SHARING SOCIAL SCIENCE DATA

Advantages and Challenges

Joan E. Sieber
editor

A SAGE FOCUS EDITION

SHARING SOCIAL SCIENCE DATA

Advantages and Challenges

Joan E. Sieber
editor

SAGE PUBLICATIONS
The International Professional Publishers
Newbury Park London New Delhi

For information address:

SAGE Publications, Inc.
2455 Teller Road
Newbury Park, California 91320

SAGE Publications Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

SAGE Publications India Pvt. Ltd.
M-32 Market
Greater Kailash I
New Delhi 110 048 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Sharing social science data : advantages and challenges / Joan E. Sieber, editor.
p. cm. — (Sage focus editions : 128)
Includes bibliographical references and indexes.
ISBN 0-8039-4082-3. — ISBN 0-8039-4083-1 (pbk.)
1. Communication in social sciences—United States. 2. Social sciences—Research—United States. 3. Information networks—United States. I. Sieber, Joan E.
H61.92.U6S53 1991
300'.72073—dc20 90-21958
CIP

FIRST PRINTING, 1991

Sage Production Editor: Judith L. Hunter

Preface

This book was written because several of us wanted to organize what we know about data sharing in the social sciences and share it with our colleagues. We are advocates of data sharing and hope to convey to others our vision of how science can be improved through the sharing of data and other materials that enable investigators to build upon the work of others.

In addition to the "several of us" who have authored chapters, there are the many whose efforts, support, and encouragement underlie this project. It is a great pleasure to acknowledge those persons:

A profound thanks goes personally from me, the editor, to my friend and colleague Conrad Taeuber, who is most fundamentally responsible for inspiring me to work on problems of data sharing. From 1977, when he first began educating me about the problems and values of data sharing, through meetings and private encouragement, he has, in his own inimitable and quiet way, made a most profound difference in my professional development. What can I say to such a great man of few words, except "Thanks, Con." Two other main supporting pillars of this project are Rachelle Hollander and Felice Levine of the National Science Foundation, whose intellectual support and encouragement, and whose hands-on participation in meetings and projects, led up to this book. Gary Melton and Bob Boruch provided some important contexts for the incubation of ideas leading up to this book. Vivian Weil's

kind and generative spirit, and her hospitality and friendship, also played a vital role at various junctures in this project. And Mark Frankel, head of the Office of Scientific Freedom and Responsibility of the American Association for the Advancement of Science (AAAS), supported this work in innumerable ways, including the hosting of an NSF sponsored conference entitled "Sharing Scientific Data" at AAAS on February 18, 19, and 20, 1988.

That conference resulted in a set of proceedings and in a symposium at the 1989 AAAS meeting in San Francisco, both of which are cited often in this volume. Both provided the impetus for this book. Warm acknowledgement and thanks are given to the conference participants: Jerry Clubb, Martin David, Nancy Flournoy, Mark Frankel, Gerald Gates, Daniel Hill, Rachelle Hollander, Arthur Kuflik, Felice Levine, Burt Pelto, Albert Reiss, Deborah Runkle, Myron Straf, Conrad Taeuber, Donald Vandeveer, and Vivian Weil. Equally warm acknowledgement and thanks are also due Roy Jenne, who participated with several of us in the AAAS symposium and educated us about data sharing among atmospheric scientists.

Another friendly force that was important in the formative stages of this effort is the National Science Foundation, whose material support via Grant BBS-8711559 is gratefully acknowledged.

As editor of this volume, I thank the contributors for being pleasant and supportive throughout, and my students Julie Robbins and Warren Zittel for assistance with manuscript preparation. I especially thank my coauthor, Bruce Trumbo, who has been a sustaining force in this work—critiquing and just being a very good friend.

Finally, it has been pure pleasure to work with C. Deborah Laughton, Sage editor.

—Joan E. Sieber

Contents

1

Introduction

Sharing Social Science Data

JOAN E. SIEBER

Openness is a familiar idea in scientific research. Open communication enables science to be self-correcting and cumulative, and to benefit from multiple perspectives. In an open scientific environment, scientists can examine and build upon the work of others, verifying, refuting, or refining that work as appropriate. With open scientific communication, any appropriately trained scientist or student can enter vigorously and productively into an area of work that has been begun by others.

Until recently, openness in science has meant publishing one's methods and results. Two exceptional circumstances in which social science data have been shared are (a) through academic, commercial, public and government archives and (b) within the exclusive networks of working scientists that constitute the *invisible college* (Crane, 1972). However, with a few very recent exceptions, archived data have been more of historical than of current research interest, have often not been available in well-documented or user-friendly form, and have been available only within a limited range of social science subdisciplines (e.g., voting patterns, consumer behavior). As for sharing within the invisible college, many social science faculty are not part of any exclusive network of working scientists; because these teachers have no important data to offer, the norms of reciprocity make them hesitant to ask for data from scientists working at the cutting edge in their field.

Now, at the urging of some funders, editors, and scientific societies, many scientists are sharing actual raw data, documented as to its characteristics and idiosyncrasies so that others can understand and use it correctly. Although a relatively new practice in many parts of

science, data sharing has already brought to the social sciences important new solutions to old problems. As examples:

Demography employs large samples to provide a broad description of populations, whereas psychology employs small samples, experimentation, repeated measures, or longitudinal design to probe the dynamics of behavior. Important scientific and practical questions (e.g., cross-cultural questions about family dynamics, fertility, health, and education) would benefit from a combination of these two approaches. Two data sharing procedures, described by V. Jeffery Evans (Chapter 2), enable scientists to achieve this feat. They either graft some psychological techniques onto demographic studies, or adapt small-scale psychological techniques for use on diverse populations. One suspects that various techniques for oversampling and multistage sampling will evolve to allow researchers to focus in-depth behavioral research efficiently on selected groups. Similarly, *in a world that is witnessing large-scale migration, how can psychology, ethnography, and demography combine to examine the individual experience, attitudes, and behavior of migrants in relation to characteristics of the sending and receiving cultures?* In Chapter 2, Evans describes another project, based on data sharing, that answers this question.

Scientific hypotheses about major processes affecting contemporary and past human populations have been largely untestable until recently. Now, they are testable through current anthropological developments in comparative methodology such as modeling of spatially distributed data and time series. As Douglas White describes in Chapter 3, linguistic, ethnographic, primatological, paleontological, archaeological, and ecological data are being assembled and distributed for wide use in instruction and secondary analysis. These data and methods are now being combined with those of other disciplines: earth and environmental sciences, remote sensing, world-system demography, and development impact studies. Through data sharing and electronic methods, our understanding of the development and transformation of cultures and societies is becoming far more complete. These methods make new knowledge available not only to scientists and policymakers, but to students via classroom use of computerized data.

The complexities of conducting controlled experiments in field settings can overwhelm the individual researcher and result in poor quality data. In Chapter 4, Robert F. Boruch, Albert Reiss, Jr., *et al.* show how simultaneous sharing and consulting among investigators at different sites can improve the quality of data and inference.

How can someone who didn't do the initial research understand the data? Before standards of data documentation were developed,

misunderstanding and unhappy outcomes were likely to mar data sharing relationships. Now, data sharing standards such as those presented by Martin David (Chapter 5) can solve this problem and three others: (a) The description enables others to understand the data. (b) It allows the initial investigator to return to the data long after the needed details have faded from memory. (c) It forces the initial investigator to be systematic and rigorous in understanding the limitations of the data (e.g., details of the sampling procedure, reliability of the instruments, details of the original research design, and any deviations). And (d) it provides a basis for more systematic building on a sample, a procedure, or a body of knowledge.

When seeking to use data generated by others, investigators worry that the data will be hard to obtain, that the data will turn out to be of poor quality with poor documentation, or will involve measures that turn out upon closer scrutiny to be irrelevant to one's own concern. Not all of the related data sets that should be compared or combined may be available. Federal agencies are now funding the development of user-friendly archives that preclude these frustrations and inefficiencies. In Chapter 6, psychologists Josefina J. Card and James L. Peterson, of Sociometrics Corporation, describe how a centralized data archive meets the needs of data donors and recipients, broadening the use of high quality studies, and producing a vast increase in scientific knowledge at a greatly lowered cost per publication.

The formality of statistical theory and the tedium of statistical computation overwhelm many social science students and focus their attention on artificial models. Many fail to learn how to work intelligently with real data, or how to use good statistical intuition about when statistics are telling the truth. In Chapter 7, Bruce E. Trumbo and I describe how students' attention is refocused in a computer-based curriculum using important real data files and instruction in exploratory data analysis. Theory and computation do not dominate, but are understood in the service of intelligent statistical exploration and analysis of real data from which students can draw valid inferences about the real world and grasp the scientific and policy implications of their work. *A second significant problem in teaching research methodology has to do with teaching the skills and values underlying openness and peer review in science.* The desire to get a good grade provides little incentive for students to practice skills of critical evaluation of data and conclusions when the research was done by the student or the professor. But, as Chapter 7 illustrates, an entirely different set of goals, skills, and values can be learned when initial lessons of critical examination

and data documentation are taught with documented data generated by someone not known personally by the student or the professor.

Those opposed to data sharing have emphasized worst cases, whereas those in favor have pointed to the obvious advantages. As data sharing evolves from a voluntary activity among members of the invisible college to a funder-mandated activity, social scientists and science administrators have reason to wonder what challenges await them. Systematic examination of the emerging challenges is provided in Part III. In Chapter 8, I examine the challenges that confront individual investigators; in Chapter 9, Vivian Weil and Rachelle Hollander examine the challenges and questions that data sharing poses to research institutions.

A Shift Toward Data Sharing

Until recently, data sharing and many of the advantages of openness have been ideals, not norms, of science. They have been honored more often in the breach than in practice in many of the sciences—including the social sciences, which are the focus of this book. Research journals and monographs rarely include the raw data or other details that students or professionals would require in order to examine the work fully. Yet, anecdotal evidence suggests that most scientists have not expected other investigators to provide access to their important or current data, nor have they wished to share their own data.

Reasons social scientists have given for not sharing data have been examined by a committee of the National Academy of Science (Fienberg, Martin, & Straf, 1985), and by others (e.g., Ceci, 1988; Hedrick, 1988; Melton, 1988; Sieber, 1988; Stanley & Stanley, 1988). Frequently cited concerns center upon the condition and documentation of the data set, the effort and costs of sharing data, the uses to which the data will be put (including fear of criticism), and the qualifications of data requesters. Although these objections are far from insurmountable, they are firmly rooted in the traditional ways of science. However, the practices and traditions of science are changing in ways that foster openness and data sharing.

Several factors are currently contributing to this shift toward openness and sharing. An ethical consensus is emerging that scientists are obligated to make all of their data available for peer scrutiny, especially when findings are new or controversial (Fienberg, Martin, & Straf, 1985). The technology for storing and distributing large amounts of data has recently become inexpensively available in such forms as high

density diskettes or compact disk read-only memory (CD-ROM). Statistical analysis programs of considerable power and often with self-documenting features are widely available for use on microcomputers. Examples of well-documented data sets and examples of documentation standards have become available to researchers. Finally, data sharing is now required by key funding agencies, urged by some professional associations, and recommended by the National Academy of Sciences.

This monograph presents the ideas and contributions of social scientists who have enhanced our understanding of the ways in which data sharing can improve science. In Chapters 2 through 7, innovations in data sharing are illustrated in a variety of examples from anthropology, sociology, psychology, economics, demography, criminology, and human development. These examples present newcomers to data sharing with a range of information about key archives and sharing relationships, and with a sense of the opportunities that lie ahead for the individual social scientist and for the social sciences. Chapters 8 and 9 examine the challenges that remain to be resolved by individual scientists and by scientific institutions.

How Data Sharing Came About

Data sharing has a relatively long history in certain disciplines where useful primary data are more feasibly gathered by organizations or groups of scientists than by individuals working alone. Three brief examples illustrate some well-established data sharing practices motivated by such considerations:

Example 1: Academic economists have neither the funds nor the access to obtain the quantity or quality of data that are gathered by government and business. Consequently, even before computers simplified data sharing, economists performed secondary analyses on economic data obtained from government and business reports, newspapers, and journals. Today, the sharing of economic data is vastly simplified because government and industry make machine-readable data available to academic economists. In turn, economists who increase the value of data files by merging data sets or by weighting or indexing certain variables in accord with theory make those value-added resources available for tertiary analysis by other economists. Journals and textbooks in economics, and increasingly in other sciences as well, sometimes include diskettes containing some of the data described therein so that readers may readily perform secondary analyses of interest to them (David, 1988).

Example 2: For over a century, meteorologists have freely engaged in an international exchange of temperature and rainfall data, without asking questions about how they are used. These data are based on an international standard for weather observations adopted by the Vienna Congress of 1873. The number of participating weather stations has increased steadily. For example, 1,632 Indian stations provided daily precipitation data in 1901; 2,536 stations in 1970 (Jenne, 1989). The greatest challenge to international sharing of meteorological data has been the cost of storing this immense archive—a challenge that diminishes as computer technology advances. For example, the cost of magnetic tape for storing a terabit (10^{12} bits) of data was $175,000 in 1960, but by 1988 this had been reduced to $268 (Jenne, 1989).

In contrast to typical concerns about possible inaccuracy of data gathered by others, international sharing of meteorological data has demonstrated that imperfect data are better than no data at all. And security concerns have paled in contrast to the advantages of obtaining weather data daily from other nations throughout the world. These data have been used in all of the other geophysical sciences (e.g., solar terrestrial physics, meteorology, oceanography, glaciology, paleoclimatology, and solid earth geophysics), as well as in biology, agriculture, demography, geography, economics, and in the study of such current world concerns as El Niño, global warming, and droughts.

In contrast to the collection of temperature and rainfall data, which can be conducted through simple, inexpensive daily measurements, other geophysical sciences—astronomy, oceanography, space exploration—involve extremely expensive equipment and travel. Essential to cost-effective exploration and discovery are the sharing of equipment (e.g., telescopes), of space aboard the vehicles from which data are gathered (e.g., space or deep-sea habitats), of specimens (e.g., moon rocks, a slice of the core taken from polar ice or from the earth's mantle), and of data.

Example 3: The Human Genome Project, a scientific mission to probe the genetic structure of the human organism, is also establishing ground rules whereby geneticists, internationally, are pooling and sharing their data. Even more than most scientific activities, the success and usefulness of the Human Genome Project depends upon sharing. One goal of this basic research project is to develop a correct and complete map of human gene sequences. It is expected that many important applications will be discovered once geneticists obtain a valid description of these gene sequences (Roberts,1990).

Data Sharing in the Social Sciences

Within the last 25 years, computers have made it possible for social scientists in some disciplines to use shared data as a way to build efficiently on prior research. Data from many large-scale studies are now archived and available to secondary users in a well-documented form.

Many major studies of social and political behavior are archived at the Inter-university Consortium of Political and Social Research (ICPSR) at the University of Michigan, and provided by annual subscription to about 300 universities. Any faculty member or supervised student at a subscriber university may obtain the documented data at no additional charge. Researchers not located at ICPSR member institutions pay a modest fee for access based on the size of the data collection requested. ICPSR data are available on magnetic tape, floppy disk, and printout. ICPSR data are "graded" on a four-point scale according to the extent to which they have been checked, corrected, and otherwise rendered more user friendly by ICPSR staff. ICPSR provides periodic catalogues describing its holdings—for example, *Data Collections from the National Archive of Computerized Data on Aging* (ICPSR, 1987), and *Data Collections from the Criminal Justice Archive and Information Network* (ICPSR, 1988). Summer workshops are provided by ICPSR to train scientists in quantitative methods with special attention to some of the kinds of data archived at ICPSR.

Another major social science data sharing resource is the General Social Survey (GSS). Conducted annually since 1972 by the National Opinion Research Center (NORC), the GSS is a key source of data to many researchers in sociology, political science, psychology, demography, economics, mass communication, business, education, health care, criminal justice, and other applied social sciences. The GSS archive is distributed directly by the National Opinion Research Center, as well as by other archives such as the Roper Center at the University of Connecticut, and ICPSR. Smith and Crovitz (1988) document over 1,700 social science publications produced between 1976 and 1986 that are based on secondary analysis of the General Social Survey.

More recently, the personal computer has made it feasible to create and disseminate entire archives on floppy disks, or a CD-ROM. To increase the amount of knowledge and the educational value that can be derived from major studies in an important area of research, funders such as the National Institute of Child Health and Human Development (NICHD) and the National Institute of Justice (NIJ) now solicit proposals from commercial archiving firms to process sets of the most

outstanding studies in designated fields of interest onto a single CD-ROM. Studies on a given topic are selected for inclusion by a panel of experts. The firm that is awarded funding then obtains the data, prepares uniform documentation and formatting, and processes the data into a user-friendly archive.

Since 1982, the major social science archiving corporation performing this work has been Sociometrics Corporation. In Chapter 8, the president of Sociometrics, Josefina J. Card, with coauthor James L. Peterson, discusses the archiving and marketing of such archives. Major archives produced by Sociometrics include "Adolescent Pregnancy and Pregnancy Prevention" sponsored by the Office of Population Affairs, "Family Data Archive" sponsored by the National Institute of Child Health and Human Development, "Archive of Social Research of Aging" sponsored by the National Institute of Aging, "Philippine Population Data Archive" sponsored by the International Development Research Center (Canada), and the "Criminal Justice Archive" sponsored by the National Institute of Justice. In addition, Sociometrics is developing new ways of searching for and retrieving information from social science data files, under funding from the National Science Foundation.

In contrast to the task of archiving sets of related studies in a designated field of interest, economists have encountered the challenge of organizing into user-friendly form the massive data from a single, ongoing national survey. Economist Martin David, author of Chapter 5, is the principal designer of an archive for data from the National Survey of Income and Program Participation (SIPP). Because the SIPP is one of the most complex sets of social measurements ever undertaken, few could master the data and develop scientific analyses from them without assistance. The archive links applications of the data and discoveries of users into the system. It provides several mechanisms for communication among data users and between data users and collectors. The retrieval and documentation of data is an ongoing process; as users discover new properties of the data, this information is added to the documentation.

In addition to the data that may be obtained from these and other major archives, the data from many individual studies are now available from initial investigators upon the request of qualified secondary users. Available data in the field of developmental psychology, especially longitudinal studies, have been cataloged by the Social Science Research Council. *An Inventory of Longitudinal Research on Childhood and Adolescence* (Verdonik & Sherrod, 1984) describes 116 studies. The catalogue description of each data file on childhood and

adolescence includes: (a) title of the study, (b) names, affiliations, and addresses of the investigators and designation of the contact person, (c) topical domains or disciplines covered, (d) substantive topics or research questions covered, (e) characteristics of the original sample, (f) years of completed waves, (g) information on sample attrition, (h) instruments used for data collection and constructs measured, (i) future data collection, plans, (j) representative references, and (k) the current status of the study, that is, whether it is active, whether the sample is available for further study to the original principal investigators who collected the data, whether the data are available for secondary analysis,and whether the data are on computer tapes. A second catalogue, *An Inventory of Longitudinal Studies of Middle and Old Age* (Migdal, Abeles & Sherrod, 1982), describes 76 studies and is slightly less comprehensive.

Other obvious sources of data are investigators who have been funded by foundations that require data sharing. Since 1985, the National Science Foundation has required that social science investigators agree to deliver their data, along with adequate documentation, to some designated public archive within one year after project completion. Thus a two-year study funded in 1985 should be archived by 1988. Lists of projects funded each year by a given NSF program may be obtained from that program. The NSF (1989) statement stipulates that:

> *Unless otherwise provided in the grant letter, data blanks and software,* produced with the assistance of NSF funds, having utility to others in addition to the grantee shall be made available to the user, 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 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.

In contrast, the National Institute of Justice, which began to require data sharing in 1976, builds the cost of data sharing into its contracts. It requires that its contractors deliver to NIJ, upon project completion, computer-readable copies and adequate documentation of all data bases and programs developed or acquired in connection with the research. The grantee must remove individual identifiers from any data and

programs prior to submission to NIJ. Most of these data files are then archived at ICPSR.

This discussion of noteworthy social science data archives is intended to illustrate the diversity of well-organized commercial and academic archives currently in existence, not to provide an exhaustive listing of social science archives. [A list of 56 key social science archives is to be found in pp. 77-79 of *Sharing Research Data* (Fienberg, Martin, & Straf, 1985).] As the above examples suggest, the diversity of archiving arrangements is responsive to a variety of factors including the nature and importance of the data, and the arrangements funders make to support data documentation, improvement, and archiving.

Finally, data are becoming more readily available from individual scientists even if these data are not cataloged or specially archived anywhere. The norm of data sharing is becoming established in the social sciences, and some journals require authors to retain their data for sharing for at least five years. Hence, it is reasonable to expect that many investigators whose data are not available under any of the arrangements described above would respond favorably to requests for their documented data.

Federal Mandates

In the latter part of the 1980s data sharing has received a dramatic impetus—congressional concern about fraud and carelessness in scientific research. In testimony before the Subcommittee on Investigations and Oversight of the Committee on Science, Space, and Technology of the House of Representatives on June 28, 1989, Dr. Lyle Bivins, Director of the Office of Scientific Integrity Review, Department of Health and Human Services, expressed his interest in resolving problems surrounding data derived from Public Health Service research. These include issues of ownership and retention of data and of ways in which the scientific community can facilitate data sharing. Recently, various committees have been appointed and conferences have been held to grapple with these questions.

In April 1989, the National Science Foundation greatly expanded its sharing requirements to include all of the disciplines it funds, and to include not just "data" in the usual sense of the word, but all materials that others might need in order to build upon the work of the original investigator. Erich Bloch, Director of the National Science Foundation, sent a memo to presidents of colleges and universities and heads of