TM-3729/000/00

SOCIAL IMPLICATIONS OF SOCIAL SCIENCE DATA ARCHIVES

7 November 1967

TM-3729/000/00

TECHNICAL MEMORANDUM

(TM Series)

SOCIAL IMPLICATIONS OF SOCIAL SCIENCE DATA ARCHIVES

by

Douglas K. Stewart*

SYSTEM
DEVELOPMENT
CORPORATION
2500 COLORADO AVE.
SANTA MONICA
CALIFORNIA
90406

*University of Pittsburgh; formerly consultant to System Development Corporation

A-2348 4/65

(Page 2 blank)

ABSTRACT

Recent development of large-scale collections of information relevant to the Social Sciences is significant for society as well as for the Social Sciences. This paper describes some of the present and possible future aspects of Social Science data archives, emphasizing associated problems and their ramifications.

(Page 4 blank)

TABLE OF CONTENTS

BLANK PAGE

SOCIAL IMPLICATIONS OF SOCIAL SCIENCE DATA ARCHIVES

by

Douglas Stewart*

I. INTRODUCTION

The recent development of large-scale collections of information relevant to the social sciences (commonly referred to as "social science data archives") is significant for society as well as for the social sciences. It is the purpose of this paper to describe the present and possible future status of social science data archives and to call attention to some of the current problems and possible ramifications.

II. HISTORY OF ARCHIVES

Much research conducted by social scientists consists of analysis of the data generated by a complex society. Both demography and economics, for example, are largely dependent upon government-collected information. In addition, data are often collected by researchers to deal with a specific problem.

In the past, data gathered for a specific research purpose were analyzed by the investigator and then discarded or filed away. It soon became apparent that the loss of this information was doubly wasteful. First, as Cantril (1948) noted, public opinion data constitute important information sources for future historians; data discarded are forever lost to future generations. Second, the initial acquisition of social science data is expensive (interview studies typically expend $10 to $20 per case for interviewing, coding and punching); furthermore, data are seldom completely exploited by the primary investigator. Clearly, data storage and sharing are worthwhile, since secondary analysis is likely to have high utility relative to the marginal cost (data costing $10,000 to collect might cost $10 to reproduce).

With these considerations in mind, scholars established bilateral arrangements for sharing data. (These arrangements were sometimes, of necessity, quite informal; it it not uncommon to see an individual arriving at a professional meeting carrying a box of punch cards or a reel of computer tape for some friend who will also be in attendance.) Data sharing based on such casual arrangements did not ensure preservation of historically relevant data, however; and the necessity for reliable methods of retrieval, together with advances in information processing methodology--greatly increasing the demand for data--made a more formal system.

As the flow of social science data has increased, practices of data sharing which were previously acceptable are no longer adequate. (How, for example, can a social scientist know who has data relevant to his interests? Will an individual schclar enjoy the burden of responding to requests for copies of his data as the frequency of these requests increases?) There are problems even when the data are held by a large research institute, if that institute is not designed as a "data library." The response to these problems has been the formation of social

---

*This paper was produced as part of the Computers and Privacy Project at SDC while Dr. Stewart acted as a consultant. Currently, he serves as Assistant Professor of Sociology and Director of the Social Science Information Center at the University of Pittsburgh.

science archives, which we here define as archives gathered by organizations committed to storing and disseminating machine-readable data to social scientists. The establishment of archives, in turn, has led to the formation of the Council of Social Science Data Archives, which now has more than twenty members and is concerned with the solution of common problems.

Several worthwhile discussions of social science archives are available; among the more recent are Bisco (1966) and Glaser and Bisco (1966).

III. KINDS OF ARCHIVES

There are various dimensions along which archives may be described. One of these, of course, is the size of the collection. The largest archives serve a set of national or international users. At the other end of the spectrum, numerous archives exist within institutions, primarily for the convenience of scholars associated with those institutions.

Another way of classifying archives is with regard to the kind of data they contain. The contents of archives vary according to methods of data collection (making sample surveys, or drawing upon data already collected, such as national economic indicators and voting data) or topical foci (geographic areas, or such subject areas as mental health and education).

Still a third distinction may be made between the functions of the various archives: While some are essentially warehouses of data limiting their services to retrieving data sets and producing the accompanying documentation, other archives offer a wide range of services such as checking for errors ("cleaning"), reformatting, and analysis.

IV. SOURCES OF INFORMATION FOR ARCHIVES

Social science archives receive data sets from numerous suppliers. Many archives exist as appendages to research institutes and are intended primarily to dispense data collected by the parent institute. The more general survey data, on the other hand, are largely supplied by commercial and academic survey research operations. Other sources include not-for-profit corporations and government agencies. In addition, several archives are putting into machine-readable form information derived from more conventional historical archives (e.g., election statistics and biographical information).

V. USAGE OF ARCHIVES

Academic researchers are among the most frequent users of social science archives, followed by not-for-profit corporations and government agencies. In general, commercial users are not granted access to archive holdings. It seems safe to assume that the intelligence community has some interest in archival holdings.

(We will return to this point, but there are numerous tales--including a report of one archive that received an overt request for a data set from a foreign embassy. Other archives have received unlikely requests from unlikely users.)

Data supplied by archives are most commonly used for the simple reanalysis of a single data set for research purposes not pursued by the original investigator. One such case of reanalysis of a single data set occurs during the training of students in empirical social analysis. This usage has the beneficial effect of introducing the student to large-scale, high-quality data and generally more ambitious research problems than would be the case if he were to depend on personally collected data. An increasingly common practice, this trend will presumably be augmented by the introduction of time-shared computer systems.

VI. BENEFITS OF ARCHIVES

The formation of a new data base through the concatenation of several data sets offers the opportunity for powerful research. Data from several studies undertaken at different points in time may be combined in order to study changes within a society. Similarly, data obtained from several countries may be brought together so that cross-national differences can be explored. In addition, certain social types are sufficiently infrequent that very few appear in any given data set; an increased number of such a type (sometimes required for analytic purposes) can often be formed by pooling across studies. Moreover, it is sometimes possible to "enrich" one data set with information from another source (e.g., census data may be merged with election statistics). It is assumed that the analysis of multiple data sets will become increasingly common as data processing capabilities reduce the cost and technical skill required.

Numerous benefits have already accrued from the existence of archives, in addition to the economies effected. Since most social science data are expensive to collect and relatively inexpensive to analyze, secondary analysis (where appropriate) constitutes a wise allocation of scarce research resources.

An additional payoff has profound ramifications for the future of smaller colleges. In the past, a scholar interested in large-scale social data felt required to locate at a large, research-oriented university. This necessity has had predictable effects on academic recruitment patterns. The existence of social science archives can be expected to counter this tendency. There is already evidence to this effect, although most small colleges still lack the necessary data processing facilities. As regional (if not national) computer systems are developed, smaller schools will have access to the computing power necessary for the analysis of large-scale data bases and the "small school" problem may be expected to diminish.

In these various ways, the advent of social science data archives has had significant effects which most would view as desirable. In general, this collective allocation of scarce research resources has effected a relatively high return.

VII. CURRENT PROBLEM AREAS

There exist several problem areas regarding social science data archives which, though not currently considered serious, may indicate future needs for clear-cut policies and practices. Many, though not all, of these problems are concerned with the protection of sensitive data.

A. INDIVIDUAL PRIVACY

Guarantees of confidentiality or anonymity are so common in social research that they are often (and sometimes mistakenly) assumed where they are not explicitly offered. Confidentiality may be compromised in several ways--some, apparently, almost unavoidable. For example, it is often necessary to retain identifying information for individual files so that additional information may be appended. (To illustrate, in an archive with a research mission engaged in a longitudinal study of individuals, it is necessary to merge data collected at various points in time.) Loss of confidentiality may also result from accident, even in the best-guarded situations. In one such instance, during a casual conversation with a member of a research staff, a nationally prominent individual was mentioned as a member of the sample under discussion. While the matter was not pursued, it would appear that this archive depends on the individual integrity of its research staff to preserve confidentiality.

Deliberate internal subversion, of course, is not a problem unique to archives. This writer does not know of any instance in which an individual's trust has been intentionally compromised. It should be added, however, that as the amount of information available in one location increases, new procedures will be required to protect it.

B. CONFIDENTIALITY BETWEEN RESEARCHER AND ARCHIVES

Another problem that has received little attention is the conflict between the need of the archive, on the one hand, to preserve confidentiality of data, and the need of the researcher, on the other hand, to obtain sufficiently usable and flexible information. An individual researcher may collect, for some set of individuals, data that he then wishes to supplement with information from an archive. The archive might agree to match records on name or social security number, remove the identification fields, and return the enriched (but sanitized) data to the researcher. Further consideration indicates, however, that a "secret" identification variable could be embedded in the record supplied to the archive--in fact, the identifying "signature" could often be supplied by

variables in which the total set of values across all variables was idiosyncratic for each individual. Various technical approaches have been suggested for this problem, but none, to my knowledge, is fully acceptable in terms of both protection of privacy and analytic flexibility--short of having analyses conducted by the archive.

C. PRIVACY OF ORGANIZATIONS

In discussing privacy-related issues, it should be noted that decision units other than individuals are often guaranteed the anonymity of information which they permit to be collected. Educational researchers often guarantee a kind of anonymity to school districts (or religious bodies that maintain schools); i.e., they guarantee that no contrast between their schools and others will be evident in the published results. Similarly, corporations that supply information to the federal government are normally guaranteed anonymity. Although they are beyond the scope of this paper, such practices raise interesting questions on the right of privacy for organizations: If, in fact, certain collections of individuals are to be guaranteed privacy, which collections are they--voluntary membership associations, neighborhoods, ethnic groups, etc.?

In general, because of the newness of the field, social science archives are without good policy guidelines in this area, although their practices have been conservative in these matters.

D. PROBLEMS OF ACCESS

Another area in which archives have even less precedent or policy revolves about the question of who should have access to archival data. Many academically based research institutes operate under the restriction that no research be undertaken for commercial clients and the same interdiction applies to archival access. Beyond this restriction, most archives operate without policy on access (except for the ubiquitous "common interest" clause).

The Simulmatics Project has caused some comment in regard to access, and raised relatively new questions. (This project was undertaken by a group of social scientists for the Democratic Party during the presidential campaign of 1960. Its purpose was to simulate, via computer, the outcomes of campaigns in which various sets of issues became dominant. The data base for this project was constituted of numerous public opinion surveys supplied by the Roper Public Opinion Research Center. For details on this project, see Pool, et al., 1964.) Through archives, a massive data base may be formed economically--indeed, this is the purpose of archives. Many perceive a danger when increasingly sophisticated social science techniques are combined with large-scale data bases. We may well ask to what degree social engineering by private groups will be possible and, if such manipulation is possible, to what extent will this constitute a danger.

E. CONTRACTUAL CONSIDERATIONS

Care must be taken in the wording of contracts involving acquisition of sensitive data. The need for contractual safeguards against unwarranted access to such data is made clear by the following case: An archive with a research mission undertook a study for a federal agency and submitted the required report. The funding agency demanded the raw data, including identification fields. The archive responded that it had guaranteed confidentiality to the sample members. The agency responded that it would "accept full responsibility," and noted that the contract executed between the archive and the agency stated that the data file was the property of the agency. The research archive thus was required to submit full data to the funding agency. We need not impugn the motives of the agency to find this practice alarming.

F. INTERNATIONAL PROBLEMS

Given the existence of archives of foreign data, we may ask if all these problems apply in the international context. Many nations view public opinion data as sensitive and do not always approve of U. S. archiving attempts. For example, Project Camelot (cf. Horowitz, 1967), which included the collection of social data in Latin America in its design, aroused sharp controversy and was disbanded. A significant portion of the criticism of Camelot was directed at the proposed analysis (cf. Galtung, 1967). However, once data are made available to scholars, policing the uses to which they are put seems impractical (and perhaps undesirable). It seems unlikely that archives would be successful if they were to restrict the analyses undertaken on their data.

The controversy generated by Camelot is indicative of the questions that may be raised when archives attempt to accumulate data from foreign countries.

G. QUALITY CONTROL

An additional problem may be described as the "halo effect." That is to say, the user of a research report may associate the prestige of the data source with the finished research. Certain research institutes in this country have deserved reputations for high-quality data and analysis. A research report based on secondary analysis of data generated by one of these institutes may be viewed as somehow sharing in their prestige. Since the original data producer has no control over the analyses conducted secondarily, the consumer of the report should be explicitly forewarned. This caveat could take the form of a standard footnote informing the reader that data were supplied by an archive, but that neither the archive nor the original supplier is responsible for the quality of secondary analysis and interpretation.

VIII POSSIBLE FUTURE STATUS OF ARCHIVES

In discussing future archives we make the assumption that recorded information which is of interest to social scientists has a fair probability of being archived. Thus archives of the following types seem to be reasonable projections for the next decade.

- Textual archives including print media. As the practice of computer-controlled typesetting increases, we may expect increasing amounts of machine-readable textual material to be available. In addition, the wide use of optical scanners capable of reading standard printed text will eventually, in part, transform conventional libraries into archives of machine-readable text.

- Archives of educational data with very extensive holdings are obviously anticipated. The New England Educational Data System (NEEDS) and Project Talent are precursors of future development in this line.

- Archives of consumer behavior information. These seem reasonable given the ubiquity of credit cards. The exact form of such archives is not clear. It may be assumed that social scientists will consider access to such data desirable.

- Archives of library-usage behavior. As libraries increasingly automate circulation information, these data will become available for archiving. Information on who checks out what books would be valuable to those interested in individual behavior, in the spread of ideas (the sociology of knowledge), and in the relative popularity of various sets of books over time (the history of ideas).

- Archives of biographical data on elites. These already exist and can be expected to increase in scope and intensity.

- Archives of the type of information currently sold by mailing list companies. Information contained in such archives would include income, consumer behavior, contributions to charities, association memberships, etc.

- Archives of the financial behavior of individuals (stock transactions, etc.). These seem likely as social scientists attempt to understand the behavior of individual decision units in financial areas.

- Medical-psychiatric archives. These can be anticipated with increased record keeping prompted by proliferation of insurance plans (in both private and public sectors).

- Election data on smallest reporting units. The Inter-University Consortium for Political Research is already involved in collecting and organizing such information.

- Archives of legal record data, made possible and likely by automation of record-keeping in the legal area.

- Data on television-viewing behavior. Information on this is currently being collected covertly as well as overtly. There is no reason to believe that these data will not be archived in the future.

It should be emphasized that the data described above currently exist in recorded form. In each case they are accessible to many people and may be in some cases already in the public domain. The archive function would be directed simply at making the data more readily usable by social scientists.

As a science progresses we may expect its data base to expand (and conversely, as a data base expands we may expect its science to progress); thus we may look somewhat farther into the future and make some predictions based upon probable developments in science and technology. In the life sciences, for example, we may project the following possible archival efforts:

- Archives of individual genetic structures may be anticipated as scientists break the genetic code.

- Archives of physiological data may be developed as telemetry techniques become increasingly sophisticated. Functions of the brain, heart, etc., can now be monitored by attached sensors and may well be monitored in the future by remote sensors.

- Electronic "sniffers" can be expected to supply information on ingested substances, including drugs.

IX. FUTURE PROBLEM AREAS

Social Science archiving efforts are currently beset by problems stemming from a lack of policy. As the scope and magnitude of social science data bases increase, these problems can be expected to become more pressing.

A society may be viewed as a system of communication channels and described in terms of the flow of information. From this viewpoint it is immediately clear that information is seldom uniformly diffused within a society. Certain social values are inextricably involved with the dispersal--or bounding--of communication. On the one hand, an informed citizenry is considered desirable; i.e., it is considered desirable that citizens (essentially equivalent to residents) be informed about various issues and occurrences. This would suggest no bounding of the communication of these types of information. On the other hand, the tradition which defines the attorney-client relation as privileged provides an example of a tightly bounded channel. In general, we may argue that communication channels are maintained through two mechanisms: formal rules (legal and otherwise) and cost/benefit ratios. Changes in technology can modify a cost/benefit ratio such that a channel which was maintained through this mechanism is no longer so maintained. If the continued maintenance of the channel boundary is deemed desirable, formal rules must be instituted to take the place of the cost/benefit ratio for boundary maintenance functions. (It should be noted that rules may be considered a special case of defining "cost," but there is no necessity to amplify this distinction here.) Various legal steps are taken to confine the transmission of certain kinds of information. For example, the Securities and Exchange Commission monitors the activities of relevant advisory services; copyright and patent laws act to restrict information diffusion; the Federal Communication Commission places various constraints on information flow with statutes regarding gambling information, political propaganda, etc. Similarly, the Postmaster General places restrictions on what may pass through the mails. In addition, laws protecting privacy (e.g., proscriptions against wire tapping) constitute rules which have as their function the maintenance of certain communication channel boundaries. Given such precedents for state intervention to restrict the transmission of information in the interest of one or more social values, we may ask what legislation may be anticipated. To do this, we must first consider some of the issues involved.

Confidentiality

As noted above, much information is collected in which the anonymity of the reporting unit is guaranteed. However, no clear-cut policies or regulations on confidentiality presently exist outside the federal government. There are two issues to be raised in this regard: First, confidentiality as a value or end in itself would seem to require something more than a set of ethics in order to be safeguarded. Second, the guarantee of anonymity is often necessary to the acquisition of information. That is, people are expected to give more truthful answers to sensitive questions if they are guaranteed anonymity. If the public believes (correctly or not) that confidentiality is being compromised, data may become increasingly difficult to procure.

A complex society needs much information about its individual citizens. A statement that 100,000 people in a certain area are unemployed does not supply information on how long they have been unemployed or whether their skills are needed in another area, etc. A national level of 5 per cent of the work force unemployed means one thing; 25 per cent unemployment within a certain age-race stratum has other implications for policy. In short, if the decision units (individuals, corporations, etc.) lose faith in the information-collection agencies of society, administration will be less effective. Not only must confidentiality be maintained, but the fact that it is "secure" must be communicated to the public. Dependence solely on the integrity of those individuals with access to sensitive information (as a guarantee that the data will not be compromised) is unlikely to convince outsiders.

The need for confidence in anonymity is considered sufficiently important to be given a prominent position in a ranking of problems regarding data banks, archives and centers.

Implications for Social Engineering

Two points may be raised here. First, if access to large amounts of social data does in fact enable individuals to have a significant impact on society, should some limitation be placed on who has such access? Second, if it seems that increasing the amount of social information increases the efficacy of social engineers, and if this effect is considered undesirable, information acquisition will become increasingly difficult (if not impossible in certain areas). Just as a lack of faith in preservation of anonymity might lead to difficulty in data procurement, so the belief that responding to a questionnaire would strengthen "Big Brother" might have the same result.

Unplanned Effects of Information Dissemination

Predictions intended for general interest may have unanticipated effects. Individuals endeavor to anticipate the future and to act according to their expectations. If a prediction of things to come is made available to an individual, his behavior may be modified. He may buy a stock if he expects the price to increase. If the outcome of an election is not in question, he may not bother to vote: Predictions of a presidential election outcome may therefore cause people to stay home and may change the outcome of a lesser contest (if not the presidential outcome itself). Similarly, predictions of elections in which more than two individuals are contesting an office can cause profound switching. The point to be made is simple: Social information has social ramifications.

The publication of certain socially relevant projections may be controlled in the future. Various possibilities exist. A minimum requirement might be a statement of the probability associated with the prediction. Other alternatives, short of censorship, include the parallel publication of other responsible predictions. Perhaps the most likely restriction will simply be the complete specifications of the boundary conditions and publication of statements regarding the validity of the prediction.

Possible Government Regulation of Archives

The federal government is currently the largest single data-collection agency in the United States. It is also the largest single source of funds for the collection of social information by academic researchers, and thus it exercises control over what information is to be gathered. It also controls dissemination of information so collected. It is not clear whether the federal government will move to control the collection and dissemination of information when funding has not been federally supplied. There may well be a move to institute standard procedures for the preservation of confidentiality. This might take a form similar to regulations dictating procedures for agencies, both within and outside the Federal Government, dealing with classified information. (The Bureau of the Census supplies one example of such an agency.) These procedures might be made mandatory whenever aconymity is guaranteed by a data collector. If a guarantee of confidentiality is not explicitly offered, a data collector might be required to inform a respondent that none is to be expected. If such procedures were instituted, an educational campaign might be initiated to inform the public.

Applications of Social Archives

The most difficult question is that of limiting the uses to which social information may be put. Almost no social research is "pure" in the sense that it has no applications. It would be ironical if at some future time social science was considered so powerful as to be dangerous; or if social analysts of large-scale data bases were required to have both a clearance and a "need to know." This is not the current state of affairs, although there have been classified opinion polls; and, as noted above, response to Project Camelot made it clear that some countries view social data as too sensitive for public consideration.

X. SUMMARY

It has been the purpose of this document to describe briefly the current and future status of social science data archives. It is hoped that some of the questions raised will receive thorough discussion.

REFERENCES

Bisco, Ralph L. Social Science Data Archives: Progress and Prospects. Social Sciences Information, 1967, 6, 39-74.

Brictson, Robert C. Some Thoughts on the Social Implications of Computers and Privacy. Santa Monica: System Development Corporation (SP-2953), 1967.

Cantril, Hadley. Opinion Trends in World War II. Public Opinion Quarterly, 1948, 12, 30-40.

Dennis, Robert L. Security in the Computer Environment. Santa Monica: System Development Corporation (SP-2240/000/01), 1966.

Galtung, Johan. After Camelot. In Horowitz (Ed.), The Rise and Fall of Project Camelot. Cambridge, Mass.: M.I.T. Press, 1967.

Glaser, William, and Ralph Bisco. Plans of the Council of Social Science Data Archives. Social Sciences Information, 1966, 5, 71-96.

Horowitz, Irving Louis (Ed.). The Rise and Fall of Project Camelot. Cambridge, Mass.: M.I.T. Press, 1967.

Pool, Ithiel de Sola, Robert P. Abelson, and Samuel L. Popkin. Candidates, Issues and Strategies. Cambridge, Mass.: M.I.T. Press, 1964.

Rokhan, Stein (Ed.). Data Archives for the Social Sciences. Paris: Mouton, 1966.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

System Development Corporation
Santa Monica, California

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

Social Implications of Social Science Data Archives

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(First name, middle initial, last name)*

Douglas K. Stewart

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
7 November 1967	17	9

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Independent Research	TM-3729/000/00
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

The recent development of large-scale collections of information relevant to the social sciences (commonly referred to as "social science data archives") is significant for society as well as for the social sciences. It is the purpose of this paper to describe the present and possible future status of social science data archives and to call attention to some of the current problems and possible ramifications.

DD FORM 1 NOV 65 1473

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Social sciences Data archives						