Sharing Research Data

Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf, Editors; Committee on National Statistics, National Research Council

ISBN: 0-309-54204-9, 240 pages, 6 x 9, (1985)

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

COMMITTEE ON NATIONAL STATISTICS

SUBCOMMITTEE ON SHARING RESEARCH DATA

STEPHEN E. FIENBERG (Chair), Department of Statistics, Carnegie-Mellon University
CLIFFORD G. HILDRETH, Department of Economics, University of Minnesota
LESLIE KISH, Institute for Social Research, University of Michigan
EDWARD R. TUFTE, Department of Political Science, Yale University
MARGARET E. MARTIN, Staff
MIRON L. STRAF, Staff

PREFACE

This report originated from a letter sent in May 1979 by Professor Melvin Reder of the University of Chicago School of Business to the executive director of the Committee on National Statistics (CNSTAT). Professor Reder proposed a conference on the sharing of social science research data to examine and discuss the conflicting pressures affecting researchers regarding the disclosure to others of data and preliminary analyses.

Such a conference, chaired by Clifford Hildreth, was held in October 1979. The participants raised many points and recommended further work by CNSTAT. The committee expresses its thanks and appreciation to the participants, who are listed in the appendix to this volume. In response to the conference recommendation, the Sloan Foundation provided the committee with a grant to work toward the development and dissemination of guidelines for the sharing of scientific data, and the System Development Foundation provided a further grant for work on this report. The study was also supported by a consortium of federal agencies that provide funding for the general activities of CNSTAT.

A subcommittee of CNSTAT members was appointed to oversee the project; it was responsible for obtaining and reviewing commissioned papers, developing a set of guidelines for sharing data, and preparing this report for the committee. Although some of their terms of appointment on the full committee

expired, all subcommittee members continued to serve throughout the study.

We were fortunate to obtain the services and cooperation of several scholars who prepared papers following a general outline developed by the subcommittee. The commissioned papers are Part II of this volume and represent different vantage points on the issues of data sharing. The sub committee is especially appreciative of the detailed materials and suggestions contained in these papers and has relied heavily on them in formulating and structuring the discussion of the costs and benefits of data sharing as well as in developing its recommendations.

The first paper, prepared at the Inter-university Consortium for Political and Social Research at the University of Michigan by Jerome M. Clubb with coauthors Erik W. Austin, Carolyn L. Geda, and Michael W. Traugott, deals primarily with large social science data sets. The other four papers deal with the advantages and disadvantages of data sharing more broadly. The paper by Robert F. Boruch of the Department of Psychology at Northwestern University describes products of data sharing. The paper by Terry E. Hedrick of the Institute for Program Evaluation of the U.S. General Accounting Office discusses justifications for and obstacles to data sharing. The paper by Joe Shelby Cecil of the Federal Judicial Center and Eugene Griffin of Northwestern University discusses legal issues relevant to data sharing and provides an important analysis of current pertinent law. And the paper by Robert F. Boruch and David S. Cordray of the Department of Psychology at Northwestern University suggests professional codes and guidelines for data sharing.

Margaret E. Martin and Miron L. Straf served as staff of the subcommittee and coeditors of this report. Lenore Bixby prepared a report of the early conference that led to the development of this study. Eugenia Grohman contributed greatly in editing our manuscript and guiding it toward publication. Valuable assistance was provided by Roberta Pirosko in bibliographic work and in typing and by Diane Goldman in proofreading and manuscript preparation. Using the computer for word processing, telecommunications, and typesetting, Lee R. Paulson prepared many versions of our manuscript; she also provided bibliographic and other research assistance. Reviewers and many others offered valuable comments and suggestions for our report. To all who have worked with us or otherwise contributed, we are very grateful.

The committee views this report as an initial examination of some of the issues of data sharing, on which readers are invited to comment.

STEPHEN E. FIENBERG, CHAIR
COMMITTEE ON NATIONAL STATISTICS
May 27, 1985

CONTENTS

PART I:

REPORT OF THE COMMITTEE ON NATIONAL STATISTICS

Issues and Recommendations

INTRODUCTION

Data are the building blocks of empirical research, whether in the behavioral, social, biological, or physical sciences. To understand fully and extend the work of others, researchers often require access to the data on which that work is based. Yet many members of the scientific community are reluctant or unwilling to share their data even after publication of analyses of them. Sometimes this unwillingness results from the conditions under which data were gathered; sometimes it results from a desire to carry out further analyses before others do; and sometimes it results from the anticipated costs, in time or money or both.

The Committee on National Statistics believes that sharing scientific data with colleagues reinforces the practice of open scientific inquiry. Cognizant of the often substantial costs to the original investigator for sharing data, the committee seeks to foster attitudes and practices within the scientific community that encourage researchers to share data with others as much as feasible.

Some examples illustrate the benefits, problems, controversies, and other consequences of sharing research data.

Reanalysis of shared data may lead to a conflicting conclusion. Because an original investigator published his raw data on measurements of human cranial capacity by race and described his procedures and methods of summarization, reanalysis of the data was possible. A reanalysis more than 120 years later overturned the original investigator's conclusions (Gould, 1978).

Confidentiality may be breeched by legally imposing sharing data. Despite promises of confidentiality to respondents, researchers may be in jeopardy of arrest if police or the courts request or demand data. A study headed by James Carroll at Syracuse University on the confidentiality of social science research sources and data identified many such cases (Carroll and Knerr, 1975); one was the Office of Economic Opportunity's New Jersey negative income tax experiment, in which a local prosecutor issued 14 subpoenas requesting the names of welfare families receiving excess payments (Kershaw and Fair, 1976).

When data are not shared, an investigator's results may have a greater influence on public policy than if the data are analyzed by others. An economist prepared a paper on the deterrent effect of capital punishment, in which he concluded that one execution prevents eight murders. A draft version of this paper was used by the Solicitor General of the United States as an appendix to the government's pro-capital punishment brief in a case before the Supreme Court. Detailed data were not available for reanalysis. Other researchers have now assembled what are believed to be virtually identical data sets, and many analysts believe the data do not support the deterrence hypothesis.

Marketing of biomedical research militates against data sharing. Several university researchers have refused to share with colleagues the exact details of how they did experiments that were reported in papers submitted for publication because such details might compromise the profit-making potential of their work.

Sharing proprietary data may be forbidden by the originator of the data. A distinguished professor of business is carrying out research based on data from a firm that not only does not want others to see the data, but is not even willing to be identified. The professor considers the research useful, but is disturbed because the conditions under which he obtained the data preclude the possibility of anyone verifying his statistical analyses.

These and other situations fuel an ongoing debate in the research community on what are appropriate principles and practices of data sharing.

Issues in Data Sharing

The Committee on National Statistics convened a conference on sharing social science research data in October 1979, chaired by Clifford Hildreth (see Committee on National Statistics, 1980; see the appendix for a list of participants). The participants were in substantial agreement regarding the exigencies faced by social science researchers and how these often conflicted with goals of greater access to data. The issues they considered included whether there is ever justification for refusing or unduly postponing access to data; the impact on data access of data collectors' responsibility for maintaining the privacy of respondents and the confidentiality of records; the professional responsibility of researchers to promote access; and procedures under which basic data should be released to others.

The conference participants presented the Committee on National Statistics with the following conclusions:

1. Guidelines on data sharing need to be developed. Desirable practices may vary with the source of the data and whether the research is publicly or privately funded.
2. A variety of institutions could be helpful in promulgating guidelines for desirable practices. The institutions include professional associations and their journals, consortia for data archiving, and foundations and other organizations that fund research.
3. Government policy on access to data is important. Much social science research relies heavily on data provided by the government directly or indirectly through grants and contracts for research.
4. Many problems of access to data in the natural sciences are similar to those in the social sciences.
5. Standards for classifying, documenting, and archiving data would greatly facilitate access to data.

In response to the conclusions of the conference, this report suggests guidelines for appropriate sharing of data and how government agencies and other institutions can encourage and foster such sharing of data.

Scope of the Report

The exploratory conference focused on the sharing of *social science* research data. Most people believe that natural scientists have fewer problems in sharing data than do social scientists. The need for shared data may be less acute for natural science experiments, which usually are replicable—a situation that occurs more rarely in the social sciences. Nonetheless, data-sharing problems

have existed in the natural sciences that are really not much different from those in the social sciences, such as instances in which only some observations are reported rather than all.

Selective reporting of experimental results in the physical sciences is not uncommon. For example, Millikan's 1910 *Science* paper on the oil drop experiment (see Holton, 1978) gave results based on 27 observations, although 40 observations were available; the most extreme 13 values were dropped. Similarly, in a 1919 report to the Astronomical and Royal Societies on expeditions to test predictions of Einstein's general theory of relativity, Eddington chose not to mention the results of one complete set of measurements that produced a value for the deflection of starlight consistent with the Newtonian, rather than the Einstein, prediction (see Earman and Glymour, 1980).

Some data-sharing problems in the biomedical sciences are also similar to those in the social sciences: for example, problems associated with large-scale, controlled clinical trials closely resemble those associated with large-scale social surveys. For these reasons, and because of the interests of the Committee on National Statistics in areas such as clinical trials, public health, and environmental monitoring, this report looks beyond the social sciences and addresses the issues of data sharing more broadly. The emphasis of the report remains on problems and practices in the social and behavioral sciences, but occasional links and parallels to the natural and biomedical sciences are identified and pursued.

This report specifically does not address two kinds of research. The first is research with nonquantitative data. Researchers often depend on materials other than quantitative information, such as anthropological field notes, oral histories, photographs, or videotape records. Problems of access to research archives in university libraries have occurred (see, for example, Halberstadt, 1982). Although such materials are research data, the principles and practices recommended in this report are not intended to cover them, primarily because their consideration was beyond the resources of the committee. It does not mean, however, that access to such research materials is not important or that this report may not help in clarifying relevant issues.

The second kind of research not specifically addressed is research pertaining to national security matters. Recently the National Security Agency has requested that some scientists who are not employed by the government submit their papers on the mathematical theory of codes to the agency for review prior to publication. The purpose of such reviews is to prevent the publication of information damaging to national security. One government spokesman has proposed that reviews be extended to fields such as computer hardware and software and crop projections (Hilts, 1982a, 1982b). Although prior review militates against free and open research, the Committee believed that to recommend guidelines for such review was beyond its scope. This report,

however, notes the existence of such pressure affecting the environment in which data sharing occurs.

The sharing of research data occurs in many ways. Sometimes data are published as appendices to papers and books. Sometimes data are made available in response to requests from other investigators. More formal methods for exchange often involve archives and data libraries, which may be particularly appropriate for the massive data files from surveys and experiments. Careful documentation is important to facilitate data sharing. Poor documentation or its absence inhibits replication and thereby allows some researchers to make bolder claims than they otherwise might. This report pays special attention to the needs for and costs of good documentation, but the formal technical aspects of data archives and the documentation required to make data of use to others are not covered.

The principles and guidelines for data sharing in this report are addressed not only to researchers in academia and government but also to institutions that provide funds for research. Over the past 20 years, government agencies and private and public foundations have underwritten social science research to collect and analyze substantial bodies of data. Social science data collected by the government in particular have been analyzed extensively by many researchers. This report, however, does not treat the special case of transfer of large data sets—usually general-purpose statistics or data from administrative records—among different agencies of the federal government, although many of the findings and suggestions in the report may be applicable. Such transfers were not included in the scope of this study because they are governed by specific statutes and regulations.

This report summarizes some of the benefits and costs of sharing research data with qualitative statements based on judgment that is bolstered by anecdotal evidence. Although quantitative estimates of benefits and costs are highly desirable, the committee unfortunately did not have the time or resources for assembling such estimates. Quantitative estimates of the benefits of data sharing are related to an assessment of the benefits of data generally, an issue that the committee has been and will continue exploring (National Research Council, 1976; Committee on National Statistics, 1980).

Parties to Scientific Research

Many different parties are involved in or affected by scientific research, from the initial investigator to the public. These parties have different, sometimes conflicting interests.

Initial investigators—scientists who first collect data for analysis. These scientists may work alone or in teams and in academic, commercial, nonprofit, or government settings. They have an interest in being the first to examine

and analyze their data and to publish results of their research.

Subsequent analysts—scientists who analyze one or more data sets collected by others, for purposes of verification of the original analysis as well as for analysis of new problems.[1] These scientists have an interest in obtaining data of others for analysis.

Scientific community—all scientists who engage in research. Their interest in the advancement of science through new knowledge is promoted by the sharing of data.

Agencies and foundations that fund research—public and private groups that give grants or contracts for research to be performed by others. Their interest is in advancing science rather than in commercial gain.

Organizations that conduct research—universities, nonprofit institutions, commercial organizations (such as biopharmaceutical concerns), individuals, and government agencies that conduct research, whether they use their own funds or are supported by others. Their interest in sharing data can be those of initial investigators, subsequent analysts, the scientific community, or any combination of them.

Respondents to surveys and participants in experiments—those who agree to participate in a survey or experiment, whether voluntarily or whether they receive remuneration or other direct benefit. Respondents have an interest in the protection of the confidentiality of information they have given, in limiting the invasion of their privacy, in reducing their time and effort required to participate in surveys and experiments, as well as in the advancement of science resulting from such investment of time and effort.

The public—society generally. The public interest is served by open, free, productive, and efficient science.

The different parties involved in or affected by scientific research have different and sometimes conflicting interests when it comes to issues of data sharing. The report and the papers in this volume address the interests of these groups, and many of the committee's recommendations reflect a balancing of conflicting interests.

Occasionally in the report and frequently in the papers, cases are mentioned in which data were shared or in which unsuccessful attempts were made to obtain data from principal investigators. These cases are included to illustrate various aspects of data sharing—the benefits, the costs, the barriers. The cases are not included to assess blame on particular principal investigators or other parties. Sometimes an incomplete account is given; sometimes the

[1] By this definition, subsequent analysts include secondary analysts. A definition of secondary analysis is provided by Hyman (1972:1): “extraction of knowledge on topics other than those which were the focus of the original surveys.”

same occurrence is treated in more than one paper and from varying perspectives. As this report frequently points out, different participants in the research process have different and sometimes conflicting interests. Even the same individual may view data sharing differently at different times, depending on whether he or she is acting as a primary investigator or a subsequent analyst or, for example, whether the issue is the completion of a research project or the protection of respondent privacy.

BENEFITS OF DATA SHARING

That sharing data has benefits is manifestly clear and widely accepted. But a brief recounting of its benefits is useful, in particular in weighing them against costs. This section presents a brief summary of some of the major benefits.

A variety of terms are used here in connection with the sharing of data.[2] A *reanalysis* studies the same problem as that investigated by the initial investigator; the same data base as that used by the initial investigator may or may not be used. If different, independently collected data are used to study the same problem, the reanalysis is called a *replication.* If the same data are used, the reanalysis is called a *verification.* In a *secondary analysis,* data collected to study one set of problems are used to study a different problem. Secondary analysis frequently, but not necessarily, depends on the use of multipurpose data sets. Data sharing is essential for all verifications and all secondary analyses; it may or may not be involved in replications.

Reinforcement of Open Scientific Inquiry

If all science were conducted according to an ideal, referred to by Robert Merton (1973) as the “ethos of science,” then scientific findings would be made available to the entire scientific community. Since the purpose of this availability is to allow others to assess the merits of the research, the need for careful description of study procedures is implicit. We believe that, in addition, the availability of the data for scrutiny and reanalysis should be part of the presentation of results. In the past, among the best investigators and with a journal practice open to extensive description, providing data was an honored tradition. Cavendish's classic paper on the density of the earth is a prime example (Cavendish, 1798).

Scientific inquiry must be open, and sharing data serves to make it so. Disputes among scientists are common; without the availability of data, the

[2] The committee acknowledges the assistance of H.H. Hyman on terminology pertaining to data sharing.

diversity of analyses and conclusions is inhibited, and scientific understanding and progress are impeded.

Verification, Refutation, or Refinement of Original Results

When data are shared, they may be used in reanalyses that provide a direct check on reported results. In addition, supplementary or alternative analyses can be done to determine whether conclusions are robust to various assumptions. This type of verification can work to bolster the findings of the initial investigator. An attempted reanalysis, however, may expose errors or inconsistencies in the data that cast doubt on the validity of the findings. The latter was the case in the research of Ehrlich (1975) on the deterrent effect of capital punishment: several other investigators (Bowers and Pierce, 1975; Passell and Taylor, 1977; Klein, Forst, and Filatov, 1978; Brier and Fienberg, 1980) subsequently pointed out shortcomings in Ehrlich's analyses.[3]

Refinement of original results is also a possible outcome of data sharing. Alternative analyses can lead to better adjustment for background variables and to stronger inferences of effects of treatments in experimental or quasi-experimental studies.

Promotion of New Research Through Existing Data

Another form of reanalysis is testing the generality of research findings (see, for example, Smith and Rowe, 1979). Investigators need to compare analyses on different data sets—across time or across locations—in order to generalize findings about social phenomena. Existing data from several sources may be reexamined from a cross-temporal or international perspective. Treiman (1977:xvi), for example, examined 85 occupational prestige studies from 53 countries and concluded that occupational evaluations are fundamentally the same throughout the world: he contended that "now, and for the foreseeable future, wide ranging secondary analysis of existing data is the only way we will have of achieving a valid comparative sociology."

The same data that were gathered by researchers to answer one set of questions can be used by others to answer a new set. This utility especially applies to large-scale data collection. Mason, Taeuber, and Winsborough (1977) summarized ideas of several social scientists for new research based on public-use samples from the 1940 and 1950 censuses and from the Current Population Surveys since 1960.

[3] The data for Ehrlich's research were shared in only one known instance; others had to reconstruct them.

Sometimes several different data files can be linked to create a new enlarged data base that allows researchers to develop and test new theories. For example, Albert Reiss, Jr., of Yale University, merged the quarterly collection tapes from the National Crime Survey to provide longitudinal information on victimization over several years. This new longitudinal data base allowed Reiss (1980) and Eddy, Fienberg, and Griffin (1981) to develop new models and analyses of criminal victimization that may improve data collection and reporting.

Encouraging More Appropriate Use of Empirical Data in Policy Formulation and Evaluation

In policy settings, the models and methods of analysis used for data are often shaped and structured by expectations associated with particular advocacy positions. When errors or incomplete analyses lead to policy conclusions that agree with those expected, the errors may go undetected, and the analyses remain incomplete. In an evaluation of programs for chronic juvenile offenders, Murray and Cox (1979) reported a large "suppression effect" of criminal behavior that results from incarceration. Their analyses purported to control for alternative explanations of this effect, such as mortality, maturation, and regression. Long before the report was published, it was used to support legislative changes in treatment of juvenile offenders in Illinois and other states. Based on a reanalysis of the basic data, which was commissioned by the National Institute for Juvenile Justice and Delinquency Prevention, other researchers claimed that the original analyses were faulty and the observed effect could be attributed to other causes. Still others argued that the original and alternative analyses were flawed and that the basic data were of low quality and unsuitable as the basis for a policy decision. If data sharing were anticipated, researchers would have greater motivation to plan studies carefully to avoid possible rejection of their data or analyses.

Some program evaluation experts have suggested that statistical analyses be carried out by independent teams of evaluators before a program evaluation report is prepared. Alternative analyses may not only confirm findings of the initial evaluators but also detect effects not found by them. The practice, of course, requires data sharing before publication. We believe that such independent reanalyses should be common practice, especially when important public policies may be affected.

Alternatives to complete analyses conducted independently are critical reviews of the analyses of the original investigator by other experts who have access to the data. An example is a review of the statistical methodology of the draft report, *Public and Private Schools,* by James Coleman et al. The Committee on National Statistics convened a meeting of experts to advise

Coleman on the strengths and adequacy of the sample and the analytical methods used for inferences in the report and to suggest further analysis and interpretation of the data (Straf, 1981). Coleman found the experience valuable and suggested that the Committee consider institutional procedures for review of reports relevant to public policy before they are publicly released.

Improvements of Measurement and Data Collection Methods

When the methods of data collection as well as the data from empirical investigations are scrutinized by scientists other than the original investigators, suggestions for improved measurement and collection methods often follow. For example, Turner and Krause (1978) compared allegedly equivalent measurements of public confidence in national institutions made by two survey organizations and found substantial discrepancies in levels of reported confidence and changes over time. Selected analyses of the data suggested that the differences were due not to technical aspects of the sample design, but probably to the result of differences in measurement techniques, questionnaire design, or field procedures.

Longitudinal studies have benefited from suggestions made by subsequent analysts. Recommendations from scientists who reanalyzed data from the National Crime Survey are partly responsible for current plans to redesign the survey. Two more examples are the national longitudinal surveys of labor force behavior, which is conducted by the Census Bureau for the Department of Labor and planned and analyzed by the Center for Human Resource Research at Ohio State University, and the various waves of interviewing for the negative income tax experiments undertaken in the late 1960s. In these three surveys, early availability of public-use tapes was planned, and comments and suggestions by other analysts were encouraged. The sharing of research data increases the likelihood of suggestions for improvements. This feedback is of special value in continuing surveys, whether cross-sectional or longitudinal.

Development of Theoretical Knowledge and Knowledge of Analytic Technique

Wider data sharing with better documentation of data sets should contribute to better theories and analytic techniques. Ideas for constructively changing or refining concepts and methods would be obtained sooner and more frequently, and the interplay between theories and data would be stimulated if well-documented observations were generally at hand.

Some of these possibilities are illustrated in trials performed by Hildreth and Lu (1960) on 17 data sets that had been used by earlier authors to estimate

demand relations. A technique to allow for first-order serially correlated disturbances was applied to relations previously estimated by a least-squares fit. The results offered useful evidence of the importance of serial correlation, of the possibility of negative serial correlations, and of the inadequacy of routinely using first differences or trends; they also suggested the possibility of higher-order correlations in some cases.

Applying new theories to existing data may lead not only to new knowledge but also to improvements in future data collections. When existing data sets are not adequate for applying and testing new theories, the theories may suggest what kinds of data sets would be more useful. Wider data sharing combined with existing and developing computer technology creates opportunities for comparing results of various techniques on given data as well as results of a given technique on various data. With wider data sharing, more could be learned and in a more timely fashion (Hyman, 1972, 1975).

Encouragement of Multiple Perspectives

When data bearing on a variety of topics are generally available and well documented, researchers may find information important to their inquiries in data obtained by researchers in other disciplines. Using data from another discipline often proves to be stimulating, especially when it leads to direct contacts between the researchers involved, and significant influences on one field from another can be expected.

Users of previously collected data need to know more than just the mechanics of how information was gathered and processed. The concepts that the collectors tried to quantify and the relevant assumptions underlying their interpretations are important to users in judging the appropriateness of data for their purposes. Insofar as it is practical, these matters should be explained in the documentation. Documentation, however, will not always be sufficient for this purpose, and a potential subsequent analyst may need to consult with those who collected the data or other scientists in the same discipline. The subsequent analyst may then learn some alternative viewpoints and approaches of the other discipline.

Initial investigators also have an interest in the results of secondary analysis of their data. When some of this analysis involves scientists from other disciplines, useful stimulation and exchanges of conceptual frameworks and techniques across fields can result.

Provision of Resources for Training in Research

The availability of a variety of carefully documented data sets can be a great asset to research training. Data on real phenomena provide interesting examples

from which students can learn in two ways. First, the process of collecting the data can be studied with regard to accuracy, relevance to policy or scientific questions, and efficiency of design. Second, the data can be used as exercises in applying different analytic techniques, in drawing inferences, and in encouraging original approaches to analyses.

Multiple use of data sets can clearly reduce the number of data collections that are undertaken, saving the time and effort of respondents who furnish information as well as the time and money of researchers who gather it. In much social science research, expenses for data collection are the predominant research cost. Avoiding such expenses allows research funds to go further. Even when new data are needed, review of existing data and preliminary analyses may make for a more efficient collection plan.

Protection Against Faulty Data

One of the worst frustrations of scientists and decision makers is caused by a revelation or strong suspicion that information that was presumed correct and on which results, recommendations, or decisions were based is faulty. Reactions are particularly bitter when willful fabrication, falsification, or distortion of data is involved. The whole basis for applying knowledge and careful inquiry to decision making is negated. The waste of professional resources is serious, but the consequences of false conclusions or damaging decisions may be much worse. People may be hurt by misguided actions, and differences of opinion on public questions may be acerbated. Public confidence in the research community will almost certainly be diminished.

Data sharing cannot eliminate these problems, but it could provide a definite, perhaps strong, preventive influence. Faulty data, whether fraudulent or due to inept collection or processing, are much more likely to be detected if studied by more than one analyst. If several data sets relating to closely related phenomena can be compared, unexpected or unreasonable discrepancies should lead to careful reexaminations. The expectation that further analyses and comparisons will be conducted should discourage dishonest manipulations. More important, such expectation would encourage greater care in the original analysis.

Climate in Which Scientific Research Confronts Decision Making

The principal benefits that would result from wider data sharing are that science would be more efficiently advanced and more effectively applied to making decisions. Wider data sharing must, however, be carefully developed. Feasible arrangements for data sharing might lead to many improvements. Our discussions with a number of scientists and administrators indicate

universally strong interest in wider data sharing and strong convictions that, if data sharing were properly developed, substantial benefits would ensue. The benefits could change the environment in which researchers work. (Expected benefits are discussed further in the papers in Part II.)

Some investigators regard their work as definitive. Results are sometimes made to sound more sweeping than is justified. Trial analyses that do not look good may not be reported. Possible weaknesses in data and methods may be ignored, if they are not generally known, and otherwise may be treated as peripheral. The possibility that other researchers will subsequently find ways to collect more informative data and perform more incisive analyses is not contemplated. Investigators may defend and amplify what they regard as theirs, sometimes to the point of misrepresentation. Few areas of research achieve such definitive results that improvements are not possible. Breakthroughs occur, but they are usually not fully understood or developed for some time. Meanwhile, less spectacular but still vital accretions of knowledge proceed. Data sharing would surely help some people overcome narrow views and pretentious habits. An improved spirit of research would benefit the products.

COSTS OF DATA SHARING

Data sharing involves costs as well as benefits. The costs may at times outweigh the benefits. And those who pay the costs often do not share in the benefits.

Most of the difficulties of data sharing could be overcome if the scientific community and funding agencies were to commit substantial resources to data sharing and if scientific recognition were given to researchers who shared their data. But the scientific community, funding agencies, and especially individual researchers have a good many other—and often higher—priorities. An appreciation of the obstacles to and costs of data sharing may suggest some remedies as well as help in constructing some reasonable and workable principles for data sharing. This section summarizes some of the obstacles and costs.

Technical Obstacles

Technical obstacles to sharing computer-readable data include incompatibilities in machine and software systems and data file structures. In early computer technology, technical factors sometimes constituted nearly insurmountable barriers to transferring data from one computer to another. Now, however, difficulties encountered in transferring data are largely due to the practices of data collectors and processors rather than to technical factors.

Data collectors should, therefore, anticipate that data may be shared and make necessary plans. Although the technical requirements and characteristics of computer programs and systems for data management and analysis do not prevent data sharing, they may complicate it. For example, data organized for analysis using the Statistical Package for the Social Sciences (SPSS) cannot be analyzed using the Statistical Analysis System (SAS) without reformatting and reorganizing the data. Most data-base dictionaries in use in the social sciences are tied specifically to certain software packages such as SPSS, OSIRIS (organized set of integrated routines for the investigation of social science data), or SAS; their conversion for use by other packages usually is not straightforward. Thus, researchers attempting to use data prepared by others often must forgo direct use of information contained in the "foreign" data-base dictionary. Researchers can facilitate data sharing by assimilating data in machine-and program-compatible formats.

Documentation

Typically, data sets are poorly documented. Researchers keep the details of data collection, variable construction, and particular quirks of the data in their memories and do not put them in writing. Data collectors sometimes prefer data preparation and documentation practices with which they are familiar, although these practices may be at odds with accepted standards. Accomplishment of the particular research goals of initial investigators may not require fully cleaned tapes and well-documented data; data are collected primarily to achieve these research goals, not to serve the purposes of data sharing and secondary analysis. The documentation requirements of research and scientific publication usually differ from those of data sharing. Moreover, available financial resources often are seen as inadequate to support data collection and analysis and certainly inadequate for elaborate data preparation and documentation. Consequently, the documentation required for effective sharing is not done.

A distinction should be made between technical and substantive documentation. Basic standards for technical documentation have been established and are in use in the preparation of many research data collections (Geda, 1979; Roistacher, 1980). Less clear are the standards for matters such as descriptions and explanations of sampling procedures; the original design of the data collection and any deviations; the assumptions that underlie particular questions, combinations of questions, and derived measures; and the degree to which instruments were pretested and the results of those pretests.[4]

[4] Derived measures, such as scales or recodes that collapse variables, are often poorly documented. Sometimes, in order to maintain confidentiality, the actual data collected cannot be shared, but aggregates or derived m easures can be. It is particularly important in such cases to document for subsequent analysts how the combinations were put together.

Practices in this area are less consistent and probably generally less adequate than in the case of technical documentation. Yet these aspects of documentation are essential for the effective secondary if substantive documentation is inadequate, data are subject to inadvertent use of data collection. Data may be in perfect technical order, but misuse with the result of misleading or erroneous findings.

Costs to the Original Researcher

Although it serves science for researchers to share their data and permit reanalysis and replication, it is often not in their interest to do so. Researchers face the costs of documentation for the use of others, of storing and transferring data, and of conducting tutorials so that subsequent analysts understand the data.

Other costs are less susceptible to monetary valuation and to recompense but are no less real. Researchers face the possibility that errors in their original analyses will be exposed. Initial investigators may also fear that subsequent analysts may publish results before they do, a problem that is particularly vexing with panel studies. And researchers know that those who reanalyze data will be able to publish only if the reanalysis contradicts or goes beyond the original work.

Researchers may be concerned about the qualifications of investigators requesting data and fear that poor reanalysis may require burdensome rebuttal or reflect adversely on original research. Initial investigators may fear criticism that, even if unwarranted, may be detrimental. Researchers may even fear that data made accessible during the peer review process may be published by others. Sharing data involves loss of control over data, the purposes for which they are used, and the methods of analysis. That requests for the sharing of data are often met with delays and noncooperation is not surprising (see Wolins, 1962; see Hedrick, in this volume, for a detailed discussion of these issues.)

Costs to Subsequent Analysts

Subsequent analysts also encounter some costs. Despite more compatible equipment and careful planning by original collectors, not all data may be shared easily. Sharing may be time-consuming and expensive to the subsequent analyst as well as to the initial researcher, particularly if the data set is

large. Data organized in complex file structures may need to be converted to simpler structures by the subsequent analyst. The data-base dictionary may be tied to an incompatible software package and require conversion. The original data collectors may not have used standard data preparation and documentation practices. The data documentation may be inadequate; the codes may be undocumented, inconsistent, or erroneous. Undiscovered errors are inevitable.

These costs can be reduced if data sharing is recognized as a goal by initial data collectors. And the costs may be shared if data tapes are transferred to an intermediate archive that takes responsibility for editing and documenting them.

Sharing Costs

One strategy for encouraging data sharing is to impose a cost for not sharing data. A public statement that a researcher was withholding data may encourage the researcher—and others—to share their data. Reinforcing data sharing as a scientific obligation may be fruitful in promoting data sharing more widely.

The practice of data sharing probably will become more widespread if the costs are not borne exclusively by the initial researcher. Data sharing, then, must also be cost sharing; subsequent analysts should contribute appropriately to the costs of documentation and pay the costs to transfer data.

Sharing data primarily benefits science and society; the costs are borne mostly by the initial investigators. Yet most scientists are willing to share their data to some extent despite this relationship. One reason is that recognition of the initial investigator usually is provided by subsequent analysts. Another reason is that scientific institutions do foster data sharing through peer recognition of altruistic behavior that advances science.

THE CHANGING ENVIRONMENT FOR DATA SHARING

Developments in computers and software, changes in research practices, the different rewards and incentives for research, and new laws and regulations may all affect the sharing of data. This section describes how a few of these changing circumstances may affect the propensity of researchers to share their data.

Use of Computers

The widespread use of computers for recording, summarizing, and analyzing research data facilitates sharing data. The use of computers avoids time-consuming clerical work and permits the transfer of large data bases that would not have been feasible in the past. Large machine-readable data files are a research resource in the social sciences analogous to large-scale instrumentation in the physical sciences.

Transfer of machine-readable data is hindered by incompatibility of computer equipment and software. Help to overcome such technical problems may come from the acceptance of common conventions for the internal storage and representation of data, from the development of standard analytic packages, and the development of conversion capabilities to move from one system to another. More burdensome to an initial investigator are the time-consuming tasks of file cleaning, preparation of data-base dictionaries and other appropriate documentation, and dissemination. As the importance of these activities has become more widely recognized, some aids have been developed; more are expected in the future. The literature on computer file management, standards for file documentation, and similar matters is growing. Moreover, institutions have been organized that specialize in the collection, maintenance, and dissemination of machine-readable data files. Some of these institutions are international in scope. Both the technical guidelines for data documentation and the number of institutions that serve as intermediaries to transfer data are growing (see Clubb, in this volume, for a further discussion on using computers for data sharing).

Privacy and Confidentiality

Confidentiality refers to not disclosing responses to questions that could be identified as belonging to an individual organization or person. Privacy refers to the right of an individual not to make personal information available to another. Confidentiality is obviously relevant to data sharing. Privacy is also relevant: as the public has become more concerned about invasion of privacy, researchers have attempted to overcome respondent hesitation by making stronger promises of confidentiality. Legal protections for privacy attempt to protect privacy by maintaining confidentiality of records, and in many cases, restricting their use to the agency to which the respondent provided information.

Growing concerns about confidentiality and the protection of privacy have affected research involving information about individuals and the conditions under which data may be shared, especially if the research is undertaken under

federal contract. As a result, more attention is paid to maintaining the confidentiality of records, whether legally required or not; to removing identifiable information from records before data are shared; and to using other disclosure avoidance techniques.

Paralleling the burgeoning use of computers in business and government, public awareness of issues of privacy and confidentiality has increased during the past two decades. Respondents express concern over invasion of privacy and are skeptical of assurances that confidentiality will be protected (see, for example, National Research Council, 1979). Also, the public is apprehensive of the growth of large-scale computerized data banks that contain personal, individually identifiable information. Investigators have become more sensitive to issues of privacy and confidentiality because of this public discussion and respondent reactions.

The public concerns have led to enactment of statutes designed to protect privacy and ensure the confidentiality of data concerning individuals (see Cecil and Griffin, in this volume). A major federal statute is the Privacy Act of 1974. Designed to protect the confidentiality of records collected and maintained by the federal government, it provides, with certain exceptions, that identifiable information about individuals may not be disclosed outside the agency that collected the information unless the prior consent of the individuals concerned is obtained.[5] A key characteristic of this statute is that it does not distinguish between data for administrative purposes and data for research or statistical purposes. The provisions of the law apply directly to investigators whose research or surveys are undertaken under a contract with a federal agency, as are, for example, most evaluations of federal programs. Such investigators must observe the provisions of the Privacy Act in sharing data by deleting identifying names and numbers from individual records; sometimes, other disclosure-avoidance techniques are used.

These rules may hamper and at times prevent the matching or linking of data files. In some research requiring access to federal data, identification of individuals is essential. In epidemiological studies, for example, it may be necessary to know the names of persons exposed to certain suspected hazards over long periods in order to match these with records of death or disease at a later time. Unless such epidemiological research is considered “routine use” under the terms of the Privacy Act, access to this information may be restricted.

Biomedical researchers in particular are affected by federal regulations governing research on humans that require review of research plans by institutional

[5] In addition to federal law, several states have enacted statutes to protect privacy that may also affect research.

review boards. In some cases, such boards may go beyond the requirements of the Privacy Act and so have an effect on the ability of researchers to share data.

The Privacy Protection Study Commission, called for by the Privacy Act of 1974, urged among other recommendations that the Act be revised to distinguish between data for research purposes and those maintained for administrative purposes (Privacy Protection Study Commission, 1977: especially pages 567–604). If the law is changed, investigators might find fewer restrictions on access to individually identifiable federal data for research purposes. It is certain, however, that there would still be strong injunctions and safeguards calling on researchers to protect the confidentiality of data.

Freedom of Information

Another federal statute, the Freedom of Information Act, enacted in 1966, which provides for greater public access to many kinds of federal data, has had the opposite effect of the Privacy Act (see Cecil and Griffin, in this volume). There are two specific exemptions to access in the Freedom of Information Act that are most relevant to research data: "personnel and medical and similar files the disclosure of which would constitute a clearly unwarranted invasion of privacy" and "trade secrets and commercial or financial information obtained from a person and privileged or confidential." An investigator whose contract with a federal agency calls for transfer to the agency of microdata that do not qualify for these exemptions should expect that the data may be shared with others, researchers or not, under the Freedom of Information Act. The act does not appear to apply, however, to data maintained solely under the control of the investigator. Even investigators working on funds from private sources may be subject to the Freedom of Information Act should they submit data to a federal agency for advice or checking. For example, a privately sponsored survey that used computer assistance from the federal Centers for Disease Control was ruled subject to the Freedom of Information Act (Dickson, 1980).

Patents, Profits, and Proprietary Data

The possibility that a research effort may lead to the development of a patentable product or process may affect the willingness of investigators to share their data. Patent laws may also delay publication of research results and, therefore, may delay data sharing. A recent change in the U.S. patent law, for example, led the Office of Management and Budget to suggest that federal agencies require notification of any potentially patentable results at least three months before research reports are submitted for publication. The rule would

apply to federally sponsored research in universities and small business and is intended to allow time to apply for patent rights in certain European countries. In the United States, patents can be applied for up to one year following publication of research results, but in some European countries patent rights may be forfeited by publication. In commenting on these developments, Dickson (1981:501) noted: "The proposed rule has already created a storm of protest from the U.S. research community, which claims that, by threatening to deny a scientist patent rights to a discovery if the procedure is not followed, it could seriously impede scientific communication."

The Copyright Act is also relevant to data sets developed by researchers. Under that act, the proprietary rights of a person who has developed information are balanced against the public benefits from distribution of the information. Interpretations of the Copyright Act, which was significantly amended in 1976, may affect the extent to which data are shared. The doctrine of fair use, which limits the exclusive rights of copyright owners in order to permit reasonable use by others for purposes such as criticism, news reporting, teaching (including multiple copies for classroom use), or research, was expanded in the Copyright Act amendments (see Cecil and Griffin, in this volume). Scholarly journals that insist on copyrighting all articles may impede reanalysis of previously published information by requiring secondary analysts to obtain copyright releases from original researchers, although the fair use provision makes this requirement unnecessary.

Recent applications of research on DNA have drawn dramatic attention to the potential profitability of some research. Academic research scientists and private firms engaged in developing profitable applications have sometimes found themselves with very different interests. A report in *Science* of a dispute between the University of California and the pharmaceutical firm of Hoffmann-La Roche concerning a human gene containing the genetic information for the synthesis of interferon carried the following headline: "University and Drug Firm Battle Over Billion-Dollar Gene: A lawsuit over interferon may change the informal ways by which researchers exchange materials" (Wade, 1980). Donald Kennedy, president of Stanford University, commented: "Scientists who once shared prepublication information freely and exchanged cell lines without hesitation are now much more reluctant to do so" (Roark, 1981). And the *New York Times* (1981) editorialized: "The values of the marketplace have so invaded the campus that on several occasions researchers have refused to share with their colleagues the exact details of how they did their experiments. Such attitudes are incompatible with the ethos of a scholarly community." Similar views were expressed in a *Nature* (1980) editorial. Potentially lucrative applications of scientific research

are not widespread, but, in the scientific disciplines in which they occur, the effect on data sharing is significant.

At a recent meeting of university and company officials, the need for faculty freedom to report research was discussed, and it was agreed that research contracts or licensing agreements between universities and private companies should avoid secrecy (*Chronicle of Higher Education,* 1982:12). The joint statement included, under the heading "Open Communication Encouraged," the following:

> The traditions of open research and prompt transmission of research results should govern all university research, including research sponsored by industry. Those traditions require that universities encourage open communication about research in progress and research results. However, it is appropriate for institutions to file for patent coverage for inventions and discoveries that result from university research. This action may require brief delays in publication or other public disclosure.
>
> Receipt of proprietary information from a sponsor may occasionally be desirable to facilitate the research. Such situations must be handled on a case-by-case basis in a manner which neither violates the principles stated above nor interferes with the educational process. Any other restrictions on control of information disclosure by institutions are not appropriate as general policy.

Restrictions on International Sharing of Data

Restrictions on the sharing of data across national boundaries are likely to fluctuate with international political tensions and changes in perceived national interests. Such restrictions may apply not only to defense-related technology, but more broadly to research that is deemed to be of advantage to other nations. The Export Administration Act of 1979, administered by the U.S. Department of Commerce, requires that export controls be used where necessary "to restrict the export of goods and technology which would make a significant contribution to the military potential of any other country or combination of countries which would prove detrimental to the national security of the United States."

In the United States, restrictions on sharing data with other countries apparently are being tightened. Examples include:

(1) Proposed revisions in the 1972 International Traffic in Arms Regulations, published in preliminary form in the *Federal Register* (December 19, 1980), require that an export license be obtained for transfer to a foreigner of technical data that may have a defense application.

(2) During 1981, an amendment was proposed to the Arms Export Control

Act (H.R. 109) to tighten restrictions on exchange of information in such fields as computer technology (Kolata, 1981).

(3) It has also been proposed to have scientific work reviewed by federal agencies on a voluntary basis prior to publication. Such a voluntary review system is now in effect in the field of cryptanalysis.

Although published unclassified data are exempt, researchers fear restraint of scholarly inquiry, and professional societies, among others, are objecting, since information presented at scientific meetings may not be exempt (Marshall, 1981).

The conflicting pressures of national security and open science have recently aroused much interest in the general press as well as in scientific circles. The National Academy of Sciences announced in March 1982 the appointment of a broadly based panel of senior policy makers and researchers to examine the relationship between university research and national security in light of the growing concern that foreign nations are gaining military advantages from American research. The panel's September 1982 report recommended guidelines that would allow government-funded, academically based scientific research to be performed without restriction, except for research in narrowly defined areas of technologies that could not justifiably be either classified or completely open (Committee on Science, Engineering, and Public Policy, 1982). In an assessment of policy developments 18 months after the panel report was issued, Wallerstein (1984) concluded that “the reach of restrictions either proposed or in force go considerably beyond the panel's recommendation.” Since then, the Department of Defense has indicated that it would not further restrict publication of militarily sensitive but unclassified research: control of fundamental research in science and engineering at universities and federal laboratories is to be achieved through classification. Some scientists fear, however, that more research will be classified (Goodwin, 1984).

CONCLUSIONS AND RECOMMENDATIONS

“… the best security for the fidelity of mankind is to make their interest coincide with their duty.”
—Alexander Hamilton
The Federalist Papers, No. 72

Most scientific advances are not solely the result of separate, individual efforts. As society turns to science with ever more problems, solutions are interdisciplinary and require the contribution of many investigators. At the same time, scientists are becoming more specialized. Sharing data can provide opportunities for interdisciplinary approaches to problems and, even

within the same discipline, the sometimes synergistic result of different people thinking about the same or similar problems.

Because of the promise for eventual solutions to important problems, as well as the benefits of increased knowledge and understanding, society supports science. Sharing data offers efficient use of research funds by allowing further discoveries to be recovered from data that have already been collected at great expense and that otherwise would not be used further. There are many other important benefits to science from sharing data. A primary one is that sharing data provides for further theories, methods, and results. Sharing data also tends to correct inadvertent error and to discourage fraud.

But there are potential costs for an investigator who provides data to others: costs of time, money, and inconvenience; fears of possible criticism, whether justified or not; possible violations of trust by a breach of confidentiality; and forgoing recognition or profit from possible further discoveries.

In some circumstances initial investigators are required to share data in accordance with the rules of their employing institutions or the terms of their grants. In many cases, however, whether data are shared and the extent to which they are shared depend on the decisions of individual scientists. Professional societies, organizations that publish scholarly journals, research institutions, and foundations and other organizations that fund research can encourage, facilitate, and even reward the sharing of data, although they seldom prescribe the behavior of individual scientists.

These considerations led the Committee on National Statistics to make the following general recommendations.

Recommendation 1. Sharing data should be a regular practice.

The advantages of data sharing are sufficient to warrant considerable attention to ways to share data without imperiling privacy or breaching the confidentiality promised to data providers. We share the views of Jowell (1981:14):

> Flaherty (1979, p. 307), in his definitive international survey of measures to enhance the confidentiality of microdata, concludes that an "ultimate goal of public policy in every country should be to encourage custodians to disseminate data and researchers to use it." As long as the individual is adequately protected, wider access to data will surely serve rather than threaten the interests of civil liberties and open government.

The Committee recommends a number of guidelines for researchers, for funding agencies, for professional journals, for research training institutions, and for other participants in research that should facilitate and encourage sharing data for research purposes.

Recommendations for Initial Investigators

When to Share Data

Data are collected in a variety of circumstances—in controlled laboratory experiments, by observation in the field, through interviews, from accumulations of records, or by combinations of these methods. In some cases, data to which access is desired may have developed through one investigator's efforts and be entirely at his or her disposal to share. In other cases, the nature of the data, promises of confidentiality, laws or regulations, contractual requirements, or proprietary rights may preclude or at least militate against sharing. In still other cases, raw data may be available to all (for example, from public records or from public-use tapes, which are samples of anonymous statistical data specifically designed for widespread research use), and the researcher's contribution may be in the compilation procedures and methods of analysis. In the latter instance, it is the edited and categorized data, an explanation of the analytical methods used, and documentation of how the data were handled to which access may be requested.

Analyzing data and reporting discoveries are clearly more glamorous tasks to many scientists than collecting data. The motivation of possible discoveries is needed even to contemplate data collection, and science is served well by this motivation. Thus, initial investigators are entitled to be the first to examine, summarize, and analyze their data. There may, however, be exceptions, for example, when data collection is a joint effort or when public funds are used to pay for data collection with the intent that the data be available to many in a timely manner. Although scientists surely deserve, in most cases, first claim to data compiled under their direction, the practice of withholding data until all possible analyses are exhausted is unnecessarily restrictive and too self-serving to advance science. A balance is needed.

Recommendation 2. Investigators should share their data by the time of publication of initial major results of analyses of the data except in compelling circumstances.

It should also be noted that, if data are made available when the results of research are submitted for publication, the submitted manuscript can be more carefully and more fully reviewed. The benefits of sharing data appreciably increase upon publication, since other researchers can then test the same and other theories and methods. We encourage researchers to make every effort to share data as soon as it is feasible.

Data Relevant to Public Policy

Scientists have a special responsibility to share data as quickly and as widely as possible when the data are or will become relevant to public policy. Withholding such data risks the use of wrong results or of ineffective analysis of important issues.

Recommendation 3. Data relevant to public policy should be shared as quickly and widely as possible.

This recommendation is not intended to support the public release of analyses prior to appropriate review.

Planning for Data Sharing as Part of Research

Researchers can more effectively share data if they keep that objective in mind in all stages of their research. Planning to share data from the outset not only helps achieve the goal of data sharing but also may improve the quality of the research. For example, adequate documentation of data helps initial investigators as well as subsequent analysts. Data files should include the unedited raw data as well and documentation on edits, handling of nonresponse, and similar problems (see Straf, 1981; Madow et al., 1983).

Not all data can be shared in a situation in which confidentiality must be preserved. For example, photographs, oral histories, detailed notes on interviews of well-known people, and some types of proprietary information are data that could not be shared if confidentiality is to be maintained. Some persons or organizations may be unique or come from such a small group that it may be impossible to share data and not identify them. There are, however, ways to share many types of data and still maintain confidentiality (see Campbell et al., 1975).

Recommendation 4. Plans for data sharing should be an integral part of a research plan whenever data sharing is feasible.

Researchers might benefit by first considering whether they could be subsequent analysts: data might already have been collected that are sufficiently useful to warrant forgoing new data collection.

Keeping Data Available

Part of research plan should include maintaining the data for a reasonable period following the completion of research for possible use by subsequent analysts. Some data collections may be small or so specialized that only limited use by others can be expected, and the initial investigator can handle requests without undue burden. Other data sets may be of such general purpose and in such demand over a considerable period that the initial investigator may find it difficult or impossible to handle the requests of subsequent analysts. Particularly in the latter case, researchers might consider submitting data to an appropriate archive that not only would assume responsibility for much of the handling of data to be shared, but also would encourage further use of the data by bringing them to the attention of a wider community of researchers. Cataloging of machine-readable data files and citing such data in a standard way (Dodd, 1982) would also encourage further use.

Recommendation 5. Investigators should keep data available for a reasonable period after publication of results from analyses of the data.

Recommendations for Subsequent Analysts

It is neither practical nor equitable to expect initial investigators to pay all costs of transferring their data to others. It is reasonable to expect subsequent analysts to reimburse initial investigators at least for the extra costs involved in data transfer.

Recommendation 6. Subsequent analysts who request data from others should bear the associated incremental costs.

Recommendation 7. Subsequent analysts should endeavor to keep the burdens of data sharing on initial investigators to a minimum and explicitly acknowledge the contribution of the initial investigators.

Explicit acknowledgment of the initial investigators and their contributions would encourage data sharing.

Subsequent analysts who discover errors in data should inform the data collectors or the appropriate archive so that the data may be corrected for the use of others. Criticism of a data collection or analysis should be made in a professional manner. With few exceptions, it is desirable that subsequent analysts also inform initial investigators or data archives promptly of the results of new analyses, even those that are unrelated to the original analysis. This

scientific courtesy may also help to avoid future duplications of efforts.

Recommendations to Institutions that Fund Research

A scientists is recognized and rewarded through the scientific community and its institutions. Researchers will have greater incentive to share data if the community and its institutions foster the idea that the practice advances science and is part of what is recognized as necessary and proper scientific behavior. We suggest that foundations, federal agencies, and other organizations that fund research provide encouragement and rewards for sharing data.

In many instances, funding organizations would be justified in requiring that data be shared. Government funding agencies, in particular, should require applicants to guarantee data sharing or to justify explicitly in their proposals why sharing would be inappropriate. Unless data sharing is a condition of a grant or contract—whether of public or private funds—applicants who have budgeted to share are at a disadvantage when costs are compared with the budgets of those who have not.

If plans to share data are given as much weight as the sample design, methods of analysis, and other aspects of proposed research in deciding on an award, researchers would then plan for sharing data at an early stage. A researcher might request funds to make important data available to others. In any case, he or she could be encouraged to describe in the application how the content and structure of the data would be documented, how invitations for subsequent analysis would be extended, and how requests for data could be honored at minimal cost. The referees of the research proposal could judge the importance of support for making the data available to others.

For research projects involving large data sets, investigators could request funds for a person with responsibility to document data files; update and correct data entries; produce data files for those who request them; consult with users on interpretations, limitations, and other important aspects of the data; and preserve the confidentiality of respondents. Even for small data sets, however, a funding organization that encourages reasonable standards for documentation will aid not only subsequent analysts, but also the initial investigators.

Funding organizations that require, in rules or by contracts, unnecessarily excessive protection of privacy and confidentiality hinder the sharing of data. Society benefits from the accessibility of data as well as from the protection of privacy and confidentiality. A reasonable balance between these often conflicting values cannot be achieved by exclusive attention to one.

When funding agencies anticipate that research results will be directly relevant to public policy, the agencies should be alert to the need for sharing data so that conclusions can be verified or contested through reanalysis. Federal

funding organizations can ensure the availability of data for such uses by including in original contracts or grants a requirement that, on completion of research, data will be delivered to the sponsoring agency. The data would then be subject to the Freedom of Information Act.

Recommendation 8. Funding organizations should encourage data sharing by careful consideration and review of plans to do so in applications for research funds.

Initial investigators whose data sets prove to be of wide interest to subsequent analysts may not be in a position to manage and disseminate data to many others for a long time. Even if initial investigators are paid for the additional time and other costs involved, sharing data may impinge too severely on other scientific activities. Intermediate research archives have been developed in some fields to meet this problem (see Clubb, in this volume, for more details). Organizations funding large data collections that are expected or later found to be of considerable general interest should be alert to this problem. If existing data archives are not suitable or are inadequately funded, funding organizations should consider supporting appropriate ones.

Recommendation 9. Organizations funding large-scale, general-purpose data sets should be alert to the need for data archives and consider encouraging such archives where a significant need is not now being met.

Recommendations to Editors of Scientific Journals

The editorial policies of scientific journals have a significant effect on scientific practice, since the publication of research results in respected, refereed journals is one of the principal rewards of scientific research. Journal editors should adopt editorial policies designed to encourage data sharing.

Providing Access to Data for Peer Review

Access to data during the review process, a practice already in use by some journals, provides reviewers an opportunity to replicate the analysis and discover possible errors. Reviewers can use alternate assumptions or analytic models to test the robustness of authors' conclusions.

Recommendation 10. Journal editors should require authors to provide access to data during the peer review process.

Publishing Reanalyses and Secondary Analyses

If researchers know that reports of replications, whether confirmatory or not, and of secondary analyses will be welcomed under journal editorial policies, such research would be encouraged.

Recommendation 11. Journals should give more emphasis to reports of secondary analyses and to replications.

Giving appropriate credit to data collectors should serve to encourage others to share data as a matter of good scientific practice. Criticism of the original data collection should be factual, temperate, and made in the light of reasonable standards of data collection.

Recommendation 12. Journals should require full credit and appropriate citations to original data collections in reports based on secondary analyses.

Encouraging Accessibility to Data

It should be standard practice for small data sets to be published with the research reports that use them. For larger sets, the availability might be announced in the research report with an explanation of where the data may be obtained: from the journal editor, from an intermediate archive, from the original investigator, or elsewhere.

Recommendation 13. Journals should strongly encourage authors to make detailed data accessible to other researchers.

Recommendations to Other Institutions

Other participants in the scientific research process can promote data sharing. Academic institutions can exercise leadership in encouraging data sharing both in training future scientists and by example. Professional associations can also play a part, as can funding agencies and archives.

Providing Training for Sharing Data

Instruction and training on data-sharing policies and practices should be included in the education of many research scientists. Professional societies might organize meeting sessions or workshops on data sharing. The technical

aspects of data sharing, especially documentation and archiving methods, should be taught in specialized courses either as a part of academic curricula or in continuing education programs. Instruction in data sharing should also include how to find and adapt existing data for research (Myers and Rockwell, 1984) and how to prepare data for secondary analysis (Fortune and McBee, 1984). In some disciplines, emphasis on sharing data could be a recognized part of graduate training.

Recommendation 14. Opportunities to provide training on data sharing principles and practices should be pursued and expanded.

Researchers should be encouraged to use data collected by others for scholarly research when appropriate. Actual data should be used in teaching whenever practical, a practice that depends on data being shared.

Reference Service for Social Science Data

A centralized reference service for computer-readable social science data would promote the use of data already collected. A start can be made with existing archives and with some federal statistical agencies. The Social Science Research Council (1983) has recently issued a compendium of brief descriptions of about 100 national data bases available for use in social science research. By allowing sufficient funds for adequate documentation of original studies and by funding research based on the use of shared data, funding agencies could foster the growth and efficient use of such a service. The National Science Foundation might take a leading role in promoting it.

Recommendation 15. A comprehensive reference service for computer-readable social science data should be developed.

Providing Recognition for Data Sharing

The scientific reward structure could be strengthened to achieve more sharing of data and more innovative subsequent analyses. In addition to our recommendations to journal editors, we suggest that academic institutions encourage data sharing by granting appropriate professional recognition to the data-sharing activities of teaching and research staff members in such matters as salary and promotion policies.

Recommendation 16. Institutions and organizations through which scientists are rewarded should recognize the contributions of appropriate data-sharing practices.

REFERENCES

Bowers, W. J., and Pierce, G. L. 1975 The illusion of deterrence: a critique of Isaac Ehrlich's research on capital punishment. *Yale Law Journal* 85:187–208.

Brier, S. S., and Fienberg, S. E. 1980 Recent econometric modeling of crime and punishment: support for the deterrence hypothesis? *Evaluation Review* 4(2):147–191.

Campbell, D. T., Boruch, R. F., Schwartz, R. D., and Steinberg, J. 1975 Confidentiality-preserving modes of access to files and to interfile exchange for useful statistical analysis. Appendix A in *Protecting Individual Privacy in Evaluation Research,* Committee on Federal Agency Evaluation Research, Assembly of Behavioral and Social Sciences . Washington, D.C.: National Academy of Sciences.

Carroll, J. D., and Knerr, C. R. 1976 The APSA confidentiality in social science research project: a final report. *PS* (The American Political Science Association) 9(4):416–419.

Cavendish, H. 1798 Experiments to determine the density of the earth. *Philosophical Transactions of the Royal Society* 88:469–526.

Chronicle of Higher Education 1982 College and industry leaders' recommendations on commercial use of research. *Chronicle of Higher Education* 24(6):12–13.

Committee on National Statistics 1980 Sharing of Social Science Research Data: An Exploratory Conference Convened by the Committee on National Statistics. October 19, 1979. Assembly of Behavioral and Social Sciences, National Research Council, National Academy of Sciences, Washington, D.C.

Committee on Science, Engineering, and Public Policy 1982 *Scientific Communication and National Security.* Panel on Scientific Communication and National Security. Washington, D.C.: National Academy Press.

Dickson, D. 1980 Research data: private property or public good? *Nature* 284:292.

1981 Patents to mean slow publication? *Nature* 293:501–502.

Dodd, S.A. 1982 *Cataloging Machine-Readable Data Files: An Interpretative Manual.* Chicago: American Library Association.

Earman, J., and Glymour, G. 1980 Relativity and eclipses: the British eclipse expeditions of 1919 and their predecessors. Pp. 49–85 in G. Heilbron et al., eds., *Historical Studies in the Physical Sciences,* Vol. 11. Berkeley: University of California Press.

Eddy, W.F., Fienberg, S.E., and Griffin, D.L. 1981 Estimating victimization prevalence in a rotating panel survey. *Bulletin of the International Statistical Institute* 43(2):719–731.

Ehrlich, I. 1975 The deterrent effect of capital punishment: a question of life or death. *American Economic Review* 65:397–417.

Fortune, J.C., and McBee, K. 1984 Considerations and methodology for the preparation of data files. Chapter 2 in David J. Bowering, ed., *Secondary Analysis of Available Data Bases.* San Francisco: Jossey-Bass.

Geda, C.L. 1979 *Data Preparation Manual.* Ann Arbor, Michigan: Institute for Social Research.

Goodwin, I. 1984 Pentagon lowers heat on science secrecy—maybe. *Physics Today* 37(7):57–59.
Gould, S. J. 1978 Morton's ranking of races by cranial capacity. *Science* 299:503–509.
Halberstadt, J. 1982 The "creative editing" of Thomas Wolfe. *Harvard Magazine* (Jan.-Feb.):41–46.
Hildreth, C., and Lu, J. Y. 1960 Demand Relations with Autocorrelated Disturbances. Technical Bulletin 276, Michigan Agricultural Experiment Station, Michigan State University, East Lansing, Michigan.
Hilts, P. J. 1982a Scientists urged to submit work for U.S. review. *The Washington Post.* January 8.
1982b Scientists call research censorship idea a "nightmare." *The Washington Post.* January 9.
Holton, G. 1978 Subelectrons, presuppositions, and the Millikan-Erenhaft dispute. Pp. 25–83 in R. McCormmach et al., eds., *Historical Studies in the Physical Sciences,* Vol. 9. Baltimore, Maryland: The Johns Hopkins University Press.
Hyman, H. H. 1972 *Secondary Analysis of Sample Surveys: Principles, Procedures, and Potentialities.* New York: Wiley.
1975 *The Enduring Effects of Education.* Chicago: University of Chicago Press.
Jowell, R. 1981 A Professional Code for Statisticians? Some Ethical and Technical Conflicts. Paper presented at the 43rd Session of the International Statistical Institute, Buenos Aires, Argentina, November 30-December 11 .
Kershaw, D. and Fair, J. 1976 *The New Jersey Income-Maintenance Experiment. Volume I: Operations, Surveys, and Administration.* New York: Academic Press.
Klein, L. R., Forst, B. E., and Filatov, V. 1978 The deterrence effect of capital punishment: an assessment of the estimates. Pp. 336–360 in A. Blumstein et al., eds., *Deterrence and Incapacitation: Estimating the Effect of Criminal Sanctions on Crime Rates.* Washington, D.C.: National Academy of Sciences.
Kolata, G. B. 1981 House bill would classify much computer research. *Science* 213:1343.
Madow, W.G., Nisselson, H., Olkin, I., Rubin, D., editors 1983 *Incomplete Data in Sample Surveys.* 3 vols. Panel on Incomplete Data, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council. New York: Academic Press.
Marshall, E. 1981 Export law affects scientific meetings. *Science* 211:1326.
Mason, W. M., Taeuber, K. E., and Winsborough, H. H., eds. 1977 Old Data for New Research. Report of a workshop on research opportunities and issues in the design and construction of public use samples from the 1940 and 1950 censuses and from Current Population Surveys from 1960 forward, held in Madison, Wisconsin, June 1976. Working paper 77-3, Center for Demography and Ecology, University of Wisconsin, Madison .
Merton, R. 1973 *The Sociology of Science.* Chicago: University of Chicago Press.
Murray, C. A., and Cox, L. A., Jr. 1979 *Beyond Probation.* Beverly Hills, California: Sage Publications.

Myers, D.E., and Rockwell, C 1984 Large-scale data bases: who produces them, how to obtain them, what they contain. Chapter 1 in David J. Bowering, ed., *Secondary Analysis of Available Data Bases.* San Francisco: Jossey-Bass.

National Research Council 1976 *Setting Statistical Priorities.* Panel on Methodology for Statistical Priorities, Committee on National Statistics. Washington, D.C.: National Academy of Sciences.

1979 *Privacy and Confidentiality as Factors in Survey Response.* Panel on Privacy and Confidentiality as Factors in Survey Response, Committee on National Statistics, Assembly of Behavioral and Social Sciences. Washington, D.C.: National Academy of Sciences.

1980 *Estimating Population and Income of Small Areas.* Panel on Small-Area Estimates of Population and Income, Committee on National Statistics, Assembly of Behavioral and Social Sciences. Washington, D.C.: National Academy Press.

Nature 1980 Editorial—Should Academics Make Money Outside? 286(July 24):319–320.

Nelkin, D. 1984 *Science as Intellectual Property: Who Controls Scientific Research.* Published for the American Association for the Advancement of Science. New York: MacMillan.

New York Times 1981 Editorial—Pure Science and Pure Profit. November 16.

Passell, P., and Taylor, J. B. 1977 The deterrent effect of capital punishment: another view. *American Economic Review* 67:445–451.

Privacy Protection Study Commission 1977 *Personal Privacy in an Information Society.* Washington, D.C.: U.S. Government Printing Office.

Reiss, A. J., Jr. 1980 Victim proneness by type of crime in repeat victimization. Pp. 41–53 in S. E. Fienberg and A. J. Reiss, Jr., eds., *Indicators of Crime and Criminal Justice: Quantitative Studies.* Washington, D.C.: U.S. Government Printing Office.

Roark, A. C. 1981 Professors urged to use caution in responding to commercial overtures on genetic research. *Chronicle of Higher Education* 21(20):1, 4.

Roistacher, R.C. 1980 *A Style Manual for Machine-Readable Data and Their Documentation.* Report No. SD-T-3, NCJ-62766. Bureau of Justice Statistics. Washington, D.C.: U.S. Department of Justice.

Smith, K. W., and Rowe, J. S. Using secondary analysis for quasi-experimental research. *Social Science Information* 18(3):451–472.

Social Science Research Council 1983 *National Social Data Series: A Compendium of Brief Descriptions.* Washington, D.C.: SSRC Center for Coordination of Research on Social Indicators.

Straf, M. L. 1981 Report of the Meeting of a Panel to Review the Statistical Methodology of the Report *Public and Private Schools.* July 23, 1981. A staff paper prepared for the Committee on National Statistics, National Research Council, National Academy of Sciences, Washington, D.C.

Treiman, D. J. 1977 *Occupational Prestige in Comparative Perspective.* New York: Academic Press.

Turner, C., and Krauss, E. 1978 Fallible indicators of the subjective state of the nation. *American Psychologist* 33:456–470.

Wade, N. 1980 University and drug firm battle over billion-dollar gene: a lawsuit over interferon may change the informal ways by which researchers exchange materials. *Science* 209:1492–1494.

Wallerstein, M.B. 1984 Scientific communication and national security in 1984. *Science* 224(4 May 1984):460–466.

Wolins, L. 1962 Responsibility for raw data. *American Psychologist* 17:657–658.

APPENDIX: EXPLORATORY CONFERENCE ON SHARING OF SOCIAL-SCIENCE RESEARCH DATA

Participants

CLIFFORD HILDRETH (Chair), Center for Economic Research, University of Minnesota

MURRAY ABORN, Division of Social Sciences, National Science Foundation

BARRY B. BOYER, Law School, State University of New York at Buffalo

JEROME M. CLUBB, Inter-university Consortium for Political and Social Research, University of Michigan

ALAN E. GROSS, Department of Psychology, University of Maryland

TAMARA HAREVEN, Department of History, Clark University

TERRY HEDRICK, Department of Psychology, Kent State University

EVELYN JACOB, Center for Applied Linguistics, Arlington, Virginia

MARTIN L. LEVIN, Department of Sociology, Emory University

HERBERT S. PARNES, Faculty of Labor and Human Resources, Ohio State University

MELVIN W. REDER, Graduate School of Business, University of Chicago

DONALD J. TREIMAN, Assembly of Behavioral and Social Sciences, National Research Council-National Academy of Sciences

Staff for the Committee on National Statistics

EDWIN D. GOLDFIELD, Executive Director

MIRON L. STRAF, Research Director

MARGARET E. MARTIN, Senior Research Associate

LENORE E. BIXBY, Consultant

ELEANOR M. MARTIN, Administrative Assistant

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 categorically

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.

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 welldesigned 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 data collection 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 to 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 recording, 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 back-tracking 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 non-government 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-terrestrial 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 hand-written 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 of 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 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.

Definitions, Products, and Distinctions in Data Sharing

Robert F. Boruch

For simplicity's sake, data sharing here is defined as the voluntary provision of information from one individual or institution to another for purposes of legitimate scientific research. In practice there are, of course, a great many variations on this theme. Some of the variations are suggested by the factors that influence data sharing and its products.

THE PURPOSES AND PRODUCTS OF DATA SHARING

The products of data sharing can serve a variety of beneficial purposes, including:

- verifying, failing to verify, or examining the conclusions of earlier analyses,

Robert F. Boruch is a professor in the Department of Psychology and the School of Education and codirector of the Center for Statistics and Probability, Northwestern University. Background research for this paper was supported by a stipend from the National Research Council and a grant from the National Science Foundation to Northwestern University, Center for Statistics and Probability.

as in public program evaluation or economic research on subsidy programs;
- facilitating education and training through active examples;
- testing new hypotheses or theories using existing data, as in a good deal of economic research;
- facilitating tests of new methods of analysis when the original data are well understood, as in attempts to better estimate cognitive ability using Rasch models or mortality using dual-system estimates;
- using the data collected in one study or series to design other studies and programs, for example, in social programs or for physical or chemical constructions in engineering;
- combining several sets of data to facilitate syntheses of knowledge, decision making, establishing limits or bounds on generalization, as in psychological and other research.

The expected products of data sharing will not always appear, of course, and may not fulfill their purposes when they do. For example, poor research can often be identified from reports or tables, reducing the need for access to raw data. Replicating a study independently is often far more important than reanalyzing the results of the original effort, and this approach also reduces the need for access to raw records. Even when the information is pertinent to a scholar and is of good quality, the stress on sharing can be dysfunctional, for several reasons. The products may be pedestrian, for example, because it can be hard to reason ably and in original ways from data that have already been well analyzed. The process of sharing may lead to self-interested or inept assaults on adequate work, as it has in the past.

Perhaps more important, the stress on repeated analysis of observational data, from surveys for instance, can divert resources from the collection of better data, say, from field experiments, that could yield less ambiguous conclusions. Data may be analyzed because they are available rather than because they are interpretable and clearly material to a problem at hand, producing work that is pedestrian or wrong repeatedly. And so on.

In summary, it is reasonable to expect a variety of outcomes, positive and negative, from data sharing. The position taken here is that sharing in principle is warranted simply because it is part of a durable scientific ethic to communicate results in a way that permits confirmation. In practice, its appropriateness, feasibility, and utility depend on other factors.

Voluntary Versus Involuntary Sharing

There are cases of forced data sharing, in responses to demands for disclosure of information by a court, in the interest of assessing a scientist's claim.

Time and resources are not sufficient to examine such sharing in detail here, but a couple of cases do deserve brief attention: the Longs' efforts to obtain data from the Internal Revenue Service for research purposes and *Forsham v. Harris.*

Dr. Susan Long, a sociologist at Princeton, and Mr. Philip Long, head of a business, have for the past 10 years been involved in efforts to secure statistical and other data from the U.S. Internal Revenue Service (IRS) for research purposes. Susan Long's professional interest lies partly in examining consolidated administrative data and IRS procedures manuals to determine how administrative discretion is used in applying tax law, i.e., how rate of audits varies by geographic region, income level of the taxpayer, etc. (see Long, 1979, 1980a, 1980b). The Longs maintain that the information they request falls within the coverage of the Freedom of Information Act (FOIA). Moreover, since the records on individuals that they need are anonymous, acceding to their request violates no privacy statutes. The IRS has disagreed, refusing, for example, to disclose counts of audits by income category and internal documents on operating procedures for audits. In different court cases dealing with the requests, the IRS has maintained that disclosure of the data tapes or procedures would help taxpayers avoid audits, that the FOIA is not relevant, and that the privacy law will be violated, so the information should not be disclosed. The Longs have also attempted to obtain information on the sampling frame and results of studies generated in IRS probability sample audits; this information has also been refused. They have brought a number of such cases to the courts, winning access to some data under the FOIA in the lower courts. In particular, federal circuit courts have ruled that data tapes were discloseable under FOIA and not subject to laws governing disclosure of IRS records (26 U.S. Code &6103) when identifiers are deleted and the risk of deductive disclosure cannot be shown to be appreciable. The Longs have testified before Congress on the need to make such information more accessible (P. Long and S. Long, 1974a, 1974b, 1976; S. Long and P. Long, 1973).

In *Forsham* v. *Harris,* which was heard by the Supreme Court in 1979–1980, researchers were trying to obtain and reanalyze data generated in the University Group Diabetes Project (UGDP). The project, supported by the National Institutes of Health (NIH), was designed as a randomized field test of alternative methods of treating diabetes and resulted in the conclusion that one of the drugs tested, a popular one, appeared to have had strong negative effects. The study generated a good deal of controversy, and the results were debated by the companies that produce the drug, physicians using it for treatment, interest groups, and statisticians. The original investigator refused requests by independent investigators for the data. The requesters filed suit under the Freedom of Information Act, maintaining that the data were collected under a federal contract for the National Institutes of Health and so

must be regarded as public, except for identification of individual participants in the study. In a 7-2 decision, the court ruled against disclosure. Writing for the majority opinion, Justice William Rehnquist maintained that the law applies to records actually in the government's hands. Because NIH had not asked for the data (at that time), the agency could not be used as a vehicle for getting the data. See Cecil and Griffin (in this volume) for details.

Research Versus Administrative Functions of Data

The emphasis in this paper is on sharing information for scientific research purposes. There is much less stress on sharing for commercial purposes, and no attention is given to data shared in the interest of making specific administrative or judicial decisions about an individual. The distinction between research function and the administrative function of information here is important. It parallels one drawn by the Privacy Protection Study Commission (1977) and adopted in some recently proposed bills on privacy in research and statistics.

The distinction is important since the rules that govern access to records for purposes of making a decision about an individual must differ from those governing access for research. For instance, access for administrative purposes can carry major consequences for an individual, as in credit reporting and criminal records. Access by researchers generally carries no such direct consequences. To judge from evidence obtained by the Privacy Protection Study Commission, abuses are more likely in the administrative use of data; thus, the focus of government rules and professional codes needs to differ depending on who collected the records, who has access to them, and what the purpose of access is.

Despite differences in function, administrative records can often be used for research purposes:[1] see, for example, Chelimsky (1977) on the use of criminal justice records in evaluating crime control programs; Conger et al. (1976) and Peng et al. (1977) on using records to assay accuracy of response in educational surveys; Del Bene and Scheuren (1979) on statistical uses of administrative records from Social Security Administration and other government files for studies of disability; and Bishop (1980) on energy consumption experiments. The uses of administrative records in public health research are sufficient to justify an annual conference on records and statistics that is sponsored by the National Center for Health Statistics and other agencies.

Access to administrative records for research purposes can be at least as important as sharing data originally collected for research purposes, but it raises different problems. The laws or rules governing confidentiality of administrative records on individuals or institutions, for example, can impede researcher access unless special exemptions are created. Such exemptions do

appear for certain kinds of research in the Privacy Act of 1974, governing federal records, and similar exemptions appear in the laws of other countries (Mochmann and Muller, 1979; Flaherty, 1980). However, the opportunity for access to addresses of taxpayers maintained by the IRS virtually disappeared with the Tax Reform Act. Rules in the commercial sector vary considerably and decisions about permitting access appear to be mostly ad hoc, systematic only for the larger companies. Because the situation for private companies is so poorly explored—very little data on access practice exist for administrative records—most of the material here focuses on public administrative or research records.

Contract and Grant Requirements

Two common funding mechanisms for publicly supported research are contracts and grants. Contracts can be and often are written to ensure that the products, a report and the information on which it is based, are provided to government at the contract's end. The idea of data sharing emerges most often in contract work, where the data belong, at least in principle, to the government agency that asked for them. In practice, the accessibility may be explicit in contract provisions (Garner, 1981), but it may be debated in the courts regardless of such provisions.

Research grants also result in data that can be shared, but there has been little stress on routine sharing of such data partly because the data have been treated as property of the principal investigator. Another reason for less attention to data sharing in grants is that most grants are for the support of laboratory research in which replication of the research rather than reanalysis of individual records is paramount.

Precedents for contract requirements to share data are easy to find. The data used in the Coleman et al. (1981) analysis of the relative effectiveness of private and public schools are part of a national longitudinal study of high school students conducted for the National Center for Education Statistics (NCES) by the National Opinion Research Center (NORC). The contract between NCES and NORC specifies that data would be turned over to NCES for storage and distribution. However, although the data were available to other analysts when controversy erupted over the Coleman et al. work, few critics had actually reanalyzed the raw data. Since then, other analysts have worked with the data (Antonoplos, 1981). Of course, access alone will not resolve some policy arguments about the work. For example, measurement of family income was based on children's responses to multiple choice questions, a process that warrants special attention and defense.

Analogous provisions were incorporated into Department of Energy requirements for 16 recent public utility demonstration projects on peak-load

pricing. The data produced and their documentation must be furnished to the department (Federal Energy Administration, 1976) for synthesis and reanalysis. Provisions to ensure that information will be made available to the research community have also been incorporated into contracts by the National Institute of Education for the National Assessment of Education Progress, an annual survey of student performance conducted for the Education Commission of the States, and by the National Center for Health Services Research for Michigan State University's Data Center on Long Term Health Care (Katz et al., 1979), and others.

THE NATURE OF SHARED INFORMATION AND VEHICLES FOR SHARING

The nature of the information that is made accessible varies a great deal. Alloy phase diagram data are consolidated and made available to scientists and engineers through the National Bureau of Standards and the American Society for Metals. The Materials Properties Data Center stores and disseminates machine-readable data on tests on metals and ceramics to government, commercial, and academic users through a facility at Battelle Laboratories, and analogous on-line facilities are under development by the Copper Development Association, the Materials Properties Council, and others (see National Research Council, 1980). The National Bureau of Standards has a major brokerage role in these and in the Fundamental Particle Data Center, Diffusion in Metals Center, the Data Center for Atomic Transition Probabilities, and the Crystal Data Center, to which physical scientists and engineers contribute.

Videotapes of selected commercial and public television broadcasts are accessible to communications researchers, historians, and others at the National Archives, in specialized libraries at George Washington University, Vanderbilt University, and elsewhere (Adams and Schreibman, 1978). Oral history tapes are maintained at Columbia University and elsewhere. Results of acoustic tests are shared, too. One of the dramatic recent examples of the latter involves Bell Telephone Laboratory's audio recordings, generated as part of research under Arthur C. Keller during the 1930s, which recorded, among others, the Philadelphia Orchestra under Leopold Stokowski. The audio products of these technical tests on stereophonic recording methods, amplification, processing, and the like are maintained at the Rogers and Hammerstein Archives at the New York Public Library.

Educational data from large-scale surveys are often made available through a variety of private and public agencies, as are health statistics and welfare data from surveys and social experiments (see below). Such data have been used in basic sociological research to test theroretical models, but they are

probably used more often in applied research to anticipate or estimate the effects of changes in tax law, Social Security and welfare rules, and the like. The administrative vehicles for distribution of these data include general government facilities, such as the National Archives (see Dollar and Ambacher, 1978), specialized ones, such as the Bureau of Labor Statistics, National Center for Health Statistics, and others (see the review by Duncan and Shelton, 1978), academic data banks at the University of Michigan, the University of California, the University of North Carolina, and elsewhere, and private distributors such as DUALabs.

The U.S. National Oceanic and Atmospheric Administration (NOAA) operates a variety of agencies that facilitate or serve as a vehicle for sharing numerical information internationally. At the National Oceanographic Data Center, for instance, routine observation data from private and public sources continually are pooled and updated. The National Geophysical and Solar Terrestrial Data Center archives and distributes data relating to solid earth physics, e.g., volcanoes and earthquakes, geothermics, etc. The National Geodetic Survey Information Center distributes mapping information in machine-readable and other forms to federal, state, and local agencies and scientists.

Whatever the administrative vehicles for sharing data and the nature of the shared data, the process can be remarkably interdisciplinary. For example, economists Cain and Watts (1970) have reanalyzed data produced by educational researchers Coleman et al. (1966) to reach conclusions about the effectiveness of compensatory education. Criminal sociologists Bowers and Pierce (1975) have rebutted Ehrlich's (1975) econometric analyses of the effect of capital punishment on homicide rates, based on publicly available data. Anthropologists have used satellite photos that were initially archived for agricultural and geophysical research to understand herd migration and the effect of new wells in North Africa. The productivity of cross-discipline conversations is also reflected in reanalyses of meteorological experiments (e.g., Braham, 1979, and his critics).

Of course, the feasibility of storing and distributing data depends on the information's character. It seems fair to say that machine-readable numerical data tapes are more suitable for routine sharing and that more is understood about efficiency in their production and distribution that for some other kinds of information, such as videotapes, partly because experience with others is more recent. The problem of ensuring individual privacy and confidentiality has received more attention and appears to be more tractable for statistical research data than for other information. For example, blocking out faces is possible in videotape research on behavior of children or adults in classrooms, but it is difficult. Voiceprint analysis and other methods may make identification possible in analysis of videotapes and audio-taped oral histories.

Because of the diversity of the kinds of data that research on for scientific purposes have to recognize major differences in the nature of information that is shared.

Source Lists

There is of course no universal list of the information that is routinely made available for scientific analysis, although archives that handle machine-readable data often issue regular reports on the data maintained. For instance, the National Technical Information Service (NTIS) and the Office of Statistical Policy and Standards (OFSPS) have regularly issued a *Directory of Federal Statistical Data Files* to assist users in locating what they need. Similar lists are issued by operating agencies for special user groups, e.g., the *Directory of Federal Agency Education Data Tapes* (Mooney, 1979). The problem of maintaining useful inventories of data tapes that can be shared is complicated and severe enough to have received the attention of President Carter's Reorganization Project on the Federal Statistical System. At least one commercial directory is available, the *Encyclopedia of Information Systems and Services* (Kruzas and Sullivan, 1978), which covers bibliographic as well as numerical machine-readable data, but it is not as thorough in coverage as the government listings noted above.

Such lists pertain to data that are stored and distributed by standing archives rather than by individual scientists. To identify new data that may eventually be shared, formally or informally, by scientists or institutes, the annual reports of research supported by private foundations or public agencies can be helpful. Catalogs of applied research and evaluation projects are issued regularly by the U.S. Department of Education and the U.S. Department of Health and Human Services (for example, 1983), the NTIS, and others. The U.S. General Accounting Office issues the *Federal Information Sources and Systems* (for example, 1976, 1980b) describing about 1,000 federal systems bearing on fiscal, budgetary, and program-related data, and *Federal Evaluations* (for example, 1980a), covering over 1,700 reports on specific programs. Either of these reports can be used to guide searches to numerical data that can be reanalyzed by independent researchers.

The final broad class of sources includes statistical reports issued by the government or commercial vendors. The federal government, for instance, serves as a broker in consolidating statistics from disparate sources in such periodicals as *Copper: Quarterly Report, Forest Products Review, Printing and Publishing Quarterly Report, Condition of Education,* and others. Some of the statistics in these publications are based on microrecords that are available from government agencies, such as the Social Security Administration, and from commercial sources, such as Dun and Bradstreet and McGraw-Hill

Information Systems Company. No formal research appears to have been published on the utility of directories such as these, nor have there been any published critiques of the documents.

International Aspects

Data sharing is not confined to researchers in the United States, of course. Danish and German data archives, for instance, serve European social scientists with an interest in accessing and storing social data from field studies (see, e.g., Kaase et al., 1980). New professional organizations such as the American Society of Access Professionals, the International Association for Social Science Information Service and Technology, and the International Federation of Data Organizations have helped to consolidate social scientists' interests in analyzing machine-readable data (Mochmann and Muller, 1979). The International Federation of Television Archives was created by representatives of the broadcasting companies' Television archives, and membership is extended to university-based and other TV archives (Schreibman, 1978). International organizations such as the Organization for Economic Cooperation and Development and UNESCO have begun to try to establish guidelines on data sharing. International exchanges are not uncommon in engineering, to judge from the American Society for Metals/National Bureau of Standards joint effort on data sharing for construction of alloy phase diagrams. A collaborative 12-country effort to assay academic achievement of students, the International Educational Assessment (Jaeger, 1978; Postlewaite and Lewy, 1979), illustrates a similar cooperative effort in educational research.

At the level of the individual researcher, examples of sharing across national boundaries are not difficult to find. The randomized field experiments on nutrition and educational enrichment in Colombia are, for instance, something of a milestone in demonstrating effects of such programs (McKay et al., 1978), and a small group of Colombian and U.S. researchers continue to reanalyze machine-readable results. Exchanges and cooperative projects are not as frequent as they ought to be in engineering, according to the National Research Council (1980) because of problems in nonuniform nomenclature and testing and reporting methods, quality of input, and language. Similar problems doubtless affect sharing in the social and behavioral sciences. Aside from single projects such as the International Educational Assessment and sporadic individual sharing, the stress in social, behavioral, and educational research is on one's own country data. Rules governing international information flows are generally designed for commercial record systems, but they may also apply to scientific data (see Boruch and Cordray, below).

Consolidation Level of Statistical Data

The level of consolidation of the data that are shared also varies. In education, for instance, some archives store individual (and anonymous) student responses to items in achievement tests and make the data available for reanalysis along with other information: for example, Gomez (1981) on tests of ability measures for children of Colombian barrios. More commonly, however, test data on individuals are consolidated to produce a total score. Such totals—for achievement tests, indices of functional or social mobility, or indices of income—have typically been available for reanalysis in educational, psychological, and sociological research.

In the archives that make institutional data available, on banks for example, the data may be aggregated in such a way as to prevent analysis of individual banks, since disclosure of confidential information on the institution may be illegal or unethical. Rather, the independent analyst has access only to summary data on a sample of small clusters of banks, as in the Wisconsin Income and Assets File (Bauman et al., 1970), or on data aggregated to regional or state level, as in most published reports of the U.S. Census Bureau. In still other cases, the data may be made available as summary statistics, obtained from a facility that analyzes the raw data according to prescription of the data requester, e.g., some research on Social Security Administration files (Alexander and Jabine, 1978) and on Internal Revenue Service files under the Tax Reform Act of 1976 (Alexander, 1981).

Much less fine-grained data are customarily available as the summary statistics published in research reports or journal articles, and a good deal can be learned from these. Indeed, what is learned may eliminate or reduce the need for access to the raw data. To the extent that tabulated statistics are designed to exploit all the information in a sample and one is willing to trust that the analysis is appropriate and carried out as described, there may be no need for the raw data from a particular study. That journal publication of even crude details of analysis can be useful in detecting errors in analysis and that some errors will be important and warrant obtaining original data is clear, however: see, for example, Good (1978) and Wolins (1982) for lessons, based on journal articles, about mistakes in analysis and inference.

There are no generally accepted guidelines on what to publish and, partly as a consequence, practice is not uniform.[2] In the interest of ensuring that readers can understand the original analysis and can verify it or not, at least superficially, suggestions on what to publish have been developed by Kruskal (1978) for science indicators, Mosteller et al. (1980) and Chalmers et al. (1981) for journal editors, and the U.S. General Accounting Office (1978) for federal evaluation reports. Such guidelines stress including information

about the nature of samples and randomization, statistical power and significance levels for tests, confidence intervals, the model underlying analysis, and so on.

PRIVACY AND CONFIDENTIALITY AND PROPRIETARY INTERESTS

Two issues in data sharing are debated often. They bear on confidentiality of information and privacy of individuals on whom records are shared and proprietary interests in capitalizing on data. The value of the data themselves, less often debated, is at least as important as are other matters treated in the remainder of the report.

Privacy and Confidentiality

If the information shared for scientific purposes bears on individuals or institutions, then privacy may be a critical issue. Partly as a consequence, a good deal of work has been done on understanding when information on identifiable individuals should remain confidential and how to ensure confidentiality. The work is international, having been undertaken in the United States, Canada, Germany, Sweden, and elsewhere. It spans disciplines, solutions to related problems being developed by statisticians, lawyers, social and behavioral scientists, and others. The following sketch of some developments is based on Boruch and Cecil (1979).

General strategies for ensuring confidentiality can be classified into three broad categories: statistical, procedural, and legal. Statistical strategies include those used in initial data collection, e.g., randomized response, contamination, response aggregation methods, and so ameliorate problems of later data distribution. They also include methods used in the data distribution process to protect against deductive disclosure of information about identifiable individuals based on nominally anonymous records. Deductive disclosure here refers to the possibility of deducing that a particular record, stripped of identifiers, belongs to a particular known individual, or deducing that identified individuals have certain characteristics from published statistical tables (or public-use tapes) and collateral information on the individual. Staff of census bureaus in the United States, the United Kingdom, and Sweden, for instance, have developed algorithms to determine if deductive disclosure is possible in releases of series of statistical tables. The strategies developed to reduce the likelihood of such disclosure include special numerical rounding techniques, error inoculation, and repeated subsampling.

Procedural strategies generally involve nontechnical approaches to reducing privacy or confidentiality problems. The simplest include not obtaining

identification at all at the data collection stage or eliminating identifiers from records at the data distribution stage. More elaborate strategies have been developed to permit linking records (of the same individuals or institutions) from different archives without violating promises of confidentiality made to the individuals on whom records are maintained or laws or rules governing access. Such strategies have been used in small and large linkages, in the private and public sectors, to produce linked records that are more useful for research than the individual files. Applications have been made in marketing, law and sociology, psychology, education, welfare, criminal justice, and other research.

Legal strategies generally focus on the problem of ensuring that research data on identifiable individuals are used only for research purposes. They include so-called testimonial privilege statutes that prevent the courts and administrative agencies from appropriating research records on individuals for the purpose of legal prosecution, and some court decisions are oriented in the same way. Most existing statutes apply to records on individuals, not to records on institutions, though protection of institutional records is in fact older in census law. The new bills in this genre (Privacy of Research Records Act, Confidentiality of Statistical Records Act) would be helpful to a researcher with interest in analyzing another researcher's data, permitting access to records on identifiable individuals under specified conditions.

There are major gaps in the existing legal protection for privacy and in the associated provisions for data sharing. As noted above, most laws apply to individuals, not institutions. Consequently, a researcher working on police departments could offer no statutory assurance that research data on individual departments would be used only for research purposes. Similarly, hospitals that cooperate in epidemiological research have no general protection against the problem of an outsider suing the government to obtain data for nonresearch purposes. An exception involves work covered by the Public Health Service Act.

More important perhaps, the current laws are fragmented, covering special areas such as criminal justice or mental health research; the Census Bureau, the Social Security Administration, and a few other agencies have different specialized statutes. Bills to ensure individual privacy and researcher access, such as the proposed Privacy of Research Records Act, would help to make the law more uniform, but not much work is being done on them. A third major limitation in existing laws is that they usually apply only to federally supported research.

Proprietary Interests

Two kinds of proprietary interests are important. The first concerns individual scientists and the "right" to analyze data, especially data collected by oneself. The second concerns institutional interests in a particular data set and the "right" to control who analyzes it.

In some research, individual interests are often negligible. For instance, individual proprietary interest is now unimportant in a good deal of economic research because the work often relies on public-use tapes or published statistics. The reanalysis of National Bureau of Economic Research studies by Feldstein (discussed below) illustrate the type. One might argue, however, that publication has become routine because proprietary interests have in the past prevented access to individual records on competitive institutions. And exceptions do occur. For instance, it was not possible for some analysts to reanalyze Ehrlich's work on the impact of capital punishment on homicides because consolidated files of the public data he actually used were unavailable; the file had to be reconstructed by Bowers and Pierce (below).

Individual interests are less important when government requirements, represented in contracts or in statements of regulations about grants, specify that the government is entitled to the data. This broad class does not apply to all government agencies, but is material to some important ones. Both the National Institute of Justice (NIJ) and the National Science Foundation (NSF) maintain provisions that require grantees to make data available to other scientists, though conditions of disclosure differ a bit. A number of agencies regularly include provision for construction of public-use data tapes in contracts for surveys, e.g., the National Center for Education Statistics (see below). And other agencies have similar contract provisions for irregular special research, e.g., the graduated work incentive experiments of the U.S. Department of Health and Human Services and the energy consumption experiments of the U.S. Department of Energy.

Individual interests are also less material for research areas in which knowledge is advanced better through replication of experiments or reanalysis of summary statistics (e.g., covariance matrices) than through reanalysis of raw records. Much laboratory research in psychology falls into this category (see the *Journal of Experimental Psychology);* the same is true, though not to the same degree, for X-ray crystallography in chemistry.

Individual proprietary interests emerge more often in research supported by public or private sponsors that have neither policy on access nor consistent practice. More important perhaps, the ability of a researcher to analyze data he or she collects before anyone else does so is regarded as a privilege or right by scientific custom. This tradition has been reiterated explicitly by, among

others, the Committee on Scientific Freedom and Responsibility of the American Association for the Advancement of Science and its chair, John Edsall (Dickson, 1980). Indeed, a tradition of not sharing or of sharing very selectively seems to be not uncommon in the history of science, and secrecy has not always served only selfish interests.

During the seventeenth century, for instance, John Graunt wondered in *Bills of Mortality* whether it is wise to make statistical data on health known generally, though the interest in advancing a new quantitative political science is clear. Earlier, of course, Copernicus and Galileo were catapulted out of a "pythagorian privacy of research" (DeSantillana, 1955), a privacy that had some implications for self-advancement and self-preservation as well as for the advancement of science.

This earlier custom has changed, partly because of research sponsors' interests in the products of their investment, as the NSF and NIJ policies suggest. The occasional but dramatic episodes of fraud may also be pertinent. The change too may stem from a gradual enlargement of what scientists view as an adequate level of communication in science, an ethical matter for Pigman and Carmichael (1950), among others. Jeremy Bernstein (1978), for example, appears to be astonished that Rosalind Franklin and her assistant, Gosling, did not publish their work on DNA structure: "They simply treated it as private and personal data" (p. 154). The idea of making more information available, including raw data, is reflected as well in recent editorial policies for some professional journals and some codes of conduct.

There is in the professional literature a demarcation between individual proprietary interests before and after a report is issued. That is, the privilege of first analysis ends with publication of findings in a scholarly journal. After publication, it is argued that scientists have an obligation to submit results to confirmation, openness to criticism being implicit in publication of a scientific article. Some forms of confirmation or criticism are simply not possible if based on the published material alone. It is partly for this reason that some professional codes and journal policies that encourage data sharing hinge on publication.

Data sharing is acceptable, even encouraged, for some research supported by nongovernment organizations. The American Chemical Society journals, which include many articles by authors in the private sector, make data supplements available to permit independent appraisal of conclusions in published articles. Contributions by commercial laboratories to cooperative efforts, such as the American Society for Metals/National Bureau of Standards alloy phase diagram project, reflect the same spirit. There is not enough evidence on sharing of scientific data in the commercial sector to make any generalization about its frequency. In social science research, private foundations such as the Russell Sage Foundation have supported secondary analyses

of data (e.g., from evaluations of "Sesame Street") and so encourage data sharing in some measure. But most private foundations appear to ignore the matter entirely.

Whether data are shared or indeed can be shared when publication is based on business-supported research varies a good deal. As noted above, some data from independent laboratories, including commercial ones, are pooled for common use in the alloy phase diagram project of the American Society for Metals/National Bureau of Standards, in the Materials Properties Data Center, in some American Chemical Society journals, and others. On the other hand, evidence on toxic chemicals, radiation, pollution risks, and other sensitive topics have often been difficult to obtain. Even reports containing only summaries of evidence are at times impossible to extract (see von Hippel and Primack, 1972; Pigman and Carmichael, 1950). Some of the difficulty in getting data concerns unpublished work or administrative data. But this does not make them any less useful for research, especially when such data are labeled as scientific evidence in public hearings (National Research Council, 1977).

Other difficulties involve potential disclosure of institutional imperfections or what could be exploited in commercial competition. For instance, at least a few contributors to the Materials Properties Data Center are wary about subsequent disclosure of the fact that they supplied certain information because it may put the materials they sell in a bad light. Finally, institutional interests may only be a smoke screen. For example, if data prepared by a company's research unit on quality of work life experiments are found to be poor by independent analysts, individual careers may be negatively affected. In any case, it is a burden to supply information, and the benefits to an enterprise may not be worth the trouble.

ENCOURAGING DATA USE

It is obvious that merely making data accessible does not guarantee that they will be used. Scholars or other potential users may need instruction in how to obtain access to the information and how to use and evaluate it. They may need guidance about exemplary uses and critical review of their own analyses. Incentives may be needed to encourage better exploitation of the data. Most of these problems have been identified elsewhere by specialists in machine-readable data archives (e.g., Robbin, 1981a), engineering (Mindlin and Kovacs, 1979), and the social and behavioral sciences and education (Boruch et al., 1981b). The following remarks illustrate a few approaches to encouraging data use.

The National Assessment of Educational Progress has for the past 10 years conducted annual surveys of student proficiency in conventional academic

subjects such as arithmetic and reading and less conventional ones such as music and visual arts. The achievement tests and sample on which the surveys are based are well designed, judging from commentary on the project. The information generated has been used at local, state, and national levels to understand student performance. But until recently, the raw data on student responses to tests, though available, have not been exploited well by academic researchers. Partly to understand the utility of the data, the National Science Foundation has supported cooperative institutional research on the topic. So, for instance, exemplary analyses have been undertaken by well-known researchers to provide models of what can be done. Workshops have been organized for interested researchers to learn about the data and about new methods of analysis. The most recent round of such workshops in 1981 included participation by science educators, economists, educational researchers, psychologists, and others. The workshops are set up so that participants in the first round prepare their own analyses and present the work for criticism in a second, and the better papers are published in an edited monograph (see Walberg et al., 1981a, 1981b, for details).

Variations on the workshop approach have been tried by the National Opinion Research Center, which developed workshops in 1980 on analysis of longitudinal data available from itself and elsewhere. Short courses on obtaining and analyzing machine-readable data files have been developed by the University of Wisconsin's Data Center (David et al., 1978), the University of Essex Data Archive (*SSRC Data Archive Bulletin,* January 1983), and by other institutions.

An approach to encouraging data use, supporting research that involves reanalysis of existing data, is natural for many foundations. No special efforts to focus solely on the topic seem to have been undertaken, but support under more general competitive grant programs and in special contracted research are not difficult to find. Illustrations include: work on verifying program evaluations in education, e.g., Wortman et al. (1978) on the voucher experiments in Alum Rock and Boruch and Wortman (1979) more generally, supported by the National Institute of Education; research supported by the Agency for Children, Youth, and Families that involves pooling different data sets in the interest of understanding child and family support systems; grants for secondary analyses of publicly supported research by private foundations such as the Russell Sage Foundation, e.g., Rossi and Lyall (1976) on the New Jersey negative income tax experiments and Cook et al. (1975) on the children's television program “Sesame Street.”

A third set of approaches applies to public policy research and other endeavors for which replication is difficult or impossible and the need for independent competing analyses takes precedence over proprietary interests. So, for example, a recent report to the Congress and the U.S. Department of

Education recommended that major policy research data be subject to simultaneous independent analysis in the interest of balanced information (Boruch et al., 1981a). The controversy sometimes produced by primary analysis in policy research can itself influence reanalysis, as shown in the efforts to acquire and analyze data used by Coleman et al. (1981) in their work on private and public schools (Antonoplos, 1981).

The fourth approach is to depend on professional societies for reporting research. Journals can establish policies that ensure that data are accessible and that capable reanalyses are published (see Boruch and Cordray, in this volume). Indeed a fair number of journals in economics do publish competing analyses of the same data. Other disciplines stress original data collection more heavily, however, and are less inclined to publish reanalyses that confirm already published findings, even in short notes. Journals can carry notes on availability of new data sets and can legitimately require full citations when the set is used as a basis for an article. Government publications that summarize data, such as *Condition of Education, Social Indicators,* and *Science Indicators,* can also do better in informing interested readers which agency maintains the data so as to encourage reanalysis (see Kruskal, 1978).

WELL-PUBLICIZED EXAMPLES OF DATA SHARING AND NOT SHARING

Some cases of data sharing or failure to share have been dramatized in the popular and professional press. The following briefly describes a few cases and the lessons that might be drawn from them.

Sociology and Education: Public and Private Schools

High School and Beyond is a longitudinal study of students based on a national probability sample, conducted for the National Center for Education Statistics (NCES) by the National Opinion Research Center in Chicago and directed by sociologist James Coleman. Begun in 1980, the project's main purpose is to follow the progress of young people during the critical transition from high school to work, college, and family, with follow-up data being collected every 2 or 3 years. It is a massive undertaking, involving over 50,000 adolescents and 1,000 schools in the initial sample, with oversampling of special groups, such as Hispanics. NCES makes resulting data accessible to educational, economic, and other researchers (National Center for Education Statistics, 1981a). This includes storage and distribution of raw microdata tapes, tape files that are tailored to commonly used statistical analysis packages, and files that are constructed for special uses.

Coleman and his colleagues issued a draft report in April 1981 containing

analyses of private and public schools, based *only* on the first survey wave (i.e., a cross-section), that generated considerable controversy (National Center for Education Statistics, 1981b; Coleman et al., 1981). The draft report suggested that private schools have a greater impact on student performance even after one accounts for differences in background characteristics of students, geography, and other obvious influences. Arguments against the analyses were made in the popular press, e.g., the *New York Times* and the *Washington Post,* as well as in professional conferences sponsored by the National Institute of Education (Antonoplos, 1981) and the National Research Council (1981).

The data on which these analyses were based were made available in March 1981 by NCES. Until the controversy emerged, however, no major analyses had been undertaken. The controversy appears to have spurred faster partial analyses of published statistics, notably on adequacy of sample size, on measures of academic achievement, and at least one competing analysis of the raw data (Page and Keith, 1981). There seems to be good argument for contracting for several simultaneous competing analyses for such policy-sensitive cases.

Economics: Feldstein and the Effects of Social Security

Feldstein (1974) concluded that the Social Security system discourages household savings considerably, thereby decreasing the money available nationally for investments. The report, published in a premier journal, was called “one of the most influential” of such works, “part of the conventional wisdom” of social security economics, and “important” by popular and professional writers. The work was widely publicized and cited by economists and influenced federal policy. Other analyses of similar data had been undertaken, of course, some making similar conclusions about direction though not magnitude of effect (e.g., Darby, 1979) and others finding no effect (e.g., Munnell, 1974, of the Federal Reserve Bank of Boston).

Several years after the publication of Feldstein's work, D.R. Leimer and S.D. Lesnoy (1980) of the Social Security Administration undertook a critical reanalysis. They initially planned to examine what they regarded as implausible assumptions in the complex set of models that Feldstein used as a basis for analysis of time-series data on savings[3]. To initiate their examination of the sensitivity of the Feldstein models to alternative assumptions, Leimer and Lesnoy attempted to replicate the original analysis, and they discovered a programming error in the original analysis—an error whose correction dramatically changed the nature of the estimated relationship. The error involved the definition and numerical computation of gross social security wealth, an indicator of retirement benefits anticipated by present and future beneficiaries.

Subsequent analyses of corrected time series suggested the effect of gross and net worth on savings is negligible for the data available 1930–1974 and the models tested, including Feldstein's. Leimer and Lesnoy also suggested that their conclusions remain unchanged if the original models are modified to incorporate alternative benefit and tax perceptions, are applied to different time periods, or use different models and indicators of Social Security wealth. Feldstein, who was chairing the American Economic Association meetings session at which the results were presented, concurred that an error had been made but made no statements that suggested a change in his beliefs about direction of the influence of social security on savings.

Feldstein assisted in discovery of the error by making available both published and unpublished reports to Leimer and Lesnoy, as indeed he should, and provided advice and reactions to the authors' questions about why their results differed from his (Leimer and Lesnoy, 1980). The authors also went a couple of steps beyond Feldstein's original analysis with corrected data to ensure that the analyses are not sensitive to plausible alternative conclusions, and those steps are distinctive. They also recognized that other model specifications may yield still different results, though they present no options.

Zoology and Genetics: The Kammerer Affair

During 1910–1920, zoologist Paul Kammerer issued a series of reports, summarized later in book form, on experiments that purported to show that acquired characteristics could be inherited. Particular experiments involved production of midwife toads, a species that normally does not have thumb pads, but did appear to inherit them following Kammerer's techniques, and salamanders with other characteristics.

The work was challenged by William Bateson, who tried between 1917 and 1926 to examine Kammerer's specimens. According to Zirkle (1954), he was not successful in doing so until 1926. Upon succeeding, his finding that “the acquired characteristics, which Kammerer claimed to have made hereditary, turned out to be India ink” (p. 189) was eventually published in *Nature* (cited in Zirkle, 1954). Kammerer eventually published retractions of his claims, maintaining that the specimens had been altered by an assistant.

The controversy appears to have clouded some writers' vision in that they maintain the Kammerer work was legitimate but not replicated. It fed politicized science in the sense of lending support to Lysenko and other Soviet geneticists for their views. The important consequence for science is identifying what was not true, i.e., that there was no evidence for the contention that species could be made to inherit characteristics acquired by their antecedents.

Pathology and Experimental Biology: Hodgkin's Disease Cell Cultures

Researcher John Long claimed success in establishing cell cultures from patients with Hodgkin's disease in work at the Massachusetts General Hospital (cited in Dickson, 1981). Recent work has concluded that the drug cultures are not authentic, and are almost certainly derived from owl monkey tissue. The importance of the claim and its subsequent rejection lies partly in the need to develop such cultures to understand how to treat the disease, in the failures of other laboratory attempts to establish the culture, and the frequent problem of contamination, i.e., original cells being supplanted by a contaminant.

Long has said that he believed the cell lines authentic but now believes they were contaminated. The problem is not uncommon, to judge from frequent contamination of cells by HeLa cells, and identification of the change is difficult. Complicating the matter, however, is the contention that the original investigator has forged data, and admitted to forgery, in a major grant application (see Dickson, 1981).

The discovery that cells were contaminated was made after four major papers on the topic were published, the papers being cited frequently in the professional literature. Discovery was possible in part because samples could be and indeed were available for independent analysis. In particular, the head of the hospital pathology department at which the investigator worked sent samples to UCLA's cell culture laboratory, and their work was later confirmed by the New England Regional Primate Center. The contaminant had indeed been used for virus research in the same laboratory. An independent audit of the work undertaken by the hospital research staff also confirmed that three of the four cell lines were nonhuman, the third being human but not clearly linked to Hodgkin's disease tumors (Harris et al., 1981). The results are important in understanding that cell lines have not yet been established. But it is not yet clear how much theory, constructed on the basis of spurious data from work with the cell lines, will be affected.

TYPICAL EXAMPLES OF DATA SHARING AND THEIR PRODUCTS

The controversies are interesting but do not reveal much about how data sharing is accomplished or about the product of the effort. The cases discussed briefly here have been selected for their diversity, including the size of the original project and disciplinary area and the lessons they teach. (The topical categories here overlap with those of Cecil and Griffin, in this volume, but examples and substance differ.)

Education and Training

Of all purposes of shared data, the pedagogical one is perhaps the most obvious. Datta's (1977) history of research on Head Start preschool programs, for instance, presents persuasive evidence that both original analyses and reanalyses of the original evaluations have been used heavily in graduate training. The idea is not new, of course. Judging from the use of small sets of raw data from actual studies in classical textbooks by Kendall and Stuart and by Snedecor and Cochran, reanalysis of data is commonplace in college and university training. But there are no statistics on frequency of use. Very limited evidence from experience in a Northwestern University program suggests that half of the published papers on reanalysis are done by graduate students or postdoctoral fellows (see Boruch and Wortman, 1978). From lists of papers catalogued as products of reanalysis in the Labor Department's longitudinal surveys of labor markets, in the NCES's national longitudinal studies of high school students (Peng et al., 1977), and in Project Talent, at least 10 percent have appeared as graduate theses or dissertations. Some special training programs and short courses are built around a particular data set: e.g., NSF has sponsored competing analyses of data from the National Assessment of Educational Progress (Walberg et al., 1981a, 1981b). Those efforts that have involved student and faculty collaboration and have resulted in published products include the Moynihan and Mosteller (1972) edited volume on reanalysis of the Equality of Opportunity Surveys and the Cook et al. (1975) reanalyses of data generated in field tests of the children's television program "Sesame Street."

Verification, Disconfirmation, and Robustness Analyses

Partly because weather control experiments are expensive and time consuming, partly because the inferences drawn can have dramatic implications for social policy, at least some of these projects have been subjected to intensive reanalyses. Among others, Project Whitetop (Braham, 1979) has received considerable attention because original work suggested that silver iodide seeding has negative effects on precipitation. Some reanalyses (e.g., Dawkins and Scott, 1979) appear to confirm this and illuminate the reasons for it. Others are skeptical that the effect is real. This particular work is due in no small measure to remarkable record-keeping in the original experiments and agreement among the original investigators to share the data.

In economic research, the Leimer and Lesnoy (1980) examination of Feldstein's (1974) original work on the effect of Social Security on capital

stock (discussed above) was initiated to determine if relaxing certain implausible assumptions had any effect on conclusions. These analyses relied on data available from published statistical abstracts, as did reanalyses of Ehrlich's (1975, 1981) work on capital punishment by Bowers and Pierce (1975, 1981), and others.

In educational research, Moskowitz and Wortman (1981) have reanalyzed the Riverside School desegregation data on reading achievement of Mexican-American children, and their results agreed with the original analyses, despite multiple analyses with more sophisticated methods. Bejar and Rezmovic (1981) reexamined data generated in the Cali, Colombia, randomized experiments in education for impoverished preschool children to corroborate original findings (by McKay et al., 1978) that enrichment programs did indeed exert substantial influence on children. In manpower economics, Director (1981) among others has reanalyzed early work by Mangum (1973) and the U.S. Department of Labor's 1975 studies to argue that gains exhibited by disadvantaged enrollees are almost certainly due to regression to the mean, rather than to the programs, a view that differs notably from the original analyses.

Methodological Studies

Recent Colombian experiments involve field tests of different levels of an educational enrichment program for impoverished children, augmented by a nutritional supplement program (McKay et al., 1978). The original analyses were based partly on standardized ability tests adapted to the Spanish-speaking children. The properties of resulting statistical estimates of ability are not completely understood, though it is clear that the treatments have a differential impact on performance as registered by the tests. In order to understand whether newer methods of estimating ability could be more informative, Gomez (1977, 1978, 1981) exploited the original data using so-called Rasch models, a mathematical representation positing that observed test scores for any individual are a function of latent ability and test item difficulty, each independently estimable. The model appears to yield estimated ability scores on an interval scale and with reasonable statistical properties. It does not change substantive conclusions on the remarkable effects of the education program.

Shared data from large-scale surveys and social experiments, and perhaps also from physical and engineering studies, are a natural vehicle for studies in reliability and validity of reporting, calibration, and the like. Some of these are enumerated in Peng et al. (1977) for educational research, and Bielby et al. (1977) for manpower training. Judging from the Peng et al. report, the relative frequency of such methodological papers is notable: 15–30 percent of all published work, depending on one's definition of methodological study.

Use in Design of Studies, Programs, or Constructions

The tradition of exploiting data in handbooks is a sturdy one in the physical sciences and engineering, and there has been recent broadening of the interest, for metals and alloys at least, in pooling data for fast-moving technology. The interest is reflected in the report of the Panel on Material Properties of Data (National Research Council, 1980) and creation of vehicles for sharing, such as the Material Properties Data Center and the American Society for Metals/National Bureau of Standards alloy phase diagram project, discussed above.

The panel's 1980 report suggests that there is strong interest among industries, government, and academic research institutes in having access to data on mechanical properties of metals and ceramics. The interest appears to be strongest for materials used in aerospace projects, nuclear, solar, and other energy production, transportation, and copper use. The data are used in materials selection, design of configuration and size of components, manufacturing and fabrication, life estimation, life testing, and failure analysis. There are significant efforts already under way to disseminate such data according to the report (see below), but "a broad need [still] exists for coordinated up to date reliable data bases that are accessible to different types of users through the various classical and modern methods of dissemination" (National Research Council, 1980:9). In the developing areas, there are still substantial problems of coordination (including cooperation), standardization of methods for soliciting, reporting, and accepting data, and quality control.

One specialized effort in this area is the Mechanical Properties Data Center (MPDC), designed to acquire and distribute information about properties of materials, especially aerospace materials. When possible, raw data are entered into the system based on test results supplied by private contractors and laboratories and publicly supported research, along with information on the nature of tests that lead to the data and metal processing history and composition. Results of 1.5 million tests are said to be available, and some 200 "new specimens" are added each month (Battelle Columbus Laboratories, 1980a, 1980b, 1980c). There is considerable attention to making the data available to users. The vehicles include a computer-based retrieval system based on alloy condition and form, the type of test (e.g., for compression or tensile strength) and testing variables such as temperatures and load rate. The product data are supplied in a variety of forms including statistical summaries and graphs, individual test results, and reports, and some are consolidated in handbooks and proceedings that are updated periodically.

The generic problems in the project are startlingly similar to those encountered in similar projects in social and behavioral data archives. According to

Mindlin and Kovacs (1979), the difficulties include: (1) obtaining access to data, especially in view of proprietary interests; (2) reformatting input data to accord with output criteria; (3) instructing potential users about the system; (4) user suspicion of data that were not generated by the user's agency; and (5) marketing.

The alloy phase diagram program is a joint venture of the American Society for Metals and the National Bureau of Standards. It is dedicated to acquiring, evaluating, and distributing data on microstructural change in alloys as a function of temperature and alloy composition, the data being summarized in standardized phase diagrams. Both private and publicly supported research laboratories supply the basic data. Cooperation among a variety of institutions is necessary since it is impractical for any single institution to undertake production of data on all types of alloys. The effort is international, involving research units in the United States, Germany, Japan, and other industrialized countries.

The phase diagram program stresses distribution heavily. Diagrams are published as final or provisional in a journal, *Bulletin of Phase Diagrams,* whose editorial board is international. The journal also carries information on how to use the diagrams, references to source articles, and reports and related information (see Bennett, 1980; National Bureau of Standards, 1980a, 1980b).

Combining Studies

Pooling data on the same topic from several sources or examining several studies simultaneously can be an effective vehicle for better understanding of the topic, though technical problems can be severe (e.g., nonindependence of the separate data). In the simplest case, of course, a review of literature constitutes one kind of common pooling. The more numerically oriented combinations take several forms (Glass, 1976).

In some research, for example, one level of combination addresses only statistics available in published articles, not raw microrecords. The approach has been used by Gilbert et al. (1977) to understand likelihood of success in innovative surgery, by Light and Smith (1971) to reconcile conflicting results, by Smith and Glass (1977) to assay distribution of successful and unsuccessful methods of psychotherapy, by Gordon and Morse (1975) in examining likelihood of success and failure in public programs, and elsewhere. There are many such routine uses in engineering. As described above, data on properties of materials and phase diagrams are constructed from data supplied by a variety of sources (National Research Council, 1980).

Combining raw data on individuals from surveys and social experiments with records from administrative archives is not common, but it has become

more so over the past 10 years, partly because the results are illuminating. Some of the effort is designed to understand the structure of error in administrative records or survey responses or both. The interagency linkage study conducted by the Census Bureau, Social Security Administration, and Internal Revenue Service illustrates the type, though sharing is confined to the federal agencies; the same is true for some program evaluations in health care and social welfare (Boruch and Cecil, 1979). In social research, the purpose of combining data sets often is for policy research. Michigan's Archive on Long Term Care, which acquires data on long-term care field experiments, puts it into uniform format, and makes the files available for policy analyses (Katz et al., 1979), and Columbia's Housing Survey Project (Beneridge and Dhrymes, 1981) fall into this category. So do recent contracts of the U.S. Department of Energy with Research Triangle Institute for compilation and standardized analysis of state utility demonstration projects on peak-load pricing that were analyzed earlier in nonuniform, different ways by the individual state utilities (Research Triangle Institute, 1978).

EVALUATION OF DATA-SHARING EFFORTS

While the idea of data sharing in principle is agreeable to many scientists, at least for publicly supported research, what good the sharing does is not often assayed systematically.

To be sure, peer review constitutes a kind of immediate evaluation when plans for large-scale sharing are drawn up and projects that hinge on data sharing are proposed. But these reviews are often neither open to scrutiny nor, more importantly, directed at the utility of the product. The more arrogant directors of a data collection effort may not say that the worth is self-evident, but the implication is there insofar as very little hard evidence on utility of the information is published. The problem of evidence has become more crucial for federally supported work as a consequence of restrictions in resources for collecting new information and increasing congressional and administrative interest in evaluating basic and applied research programs. Apart from political incentives, the problem of understanding how to evaluate the product of data-sharing systems, how to improve them, and when to encourage or terminate them seems a reasonable intellectual problem.

The state of the art in evaluation of information collection efforts, including the product—data sharing—is underdeveloped. Systematic theory on cost/benefit analysis of data has only recently been developed (Spencer, 1980) and only for social survey data used in allocating resources by the Congress. Use of data, much less its value, is difficult to measure even when mission-oriented research is carried out and the resulting data subjected to competing analyses (Boruch and Cordray, 1980). Nonetheless, some crude methods are

available, and they could be applied and refined.

The documents issued as a result of analyses of shared data constitute one indicator of productivity of an archive. But frequency counts of publications and bibliographies that summarize the products and why they are important are not common. Exceptions include Peng et al. (1977) and Taylor et al. (1981) on NCES's national longitudinal studies, Bielby et al. (1977) on the Department of Labor's national study of labor supply, Postlewaite and Lewy (1979) on the international educational assessment, and related products issued by the NRC medical follow-up study, Project Talent, Northwestern's Project on Secondary Analysis, etc. Logs sometimes maintained by data-sharing institutions, such as those of the National Assessment of Educational Progress and National Center for Education Statistics, on requests for tape files, documentation, etc., constitute a major vehicle for tracing further products and their utilization (see Peng et al., 1977). Frequency counts are at least partly corruptible and insufficient. The corruptibility is fair game for measurement research. Sufficiency might be achieved with other indicators.

Quality of the product is important, but systematic research on this is even less common. Exceptions are confined to a few evaluations of medical and oceanographic research programs, and of the use of peer ratings and citation counts as bases for judging the adequacy of institutional work (National Research Council, 1981). The strategies developed in those approaches are generalizable perhaps to products generated by data archives but have not been applied.

The process of sharing data, as well as products such as reports, can also be evaluated in some sense. The questions that might be addressed include: How easy is it to find out about data? How easy and efficient is the process of acquisition or distribution? What are the costs and are they reasonable? How well are data updated, corrected, documented? And so on. Managerial questions such as these are examined at times within archives. But the experience itself is not often discussed in published papers, seems to be less orderly than it might be, and probably would profit from more concerted attention. There are a sufficient number of efforts to develop standards of documentation by Robbin (1981b) and others to make some evaluations of this sort possible. But evaluations of processes of other sorts and especially of product utility are likely to be more difficult.

Vehicles for simple routine monitoring of extent and nature of sharing are sometimes available. For instance, the American Chemical Society's journals department head, Charles Birch, maintains records, for articles published since 1974, on the provision of supplements by authors (e.g., raw data) to the *Journal of the American Chemical Society.* The supplements in microfiche form are available through subscription or ad hoc requests, and estimates of rates of requests for various ACS journals are available (see Boruch and

Cordray, in this volume). Not all journals have a data supplement service of this type, and a monitoring system for those that simply require authors themselves to make data available would have to be invented.

Establishing the impact of sharing regardless of quality or number of the physical products and regardless of the process is likely to be most difficult. Most management decisions based on such data, e.g., at the level of the Assistant Secretary for Policy in the U.S. Department of Health and Human Services, are likely to be barely visible and tangled with other information. Consequently, making an inference about whether the data actually influenced the decision is risky.

Design decisions in engineering and experimentation are typically small and forgettable, and utility of information hard to obtain. Deciding whether a scholarly paper, published on the basis of shared information (or for that matter on unshared information), is a distinctive contribution to scholarship is frustrating, difficult, and will be impossible for some. The whole matter becomes much more difficult with multiple users, of course, when users are barely identifiable.

In summary, formal evaluations of data-sharing efforts are not common, the state of the art in evaluation is underdeveloped, formal evaluation may be warranted to understand the worth of the activity, and a variety of types of evaluation may be possible.

NOTES

1. Some formal research on levels of accessibility of administrative records has been done by Gordon and Heinz (1979) and Sasfy and Siegel (1982) to understand the influence of practice and policy of government agencies and the nature and source of demand for information.
2. The quality of reporting summary data and other aspects of research seems to have improved considerably since Pigman and Carmichael (1950) identified good reporting as an ethical obligation of scientists (p. 644): “Even casual inspection (showed) that many articles are not written so that the work can be repeated.”
3. The time-series data underlying Feldstein's work and used by Leimer and Lesnoy are accessible in published statistical abstracts, e.g., *Annual Statistical Supplement to the Social Security Bulletin, Handbook of Labor Statistics, Current Population Reports of the Census Bureau,* and others (see Leimer and Lesnoy, 1980, Appendices D and E).

References

Adams, W., and Schreibman, F.C., eds. 1978 *Television Network News: Issues in Current Research.* Washington, D.C.: School of Public and International Affairs, George Washington University.

Alexander, L. 1981 Proposed Legislation to Improve Statistical Research Access to Federal Records. Unpublished report. Social Security Administration, U.S. Department of Health and Human Services, Washington, D.C.

Alexander, L., and Jabine, T. 1978 Access to Social Security microdata files for research and statistical purposes. *Social Security Bulletin* 41:3–17.

Antonoplos, D., ed. 1981 Proceedings of the National Institute of Education Conference on Conflicting Research Results. National Institute of Education, Washington, D.C.

Battelle Columbus Laboratories 1980a *Metals and Ceramics Information Center List of Technical Publications.* Columbus, Ohio: Battelle.

Battelle Columbus Laboratories, Mechanical Properties Data Center 1980b Descriptive brochure. Battelle, Columbus, Ohio.

Battelle Columbus Laboratories, Metals and Ceramics Information Center. 1980c Descriptive brochure. Battelle, Columbus, Ohio.

Bauman, R.A., David, M.H., and Miller, R.F. 1970 Working with complex data files: the Wisconsin assets and income studies archive. Pp. 112–136 in R.L. Bisco, ed., *Data Bases, Computers, and the Social Sciences.* New York: Wiley-Interscience.

Bejar, I., and Rezmovic, V. 1981 Assessing educational and nutritional findings in the Cali experiment. In R.F. Boruch and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Beneridge, A.A., and Dhrymes, P.J. 1981 Annual Housing Survey Project. Center for the Social Sciences, Columbia University.

Bennett, L. 1980 Editor's corner. *Bulletin of Alloy Phase Diagrams* 1(1):5.

Bernstein, J. 1978 *Experiencing Science.* New York: Basic Books.

Bielby, W.T., Hawley, C.B., and Bills, D. 1977 *Research Uses of the National Longitudinal Surveys of Labor Market Experience.* Madison, Wisc.: Institute for Research on Poverty.

Bishop, L. 1980 Consideration in Analyzing and Generalizing from Time of Use Electricity Pricing Studies. Paper presented at the Electric Rate Demonstration Conference, Denver.

Boruch, R.F., and Cecil, J.S. 1979 *Assuring the Confidentiality of Data in Social Research.* Philadelphia: University of Pennsylvania Press.

Boruch, R.F., and Cordray, D.S., eds. 1980 An Appraisal of Educational Program Evaluations: Federal, State, and Local Agencies. Report to the Congress. Office of the Assistant Secretary for Management, U.S. Department of Education, Washington, D.C.

Boruch, R.F., Cordray, D.S., Pion, G., and Leviton, L. 1981a A mandated appraisal of evaluation practices: digest of recommendations to the Congress and to the Department of Education. *Educational Researcher* 10(April):10–13, 31.

Boruch, R.F., Wortman, P.M., and Cordray, D.S., eds. 1981b *Reanalyzing Program Evaluations.* 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.

1979 Implications of educational evaluation for evaluation policy. In D. Berliner, ed., *Review of Research in Education* 7:309–361.

Bowers, W.J., and Pierce, G.L. 1975 The illusion of deterrence in Isaac Ehrlich's research on capital punishment. *Yale Law Journal* 85:186–209.

1981 Capital punishment as deterrent: challenging Isaac Ehrlich's research. Pp. 237–261 in R.F. Boruch, P.M. Wortman, and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Braham, R.R. 1979 Field experimentation in weather modification. *Journal of the American Statistical Association* 74:57–104.

Cain, G.G., and Watts, H.W. 1970 Problems in making policy inferences from the Coleman report. *American Sociological Review* 35:228–242.

Chalmers, T.C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, A., and Ambroz, A. 1981 A method for assessing the quality of a randomized controlled trial. *Controlled Clinical Trials* 2(1):31–50.

Chelimsky, E. 1977 The need for better data to support crime control policy. *Evaluation Quarterly* 1 (3):439–474.

Coleman, J.S., Campbell, E.Q., Hobsen, C.J., McPartland, J., Mood, A., Weinfeld, F.D., and York, R.L. 1966 *Equality of Educational Opportunity.* Washington, D.C.: U.S. Government Printing Office.

Coleman, J., Hoffer, T., and Kilgore, S. 1981 Private and Public Schools: Report to the National Center for Education Statistics. National Opinion Research Center, Chicago.

1976 *Reliability and Validity of National Longitudinal Study Measures.* Research Triangle Park, N.C.: Research Triangle Institute.

Cook, T.D., Appleton, H., Conner, R., Schaffer, A., Tomkin, G., and Weber, S.J. 1975 *Sesame Street Revisited.* New York: Russell Sage Foundation.

Darby, M.R. 1979 *The Effects of Social Security on Income and the Capital Stock.* Washington, D.C.: American Enterprise Institute

Datta, L.E. 1977 The impact of the Westinghouse/Ohio evaluation on the development of project Head Start: an examination of the immediate and long-term effects and how they came about. In C.C. Abt, ed., *The Evaluation of Social Programs.* Beverly Hills, Calif.: Sage.

David, M., Robbin, A., et al. 1978 Instructional Materials for Microdata Collection Methods in Economics. Economics and Library Science Data and Computation Center, University of Wisconsin, Madison.

Dawkins, S.M., and Scott, E.L. 1979 Comment. *Journal of the American Statistical Association* 74:70–77.

Del Bene, L., and Scheuren, F., eds. 1979 *Statistical Uses of Administrative Records with Emphasis on Mortality and Disability Research.* Social Security Administration, Office of Research and Statistics. Washington, D.C.: U.S. Department of Health and Human Services.

DeSantillana, G. 1955 *The Crime of Galileo.* Chicago: University of Chicago Press.

Dickson, D. 1980 Research data: private property or public good. *Nature* 284:292.

1981 Contaminated cell lines. *Nature* 289:227–228.

Director, S. 1981 Examining potential bias in manpower training evaluations. Pp. 354–361 in R.F. Boruch, P.M. Wortman, and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Dollar, C.M., and Ambacher, B.I. 1978 The national archives and secondary analysis. *New Directions for Program Evaluation: Secondary Analysis* 4:1–6.

Duncan, J.W., and Shelton, W.C. 1978 *Revolution in United States Government Statistics.* Washington, D.C.: Office of Statistical Policy and Standards, U.S. Department of Commerce.

Ehrlich, I. 1975 The deterrent effect of capital punishment: a question of life and death. *American Economic Review* 65:397–417.

1981 Capital punishment as deterrent: challenging reanalysis. Pp. 262–282 in R.F. Boruch, P.M. Wortman, and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Federal Energy Administration, Regulatory Institutions Office 1976 Experiment Guidelines for Electric Utility Demonstration Projects. Unpublished memorandum, November 8. U.S. Department of Energy, Washington, D.C.

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

Flaherty, D.H. 1980 *Privacy and Government Data Banks: An International Perspective.* London: Mansell.

Garner, J. 1981 National Institute of Justice access and secondary analysis. Pp. 43–49 in R.F. Boruch, P.M. Wortman, and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Gilbert, J.P., McPeek, B., and Mosteller, F. 1977 Progress in surgery and anesthesia: benefits and risks of innovative therapy. Pp. 124–169 in J.P. Bunker, B.A. Barnes, and F. Mosteller, eds., *Costs, Risks, and Benefits of Surgery.* New York: Oxford University Press.

Glass, G.V. 1976 Primary, secondary, and meta-analysis of research. *Educational Researcher* 5 (10):3–8.

Gomez, H. 1977 Evaluating Longitudinal Data with the Use of Rasch Model. Paper presented at the 41st Session of the International Statistical Institute, December 4–15, New Delhi, India.

1978 The Analysis of Growth. Ph.D. dissertation, psychology department, Northwestern University. (Available from University Microfilms, Ann Arbor, Mich.)

1981 Reevaluating educational effects in the Cali experiment. Pp. 283–295 in R.F. Boruch, P.M. Wortman, and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Good, I.J. 1978 Statistical fallacies. Pp. 337–349 in W.H. Kruskal and J.M. Tanur, eds., *International Encyclopedia of Statistics* (Vol. 1). New York: Free Press.

Gordon, A.C., and Heinz, J.P., eds. 1979 *Public Access to Information.* New Brunswick, N.J.: Transaction.

Gordon, G., and Morse, E.V. 1975 Evaluation research. *Annual Review of Sociology* 1:339–362.

Harris, N.L., Gang, D.L., Quay, S.C., Poppema, S., Zamecnik, P.C., Nelson-Rees, W.A., O'Brien, S.J. 1981 Contamination of Hodgkin's disease cell cultures. *Nature* 289:228–230.

Jaeger, R.M. 1978 About educational indicators: statistics on the conditions and trends in education. *Review of Research in Education* 6:276–315.

Kaase, M., Krupp, H., Pflanz, M., Scheuch, E.K., and Simitis, S., eds. 1980 *Datenzugang und Datenschutz.* Mannheim, Germany: Athenaum. (In German)

Katz, S., Hedrick, S.C., and Henderson, N. 1979 The measurement of long-term care needs and impact. *Health and Medical Care Services Review* 2(1):1–21.

Kruskal, W.H. 1978 Taking data seriously. In Y. Elkana et al., eds., *Toward a Metric of Science: The Advent of Science Indicators.* New York: John Wiley & Sons.

Kruzas, A.T., and Sullivan, L. V., eds. 1978 *Encyclopedia of Information Systems and Services* 3rd ed. Detroit: Gale Research Co.

Leimer, D.R., and Lesnoy, S.D. 1980 Social Security and Private Savings: A Reexamination of the Time Series Evidence Using Alternative Social Security Wealth Variables. Paper presented at the 93d Meeting of the American Economic Association, September 6, Denver, Colo.

Light, R.J., and Smith, P.V. 1971 Accumulating evidence: procedures for resolving contradictions among different research studies. *Harvard Educational Review* 41:429–471.

Long, P., and Long, S. 1974a Statement before the Senate subcommittee on administrative practice and procedure, February 28. Washington, D.C.

1974b Statement at hearings before the Senate subcommittee of the Committee on Appropriations for the Treasury, April 10. Washington, D.C.

1976 Statement at hearings before the Senate subcommittee of the Committee on Appropriations for the Treasury, April 22. Washington, D.C.

Long, S.B. 1979 The Internal Revenue Service: Examining the Exercise of Discretion in Tax Enforcement. Paper presented at the Annual Meeting of the Law and Society Association, May 11. San Francisco.

1980a *The Internal Revenue Service: Measuring Tax Offenses and Enforcement Response.* Washington, D.C.: U.S. Department of Justice.

1980b Measuring White Collar Crime: The Use of the "Random Investigation Method for Estimating Tax Offenses." Paper presented before the Annual Meeting of the American Society of Criminology, November 5. San Francisco.

Long, S., and Long, P. 1973 Statement before the Senate subcommittee of the Committee on Appropriations for the Treasury, February 28. Washington, D.C.

Mangum, G.L. 1973 *A Decade of Manpower and Development and Training.* Salt Lake City: Olympus.

McKay, H., Sinisterra, L., McKay, A., Gomez, H., and Llorenda, P. 1978 Improving cognitive ability in chronically deprived children. *Science* 200:270–278.

Mindlin, H., and Kovacs, G.J. 1979 The Mechanical Properties Data Center and numeric on-line information systems.

Pp. 13–22 in J.A. Graham, ed., *Use of Computers in Managing Material Property Data.* MPC-14. New York: American Society of Mechanical Engineers.

Mochmann, E., and Muller, P.J., eds. 1979 *Data Protection and Social Science Research.* Frankfurt, Germany: Campus Verlag.

Mooney, E.D. 1979 *Directory of Federal Agency Education Data Tapes.* Washington, D.C.: National Center for Education Statistics.

Moskowitz, J., and Wortman, D.M. 1981 Reassessing the impact of school desegregation. Pp. 322–340 in R.F. Boruch, D.M. Wortman, and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Mosteller, F., Gilbert, J.P., and McPeek, B. 1980 Reporting standards and research strategies for controlled trials: agenda for the editor. *Controlled Clinical Trials* 1:37–58.

Moynihan, D.P., and Mosteller, F., eds. 1972 *On Equality of Educational Opportunity.* New York: Vintage Books.

Munnell, A. 1974 *The Effect of Social Security on Personal Savings.* Cambridge, Mass.: Ballinger.

National Bureau of Standards, Alloy Data Center 1980a ASM/NBS Alloy Phase Diagram Program. Unpublished memo. Washington, D.C.

1980b Alloy Data Center Publications. National Bureau of Standards, unpublished bibliography , September. Washington, D.C.

National Center for Education Statistics 1981a Policy seminar sponsored by NCES in conjunction with the Horace Mann Learning Center. NCES Announcement, April.

1981b NCES data tapes now available for High School and Beyond. NCES 81–226a. NCES Announcement, February.

National Research Council 1977 *Perspectives on Technical Information for Environmental Protection.* Steering Committee for Analytic Studies for the U.S. Environmental Protection Agency. Washington, D.C.: National Academy of Sciences.

1980 *Mechanical Properties Data for Metals and Alloys: Status of Data Reporting, Collecting, Appraising, and Disseminating*. Panel on Mechanical Properties Data for Metals and Alloys, Numerical Data Advisory Board. Washington, D.C.: National Academy Press.

1981 Synopsis of the *Ad Hoc* Meeting on Private and Public Schools. Unpublished report. Committee on National Statistics, National Academy of Sciences, Washington, D.C.

Page, E.B., and Keith, T.Z. 1981 Effects of U.S. private schools: a technical analysis of two recent claims. *Educational Researcher* 10:1–7.

Peng, S.S., Stafford, C., and Talbert, R.J. 1977 *Review and Annotation of Study Reports: National Longitudinal Study.* Washington, D.C.: National Center for Education Statistics.

Pigman, W., and Carmichael, E.B. 1950 An ethical code for scientists. *Science* 111:643–645.

Postlewaite, T.N., and Lewy, A.

1979 *Annotated Bibliography of IEA Publications (1962–1978).* Stockholm: International Educational Assessment, University of Stockholm.

1977 *Personal Privacy in an Information Society.* Supt. Doc. No. 052-003-00395-3. Washington, D.C.: U.S. Government Printing Office.

Research Triangle Institute 1978 Project Pooled Analyses: Feasibility of Combining Data from Several Electric Utility Rate Demonstration Projects. Report prepared by the U.S. Department of Energy, Office Utility Systems. Research Triangle Institute, Research Triangle Park, N.C.

Robbin, A. 1981a Strategies for improving utilization of computerized statistical data by the social scientific community. *Social Science Information Studies* 1:89–109.

Robbin, A. 1981b Technical guidelines for preparing and documenting statistical data for secondary analyses. In R.F. Boruch, P.M. Wortman, and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Rossi, P.H., and Lyall, K.C. 1976 *Reforming Public Welfare: An Evaluation of the New Jersey Income Maintenance Experiment.* New York: Russell Sage.

Sasfy, J., and Siegel, L. 1982 *A Study of Research Access to Confidential Criminal Justice Agency Data.* McLean, Va.: Mitre Corp.

Schreibman, F.C. 1978 Television news archives: a guide to major collections. Pp. 89–110 in W. Adams and F. Schreibman, eds., *Television Network News: Issues in Current Research.* Washington, D.C.: School of Public and International Affairs.

Smith, M.L., and Glass, G.V. 1977 Meta-analysis of psychotherapy outcome studies. *American Psychologist* 32:752–760.

Spencer, B.D. 1980 *Benefit-Cost Analysis of Data Used to Allocate Funds.* New York: Springer-Verlag.

Taylor, M.E., Stafford, C.E., and Place, C. 1981 *National Longitudinal Study of the High School Class of 1972 Study Reports Update: Review and Annotation.* Washington, D.C.: National Center for Education Statistics.

U.S. Department of Health and Human Services 1983 *Compendium of HHS Evaluation Studies.* Washington, D.C.: U.S. Department of Health and Human Services.

U.S. General Accounting Office 1976 *Federal Information Sources and Systems: A Directory for the Congress.* Washington, D.C.: U.S. General Accounting Office.

1978 *Assessing Social Program Impact Evaluations: A Checklist Approach.* Washington, D.C.: U.S. General Accounting Office.

1980a *Federal Evaluations: A Directory Issued by the Comptroller General* . Washington, D.C.: U.S. General Accounting Office.

1980b *Federal Information Sources and Systems: A Directory for the Congress.* Washington, D.C.: U.S. General Accounting Office.

von Hippel, F., and Primack, J. **1972** Public interest science. *Science* 177:1166–1171.

Walberg, H., Anderson, R.E., Miller, J.D., and Wright, D.J. 1981a Policy Analysis of National Assessment Data. University of Illinois, Chicago Circle.

1981b Probing a model of educational productivity in science with national assessment samples of early adolescence. *American Educational Research Journal* 18(2):233–249.

Wolins, L. 1982 *A Critical Commentary on Research in the Social and Behavioral Sciences.* Ames, Iowa: Iowa State University Press.

Wortman, P.M., Reichardt, C.S., and St. Pierre, R.G. 1978 The first year of the education voucher demonstration: a secondary analysis of student achievement scores. *Evaluation Quarterly* 2:193–214.

Zirkel, C. 1954 Citation of fraudulent data. *Science* 120:189–190.

Justifications for and Obstacles to Data Sharing

Terry Elizabeth Hedrick

INTRODUCTION

Several types of data sharing have been described in the preceding paper by Boruch, from relatively passive efforts to intensive efforts involving the provision of large computerized data files and extensive accompanying documentation. That the sharing of data in many instances has led to significant benefits is easy to document. Yet a simple, unqualified endorsement of the practice would be both unrealistic and irresponsible. Many parties have interests at stake when data are shared, and the appropriate balancing of these interests is not always clear. The complexity of the issues is attested to by controversies over data release and reanalysis described in the popular press and the

Terry Elizabeth Hedrick, a social psychologist specializing in program evaluation, is a group director with the U.S. General Accounting Office. The views expressed in this paper are the views of the author and do not necessarily reflect the policies of the U.S. General Accounting Office.

scientific literature (see, e.g., *Nature,* 1980; Feldstein, 1980; Hedrick et al., 1978; Wolins, 1962).

This paper is based on the general premise that data sharing is a desirable and worthwhile practice. Thus it is organized around discussions of justifications for data sharing and obstacles that impede it.[1] When possible, actual examples are provided. The lack of empirical information on benefits and problems associated with data sharing means that these discussions may be somewhat biased toward the more controversial cases that have received public attention; in addition, in some cases, only one side of a controversy may have been fully documented. Therefore, this paper should be read as an attempt to identify, rather than to quantify, demonstrated or anticipated benefits from data sharing and obstacles to an across-the-board institutionalization of such a practice.

In the paper, the interests of the following parties are identified and discussed:[2]

Primary researchers—persons originally responsible for collecting or analyzing th e data or in some cases for funding its collection and analysis.

Research participants—persons or units from whom data have been collected: people, firms, towns, states, schools, etc.

Data requesters—researchers or other persons requesting release of data.

Scientific community—all members of the research community.

Society—all persons.

As will be seen, the benefits and burdens of data sharing are not evenly distributed across these parties, and their interests can vary according to the characteristics of each particular case. Guidelines on data sharing must be responsive to the diverse interests and circumstances.

JUSTIFICATIONS FOR DATA SHARING

Justifications for data sharing are based on demonstrated or anticipated benefits for specific parties. To a large degree, the beneficiaries of data sharing are the scientific community, data requesters, and society; to a lesser degree and under some circumstances, primary researchers and research participants may also realize gains. A variety of benefits associated with data sharing are discussed below.

Reinforcement of Open Scientific Inquiry

One of the most widely held tenets of science is that research should be conducted and reported in a manner that yields sufficient information to enable people other than the original researchers to assess its merits and to replicate

it. While the majority of researchers are likely to interpret this tenet as referring to the provision of careful descriptions of study procedures—as provided for in most journal articles—the provision of data for reanalysis can serve similar functions. The establishment of a policy of data sharing by professional organizations, journals, research institutions, and government could serve to reinforce the openness of scientific inquiry, thereby benefiting the scientific community and society.

Verification, Refutation, or Refinement of Original Results

Probably the most significant benefits realized through the sharing of data stem from reanalyses by other researchers. These benefits include the verification and refinement of original findings and the refutation of them. Secondary analysts may reanalyze data by following the original researcher's methods, thus checking the accuracy of the reported results, or by using competing analytic techniques or sets of assumptions, thus testing the robustness of the original conclusions to alternative approaches. If independent reanalyses are done conscientiously and with visibility, the credibility of the original research may be enhanced.

When research results have entered the policy process, the sharing of data to permit reanalysis is extremely important. Analyses that confirm the original results can help combat political pressures to deny or bury them.

The Wortman et al. (1978) reanalysis of data from the Alum Rock Education Voucher Demonstration Program is a good example of a reanalysis that refined the original work. Initial reports on the voucher demonstration posited a relative loss or no gain in reading achievement for the six voucher schools (Barker, 1974; Klitgaard, 1974). Wortman et al. used a quasi-experimental design with multiple pretests and individual-level data and concluded that the deleterious effect reported earlier was confined to a few non-traditional programs within the six schools.

The complex analyses and large data sets now used in much social science research have increased the susceptibility of findings to statistical and programming errors, errors unlikely to be detected without intensive review or reanalysis of the data. As Martin Feldstein said (1980:96):

> When economists deal with large data sets and complex econometric operations, there will be mistakes. If anyone relies on one study, he runs the risk of being misled by an error or statistical fluke. Indeed all models are untrue in the sense that they are crude approximations to the real world.

A dramatic illustration of this susceptibility comes from a reanalysis of Feldstein's own early work, exploring the effect of Social Security on personal

saving behavior, by two analysts of the Social Security Administration, Dean Leimer and Selig Lesnoy. At a 1980 conference of the American Economic Association, Leimer and Lesnoy showed that an elementary computer programming error led Feldstein to greatly overestimate the negative impact of Social Security on saving behavior. Although Feldstein later took issue with the Leimer and Lesnoy claim that the introduction of Social Security has not substantially reduced personal saving, he acknowledged the programming error and stressed that such replication studies are at the core of the scientific tradition (Feldstein, 1982).

The Ehrlich research on the deterrent effect of capital punishment is a classic case of research findings quickly entering the policy process without provision for timely reanalysis by other interested parties. In 1975 Ehrlich published an article claiming that between 1935 and 1969, each execution in this country prevented seven to eight murders (Ehrlich, 1975). At the time of publication, the Supreme Court (in *Fowler v. North Carolina*) was reconsidering its 1972 decision declaring capital punishment unconstitutional, and U.S. Solicitor General Robert Bork used the study results in an amicus curiae brief filed by the Justice Department to argue for the reinstitution of capital punishment. The data on which the research was based were not immediately available to other researchers, so it was impossible for other parties to determine the quality of the work.[3]

Examples of errors in analyses or assumptions that led to distorted or incorrect results are not difficult to locate. Steven Director's (1979) work in the area of evaluations of employment and training programs, for instance, demonstrated that past evaluations had used approaches that probably underestimated the impacts of employment and training programs on postprogram earnings of enrollees. Campbell and Erlebacher (1970), using a simulation technique, concluded that many evaluations of compensatory education programs were likely to have suffered from similar problems. Magidson's (1977) and Rindskopf's (1978) applications of competing analytic techniques to Head Start and Title I data were based on similar concerns.

The evaluation field is not necessarily more prone to these problems than other fields. Wolin's efforts in the early 1960s to acquire and reanalyze several small data sets underlying articles in psychology journals were based on a suspicion that the original analysts had used inappropriate analytic techniques (Wolins, 1962). More recent work by Wolins (1978), a secondary analysis of Bayer and Astin's (1975) data on faculty salaries, was concerned with problems of nonadditivity and irrelevant variance in the predictors and challenged the conclusion that the data supported a finding of a sex differential in the academic reward system.

These kinds of concerns have motivated several observers to call for simultaneous or serial analyses of evaluative data sets, arguing that data with significant

potential for influencing public policy should undergo analysis by several different researchers (Cronbach et al., 1980; Boruch and Cordray, 1980; Raizen and Rossi, 1981). Cronbach's 65th thesis of program evaluation states: "In any primary statistical investigation, analyses by independent teams should be made before the report is distributed" (Cronbach et al., 1980). That such a prepublication reanalysis policy can be beneficial is corroborated by the experience of Stephen Fienberg, who, as editor of the *Journal of the American Statistical Association (JASA),* required authors of manuscripts to simultaneously submit copies of data. In Fienberg's judgment, many of the articles submitted for review were subsequently strengthened by alternative analyses conducted by journal referees.[4] Bryant and Wortman (1978) have proposed that similar procedures should be adopted to govern submissions to psychology journals.

The benefits to the public from sharing policy-relevant data to permit verification and refutation of the original conclusions are fairly obvious. Benefits to the scientific community may also result to the extent that faulty studies are not published and, therefore, do not lead other researchers astray, shaping the directions of future research until sufficient numbers of conflicting studies terminate that avenue of inquiry. Public confidence in the worth of research might also be improved. Finally, as Fienberg's experience with *JASA* illustrates, this is one circumstance in which primary researchers may also profit from the sharing of data.

Replications With Multiple Data Sets

Conclusions drawn from the analysis of a single data set are heavily dependent on the quality of that data set and are subject to distortion from its idiosyncrasies—its scope, format, method of collection, etc. The confidence one places in research conclusions can be greatly increased by consistency of results across data sets; conversely, inconsistencies in results across data sets lead one to view research results with skepticism and to engage in a careful exploration of possible reasons underlying those inconsistencies.

The advancement of knowledge, especially in the social sciences, has been hindered by single studies that capture the fancy of a discipline and send researchers off on extended efforts to replicate, refute, or refine the findings of the original study. The pressure for academic researchers to publish is one cause of this reactive approach. Researchers are sorely tempted to publish each study, rather than to pursue a systematic line of inquiry through the execution of multiple studies. In partial response to the proliferation of one-shot studies (and an extremely high submission rate), the *Journal of Personality and Social Psychology* in 1976 instituted an editorial policy that encouraged more systematic research efforts and stronger support for conclusions

(Greenwald, 1976).

Researchers should be encouraged to complement their collection and analysis efforts with analyses of other existing data appropriate for addressing the same questions. To the extent that a policy of sharing data makes researchers aware of other data sets suitable for their needs and encourages the publication of articles that demonstrate similar findings across multiple data sets, sharing data can benefit the scientific community and increase public confidence in research findings.

Exploration of New Questions

In many cases, especially with large surveys or evaluative data sets, the primary researcher's interest in a data set may encompass only a small part of the data set's potential usefulness. Providing other analysts access to such data sets would permit additional benefits to be obtained from the original investments in data collection. In this respect, evaluative data sets collected by private research firms under government contract are one of the most underused sources of information. Contract research firms necessarily must direct their analytic efforts to address specific questions posed by the sponsor agency; time and resource constraints are likely to prevent analysts from branching out and exploring additional questions when the sponsor considers these questions subsidiary and outside the original scope of work. Consequently, these kinds of data frequently pass into oblivion without other parties being aware of them. Hedrick et al. (1978) have provided a discussion of four general categories of obstacles to the acquisition of evaluative data: problems in locating a particular data set and authority for its release; insufficient documentation; inappropriate aggregation; and delays and refusals to data requests. Many of these obstacles are applicable to the present discussions.

A positive example of an effort to increase returns from investments in data collection can be found in the Employment and Training Administration's (ETA) dissemination and support activities with respect to the production of public-use tapes from the Continuous Longitudinal Manpower Survey. This survey collected information from quarterly samples of enrollees in CETA programs (employment and training services delivered under the Comprehensive Employment and Training Act) and includes, or will eventually include, three years of postprogram labor force and welfare participation data, as well as Social Security earnings information over an extended period. ETA's interest in the survey was initially largely confined to descriptions of the characteristics of CETA enrollees and estimates of earnings gains from CETA participation, but other researchers have been encouraged to exploit the data set for other purposes. From such data-sharing efforts, benefits accrue to data requesters, the scientific community, and society.

Creation of New Data Sets Through Data File Linkages

Another benefit obtainable through the sharing of data is the opportunity to create new data sets by linking two or more existing sources of information. As will be discussed in the section on obstacles to data sharing, this procedure can raise problems of violations of confidentiality, possibly even of privacy (through outright or deductive disclosure of identities), but the potential exists for researchers to address new questions or refine their inquiries into old ones by expanding the kinds and amounts of information available.

The Continuous Longitudinal Manpower Survey is also a good example of data linkage since it involves linking information from CETA agency files, enrollee interviews, and Social Security records to create a single data file rich in detail about CETA participants. On a smaller scale, researchers have combined media reports of daily pollution levels in Los Angeles with Blue Cross of California records of cause, frequency of admission, and length of hospital stay to assess the effects of air pollution on urban morbidity (Sterling et al., 1969). Keesling (1978) creatively merged his own data on school attendance rates with information on reading test scores to examine the contribution of school attendance to achievement-test performance. Again, data requesters, the scientific community, and society are the major beneficiaries of this type of data sharing.

Encouragement of Multiple Perspectives

Every scientific discipline has its own blinders with respect to methodologies, analytic techniques, and the phrasing of research questions. Even the selection of outcome indicators often involves making value judgments about the desirability of certain types of behavior or the characteristics possessed by certain groups of people (Cochran, 1979; Johnston, 1976). The findings of marital instability from the Negative Income Tax Experiments are a case in point (Groeneveld et al., 1980): increases in divorce rates are viewed by some as a positive indicator of women's emancipation; others view them as a negative indicator of the breakdown of the traditional family. Thus analysts may interpret identical variables from different perspectives. Of course, they may also select different variables to address the same questions.

The analytic techniques employed by researchers may also be a function of disciplinary background. Unfortunately, decisions to employ input-output models, analysis of covariance, multiple regression, causal modeling, or other techniques frequently derive less from the nature of the question at hand or the appropriateness of the technique for the data than from a researcher's personal training or past experience. Researchers are most comfortable with analytic techniques that are familiar to them and for that reason they can be indiscriminate

in their use of the techniques.

Data sharing, if it can be extended across disciplinary lines—a large if—has the potential to benefit almost all parties. Sharing data may encourage cross-disciplinary work, permitting questions to be viewed from diverse viewpoints, and it may broaden the perspectives of researchers, including primary researchers, by exposing them to new viewpoints, methodologies, and analytic techniques. To the extent that a broader perspective is taken and the development of knowledge is enhanced, society should profit from the sharing of data.

Reductions in the Incidence of Faked and Inaccurate Results

The existence of dishonesty in science is becoming more and more difficult to ignore. In the past few years, disclosures of hoaxes such as the Piltdown man and Cyril Burt's fabricated data on the inheritance of intelligence have sensitized the scientific community and society to the issue of dishonesty in science. More recently, controversies concerning Dr. John Long's cultures of human Hodgkin's disease cells, which turned out to be monkey cells (Harris et al., 1981), and the 30 drug researchers discovered by the Food and Drug Administration to be faking data or otherwise being dishonest (Hilts, 1981) highlight weaknesses in the current system of peer review and promise to keep public attention on the issue. The costs of errors or dishonesty in science go beyond public loss of confidence in the objectivity of science; in the Long case, several researchers wasted considerable time working with cultures of monkey instead of human cells. Not only did the researchers bear costs in terms of advancement of their careers, but society may have borne costs through time lost in research on cancer. When dishonest science leads other researchers astray or influences policy decisions on public programs, it can have negative effects on the welfare of society that range far beyond the scope of the original research.

Again, motivations for dishonest behavior must be at least partially attributed to the intense competition to publish and to obtain grant money. A policy of open access to data, while far from a complete solution to the problem, might serve as a deterrent to the faking of data and dishonest reporting of research results. The extent of the deterrent potential of open access is unknown, but if even a few researchers are discouraged from dishonesty by fear of discovery and exposure, a data-sharing policy may be cost effective.

Unintentional mistakes are a wholly different problem, and respect is warranted for those researchers who, when shown errors in their work, acknowledge the problems. Although researchers are assumed to carefully check the accuracy of quality control, the pressure to work and publish quickly, the

complexity of many analytic techniques, and simple human fallibility not surprisingly sometimes result in errors. Once again, an acknowledged policy of open access to data and the attendant risk that one's mistakes could be publicly exposed might increase the attention researchers give to their work and, therefore, might improve its quality. Both the scientific community and society would benefit from any reductions in errors resulting from an open-access policy.

Development of Knowledge About Analytic Techniques and Research Designs

Secondary analysis is a fruitful activity for the production of information on analytic techniques and research designs. “To the conscientious analyst, there is often no single, generally accepted way to deal with data stemming from an evaluation” (Rindskopf, 1978:15). The further the study design deviates from the classical experiment, the more confusing the choice of analysis approaches becomes. In evaluation, and presumably in most areas of science, acquiring data sets to explore new analytic approaches promises future benefits to many parties through the development of information on the strengths and weaknesses of analytic approaches. It can also provide justifications for better research designs to detect what are frequently small and elusive treatment (program) effects. At Northwestern University, participants in the Project on Secondary Analysis have been engaged in efforts to advance the state of the art in analyzing effects of education programs since 1974, identifying problems in drawing causal inferences, investigating methods of using multiple analytic approaches to provide evidence of convergent validation, and exploring the biases that result when assumptions of various analytic methods are violated (Boruch and Wortman, 1978).

Provision of Resources for Training

The availability of data for secondary analysis offers benefits for training students, especially in statistics and methodology. Boruch and Reis (1980) have documented a wide variety of payoffs to students from engaging in secondary analysis: reduction in the time necessary to get to the analysis stage of research, lower research costs, gains in knowledge about the nature of evidence, increased experience with analytic procedures, early exposure to the untidy world of applied research in comparison with the world of the textbook, and early entry into discussions of public policy. Fields such as economics have traditionally relied heavily on data collected by others for training graduate students. For postdoctoral programs, which may allow only a one-year training period for exposure to a new area of inquiry, access to

data collected by others may be a necessity if a trainee is to be able to follow a project to completion.

Improved training is a benefit in and of itself, but in many cases, students engaging in secondary analysis have made significant contributions to theory, methodology, and statistics. Magidson's (1977) work with competing analyses of Head Start data, Rindskopf's (1978) analyses of Head Start and Title I data, and Rezmovic and Rezmovic's (1981) efforts to test theories underlying the measurement of psychological traits are but a few examples of high-quality and useful secondary analysis work by students or postdoctoral trainees.

Student requests for the sharing of data are probably more likely to run into obstacles or outright refusals then requests from more established researchers. A policy of open access to data should reduce these obstacles and increase both the quality of student training and students' resources for making significant contributions to their professions and to public policy. A side benefit is that early experiences with data sharing may sensitize these new professionals to both the legitimacy of others' data requests and the need for thorough documentation for their own data.

Reduction of Respondent Burden

A concern of many researchers is to reduce the response burden on research participants, whether participants are students in undergraduate survey courses or people in government programs. This concern is especially salient in government research, for which the clearance procedures of the Office of Management and Budget call for budgets (in terms of hours) to be submitted estimating the amount of respondent burden associated with ongoing or planned data collection efforts. If appropriate data already exist that are suitable for answering a researcher's questions, there is little justification for imposing additional respondent burden. Uncontrolled data collection runs the risks of depleting research subject pools and endangering the future cooperation of participants. The sharing of data, by preventing redundant data collection, can benefit both the scientific community and society.

OBSTACLES TO DATA SHARING

The variety of obstacles to data sharing range from clearly illegitimate refusals for data access by primary researchers who fear criticism to legitimate refusals based on national security considerations. In between are many gray areas in which the legitimacy of refusing access is not easily resolved or in which resources and effort are needed before data sharing can become possible. Low-cost solutions are readily apparent for some of these impediments,

moderate- or high-cost solutions will resolve others, and a few appear intransigent.

The discussion of justifications for data sharing demonstrated that most of the actual or anticipated benefits are received by three parties—data requesters, the scientific community, and society. Primary researchers and research participants are by definition members of these groups and, therefore, they also receive benefits. However, as the following discussion of obstacles indicates, primary researchers and research participants bear the brunt of the costs and risks attendant to data sharing, much more so than other parties. Clearly, the costs and benefits of data sharing are not distributed evenly among the parties involved.

Concern About the Qualifications of Data Requesters

When a primary researcher has reservations about the qualifications of a data requester, he or she may be reluctant to share data for several reasons. First, the primary researcher may anticipate that the data requester will require extensive assistance in developing specifications for the exact variables desired and the most appropriate format for transfer of the data. Second, if the data requester does not have experience with comparable data sets, the primary researcher may anticipate having to respond to repeated requests for guidance concerning interpretation of variables, computer programming, and analytic procedures. Third, and perhaps most importantly, the primary researcher may fear that analyses performed by the requester will be of poor quality and that significant amounts of time will be necessary to review, critique, and re-but those analyses. Finally, the researcher may fear that the data set itself and the original analyses will lose credibility if poor reanalyses are performed that elicit criticism in the scientific community. (These concerns may be exacerbated if the researcher perceives the requester to have personal interests in analyzing the data to demonstrate a particular outcome.) The criticism issue is broached again later in this paper, but it should be noted that there is no evidence that incompetent reanalyses come to overshadow competent primary analyses. Also, the same problem of time-consuming debate between investigators critiquing each others' analyses exists for research that does not involve reanalyses of data; therefore, this problem is not peculiar to secondary analysis.

National Security Considerations

Most researchers recognize that national security can sometimes be a compelling reason for nonrelease of data or even nonpublication of results, although some have argued that even under national security constraints, data

should be held in confidence only for a limited period (Edsall, 1975). The Public Cryptography Study Group, a committee of scholars formed by the American Council of Education at the request of the Defense Department's National Security Agency (NSA), recently devised a review system aimed at limiting publication of computer research on cryptography (Public Cryptography Study Group, 1981). NSA had argued that national security required limits on publication, since other countries might use research results to break U.S. codes and ciphers. Researcher compliance with the Study Group's review procedures is voluntary, but the Study Group presumably agreed with the National Security Agency's position on potential threats to national security.[5]

A rather different type of controversy has arisen with respect to presentation of certain kinds of research at international conventions (see *The Economist,* 1981). New regulations on international traffic in arms specify that an export license must be obtained by anyone giving a foreigner technical data with possible defense applications. Unclassified data are exempt from such a requirement if they have been published because publication puts them in the public domain. The dilemma arises from interpretations of "defense application" and "public domain." The president of the American Vacuum Society has argued that almost all data could be construed to have at least a remote relevance to defense applications. Indeed, this society was forced to un-invite representatives of the Soviet Union, Eastern Europe, and China from a meeting on computer bubble memories in 1980. The society is proposing that convention presentations, like publications, should be considered in the public domain and, therefore, be free from regulation. Freeing convention presentations from control doesn't necessarily solve the problem of information being communicated that could be detrimental to national security, but the labeling of convention presentations as public domain has implications for the timing of access to data by other parties. If a researcher has not yet published about a particular line of work but speaks briefly of it at a convention, it could be argued that other researchers then have the right to request access.

Less obvious in terms of its immediate relevance to national security, but of relevance to the government's long-term ability to enforce federal tax laws, is the dispute between Philip and Susan Long and the Internal Revenue Service (IRS). The Longs requested copies of computer tapes containing some 50,000 IRS audits done between 1972 and 1976. The Longs claim that the information is safely out of data and that privacy arguments are not relevant because identifiers have been removed; the government claims that privacy is an issue and that release of the tapes could enable users to derive the formula the IRS uses to identify taxpayers for audits. (An extended discussion of this case is provided in the paper by Boruch.)

National security considerations differ from most others in that society bears most of the costs of data release, while individual researchers have the most to gain.

Data With Special Problems

Many types of data can be stripped of individual identifers and shared with other parties without affecting their usefulness or violating the confidentiality of respondents; other types obtain a portion of their value from the wholeness of the picture they present. The general issue of violations of confidentiality through data sharing is discussed below, but it must be acknowledged that some types of data have special problems.

One such problem involves the sharing of data that include photographs, videotapes, audiotapes, oral histories, or diaries, for instance, that can result in disclosure of research participants, as can detailed case notes from anthropological studies. While identities of individuals can sometimes be camouflaged by such techniques as pseudonyms, voice distortion, or blackout of faces on film, part of the value of the materials may be their completeness. When researchers have given unqualified pledges of confidentiality to participants, access to the materials by others may effectively be barred unless participants can be recontacted for permission. In other situations, it may be possible to share these types of data selectively by requiring requesters to thoroughly document their interest and need for the data, requiring them to assume responsibility for protecting the confidentiality of respondents and preventing access by third parties, or requiring them to work only with the copy provided and to return that copy when their research is completed. These requirements would be far from infallible, but they could be given teeth through written contracts and the cooperation of journal editors.

A second special problem is that, in certain fields, a data set can represent the fruits of an individual researcher's life work. Not only is it unlikely that such a researcher would want to give away data with so much personal investment, but part of the value of the data may lie in the researcher's exclusive access. An anthropologist, for example, may have spent several years living in a remote culture and may have planned many more years of analysis and writing from the case notes. Under such circumstances, agreements to share case notes are likely to be informal, based on familiar networks of researchers working in the same area or casual assurances that planned fields of inquiry are not overlapping.

Guidelines for data sharing must be flexible enough to encompass these types of special problems. The costs of data sharing are potentially heavy for research participants and, in some circumstances, prohibitively heavy for primary researchers.

Loss of Control of Data

Regardless of written agreements or verbal assurances, whenever a primary researcher releases data to other parties, a loss of control occurs. The original researcher cannot monitor whether the requester uses the data for the purposes originally stated, nor can he or she prevent the data from being passed on to third parties. Risks to the researcher and the research participants are associated with both of these possibilities.

If the primary researcher has not completed analysis of a data set, agreements to share data may be contingent upon the requester using the data to address only questions unrelated to the primary researcher's work. Yet there is no way to enforce such agreements. Releasing the data to a second party risks having that party publish first or competitively submit to the same journals and meetings. Although no incidents of this type have been identified through a cursory review of the literature, at least one individual has expressed concern about it (Shapiro, 1979:1):

> There are a number of reasons why the premature disclosure of research data may be undesirable. One is the fact that a scientist's ideas are his only "stock-in-trade." The scientist has a proprietary interest in being allowed to develop his or her own ideas and to control the release of those ideas. Ultimately, those ideas, like those of an artist or author may deserve commercial protection.

A second risk attendant to releasing data is that secondary analysts may use the data for purposes that differ significantly from the original investigator's stated purpose when obtaining consent from research participants. Many scientists have argued that secondary analysis presents difficult ethical problems in that it is not possible to obtain informed consent for unanticipated uses of data (Menges, 1973; Ware, 1974; Ruebhausen and Brim, 1966). Sociologists appear to have been more sensitive to this issue than other social scientists, perhaps stemming from a long tradition of concern for the ways in which the findings of social science can be used to the detriment of study participants—"class risk." Rainwater and Pittman's (1967) concern with secondary analysis of data from a study to assess pathologies in the Pruitt-Igoe public housing projects is one discussion of the dilemma.

Allowing a second party access to a data set also involves a loss of control in that the second party may release the data to yet other researchers. For controversial topics, the possibility of an active research underground developing is not farfetched. The primary researcher may obtain written or verbal agreements from data requesters barring further release, but, if more than one secondary analyst has access to the data, responsibility for unauthorized release may be impossible to determine. Penalties, such as rejections by journals of submissions based on unauthorized use of data, may be necessary to enforce such agreements.

Fear and Costs of Criticism

Two different bases can underlie a fear of criticism from secondary analysis: fear that purposive distortions of study results or faked data will be exposed, and fear that the secondary analyst will find fault with the original researcher's work.

From society's standpoint, exposure of faked data and correction of inaccurate results are appropriate justifications for the sharing of data; from the primary researcher's standpoint, sharing data is a risky enterprise. It is always uncomfortable to have someone checking your work. Fear of criticism by dishonest researchers is easily understood, but there are also reasons for honest and careful researchers to experience trepidation at the thought of secondary analysis. Regardless of the care given to the original analyses and interpretation, the complexity of many analytic techniques and the rapid advent of new methods of analysis make criticism by a secondary analyst a not unlikely possibility. Such criticism may at times be warranted; in other cases, it may be shoddy and unjustified. It can have two kinds of costs for the researcher. First, it may threaten his or her reputation and esteem and, to an unknown degree, interfere with obtaining money for future work. Second, as mentioned earlier, responding to criticism may become a significant drain on a primary researcher's time.

It is unclear what impact criticism and debate have on society. For instance, the debates over econometric analyses of the deterrent effects of capital punishment have been vigorous since Ehrlich's original work. But we do not know what effects that controversy had, if any, on legislators, their staffs, or the public. For that matter, it is unclear whether distinctions are ultimately made between poorly informed criticism and sound criticism. Does the best work sift out by the end of a debate? What is clear is that there is concern that sharing data will yield criticism entailing substantial costs. Objections have been raised, for example, to the policy recommendations that all major program evaluations be subjected to secondary analyses. The argument by federal managers responsible for evaluation is that the opponents of the results of an evaluation will exploit the opportunity to attack even when faced with a good product. Few incentives are perceived to exist for confirming the results of the original researchers; instead, secondary analysis may be viewed as an opportunity to make a reputation by refuting the work of others.

As more information is shared, research often becomes more vulnerable to criticism. The University Group Diabetes Program, a research effort to contrast multiple therapies for mild diabetes in middle-aged and elderly people, responsibly reported detailed information on baseline variables for each type of therapy group. By doing so, it opened the door to critics (and opponents). As Meier (1975) and others (Jablon, 1979; Sterling and Weinkam, 1979) have noted, twentieth-century science may have more of an adversarial atmosphere

than one of dispassionate, open inquiry (Meier, 1975:521):

> Where the study is conducted with even greater care, on the other hand, and many baseline variables are reported, demands for endless subanalyses and recombinations, not to mention recriminations, seem to be the inevitable result. Of course, I regard more information to be better than less, and I certainly do not mean to suggest that one should refrain from reporting the values of all variables of interest. My complaint is with the carnivorous appetites among critics which seem to be generated by this kind of raw meat.

Poor Communication

There are two very different types of communication problems that are of concern: problems of identifying and locating a data set appropriate for one's needs and problems of resolving disputes when secondary analyses lead to different results.

To begin with, an investigator interested in a particular topic must be aware that a data set appropriate for his or her needs exists; the lack of mechanisms for matching research needs to existing data is a major impediment to secondary analysis (Finifter, 1975). Guidelines from the Office of Federal Statistical Policy and Standards that require federal agencies to submit standardized abstracts summarizing public-use, machine-readable data files for input into the Department of Commerce's *Directory of Federal Statistical Data Files* promise a partial solution to this problem. Focused efforts, such as the archive of data on long-term care at Michigan State University (Katz et al., 1979), can also contribute. A variety of data archives and mechanisms to identify data sets are discussed by Clubb et al. (in this volume). Robbin (1981), for instance, has proposed guidelines for bibliographic forms to ensure access to information on machine-readable data files. Even so, if data sets were not originally intended for use by others and were not products of federal agency collection efforts, the matching of researcher interests and data sources is likely to be a hit-or-miss activity relying on informal communication networks.

If a researcher does identify a data set appropriate for his or her needs, the next step is to locate the people with authority to grant release of the data. While this may sound so straightforward as to not be worth mentioning, experience has shown otherwise (see Hedrick et al., 1978). Even though data collected with public funds are assumed to be public property, federal agencies do not always obtain and archive federally funded data sets. If a data set has been left in the control of the original researchers, there is no guarantee that it will still exist at the time of the access request. Or, if a data set does exist, the people most closely involved with data collection, data processing, and analysis may have changed interests or jobs, may have died, or may not

be willing to expend their time to organize, explain, and transfer the data.

The second problem of communication between users, that arising from challenges to the accuracy of the original analysis, has yet to be addressed in any systematic fashion. As mentioned previously, some researchers have characterized twentieth-century science as an adversary process (Jablon, 1979; Sterling and Weinkam, 1979). For instance, Sterling and Weinkam have described in detail their efforts to seek corrections and clarifications regarding misclassifications of persons in the data of the Dorn Study of Mortality among U.S. veterans. While their account gives only one side of the controversy, it does not breed optimism with respect to the resolution of such disputes (Sterling and Weinkam, 1979:1):

> Our experience indicates that reactions to discovery of and attempts to correct errors in scientific studies are similar to those met by consumer attempts to deal with errors in large commercial computerized procedures. Because scientific "management" appears to opt for an adversary rather than cooperative mode of responding to discovery of errors, much of the value may be lost which secondary analysis has for verifying the validity of past work.

When a secondary analyst has sufficient information to replicate the original analysis and obtains different results, the solution to the communication problem may lie in simply offering the primary analyst (and funding agency) an opportunity to review and comment on the new analysis. If no comment is forthcoming, the secondary analyst then proceeds with publication. When, however, additional information from the primary researcher is required to identify the source of the discrepancy and an adversarial attitude wins out over a cooperative one, the benefits associated with the sharing of data may not materialize.

Data Set Inadequacies

A number of obstacles to data sharing stem from inadequate preparation and retention of data sets. If data are located and agreement is forthcoming for data release, the lack of good documentation can reduce or completely prevent exploitation of a data set by a secondary analyst. Here documentation is meant in its widest sense: sampling frames and study design, copies of the original data collection instruments, data collection procedures, validity information, data transformations, aggregation procedures, procedures followed in creating new variables, etc. Having information on the physical format of data tapes is not enough; some of the most valuable information for a secondary user of data is information about the strengths or weaknesses of data set items. Unfortunately, this type of information is often not treated formally and exists only in the memory of the original researcher. To the extent that formal documentation is poor or not available and the primary researcher is

uncooperative or unavailable, major obstacles exist to the productive use of the data by other parties. It may be next to impossible to create documentation after the fact, or the time necessary to disentangle formats may prove prohibitive.

If use of data by other parties has not been anticipated by the primary researcher, data requesters may find that the data have not been properly archived and maintained. Information recorded on magnetic tapes or disks can deteriorate rapidly unless stored under proper conditions. Placing a data tape on a shelf in one's office provides far from optimal storage conditions; storing card copies of data in a garage or basement is likewise undesirable. Backup copies of data sets should be created and carefully maintained as a matter of normal research procedures.

Recognition and Proprietary Concerns of Primary Researchers

A researcher who invests time and resources in the collection and processing of data deserves the first opportunity to analyze those data and make a contribution to his or her field. Release of data before a primary researcher has had a reasonable opportunity to capitalize on those efforts would be an enormous disservice to researchers and would discourage future data collection. In many cases, the concerns of primary researchers for recognition will not be an obstacle to data release since data requesters will have learned of the data set's existence from the published work of the original researcher. Even in this circumstance, however, problems arise when the requester desires the data for a purpose that overlaps with the data collector's future research plans. Questions concern the scope of data available to other parties—the entire data set or only the portion supporting published material; the specification of a reasonable time period for the primary researcher to regain control of the data; the extent of the primary researcher's obligation to release data for purposes that may overlap with future research plans; and the definition of when data enter the public domain—upon publication, upon presentation at a convention, upon use in court, upon communication in some form of professional correspondence, etc.

The proprietary issue is treated in detail in the Cecil and Griffin paper in this volume, and the complex issues of private versus public ownership are not discussed further here. It is worth noting, however, that the reward structure for scientists is undergoing a significant change. Researchers working in such fields as genetic engineering and econometrics are finding that there are commercial rewards for their work that may supersede the goal of the cooperative pursuit of knowledge (Nelkin, 1981). Since constrained access to research results can increase their commercial value, this shift in the reward

structure may become a major obstacle to data sharing.

A controversy between a public interest group, the Interfaith Center on Corporate Responsibility (ICCR), and Abbott Laboratories and Mead-Johnson (see *Nature,* 1980) illustrates several of the recognition and proprietary issues. ICCR conducted a nationwide survey of infant feeding practices and sought assistance from the Center for Disease Control (CDC) in providing computer resources for analysis of the survey responses. CDC's records, as a public agency, are open to scrutiny by others under the Freedom of Information Act, and Abbott Laboratories and Mead-Johnson requested access to the data. The nutritional quality of baby foods is a controversial topic and one of extreme economic importance to the data requesters. ICCR argued that it should be able to retain control over the data until its analysts analyzed and published their findings; the data requesters argued for immediate access to public records, presumably to get a jump on the ICCR findings.

Similar dilemmas occur with respect to access to company test data on product safety. Companies argue that test data are protected as trade secrets; their release could endanger a company's competitiveness in the marketplace. Other parties see independent analyses of test data as necessary for the development and promulgation of regulations to promote public safety.

There are no obvious resolutions to these kinds of disputes. One party or the other stands to lose, and the degree of loss is likely to vary substantially with the particulars of each case. Guidelines on sharing data must be sensitive to protecting the investments of primary researchers, yet when data have immediate relevance to public policy, the interests of the public in having the best information possible available for use in the decision-making process may be judged to outweigh costs to primary researchers. The case for access is even stronger when the collection of the data was supported by public funds.

Violations of Confidentiality

With the increasing number of surveys, the growth in administrative record-keeping, and the potential to link data files, there is growing concern with protecting the identities of research participants. Sharing data can create problems of violations of confidentiality and even lead to threats to privacy if identified data are transferred, if cross-tabulations are run on variables of low frequency (which can result in deductive disclosure), or if a data set is linked with other data or information sources.

Consider, for example, the case of an education evaluation data set containing information on principal and teacher attitudes and student performance. Even if personal identifiers such as Social Security number and name are deleted before data are publicly released, a second party with knowledge of the

location of the research and publicly available information on teaching assignments may be able to identify individuals through cross-tabulations of variables such as race and sex of teacher and grade level of class. If there is only one black female math teacher at the sixth-grade level, her responses regarding her school, her principal, and her students could become public knowledge. So, perhaps, could the responses of her principal and students.

Another situation in which researchers face problems in reporting results or sharing data is in research conducted in private firms. Access to collect data in a private company may be contingent upon a promise never to identify the company when reporting results and to deny access to the data by other parties. This puts the researcher in an awkward position of not being able to permit others to verify his or her analyses. A recent article on sex discrimination in a private company illustrates the difficulty of camouflaging the identity of a private firm (Hoffman and Reed, 1981). The company asked not to be identified, but the article described it as a Fortune 500 company and gave descriptive information on the number of employees, number of branch offices, and administrative organization to a degree that may have compromised the pledge of confidentiality.

Threats to confidentiality, therefore, should be recognized as an obstacle to data sharing. While these examples and the previous discussion of data with special problems indicate that in some circumstances data may be difficult to share, it is likely that this obstacle is cited much more frequently as an excuse for denying access than is warranted. There are a variety of mechanisms to solve the confidentiality problem, including deletion of identifiers, use of cruder report categories, random subsample release, microaggregation, and error inoculation (see Campbell et al., 1975).

Administrative Inconvenience and Cost

Lastly, there are administrative and cost burdens associated with data sharing that must be balanced against the benefits of access. These burdens largely fall on primary researchers.

Documentation of data is a major burden; it is a labor-intensive activity, and if done properly, involves much more than the provision of a simple list of variables and their format on a computer tape. Robbin (1981) has listed five major parts to documentation: (1) a general study description, (2) a history of the project, (3) a summary of the data processing history, (4) a codebook, and (5) appendices with error listings, glossaries, list of publications, and instructions for using the data file. For large, complex data files with multiple waves of data, the documentation task can be expensive and time consuming.

After documentation is complete, researchers who are willing to share data must either notify the research community of the data's availability or make

arrangements to transfer it to an archive. In addition, there are costs of storage and maintenance, as well as costs of updating the file with new information or correcting errors. When data requests are received, there are still more costs associated with providing copies of the data in a form usable by the requester and responding to inquiries for clarification. These issues are not treated comprehensively here; estimating time or dollars associated with these tasks is beyond the scope of this paper and would depend heavily on the nature of the data set. Nevertheless, researchers who are impatient to proceed with their own work may find the requirements of data sharing to be an unwelcome interference.

SUMMARY

This paper has identified a variety of actual and anticipated benefits and obstacles associated with a policy of data sharing. From these discussions, it has become evident that the benefits and costs of data sharing are not evenly distributed across all parties. Data requesters, the scientific to bear most of the costs. The extent to which these benefits and costs will materialize if a policy of sharing data is institutionalized is an empirical question, in many respects not yet answerable. Except for large-scale survey efforts, the scientific community's experiences with data sharing have been spotty and unchronicled; case studies constitute the major source of information and often document only one side of a controversy.

NOTES

1. It is worth noting that there are also some arguments for *not* sharing data. First, some data are not worth the costs involved in sharing them. In these cases, science might be better served by committing such data to obscurity rather than by their occupying the time and resources of other researchers. Second, encouraging reliance on already existing data sets through data sharing may have negative effects by reducing efforts to collect new data. This can have drawbacks to the extent that science progresses by analyses of independent data and to the extent that researchers become insensitive to the difficulties of collecting good-quality data and simply take variables on existing data sets at face value. Other justifications for not sharing data, such as time and resource burdens on primary researchers, are embedded in the discussions of obstacles to data sharing.
2. Other distinctions, such as those between primary researchers and research funders and between researchers in private and in public institutions, can also be important. For reasons of parsimony, these parties have been treated as one group in this chapter. Other papers in this volume, particularly that of Cecil and Griffin, frequently accord them separate treatment.
3. Subsequently, several other studies refuted or at least failed to confirm Ehrlich's results.
4. Letter to Clifford Hildreth from Stephen Fienberg regarding *JASA* policy on publication of data, October 8, 1979.
5. A minority report was also filed by the Public Cryptography Study Group. It argued against restraints on the publication of nongovernmental cryptography research on several

grounds: national security interests are broader than the interests of NSA; restraints will have negative effects on research in other fields; unconstitutionality; international complications; legal complications; the ineffectiveness of such restraints; and a low perceived threat to NSA's cryptosystems from publication of such material (see Davida, 1981).

References

Barker, P. 1974 Preliminary analysis of metropolitan achievement test scores, voucher schools and Title I schools. Pp. 96–104 in D. Weiler, ed., *A Public School Voucher Demonstration: The First Year at Alum Rock, Technical Appendix.* Technical Report R-1495/2-NIE. Santa Monica, Calif.: Rand Corporation.

Bayer, A.E. and Astin, H.S. 1975 Sex differentials in the academic reward system. *Science* 188:796–802.

Boruch, R.F. and Cordray, D.S. 1980 *An Appraisal of Educational Program Evaluations: Federal, State, and Local Agencies.* Washington, D.C.: U.S. Department of Education.

Boruch, R.F. and Reis, J. 1980 The student, evaluative data, and secondary analysis. In R.F. Boruch, ed., *New Directions for Program Evaluation* 8:59–72. San Francisco: Jossey-Bass.

Boruch, R.F. and Wortman, P.M. 1978 An illustrative project on secondary analysis. In R.F. Boruch, ed., *New Directions for Program Evaluation.* Vol. 4. San Francisco: Jossey-Bass.

Bryant, F.B. and Wortman, P.M. 1978 Secondary analysis: the case for data archives. *American Psychologist* 33:381–837.

Campbell, D.T., and Erlebacher, A.E. 1970 How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth, ed., *The Disadvantaged Child.* Vol. 3. New York: Brunner/Mazel.

Campbell, D.T., Boruch, R.F., Schwartz, R.D., and Steinberg, J. 1975 Confidentiality-preserving modes of access to files and to interfile exchange for useful statistical analyses. In *Protecting Individual Privacy in Evaluation Research.* Committee on Federal Agency Evaluation Research, Assembly of Behavioral and Social Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences.

Cochran, N. 1979 On the limiting of properties of social indicators. *Evaluation and Program Planning* 2:1–4.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F., Weiner, S.S.S. 1980 *Toward Reform of Program Evaluation.* San Francisco: Jossey-Bass.

Davida, G.I 1981 The case against restraints on non-governmental research in cryptography . (A minority report of the Public Cryptography Study Group prepared for the American Council on Education.) *Communications of the Association for Computing Machinery* 24 (7):445–450.

Director, S.M. 1979 Underadjustment bias in the evaluation of manpower training. *Evaluation Quarterly* 3(2):190–218.

The Economist 1981 Science and Technology: Export Law Affects Scientific Meetings. January 24:1326.

Edsall, J.T. 1975 Scientific freedom and responsibility: report of the AAAS Committee on Scientific Freedom and Responsibility. *Science* 188:687–691.

Ehrlich, I. 1975 The deterrent effect of capital punishment: a question of life or death. *American Economic Review* 65:397–417.

Feldstein, M. 1980 *Business Week* October 6:96.

1982 Social Security and private savings: a reply. *Journal of Political Economy* 90(3):630–642.

Finifter, B. 1975 Replication and extension of social research through secondary analysis. *Social Science Information* 14(2):119–153.

Greenwald, A.G. 1976 An editorial. *Journal of Personality and Social Psychology* 33:1–7.

Groeneveld, L.P., Tuma, N.B., and Hannon, M.T. 1980 Marital dissolution and remarriage: the expected effect of a negative income tax program on marital stability. In P.K. Robins, R.G. Spiegelman, and S. Weiner, eds., *A Guaranteed Annual Income: Evidence from a Social Experiment.* New York: Academic Press.

Harris, N.L., Gang, D.L., Quay, S.C. Poppema, S., Zamecnik, P.C., Nelson-Rees, W.A., and O'Brien, S. 1981 Contamination of Hodgkin's Disease cell cultures. *Nature* 289:228–230.

Hedrick, T.E., Boruch, R.F., and Ross, J. 1978 On ensuring the availability of evaluative data for secondary analysis. *Policy Sciences* 9:259–280.

Hilts, P.J. 1981 Research results falsified. *The Washington Post* February 17:A1, A4.

Hoffman, C. and Reed, J.S. 1981 Sex discrimination? The XYZ affair. *The Public Interest* 62 (Winter):21–39.

Jablon, S. 1979 The Uses of Third-Party Data Sets: A View From the Fence. Paper presented at the annual meetings of the American Statistical Association, Washington, D.C.

Johnston, D.F. 1976 The OMB report: social indicators. Pp. 100–105 in *Proceedings of the American Statistical Association.* Part I. Washington, D.C.: American Statistical Association.

Katz, S., Hedrick, S.C., and Henderson, N. 1979 The measurement of long-term care needs and impact. *Health and Medical Care Services Review* 2(1):1–21.

Keesling, J.W. 1978 On school attendance and reading achievement. In R.F. Boruch, ed., *New Directions for Program Evaluation.* Vol. 4. San Francisco: Jossey-Bass.

Klitgaard, R. 1974 Preliminary analysis of achievement test scores in Alum Rock voucher and nonvoucher schools. Pp. 105–119 in D. Weiler, ed., *A Public School Voucher Demonstration: The First Year at Alum Rock, Technical Appendix.* Technical Report R-1495/2-NIE. Santa Monica, Calif.: Rand Corporation.

Magidson, J. 1977 Toward a causal model approach for adjusting for pre-existing differences in the nonequivalent control group situation: a general alternative to ANCOVA. *Evaluation Quarterly* 1:399–420.

Meier, P. 1975 Statistics and medical experimentation. *Biometrics* 31:521.

Menges, R.J. 1973 Openness and honesty versus coercion and deception in psychological research. *American Psychologist* 28:1030–1034.

Nature 1980 Research Data: Private Property or Public Good? 284(March):292.

Nelkin, D. 1981 Intellectual Property: The Control of Scientific Information. Paper prepared for the AAAS Committee on Scientific Freedom and Responsibility.

Public Cryptography Study Group 1981 Report of the Public Cryptography Study Group. Prepared for the American Council on Education. *Communications of the Association for Computing Machinery* 24(7):435–444.

Rainwater, L. and Pittman, D.J. 1967 Ethical problems in studying a politically sensitive and deviant community. *Social Problems* 14:61–72.

Raizen, S.A. and Rossi, P.H. 1981 *Program Evaluation in Education: When? How? To What Ends?* Committee on Program Evaluation in Education, Assembly of Behavioral and Social Sciences, National Research Council. Washington, D.C.: National Academy Press.

Rezmovic, E.L. and Rezmovic, V.A. 1981 A confirmatory factor analysis approach to construct validation. *Educational and Psychological Measurement* 41:75–88.

Rindskopf, D.M. 1978 Secondary analysis: using multiple analytic approaches with Head Start and Title I data. In R.F. Boruch, ed., *New Directions for Program Evaluation* 4:75–88. San Francisco: Jossey-Bass.

Robbin, A. 1981 Technical guidelines for preparing and documenting data. In R.F. Boruch, P.M. Wortman, and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Ruebhausen, O.M., and Brim, O.G. 1966 Privacy and behavioral research. *American Psychologist* 21:423–444.

Shapiro, S. 1979 Release of Primary Data Sets From the Producer's Point of View. Paper presented at the annual meetings of the American Statistical Association, August 1979.

Sterling, T.D., Pollack, S., and Weinkam, J.J. 1969 Measuring the effect of air pollution on urban morbidity. *Archives of Environmental Health* 18:485–494.

Sterling, T.D., and Weinkam, J.J. 1979 What Happens When Major Errors are Discovered Long After an Important Report has been Published? Paper presented at the annual meetings of the American Statistical Association, August 1979.

Ware, W.H. 1974 Computer privacy and computer security. *Bulletin of the American Society for Information Science* 1:3.

Wolins, L. 1962 Responsibility for raw data. *American Psychologist* 17:657–658.

1978 Secondary analysis: in published research in the behavioral sciences. In R.F. Boruch, ed., *New Directions for Program Evaluation.* Vol. 4. San Francisco: Jossey-Bass.

Wortman, P.M., Reichardt, C.S., and St. Pierre, R.G. 1978 The first year of the education voucher demonstration: a secondary analysis of student achievement test scores. *Evaluation Quarterly* 2(2):193–214.

The Role of Legal Polices in Data Sharing

Joe Shelby Cecil and Eugene Griffin

INTRODUCTION

As an abstract principle, the sharing of research data is a noble goal and meets with little opposition. However, when data sharing is attempted in a particular circumstance, the conflicting interests of the parties can thwart the exchange. A glance at the benefits and obstacles to data sharing discussed by Hedrick (in this volume) reveals the reason: few of the benefits and most of the burdens fall to the possessor of a data set. Of course, if the person seeking

Joe Shelby Cecil is at the Federal Judicial Center, Washington, D.C.; Eugene Griffin is in the Department of Psychology at Northwestern University, Evanston, Illinois. We wish to thank Hugh O'Neill, Gilbert Beebe, and other members of the American Society of Access Professionals for assisting us in sorting out the policies of the various federal agencies in disclosing research data. Since we did not accept all of their suggestions, it may be assumed that the errors that remain are our own. This manuscript was prepared for consideration by the Subcommittee on Sharing Research Data of the Committee on National Statistics at its meeting in 1982.

the data set and the person possessing it are colleagues or if the sharing of data is seen by the possessor as beneficial, then the exchange usually takes place without difficulty. But if the possessor does not view the exchange as beneficial, discussion of data sharing can turn quickly to conflict and allegations of the rights and responsibilities of the various parties.[1]

Conflict is rarely over the simple right of possession. More likely, it is conflict in defining the limits of the proprietary interest in the data set retained by the one who develops it. Clearly, one who devotes time and effort to develop a data set has a right to capitalize on the investment through publication of findings based on the data, and an adequate return on this investment may require several publications over a period of time. However, others may wish to verify the initial findings, a purpose well grounded in the traditions of science. As a conflict sharpens, the parties may look to the law in an effort to define the extent of their rights. This paper discusses several areas of the law that are relevant in defining the balance between these conflicting interests.

There is no specific body of law that addresses the sharing of research data. In fact, most relevant legal standards fail to acknowledge the unusual nature of research records.[2] Researchers must turn to statutes and case law developed for administrative records[3] and to literature[4] for the standards to resolve their differences. Since these standards fail to consider the unique characteristics of research data, the results are awkward and unsatisfying. Little effort is made to balance the proprietary rights of the primary researcher and the rights of data requesters. In some circumstances the legal standards do not permit adequate disclosure for data sharing, and in other circumstances they permit such open disclosure that the interests of primary researchers in receiving recognition for their work are threatened. By specifying the legal relationships among the parties, however, a role for professional standards and guidelines can be seen. Professional standards will be most effective in defining data-sharing practices in areas that are unregulated or where federal regulations permit but do not require disclosure.

As discussed in Hedrick (in this volume), data sharing affects the interests of at least five parties: the possessor of the data set, usually the person who developed it; the data requester; the research participants; the scientific community; and society. Frequently the interests of these parties are in conflict. The interests of data requesters and society generally favor access while the interests of the primary researcher and the research participants generally oppose access (see Hedrick, in this volume).

While the law has not specifically attended to the problem of access to data for research purposes, it has acknowledged in other contexts some of the interests of some of the parties. The proprietary interests of primary researchers are recognized through copyright laws.[5] The interests of data requesters are acknowledged in exceptions to copyright protection and in statutes and case

law allowing access to federal records.[6] However, neither of these private interests have received the legal recognition accorded to the public interest in developing and having available accurate information for decision making. Across a wide variety of situations, this underlying public interest in accurate information guides the legal resolution of disputes between persons who seek information and persons who possess it.

The legal standards governing access to research information vary with the public or private employment status of the primary researcher and the source of funding for development of the research record system. This paper considers three circumstances. The first circumstance involves access to research records developed with private funds and in the possession of a researcher supported by a private institution. This is the most basic circumstance, since the proprietary rights of the primary researcher are not affected by public funding of the research. Some proprietary rights are recognized through copyright protection; however, copyright law offers less protection of proprietary interests of a primary researcher than is available through simply with-holding the data set. In this circumstance, professional standards can be most useful in encouraging data sharing.

The second circumstance involves the other end of the spectrum, research records developed and maintained by federal agencies. Legal standards in this area are well developed. The Freedom of Information Act provides a mechanism for data requesters and others to gain access to anonymous federal records. Access to identifiable federal records is more problematic since the restrictions of the Privacy Act of 1974 must be considered. The extent of federal regulation of agency records suggests a limited role for professional standards in this circumstance.

The third circumstance, combining elements of the first two, concerns access to research records developed and maintained by private researchers but sponsored through public funding of the research. This is the most difficult circumstance since the proprietary interest of the primary researcher must be balanced against the broader interest of society, an interest derived from the public sponsorship of the research. This is also an increasingly important circumstance considering the extent of federal support for scientific research.[7] The lack of an effective mechanism for obtaining access to research records in this circumstance also suggests a role for professional standards.[8]

ACCESS TO RESEARCH RECORDS MAINTAINED BY A PRIVATE RESEARCHER SUPPORTED BY PRIVATE FUNDS

The first circumstance occurs when there is a request for access to a data set developed by an independent researcher supported by private funds. This situation usually occurs when a data set is developed through an inexpensive laboratory

study or from publicly available documents without federal funding for the research. The lack of federal support for either the researcher or the individual research project is the essential characteristic of this circumstance. In such a situation the right of the researcher to control access to the data is strongest, since the researcher's proprietary interest in the data is not compromised by public funding of the data collection. Such a researcher may retain and use the information he or she develops just as any individual may exercise a private right over personal information. The rights of those seeking access to data and the rights of research participants are very limited.[9]

Since there is no specific case law or legislation discussing proprietary rights in privately developed research data, those rights must be deduced from the general protection offered to intellectual property by the copyright laws. However, formal copyright protection is not the only means researchers have of protecting their investments. Unlike authors or composers who must publicly distribute their intellectual products to gain from their creation, researchers can benefit from the creation of a data set through publication of analyses without distribution of a data set itself.[10] Protection under the copyright laws, however, must offer incentives for public distribution that outweigh the benefits of private possession if researchers can be expected to take advantage of them.

Public Benefit as the Basis of Copyright Protection

When researchers create a data set, they create objects of value, objects in which they can claim a property right. But it is a property right that can be difficult to protect since the property right is in information rather than in some tangible good.[11] Copyright protection, developed to meet the needs of authors and composers, can provide similar protection to researchers. The foundation of copyright protection is in the Constitution, which gives Congress the power to pass legislation (art. I, §8): “To Promote the Progress of Science and Useful Arts, by Securing for Limited Times to Authors and Inventors the Exclusive Right to their Respective Writings and Discoveries.” “Science” has typically been associated with copyright protection while “Useful Arts” has been associated with patent protection.

This passage can be misinterpreted to imply a general property right in the products of intellectual endeavors. However, the primary purpose of such constitutional protection is to obtain “the general benefits derived by the public from the labors of the authors” (Nimmer, 1980).[12] The Constitution seeks to further the public benefits in “Science and Useful Arts” by guarding the economic rights of authors and inventors (and researchers) in the intellectual property they create.[13] When the private interests of authors or researchers in controlling dissemination of their intellectual product cannot be justified as a

means of obtaining an ultimate public benefit, those private interests will not be sanctioned by the Constitution.

Copyright Protection of Research Data Sets

The specific policy for obtaining the public benefits is expressed in the Copyright Act, which reflects a congressional determination of the optimal balance between the proprietary rights of those who create the information and the public benefits from distribution of that information.[14] According to the Copyright Act (§102), a copyright may be obtained for "original works of authorship fixed in any tangible medium of expression."[15] As implied by the language of the statute, the only two necessary characteristics for copyright protection are originality and tangible expression. A research data set can meet both of these requirements. Research data expressed in any tangible form will qualify for protection, including data on computer tape, disks, paper cards, or even scribbled data in a lab book.[16] While some forms of expression may make it more difficult to obtain copyright protection, the form of the expression will not bar the copyright as long as the expression "can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or a device."

Originality is the more fundamental requirement, since the Copyright Act (§102) restricts protection to "original works of authorship" [emphasis added]. Only a minimum level of originality is required, permitting copyright even if the resulting work is substantially similar to a work previously produced (Nimmer, 1980:2.01[A]). Any nontrivial "distinguishable variation" that results from an author's independent intellectual effort will offer sufficient originality to support a claim for copyright (Nimmer, 1980:2.01[B]; Denicola, 1981).

If a data set is an original expression of a researcher, as described above, the data set can be protected as a "compilation," defined by the Copyright Act (§101)[17] as:

> A work formed by the collection and assembling of preexisting materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship.

Examples of works that have been protected as compilations include city telephone directories,[18] interest and discount tables,[19] and other utilitarian collections of facts (Nimmer, 1980:2.04[B]). Although no instance in case law was found, a scientist's collection of data, arranged in such a way as to permit some meaningful analysis, would certainly qualify as a compilation under the Copyright Act.[20] Since a data set can be eligible for copyright protection, the issue becomes whether the copyright law offers sufficient control over release

and distribution of the data to encourage researchers to choose this form of protection over simple secrecy.

The copyright laws attempt to promote disclosure while protecting the proprietary interest of the creator of a work. Consequently, copyright protection extends only to "original works of authorship" (Copyright Act, §102(a); see also Nimmer, 1980:2.01). This seemingly innocent phrase has been interpreted in ways that do not suit the needs of primary researchers. The emphasis on the original work of the author or artist means that protection extends only to the original *expression* of facts and ideas, not the facts and ideas themselves; a copyright on a data set will not give an exclusive right to the information itself (Nimmer, 1980:2.01; Squires, 1979:205,213).[21] Thus, the copyright will not bar another researcher from creating an identical data set containing the same facts and based on the same ideas if the second data set is developed as an independent effort.[22] This is true even if the purpose of the second researcher is to duplicate the work of the primary researcher.[23] While the policy of the copyright law favoring dissemination may be met, a researcher's interest in retaining control over distribution of the work product may be lost.[24]

Even if the second data set is developed directly from the facts presented in the copyrighted data set, either for a replication of the original analysis or for a novel analysis, there may be no infringement of the copyright held by the primary researcher.[25] In some circumstances, even direct copying of a copyrighted data set will not be an infringement of the rights of the primary researcher. This apparent infringement is justified by the doctrine of "fair use," defined by one commentator as a "privilege in others than the owners of a copyright to use the copyrighted material in a reasonable manner without his consent, notwithstanding the monopoly granted to the owner by the copyright" (Ball, 1944, quoted in Freid, 1979).[26]

The fair use doctrine was first developed by the courts as a means of avoiding unnecessary hindrances to progress in the development of the arts and sciences that could result from a strict interpretation of a copyright owner's exclusive rights (Freid, 1979).[27] The Copyright Act recognizes fair use of a copyrighted work, by limiting the exclusive rights of the copyright owner (§107):

> [T]he fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by [section 106], for purposes such as criticism, comment, newsreporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—

(1) the purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes;

(2) the nature of the copyrighted work;
(3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
(4) the effect of the use upon the potential market for or value of the copyrighted work.

The four factors listed in section 107 form the test for determining if a use that otherwise might be an infringement may be permitted as a fair use of a copyrighted work. Two of the factors, the nature of the use and the economic consequences of the use, seem to be most important in determining whether a use qualifies for the exception to copyright protection (Freid, 1979:466-7; Squires, 1979:216,232).

In general a use that would otherwise be an infringement will be permitted if the use is for a noncommercial educational purpose and results in no apparent economic injury to the copyright holder (Freid, 1979:469). Scholarly and educational uses of copyrighted material have received great deference in determining if a use is to be permitted: courts have permitted liberal use of copyrighted material if science and the arts are furthered.[28] One case noted that the doctrine of fair use will be given broader scope when a "field of learning" is concerned, and a narrower scope when the use is solely for commercial purposes.[29] This deference to scholarly uses is also evident in the legislative history of the Copyright Act.[30]

Some commentators claim that the fair use doctrine can be explained solely by looking to the economic consequences to the copyright holder; if there is no detrimental effect the use will be permitted (Squires, 1979:216,232). The test for determining if the use has an adverse economic effect is prospective: Does the use of the copyrighted work "tend to diminish or prejudice the potential sale of the plaintiff's work?" (Nimmer, 1980:13.05). The relevant comparison is between the actual market for the copyright holder's work and the market that would have existed had the use not occurred (Freid, 1979:472). While it is always difficult to prove that this hypothetical market exists, there must be some evidence that the use diminished the market value of the copyrighted work.[31]

When the use of copyrighted work furthers the constitutional purpose of promoting "the progress of Science and the Useful Arts," without diminishing the market value of the copyrighted work, the courts have little trouble finding that such a use is permitted under the fair use standard. An example of such an instance is *Rosemont Enterprises, Inc.* v. *Random House, Inc.*,[32] in which copyrighted information was used in a biography. The court permitted the use after finding that the use served a public purpose and that the copyright owner did not suffer any detrimental economic effects from the use. Since it will be difficult for the copyright holder of a data set to show a diminished market for the data set if it is used for other scholarly purposes, it is likely that

the courts would find it to be a fair use.

Even if there is some likelihood of demonstrating economic injury, the courts have not been willing to find that scholarly use of a copyrighted work is an infringement of the copyright protection. In *Williams and Wilkins Co.* v. *United States,*[33] a publisher brought an action for infringement against a number of federal medical libraries that had been engaged in photocopying and distributing various copyrighted articles from medical journals and books to agency researchers and other libraries. This case posed a more difficult issue, since the public benefit would seem to be offset by a more obvious economic detriment to the publisher. But in this case, too, the Supreme Court permitted the use, relying heavily on findings that the photocopying practice benefited medicine and research, thereby furthering the constitutional purpose of the protection. Though interpretation of the Court's standard of proof of economic injury is somewhat confused,[34] it seems clear that the holder of a copyright will have a difficult time of proving infringement when the copyrighted work is used is a way that furthers a noncommercial scholarly or educational purpose.

The fair use section of the Copyright Act, along with its legislative history and judicial interpretations, suggests that the use of a copyrighted data set by a researcher for purposes of reanalysis or some other noncommercial scholarly pursuit will not be considered an infringement of the copyrighted work; the difficulty a primary researcher would have in demonstrating a market for the data set, much less a diminution in market value in the data set as a result of its use for research purposes, suggests that a broad range of scholarly uses of the copyrighted work will be permitted without resulting in an infringement of the copyright protection afforded the primary researcher.

Though an individual researcher may have little personal incentive to seek copyright protection for a data set, the publisher of the research may insist on an exclusive copyright to all of the material in the publication, perhaps including published portions of the data. Apparently, scholarly journals, which rely on profits from selling reprints to subsidize publication costs, are particularly eager to bargain for exclusive rights to as much of a scholar's work as possible (Patton, 1980). A publishing contract will specify those rights that are transferred from the scholar to the publisher; researchers under great pressure to publish may have little leverage or interest in bargaining on behalf of others for broad access to the data.

In some circumstances the practices of publishers of scholarly journals may discourage dissemination of research data by undercutting the fair use provisions of the copyright laws. If data are published, the fair use provisions of the copyright law will permit other interested parties to use the data; but when a secondary researcher seeks to publish a reanalysis of a data set previously published, the publisher of the reanalysis may choose not to rely on the fair

use exemption and insist that the secondary researcher obtain a copyright release from the original publisher of the data set. If the original publisher is reluctant to provide such a release or if the process of obtaining such a release is too time-consuming, the reanalysis may remain unpublished. While this is less a legal issue than one of customary practices among publishers, it may still unnecessarily restrict the dissemination of previously published information.

ACCESS TO RESEARCH RECORDS MAINTAINED BY FEDERAL AGENCIES

Records maintained by federal agencies can be a rich source of research data.[35] However, obtaining access to agency records can be a difficult problem.[36] Unlike data sets developed by private researchers, records maintained by federal agencies are governed by a web of federal statutes that are "inconsistent at best and chaotic at worst" (Commission on Federal Paperwork, 1977). These statutes determine the rights of researchers who seek access to federal records.

The basic policy governing access and distribution of federal records is found in the Federal Records Act of 1950,[37] part of the Administrative Procedures Act.[38] These general policies have been modified by the Freedom of Information Act (FOIA),;[39]: and the Privacy Act of 1974.;[40]: Both statutes attempt to establish standards for appropriate disclosure of federal records. However, each was drafted to control abuses from administrative misuse of records and fails to distinguish between access for administrative purposes and access for research purposes. Consequently, researchers seeking federal records must frame their requests within the regulations and standards that do not consider the needs of research.

Not all federal records are accessible through these statutes. Both the Freedom of Information Act and the Privacy Act extend only to federal executive "agencies," defined as:

> Any executive department, military department, Government corporation, Government-controlled corporation, or other establishment in the executive branch of the government … or any independent regulatory agency.[41]

This definition is important for what it omits. The Freedom of Information Act and the Privacy Act do not extend to either the legislative or judicial branches of government, whose agencies generally follow more restrictive policies of disclosure. The General Accounting Office, a congressional agency, has adopted policies that comply with the spirit of the Freedom of Information Act,[42] but this compliance remains a matter of agency discretion rather than a statutory right. Agencies of the judicial branch are not within

the scope of the acts,[43] and no independent statement suggests adoption of these policies.[44] Finally, the statutes do not extend to the Executive Office of the President.[45]

The next section discusses the use of the Freedom of Information Act in obtaining access to anonymous research information maintained by federal agencies. The subsequent section examines the role of the Privacy Act of 1974 in restricting access to identifiable agency records. Finally, the interaction of the Freedom of Information Act and the Privacy Act is discussed in relation to requests for identifiable information when federal agencies are unwilling to disclose the information.

Request for Anonymous Records for Research Purposes—The Freedom of Information Act

The Freedom of Information Act (FOIA) amended the Administrative Procedures Act, a statute that had allowed the government to withhold information "for good cause" or when the requesting party was not "properly and directly concerned."[46] These restrictions permitted federal agencies to interpret the Administrative Procedures Act in ways that severely limited access by private parties to federal records. The FOIA, based on a citizen's "right to know" (Comment, 1976a), was introduced to correct these restrictive practices by assuring "the free flow of governmental information 'necessary to an informed electorate'" (Note, 1976a).[47] The ambiguous "good cause" exemption was replaced by nine specific exemptions.[48] The requirement that a requesting party be "properly and directly concerned" was dropped, with information now being disclosed to "any person." [49]

The FOIA requires federal agencies to make available all information to the public unless the records come under one of the nine specific exemptions.[50] Two exemptions have been used by federal agencies in attempting to restrict disclosure of research information when that information is not already protected by some other statute.[51] Identifiable records may be protected from disclosure under exemption 6, which applies to "personnel and medical and similar files the disclosure of which would constitute a clearly unwarranted invasion of privacy."[52] Other records, including anonymous data, may be protected under exemption 4, which applies to "trade secrets and commercial or financial information obtained from a person and privileged or confidential."[53]

All exemptions to the FOIA are subject to judicial interpretation. Thus far the courts have been very conservative in qualifying information as exempt from disclosure, holding that the nine exemptions of the FOIA are to be narrowly construed.[54] Furthermore, few courts have endorsed the theory of a

court's equitable discretion in FOIA cases, which permits a court to refuse to order disclosure of information even when that information does not qualify as one of the nine FOIA exemptions.[55] Such a narrow reading of the exemptions to the FOIA generally has resulted in the release of unidentifiable records for research purposes.

The Trade Secret Exemption

The exemption most frequently invoked to thwart disclosure of unidentifiable research data is the trade secret exemption of the FOIA, exemption 4.[56] However, this effort has met with limited success. The exemption extends only to "trade secrets and to information which is commercial or financial, obtained from a person, and privileged or confidential."[57] Anonymous research data are not customarily considered to be a business "trade secret"[58] and thus the data must be protected under the second part of the exemption, which has three requirements.

The first requirement, that the information be commercial or financial, has been narrowly defined.[59] For example, information has been held to be commercial or financial when it contained "knowledge of production, overhead and operating costs, levels of profit, sales and pricing data, as well as other factors."[60] Anonymous research records do not generally meet this criterion. Documents concerning the evaluation of federally funded medical services were held not to be commercial information, since they were not "data concerning fees, payment schedules, or other commercial arrangements. Furthermore, [the] studies contain no information about secret formulas or rare treatment methods; their object is the review of prevalent medical services, not esoteric experiments."[61]

Efforts to characterize the interests of researchers as commercial interests worthy of protection against disclosure have been unsuccessful. *Washington Research Project, Inc.* v. *Dept. of HEW*[62] involved a request under the FOIA for information concerning 11 research projects being funded by the National Institute of Mental Health (NIMH). The government agency argued that since the research designs had been submitted with the expectation of confidentiality and since researchers' ideas are their "stock-in-trade," such information should be considered trade secrets or commercial or financial information.[63] The federal appellate court held that the initial grant applications, as well as any continuation, renewal, or supplemental applications (both approved and pending), were not exempt from disclosure.[64] The court rejected the agency's stock-in-trade argument, holding that the reach of exemption 4 "is not necessarily coextensive with the existence of competition in any form."[65] Furthermore, the court stated that:

> It is clear enough that a noncommercial scientist's research design is not literally a trade secret or item of commercial information, for it defies common sense to pretend that the scientist is engaged in trade or commerce. This is not to say that the scientist may not have a preference for or an interest in nondisclosure of his research design, but only that it is not a trade or commercial interest…. We cannot, consistently with the Act's recognized mandate to construe exemptions narrowly, … extend them by analogies that lead so far away from the plain meaning of Exemption 4.[66]

Similarly, in *St. Paul's Benevolent Educational and Missionary Institute* v. *United States*[67] a group of privately funded researchers failed in their attempt to prevent disclosure of their data by the Centers for Disease Control (CDC). The researchers had conducted a survey concerning the feeding of infants in low-income families in the United States. The CDC was not involved in the survey or the analysis, but it assisted the researchers in release copies to others after they had published their results. Two companies that produce infant formula requested the CDC data prior to the researcher's publication, and the court ordered the CDC to disclose the information. The court agreed with an administrative hearing officer's findings that the trade secret exemption did not apply, since:

> The information in the requested materials is not confidential, commercial or financial information. [The researchers do] not argue that it is privileged. The information is certainly not financial, and [the researchers are] not engaged in any commercial enterprise.[68]

Even if research information is found to be commercial or financial, in order to be withheld under exemption 4 it must meet two additional criteria: the information must have been obtained from a person, and it must be privileged or confidential. The requirement that information be obtained from a person simply means that the agency must have obtained the commercial or financial information from a private source rather than from a government source.[69] Thus, a research contractor or grantee would qualify as a "person" under exemption 4.

The requirement that information be privileged or confidential was addressed in *National Park and Conservation Association* v. *Morton,*[70] where the court held that:

> [a] commercial or financial matter is "confidential" for purposes of the exemption if disclosure is likely to have either of the following effects: (1) to impair the government's ability to obtain necessary information in the future; or (2) to cause substantial harm to the competitive positive of the person from whom the information was obtained.[71]

Though one need only demonstrate the likelihood of substantial competitive harm,[72] it will still be a difficult burden for one who wishes to thwart the disclosure

of anonymous research data. Nonetheless, the substantial harm clause was successfully used in conjunction with the trade secret clause to prevent the release of Food and Drug Administration (FDA) data concerning manufacturers' clinical testing of artificial optical lenses in *Public Citizen Health Research Group* v. *Food and Drug Administration.*[73] There the court noted that since the manufacturers were required to submit the information to the government, the agency would not be harmed by the disclosure, but that, since the manufacturers "would sustain substantial competitive injury because their competitors would be receiving, free of charge, the benefits of this costly research and testing," the data was considered "confidential commercial information exempt from disclosure under FOIA exemption 4."

Sections of anonymous research data have also been withheld because their release might "impair the government's ability to obtain necessary information in the future." For example, in *Orion Research Inc.* v. *Environmental Protection Agency*[74] the court held that a technical proposal by a private bidder relating to the development of a monitoring system qualified as a trade secret exemption, partly because disclosure could result in a chilling effect on submission of proposals to the agency. More interestingly, in *Consumers Union* v. *Veterans Administration*[75] the court found that certain data concerning the testing of hearing aids did not qualify for nondisclosure under exemption 4, yet, by applying its "equity jurisdiction"[76] the court still decided to withhold some of the information since it might mislead the public and result in the government's receiving a more limited selection of hearing aids which, in turn, would curtail its research program.[77]

Thus, it appears that the trade secret exemption to the Freedom of Information Act will not restrict the release of agency research information when that data is not identifiable unless such release might "impair the government's ability to obtain necessary information in the future" or substantially harm a business's competitive position. In the few instances in which the courts have limited access under these interpretations, there were circumstances that are unlikely to be present when most researchers seek access to agency records.[78] If the release will not harm a business or jeopardize the government's ability to obtain information, the requested information will be released without consideration of the proprietary interest of the persons who developed the information. The same standard of disclosure will permit release of research proposals and data submitted as interim reports from ongoing research.

Other Issues

Though not dealing specifically with the trade secret exemption, three other issues have arisen in cases dealing with the release of anonymous research data. First, in order to be considered anonymous under the Act, records must have all names and identifiable numbers removed, as well as any information that might allow indirect identification.[79] Second, the removal of identifiers from a set of records does not create a new set of records.[80] Thus, an agency cannot avoid a request for anonymous records by claiming that it maintains the records only in identifiable form.[81]

Third, the courts may require raw data as well as statistical summary data to be released. In *Long* v. *I.R.S.*,[82] a party requested data that the Internal Revenue Service had compiled in a series of national studies measuring the level of compliance with federal tax law. The Internal Revenue Service released the statistical summary data but would not release the raw data on which the analyses were conducted. The Long court ordered the raw data released, stating that:

> [t]he [district] court reasoned that what was really important were the statistical tabulations previously disclosed, not the raw data, because it was only from the statistical summary that the effectiveness of the IRS could be evaluated. This conclusion is valid only if we assume that the IRS statistics encompass every useful analytic conclusion that could be drawn from the information. We find no evidence in the record to support that proposition.[83]

The Role of the Privacy Act in Regulating Disclosure of Identifiable Records Maintained by Federal Agencies

The Freedom of Information Act provides researchers with a mechanism to obtain access to anonymous federal records even if the federal agency is reluctant to release them. But some research purposes require identifiable records (see Boruch and Cecil, 1979). This section addresses the general restrictions on the ability of agencies to release identifiable information even if the agency is willing to release such information. Most of these restrictions are found in the Privacy Act of 1974.[84] Judicial interpretation of the Privacy Act has been slow to develop, and no case directly addresses the problems of access to statistical and research record systems.[85] In addition, published commentary from the research community concerning implications of the Privacy Act and other restrictive legislation has been limited.[86] With no case law and only limited commentary on this problem, observations and conclusions concerning the impact of the Privacy Act on research records must necessarily be speculative.

Overview of the Privacy Act

The Privacy Act of 1974 is the first attempt by Congress to provide comprehensive protection of an individual's right to privacy by regulating the collection, management, and disclosure of personal information maintained by governmental agencies. The Act regulates over 6,500 federal record systems, including both administrative and research record systems.[87] Before the Privacy Act was passed, federal policy toward data management practices encouraged data sharing among agencies in order to reduce the burden and expense of reporting.[88] This open-access policy was restricted only when statutes provided for the confidentiality of specific sensitive record systems.[89] The Privacy Act reversed this general policy by recognizing the right of individuals to control dissemination of information provided about themselves to federal agencies. The Privacy Act seeks to strike a sensitive balance, preserving individuals' interests in controlling identifiable information while recognizing the legitimate uses of that information.

In general, research and statistical uses of identifiable information receive no special recognition under the Privacy Act. However, the act does make a distinction in the definition of administrative records, the primary concern of the legislation, and the definition of statistical records. The term "record" is defined as "any item, collection or grouping of information about an individual that is maintained by an agency, … and that contains his name, or identifying number, symbol, or other identifying particular assigned to the individual, such as a finger or voice print or photograph."[90] This general definition is then narrowed by the subsequent definition of the term "statistical record," defined as:

> A record in a system of records maintained for statistical or reporting purposes only, and not used in whole or in part in making any determination about an identifiable individual, except as provided by Section 8 of Title 13 [authorizing certain research activities by the Bureau of the Census].[91]

Thus, statistical records are distinguished from administrative records in terms of the uses of the information and the consequences to the individual supplying the information. Recognition of this distinction suggests congressional awareness of the utility of such research record systems. However, the Privacy Act imposes the same general scheme of regulation on statistical and administrative records that are identifiable.

Briefly, the Privacy Act of 1974 requires that federal agencies must (1) grant access by individuals to their identifiable records maintained by federal agencies; (2) ensure that existing information is both accurate and timely, and limit the collection of unnecessary information; and (3) limit the disclosure of identifiable information to third parties. This third provision of the Privacy Act, forbidding the disclosure of any identifiable record without the

prior written consent of the individual,[92] is most relevant to researchers' access to federal data. This prohibition is also the crux of the right of privacy provided by the act, since an enforceable consent requirement could thwart the disclosure of identifiable information for purposes that the individual never considered and would not approve.

In recognition of legitimate needs for identifiable information, the Privacy Act carves out 11 categories of exceptions to the consent requirement. For instance, an agency may, at its discretion, disclose records without prior written consent to officers and employees of the agency who have a need for the record in the performance of their duties.[93] Other exemptions include disclosures that are required by the Freedom of Information Act; to the Bureau of the Census for planning or carrying out a census, survey, or related activity under Title 13; to the General Accounting Office to permit auditing of federal programs; and in emergency circumstances involving the health and safety of any individual. In addition, the act permits disclosure without written consent to other federal agencies for authorized civil or criminal law enforcement activities,[94] and, pursuant to a court order, disclosures to which individuals would most likely decline to consent. Perhaps it was this same concern that led Congress to include an exemption for itself.[95]

Of special interest to researchers is an exemption that permits disclosure "to a recipient who has provided the agency with advance adequate written assurance that the record is to be transferred in a form that is not individually identifiable" (§552a(b)(5)). While the practical benefits of such an exemption may be questioned (see discussion below), the exemption indicates an attempt by Congress to accommodate the need for access to agency records for research purposes. Similarly, agency records of historical interest may be transferred to the National Archives without obtaining consent (§552a(b)(6)).

Fearing that it had failed to provide for all of the legitimate needs for identifiable information that merit an exclusion, Congress also included a "safety-valve" exemption, permitting disclosure without consent for a "routine use" of the record (§552a(b)(3)). A "routine use" is "for a purpose that is compatible with the purpose for which it was collected" (§552a(a)(7)). Instead of obtaining individual consent prior to disclosure for such a use, the agency must only publish a notice of the anticipated routine uses of the record in the *Federal Register* and accept comments from the public for a period of 30 days (§552a(e)(4)(D), (e)(5)(11)).[96]

Two further points regarding disclosure should be noted. First, the requirement of prior written consent of an individual may be avoided by inserting broad waiver provisions in the original request for information. If a person signs such a waiver, identifiable information may be released for purposes consistent with the waiver.[97] Finally, the Privacy Act places no obligation on the recipients of information to maintain the confidentiality of the records

or limit subsequent disclosure. Once the records are released to a party not under the jurisdiction of the act, there is no assurance that the individual's rights will be protected.

Subject to the exemptions noted above, the Privacy Act prohibits disclosure by any agency of any record contained in a system of records to a person or to another agency without the written consent of the individual to whom the record pertains (§552a(b)). The extent of the regulation of social research by the act is determined by the manner in which it defines such terms as "record," "system of records," and "agency."

"Record" is defined as "any item, collection, or grouping of information about an individual that is maintained by an agency … and that contains his name, or the identifying number, symbol, or other identifying particular assigned to the individual …" (§552a(a)(4)). Since such a record can include "as little as one descriptive item about an individual …,"[98] identifiable research data can clearly qualify as a "record" for the purpose of the act. The Privacy Act extends to record systems maintained by "agencies" of the federal government:[99] clearly, systems of research and statistical records maintained by such agencies are regulated by the act; just as clearly, private data archives or record systems maintained by state or local governments without federal assistance are exempt from the act.

However, not all "records" maintained by "agencies" are regulated by the Privacy Act. Mindful of the administrative burden on agencies that would result if access were permitted to all identifiable information, Congress restricted the protection of the Act to records that are contained in a "system of records" (§552a(b)), further defined as "any group of records from which information is retrieved by the name of the individual or by some number, symbol, or other identifying particular assigned to the individual" (§552a(a)(5)). This definition encompasses a vast number of identifiable records maintained by federal agencies,[100] but it does not extend to those record systems in which the information is not actually retrieved by individual identifier.

Restrictions on Research Access to Agency Record Systems by the Privacy Act of 1974

The Privacy Act's prohibition on disclosure of identifiable information without the prior written consent of the individual can sharply restrict the use of identifiable federal records for research purposes. Researchers usually seek access to agency record systems either to obtain a sample of individuals for anticipated research or to supplement existing research information.[101] The consent requirement can interfere with both of these activities.

When seeking to supplement existing research data with information from agency records, in an ideal situation researchers would be able to anticipate

this need and obtain the informed consent of research participants at the time the information is gathered. But even if a research participant agrees to subsequent release of agency information, the consent may be invalid if the researcher seeks access to records in a system that did not exist at the time consent was obtained.[102] And when the need for research access to agency records was not anticipated or when the initial consent becomes invalid, a researcher may have to recontact the participants to obtain proper consent. Recontacting a participant in an earlier research study imposes special difficulties. Some target populations are highly mobile, so addresses and telephone numbers obtained at the initial encounter may be outdated. Some target populations are difficult to recruit for research, so subsequent attempts to obtain consent to release agency information will likely be expensive and subject to self-selection biases.

Even more formidable obstacles are faced by researchers who seek access to agency records to generate a sample of identified individuals to be contacted for participation in anticipated research. Since the purpose is to obtain a list of names and addresses of individuals, the researcher will be unable to contact the individuals to obtain consent for release of this information. Researchers employed by the agency maintaining the records may avoid such consent requirements by demonstrating a need for the record in the performance of their duties.[103] But some researchers outside the agency have found the consent requirement a frustrating hurdle.[104] At hearings of the Privacy Protection Study Commission, a number of researchers who rely on file linkage to conduct longitudinal research were sharply critical of the potential for disruption of their research by the restrictions of the Privacy Act,[105] although some of these problems have apparently been avoided by designation of research as a "routine use" of many of the most important systems (Bebee, 1981:661,666).[106] However, without such a designation, the Privacy Act represents a considerable obstacle to researchers who seek to use federal records to identify or locate persons they wish to include in their sample of research subjects.

Several of the exemptions to the disclosure requirements in the Privacy Act may be of some aid to researchers.[107] The most obvious example is the exemption that permits agencies to disclose information for purposes of statistical research if the record is transferred in a form that is not individually identifiable (§552a(b)(5)). In fact, this exemption offers very little: a record that is not individually identifiable is not a "record" within the definition of the Privacy Act (§552a(a)(4)) and therefore is not subject to the restrictions on disclosure imposed by the act.[108] Nevertheless, the Privacy Act will not thwart requests for anonymous research records. Such records may be useful for a number of research purposes. Statistical and administrative procedures have been developed that permit meaningful statistical analysis of data while

preserving the anonymity of the respondents.[109] For example, techniques of microaggregation permit statistical analysis of information on identified groups of individuals (Campbell et al., 1975). Release of such aggregated information would seem to be permitted under the act: since the information is not individually identifiable, disclosure is not restricted. A mutually insulated file linkage technique may even permit agencies to share large archives of aggregated data without violating the act (Campbell et al., 1975).

Merely removing the name or individual identification number may be sufficient to permit disclosure under the exemption for statistical research. The exemption states that disclosure is permitted only if the record is "in a form which is not individually identifiable" (§552a(b)(5)). The guidelines for implementation of the Privacy Act interpret this phrase to mean:

> Not only that the information disclosed or transferred must be stripped of individual identifiers, but also that the identify of the individual cannot be reasonably deduced by anyone from tabulations or other presentations of the information (i.e., the identity of the individual cannot be determined or deduced by combining various statistical records or by reference to public records or other available sources of information).[110]

This guideline implies that where the research population is small and some of the variables are also recorded with names on publicly available lists, precautions beyond the deletion of identifiers must be taken to guard against public disclosure.[111] For example, the data may be "inoculated" with random error or the reporting categories may be structured so that they do not correspond to categories available in public reports (Riecken and Boruch, 1974).[112] Unless such methods are used, the identities of the individuals may be inadvertently disclosed, in violation of the act. Many research needs cannot be satisfied by anonymous data (Boruch and Cecil, 1979). Researchers may attempt to obtain identifiable agency records by tailoring their requests to fit the exemptions of the Privacy Act. One of the most important exemptions (§552a(b)(2)) permits disclosure of agency records required to be disclosed under the Freedom of Information Act. Since the Freedom of Information Act *requires* an agency to disclose records—unlike the Privacy Act, which simply *permits* an agency to disclose if it chooses to do so—this exemption is most useful to researchers when an agency resists disclosure. The interrelationship between the Privacy Act and the Freedom of Information Act is quite complicated (see below).

Another exemption that may assist researchers seeking identifiable information permits the transfer of identifiable agency records to the Bureau of the Census for planning or carrying out a census, survey, or other related activity (§552a(b)(4)). The law also permits linkages between agency files when conducted by the Bureau of the Census for some purpose, such as to establish the credibility of these alternative sources of information. Furthermore, this

exemption allows the Census Bureau to accept identifiable agency records to perform certain statistical analyses for researchers outside the Census Bureau who are unable to gain access to these records.[113] Since such analyses appear to be a "related activity" under the exemption, the Bureau has been acting as a broker for such research (Office of Federal Statistical Policy and Standards, 1980b; Commission on Federal Paperwork, 1977).

The final exemption that might benefit researchers permits disclosure of an identifiable record for a "routine use" of such a record (§552a(b)(3)). "Routine use" is defined as a use "for a purpose which is compatible with the purpose for which [the record] was collected" (§552a(a)(7)).[114] Such ambiguity in statutory language suggests that an agency may define "statistical analysis" as a routine use of all or a selected portion of agency record systems, permitting researchers outside the agency to have access to identifiable records without gaining the consent of the individuals to whom the records pertain. In fact, a great many agency notices allow for disclosure involving statistical research programs as a routine use. The Department of Health and Human Services has been particularly thorough in identifying record systems that have research potential and publishing notices permitting research as a routine use (see, generally, O'Neill and Fanning, 1976). One version of the routine-use notice requires an assessment of the risks and potential benefits of the research and requires the recipient to sign an agreement to protect the records from subsequent disclosure.[115] This is one instance in which the discretion delegated to agencies by the Privacy Act has been used to fashion a specific set of standards to permit data sharing while maintaining the proper safeguards. However, the need to rely on the routine-use exemption to overcome the failure of the statute to provide for research and statistical access to identifiable records is an awkward solution to the problem. Without a statutory policy concerning research access to federal records, individual agencies are free to develop inconsistent regulations that may either be too restrictive or fail to offer adequate protection to the identified individuals.[116]

In summary, the Privacy Act's failure to distinguish research and statistical purposes from administrative purposes in restricting access to records may pose a major obstacle for researchers who seek identifiable information, especially when a list of names and addresses is needed to develop a sampling frame for research. In general, a researcher must structure a request for access to identifiable information to fit within one of the exceptions to the consent requirement of the Privacy Act, such as the routine-use exemption.

Request for Identifiable Information for Research Purposes When an Agency is Unwilling to Disclose

As mentioned above, both the Freedom of Information Act (FOIA) and the Privacy Act of 1974 deal with the release of identifiable records by the federal

government. However, the two statutes have opposite purposes. The FOIA is designed to encourage disclosure of agency records, while the Privacy Act is designed to protect the privacy of any individual who is the subject of a government record. Prior to the implementation of the Privacy Act, researchers could use the FOIA to obtain identifiable information within the constraints of the exceptions to disclosure. The impact of the Privacy Act on disclosure practices under the FOIA remains a matter of speculation and dispute.

A third party seeking an identifiable record without permission of the identified individual must base the claim for access either on the FOIA or on one of the Privacy Act's exemptions to the consent requirement. As noted above, the Privacy Act lists 11 exceptions to the consent requirement of prior consent, one of which is disclosure under the FOIA (5 U.S.C. &552a(b)(2)). This exception to the consent requirement for disclosure is the crucial exemption when a third party requests identifiable records and the agency resists disclosure (as discussed above). Most commentators contend this exemption indicates a congressional intention to exempt from the restrictions of the Privacy Act all personal information not constituting a "clearly unwarranted invasion of personal privacy" under the standards of exemption 6 of the FOIA as they existed prior to the implementation of the Privacy Act.[117] In other words, this interpretation suggests that the Privacy Act was not intended to change existing practices of disclosure of identifiable information under the Freedom of Information Act. Other commentators believe that the adoption of the Privacy Act indicates an additional interest by Congress in the privacy of individuals that should be considered by the courts, resulting in greater restrictions on disclosure under the FOIA (see, e.g., Arnold and Kissiloff, 1976, and M. Hulett, 1975). Unfortunately, case law has not addressed this conflict. However, under either of these interpretations release of identifiable information will have to meet at least the standards of the Freedom of Information Act.

The major exemption used to prohibit disclosure of identifiable information under the FOIA is exemption 6, which applies to "personnel and medical and similar files the disclosure of which would constitute a clearly unwarranted invasion of privacy."[118] While there have been many court cases concerning this exemption,[119] including two Supreme Court cases,[120] it is still not clear what constitutes a "clearly unwarranted invasion of privacy." Nonetheless, the Supreme Court has suggested that exemption 6 permits the withholding of information only when two requirements have been met:

> [F]irst, the information must be contained in personnel, medical or "similar" files, and second, the information must be of such a nature that its disclosure would constitute a clearly unwarranted invasion of personal privacy.[121]

The first requirement, that a record be classified as a "personnel," "medical," or "similar" record, was given a fairly broad definition in *United States Department of State* v. *The Washington Post Co.,*[122] in which the Supreme Court held that records containing passport and citizenship information qualified as "similar" files. After a review of the legislative history of exemption 6, the Court concluded that the term "similar" files was not limited to files that contain "intimate details" and "highly personal" information.[123] Instead, the Court looked to congressional statements that exemption 6 was a "general exemption" protecting information in "great quantities of files," and the Court concluded that:

> We do not think that Congress meant to limit Exemption 6 to a narrow class of files containing only a discrete kind of personal information. Rather, '[t]he exemption [was] intended to cover detailed Government records on an individual which can be identified as applying to that individual….' When disclosure of information that applies to a particular individual is sought from government records, courts must determine whether release of the information would constitute a clearly unwarranted invasion of that person's privacy.[124]

Thus, the Court quickly reached the second part of the test.

While the Supreme Court in the *Washington Post* case did not decide the issue of the clearly unwarranted invasion of privacy, the Court did in *Department of Air Force* v. *Rose*[125] examine both the probability and consequences of identification before deciding to release anonymous case summaries of honor code hearings. The Court pointed out that the exemption did "not protect against disclosure every incidental invasion of privacy—only such disclosures as constitute 'clearly unwarranted' invasions of personal privacy,"[126] and that the exemption required a balancing of the individual's right to privacy against the public's right to open government.

In conducting this balancing test, the lower courts have first looked for the possible privacy interests that might be invaded. For example, in *Washington Research Project, Inc.* v. *Department of HEW*[127] the district court pointed out that "the identity of an institutional applicant [may not] be concealed because the right to privacy envisioned in the [FOIA] is personal and cannot be claimed by a corporation or association."[128] Similarly, in *Public Citizen Health Group* v. *Department of HEW*[129] a group sought access to agency records that evaluated federally funded medical services, including information from hospital profiles, patient records, and physician profiles. The district court held that the hospitals did not have a cognizable privacy interest but that the patients and physicians did.[130]

Having found a privacy interest, that interest must then be balanced against the degree of public interest served by the disclosure. The party seeking the information must be able to point to a public interest being served by the court would not release personnel information to a group of employees, stating that

"[t]he disclosure of personnel records in the instant case would be a serious invasion of privacy" since the information sought was "personal and capable of causing embarrassment" while "practically no public interest is advanced by disclosure."[132]

Also, it is not enough for the party seeking the information to have another private interest to be balanced against the subjects' privacy. A public interest is necessary to justify the disclosure. Thus, researchers were granted access to a list of employees eligible to vote in a union election so that a study of the effectiveness of National Labor Relations Board (NLRB) election regulations could be conducted.[133] Lists of names and addresses were also disclosed when the requester wished to lobby for those persons.[134] The court ordered that an agency preserve certain records to avoid mooting an FOIA request when the plaintiff sought to identify parties for an antitrust class action.[135] However, disclosure has not been required when the court perceives a private rather than public interest furthered by disclosure. In one case an employer was denied access to union cards maintained by the NLRB since the cards were to be used by the employer to attack the validity of union registration, furthering what the court determined to be a private rather than a public interest.[136] Similarly, a business was refused access to a list of names and addresses it sought for purposes of sending out store catalogues and announcements, with the court again discerning little public interest.[137]

Several commentators have criticized a court's emphasis on the requester's purpose in seeking disclosure (Kronman, 1980; Easterbrook, 1980), especially since once the information is released to someone outside the federal government it is no longer protected under the Privacy Act or the FOIA. Thus (Easterbrook, 1980:775,781):

> A requester might justify his inquiry on the ground of scientific research and then sell the information to someone else who will use it for different purposes, because it seems clear that anyone who obtains information under the FOIA can broadcast it as he please.

Finally, just as with exemption 4 cases, agencies have argued that disclosure under exemption 6 would "impair the government's ability to obtain necessary information in the future." For example, in *Public Citizen Health* v. *HEW,*[138] HEW argued that disclosure would result in diminished participation by the medical profession in the evaluation process and in the medicare and medicaid programs. The court found that such considerations were valid, but that in this case they were simply broad speculations without supporting evidence.[139] In addition, as one commentator argues (Easterbrook, 1980:797–799), the impact of disclosure on an agency is a major consideration in the agency's decision of whether to oppose release of the information in the first place. In fact, most FOIA litigation involves only the agency that

possesses the information and the requesting party. The individual whose records are being sought is rarely allowed to participate in the litigation, if the person is informed of it at all.

In summary, the confusion regarding research access to identifiable data under the Privacy Act becomes even greater when identifiable information is sought under the Freedom of Information Act. Although the courts have ordered the release of some identifiable records under the FOIA, the relationship between the FOIA and the Privacy Act continues to be controversial. Of course, if disclosure of records is permitted, the proprietary interest of the primary researcher in the data will not be considered. The extent of federal regulation of agency records leaves little opportunity for professional standards for data sharing to influence the process. While waiting for case law and new legislation to address these issues,[140] researchers must do their best to weave their way through the conflicting standards.

ACCESS TO RESEARCH RECORDS DEVELOPED WITH PUBLIC FUNDS THAT ARE MAINTAINED BY PRIVATE RESEARCHERS

The first two sections of this paper addressed circumstances in which there is little confusion over the private or public character of the research information. However, the private or public status of a data set can be difficult to determine when research data sets are developed through public funding of private researchers. This is a common circumstance. During fiscal 1979, for example, the federal government spent almost $4 billion to fund research by universities, usually by research grants or contracts (National Commission on Research, 1980). The question is whether public funding of such data permits access through the laws regulating the federal data archives or the private status of the researcher removes these data sets from federal regulation.[141]

Scope of Federal Regulation of Research

As noted above, the two federal laws most relevant to general access to federal data archives are the Freedom of Information Act and the Privacy Act of 1974; they extend only to agency records and define "agency" as:

> Any executive department, military department, Government corporation, Government controlled corporation, or other establishment in the executive branch of the government … or any independent regulatory agency.[142]

The Freedom of Information Act authorizes the federal district courts to enjoin an "agency from withholding agency records and to order the production of any agency records improperly withheld from the complainant."[143] Recent

interpretations of the term "agency records" have been rather restrictive and not likely to aid researchers who seek access to data sets maintained by private researchers but developed with public funds through either contracts or grants.

Two recent Supreme Court cases in which scholars sought records not maintained by federal agencies suggest that the FOIA and the Privacy Act will not be an effective means of obtaining access to data sets developed through research grants. The most widely known case is *Forsham* v. *Harris,*[144] in which researchers sought access to data developed under an extended research grant. In that case a federal agency, the National Institute of Arthritis, Metabolism, and Digestive Diseases (the institute), awarded a series of research grants to the University Group Diabetes Program (the university group), a group of private physicians and scientists conducting a long-term study of the effectiveness of certain treatments of diabetes. The study was funded solely by the federal government at a cost of approximately $15 million. The study generated more than 55 million records documenting the treatment of over 1,000 diabetic patients who were monitored for 5–8 years.

As with most such research grants, the institute exercised some supervision over the research and had a right of access to the raw data, even a right to obtain permanent custody of the raw data generated by the university group. However, the day-to-day administration of the research was by the university group. The institute did not exercise its right to review or obtain custody of the raw data, which remained at all times in the possession and control of the university group.

Ultimately, the university group's reports on the results of its study indicated that the use of certain drugs for the treatment of diabetes resulted in an increased risk of death from cardiovascular disease compared with treatment by the other methods studied. These findings then led the Food and Drug Administration (FDA) and the secretary of the Department of Health, Education, and Welfare to adopt regulations to control the labeling and use of those drugs.

When the university group began releasing its preliminary findings in 1970, its conclusions were challenged by other researchers. The Committee on the Care of the Diabetic (the committee), a national association of physicians involved in the treatment of diabetes patients, was among the most persistent critics of the university group study. The committee requested access to the raw data to facilitate its review of the university group's findings, and the university group declined its request. The committee then sought to obtain access to the research data under the Freedom of Information Act.[145]

The case was further complicated by the involvement of the federal agency in authorizing a review of the findings. Although no employees of the institute reviewed the research records of the grantee, the institute did contract in

1972 with another private grantee, the Biometric Society, for an assessment of the university group study. The Biometric Society was given direct access to the raw data by the terms of its contract with the institute. The Biometric Society issued a report to the institute in 1974 concluding that the university group results were "mixed" but "moderately strong."

The researchers seeking access to the data under the Freedom of Information Act found no friend in the courts. The Court of Appeals denied the FOIA request, concluding that records of grantees are not "agency records" since the FOIA applies only to records that have been "created or obtained … in the course of doing its work."[146] In a dissenting opinion, Judge Bazelon concluded that the university group data were "agency records" under the FOIA since the government had been "significantly involved" in the study through its funding, access to raw data, and reliance on the study in its regulatory actions.[147]

The Supreme Court, in an opinion written by Justice Rehnquist, made clear that in normal circumstances the FOIA does not extend to records developed by research grants:

> We hold here that written data generated, owned, and possessed by a privately controlled organization receiving federal study grants are not "agency records" within the meaning of the Act when copies of those data have not been obtained by a federal agency subject to the FOIA. Federal participation in the generation of the data by means of a grant from HEW does not make the private organization a federal "agency" within the terms of the Act. Nor does this federal funding in combination with a federal right of access render the data "agency records" of HEW, which *is* a federal "agency" under the terms of the Act.[148]

In reaching this conclusion the court relied heavily on the legislative history of the FOIA,[149] and found that Congress chose not to confer any direct public rights of access to such federally funded project information.[150] This interpretation suggests that federal funding of a research record system alone is not sufficient to provide a right of access to the data to other researchers.

The Committee on the Care of the Diabetic advanced a second argument, contending that while in ordinary circumstances the records of a grantee are not agency records, the reliance by the federal agencies on the research data of grantees in developing their regulations was sufficient to transform these records into "agency records" within the meaning of the statute. It argued that in this case, in which the agency was actively involved in the development of the research records, had a right of access, exercised that right of access through a second grantee, and then based its regulations on the conclusions of the research, the extent of involvement and reliance of the agency on the records of the grantee was sufficient to transform them into agency records within the meaning of the statute.[151]

The Supreme Court also rejected this argument, turning to the legislative

history of related statutes that emphasize the possession and control of records in defining an "agency record."[152] The court did suggest that in some circumstances, such as if a grant created a partnership or joint venture between an agency and a grantee, there may be a right of access to records held by a grantee,[153] but it concluded that no such relationship existed in this case. Since the court found that the records were not "agency records" within the meaning of the statute, the court affirmed the action of the Court of Appeals in dismissing the FOIA request.

In a dissenting opinion, Justice Brennan, joined by Justice Marshall, showed more sympathy to the second argument of the committee. Justice Brennan contended that in some circumstances the relationship between an agency and a grantee, and the importance of the information in developing public policy, can transform the research records of a grantee into "agency records" within the meaning of the Freedom of Information Act, therefore making them available to the public. Justice Brennan noted that the purpose of the Freedom of Information Act was to "open the processes of government to public inspection," and that "[n]othing in the legislative history suggests that Congress meant to allow agencies to insulate important steps in decision making on the basis of technical niceties of who 'owns' crucial documents."[154] Justice Brennan concluded:

> Where the nexus between the agency and the requested information is close, and where the importance of the information to public understanding of the decisions or the operation of the agency is great, I believe the congressional purposes require us to hold that the information sought is an "agency record" within the meaning of the Freedom of Information Act…. The existence of this factor can be tested by examining, *inter alia,* the degree to which the impetus for the creation of the record came from the agency or was developed independently, the degree to which the creation of the record was funded publicly or privately, the extent of governmental supervision of the creation of the record, and the extent of continuing governmental control over the record.[155]

The dissenting opinion of Justice Brennan took into account a number of the issues mentioned by Hedrick (in this volume) concerning justification for data sharing. He noted the scholarly debate that surrounded the dispute and the public importance of research data when they become involved in the regulatory process. In conclusion, he noted: "If the records of such organizations [i.e., grantees], when drawn directly into the regulatory process, are immune from public inspection, then government by secrecy must surely return." [156]

The interpretation of the majority in *Forsham* suggests that the research records of a grantee cannot be obtained through the Freedom of Information Act. The fact that the research findings are controversial and that the findings are used in establishing public policy through various agency proceedings will

not be considered in determining the right of access to research records through the FOIA. This case suggests that an agency can insulate its actions from public scrutiny by funding a grant for controversial research and then basing its action on those findings. As long as the agency does not take possession or control of the records, the FOIA will not assist those who wish to challenge the findings that underlie the agency action. Of course, if the data are filed with the agency, the FOIA will be a more effective means of obtaining disclosure.

In the companion case to *Forsham, Kissinger* v. *Reporters Committee for Freedom of the Press,*[157] the Supreme Court sounded another discouraging note for researchers who seek to use the FOIA to gain access to data not in the direct possession of federal agencies. In that case, several reporters, along with representatives of the American Historical Association and the American Political Science Association, attempted to gain access to telephone notes made by Henry Kissinger while he was Secretary of State. They sought to rely on the Freedom of Information Act. Their FOIA requests were filed after the records were wrongfully removed from the State Department, transported to a private estate, then deeded to the Library of Congress with specific restrictions concerning who should have access to them.

When the State Department denied the FOIA request, the parties turned to the federal courts, contending that because the records had been wrongfully removed from the State Department, the State Department still had access to the records and should be required to retrieve them and make them available for inspection. Both the federal district court and the court of appeals ordered production of the requested information.[158] The Supreme Court reversed. Again, Justice Rehnquist wrote for the majority, and he stated:

> We hold today that even if a document requested under the FOIA is wrongfully in the possession of a party not an "agency," the agency which received the request does not "improperly withhold" those materials by its refusal to institute a retrieval action. When an agency has demonstrated that it has not "withheld" requested records in violation of the standards established by Congress, the federal courts have no authority to order production of such records under the FOIA.[159]

The majority in *Kissinger* reasoned that an agency cannot withhold records, under the meaning of the Freedom of Information Act, if it does not have possession of those records.[160] To the plaintiffs' contention that the wrongful removal of the records from the control of the agency should not defeat their right to obtain access, the court responded that Congress has established a scheme for management and disposal of records under the Federal Records Act of 1950[161] and the Records Disposal Act,[162] and that scheme did not contemplate a private right of action for wrongful removal.[163] After another extensive examination of legislative history, the court concluded that "Congress did not mean that an agency improperly withholds a document which has been

removed from the possession of the agency prior to the filing of the FOIA request."[164] In emphasizing the timing of the FOIA request, the Court suggested that it might permit access to records if it is shown that an agency purposefully routed a document out of its possession in order to circumvent a FOIA request, but it did not decide this issue. [165]

Justice Brennan, again writing in dissent, took exception to what he described as, "the Court's crabbed interpretation of 'improper withholding' under the Freedom of Information Act." Again he cited the public access purpose of the Freedom of Information Act, and thought it "plainly unacceptable for an agency to devise a records routing system aimed at frustrating FOIA requests in general by moving documents outside agency custody with unseemly haste."[166]

Justice Stevens, also writing in dissent, emphasized the wrongful removal of the documents from the agency, and noted that the decision of the majority, "… creates an incentive for outgoing agency officials to remove potentially embarrassing documents from their files in order to frustrate future FOIA requests." He interpreted the Freedom of Information Act to modify the congressional scheme expressed by the Federal Records Act of 1950 and the Records Disposal Act and to require an agency to produce requested documents if it retains a legal right to the custody of those documents wrongfully removed from its files.[167]

The majority opinion in *Kissinger* suggests that researchers' efforts to obtain research data in possession of an agency will be unsuccessful if the agency removes those documents from its custody prior to the filing of the FOIA request—even if the documents are wrongfully removed from agency custody. However, if the documents are removed after the filing of the FOIA request and the persons requesting access can show that the purpose of the removal was to thwart access, the court has implied that access may be permitted.

The *Forsham* and *Kissinger* cases indicate the importance of the timing of the FOIA request: it must be filed while the agency has custody of the data. If it is filed before custody passes from the grantee to the agency, *Forsham* indicates that there will be no right of access; if it is filed after custody passes from the agency, even if the documents are wrongfully removed, *Kissinger* indicates there will be no right of access. Together these cases provide a convenient means of permitting agencies to thwart access to research records while still basing their regulatory actions on controversial findings.

An alternative way of obtaining data may be available to some parties affected by an agency's regulatory action, even when those data are not accessible through the FOIA or the Privacy Act. In *Dow Chemical Co.* v. *Allen,*[168] the Environmental Protection Agency suspended the use of two herbicides manufactured by Dow and scheduled cancellation hearings. The emergency

suspension was ordered as a result of certain animal toxicity studies conducted at the University of Wisconsin. The researchers had voluntarily turned over the data from the study that the agency had relied on in ordering the suspension, but they refused to turn over the data from other, similar uncompleted studies. Since the data remained at Wisconsin and were not in the agency's possession, under the *Forsham* decision the data were not accessible through the FOIA. However, under section 6(d) of the Federal Insecticide, Fungicide, and Rodenticide Act,[169] an administrative law judge can issue a subpoena for documents relevant to cancellation hearings. Dow subpoenaed all of the data from the ongoing Wisconsin studies, and Dow pressed its case on appeal even though the researchers did not plan to testify at the cancellation hearings and the findings of the disputed studies were not to be introduced as evidence in the hearings. The Court of Appeals refused to enforce the subpoena, holding that it would be an unreasonable burden on the researchers to require them disclose this additional information since the incompleted studies were of little probative value and since the risk of inadvertent, premature disclosure outweighed the need for the information. Furthermore, the court noted that requiring such a disclosure would unreasonably restrict the researchers' first amendment interest in academic freedom.[170] Such a factor has not been considered by any other courts in FOIA or Privacy Act cases.

The *Forsham* case considered access to records developed through a research grant. The issue with regard to research contracts is not so clear. Typically, agencies are more involved in the conduct of research funded through contracts. The Privacy Act indicates that its regulations apply when "an agency provides by contract for the operation by or in behalf of the agency of a system of records to accomplish an agency function…."[171] Since a contract to conduct research would appear to be a contract to "accomplish an agency function," the record systems necessary to conduct research supported by federal contracts would seem to fall within the terms of the Privacy Act.

However, for some time after the passage of the Privacy Act and the development of clarifying guidelines,[172] some agencies contended that contractors performing statistical surveys were not subject to the act, even though it may be necessary to establish a system of records to perform the contract.[173] Apparently this situation is still somewhat confused. The Commission on Federal Paperwork (1977) suggested that federal contractors and grantees should be required to comply with the Privacy Act (see also Privacy Protection Study Commission, 1977, and Office of Federal Statistical Policy and Standards, 1980b), and apparently there has been increasing compliance by contractors. However, to the extent that federal contractors and grantees avoid complying with the Privacy Act, federal agencies will be able to avoid responsibility for record keeping practices, even though the records come into existence only because of federal funding. As long as the research data are

not filed with the agency, the records remain beyond the reach of the Freedom of Information Act and may be disclosed at the discretion of the primary researchers without regard to the Privacy Act's restrictions on disclosure. Professional standards for data sharing can play an important role in directing disclosure practices of research contractors and grantees.

SUMMARY AND CONCLUSIONS

Legal standards for access to research data vary with the status of the person or institution possessing the data. This paper examined access to data in three circumstances: (1) a data set developed by a private researcher with private funding, (2) a data set developed by a federal agency, and (3) a data set developed by a private researcher with federal funding.

The first circumstance, in which the primary researcher both develops and maintains the data set with private resources, is the simplest case. Since possession of the data by the primary researcher is not compromised by federal funding, he or she may exercise great discretion in restricting access. Control over data allows primary researchers to adequately protect their own interests. From this perspective, the available legal protection can offer few advantages over simple secrecy. And although primary researchers may choose to copyright the data set, such protection may be ineffective in restricting disclosure to those persons they choose. Once data are published, the fair-use provisions of the copyright laws will permit broad use of the data by others for scholarly purposes.

The absence of effective legal standards in this circumstance suggests an important role for professional guidelines. Such guidelines can recognize the rights of primary researchers to be the first to publish their findings, while encouraging data sharing once those proprietary interests have been realized. Such guidelines can be tailored to individual professional associations with specific needs and customs concerning data access. Professional guidelines may also be useful in encouraging publishers to adopt less restrictive practices in permitting republication of previously published data sets. Standards for data access established by scholarly journals affiliated with professional associations may be able to lead the way in reforming overly restrictive publishers' practices.

The second circumstance, in which data sets are developed and maintained by federal agencies, is the most heavily regulated area in which data sharing occurs. This circumstance may be of particular interest since it includes many data sets that are relevant to the study of public policy issues. Since federal laws and regulations were drafted to correct abuses of administrative records, those standards do not address the needs of the research community, permitting overly broad disclosure in some instances and unnecessarily restricting disclosure in others.

The legal standards governing access to records possessed by federal agencies turn on whether or not identifiable records are requested. If unidentifiable records are sought, the Freedom of Information Act usually requires release of the information. The broad disclosure provisions of the FOIA can be helpful in obtaining a range of unidentified federal agency records, even if the agency resists disclosure. Unfortunately, these same broad disclosure provisions may be used to obtain information that trespasses on the proprietary rights of those who submit research proposals to federal agencies. In addition to research data from completed studies, such proprietary information includes literature reviews, preliminary theoretical analyses, and even clinical trial data from ongoing research. In general, if the information is not identifiable and is maintained by a federal agency, the FOIA requires the agency to disclose the information, whether or not such disclosure interferes with the proprietary rights of the person who gathered the information.

In instances in which identifiable records are sought for research purposes, federal policy is too restrictive. The Privacy Act of 1974 places sharp limitations on release of identifiable records, permitting no exception for research. The resulting restrictions on access to identifiable federal records can be a persistent problem faced by researchers who need to develop sampling frames. Though the Privacy Act offers a few exemptions that may permit release of identifiable records for research (such as the "routine-use" exemption), these exemptions are exercised at the discretion of agencies. Without agency cooperation, and sometimes even with it, researchers will face a difficult time obtaining access to identifiable federal records.

When an agency does not wish to provide access to identifiable federal records, a researcher may attempt to obtain the records under the Freedom of Information Act. In cases in which the courts have ordered disclosure of identifiable records, an awkward balance has been struck between the conflicting goals of the Freedom of Information Act and the Privacy Act. Legal standards in this area remain unclear. However, a researcher who seeks identifiable records without the support and assistance of the federal agency that has them should anticipate a long and frustrating struggle.

While federal regulation of agency records leaves little opportunity for professional guidelines, there is a clear need for researchers to contribute to restructuring the legal standards in this area. Though the Freedom of Information Act is quite helpful in obtaining a wide range of anonymous federal records, it also permits access to research proposals and other information that may trespass on the proprietary rights of those who are seeking federal funding for their research, even if no such funding is granted. Disclosure of anticipated research plans and data from research in progress seems to go beyond appropriate sharing of information. While arguing for a more restrictive disclosure for research proposals under the Freedom of Information Act, researchers should also seek broader disclosure exemptions for research purposes

under the Privacy Act. This legislation was intended to correct administrative abuses of identifiable records; no instance of research abuse of identifiable records was cited. Yet the same restrictive interpretations are extended to requests for identifiable data for research purposes.

The third circumstance involving data sets, those developed through federal funding of private researchers, presents the most complicated area of regulation. In this circumstance the proprietary rights of the private researcher are compromised only by the source of funding for the research, since possession of the data is retained by the private researcher. Recent Supreme Court decisions suggest that unless the research records are directly maintained by the federal agency, the FOIA will be an ineffective tool in obtaining access. Though the precedents are confusing and regulations vary from agency to agency, it appears that if an agency does not take possession of the research data, the agency can fund the research, participate in the design and development of the research, permit access by third parties to the data, base regulatory findings on the conclusions of the research, and yet thwart access to the records by persons and organizations the agency does not wish to have them. Even if the agency once possessed the data and wrongly forfeited possession before access was sought, no access to the records is required. This standard invites agencies to structure their relationships with research grantees and contractors in such a way that controversial or sensitive federal research records relied on by the agencies will be beyond public scrutiny. One solution to this problem may be to require agencies to take possession of research data sets they rely on in setting policy. Once the records are in the possession of the agency, the FOIA may be used to compel disclosure.

As indicated throughout this paper, the statutes and case law relevant to access to research records have been developed without consideration of the unique aspects of research records. In some areas access is too broad; in other areas access is too restrictive. In those areas in which access is permitted but not required, professional guidelines and standards may encourage agencies to adopt more open disclosure policies. In other areas, in which statutes and regulations result in access that is either too broad (e.g., release of preliminary research data) or too restrictive (e.g., no release of identifiable records, no release of data from grantees and contractors, even if the findings are used by the agency), modifications of existing legal standards should be sought.

NOTES

1. Nelkin (1982) offers an excellent overview of the variety of disputes that involve control of research data and findings.
2. The unique character of research data is discussed in Privacy Protection Study

Commission (1977).

3. Freedom of Information Act, 5 U.S.C. §552 (1976); Privacy Act of 1974, 5 U.S.C. §552a (1976); see below.

4. Pub. L. No. 94–553, §102(a) (1976), encoded as 17 U.S.C. §101 *et seq.* (1976); see below.

5. Pub. L. No. 94–553, §102(a) (1976), encoded as 17 U.S.C. §101 *et seq.* (1976); see below.

6. Freedom of Information Act, 5 U.S.C. §552 (1976); Privacy Act of 1974, 5 U.S.C. §552a (1976); *Forsham* v. *Harris,* 445 U.S. 169 (1980); see below. Though Hedrick (in this volume) identifies a fifth group, the scientific community, it has received no special recognition under the law. When the interests of the scientific community are considered, it is in the context of furthering some general societal purpose. Rights of research participants are also given little recognition in law governing access to data, but are acknowledged in statutes and regulations specifying the proper conduct of federally funded research. 42 U.S.C. §242(a) (1976); 21 U.S.C. §1175a (1976); 42 U.S.C. §3771 (1976); 42 U.S.C. §242m(d) (1976); Final Regulations Amending Basic HHS Policy for the Protection of Human Research Subjects, 45 C.F.R. §46 (1981). Consequently, this paper focuses on the rights of the primary researcher and the data requester, interpreting these rights in the context of the benefits to society, the interest given greatest deference under the law.

7. The federal government spent approximately $12 billion on basic and applied research in 1980 (National Science Board, 1981). Reductions in federal funding for scientific research may result in renewed attention to the data sets developed with federal funding in past years. The policy of the Reagan administration toward funding of social and behavioral sciences is discussed by Holden (1981), Mosteller (1981), Prewitt and Sills (1981), Holden (1982), Norman (1983a, 1983b).

8. There is a fourth circumstance that may be considered as a separate category, arising when a privately developed data set is maintained by a federal agency; see, e.g., *St. Paul's Benevolent Educational and Missionary Institute* v. *U.S.,* 506 F. Supp. 822 (N.D. Ga. 1980). This situation is very rare and appears to be governed by the standards discussed under circumstances involving access to agency records under the Freedom of Information Act (see below and Dickson, 1980).

9. No general right of access to privately developed research data exists outside the context of litigation. The rights of those seeking access to data arise only in those rare circumstances in which both the primary researcher and the person seeking access to the data are parties to litigation. Even then only limited access may be permitted to meet the needs of the litigation. For an example of an instance in which the court refused to permit access to confidential research records developed by a researcher who was not a party to litigation, see *Richards of Rockford* v. *Pacific Gas and Electric,* 71 F.R.D. 388 (N.D. Cal. 1976). The general rights of research participants to control access to research records is even more limited. The legal relationship between a researcher and research participants who have received promises of confidentiality is discussed by Teitlebaum (1983); see also Boruch and Cecil (1979).

10. “Secrecy is likely to work most effectively where the uncopyrighted product is an intermediate good which is itself used in the production of other goods, e.g., if it is a computer program used to process other people's data. When this is the case it is easier for this investor to bear the cost of his investment without revealing his secret because his product need never leave his hands” (Braunstein et al., 1979).

11. Unlike other products, which are consumed upon use, information is not diminished by use and requires special protection. It is this characteristic that causes information to be “a somewhat recalcitrant economic good” (Thompson, 1979:30). Posner has summarized the relationship between this unique property of information and the need for legal protection as follows (Posner, 1979:1193): “The underlying problem of information production is the difficulty of appropriating private profits from any of the social benefits that the disseminator of information

creates…. [I]f I sell you an idea, and you use it to produce something that reveals the idea, anyone else can use the idea without dealing with me.... The point, however, is that some legal intervention or other 'artificial' restriction is necessary to make an idea a saleable commodity." Hammond (1981) has found this conception of information as a private commodity to be outdated and has urged a reformulation of information property rights that would recognize the character of information as a "collective economic good". He believes that traditional Western legal systems "have been too much concerned with the creation of a sufficient stock of information and too little concerned with usage and access (p. 55)." Nevertheless, it is access to and dissemination of information that must be controlled if researchers are to realize the value of the objects they create.

12. In *Twentieth Century Music Corp.* v. *Aiken,* 422 U.S. 151, 156 (1975) the Court said: "Creative work is to be encouraged and rewarded, but private motivation must ultimately serve the cause of promoting broad public availability of literature, music, and the other arts. The immediate effect of our copyright law is to secure a fair return for an 'author's' creative labor. But the ultimate aim is, by this incentive, to stimulate artistic creativity for the general public good." See also *Fox Film Corp.* v. *Doyal,* 286 U.S. 123, 127 (1932).

13. This economic incentive system was acknowledged by the Supreme Court in *Mazer* v. *Stein,* 347 U.S. 201, 219 (1954): "The economic philosophy behind the clause empowering Congress to grant patents and copyrights is the conviction that encouragement of individual effort by personal gain is the best way to advance public welfare through the talents of authors and inventors in 'Science and Useful Arts.'" Of course, a similar incentive process is at work in the scholarly community.

14. Copyright protection is part of a larger body of law, known generally as "intellectual property." This broad area also includes patent protection and trade secret protection. Neither patent protection nor trade secret protection is appropriate for guarding the proprietary interest of a researcher who develops a data set. Patent protection may be extended to "[w]homever invents or discovers any new and useful *process,* machine, manufacture, or composition of matter, or any new and useful improvement thereof …" (35 U.S.C. §101 (1976)) [emphasis added]. The term "process" has been given a narrow interpretation that would not include development of a data set, even under the recently expanded standards set by the Supreme Court in *Diamond* v. *Diehr,* 450 U.S. 175 (1981). See also *Gottschalk* v. *Benson,* 409 U.S. 63 (1972). Although in *Diamond* v. *Diehr,* the Supreme Court indicated that a process that involves a scientific truth *may* be eligible for patent protection, the development of a data set is not such a process, since there is no "[t]ransformation of an article to a different state or thing" (*Gottschalk* v. *Benson,* 409 U.S. 63, 700 (1972)). Trade secret protection is also inappropriate to general research data sets, since this form of protection is extended by statute to a limited number of areas that require federal reporting of commercial information: see, e.g., Energy Supply and Environmental Coordination Act of 1974, §11(d), 14 U.S.C. §796(d) (1976); and the Flammable Fabrics Act, §4(c), 15 U.S.C. §1193(c) (1976). Many of these statutes simply reference the general protection offered by the Trade Secrets Act, 18 U.S.C. §1905 (1976). For a discussion critical of trade secret protection of health and safety testing information concerning drugs and pesticides, see T.O. McGarity and S. Shapiro (1980). A final alternative for a researcher who believes that his or her data set has been misused is to bring a claim in a state court under the common law tort of misappropriation. This doctrine has been extended to protect compilations of facts distributed for commercial purposes, such as news gathering services: e.g., *International News Serv.* v. *Associated Press Inc.,* 248 U.S. 215 (1918); *Veatch* v. *Wagner,* 116 F. Supp. 904 (D. Alaska 1953). However, in the circumstance presented here of a data set being copied for use in subsequent analyses, it is likely that the courts would find that the state misappropriation doctrine is preempted by the federal copyright law: see Denicola (1981) and *Comment* (1977b). Therefore, this paper focuses on the protection afforded data sets under the copyright law of the federal government.

15. Pub. L. No. 94–553, §102 (1976); codified at 17 U.S.C. §1 *et seq.* (Supp. 1977). When

this act became effective on January 1, 1978, all other forms of copyright protection for new works were preempted. Copyright protection has been extended to a wide variety of tangible expressions, including computer programs, pantomimes and choreographic works (when expressed in tangible form), blank forms, and color arrangements (Nimmer, 1980:2.04–2.19).

16. The requirement of expression in a tangible form is also found in the constitutional provision that gives authors an exclusive right to "their respective *writings* and discoveries" U.S. Constitution, art. I, §8 [emphasis added]. For a discussion of the difficulties in extending traditional doctrines of copyright protection to computer data bases, see Denicola (1981:531).

17. See also Denicola (1981).

18. *Leon* v. *Pacific Tel. §Tel.,* 91 F.2d 484 (9th Cir. 1937).

19. *Edward § Deutsch Lithographing Co.* v. *Boorman,* 15 F.2d 35 (7th Cir. 1926).

20. The process of copyrighting a data set expressed as a list of numbers on sheets of paper is straightforward. Copyright protection attaches to the data set as the data are "fixed in a tangible medium of expression" (i.e., entered onto the data sheet). However, when the data set is distributed or made public, the author's rights will be protected only if: the copyright notice is affixed to the data set in such a way as to give fair notice of the author's claim to the copyright; the data set is registered with the copyright office; and two copies of the data set are filed with the Library of Congress within three months of publication. The copyright notice must contain the following: the letter "c" in a circle, or the word "copyright" or the abbreviation "copr."; the year of first publication of the data set; and the name of the owner of the copyright. The filing of copies with the Library of Congress is usually accompanied by registration of a claim to copyright with the Copyright Office. Failure to register a claim to copyright and deposit the copies with the Library of Congress limits the damages that the copyright holder can obtain from an infringer. Of course, once a document is registered with the Library of Congress, it becomes a publicly available document (Copyright Act §§101, 401 *et seq.* (1976)); see also *Note* (1982). While the notice and registration requirements are only a nuisance for hard-copy data sets, they pose more of a difficulty for large data sets in machine readable form. The statute seems to contemplate copyright of machine readable data sets, since copyright protection attaches to works, "in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced or otherwise communicated, either directly or with the aid of a machine or device" (Copyright Act §§102, 117 (1976)). Yet the notice provision in the statute remains ambiguous as to the notice necessary for copyright of machine readable data sets that are not also available in hard copy (Copyright Act §§117, 401 *et seq.* (1976)).

21. Compare *International News Service* v. *Associated Press,* 248 U.S. 215 (1918).

22. See Squires, 1979; and Nimmer, 1980:2.01[A]. Note especially the discussion by Squires (1979:214–215) of "slipping" where a compilation of information is used to guide the fact-gathering in the development of a compilation of the same information. The difficulty and importance of distinguishing between facts and the expression of facts in determining the limits of copyright protection is discussed in *Miller* v. *Universal City Studios,* 650 F.2d 1365 (5th Cir. 1981). For an argument that copyright protection should be expanded to protect an "author's research," in circumstances where the selection of facts is the essence of the creative contribution of the author, see <Comment (1982).

23. Such a circumstance is likely to occur if the principal investigator develops a data set from publicly available sources. For an example of such a replication, see Passell and Taylor (1977), attempting to replicate the study by Ehrlich (1975). Passell and Taylor relied on an unpublished list of data sources made available by Ehrlich. See also <Comment (1977a).

24. As discussed above, the law offers its full protection only if the the copyrighted work is registered with Library of Congress and made available to the public. Therefore, if researchers wish to enforce their rights under the Copyright Act, they must register their data sets in such a way that the public will have access to them. This registration requirement, which is quite suitable for authors and composers, is fundamentally inconsistent with the interest of researchers in

retaining control over data sets.

25. Protection of the original contribution of the copyright holder amounts to "… little more than a prohibition of actual copying…. Any 'distinguishable variation' of a prior work will constitute sufficient originality to support a copyright if such variation is the product *Alfred Bell § Company* v. *Catalda Fine Arts, Inc.,* 191 F.2d 99 (2d Cir. 1951)). Since facts themselves cannot be copyrighted, a principal researcher's original contribution is in the expression of the facts through their arrangement. As discussed above, if a second researcher copies the facts in the principal researcher's data set, but varies the arrangement, perhaps to permit some novel analysis, this variation in selection and arrangement of facts may be an original contribution of the second researcher and may not infringe on the copyright of the principal investigator. The quantum of originality present in the work of the second researcher required to remove it from the constraints of copyright protection is not great. For a more restrictive interpretation, see Denicola (1981:522).

26. Freid (1979) offers an excellent review of the fair use provisions of the Copyright Act.

27. See also *Encyclopaedia Britannica Educational Corp.* v. *Crooks,* 447 F. Supp. 243 (W.D.N.Y. 1978).

28. As Freid (1979:468) notes: "If copyrighted material is used in such a way that the arts and sciences are benefited, the purposes of the copyright laws are being furthered, despite the apparent invasion of the copyright owner's 'exclusive' rights." Freid cites *Rosemont Enterprises, Inc.* v. *Random House, Inc.,* 366 F.2d 303 (2d Cir. 1966), *cert. denied,* 385 U.S. 1009 (1967) and *Time Inc.* v. *Bernard Geis Associates,* 293 F. Supp. 130 (S.D.N.Y. 1968), and he notes that these cases also support the proposition that the arts and sciences should be interpreted in broad terms, since the first case involved the use of a copyrighted article in a biography about Howard Hughes, and the second case involved the use of copyrighted frames of a motion picture in a book about the assassination of President Kennedy.

29. *Loew's Inc.* v. *Columbia Broadcasting System,* 131 F. Supp. 165 (D.C. Cal. 1955), *affirmed,* 356 U.S. 43 (1957), *rehearing denied,* 356 U.S. 934 (1957). See also *Williams § Wilkins Co.* v. *United States,* 487 F.2d 1345 (Ct. Cl. 1973), *aff d by an equally divided Court,* 420 U.S. 376 (1975), discussed below.

30. When the law was being revised, the House judiciary subcommittee recognized the special needs of teachers to copyrighted material and requested that the educational community draft guidelines for "educational" fair use for printed material and music (H. R. Rep. No. 1476, 94th Cong., 2d Sess. 66 (1976)). The resulting guidelines were then adopted by the House committee as "a minimal interpretation of the standard of 'fair use'" and incorporated into its report (p. 72). It remains open to the courts to decide that minimal acts of photocopying by a teacher, though they go beyond the guidelines, nevertheless constitute fair use (Nimmer, 1980:1305[E]).

31. *Williams § Wilkins Co.* v. *United States,* 487 F.2d 1345 (Ct. Cl. 1973), *aff d by an equally divided Court,* 420 U.S. 376 (1975); but see Freid (1979:474–477).

32. 366 F.2d 303 (2d Cir. 1966), *cert. denied,* 385 U.S. 1009 (1967). The court also noted that, "… the arts and sciences should be defined in their broadest terms, see *Sampson § Murdock Co.* v. *Seaver-Radford Co.,* 140 F. 539, 541 (1st Cir. 1905), particularly in view of the development of the field of social sciences" (366 F.2d at 307).

33. 487 F.2d 1345 (Ct. Cl. 1973), *aff d by an equally divided Court,* 420 U.S. 376 (1975); see also discussion by Freid (1979:467–477).

34. *Williams § Wilkins Co.* v. *United States,* 487 F.2d 1345, 1359 (Ct. Cl. 1973), *aff d by an equally divided Court,* 420 U.S. 376 (1975); see also Freid (1979:472, 475). This also appears to be true for large machine-readable data sets (Keplinger, 1977).

35. See, generally, Office of Federal Statistical Policy and Standards (1980b). Another common source of research data from federal agencies is public-use data tapes, developed from the records of federal agencies. For a statement of basic federal policy and a list of federal statistical data file catalogs and directories, see Sprehe (1981). For examples of studies that have relied on

public-use data tapes prepared by the National Center for Health Statistics and the Social Security Administration, see Flaherty (1979ÄAppendices 1 and 2).

36. See Privacy Protection Study Commission (1977), and Sprehe (1981). For a discussion of state statutes, see Braverman and Heppler (1981). For an international perspective on the problems of obtaining research access to governmental records, see Mochmann and Muller (1979), Flaherty (1979), and Rozsa and Foldi (1980).

37. 44 U.S.C. §2901 *et seq.* (1976); see also The Records Disposal Act, 44 U.S.C. §3314 (1976).

38. Administrative Procedure Act, ch. 324 §3, 60 Stat. 238 (1946) (amended 1966).

39. 5 U.S.C. §552 (1976). The original version of the Freedom of Information Act was passed in 1967, Pub. L. No. 90-23 (1967); the 1974 amendments were contained in Pub. L. No. 93–502 (1974).

40. 5 U.S.C. §552a (1976).

41. 5 U.S.C. §552(e) (1976).

42. 4 C.F.R. §81 (1980).

43. According to an opinion issued by Carl H. Imlay, general counsel, Administrative Office of the United States Court, November 16, 1978: "[C]ourts of the United States and the Administrative Office are exempt from coverage under [the Freedom of Information Act]." See also 5 U.S.C. §§551(1)(B), 552(e) (1976); *United States* v. *Dingle,* 546 F.2d 1378, 1380-81 (10th Cir. 1976); *United States* v. *Caniff,* 521 F.2d 565, 573 (2d Cir. 1975); *Cook* v. *Willingham,* 400 F.2d 885 (10th Cir. 1968). The courts are frequently faced with requests for disclosure of presentence reports. Rule 32(c)(3) of the Federal Rules of Criminal Procedure provide for the disclosure of the report to the defense, although judges retain broad discretion to prohibit or limit the defendant's access to the report. Circumstances in which presentence information can be disclosed to third parties are quite rare and within the discretion of the court (Volume X of the *Guide to Judiciary Policies and Procedures,* §3004 (1978)). For a discussion of the disclosure practices of presentence reports by federal judges, see Fennell and Hall (1980).

44. From time to time, the U.S. Probation Office, a part of the judiciary, receives requests from researchers for access to probation records. In the past these requests have been reviewed on a case-by-case basis. However, a policy statement by the U.S. Probation Office suggests that access to records maintained by the U.S. Probation Office for the District of Columbia can be obtained if the records remain anonymous or, if identifiable records are necessary, if the informed consent of the research participant is obtained. The Probation Office also reviews research proposals to determine that they are of scientific merit (memorandum and attachments, Mr. Stephen J. Reynolds, U.S. Probation Office, October 6, 1978). See also, Federal Judicial Center Advisory Committee on Experimentation in the Law (1982).

45. *Kissinger* v. *Reporters Committee for Freedom of the Press,* 445 U.S. 136, 155–157 (1979).

46. Administrative Procedure Act, ch. 324 §3(c), 60 Stat. 238 (1946) (amended 1966).

47. Some have argued that a similar public right to information resides in the Constitution; see Lewis (1980).

48. A number of proposals have been introduced that would restrict access to information under the Freedom of Information Act: see, e.g., Hill Panel Votes Bill to Restrict Information Act (*Washington Post,* December 15, 1981ÄA6).

49. 5 U.S.C. §552(b) (1976).

50. 5 U.S.C. §552(a)(3) (1976).

51. Once an agency receives a request that "reasonably describes" the desired records, the agency has ten working days to decide if any or all of the material comes under any of the nine exemptions. The requesting party must be properly informed of the agency's decision, and, if applicable, of the agency's reasons for denying access to the information. An individual denied any information may pursue an administrative appeal for disclosure. Should the appeal fail, the re

quester can then file for disclosure in federal district court. The court decides the matter de novo and may examine any documents in determining whether they are, in part or in whole, exempt. The agency has the burden of proving that the materials in question come under a particular exemption: 5 U.S.C. §§552(a)(3) (a)(6)(A)(i) (a)(4)(B) (1976).

52. If the research data are exempted from disclosure by some other statute, they need not be disclosed under the Freedom of Information Act: 5 U.S.C. §552(b)(3) (1976). An amendment to the Freedom of Information Act in 1977 limited this exemption to records governed by confidentiality statutes that *require* the records to be withheld from the public; confidentiality statutes that *permit* the exercise of discretion in withholding records are not adequate to meet the standards of this exemption: Government in the Sunshine Act, Pub. L. 94–409 (1976). The effect of this change was to substantially narrow the applicability of this exemption and shift attention to other relevant exemptions (Office of Federal Statistical Policy and Standards, 1978). See also *Comment*(1981).

53. 5 U.S.C. §552(b)(6) (1976).

54. 5 U.S.C. §552(b)(4) (1976).

55. See, for example, *Washington Research Projects* v. *DHEW,* 504 F.2d 238 (D.C. Cir. 1974). This policy of narrow construction of the exemptions was further aided by the policy of the Justice Department in the previous administration of defending Freedom of Information suits only when "disclosure is demonstrably harmful, even if the documents technically fall within the exemptions in the Act." Letter from Attorney General Griffin Bell to Heads of Federal Departments and Agencies (May 5, 1977). The current administration has apparently reversed this policy (*New York Times,* May 5, 1981ÄA18).

56. Compare *Consumers Union* v. *Veterans Administration,* 301 F. Supp. 796 (S.D.N.Y. 1969), dismissed as moot, 436 F.2d 1363 (2nd Cir. 1971), with *Getman* v. *NLRB,* 450 F.2d 670 (D.C. Cir. 1971); see also *Note* (1975). In fact, a recent Supreme Court case took a contrary stand, holding that an agency can decide to release information even when that information qualifies as being exempt from disclosure: *Chrysler Corp.* v. *Brown,* 441 U.S. 281 (1979).

57. 5 U.S.C. §552(b)(4) (1976).

58. 5 U.S.C. §552(b)(4) (1976). See also *Note* (1976c); *Comment* (1976a); *Braintree Elec. Light Dept.* v. *Dept. of Energy,* 494 F. Supp. 287 (D.D.C. 1980); *National Park § Conservation Assn.* v. *Morton,* 498 F.2d 765 (D.C. Cir. 1974); *Consumers Union* v. *Veterans Administration,* 301 F. Supp. 796 (S.D.N.Y. 1969), dismissed as moot, 436 F.2d 1363 (2nd Cir. 1971).

59. While the FOIA does not define "trade secret," an older edition of the *Restatement of Torts* (American Law Institute, 1938:§757(b)), which is often cited for its definition, states that "a trade secret may consist of any formula, pattern, device or compilation of information which is used in one's business, and which gives him an opportunity to obtain an advantage over competitors or suppliers who do not know or use it. It may be a formula for a chemical compound, a process of manufacturing, treating or preserving materials, a pattern for a machine or other device, or a list of customers." See also *Public Citizen Health Research Group* v. *Food and Drug Administration,* 539 F. Supp. 1320, 1325 (D.D.C. 1982); Connelly (1981); and Stevenson (1982).

60. "Lacking any legislative history defining the scope of the terms 'commercial' and 'financial,' the courts have given them their ordinary meanings": *Board of Trade* v. *Commodity Futures Trading Co.,* 627 F.2d 392, 403 (1980).

61. *Fisher* v. *Renegotiation Board,* 355 F. Supp. 1171, 1175 (D.D.C. 1973).

62. *Public Citizen Health Research Group* v. *DHEW,* 477 F. Supp. 595, 605 (D.D.C. 1979). The district court order to release the information was later reversed following the decision by the Supreme Court in *Forsham* v. *Harris.* After an extensive examination of the legislative the court of appeals held that the medical foundation did not have history of the act establishing Professional Standard Review Organizations, to turn over the documents since it, serving as a Professional Standard Review Organization, was not an "agency" under the FOIA: *Public Citizens Health Research Group* v. *HEW* 668 F.2d 537 (D.C. Cir. 1981).

63. 504 F.2d 238 (D.C. Cir. 1974).
64. 504 F.2d 244 (D.C. Cir. 1974).
65. 504 F.2d 253 (D.C. Cir. 1974).
66. 504 F.2d 244 (D.C. Cir. 1974). The federal appellate court did hold that the NIMH summary statements and site visit reports were exempt from disclosure since they were intra-agency memoranda. Intra-agency memoranda are exempt from disclosure under the FOIA: 5 U.S.C. §552(b)(5) (1976).
67. 504 F.2d 244–245 (D.C. Cir. 1974). The court adds, in a footnote to this section, that "[o]nly an individual grantee engaged in profit-oriented research, or a non-profit organization that engages in profit-making ventures based on biomedical research, could conceivably be shown to have a commercial or trade interest in his research design." Some commentators have contended that the FOIA and its interpretation in *Washington Research Products* v. *DHEW* greatly interfere with the proprietary rights of researchers (see, e.g., Morris et al., 1981).
68. 506 F. Supp. 822 (N.D. Ga. 1980).
69. 506 F. Supp. 830 (N.D. Ga. 1980).
70. See *Board of Trade* v. *Commodity Futures Trading Co.* 627 F.2d 392, 403 (1980) and *Consumers Union* v. *Veterans Administration,* 301 F. Supp. 796 (S.D.N.Y. 1969), dismissed as moot, 436 F.2d 1363 (2nd Cir. 1971). 5 U.S.C. §551 (2) defines "person" as including individuals, partnerships, corporations, or associations. Information produced by a government agency would not be protected under exemption 4, but it may be protected as interoffice memoranda, another FOIA exemption.
71. 498 F.2d 765 (D.C. Cir. 1974).
72. 498 F.2d 770. *Accord, Pacific Architects § Engineers, Inc.* v. *Renegotiation Board,* 505 F.2d 383 (D.C. Cir. 1974) See also *Comment* (1976a). At least one commentator maintains that, even though the information in *Washington Research Projects* v. *DHEW* was not considered commercial or financial, it would have passed the confidentiality test set up under the *National Park* v. *Kleppe* case (*Note,* 1976c).
73. Substantial competitive harm is a factual question. The party claiming that it will be harmed usually attempts to prove the likelihood of the harm through expert testimony or affidavits. While detailed economic analysis or elaborate antitrust-type market analyses are not required, the resisting party usually must show the likelihood of a specific harm: *National Parks* v. *Kleppe,* 547 F.2d 673 (D.C. Cir. 1976); *Sears, Roebuck and Co.* v. *General Services Administration,* 553 F.2d 1378 (D.C. Cir. 1974).
74. 539 F. Supp. 1320 (D.D.C. 1982).
75. 615 F.2d 551 (1st Cir. 1980).
76. 301 F. Supp. 796 (S.D.N.Y. 1969), dismissed as moot, 436 F.2d 1363 (2nd Cir. 1971).
77. The court held: "Even though the records sought are not exempt, the court is not bound under the Act to automatically order their disclosure. In exercising the equity jurisdiction conferred by the Act, it must, according to traditional equity principles, weight the effects of disclosure and non-disclosure and determine the best course to follow at the present time. In an action under the Freedom of Information Act, which shifts the burden of proof to the defendant, the balance of the equities is presumptively on the side of disclosure. The rule that will be followed, therefore, is this: where agency records are not exempted from disclosure by the Freedom of Information Act, a court must order their disclosure unless the agency proves that disclosure will result in significantly greater harm than good. Because the Act was intended to benefit the public generally, it is primarily the effects on the public rather than on the person seeking the records that must be weighted": 301 F. Supp. 806 (S.D.N.Y. 1969). On appeal the agency abandoned its assertion of a public interest in withholding the information, and the case was dismissed as moot after the agency released the requested information: 436 F.2d 1336, 1365 (2nd Cir. 1971). Compare with *Getman* v. *NLRB,* 450 F.2d 670 (D.C. Cir. 1971), which denies that the courts have such equity jurisdiction.

78. 301 F. Supp. 808 (S.D.N.Y. 1969); see also Kronman (1980).
79. In *Consumers Union* the court permitted release of the raw data, but refused to disclose the scoring system used to assess the raw data. While such a precedent may make it difficult for researchers to replicate findings based on raw data, it would not stand in the way of developing competing analyses. The precedent in *Orion Research Inc.* v. *Environmental Protection Agency,* 615 F.2d 551 (1st Cir. 1980), extended to technical information in a proposal identifiable to an individual bidder and is unlikely to be extended to requests for anonymous research data.
80. *Long* v. *I.R.S.,* 596 F.2d 362 (9th Cir. 1979).
81. *Long* v. *I.R.S.,* 596 F.2d 362, 366 (9th Cir. 1979).
82. See *Forsham* v. *Harris,* 445 U.S. 169 (1980); also see *Note* (1978), which distinguishes *Washington Research Projects* from *Forsham* on the grounds that in *Washington Research Projects* the government already possessed the private report while in *Forsham* the data remained with the researcher.
83. 596 F.2d 362 (9th Cir. 1979).
84. 596 F.2d 369 (9th Cir. 1979).
85. 5 U.S.C. §552a (1976). However, there are areas in which more specific statutes and regulations control access to agency data. In these circumstances, the presence of a specific regulatory scheme may better accommodate the needs of researchers. Sasfy and Siegel (1982) examined the practices of a number of criminal justice agencies in permitting research access to agency records and found that there was no general "chilling effect" on criminal justice research due to the Privacy Act and related privacy and confidentiality statutes. Access to such records is typically governed by specific statutes and regulations that apply to individual agencies and agency records. They found that there may be "chilling effects" on research access in specific agencies if the statutes governing records in these agencies do not contain provisions permitting access for research in the statutes governing agency records. Sasfy and Siegel's work is one of the few studies of such disclosure practices, and it suggests that if research needs are anticipated by the statute, research access to agency records can proceed without difficulty.
86. An early statement of difficulties resulting from the regulation of research records by the Privacy Act is found in "Notice of Hearings and Draft Recommendations: Research and Statistics," 41 *Fed. Reg.* 55007 (proposed, December 16, 1976).
87. See Privacy Protection Study Commission (1977), Mochmann and Muller (1979), and Flaherty (1979); for early discussion of this issue, see D.T. Hulett (1975), and Martin (1974).
88. The Privacy Protection Study Commission (1977, Appendix 4:11) found that as of December 21, 1975, there were 6,723 systems of records of varying size containing 3.8 billion individual records.
89. 44 U.S.C. §3501–3512 (1976).
90. For example, federal tax returns and information have specific statutory protection against disclosure: 26 U.S.C. §1603(a) (1976); see also discussion of Freedom of Information Act, above.
91. 5 U.S.C. §552a(a)(4) (1976).
92. 5 U.S.C. §552a(a)(6) (1976).
93. 5 U.S.C. §552a(b) (1976). For an overview of the Privacy Act of 1974, see *Note* (1976a); *Note* (1976b); Davidson (1976); Eastman (1975); and *Project*) (1975).
94. 5 U.S.C. §552a(b)(1) (1976).
95. The act requires the head of an agency or instrumentality to make a written request to the agency maintaining the record specifying the particular portion desired and the law enforcement activity for which the record is sought.
96. Apparently this exemption was intended to permit access for resolving problems of constituents, but through a drafting error the exemption was extended to Congress as a body rather than individual members. Access to members of Congress for solving constituents' problems is now considered to be a routine use of most record systems (Privacy Protection Study

Commission, 1977:519–520). All of these exemptions are found in 5 U.S.C. §552a(b) (1976).

97. There is general agreement that this publication requirement has been an ineffective means of notifying the public (Commission on Federal Paperwork, 1977).

98. While no court has considered such a practice, one commentator (*Note,* 1976a) has suggested that the courts should construe the consent provision narrowly and reject an agency's claim of prior consent, absent a clause in the original request specifically stipulating not only the anticipated uses, but also the potential recipients of the data.

99. 20 *Congressional Record* H. 12246 (December 18, 1974).

100. According to 5 U.S.C. §552a(a)(1) (1976), the term "agency" means agency as defined by the Freedom of Information Act, 5 U.S.C. §552(e) (1976); see also discussion of the FOIA, above.

101. See note 88. The distinction favoring records retrieved by individual identifiers seems to presume a manual rather than a computer-based information system, but most federal records are contained in automated record systems. Of those personal data systems reported to the Office of Management and Budget (OMB), only 21 percent are fully or partially automated; but 81 percent of the total number of individual records are maintained in these systems (*Federal Personal Data Systems Subject to the Privacy Act of 1974, First Annual Report to the President, Calendar Year 1975,* page 2, cited in Privacy Protection Study Committee, 1977, Appendix 4:11). Some people have questioned whether this "systems of record" definition is adequately broad to serve as a triggering mechanism for the protections of the Privacy Act (Appendix 4:6). Computer technology permits identification of an individual's record based on some combination of attributes or characteristics, as well as by individual identifiers. Yet, without regard to the ease with which an attributional search by computer can be made, an agency may place the record system beyond the scope of the act by retrieving the records by some means other than individual identifiers.

102. The Commission on Federal Paperwork (1977:115) strongly endorsed the use of administrative records for research and statistical purposes. For examples and a discussion of agencies sharing administrative records for research purposes, see Privacy Protection Study Commission (1977:588).

103. For commentary regarding standards of consent required by the acts, see *Project* (1975:1309–1310) and *Note* (1976a:682).

104. 5 U.S.C. §552a(b)(1) (1976). Some people have contended that the designation of large organizations, such as the Department of Health and Human Services (HHS), as a single agency has permitted improper and unmonitored transfer of sensitive records to diverse units (Commission on Federal Paperwork, 1977:67). Of course, disclosures within an agency may be restricted by a number of other statutes. For example, the Tax Reform Act of 1976 will not permit disclosure of tax information by the Social Security Administration to other researchers in HHS, even though the Social Security Administration is part of HHS (Office of Federal Statistical Policy and Standards, 1980b:97). There have been a number of legislative proposals to improve the interagency sharing of information for research purposes. The proposals vary in their details, but nearly all involve some "functional separation" of statistical and administrative records with greater centralization of research and statistical responsibilities (see, e.g., Alexander, 1983). Since these proposals are intended to improve the interagency sharing of research information—rather than to improve directly the opportunity for persons outside the federal government to obtain access to this information—these proposals are not addressed in this paper.

105. The Privacy Protection Study Commission (1977, Appendix 4:67) found that the Privacy Act has resulted in a modest overall decline in the amount of individual information agencies disclose to others, but that impact has been greatest at the margins of agency duties, such as support for nonfederal research. Researchers and statisticians who have received identifiable information are mostly federal agency employees or contractors, some grantees, and a relatively small number of persons who have neither contracts nor grants. Disclosure usually consisted of a list of names and addresses (Privacy Protection Study Commission, 1977:590).

106. See the testimony of Drs. Leonard T. Kurland and Lee Robins before the Privacy Protection Study Commission (June 11, 1976 and September 20, 1976). At least one epidemiologist contended that restrictions such as those contained in the Privacy Act "will spell a virtual end to population-based studies directed toward solving public health problems" (statement by Helen Chase of the Joint Committee on National Data Resources of the American Public Health Association, cited in the testimony of Leonard Kurland); see also Curran (1978), and Kelsey (1981).

107. Apparently the greatest difficulty now facing researchers who wish to link archival records for longitudinal studies is the maze of state and federal privacy legislation that followed the Privacy Act, which extends to specific agencies or specific kinds of records, such as medical records (Beebe, 1980). For a review of the effects of recommendations of the Privacy Protection Study Commission on longitudinal research, see Robins (1978).

108. Some commentators contend that the exemptions, coupled with the ineffective scheme for enforcement, largely defeats the requirement of obtaining informed consent prior to disclosure of personal records (*Note,* 1976a:691).

109. The Commission on Federal Paperwork (1977:111) took exception to this limitation on disclosure and noted that if the record "is to be used solely for statistical purposes, there seems no need for requiring that it be transferred 'in a form that is not individually identifiable.' Such restrictions have severely limited not only the interagency flow of information but the release to the public of much worthwhile information, such as that contained in statistical microdata files."

110. A review of these techniques is found in Boruch and Cecil (1979); see also Office of Federal Statistical Policy and Standards (1978, 1980a).

111. Office of Management and Budget (1975); see also Privacy Protection Study Commission (1977:571).

112. Examples of injuries to individuals from improperly disclosed research data are difficult to find. Efforts by the Office of Federal Statistical Policy and Standards (1978) and by the Privacy Protection Study Commission to identify instances of injury resulting from improperly disclosed federal records turned up no examples. Of course, individuals may be adversely affected by interpretations of data that identify characteristics of a group of which they are a member. See Morris et al. (1981), for an example of embarrassment to a group of teachers resulting from publication of statistical characteristics of the group based on improper interpretation of personnel test data. Similarly, the only "injury" found by the Office of Federal Statistical Policy and Standards (1978:34) involved complaints by several persons that "release of population census summary data by zip-code area has contributed to their increasing receipt of junk mail." However, even if the identities of individuals are withheld, it may be possible to deduce their identities from the public information that is released. For example, Nelson and Hedrick (1983:34) sought to identify researchers who received grants of confidentiality under the Drug Abuse Act of 1970. Although their FOIA request for the names of the researchers was denied (with misplaced reliance on the Privacy Act), some information concerning the general nature of the research project was released. While the agency's decision to withhold the names of the grantees was being appealed (an appeal that was ultimately successful), the names of 76 percent of the grantees were identified by matching the released information (contract numbers, telephone numbers, etc.) with other publicly available information.

113. For a list of specific epidemiological studies that would have been "virtually impossible" to conduct without identifiable information, see Gordis et al. (1977).

114. In the past some academic researchers and personnel from other agencies have been sworn in as Census Bureau officials to conduct special analyses (Martin, 1974:265). Sasfy and Siegel (1982) also found the use of "temporary employees" to be a common practice of criminal justice agencies.

115. The routine uses of a record must be listed in the annual system notices and must be published for comment in the *Federal Register* at least 30 days before they are included for the first

time in the annual system notice, 5 U.S.C. §552a(e)(11)(1976).

116. A routine-use provision that permits access to identifiable records for research, which appears in many record systems notices of the Department of Health and Human Services, reads as follows: "A record may be disclosed for a research purpose, when the Department: (A) has determined that the use of disclosure does not violate legal or policy limitations under which the record was provided, collected, or obtained; (B) has determined that the research purpose (1) cannot be reasonably accomplished unless the record is provided in individually identifiable form, and (2) warrants the risk to the privacy of the individual that additional exposure of the record might bring; (C) has required the recipient to (1) establish reasonable administrative, technical, and physical safeguards to permit unauthorized use or disclosure of the record (2) remove or destroy the information that identifies the individual at the earliest time at which removal or destruction can be accomplished consistent with the purpose of the research project, unless the recipient has presented adequate justification of a research or health nature for retaining such information, and (3) make no further use or disclosure of the record except (a) in emergency circumstances affecting the health or safety of any individual (b) for use in another research project, under these same conditions, and with the written authorization of the Department (c) for disclosure to a properly identified person for purpose of an audit related to the research project, if information that would enable research subjects to be identified is removed or destroyed at the earliest opportunity consistent with the purpose of the audit, or (d) when required by law; (D) has secured a written statement attesting to the recipient's understanding of, and willingness to abide by these provisions." For examples of such notices of routine use for research purposes, see the Department of Health and Human Services' annual publication of the systems of records, 46 *Fed. Reg.* 52693, 52697, 52700 (various Medicare, Medicaid, and health insurance record systems), 52781 (mental health record systems), 52809, 52782, 52794, 52809, and 52867 (various clinical research record systems) (October 27, 1981). Many other record systems permit research access to "approved or collaborating researchers, including HHS contractors and grantees." For examples of such notices, see the Department of Health and Human Services' annual publication of the systems of records, 46 *Fed. Reg.* 52796, 52797, 52798 (October 27, 1981). For even more general notices of research as a routine use, see 41 *Fed. Reg.* 39719, 39720 (September 15, 1976) (personnel records maintained by the Federal Trade Commission), and 41 *Fed. Reg.* 55568 (December 14, 1976) (personnel records of the Civil Service Commission). Of course, the first version of the notice is preferable, since it permits disclosure to those who are not collaborating researchers while establishing the necessary safeguards to protect the identified individuals.

117. Reliance on the routine-use provision of the Privacy Act to permit sharing of identifiable research data is also risky for another reason. In examining agency practices, the Commission on Federal Paperwork (1977:66-7) found that in many instances, "agency 'routine use' notices authorize transfers for purposes which, by no stretch of the imagination, could be considered 'compatible' with the purpose for which it was collected. Typical of these is the practice of many agencies to share medical information with law enforcement agencies" [footnotes omitted]. It seems that such excesses may make the routine-use exemption ripe for reform. The relevant House committee in its initial report promised vigorous oversight of agency practices in this area (H. Rep. No. 93–1416, 93rd Cong., 2d Sess. 12, 1974). A well-tailored routine-use exception permitting access for research to specific record systems seems proper under the Privacy Act. However, if the routine-use section of the Privacy Act is restricted without consideration of the consequences to research that relies on the current exemptions, one of the few mechanisms for permitting access to identifiable records may be lost.

118. See *Comment* (1976b), *Note* (1976b), *Note* (1976a), and *Project* (1975:1337). This interpretation is consistent with the analysis offered in *Comment* (1976a:135, 140), which maintained that: "the important point is that the FOIA is the parent act and ultimately governs access to information. The Privacy Act is relegated to the backseat when a successful disclosure request is made under the FOIA. Thus, even if a record has been declared exempt under the Privacy Act,

access may still be sought under the FOIA with its nine exemptions. If the record is available under the FOIA, access must be granted, the Privacy Act notwithstanding."

119. 5 U.S.C. §552(b)(6) (1976).

120. For a review of these cases, see Kronman (1980).

121. *Dept. of Air Force* v. *Rose,* 425 U.S. 352 (1975); *U.S. Dept. of State* v. *The Washington Post Co.,* 456 U.S. 595 (1982).

122. *U.S. Dept. of State* v. *The Washington Post Co.,* 456 U.S. 595, 596 (1982), citing the standard used by the court of appeals.

123. 456 U.S. 595 (1982).

124. Prior to the *Washington Post* decision, many lower courts had used this stricter standard: *Getman* v. *NLRB,* 450 F.2d 670 (D.C. Cir. 1971); see also *Note* (1975); *Robles* v. *E.P.A.,* 484 F.2d 843 (4th Cir. 1973); *Rural Housing Alliance* v. *Dept. of Agriculture* 498 F.2d 73, 77 (D.C. Cir. (1974); *Sims* v. *C.I.A.,* 692 F.2d 562 (D.C.C. 1980).

125. 456 U.S. 595, 602 (1982).

126. 425 U.S. 352 (1975).

127. 425 U.S. 352, 382 (1975).

128. 366 F. Supp. 929 (D.D.C. 1973).

129. 366 F. Supp. 929, 937-38 (D.D.C. 1973).

130. 477 F. Supp. 595 (D.D.C. 1979), *rev'd on other grounds,* 668 F.2d 537 (D.C. Cir. 1981); see note 62.

131. The court also found that the patient's privacy interest was protected by the removal of personal identifiers and the doctors' privacy interest, while more substantial, still did not make disclosure "clearly unwarranted" 477 F. Supp. 595, 604–605 (D.D.C. 1979).

132. 539 F.2d 58 (10th Cir. 1976).

133. 539 F.2d 58, 62 (10th Cir. 1976).

134. *Getman* v. *NLRB,* 450 F.2d 670 (D.C. Cir. 1971).

135. *Disabled Officers Associated* v. *Rumsfeld,* 428 F. Supp. 454 (D.D.C. 1977).

136. *Ditlow* v. *Schultz,* 517 F.2d 166 (D.C. Cir. 1975).

137. *Committee on Masonic Homes* v. *NLRB,* 556 F.2d 214 (3d Cir. 1977).

138. *Wine Hobby U.S.A.* v. *Internal Revenue Service,* 502 F.2d 133 (3d Cir. 1974).

139. 477 F. Supp. 595, 605 (D.D.C. 1979), *rev'd on other grounds,* 668 F.2d 537 (D.C. Cir. 1981).

140. 477 F. Supp. 595, 604–605 (D.D.C. 1979).

141. There have been a number of legislative proposals to amend the Privacy Act to permit greater access for research purposes; see, for example, the Privacy of Research Records Act, introduced in the House as H.R. 3409, 96th Cong., 1st Sess. (1979), and in the Senate as S. 867, 96th Cong., 1st Sess. (1979); the Privacy of Medical Information Act, S. 865, 96th Cong., 1st Sess. (1979); and the Confidentiality of Statistical Records Act, which was never introduced but was intended to be part of the Paperwork Reduction Act, Pub. L. No. 96–511 (1980). See also the recommendations of the Privacy Protection Study Commission (1977).

142. This discussion assumes that an agency has not made some provision for release of information at the time it awards funds to contractors and grantees. In fact, several agencies have established policies to ensure that research data collected through funds provided by the agency will become available to the public at the termination of the grant or contract. For example, the National Institute of Justice includes in its research grants a condition that requires the grantee to furnish the Institute a documented, computer-readable copy of all data sets and programs developed in connection with the project; these data sets are maintained by the agency and other data archives (Garner, 1981). For an account of the frustrations faced by researchers who must share data with the federal sponsors of the research, see Dawber (1980).

143. 5 U.S.C. §552(e) (1976). The Privacy Act adopts this definition of "agency," 5 U.S.C. §552a(1) (1976); see, generally, *Note* (1981).

144. 5 U.S.C. §552(a)(4)(B) (1976).
145. 445 U.S. 169 (1980). There was also an earlier FOIA suit for the same information by a pharmaceutical manufacturer: *Ciba-Geigy* v. *Matthews,* 428 F. Supp. 523 (S.D.N.Y. 1977).
146. The Committee on the Care of the Diabetic also sued the FDA to enjoin the proposed labeling of the controversial drugs. The First Circuit remanded the case to the FDA for exhaustion of administrative remedies: *Bradley* v. *Weinberger,* 483 F.2d 410 (1st Cir. 1973). The administrative law judge then found that one of the drugs, phenformin hydrochloride, was not shown to be safe and ordered it withdrawn from the market: 44 *Fed. Reg.* 20967 (1979). However, this decision was not based substantially on the raw data of the University Group study, but on reference to the study as the basis of an expert opinion.
147. *Forsham* v. *Califano,* 587 F.2d 1128, 1136 (D.C. Cir. 1978).
148. *Forsham* v. *Califano,* 587 F.2d 1128, 1141–1142 (D.C. Cir. 1978).
149. *Forsham* v. *Harris,* 445 U.S. 169, 171 (1980).
150. A legislative conference report indicated that Congress did not "intend to include corporations that receive appropriated funds but are neither chartered by the Federal Government nor controlled by it, such as the Corporation for Public Broadcasting": H. Conf. Rep. No. 93–1380, 93rd Cong., 2d Sess. (1974), cited by the Court in *Forsham* v. *Harris,* 445 U.S. 169, 179 (1980).
151. The court mentioned in a footnote that a number of bills seeking to expand the FOIA to federal grantees have been introduced in each Congress since the 92nd, but none has yet been reported out of committee: *Forsham* v. *Harris,* 445 U.S. 169, 179, footnote 10 (1980).
152. *Forsham* v. *Harris,* 445 U.S. 169, 182 (1980).
153. *Forsham* v. *Harris,* 445 U.S. 169, 182–187 (1980). The court stated: "Petitioners place great reliance on the fact that HEW has a right of access to the data, and a right if it so chooses to obtained permanent custody of the UGDP records. [citation omitted] But in this context FOIA applies to records which have been *in fact* obtained, and not to records which merely *could have been* obtained. [emphasis in original, footnote omitted] To construe FOIA to embrace the latter class of documents would be to extend the reach of the Act beyond what we believe Congress intended." *Forsham* v. *Harris,* 445 U.S. 185-6.
154. *Forsham* v. *Harris,* 445 U.S. 169, 180 (1980).
155. *Forsham* v. *Harris,* 445 U.S. 169, 188 (1980).
156. *Forsham* v. *Harris,* 445 U.S. 169, 188–190 (1980).
157. *Forsham* v. *Harris,* 445 U.S. 169, 192 (1980).
158. 445 U.S. 136 (1980).
159. *Reporters Committee for Freedom of the Press* v. *Vance,* 442 F. Supp. 383 (D.D.C. 1977), *aff'd,* 589 F.2d 1116 (D.C. Cir. 1978).
160. *Kissinger* v. *Reporters Committee for Freedom of the Press,* 445 U.S. 136, 139 (1980).
161. 445 U.S. 150 (1980).
162. 44 U.S.C. §2901 *et seq.* (1976).
163. 44 U.S.C. §3314 (1976).
164. *Transamerica Mortgage Advisors, Inc.* v. *Lewis,* 444 U.S. 11 (1979).
165. *Kissinger* v. *Reporters Committee for Freedom of the Press,* 445 U.S. 136, 150 (1980).
166. 445 U.S. 136, 150, note 9 (1980).
167. 445 U.S. 136, 159 (1980).
168. 445 U.S. 136, 161 (1980).
169. 672 F.2d 1262 (7th Cir. 1982).
170. 7 U.S.C. §136d(d) (1976).
171. 672 F.2d 1262, 1274–1277 (7th Cir. 1982). One of the three judges did not concur with the section on academic freedom. There is some uncertainty over the extent of this protection. According to Michael A. Liethen, the attorney representing the University of Wisconsin researchers (quoted in Broad, 1982): "Our view is that a scientist has to be free to take his inquiries where they lead him, and that a scientist should not be forced to disclose his research data until he

has results he is willing to stand behind." However, other language in the opinions suggests that if the data were the evidentiary basis of the administrative condemnation proceeding, the court may well have enforced the subpoena.

172. 5 U.S.C. §552a(m) (1976).

173. OMB Guidelines, 40 *Fed. Reg.* 28947 (July 9, 1975).

174. The general counsel of HEW (now HHS) contended that the requirements of the Privacy Act did not extend to record systems maintained by contractors, since, "[w]here the contracting agency is interested only in obtaining the results of the research or other work performed under the contract (generally in the form of a report) and does not require the contractor to furnish it with individually identifiable records, the system is not one which 'but for' the contract, the agency would have established" (memorandum from Mr. William H. Taft IV, General Counsel, to Mr. John Ottina, Assistant Secretary for Administration and Management, May 14, 1976).

References

Alexander, L. 1983 Proposed legislation to improve statistical and research access to federal records. Pp. 273–292 in R.F. Boruch and J.S. Cecil, eds. *Solutions to Ethical and Legal Issues in Social Research.* New York: Academic Press.

American Law Institute 1938 *Restatement of Torts.* Vol. 4. St. Paul: American Law Institute.

Arnold, M., and Kissiloff, A. 1976 An introduction to the federal Privacy Act of 1974 and its effect on the Freedom of Information Act . *New England Law Review* 11:463–496.

Ball, H. 1944 *The Law of Copyright and Intellectual Property,* cited in S. Freid, Fair use and the new act. Pp. 465–487 in G. P. Bush and R. H. Dreyfuss, eds., *Technology and Copyright: Sources and Materials.* Mt. Airy, Md.: Lomond Books.

Beebe, G. W. 1981 Record linkage and needed improvement in existing data resources. Cancer: Branbury Report 9. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory.

1980 Problems of long-term recordkeeping. In *Issues in Research with Human Subjects.* (NIH Pub. No. FIC80-1858). Washington, D.C.: Department of Health, Education, and Welfare.

Boruch, R.F., and Cecil, J.S., eds. 1983 *Solutions to Ethical and Legal Problems to Social Research.* New York: Academic Press.

Boruch, R.F. and Cecil, J.S. 1979 *Assuring the Confidentiality of Social Research Data.* Philadelphia: University of Pennsylvania Press.

Braunstein, Y.M., Fischer, D.M., Ordoner, J.A., and Baumol, W.J. 1979 Economics of property rights as applied to computer software and data bases. Pp. 235–246 in G. P. Bush and R. H. Dreyfuss, eds., *Technology and Copyright: Sources and Materials.* Mt. Airy, Md.: Lomond Books.

Braverman, B.A., and Heppler, W.R. 1981 A practical review of state open records laws. *George Washington Law Review.* 49:720–760.

Broad, W.J. 1982 Court upholds privacy of unpublished data. *Science* 216(April 2):34–36.

Campbell, D.T., Boruch, R.F., Schwartz, R.D., and Steinberg, S. 1975 Confidentiality-preserving modes of access to files and to interfile exchange for useful statistical analysis. Appendix A in A. Rivlin, ed., *Protecting Individual Privacy in Evaluation Research.* Report of the Committee on Federal Agency Evaluation Policy. Washington, D.C.: National Academy of Sciences.

Comment 1976a Access to information? Exemption from disclosure under the Freedom of Information Act and the Privacy Act of 1974. *Willamette Law Journal* 13:135–171.

1976b The Freedom of Information Act's privacy exemption and the Privacy Act of 1974. *Harvard Civil Rights—Civil Liberties Law Review* 11:596–631.

1977a Copyrighted compilations of public domain facts in a directory: the criterion of infringement. *Northwestern University Law Review* 71:833–842.

1977b The misappropriation doctrine after the Copyright Revision Act of 1976. *Dickinson Law Review* 81:469–493.

1981 Applying the Freedom of Information Act to tax return information, *Georgetown Law Journal* 69:1283–1307.

1982 Copyright law—will the denial of a copyright to an author's research impede scholarship? *Western New England Law Review* 5:103ff.

Commission on Federal Paperwork 1977 *Confidentiality and Privacy.* Washington, D.C.: U.S. Government Printing Office.

Connelly, M.Q. 1981 Secrets and smokescreens; a legal and economic analysis of government disclosures of business data. *Wisconsin Law Review* 1981:207–273.

Curran, W.J. 1978 The privacy protection report and epidemiological research. *American Journal of Public Health* 68:173–176.

Davidson, J.H. 1976 The Privacy Act of 1974—exceptions and exemptions. *Federal Bar Journal* 34:323–329.

Dawber, T.R. 1980 *The Framingham Study: The Epidemiology of Atherosclerotic Disease.* Cambridge, Mass.: Harvard University Press.

Denicola, R.C. 1981 Copyright in collections of facts: a theory for the protection of nonfiction literary works. *Columbia Law Review* 81:516–542.

Dickson, D. 1980 Research data: private property or public good? *Nature* 284:292.

Easterbrook, F.A. 1980 Privacy and the optimal extent of disclosure under the Freedom of Information Act. *Journal of Legal Studies* 9:775–800.

Eastman, H.B. 1975 Enforcing the right of privacy through the Privacy Act of 1974. *Federal Bar Journal* 34:335–339.

Ehrlich, I. 1975 The deterrent effect of capital punishment: a question of life and death. *The American Economic Review* 65:397ff.

Federal Judicial Center 1981 *Experimentation in the Law: Report of the Federal Judicial Advisory Committee on Experimentation in the Law.* Washington, D.C.: U.S. Government Printing Office.

Fennell, B.A., and Hall, W.N. 1980 Due process at sentencing: an empirical and legal analysis of the disclosure of presentence reports in federal courts. *Harvard Law Review* 93:1615-97.

Flaherty, D.H. 1979 *Privacy and Government Data Banks: An International Perspective.* London: Mansell.

Fried, S. 1979 Fair use and the new act. Pp. 465–487 in G.P. Bush and R.H. Dreyfuss, eds., *Technology and Copyright: Sources and Materials.* Mt. Airy, Md.: Lomond Books.

Garner, J. 1981 National Institute of Justice: access and secondary analysis. Pp. 43–49 in R.F. Boruch, P.M. Wortman, and D.S. Cordray, *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Gordis, L., Gold, E., and Seltser, R. 1977 Privacy protection in epidemiological and medical research: a challenge and a responsibility. *American Journal of Epidemiology* 105:163–168.

Hammond, R.G. 1981 Quantum physics, econometric models and property rights to information. *McGill Law Journal* 27:47–72.

Holden, C. 1981 Dark days for social research. *Science* 211(March 27):1397.

1982 Statistics suffering under Reagan. *Science* 216(May 21):833.

1975 Confidentiality of statistical and research data and the Privacy Act of 1974. *Statistical Reporter* (June):197–209.

Hulett, M. 1975 Privacy and the Freedom of Information Act. *Administrative Law Review* 27:275–294.

Kelsey, J.L. 1981 Privacy and confidentiality in epidemiological research involving patients. *IRB: A Review of Human Subjects Research* 3(February):1–4.

Keplinger, M.S. 1977 Computer intellectual property claims: computer software and data base protection. *Washington University Law Quarterly* 1977:461–467.

Kronman, A.T. 1980 The privacy exemption to the Freedom of Information Act. *Journal of Legal Studies* 9:727–800.

Lewis, A. 1980 A public right to know about public institutions: the First Amendment as a sword. *Supreme Court Review* 1980:1–25.

Martin, M.E. 1974 Statistical legislation and confidentiality issues. *International Statistical Review* 42:265-7.

McGarity, T.O., and Shapiro, S.A. 1980 The trade secret status of health and safety testing information. *Harvard Law Review* 93:837–888.

Mochmann, E., and Muller, P.J. 1979 *Data Protection and Social Science Research.* Frankfurt: Campus Verlag.

Morris, R.A., Sales, B.D., and Berman, J.J. 1981 Research and the Freedom of Information Act. *American Psychologist* 36:819–826.

Mosteller, F. 1981 Taking science out of social science. *Science* 212(April 17):291.

National Commission on Research 1980 *Funding Mechanisms: Balancing Objectives and Resources in University Research.* Washington, D.C.: National Commission on Research. National Science Board 1981 *Science* Indicators 1980. Washington, D.C.: National Science Foundation.

Nelkin, D. 1982 Intellectual property: the control of scientific information. *Science* 216(May 14):704–708.

Nelson, R., and Hedrick, T. 1983 The statutory protection of confidential research data: synthesis and evaluation. Pp. 213–236 in R.F. Boruch and J.S. Cecil, eds., *Solutions to Ethical and Legal Problems in Social Research.* New York: Academic Press.

Nimmer, M. 1980 *Nimmer on Copyright: A Treatise on the Law of Literature, Artistic and Musical Property and the Protection of Ideas* (rev. ed.). Albany, N.Y.: Matthew Bender.

Norman, C. 1983a Administration relents on social science funds. *Science* 219(March 4)1048–1049.

1983b Congress looks fondly on science and technology. *Science* 221(July 15):246.

Note 1975 Administrative law—Freedom of Information Act—personal information exempted from disclosure—*Wine Hobby, USA* v. *IRS. Boston College Industrial and Commercial Law Review* 16:240–254.

1976a The Privacy Act of 1974: an overview and critique, 1976. *Washington Law Quarterly* 1976:667–718.

1976b The Privacy Act of 1974: an overview, 1976. *Duke Law Journal* 1976:301–329.

1976c Freedom of Information Act—Exemption (4)—research designs contained in grant applications—*Washington Research Project, Inc.* v. <Department of Health, Education and Welfare. *Boston College Industrial and Commercial Law Review* 17:91–106.

1978 Applying the Freedom of Information Act in the area of federal grant law: exploring an unknown entity. *Cleveland State Law Review* 27:294–311.

1981 The definition of "agency" under the Freedom of Information Act as applied to federal consultants and grantees. *Georgetown Law Journal* 69:1223–1255.

1982 The applicability of the Freedom on Information Act's disclosure requirements to intellectual property. *Notre Dame Lawyer* 57:561–579.

Office of Federal Statistical Policy and Standards 1978 *Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure-Avoidance Techniques.* Washington, D.C.: U.S. Department of Commerce.

1980a *Statistical Policy Working Paper 5: Report on Exact and Statistical Matching Techniques.* Washington, D.C.: U.S. Department of Commerce.

1980b *Statistical Policy Working Paper 6: Report on Statistical Uses of Administrative Records.* Washington, D.C.: U.S. Department of Commerce.

Office of Management and Budget 1975 Privacy Act implementation guidelines and responsibilities. *Federal Register* 40:28948ff.

O'Neill, H.V., and Fanning, J.P. 1976 The challenge of implementing and operating under the Privacy Act in the largest public sector conglomerate—HEW. *Bureaucrat* 5:171–188.

Passell, P., and Taylor, J.B. 1977 The deterrent effect of capital punishment: another view. *American Economic Review* 67:445ff.

Patton, W. 1980 *An Author's Guide to the Copyright Law,* 31–32, 84–85. Lexington, Mass.: D.C. Heath § Co.

Posner, R.A. 1979 Information and antitrust: reflection on the *Gypsum* and *Engineers* decisions. *Georgetown Law Journal* 67:1187–1203.

Prewitt, K., and Sills, B. 1981 Federal funding for the social science: threats and responses. *Items* 35 (September):33ff.

Privacy Protection Study Commission 1977 *Personal Privacy in an Information Society.* Washington, D.C.: U.S. Government Printing Office.

Project 1975 Government information and rights of citizens. *Michigan Law Review* 73:791–1339.

Riecken, H.W., and Boruch, R.F. 1974 *Social Experimentation.* New York: Academic Press.

Robins, L.N. 1978 The Consequences of the Recommendations of the Privacy Protection Study Commission for Longitudinal Studies. Paper presented at the Life History Research in Psychopathology Meeting, Cincinnati, Ohio.

Rozsa, G., and Foldi, T. 1980 International cooperation and trends in social science data transfer. *UNESCO Journal of Information Sciences, Librarianship and Archive Administration* 2:234–239.

Sasfy, J.H., and Siegel, L. 1982 *A Study of Research Access to Confidential Criminal Justice Agency Data.* Washington, D.C.: The MITRE Corporation.

Sprehe, J.T. 1981 A federal policy for improving data access and user services. *Statistical Reporter* 81(March):323–344.

Squires, J. 1979 Copyright and compilations in the computer era: old wine in new bottles. Pp. 205–234 in G.P. Bush and R.H. Dreyfuss, eds., *Technology and Copyright: Sources and Materials.* Mt. Airy, Md.: Lomond Books.

Stevenson, R.B., Jr. 1982 Protecting business secrets under the Freedom of Information Act: managing Exemption 4. *Administrative Law Review* 34:297-261.

Teitlebaum, L.E. 1983 A positivist approach to law and social science research. Pp. 11–48 in R. F. Boruch and J. S. Cecil, eds., *Solutions to Ethical and Legal Problems to Social Research.* New York: Academic Press.

Thompson, G.B. 1979 *Memo From Mercury: Information Technology is Different.* Montreal: Institute for Research on Public Policy.

Professional Codes and Guidelines in Data Sharing

Robert F. Boruch and David S. Cordray

INTRODUCTION

This paper reviews available information about professional codes and guidelines that are pertinent to data sharing. Our working definition of codes here includes statements of principle, conduct, or rule that bear on the rights and responsibilities of the parties involved in data sharing. Parties at interest here include primary and secondary analysts and the professional societies and associations to which they may belong; federal, state, and local agencies that sponsor or conduct research; and the editors of professional journals. While this is not a complete listing of those involved in data sharing, nor is the definition of professional practice satisfactory, we believe it suffices for this discussion.

Robert F. Boruch is a professor in the Department of Psychology and the School of Education and codirector of the Center for Statistics and Probability, Northwestern University. David S. Cordray is an assistant professor of psychology in the Division of Methodology and Evaluation Research, Department of Psychology, Northwestern University.

Background research for this paper was supported by a stipend from the National Science Foundation to Northwestern University, Center for Statistics and Probability.

International codes are discussed first, followed by specialized codes generated by disciplinary associations. Laws and regulations constitute a backdrop for any professional code, so they too are discussed briefly. Some of the illustrations of actual practice mentioned here are discussed in greater detail in the other papers in this volume. The conclusion of the paper is dedicated to a brief discussion of the adequacy of professional guidelines.

The discussion in this paper focuses on codes that concern data sharing. These codes often include provisions regarding privacy of individuals or of institutions, standards for reporting, and other related topics; they are given far less attention here than those bearing directly on data sharing. More thorough reviews of privacy-related codes and guidelines are discussed else-where, e.g., Boruch and Cecil (1979).

INTERNATIONAL COMMITTEES AND ORGANIZATIONS

Bellagio Principles

The broadest set of guidelines on data sharing appear to be the so-called Bellagio principles (see Exhibit A). The principles evolved from a conference among university and government scholars and bureaucrats from five countries: the United States, United Kingdom, West Germany, Sweden, and Canada. The meetings, organized by historian and lawyer David Flaherty at Bellagio, Italy, resulted in a statement of 18 general principles on which all participants could agree (see Flaherty, 1978). The principles have been published in at least five journals. Though they have no legal standing, they do serve as a framework for international agreements and codes of good practice.

The Bellagio principles endorse the idea of provision of government data to individual researchers or research institutions for legitimate research purposes. They were not designed to apply to the individual researcher's sharing his or her data with others, but might be considered for adoption to this as well. They cover three general aspects of data sharing: the conditions under which sharing should occur, modes of data release, and responsibilities of those who receive data. With respect to conditions, the principles endorse the idea of the broadest practicable access to government data by nongovernment researchers or research organizations, recognize the legitimacy of limited constraints on access necessary to achieve a balance between public and researcher interests, and endorse the idea of statutory privilege for data collected primarily for research purposes. Eight of the principles focus on modes of data release. They include suggestions about distribution of data at the lowest possible level of aggregation when microdata are required and distribution of public-use sample tapes with any individual identifiers removed.

Two principles consider the fact that records on identifiable individuals and complete data rather than public-use samples are essential for certain kinds of research. The use of special techniques to protect against deductive disclosure, through customized user service procedures, for example, is acknowledged. One principle addresses the matter of linking files from independent sources when privacy is an issue and one considers the distinction between administrative and research records.

The remaining principles focus attention on the responsibilities of researchers and other parties. They address the need for statisticians and researchers to contribute to policy and legal definitions of privacy and enumerate simple ways to meet public concerns about privacy and confidentiality on the collection and utilization of individual data. (Those ways include informed consent, public education, and provisions for public knowledge of data uses, among others.) One principle urges professional societies to devise codes of conduct. Another is devoted to ensuring that access is not discriminatory and that appeals processes are available in the event of conflict over access. The final principle places responsibility for proper conduct on users of microdata by encouraging researchers to sign written agreements for the protection of confidentiality.

Statements of Government-Related International Organizations

Various government-related international organizations have considered, although not necessarily adopted, guidelines on transnational data flow. Many of those guidelines are very general in that they do not distinguish between commercial exchanges or use of data and exchanges of research data. Many existing guidelines are less relevant for physical science and engineering data since those data do not concern records on identifiable individuals and the guidelines stem invariably from concerns about privacy and confidentiality. The practices and procedures specified by three organizations are summarized briefly here.

The Organization for Economic Cooperation and Development (OECD) has accepted for consideration a set of guidelines on data protection submitted by the United States. The guidelines constitute a set of principles of fair practice regarding records on individuals, and they are designed primarily to protect individuals' interest within member states. So, for example, the principles state that individuals should have a right of access to their records in personal data recordkeeping systems and right of correction and that there should be explicit limits on method of collection, use, and external disclosure of records. Personal implications of transnational data flows are considered in recommendations that privacy law or policy be created and enforced and that

codes of conduct be fostered.

The OECD statements that are most pertinent to data sharing include the following (Organization for Economic Cooperation and Development, 1981):

(5) Governments should undertake to ensure to the greatest extent possible, the uninterrupted free flow of information.

(6) Governments should undertake to avoid the unjustified disruption of international trade patterns and the creation of nontariff barriers that would interfere with transnational data flow.

(7) Government should refrain from restricting the import and export of data unless doing so is essential to national security.

The other statements call for more cooperation on policy and law governing individual privacy. Though the nature and function of the data being considered is not explicit, the draft material supplied by expert groups justifies interest in the area by arguing that transborder movements of personal data are essential to economic, scientific, educational, and social development (Katzan, 1980:148).

The Council of Europe's preliminary draft convention on automated records focuses on maintenance and privacy protection for automated administrative records. There is no explicit distinction between research and administrative functions of records, no reference to sharing such data for research, and no recognition of exchanges among researchers. The emphasis *appears* to be on commercial uses despite the lack of distinctions (e.g., a German firm's processing an Italian firm's records or vice versa).

The European Science Foundation's (1980) “Statement Concerning the Protection of Privacy and the Use of Personal Data for Research” contains six basic principles dealing with privacy. The statement itself was drawn up partly in reaction to OECD and Council of Europe directives that emphasized restrictions on data access in the interests of institutional and individual privacy. Section 1.4 of the statement (p. 5) directs special attention to data sharing: “Freedom of research presupposes the broadest possible access to information. Legislation should therefore, besides specifying the conditions under which personal data may be used for research, ensure access to the information needed.” Reuse of such data is considered in two sections (2.10, 2.11) that bear on secure storage in a centralized research archive.[1]

As can be seen, the statements considered or issued by international organizations such as the ones discussed here vary greatly. Except for the statement of the European Science Foundation, they do not recognize any special characteristics of scientific data or records and provide no special guidance on sharing this kind of information. This omission is likely to lead to problems, as evidenced by difficulties engendered by Sweden's Privacy Act of 1973 and the U.S. Privacy Act of 1974: needless restrictions on collection and disclosure of records whose function is solely research (see Mochmann and Muller,

1979).

INFLUENCE OF LAW ON DATA SHARING

Law and government rules within a country constitute broad limits on the conduct of researchers. They can impede or enhance the sharing of data generated in individual projects. More generally, they affect the extent to which independent researchers can obtain access to records maintained by government for research, policy, or management purposes. The influence of the U.S. Privacy Act and Freedom of Information Act (examined by Cecil and Griffin, in this volume) and of the Tax Reform Act and other legislation illustrate the complex nature of problems. Here we confine attention to several recent studies of relevant law in developed countries. Rules and guidelines issued by specific agencies are considered later in the paper.

Flaherty's (1979) study, supported by the Ford Foundation, examined five countries: Britain, the United States, Canada, Germany, and Sweden. It covered the legal framework and theory underlying privacy law for each country, paying special attention to legal restrictions on access to government data by social scientists. The legal and institutional mechanisms for disseminating public-use tapes and other microdata are described. The examination of U.S. rules is based on the law, regulation, and practices of the National Center for Health Statistics and the Office of Research and Statistics of the Social Security Administration (both a part of the U.S. Department of Health and Human Services).

The Mochmann and Muller (1979) monograph includes reports from the five countries covered by Flaherty as well as Norway, Denmark, Italy, Holland, and Belgium. Both of these country reports give a general description of privacy legislation and law governing researchers' access to data and indicate whether such laws distinguish between administrative and statistical records, define anonymous and identifiable records, and provide for researchers' use of data (including for sampling of individuals as a data base, linkage with other archives, etc.). Law, regulation, or practice on conditions of access are also discussed for each country.

One of Flaherty's (1979) major conclusions is that statistical information is often guarded on privacy grounds to such an extent that research needs in many fields are not being met. His conclusions and recommendations cover policy, rules, and regulation, public relations, dissemination of data, and other matters.

The largest compendium of state laws that bear on access to data for research purposes has been developed by Robbin and Jozefacki (1982). It covers vital, health, and social services records, the laws governing privacy of individuals on whom records are kept, and researchers' access to those records.

About 350 statutes are included for all 50 states. Work by Sasfy and Siegel (1982) is similar in spirit but focuses on criminal justice agencies and their record practices, including disclosure of police, court, and other records to researchers. It also covers all the states and more than 130 agencies.

As one might expect, attention to disclosure for research purposes varies considerably. Most privacy-related statutes do *not* include explicit provision for access by researchers. The minority that do most often provide for access by delegating access authority to an agency director's discretion. According to agency officials interviewed in these studies, they base their judgments on law and the nature of the research for which the records are requested (e.g., quality of research design and relevance to the agency's mission).

Some laws are explicit, partly as a result of federal models. California's Information Practices Act, for example, permits legitimate researchers access to medical, psychiatric, and psychological records, provided that the identified information is essential to the research; further disclosure is forbidden. Similarly, researchers involved in mental health work are granted access under certain restrictions to state agencies providing relevant services.

The works by Robbin, Sasfy, and others demonstrate that there are many fewer federal and state laws that provide for researcher access to data than there are laws that govern privacy of individuals, collecting and storing information, and so on. When disclosure is made legally possible, it is most often discretionary. Laws are only occasionally explicit in specifying that records may be disclosed for legitimate research purposes; however, most researcher requests appear to be honored. To judge by Robbin's (1982) surveys, few archivists are aware of the laws that provide access; to judge by lack of coverage of the topic in such journals as *American Statistician, American Sociologist,* and others, many professional groups may also not be informed.

PROFESSIONAL SOCIETIES AND ORGANIZATIONS

A variety of professional organizations and societies have issued codes of ethics or professional conduct that address data sharing at least indirectly. This section discusses the extent to which societies have explicitly acknowledged data sharing practices and summarizes the character of standards, guidelines, or codes issued by societies or professional groups.[2]

American Association for the Advancement of Science Professional Ethics Project

In December 1980 the American Association for the Advancement of Science (AAAS) issued a report on professional ethics activities in scientific and engineering societies affiliated with AAAS (Chalk et al., 1980). At the time of

the survey, 241 science and engineering societies were affiliated with AAAS. The data reported by Chalk et al. concern roughly 74 percent of the societies and cover a broad range of disciplines and society characteristics (large and small membership, new and established, etc.). While there is some ambiguity as to the number of societies that have adopted ethical rules or codes of conduct, it appears that between 50 and 60 societies have either done so or have issued advisory opinions.

Chalk et al. (1980) identify 191 distinct rules of conduct. Appendix J of their report enumerates statements appearing in these documents and the frequency of each. We have analyzed the contents of 74 statements issued by 57 societies to provide a crude characterization of the extent to which data sharing is considered by professional societies, and we reproduced below the ones that are relevant to data sharing (the frequency with which each statement is given in parentheses; the most pertinent statements are italicized):[3]

> Members shall disseminate knowledge and share experience with other colleagues and be honest, realistic and clear in presenting findings. (19)
>
> Members shall avoid and/or discourage sensational, exaggerated, false and unwarranted statements. (21)
>
> Members shall refrain from or exercise due care in criticizing another professional's work in public, recognizing that the Association provides a proper forum for technical discussion and criticism. (2)
>
> Members should not communicate their findings secretly to some and withhold them from others. (1)
>
> *Members should clarify in advance with employers or sponsors expectations for sharing and utilizing data and/or the ownership of materials or patents.* (9)
>
> *Funding agencies should include in grants a stipulation that data gathered under the grants be made available to scholars at cost after a specific time.* (1)
>
> Members shall protect clients from the misuse of information collected about them. (1)
>
> Members shall respect the privacy of their clients. (1)
>
> Information gained from research participants shall be held in confidence unless the subject's consent to release information is obtained. (5)
>
> Solicitation of research subjects should make clear the obligation, rewards, and consequences to research subjects for their participation. (4)

As can be seen, the most frequently appearing statement in codes are those directed at honesty and balanced reporting (statements 1 and 2); 40 of the 57 societies offer some advice on this matter. The frequency of sanctions against criticism or concealment, reflected in statements 3 and 4, are considerably fewer. Of particular relevance to the topic of data sharing are statements 5 and 6. Only 10 instances of statements pertinent to sharing are reported despite the frequency with which honesty and balanced reporting are advocated. The remaining statements apply to privacy and confidentiality and are more frequent, judging from the list in Chalk et al. Each of them emphasizes the roles and responsibility of the research practitioner.

In the codes of conduct issued by professional societies, less explicit emphasis is placed on conditions for release of information. Rather, the stress is on conducting honest and objective research. The completeness of any given professional society statement cannot be assessed using the Chalk et al. report, but the data are available for reanalysis. (The report includes excerpts from ethics statements of a selected group of professional societies.)

Guidelines Bearing on Statistical Research and Data Sharing

Since 1980 four professional groups have issued standards and guidelines bearing at least partly on sharing statistical research data. They are remarkable in that each dedicates explicit attention to providing access to data used as a basis for reports. The Joint Committee on Standards for Educational Evaluation (1981) issued professional standards and guidelines for evaluating educational programs, projects, and material; the Evaluation Research Society (1980) recently issued a parallel document for a wide variety of disciplines, including education (also see Rossi, 1982); the American Statistical Association (1980; 1983) has independently issued a draft code of conduct bearing on the topic of data sharing (also see Ellenberg, 1983); and the American Sociological Association (1982) issued a draft code of ethics to its members.[4]

The Evaluation Research Society (ERS) (1980) explicitly states (guideline number 7) that any restrictions on access to data generated as part of an evaluation should be established at the outset. Similarly, the Joint Committee on Standards for Evaluation (1981) acknowledges the need to negotiate access to data as part of the planning process. The ERS standards are more explicit than the joint committee on when access is and is not negotiable. Specifically, access is not negotiable when the evaluation is subject to conditions specified by the Freedom of Information Act, or it is understood that results are in the public domain. The ERS standards note that the sponsor or the evaluator are obligated to point this out. For privately sponsored research, “the client may rightfully expect confidentiality of the findings to be maintained” (ERS, guideline 7). To facilitate reanalysis of data to which access has been obtained, the ERS standards specify that a description of analysis procedures, including their assumptions and relevance to the data, should be provided (see guideline 32). Documentation should be sufficient to make the analysis replicable (guidelines 36 and 49), and methods and circumstances of data collection should be recorded for each data item (guideline 29). As in the standards report on impact evaluations of the U.S. General Accounting Office (1978), the persons responsible for release of the data should be identified.

The ERS and joint committee standards are similar in their treatment of the issues related to access and the factors that facilitate or impede access (Cordray, 1982). They differ in organization and detail. The joint committee statements bearing on these issues are spread throughout the volume, reflecting a need to consider access and limitations on access in the research design, during data collection and processing, and in reporting. There are explicit guidelines addressing access to data records (C5-1); identification of right-to-know audiences to whom summary information is to be provided (A6-B, A6-C); agreement on to whom identified data should be released (A6-H); release of evaluation procedures, data, and reports so that they can be examined (judged) by other independent evaluators (C2-2 and D3-I); and a general statement (D4-G) regarding making data, procedures, and records of analysis available to *responsibly planned* reviews. When anonymity is promised, procedures are to be devised to protect subject anonymity (C5-J). A caveat acknowledges the need to avoid making promises of confidentiality when it cannot be guaranteed and to avoid a guarantee that information will not be used beyond its stated purposes when there is the possibility that it may be released (e.g., through a court order). Under accuracy guidelines, analysts are urged to adopt and implement standard procedures for storing and retrieving data (D7-D) and to implement checks for errors, processing, and reporting data (D7-E), and weaknesses in the data are to be described and their impact on conclusions assessed (D8-B). Seven guidelines are offered on what should be reported (D4-A to D4-G).

The ERS guidelines explicitly acknowledge the need to identify those individuals who are authorized to release the data. In this respect, they are similar to guidelines issued 2 years earlier by the U.S. General Accounting Office (GAO) (1978) for assessing quality of federal program evaluations. The GAO also suggests that data files, stripped of identifiers, should be released as soon as possible after an evaluation is completed. The joint committee is less direct, treating this as a point of negotiation unless contractual or legal constraints apply. The joint committee hedges a bit by prefacing its recommendation with the phrase “make available for responsibly planned reviews” (p. 108), implying that some requests could be justifiably denied. All of the guidelines share a concern that pledges of confidentiality be offered only when it is necessary, that researchers should not make assurances of confidentiality that cannot be honored, and that restrictions on access due to pledges of confidentiality should be avoided.

The committee on Code of Conduct of the American Statistical Association (ASA) has proposed an interim set of guidelines on a 3-year trial basis (see *Amstat News,* 1981; *American Statistician* 37(1), 1983). Several items in the guidelines bear directly on data sharing and are summarized below (American Statistical Association, 1983).

First, the code recommends that statisticians "make data sources available for analysis by other responsible parties with appropriate safe-guards for privacy concerns" (p. 6). Second, it recommends that "statisticians establish their intentions where pertinent to protect the confidentiality of information … to ensure that the means are adequate to protect confidentiality to the extent pledged … and to insure that transfers of data are in conformity with pledges" (p.5). Third, it recommends that the statistician document data sources used in an inquiry and known inaccuracies. These American Statistical Association proposals articulate the spirit of suggestions made by Bentley Glass (1965) to the general scientific community. His view is that the scientist is obligated to "publish his methods and his results so clearly and in such detail that another may confirm and extend his work" (p. 1258). For some sciences, it is only by getting hold of the raw data that confirmation and extension are possible.

Some of the commentaries on the ASA guidelines pertain to data sharing and reanalysis. For E.A. Gehan (1983), a biostatistician, a crucial ethical concern is whether certain subgroups of patients in clinical trials are analyzed—the subgroups one compares may produce very misleading estimates of the effect of a drug or surgical technique. Bross (1983) registers a related concern that major government-sponsored research is analyzed in ways that produce artificially favorable results. Mosteller's (1983) concern, which is also related, is that the guidelines should not lead to relaxation of vigilance against fraudulent activity by the statistician. None of the commentators recognizes that the ASA guidelines may enhance an independent analyst's ability to reanalyze data to produce a less misleading or at least a more balanced view.

Mosteller notes the ostensible internal conflict between a guideline that advocates disclosure of data and one that warns against disclosure of a client's private information. For Kish (1983), the ostensible conflict lies between a guideline that urges researchers to collect only information that is necessary and the contemporary emphasis on omnibus surveys and data banks, in which what is "necessary" often cannot be clear for some time.

The American Sociological Association issued a draft code of ethics in 1980 and a revised code in 1981. Its membership voted approval of a final draft in 1983; enforcement procedures are still under review. The code maintains (American Sociological Association, 1982:2):

> Sociologists are obligated to report findings fully and without omission of significant data … disclose details of theory, methods, and research designs that might bear upon interpretation of research findings….
> Consistent with the spirit of full disclosure of method and analysis, sociologists should make their data available to other qualified social scientists at reasonable

> costs, after they have completed their analyses, except in cases where confidentiality or the claims of a field worker to the privacy of personal notes would be violated in doing so. The timeliness of this obligation is critical especially where the research is perceived to have policy implications.

The code is remarkable in several respects. For instance, timeliness of disclosure is recognized by no other codes that we are aware of. Other statements in the code make it plain that data generated in other countries, as well as in the United States, should be stripped of identifiers and made available for reanalysis. It also appears to be the only code that is explicit about methods for ensuring privacy of respondents (American Sociological Association, 1982):

> To the extent possible … researchers should anticipate potential threats to confidentiality. Such means as removal of identifiers, the use of randomized responses, and other statistical solutions to problems of privacy should be used where appropriate.

The proposed code is terse, but covers a variety of other topics.

Arguments Against Codes

Whether the various science organization should adopt a code of conduct at all, much less one that takes a position on data sharing, has been debated periodically. During the 1950s and 1960s, the arguments against codes included the view that scientific ethics although not codified are adhered to nonetheless and that unwritten ethics should not be codified and put to a vote (Lanz, 1963; Fosberg, 1963). Opponents of guidelines argued that such codes are usually created for legalistic reasons, and the remoteness of science from legal settings obviated this justification for codes (Cranberg, 1963). The recent court cases over access to data (such as *Forsham* v. *Harris*), government suspension of grants in cases of research fraud in medical research, and similar problems imply that this argument is no longer true (see Cecil and Griffin, in this volume).

The recent comments on the American Statistical Association guidelines by many prominent statisticians are also instructive in this respect. For many, the guidelines are a promising and fundamental vehicle for education in the profession, a set of reminders about what traps one ought to be aware of and how to avoid them (Martin, 1983; Gehan, 1983; Rice, 1983; Mosteller, 1983). For some, however, acceptance of the idea of guidelines is reluctant: they would have hoped guidelines to be unnecessary (Greenhouse, 1983; Kish, 1983). Still others believe such guidelines are, at best, gratuitous and will, at worst, be dangerous insofar as they invite sanctions against a politically unpopular view (Solomon, 1983) or detract from the development of personal

ethics, courage, and action in risk-laden work (Bross, 1983).

At least a few individuals, such as W.E. Deming (1972), have developed their own written code of conduct as part of their professional consulting activity. Other statisticians may not have written personal codes, but are clear in their public views about what codes can do. In commenting on the American Statistical Association's interim code, for example, Kish (1983:17), a sampling statistician, said that the guidelines could do "perhaps a little good and no appreciable harm." Irwin Bross (1983:13), a biostatistician, is at least as direct and goes a bit further (in this as in other published papers) to maintain that each statistician is responsible for deciding what is "the right action to take in the face of ethical challenge" especially because the right action may invite strong criticism from a variety of public interest groups. These ideas and the fact that many scientists whose professional organizations have no pertinent code still share data (e.g., economists) strengthen the notion that general codes are not sufficient. They are likely to be useful for cases that involve "frequently discordant … highly competitive social endeavors" (American Association for the Advancement of Science, 1960).

Specific Data-Sharing Professional Standards and Guidelines

The preceding description of professional societies focused on general statements of ethical conduct. All of them cover professional activities in addition to data sharing. This section discusses codes that direct primary attention to acquiring, processing, and sharing data.

Archivists have had an abiding interest in preserving and making available the materials in their custody. Those materials include machine-readable records although manuscripts, hard-copy records, photos, and the like are far more common. A code proposed recently by the Society of American Archivists (SAA) (1980) applies to all such materials. (A variation on the code, produced by Robbin (1978), focuses on machine-readable records.)

The SAA code is distinctive in more than a few respects. It is specific in encouraging quality control and use of information and in advising members to ensure that materials are placed in repositories where they will be "adequately processed and effectively utilized" (p. 414). Furthermore, the archivist should "encourage use of (holdings) to the greatest extent compatible with institutional policies, preservation of holdings, legal considerations, individual rights, donor agreements …" (p.414). It is unique too in attending to the value of materials, maintaining that archivists must "appraise records and papers with impartial judgment" (p.414), a guideline that in principle is applicable to machine-readable records, researchers' field notes, etc.

The SAA code of ethics considers timeliness of organization and distribution

of holdings, as the American Sociological Association is concerned with timely disclosure, and reiterates the need to avoid delay in the commentary that follows it (Society of American Archivists, 1980:416). Archivists are enjoined to respect the privacy of individuals on whom records are maintained, just as other codes dealing with access to research data do. However, the commentary implies that privacy refers only to living individuals, and institutions are not mentioned. Finally, the code says that the archivist takes responsibility for being informed and for informing donors of laws (e.g., copyright and tax) and provisions of access and the like, guidance that runs parallel to requirements that social researchers be aware of and inform their respondents about confidentiality and privacy provisions.

Alice Robbin (1978) has proposed a code for data archivists that is oriented toward storing, processing, and distributing research data. It covers, among other things, the responsibility of the archivist: in protecting subject records and interests of the subject; for accountability; and for confidentiality, in placing limits on access, responsibility for release, and adherence to conditions of sharing.

Robbin (1981) also delineates numerous technical and nontechnical guidelines pertaining to what should be documented for machine-readable files, how files should be documented, and when documentation is necessary. The following remarks focus on the nontechnical guidelines pertaining to practices that facilitate data sharing and access to material necessary for understanding the data.

As part of the documentation of a machine-readable data file, Robbin's standards ask that the following be provided:

1. Description of the methodology employed in the study, including (a) sources of data, (b) universe or target population characteristics, (c) type of sample, (d) characteristics of instrumentation, and (e) date(s) of data collection.
2. Summary of the purpose or scope of the file, including subject matter, special characteristics, and number of variables.
3. Description of the terms of availability, including the condition of data, restrictions/limitations, and the name and address of a contact person.
4. History of the research project and data collection effort and rationale for collecting particular data.
5. File processing history and general editing strategies, index formulation, and software used to process the data file or conduct the analyses.
6. Detailed information on index construction, estimation processes, transformations, and other data manipulations.
7. Description of the known data errors, anecdotal information on processing idiosyncrasies of the analyst, and, if applicable, estimates based on alternative

analyses that make different assumptions about the structure and functional form of the relationship.

Beyond these methodologically oriented guidelines, Robbin proposes recommendations in anticipation that data will be shared. She focuses attention on the role of planning for data use by specifying issues that should be considered by sponsors of research and researchers prior to funding a research effort, including: designation of funds for specialists in file architecture, processing, and documentation; consultation with experts on confidentiality and privacy protection; monitoring and evaluation of the data acquisition, processing, and documentation throughout the course of the study; and development of multiple public-use data files, each of which may be relevant to a particular disciplinary area.

A second society with broad interest in acquiring, storing, and distributing numerical data for social science research is the International Federation of Data Organizations (IFDO). This group includes civil servants and university researchers from most Western European countries, Canada, and the United States. Their focus is on machine-readable archives, and their general mission is to facilitate use of data through meetings, publications, research, and other initiatives. The Bellagio Principles, discussed above, were formally endorsed at the IFDO meetings in Cologne, Germany, in 1978 (Flaherty, 1978).

A Special Case: American Society of Access Professionals

The American Society of Access Professionals (ASAP) was organized in 1980 as a forum for bureaucratic, scholarly, and legal discussions about access to information produced by government. Its members include civil servants, academic researchers, lawyers and managers, professional and government researchers, and others. Federal agency representation is considerable, including the Railroad Retirement Board, Food and Drug Administration, Social Security Administration, Small Business Administration, the military services, and the Departments of Commerce and Interior.

ASAP is a young organization and has not yet issued codes of conduct for its members or formal guidelines on data sharing. However, its annual meetings have included formal presentations and discussions of relevant research by Flaherty and others cited here, and of codes and guidelines suggested by Robbin among others. The coverage of issues is broader than most societies in that information shared may be numerical, narrative, or visual; may involve medical and engineering sciences as well as social and behavioral sciences; and covers management issues as well as law and public policy.

AGENCY AND INTERAGENCY REGULATIONS AND GUIDELINES

Research sponsors can take some responsibility for ensuring that data produced with government agency funds are available to other scholars. In the United States, several agencies have developed policies on data sharing; many other agencies have informal policies and treat data sharing on an ad hoc basis. Two agencies with formal policies are the National Science Foundation (NSF) and the National Institute of Justice.

National Science Foundation

The National Science Foundation's statement of grant conditions (1983:6) specifies the following:

> 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.

We agree with Clubb et al. (in this volume) in their general comment that NSF's statement is a significant step toward routinizing data sharing practice. They also noted some serious drawbacks to the NSF statement: there is no time frame specified as to when release should be carried out, the mode of release is not specified, and alternative forms of release beyond those that preserve privacy are not indicated.

While not ideal, the NSF statement contains elements that are noteworthy. In particular, the Foundation has made a specific statement as to its role in resolving conflicts, should they arise, about release *or* use of data or software. Also, the provision for release “on request” implies that interim data may be disseminated with the proviso that the grantee be given first right of publication. Whether this latter aspect is feasible is not clear.

National Institute of Justice

Joel Garner (1981:43–44) describes the special condition added to all grants at the time of award by the National Institute of Justice (NIJ):

> Upon grant termination, grantee agrees that computer-readable copies and adequate documentation of all data bases and programs developed or acquired in connection with the analysis in this project will be submitted to NIJ, at no additional costs. These may be used by the Government, or disseminated to others for their use, for any purposes deemed appropriate by NIJ, without further compensation to the grantee. The grantee shall make no guarantee that the data collected will not be transferred or released without the prior approval of the Institute. Consistent with 28 C.F.R., Part 22, the grantee must remove individual identifiers from any data bases and programs prior to submission to NIJ.

The NIJ policy is remarkable for its attention to the details of assuring access and for its early appearance in 1976. The initial policy, developed by NIJ staff, differs from the NSF statement in that it specifies when transfer is to be accomplished, makes provision for professional storage of the data, requires it to be documented in machine-readable form, and prohibits party agreements that hinder access by others. Similarly to the NSF statement, the NIJ policy acknowledges researcher's proprietary rights to data until a project has been completed. The possibility of early release, which may be necessary for timely reanalysis or simultaneous analysis (see Boruch et al., 1981), does not receive explicit or implicit mention in the NIJ policy.

President's Reorganization Project

The Federal Statistical System Project Report, a product of the President's Reorganization Project (1981), considers how federal statistical activities could be better organized to facilitate access to federal data. The project's recommendations recognize the need for a centralized policy on statistical activities in order to ensure high quality and accessible data across agencies. Among the mechanisms offered to facilitate access are the development of a central statistical office, a federal locator service, and customized user services.

Formal and Informal Agency Policy

The National Institute of Education and the evaluation unit of the U.S. Department of Education have encouraged the reanalysis of data by providing financial support and by fostering a de facto tradition of access to evaluative research data. The National Center for Health Services Research has encouraged reanalysis by aiding the development of a policy-data archive on long-term care experiments at Michigan State University, among other projects (e.g., Katz et al., n.d.).

These informal policies are admirable. But for several reasons it seems sensible to develop formal policy. First, secondary analyses are as susceptible

to pressure as original analyses. Access to data can be impeded by politicians, bureaucrats, and scientists. In the absence of a formal policy, occasional refusals to release data will continue. Furthermore, the normal turnover of staff of agencies, contractors, and advisory boards should not affect access to data, and we believe policy can have a stabilizing influence. Of course, any policy has to be monitored to ensure that it meets the needs of those who request data.

Regulations of Operating Agencies

Rules and regulations issued by federal operating agencies as a means of implementing legislation are also pertinent to data access and sharing for research purposes. For example, federal regulations on evaluating Title I compensatory education programs have required local education agencies and state education agencies to retain all of the data used to develop their reports for a period of five years or until any pending federal audit has been completed (*Federal Register,* 1979:44). For local education agencies, "all individual scores with an identifying code" are to be maintained. However, the regulations are *not* explicit as to who should bear fiscal responsibility for data storage or about the nature and scope of the documentation. Trochim (1982) was successful in securing data on Title I evaluations from such agencies to produce useful reports on the impact of Title I. But the information was often poorly documented and not in machine-readable form; considerable communication between parties was required in order to successfully use the acquired data.

As described in the paper by Boruch, other government agencies foster data sharing through a variety of contractual rules. For example, contracts issued by the National Center for Educational Statistics that are designed to support long-term longitudinal studies generally require the production of public-use data tapes by the contractor. Similar contract provisions have been created for large-scale evaluation studies, such as the graduated work incentive experiments supported by the U.S. Department of Health and Human Services in Seattle, Denver, New Jersey, and elsewhere, and the housing allowance experiments supported by the U.S. Department of Housing and Urban Development. State and local examples are much less visible and doubtless less frequent, although they do exist, but we have no hard information on them.

PROFESSIONAL JOURNALS AND POLICY ON DATA ACCESS

In preparing this paper, an effort was made to identify when and how journal policy is structured with respect to data access issues. Time and resources

did not permit an exhaustive review. We did uncover instances of explicit policies on access and other situations in which the issue has never arisen. The prevalence of either cannot be determined at this time.

Journal of Personality and Social Psychology

Anthony Greenwald's editorial policy for the *Journal of Personality and Social Psychology* makes plain his position on data access. Two aspects of data sharing are considered. First, to aid in the editorial evaluation of a manuscript, the author is instructed to supply one copy of the summary tables for the major analyses reported in the manuscript. The second aspect is more pertinent to data sharing (Greenwald, 1976:5):

> Submission of a research report to JPSP will be interpreted as an implicit assurance that the author has records of exact procedures and of data in unanalyzed form, and that both of these types of information shall be available to investigators who would like to replicate the research or reanalyze its data, respectively…. When a manuscript is accepted for publication, the author will be asked to provide assurance that (a) the data in unanalyzed form and the exact details of the procedures will be available to other investigators for at least 5 years after publication and (b) ethical problems have been handled in accordance with current APA code unless indicated otherwise in the published article.

Greenwald has since stepped down as editor of this journal. We were unable to determine in this case the extent to which the policy was implemented, nor were we able to determine if the data made available under his editorship was actually used by secondary analysts.

Journal of the American Statistical Association (JASA)

Stephen Fienberg's policy on data sharing while editor of the applications section of JASA is similar to Greenwald's policy: authors' submissions were to be accompanied by the data so that referees could “check calculations or carry out alternative ones” and to make them available from published manuscripts so that others could conduct reanalyses.

Fienberg notes certain obstacles (massive data sets, confidentiality restrictions) he encountered in instituting the policy, but notes that provisions for summary tables rather than microdata, or statements indicating that the data are available from the author, usually solve the size problem. He reports few instances in which authors were reluctant to comply with these requests.

As indicated earlier, the American Statistical Association Committee on Code of Conduct has proposed an interim code, to be followed for a 3-year trial basis. It is more explicit about what is to be documented than is Fienberg's policy, but the code of conduct does not apply to the process of

submitting articles to journals. This seems to be an area that is left to the discretion of the individual editors.

American Chemical Society Journal Practices

Authors of papers appearing in the *Journal of the American Chemical Society, Analytical Chemistry,* and others published by the American Chemical Society (ACS) can and often do make their data available for reanalysis. In particular, ACS provides a regular subscription to a supplement, which appears annually for the *Journal,* that contains auxiliary and raw data pertinent to a selection of published articles. The interested reader can also request supplementary material about a particular article. Both the annual supplement and the supplementary material are on microfiche.

According to Charles Birch, head of the journals department of the ACS, not all authors provide data to the journal for the supplement; the decision is made by the author and journal editor. So, for example, in 1974–1980 more than 13,000 articles were published by the *Journal of the American Chemical Society,* of which about 1,400 had supplements provided. *Analytical Chemistry* published 14 articles during the same period, and all were issued with supplements. Not all supplements are used or even requested. Of the 1,400 articles in 1974–1980 in the *Journal* for which material was available, there were requests for 235 articles. According to Birch, the rate varies considerably by journal, though, partly because of topic. The *Journal* often includes articles on crystallography, for which the request rate is not high. On the other hand, requests have been made for all supplements for all articles carrying them in *Analytical Chemistry.*

The practice of making supplementary material available is discussed briefly in the *Author's Handbook* for ACS publications, but there appears to be no readily accessible formal document on the topic or the history of the practice. Birch suggested that the development of a system of microfiche supplements came about because publication of raw data in the journals themselves was too expensive.

Journals With No Policy on Data Access

Of those journal contacted, a few reported not having an explicit policy on data sharing. Managing editor Robert V. Ormes of *Science* indicated that no written policy was in effect. Instead, authors are expected to disclose data should such a request be made. He suggested that *Science* and the American Association for the Advancement of Science would facilitate data access if a primary researcher was reluctant to release the data reported in a published manuscript. In his 20 years at *Science,* Ormes did not recall any instance in

which such action was necessary. The issue of confidentiality pertaining to ideas, findings, and procedures reported in unpublished manuscripts receives explicit attention by the staff of *Science* in its instructions to reviewers.

Discussion with various staff at the American Medical Association (AMA) and American Anthropological Association revealed no policy on data sharing for their major publications. In the case of the AMA, there is a policy on release of their own physician data but it does not apply to the data reported in the *Journal of the American Medical Association* or any of the specialty journals. For both associations, the staff suggested that individual editors may prescribe their own policies.

The American Economic Association (AEA) has adopted no formal code of ethics or conduct bearing on data sharing or related topics, such as privacy and confidentiality. Nor do the journals published under AEA auspices appear to devote any special attention to the matter. A major reason for the situation is that many articles are based on published statistical data. Major articles, however, are careful to cite references to each source (see the illustration describing Feldstein's work, in Boruch, in this volume) but problems of detail do appear (Bowers and Pierce, 1975, 1981). Whether reanalyses of earlier work are published depends on the journal editor, the quality of the paper, and other factors, just as it does in the other disciplines.

MONITORING COMPLIANCE WITH CODES

Some professional societies, agencies, regulations, and journal policies have advocated the need to make data available to other scholars. But with a few exceptions, procedures for handling violations are not specified. For example, one rule proposed in the Bellagio Principles is that data sharing should be equitable and that provisions should be made for hearing and adjudicating complaints of unfair practice or charges of unfair restrictions on data access. The principle does not state how this should be carried out and by whom. The National Science Foundation's policy is more explicit. It states that the NSF will resolve any conflicts among parties over access.

In their review of professional ethics for scientific and engineering societies, Chalk et al. (1980) provide some information about how violations are handled. Roughly 16 percent of the societies responding to their survey have appeals procedures and a variety of sanctions, the most frequently mentioned being expulsion, formal censure, and informal reprimand. Over a 10-year period, 76 societies have applied available sanctions 249 times; the most frequent kind (162) was an informal reprimand. These figures are for violations of any element in codes of ethics, and the reader should recognize that data sharing is a minor part of such codes, if it appears at all. The interesting

aspect of the Chalk et al. survey results is the infrequency with which action is possible, taken, or needed. Review by professional societies represents at least one option for monitoring compliance.

Failure to comply with specified standards and guidelines for conduct should not necessarily be viewed as a transgression. Because laws, regulations, public sentiment, and the like change over time, some codes may require modification. The Joint Committee on Educational Evaluation and the Evaluation Research Society, among others, maintain standing committees for review and modification of the guidelines they have issued.

EVALUATION AND MODIFICATION OF CODES

Modifications of standards and codes of conduct are necessary at times, and the organizations that have codes also have mechanisms for their change, at least in principle. For example, the American Statistical Association's Committee on Code of Conduct proposed set of practices and a 3-year trial period, which will presumably be reviewed during and at the close of the time period.

The following criteria for evaluating codes and standards of practice have been proposed by Chalk et al. (1980:51).

1. *Applicability*—This refers to the responsiveness of the rules to specific problems. What is elegant in theory can sometimes be elusive in practice. How effectively can the rules be applied to real-world problems? Are some ethical problems not likely to be resolved by an approach based on rules?
2. *Clarity*—Are the rules sufficiently clear to provide a basis for the responsible exercise of professional authority? Ambiguity is likely to breed confusion and frustration and, as a consequence, may invite neglect. Moreover, clarity is especially important in those cases where the rules are expected to play a role in the adjudication of grievances.
3. *Consistency*—Are the rules internally consistent? Are there logical contradictions within or between rules?
4. *Ordering*—Does the statement of ethical rules provide a means for setting priorities between two or more rules which, although not *prima facie* inconsistent, when applied in practice will require the professional to choose between conflicting obligations?
5. *Coverage*—This refers to the scope of the actions and situations addressed by the rules. Are the rules silent on matters of serious ethical concern? Do they overemphasize matters of convenience, etiquette or expedience at the expense of more pressing issues?
6. *Acceptability*—Do the rules express proper ideals? Should they be accepted as ethically prescriptive?

NOTES

1. The need for international scientific exchanges of knowledge (though not necessarily raw data) has been reiterated periodically by working groups of the American Association for the Advancement of Science (AAAS): see AAAS Special Committee on Civil Liberties for Scientists (1949) in defense against political calls for secrecy and loyalty oaths; the AAAS Committee on Social Aspects of Science (1957) on information transfer; and the AAAS Committee on Science in Promotion of Human Welfare (1960) on international aspects of science.

2. The history of scientific codes of conduct, especially codes that bear on sharing information, seems not to have been well documented. Yet the transformation of ethics codes from questions of etiquette through accepted tradition and codification and training seems interesting enough to warrant the historian's attention. Pigman and Carmichael (1950) made a beginning in their appeal for a code that would improve professional relations and so “better morale and increase productivity among research men” (p. 644). Perhaps changing expectations about such codes also warrants attention.

3. The authors note that not all similar rules were identically phrased in the documents; some editorial discretion was used in the preparation of this list.

4. Some organizations that one might expect to have developed codes or guidelines relating to data sharing have not. They include the American Economic Association and others listed in the Chalk et al. (1980) report.

References

American Association for the Advancement of Science, Committee on Science in the Promotion of Human Welfare 1960 Science and human welfare. *Science* 132:68–73.

American Association for the Advancement of Science, Committee on the Social Aspects of Science 1957 Social aspects of science. *Science* 125:25–147.

American Association for the Advancement of Science, Special Committee on Civil Liberties for Scientists 1949 Civil liberties for scientists. *Science* 110:177–179.

American Sociological Association 1982 *Code of Ethics.* Washington, D.C.: American Sociological Association.

American Statistical Association, Ad Hoc Committee on Professional Ethics 1983 Ethical guidelines for statistical practice: Report of the ad hoc committee. *American Statistician* 37:5–60.

American Statistical Association 1980 *Interim Code of Conduct for the ASA: General Guidelines.* Washington, D.C.: American Statistical Association.

Boruch, R. F., and Cecil, J. S. 1979 *Assuring the Confidentiality of Data in Social Research.* Philadelphia: University of Pennsylvania Press.

Boruch, R. F., Cordray, D. S., and Wortman, P. M. 1981 Secondary analysis: why, when and how. In R. F. Boruch, P. M. Wortman, and D. S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Bowers, W., and Pierce, G. 1975 The illusion of deterrence and Isaac Ehrlich's research on capital punishment. *Yale Law Journal* 85:187–208.

1981 Capital punishment as deterrent: challenging Isaac Erlich's research. Pp. 237–261 in R. F. Boruch, P. M. Wortman, and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Bross, I.D.J. 1983 Comment. *American Statistician* 37:12–13.

Chalk, R., Frankel, M.S., and Chafer, S. B. 1980 *AAAS Professional Ethics Project: Professional Ethics Activities in the Scientific and Engineering Societies.* Washington, D.C.: American Association for the Advancement of Science.

Cordray, D. S. 1982 An assessment of the utility of the ERS standards. *New Directions for Program Evaluation: Standards for Educational Practice* 15:67–81.

Cranberg, R. 1963 Ethical code for scientists? *Science* 141:1242.

Deming, W.E. 1972 Code of professional conduct. *International Statistical Review* 40:215–219.

Ellenberg, J.H. 1983 Ethical guidelines for statistical practice: a historical perspective. *American Statistician* 37:1–4.

European Science Foundation 1980 *Statement Concerning the Protection of Privacy and the Use of Personal Data for Research.* Strasbourg, France: European Science Foundation.

Evaluation Research Society 1980 *Standards for Program Evaluation.* Potomac, Md.: Evaluation Research Society.

Flaherty, D.H. 1978 The Bellagio conference on privacy, confidentiality, and the use of government micro data. *New Directions in Program Evaluation* 4:19–30.

Flaherty, D. H. 1979 *Privacy and Government Data Banks: An International Perspective.* London: Mansell.

Fosberg, F. R. 1963 Letter to the editor. *Science* 141:916.

Garner, J. 1981 National Institute of Justice access and secondary analysis. Pp. 43–49 in R. F. Boruch, P. M. Wortman, and D. S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Gehan, E.A. 1983 Comment. *American Statistician* 37:8–9.

Glass, B. 1965 The ethical basis of science. *Science* 150:1254–1261.

Greenhouse, S.W. 1983 Comment. *American Statistician* 37:15–16.

Greenwald, A. 1976 An editorial. *Journal of Personality and Social Psychology* 33:1–7.

Joint Committee on Standards for Educational Evaluation 1981 *Standards for Evaluations of Educational Programs, Products, and Materials.* New York: McGraw Hill.

Katz, S. et al. n.d. Plan for Analysis of Existing Long-Term Data Relative to Distribution and Mix of Functional Impairment, and Effects of Care. Department of Community Health Science, College of Medicine, Michigan State University.

Katzan, H.S. 1980 *Multinational Computer Systems: An Introduction to Transnational Data Flow and Data Regulation.* New York: Van Nostrand.

Kish, L. 1983 Comment. *American Statistician* 37:17.

Lanz, H. 1963 Letter to the editor. *Science* 141:916.

Martin, M.E. 1983 Comment. *American Statistician* 37:6–7.

Mochmann, E., and Muller, P.J., eds. 1979 *Data Protection and Social Science Research.* Frankfurt: Campus Verlag.

Mosteller, F. 1983 Comment. *American Statistician* 37:10–11.

National Science Foundation 1983 *Grant General Conditions.* NSF-FL-200(4–83). Washington, D.C.: National Science Foundation.

Organization for Economic Cooperation and Development 1981 *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data.* Paris: Organization for Economic Cooperation and Development.

Pigman, W., and Carmichael, E. B. 1950 An ethical code for scientists. *Science* 111:643–647.

President's Reorganization Project 1981 Federal statistical system: access and dissemination. Pp. 21–33 in R. F. Boruch, P.M. Wortman, and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

Rice, D.P. 1983 Comment. *American Statistician* 37:9.

Robbin, A. 1978 Ethical standards and the archivist. *New Directions for Program Evaluation: Secondary Analysis* 4:7–18.

1981 Technical guidelines for preparing and documenting statistical data for secondary analysis. Pp. 84–143 in R.F. Boruch, P.M. Wortman and D.S. Cordray, eds., *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass.

1982 Ambiguity, Value Choice, and Administrative Discretion: When Policy and Practice Diverge in Public Organizations. Unpublished manuscript, University of Wisconsin.

Robbin, A., and Josefacki, L. 1982 *Public Policy on Health and Welfare Information: Compendium of State Legislation on Privacy and Access.* Madison, Wisc.: University of Wisconsin.

Roberts, H. V. 1983 Comment. *American Statistician* 37:18.

Rossi, P., ed. 1982 *New Directions for Program Evaluation: Standards for Educational Practice* 15.

Sasfy, J., and Siegel, L. 1982 *A Study of Research Access to Confidential Criminal Justice Agency Data.* McLean, Va.: Mitre Corp.

Society of American Archivists 1980 A code of ethics for archivists. *American Archivist* 43:414–420.

Solomon, H. 1983 Comment. *American Statistician* 37:15.

Trochim, W.M.K. 1982 Methodologically based discrepancies in compensatory education evaluations.

Evaluation Review 6:443–480.

U.S. General Accounting Office 1978 *Assessing Social Program Impact Evaluations: A Checklist Approach.* PAD-79-2. Washington, D.C.: U.S. General Accounting Office.

Exhibit A
The Bellagio Principles

1. National statistical offices should provide researchers both inside and outside government with the broadest practicable access to information within the bounds of accepted notions of privacy and legal requirements to preserve confidentiality.
2. Legal and social constraints on the dissemination of microdata are appropriate when they reflect the interests of respondents and the general public in an equitable manner. These constraints should be re-examined when they result in the protection of vested interests, or the failure to disseminate information for statistical and research purposes (i.e., without direct consequences for a specific individual).
3. All copies of government data collected or used for statistical purposes should be rendered immune from compulsory legal process by statute.
4. In making data available to researchers, national statistical offices should provide some means to ensure that decisions on selective access are subject to independent review and appeals.
5. The distinction between a research file, in the sense of a statistical record (as defined in the 1977 report of the U.S. Privacy Protection Study Commission), and other micro files is fundamental in discussions of privacy and dissemination of microdata. All dissemination of government microdata discussed in connection with the Bellagio Principles is assumed to be a transfer of data to research files for use exclusively for research and statistical purposes.
6. There are valid and socially-significant fields of research for which access to microdata is indispensible. Statistical agencies are one of the prime sources of government microdata.
7. Public use samples of anonymized individual data are one of the most useful ways of disseminating microdata for research and statistical purposes.
8. Techniques now exist that permit preparation of public use samples of value for research purposes within the constraints imposed by the need for confidentiality. Countries with strict statutes on confidentiality have prepared public use samples.
9. There are legitimate research purposes requiring the use of individual data for which public use samples are inadequate.
10. There are legitimate research uses which require the utilization of identifiable data within the framework of concern for confidentiality.
11. Other techniques of extending to approved research the same rights and obligations of access enjoyed by officers of the government agency need to be considered in terms of better access.
12. There is considerable potential for development of more economical

and responsive customized-user services, such as 1) record linkage under the protection of the statistical office, 2) special tabulations, 3) public use samples for special purposes. Such services must often involve some form of cost recovery.

13. Some research and statistical activities require the linking of individual data for research and statistical purposes. The methods that have been developed to permit record linkage without violating law or social custom regarding privacy should be used whenever possible.
14. Professional or national organizations should have codes of ethics for their disciplines concerning the utilization of individual data for research and statistical purposes. Such ethical codes should furnish mutually agreeable standards of behavior governing relations between provides and users of governmental data.
15. Users of microdata should be required to sign written undertakings for the protection of confidentiality.
16. Considerable efforts should be made to explain to the general public the procedures in force for the protection of the confidentiality of microdata collected and disseminated for research and statistical purposes.
17. The right of privacy is evolving rather than static, and closely related to how statistics and research are perceived. Therefore, statisticians and researchers have a responsibility to contribute to policy and legal definitions of privacy.
18. Public concern about privacy and confidentiality in the collection and utilization of individual data can be addressed in part as follows:

(1) voluntary data collection, whenever practicable;
(2) advanced general notice to respondents and informed consent, whenever practicable;
(3) provisions for public knowledge of data uses;
(4) public education on the distinction between administrative and research uses of information.

Source: Flaherty (1978:19–27).