

DATA BANKS FOR RECREATION SUPPLY AND PARTICIPATION¹

E. M. Avedon & S. L. J. Smith²

Data archives and data banks have become increasingly important as more researchers begin to examine trends. Characteristics of data banks, sources of bias in secondary data sources and important trends in data banks are described. The paper concludes with advice about using data banks.

INTRODUCTION

Studying outdoor recreation trends presumes a source of historical information upon which description and forecast of trends can be based. In the past, forecasters relied on private sources of data, on access to private and public agency records, and on published tables and statistical documents archived in research libraries for these purposes. A relatively new source of information (primarily but not exclusively quantitative) has become more accessible to recreation researchers in the last decade, namely, machine readable data and data archives.

Since World War II, the social sciences have undergone an "information explosion". This explosion is usually evidenced by the dramatic growth in the publication of books and journals. As great as this growth has been, it does not tell the full story. The social sciences, like the physical sciences have become more quantitative, and behind every scientific publication in the social sciences there are quantitative findings upon which the report is based. However, in the social sciences more data are often collected than ever find their way into published reports. This is common practice in many contemporary surveys and historical studies. In fact,

many elaborate and costly primary data gathering projects are undertaken without a clear plan for analysis of all data collected in the project.

In the past, such "excess data" would have been destroyed, but today such "excess data" represent a valuable storehouse for future research and planning and policy analyses. Information is stored without analysis or interpretation on punched cards, magnetic cards, paper tape, magnetic disk, magnetic tape, microfilm, and other mechanical or electronic media to facilitate retrieval. Unlike information stored in published form, such as the familiar printed tables of a national census, emphasis in a data archive is on rapid retrieval, custom rearrangement, processing, and summarization. Such flexibility is a boon to the myriad of potential users of any data set. Academics, planners, management consultants, entrepreneurs and others can examine and use the same data set for widely differing purposes with equal ease. Although there may be different missions and policies in different machine readable data archives, and although information may vary from bank to bank, usually a data bank does not direct or control the types of information a user seeks. The secret of success and usefulness of data banks, if there is a secret, is to emphasize technology over teleology.

¹Presented by Dr. Smith at: OUTDOOR RECREATION TRENDS - A NATIONAL SYMPOSIUM, Elliott Alumni Centre, University of New Hampshire, April 20-23, 1980.

²Authors are Director and Research Associate, respectively, of the Leisure Studies Data Bank/Banque de Données sur les Loisirs, University of Waterloo, Waterloo, Ontario, Canada.

The concept of a data bank took hold in many facets of industrial society over the past thirty years. Airline reservations systems, department store accounting records, warehouse inventories, banking statements -- are just a few well known examples of this phenomenon. However, industrial and commercial data banks are designed for single-

purpose use, by highly trained users, based upon rigidly efficient and economic systems. Within the social sciences, it was recognized that data have multiple research uses, and although some potential users may be skilled methodologists, many are relatively unsophisticated users - particularly in academic settings. Thus social science data banks could not be designed as a single purpose entity, but rather had to be designed as "banks" with multiple functions. Furthermore, data elements destined for industrial and commercial data banks were collected and prepared within the objectives of their single purpose systems. Social science data on the other hand continues to be generated through a variety of sources, and each source has its own research "perspective".

Data banks developed as a practical response to the need to handle a flood of information. One major contributor to this was the rapid development and expansion of computer technology. Social science data in the 1950's and 1960's paradoxically rendered the information generally irretrievable to the unaided researcher. To overcome this, scholars and decision-makers began to cooperate to pool information resources. They found, however, that such a project easily became time-consuming, costly, and involved great practical difficulties. If an individual did not have personal knowledge of a particular information source, the research literature often had to be scanned to track down possible sources. Once the required source of the data had been identified, it was necessary to determine if these data were still in existence, and how accessible they were to the researcher's computing resources. Often it was found that the source had not been sensitive to the possibilities of further analysis on the data collected, and as a consequence, took no pains to store the information in a manner as to allow universal retrieval and re-use.

When such data were available, they were often stored in an idiosyncratic fashion, accessible only to the source, and after a while, not comprehensible even to the source. Coding may have been ambiguous, or worse, undocumented, rendering the data useless. Formats, definitions, and terminology often reflected local, arbitrary conventions. Data documentation might contain mis-codes, missing records, and labels in a foreign language. To clarify, clean, and edit these records required more time, money, and technical expertise and sophistication than most individual researchers have. It became evident, not just to individuals, but also to universities, governments, and other agencies that retaining archivists and related personnel would be necessary to acquire, prepare, store, and retrieve machine readable social science data on a con-

tinuing basis. This recognition led to the establishment of the specialized data banks for the social sciences that exist throughout the industrialized world.

Characteristics of Data Banks

The parallel between data banks and research libraries is obvious, but there is a divergence more important than any similarities. Data banks do not attempt to archive all available information related to their specific mission. Each bank has some identified theme, and collects data in keeping with that theme. Sources of information are diverse, to the point that an "outsider" may not immediately recognize the relevance of a particular data set to the archive's mission. For example, the Leisure Studies Data Bank at the University of Waterloo primarily acquires machine readable data concerning leisure-related phenomena. This perspective normally includes survey and administrative data regarding facility or resource use, expenditures, tourist origin and destination information, performing arts audience information, park attendance information, and the like. Some information is donated by expected sources, such as the Ontario Ministry of Culture and Recreation, but data are also sought from less obvious sources such as the Federal Ministry of State for Science and Technology. In this example, we obtained a survey of the impacts of science on Canadians that included useful information about attitudes toward television viewing. Another social science data bank may acquire data only regarding political attitudes, such as the Roper or Gallup poll data, and the extent of leisure-related information in that data bank would be minimal. Some data banks acquire data for specific geo-political areas only, rather than for a specific theme. When the holdings of a bank are examined, new perspectives on both themes and data sources are possible. Collecting information on different aspects of leisure not only creates new perspectives and research possibilities, it creates a collection of skilled personnel. Data banks become the loci for contacts among people trained in data collection, data management, and data analysis. Exchanges among technicians, scholars, and policy-makers provides a rich environment for all. A data bank is more than just a warehouse of numeric information, it is a source of assistance, insight, and inspiration for the researcher.

Data are initially obtained in a variety of forms and structures, from simple tabular reports through decks of standard punch cards, complex multi-punched formats to edited and labeled variable spanned matrix system files. Most banks in archiving a data file, store information in a single medium (commonly magnetic tape), that can be read easily and

worked efficiently. Generally, data banks do not place restrictions of the size of a data set that can be archived, and although most data sets are of a manageable size for research or re-analysis, storage can become a problem when a collection grows large. Large data sets are more often a processing problem for the researcher, than an archiving problem for a data bank. Tape storage is usually in a central computer tape library under climate control and access is protected by a stringent security system. The former is to prevent physical decomposition of the tapes, the latter to prevent unauthorized use of a data file, or the inadvertent destruction of a file by an unsophisticated user. Because data are machine readable, physical distance from the tapes, or for that matter from the computer, is not a problem. Access is accomplished through electronic communication, and all a user needs is a small computer terminal, a communication link, and a telephone number. For example, although we are hundreds of miles from the Leisure Studies Data Bank -- with a terminal, a coupler, and a telephone we could access and process any of the Bank's holdings from this room! This is not unusual. At the Central Social Science Archive at the University of Cologne, users are in a building on the campus in Cologne, but the computers and the tapes are in the city of Bonn, kilometers away. At the University of Odense at the Danish Data Archive, on the island of Fyn, one of their computers and its adjacent tape library is on the island of Zeeland, in a town north of Copenhagen. At this time many researchers have access to a network of information from a number of data banks, and such practice is becoming easier as technology is developed.

Although stored numeric data form the heart of a data bank, these would be useless without supplementary documentation which explains, for example, that the fourth through the ninth columns in each record gives the total population of a region; or that specific concentrations of magnetic oxides at a certain location on a tape indicates that information concerns swimming. Documentation may be in print or may also be machine readable. In fact, the latter is becoming more common place to permit greater access to data by users from a distant terminal.

At a minimum, documentation provides the following types of information:

1. description of data structure;
2. description of data format;
3. examples of both structure and format;
4. size of data set;

5. definitions of data elements;
6. explanation of abbreviations and codes used;
7. description of sampling design and technique, substitution procedures, etc.;
8. non-response rates and weighting procedures
9. source statements that generated data, instrumentation, tests;
10. bibliographic citations for publications based on use of the data set;
11. list of related data sets;
12. names and addresses of personnel or the agency responsible for collecting the data;
13. special information regarding access or processing.

To provide data and documentation to users, data bank staff members are called upon to perform a variety of tasks. They must provide enough information about the availability and contents of holdings and how to retrieve the required data set. Behind these obvious tasks are many hours of effort that are invisible to the user. The process of archiving a data set so that it can be used has become a highly technical and exacting skill. As a bank gains visibility and its staff matures, they are expected to serve as consultants to potential users not only with respect to retrieval of data sets they have archived, but with respect to computer software, statistical procedures, and eventually even research design. In time, they are called upon to make recommendations about the process of primary data collection and storage for eventual deposit in a data archive.

Sources of Potential Bias

Because of the wealth of data available, and the pressure on archival staff to provide ever greater detailed and technical advice, data banks specialize and refine their official mission. Specialization produces an inherent bias through the type of data available, who donates data to a bank, and who would likely use a bank. Although "leisure" as a subject is a specialization within social science, some archives have close, continued relationships with certain donors, or have continued experience with certain types of users, and consequently specialize within the field of leisure studies. One bank may become stronger in the area of the sociology of leisure, another might become more skilled in the economics of leisure; some collect data only on user patterns and visitation, still others might focus only on subsets dealing with sport.

Banks may also specialize in the scope of their data. Some are depositories of national and international studies. Others might serve a national clientele, but limit holdings to provincial or state data. Often, this is not the result of a single, conscious decision by administrators, but rather the result of responses to opportunities and requests. Whatever the cause, banks generally become more specialized over time.

In addition to bias arising from objectives and specializations of a data bank, there is bias imposed by the archivists' decisions about the quality of data and documentation. A decision not to include a particular data set is based on a number of reasons. Data may be of questionable value. Records may be missing. There may be coding errors, biased sampling, ambiguous questions, or problems with instrument design. On the other hand, qualities of a file might be adequate, at least for the original purposes of research, but documentation may be inadequate or missing, and this prevents further use of data by other researchers. A decision to archive or not archive is a technical one, made on objective grounds. However, the evaluation of whether a data set or documentation meets the objective criteria is often a subjective decision made by data bank personnel; and the quality of that decision depends upon the knowledge, skill, and ability of the specific staff.

If a data set is archived and made available for use, it is not guaranteed to be free from error or distortion. The purpose of a primary data gathering project, the wording of questions, the sampling frame and design, definitions of words and terms and how substitutions were made for non-respondents, the basis upon which a test has been standardized, interpretations or shifts in meaning by the original coder, all can cause bias in reliability and generalizability. This is why documentation is so important as a part of a data bank's holdings. As a user, you should be able to assume that the data bank staff has acquired, cleaned, stored, and retrieved data properly. You cannot make any assumption, however, about the inherent quality or characteristics of the data set without examining the associated documentation.

Access rights are another source of possible bias, not so much as a distortion in the interpretation of a particular data set as it is a distortion in the information available from a data bank. The Leisure Studies Data Bank does not accept any file that is restricted to only certain users. For example, we were given a copy of a survey of violence in a specific sport. Shortly after receiving the file we were advised that this information

could not be released to all classes of potential users, and that only "approved" researchers could have access to these data. It was therefore decided to de-archive this file. The decision to include only publicly available information limits the data that can be archived, but it does ensure all potential users of availability and access. This is not always the case in other social science data banks.

The final source of possible bias is that of the donor. Here we refer not only to the types of information or questions that a collector gathers, but to the original treatment of these data. Some data are distorted before being released. In other cases data are "laundered". The process of "laundering" alters the validity and accuracy of a file. The level of aggregation of observations is another reflection of each nation's political concerns and its perceptions of privacy, social responsibility and individual rights. In North America, as in many parts of the world, it is not possible for the ordinary data user to identify specific individuals by name or address. Normally disaggregation is possible only to a subgrouping short of the individual case level. Privacy extends to protecting the identity and responses of corporations as well as individuals. Protection of identities means more than just eliminating names, addresses, case numbers, and some geocodes, it can also mean aggregating responses from small or lightly populated areas to thwart attempts to deduce the probably identity of a person or corporation. For example, if you were studying private campgrounds, and had the responses of an owner in a specific local region, state, or province that reported gross income and expenditure, you may be able to narrow identity to one or two campgrounds. To prevent this, disaggregation to a local region may be limited during the archiving process.

There is substantial variation among banks in different countries with respect to data availability. In one European nation researchers can only obtain the most generalized, averaged figures for most social statistics of their population; whereas in a neighbouring country, data files are so specific that it is possible to link individual income tax returns with responses on other social surveys to check on the validity and reliability of responses concerning income and expenditure on the social survey. Many governments retain registries of disabled persons which are available for research purposes from public agencies; this is generally not the case in North America. Some data banks regularly receive official government data for permanent archiving with the intent of providing wider access of data to researchers. Some banks have no liaison

with government and have access only to academically generated data. Still others may only have access to commercially produced data. Still others may only have access to commercially produced data. Availability of social science data for research thus differs from nation to nation, and this bias influences the scope of data a researcher has for analysis.

Trends in Data Banks

One of the most important trends is the apparently contradictory tendency to become both more specific and more general. The growing specificity of a bank is the result of continued relationships with certain donors and users. A bank that develops a good working association with particular agencies, tends over time, to specialize in the information these agencies provide and need. A subtle, but important force thus slowly influences the mission of each archive. Growth of other data banks into related subjects also encourages specialization.

At the same time, there is a broadening in the perception of data files useful to researchers in a specific subject. In past years, only data files obviously related (as indicated by title of a file) to the mission of a bank would be archived. There is now a recognition that data files from unlikely sources can be of great value. For example, we have recently obtained a copy of projections used by the local public school board for educational planning. This file contains information on the number, gender, and ages of children expected in the regional population in years to come. Although this information was collected for educational planning, it is also valuable for doing feasibility studies and needs assessments for public recreation facility planning.

Another trend in data banking is toward use of more efficient and generalized technology. The newest generation of computers, and the availability of new memory technology will speed the time necessary for data processing, reduce computing costs, and increase available storage. Software packages, such as SPSS, SAS, BMD, OSIRIS, and others have been improved to match improvements in hardware and operating systems. It is now possible to work with files that would have been considered monstrous only a few years ago. The Leisure Studies Data Bank, for example, regularly assists users in working with files that contain over 1,000 variables or that have as many as 50,000 cases.

In addition to greater power and efficiency, there is a trend to greater flexibility and compatibility. Work is proceeding to develop a universal interchange file that will facilitate linking data sets or the output of one system with any of several software

packages.

Just as data banks were developed to help researchers cope with data, archivists are beginning to see a need for providing assistance in helping potential users cope with the growing number of data banks. The first step is the development of a system to allow a potential user to query a bank's holdings for information about a specific topic, geographic region, or other characteristics of a data set. Because of the kind of specificity inherent in different thematic research approaches, efforts are underway to develop hierarchical modes of inquiry that may be shared among several cooperating data banks. A user at the Leisure Studies Data Bank for example, will one day be able to browse not only LSDB holdings, but leisure-related data that is part of the holdings of other universities in other countries. The significance of this system of hierarchical study descriptions or file precis is not only to allow an efficient search to be made, but to provide for common terminology and descriptions among several archives.

This growing cooperation is international in scope, and thus there is a trend in the polylingualization of archives. The international language of computers may be English, but file precis, variable labels, catalogues, and the like will need to become available in all the major languages of the world. Several data archive organizations have been established to encourage system and file exchange, cooperation and consistency among member archives. In 1965, as one example, the Council of Social Science Data Archives was established to further these goals among two dozen United States archives. Unfortunately, the differences of opinions among members was so great that the council collapsed. This is a problem and a challenge to data banks in the United States. In Canada, data archives, government, academic, and private are members of the Canadian Data Organization Committee of The Social Science Federation of Canada. A similar organization exists for Western Europe. One of the most important organizations promoting inter-archival cooperation today is the International Federation of Data Organizations, an associate member of the International Social Science Council - a UNESCO organization. Member archives are from both east and west Europe, the United States, and Canada.

Perhaps the last major trend to cite is the growing importance of data bank personnel as research consultants. Because of familiarity with different problems in research design and analysis, they acquire an overall perspective on the production and use of new data collections. Our own staff have provided consultation to a number of government and private organizations on the design of data gather-

ing projects, on coding and weighting of data after collection, and on other related matters. Data banks also have the potential to organize groups of individual users to pursue new lines of inquiry. Because of the potential to serve as "spokespersons" for both data and computer users, these personnel can help to provide the impetus for developing computer software systems, and can become effective spokespersons for social scientists to the computer industry.

Using a Data Bank

Users of data banks fall into a number of relatively well identified groups: the unsophisticated new researcher who has not been a data user, and knows almost nothing about computers; the researcher who has had some primary research experience and some familiarity with data, but little computing experience; the experienced researcher who has considerable methodological skill, and computer literacy; the sophisticated researcher and computer user. Each of these types of users require different types of assistance from a data bank staff. The more unsophisticated a potential user is, the more likely the first few visits to a data bank will be a "fishing expedition". The more sophisticated a user is, the more specific and technical is the use of a data bank. Preliminary visits to a data bank by any user concern documentation rather than data, regardless of the level of sophistication.

In order to ensure that users have access to the archived data that will be of most use to them, documentation is usually organized on five levels.

1. File Identification -- a user may discover that a data file exists from an entry in a library catalogue, an inventory of data sets, a data bank catalogue of holdings. These citations are often cryptic, and will often include only the name of a file, and some general identifying information.
2. File Precis -- a user, upon discovering a data file that may meet research needs, then proceeds to examine more detailed information about a file. File precis are available in some library reference rooms, in data bank offices, and many are available in a hard copy form that can be mailed from a data bank to a potential user. Some banks offer machine-readable precis that can be examined on-line through an interactive computer system at a CRT terminal. A file precis describes the data, presents sampling and weighting information, indicates types

of variables in the file, lists published reports based on these data, and provides a summary of the research which generated the data.

3. Source Documents -- after examining the file precis, a user may wish to consult some of the sources cited in the precis. Depending upon the practice of the bank, some data archives in collaboration with libraries have catalogued data files in the same manner as related source documents. Thus a user can examine related sources documents within a library or through inter-library loan by using the same catalogue classification codes. In many instances, a data bank will also have copies of these source documents, but these are usually for on-site use.
4. Variable Lists -- generally these are available for in-depth examination of a data file for use when designing a research plan. Hardcopy lists are usually available for each file and can be mailed to a potential user. Sometimes these lists include the source statements (from a questionnaire or psychological test, etc.) which generated the variables in the data set.
5. Codebooks -- offer the user who has developed a research plan for use of a specific data set detailed information regarding the characteristics of the file structure, a complete listing of code elements specifying the variables and all values for each variable. Univariate tables in a codebook specify the frequency for each value within a variable, precise information about codes needed for processing, matrix information, CPU requirements and the like. Codebooks may be in hardcopy and sent through the mail, or may be machine readable and available on-line locally or can be used at a distance from the bank with the necessary hardware and software, and software systems documentation.

A word should be said about the nature of the files that are usually available within a bank. There are three basic types of files: raw data files, edited system files, and process produced files.

1. Raw data files are files that have been cleaned for errors, wild codes, etc. These files are stored in their cleaned "raw" form. Generally they are available to a potential user that wishes to write a special analysis programme for these data, and does not wish to use a package programme.

2. Edited system files are files that have been put through the archive process and been prepared to be used with a package analysis programme such as SPSS, SAS, OSIRIS, BMD, etc. There are often different versions of these files, such as simplified editions for new student researchers that may have some of the ordinal data grouped and labeled for simplified analysis, or the file might be reduced to a sample of cases to enhance useability when there are large numbers of records. Or a special file might have been created with an inverted matrix for use in certain factor analytic programmes by experienced researchers.
3. Process produced files are computer-generated files, based upon a researcher designed programme. Input for these files will be variables from a number of different files. The final file will be a "raw" data file, distinguished from researcher-collected data because data collection is actually internal to the computer used.

The process of using a data bank is generally the same in most locales. First a user needs to have a computer account number in order to access the required data. Access in many places today can be either through batch mode or through an interactive system. A user can either do all of his own computing or can seek assistance from data bank staff. In effect, the process is analogous to designing any research project, the only difference is that after designing the project (in this instance with the use of the data documentation) and instead of going out to collect data from households or on-site, or from some administrative reports or documents, the researcher writes a computer programme and collects the required data from the computer--data that have been stored by the data bank. Although this seems somewhat oversimplified, any working researcher knows just what all that simple formula can involve.

In many instances, there really is no need to expend the time and money to collect new data, particularly concerning leisure-related topics because there is a wealth of data available today that has never been subjected to analysis. Perhaps the biggest problem leisure researchers face today is identifying where useable data reside. Because finding these data is so difficult, the tendency for researchers is to develop research designs that include new data collection. This is not only a common practice within the field of leisure studies, but throughout the social sciences. However, as data banks become more common in universities, and are

used more frequently as laboratories in the teaching of social science research, there will be a greater tendency for new researchers to think about analysis of secondary data before embarking on new data collection.