March, 1963

A proposal to the
Ford Foundation International and Comparative Grant

TOWARD THE DEVELOPMENT OF AN INTER-UNIVERSITY DATA ARCHIVE SYSTEM FOR COMPARATIVE RESEARCH

SUMMARY

Mutual concern over problems of orderly data accumulation and the scholarly access to information necessary for large-scale comparative and area research has brought together interested parties from a number of major American universities. It has been agreed that

(1) data of basic social, political and economic interest are now being collected at a rapid and increasing pace by a great variety of agencies and investigators all over the world;

(2) that while most of these data are "public" in theory, in practice there is little provision to ensure that they be made widely available for research purposes; and

(3) that availability of these data in a form manipulable by modern data-processing techniques would greatly widen the horizons for international, comparative and area research.

In short, there is an urgent need for the establishment of major data archives, parallel to conventional library facilities but "automated," to serve those behavioral sciences engaged in the study of historically significant populations.

It has been agreed furthermore that the nature and format of these kinds of data pose a series of unique problems that would best be approached by the establishment of a coordinated network of repositories associated with interested graduate schools. These repositories would divide the labor of data acquisition, organization and storage, and would work cooperatively to solve the many technical problems of data classification and retrieval, thereby facilitating genuine access to the materials for the broader scholarly community. The overall task is a large one. This initial proposal seeks limited "pilot" funds of $23,760 to permit a year of exploration of potential repository sites and the drawing up of a mutually satisfactory organizational and technical design for the network.

BACKGROUND: THE NATURE OF THE NEED

The past few decades have witnessed a dramatic upturn in the rate of systematic collection of information about significant populations around the world. We are in a period, particularly in the developing countries, where there is a recognition of a growing need for basic data of a great variety of kinds. To meet these needs, a great proliferation of governmental, academic and commercial agencies has occurred, and these agencies are now producing a great volume of worthwhile information. Despite its relatively recent invention, the sample survey is now being applied by numerous research organizations in Latin America and Asia as well as by countless agencies in the United States and Europe. Teams of social scientists have been able in many instances to collect significant bodies of systematic data even within the least developed nations.

At present there is scholarly consensus that this expanding body of knowledge represents a rich resource, yet one which has scarcely begun to be exploited. Most collections of data have rather circumscribed goals: most investigators recognize that they cannot begin to exhaust the potential analyses to which their materials might be put. Furthermore, as the numbers of such collections multiply in different areas, the possibility of confronting comparable materials across a large array of studies in different areas becomes tantalizing. Twenty years ago, the difficulty of manipulating such a great volume of information would have prohibited exploitation of this order. Today, rapid advances in data-processing techniques have given us just this technological capacity on an unforeseen scale. The remaining bottlenecks are of a different order. We simply lack any institutional provision on an adequate scale for the organization,

accumulation and effective dissemination of these scientifically worthy data. It is our belief that steps taken to render these services would do more than any other single thing to expand the efficiency, power and depth of comparative research. We feel that the need for such steps is already urgent, and becoming more so as information-gathering activities continue to mushroom.

The absence of appropriate provisions for data dissemination incurs costs in research efficiency which have become alarming. Among these costs are the following:

(1) Basic inefficiency of single-shot analyses. A very large portion of research time and financing for primary research is given over to data gathering and the organization of materials in a form susceptible to rapid processing. Once data are in this form, they can be economically preserved, duplicated and disseminated for secondary analyses at a small fraction of the original cost. The research community is not currently organized to profit from such secondary analysis.

(2) Duplication of efforts. There are an increasing number of incidents in which two or more investigators have plowed research funds into the collection and systematization of the same bodies of data.

(3) Dilution in scope of current primary research. There is little current area research which would not be enriched by comparison with comparable materials from other areas or by the historical depth which earlier information from the same area could provide. Investigators are usually aware that this is true. However, they are often unaware of extant materials or where aware, recognize that search and systematization of such background materials would be prohibitively expensive in the current state of the arts.

(4) Destruction of collected information. It is said that a number of research agencies in Europe have simply begun to destroy old data, as they lack provision either for their storage or for servicing a mounting tide of requests for information from them. While not all of this information would warrant preservation, that which does is quite irreplaceable inasmuch as it refers to a historical past which will not be recaptured. Similar disappearance of classic bodies of systematized information is occurring in the

United States with the deaths or retirement of principal investigators, most of whom would be delighted to find a place for the preservation of their data.[1]

Widespread recognition of these inefficiencies has led to a number of published proposals for remedies,[2] as well as to the establishment of several small data libraries, with others planned for the near future both in the United States and Western Europe.[3] These latter libraries have by and large been conceived to fit local research requirements. With but one or two exceptions, however, all of the primary scholars associated with these archival developments have come to feel a strong desire for coordination of efforts with other archives. Up until now

---

[1]Instances have even come to our attention of proposals to systematize obviously basic bodies of data which have been called into question simply on the grounds of no obvious locus where the data might ultimately be stored to serve the research community. After the SSRC had commissioned a study of the feasibility of a concentrated effort to collect and validate a wide range of American voting statistics, for example, this became one of the prime residual questions after feasibility of collection was demonstrated. See Walter Dean Burnham, "Pilot Study: Recovery of Historical Election Data," report to the Committee on Political Behavior, Social Science Research Council, October, 1962 (mimeo).

[2]The most recent summary of the plight along with a plea for a remedy has appeared independent of our discussions and since the time when this proposal was first drafted. See Myron J. Lefcowitz and Robert M. O'Shea, "A Proposal to Establish a National Archives for Social Science Survey Data," The American Behavioral Scientist, Vol. VI, No. 7, March 1963, p. 27.

[3]The earliest full-fledged archive in Europe, covering primarily sample survey data from Germany, was developed at Cologne by Professor Erwin Scheuch. Another archive is expected to be in operation at Cambridge shortly. A library of comparable data is being accumulated in Norway by Stein Rokkan and his colleagues, and other library possibilities in Amsterdam and with UNESCO itself in Paris are being discussed. In the United States, the Roper Center at Williams College is the earliest in point of establishment. More recently, data libraries have been set up at Yale and at the University of California (Berkeley). A library of American data is currently being organized at the University of Michigan as a resource for the Inter-university Consortium for Political Research, a cooperative research and training organization involving the membership of 24 leading graduate schools.

insufficient financing has meant nonetheless that contact has been limited to stray informal communication and a few small gatherings supported variously by UNESCO, the International Social Science Council, the Social Science Research Council, and the Inter-university Consortium for Political Research.[4]

This desire for coordination has at least two important sources. First, there are practical limitations on the scope of any archive, and often these limitations lie in the geographical scope of the data. Hence a de facto division of labor, often along geographical grounds, is inevitable; the only question is whether it can be a coordinated division of labor or not. Library directors recognize that the holdings will be more or less useful for true cross-national or comparative research in the measure that content classification, data format and other conventions are rendered compatible with those at other archives. There is enough of the arbitrary in many of the decisions that the point of prime importance is simply to have them made in some kind of close communication. However, the decisions are sufficiently numerous and technical that they require truly sustained or intensive communication, the mechanisms for which are currently lacking. Secondly, there is another order of technical decisions to be made which ideally require consultation between the social scientist and other highly-trained personnel, such as data retrieval experts, if the archive is to provide much genuine access to the scholar. Up until now, individual archives have been much too small to permit such ambitious planning, although once again technical innovations and solutions good for

---

[4]The two conferences from which this proposal has been developed were financed by the Inter-university Consortium out of concern for the problems being here discussed.

one archive would be equally useful for another. Hence, the interest in coordination has been whetted.

The utility of standardized solutions to a wide range of data organization questions between archives has its counterpart at the level of the individual investigator who is engaged in the systematization of a body of data. Even limited experience with data libraries in the United States has served to underscore the tremendous variety of ways in which common procedural and classification problems can be handled by isolated investigators. This often means that cards being absorbed in the data repository must undergo a variety of format revisions which are expensive in themselves. At times, too, such investigators have failed to capture fairly obvious aspects of their information which are irretrievable and which would have added markedly to the multi-purpose value of their labors. Occasionally, it is true that these variations reflect important intellectual choices on the part of the investigator, a freedom which we would hardly question; far too often, however, they are a simple function of hasty and arbitrary decisions, if not decisions made in ignorance that more efficient alternatives existed. Coordination between repositories should lead to publicized conventions in many of these matters which would greatly increase the compatibility of data systematized by different individuals and organizations, and should thereby reduce one whole sector of cost for the repositories in absorbing data into their holdings.

In general, then, the coordinated development of a more ambitious network of data repositories would seem to hold great promise for the increased efficiency of comparative research. It would forestall the loss of some precious bodies of data, and make them more generally available.

It would also provide an outstanding impetus to secondary analysis of extant data. That the time is ripe for such a development is signalled by the establishment of smaller data libraries here and abroad, for these provide a skeleton on which to build. Here again, however, we are impressed by the urgent need for immediate progress. The older of the young data libraries now report themselves to be arriving at a stage where many of their procedures and conventions must be "frozen" very soon. Attempts at coordination after this point, however great the good will, are going to become increasingly expensive, inasmuch as they may require undoing and redoing of steps taken across increasingly large stores of data.

THE GENERAL NATURE OF THE ENVISAGED PROGRAM

The ultimate magnitude of the task which we are proposing is great. We propose to cope with an effort of this size by the familiar tactics of a rational division of labor between a network of repositories,[5] and a phased operation which begins with modest and manageable steps, yet which is conceived from the start to take account of the likelihood of major growth.

There is at least some rough precedent for this kind of operation. Conventional research libraries across the United States have in recent

---

[5]This proposal, as well as the broader program envisaged, refers only to a network of repositories for international data at sites within the United States. However, up to this point we have worked in close cooperation with parties in Europe engaged in the discussion of comparable repository coordination there with the help of UNESCO. Indeed, Erwin K. Scheuch, Director of the Cologne Zentralarchiv and currently lecturing at Harvard, has been a party to the two meetings which have led to this proposal. We would expect this close coordination between European and American efforts to be maintained.

years coped with problems of access and coverage of foreign book publications by a cooperative arrangement (the Farmington Plan) which assigns to each participating library the responsibility for collecting materials from certain limited publication areas. The individual library naturally retains the right to collect whatever other books it sees fit, but it must acquire all publications in its area of responsibility.

We would anticipate an analogous arrangement whereby each repository would assume responsibility for certain jurisdictions, defined in terms of geographic areas, and to some degree in terms of types of data (e.g., sample survey materials or enumerative aggregate statistics; or elite data as opposed to broader population materials) and special disciplinary interests. These jurisdictions would of course be matched to existing curricular and research strengths of the graduate faculties at the host institutions.

While the several repositories could not be expected for a variety of reasons to acquire even a major proportion of the total flow of data within their respective jurisdictions, each would have the responsibility for monitoring its assigned flow. Each repository would be expected to establish relations with the more prominent sources of valuable data within its jurisdictions abroad. The individual repository could thus act as a middleman for American scholars interested in data not yet acquired, and help to sharpen the consumer's awareness of relevant data while reducing current strains which are growing up as foreign data sources find themselves burdened by numerous American requests for the same materials. Finally, and most important, the individual repository would have the responsibility of selecting new data for acquisition, guided by mutually-erected criteria

of (1) data quality; (2) multi-purpose interest; (3) cost of acquisition; and (4) parallelism of holdings referring to different areas across the repositories. With selection would come the responsibility for organizing the acquired data in a format compatible with practices of the network as a whole, thereby facilitating easy access and transfer of information between repositories.

In principle and in the long term, the repositories would be seen as fitting sites for the storage of any codable material bearing on the policy aspects of all of the social sciences. In the early stages, however, the scope of data acquisitions would be strongly influenced by immediate research demands. Furthermore, while the prospective repositories would attempt to cover the world geographically, the assigned jurisdictions would not be likely to be exhaustive in terms of types of data. It seems reasonable to envisage an initial network of from four to six major repositories. The number of repositories, along with the effective scope of the jurisdictions, could be expanded in later stages.

As a practical necessity, the choice of repository sites will be strongly influenced by existing concentrations of capital investment in the kinds of research capability and trained talent on which individual repositories must depend. This would seem to mean that minimal requirements for consideration as a site would include the presence of

(1) a graduate faculty in the social sciences which is already engaged in behavioral research with quantitative empirical data, along with substantial complements of graduate students receiving training in modern methodology; and

(2) a major computer installation.

The presence of a strong area program, particularly if it is behaviorally inclined, would be a great additional advantage. There are also some

considerations of geographic location. The site should be highly accessible, although some geographic dispersion for the network as a whole would seem desirable. Finally, areas boasting more than one graduate faculty which could contribute expertise to the local repository (as in the Boston-Cambridge or New York city areas) would have a number of distinct advantages.

Although data libraries are few at present, those which do exist are natural nuclei for more ambitious repository developments. There are three American institutions filling the requirements suggested above which already have initiated data libraries: Berkeley, Yale and Michigan. All three have great interest in coordination of efforts, have participated in our discussions and are signatories to this proposal. They would be expected to account for three of the major repository sites.

THE FACILITATION OF ACCESS TO HOLDINGS

The fact that a large store of potentially relevant information has been brought together at a repository is of help to the investigator interested in comparative research only in the degree that the repository has developed an effective cataloguing and data retrieval system. Modern information retrieval technology, as exploited by the physical and biological sciences, has now far outshadowed the traditional search capabilities of conventional libraries. This is not true as yet for social science, and it seems safe to say that even the best of the small but growing social science data libraries facilitates the search for information less well than conventional library indexing. Data classification problems are severe, and despite much individual concern, they have not been subjected to any concerted frontal assault. While such a frontal assault will be expensive

in the amount of substantive expertise and technical consultation required, such efforts will be maximally efficient if they are carried on cooperatively at the time when a number of large repositories are in their formative stages. Since the utility of these repositories hinges on facilitation of access in this sense, we propose to cope head-on with the problems involved.

Access may be limited in another sense, where the character of the archival holdings places a premium on the physical presence of the investigator at the site of the holdings, as tends to be the case with conventional libraries. Thus, while books are shuttled about to some degree on inter-library loan arrangements, physical distance from an adequate conventional research library severely penalizes the scholar. He is almost equally penalized for distance from data libraries as they are currently organized. However, in our estimation the basic currency of data repositories differs from that of conventional libraries in ways which, if properly exploited, can protect the remote investigator from penalty.

Most obviously, one does not "browse" through decks of punched cards or tapes as one browses through books. Hence, provided decks or tabulations from them are on rapid call, their physical location is immaterial. Since the supporting staff which this technology requires is more elaborate than that required for shelves of books, centralization of data holdings makes good sense. The closest thing to browsing in this modern data technology comes from a scanning of descriptive materials on holdings. While the archives themselves would be centralized, there is no reason why descriptive materials may not be reproduced and dispersed wherever there are libraries interested in handling them. This propagation would be taken for granted as one obligation of the repository network.

Some remote users, located at institutions without even a modest punched card processing facility, would in one sense remain handicapped. However, the number of such institutions is rapidly dwindling, and users in this situation could be better served than they currently are, for the repositories would be geared to provide tabulations from data stores as readily as they could produce duplicate cards or tapes.

The final practical barrier to access which affects use of current data libraries is that of cost of services. Benjamin Franklin's conception of the free public library has become so entirely assimilated that social subsidization of the machinery which ensures such access to information is never seriously questioned in the modern day. However, the conception is radical when laid against current data-library practice. One model of current practice is that in which the data library is so poorly endowed with operating support for staff and other fixed costs that it must attempt to recoup these in charges for services.

This not only raises a substantial barrier to genuine use of the holdings in sheer cost of service but entrains a variety of other inefficiencies. Thus, for example, one key characteristic which distinguishes a deck of cards from a book--its cheap and rapid duplicability--becomes an economic threat to the library rather than the tremendous force toward liberation of information which it should properly represent. The library dependent upon the "renting" of decks to cover fixed costs must take elaborate steps against the possible duplication of the materials outside the library (thereby robbing the library of a customer), and must require him to return the materials as a library requires the return of books. This is almost sheer waste, for the repository itself can in most instances

generate a new duplicate deck in better physical condition than the used deck for lesser cost than the return mailing.

We do not propose for the immediate future an operation in which all information-generating services are rendered free to each customer. However, we feel most strongly that a repository network of the type envisaged will be satisfactory only in the degree to which its fixed costs are subsidized, with the consumer paying only the immediate costs of his cards, tape or machine time. In the long run, we would suppose that such subsidization would begin to come through the institutional channels which currently support conventional research libraries. For the first five or ten years of operation, however, it seems clear that such subsidization must come from other outside institutions enjoying greater freedom to stimulate innovation.

It is apparent that each of the several practical barriers to access which we are pledged to reduce involves some additional expense. This means that an ultimate request for funds to support the development of the repository network would be a request for a substantial and broad-gauge contribution to the development of large-scale comparative research in the behavioral sciences for the United States. At the moment, however, it would be impossible to assess these costs for budget purposes without further exploration of site possibilities and more detailed consideration of a network design in which all interested parties can concur. It is to such an investigation that this initial proposal is addressed.

THE INITIAL PROPOSAL AND BUDGET

The limited funds which are requested in this proposal will be used solely for the purpose of developing an organizational and technical design

for the repository network. This will involve some preliminary technical consultation and an organizational conference between interested parties, along with staff time involved in preparation for the latter.

We would wish initially, during the summer of 1963, to broaden our contact with other scholars and other institutions who represent concentrations of demand for repository services, or who themselves are potential repository sites. We have some breadth already, inasmuch as our preliminary discussions have involved scholars from seven major universities scattered across all four regions of the United States.[6] However, it would be our intention to explore the repository plan with relevant parties at a number of other institutions, assessing the extent and character of consumer interest, and arranging a visit or other more extensive contact with a few institutions most likely to welcome a repository site in the early phases of the program.

These contacts in the summer of 1963 would lead in turn to a full-blown organizational conference, probably to be held in the fall. To this conference would be invited at least the primary representatives of units within institutions which have become, on the basis of discussions in the summer, potentially committed to participation as repository sites.

---

[6] Members of the ad hoc committee whose two conferences have led to this proposal include Professors S. M. Lipset and C. Y. Glock of the University of California (Berkeley); Karl Deutsch and Robert E. Lane of Yale; David Easton of the University of Chicago; James S. Coleman of the University of California at Los Angeles; James W. Prothro of the University of North Carolina; Erwin K. Scheuch, Visiting Professor at Harvard; and Warren E. Miller and Philip E. Converse of the University of Michigan. Intramural institutional arrangements which would facilitate the acceptance of a major repository commitment have already been completed for Berkeley and Yale. They are likely to be completed at Michigan in the near future. There is strong "demand" interest at the other universities, but further exploration must be conducted intramurally before it will be certain that the institution is prepared to accept responsibility for a site.

The conference might also include representatives of institutions who would expect to become major consumers of the network holdings, yet which for one reason or another did not find it feasible to participate as a site. In view of the working nature of the conference, we would hope to keep the number of institutions represented within a dozen.

The conference would have three main tasks:

(1) To modify and ratify a basic document describing the prospective nature of the repository obligations and their relations to one another. This document would be based in part on more detailed recommendations already developed by our group, and additional recommendations developing in the course of discussions during the summer or at the conference itself. Having thus established some collective entity, the conference would select a governing or coordinating body which would shoulder the executive functions of the collectivity for the remainder of the initial grant.

(2) To formulate a more ambitious proposal seeking the funds necessary to establish the repository network and to ensure its solvency over a longer period, probably five years. Prospective repositories would be the signatories of the final proposal, and signing would represent an institutional commitment to participate in the program if the proposed funds were granted.

(3) To make recommendations to the executive body concerning the most useful procedures to be followed in developing the technical aspects of the repository network design. These technical aspects involve a number of discriminable problems which require the attention of experts of differing backgrounds and technical skills, and include such things as the relatively substantive problems of data classification, standardization of code categories, and standards of data quality, or the more esoteric skills surrounding the newer automated information retrieval techniques.[7]

The conference would recommend either a sequence of small working meetings involving relevant experts, or the commissioning of individual

---

[7]With respect to the latter, we feel we could profit from establishing a relationship with an information retrieval expert as early as the summer of 1963, with the expectation that such an expert would become familiar with the general character of our problem and would be prepared to provide some initial technical guidance at the organizational conference.

papers developing technical recommendations. These technical departures would not necessarily be expected, during the course of the initial grant, to arrive at acceptable solutions to all of the problems involved. However, in the less recalcitrant areas they would arrive at some initial working conventions, and in the more recalcitrant areas or "dark continents" of the technical planning they would provide a realistic overview of the difficulties, thereby helping to pin down the budget items in a larger proposal which would be necessary if progress is to be made toward their solution.

We are requesting a sum of $23,760 covering a twelve-month period. The budget is as follows:

STAFF TIME	$ 2,500	
STAFF TRAVEL	1,000	
Organizational Conference (travel and expenses, 12 participants)	2,500	
Technical Consultation	1,000	
Technical Conferences (4) (honoraria, travel & expenses, 6 persons per conference)	9,600	
Secretarial	3,000	
Supplies	2,000	
		21,600
Administrative Expense		2,160
		$23,760

By common consent, the Survey Research Center of the University of California (Berkeley) would be the official recipient of the grant. The grant would begin as of June 1, 1963, and would terminate as of May 31, 1964.