeur, W.

1969 Specialized data archives as instruments of theory testing: with examples drawn from small-group leadership studies. *Social Science Information* 8:119–125.

Sprehe, J.T.

1981 A federal policy for improving data access and user services. *Statistical Reporter* 81–6:323–344.

Stewart, D.K.

1967 Social Implications of Social Science Data Archives. Technical Memorandum 379/000/00. Systems Development Corporation, Santa Monica, Calif.

Stone, P.

1980 A perspective on social science data management. In J.M. Clubb and E.K. Scheuch, eds., *Historical Social Research: The Use of Historical and Process-Produced Data*. Stuttgart, Germany: Klett-Cotta.

Taylor, C.L., and Hudson, M.C.

1972 *World Handbook of Political and Social Indicators* 2d ed. New Haven: Yale University Press.

Toxic Substances Strategy Committee

1979 Report to the President. Toxic Substances Strategy Committee, Washington, D.C.

Traugott, M.

1981 The availability of resources for public policy analysis from ICPSR.

Traugott, M.W., and Clubb, J.M.

1976 Machine-readable data production by the federal government. *American Behavioral Scientist* 19:387–408.

Traugott, M.W., and Haberkorn, S.B.

1981 The national archive of computer-readable data on aging. In P. Bagnell, ed., *Special Collections: Gerontology and Geriatrics*. New York: Haworth.

Traugott, M., and Marks, J.A.

1980 Data resources and services from the Criminal Justice Archive and Information Network. In J. Raben and G. Marks, eds., *Data Bases in the Humanities and Social Sciences*. Amsterdam: North Holland.

Trystram, J.-P.

1966 Data archives and regional planning in France. *Social Science Information* 5:81–87.

1971 From automatic documentation to the data bank. *International Social Science Journal* 23:285–292.

U.S. Congress, House Committee on Government Operations

1968 Thirty-fifth Report on Privacy and the National Data Bank Concept. House Report No. 1842. 90th Cong., 2d session.

U.S. Department of Commerce

n.d. Solar Geophysical Data. Parts I and II. 414. Boulder, Colorado.

U.S. Department of Health, Education, and Welfare

1980 Report on Request of NIH for Limited Exemption from the Freedom of Information Act. Ethics Advisory Board, U. S. Department of Health, Education, and Welfare.

Valkonen, T.

1969 Secondary analysis of survey data with ecological variables. *Social Science Information* 8:33–36.

Vandaele, W.

1978 Participation in illegitimate activities: Ehrlich revisited. Pp. 270–335 in A. Blumstein, J. Cohen, and D. Nagin, eds., *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. Washington, D.C.: National Academy Press.

Van Devender

1978 Computers, curators and catalogs. *ASC Newsletter* 6(3):25–29.

Voss, P.R.

1977 Population data in social science data archives: the survey holdings of the Roper Public Opinion Research Center 14:141–144.

Watson, J.D.

1968 *The Double Helix*. New York: Atheneum.

White, H.D., ed.

1977 *Reader in Machine-Readable Social Data*. Englewood, Colo.: Information Handling Service.

Winship, D.H., and Summers, R.W., et al.

1979 The National Cooperative Crohn's Disease Study: study design and conduct *Gastroenterology* 77:827ff.

Yesley, M.

1980 The Ethics Advisory Board and the Right to Know. *Hastings Center Report*. October:5–9.