APPLICATIONS OF INFORMATION SCIENCE TO SOCIAL MEASUREMENT

by

Murray Aborn

National Science Foundation

This paper was presented at the IASSIST ANNUAL CONFERENCE, May 19-22, 1983, in Philadelphia, Pennsylvania.

In the 1960's, the growing influence of the computer caused dramatic changes to take place in the concept of scientific data and the character of data analysis. Among these changes was the onset of a shift away from single-purpose data collections and analyses based on relatively small data sets, toward large-scale data collections and analyses based on data banks serving multiple applications and possessing widely accessible storage and retrieval systems. In social science this led to the establishment of data archives and an early attempt to regulate their functions. Additional purposes were to keep these facilities abreast of a rapidly advancing technology, and enable them to remain *au courant* with increasingly sophisticated management schemes for operating over larger and larger bodies of data (1). This paper briefly traces the role of the National Science Foundation (NSF) in these developments, discusses the current state of affairs with respect to social science data resources, and questions whether continued reliance on sheer data amassment is the true path to the further intellectual progress of the field.

THE BUILDING OF NSF DATA PROGRAMS

In the years immediately following the advent of the computer-based data archive, NSF involvement in the expansion and upgrading of the major sources of social data grew and intensified. To increase the research-return from the enormous investment society makes in the collection of social statistics, projects were supported to enhance the researcher access to them. To fill the gaps in the social science data base, projects were supported to maintain data series not covered by the federal statistical system but needed for the monitoring of social and economic trends and the modeling of long-term social change. Direct support was afforded to archival facilities to help them expand their holdings and degray the costs of dissemination. On the user side, projects were supported to increase the research utilization of stored social data across disciplines, including projects to introduce bibliographic-type control over machine-readable data files in order to reduce duplicate data collection and help prevent incomplete data analyses. And alongside these data programs, projects were funded to improve existing methods and create new tools of broad utility in analyzing the growing stock of social and economic data becoming available to the research community.

In 1976, a committee of the National Academy of Sciences surveying the social sciences at NSF acknowledged the role NSF programs were playing in the sphere of data resources. It declared:

> *It is generally felt, and reflected in the long-range plans of several of the social science programs in NSF, that deficiencies in the available base of social science data are seriously impeding the progress of research* (2).

The Committee did not refute this outlook; in fact, it ultimately recommended that such planning continue and include greater support for longitudinal studies over extended time periods and national facilities for survey research and large data bases.

Financial backing for the data programs described above was provided not only by NSF's Division of Social and Economic Science, but by sections in other parts of the Foundation, such as Computer Science and Information Science. Over the next six years, however, Computer Science and Information Science turned inward, concentrating on their own disciplinary development and gradually eschewing applicational extensions to other fields of science. But in social science, the work went on. Programs were maintained that to this day continue to build a data resource infrastructure capable of sustaining the large empirical research tradition which characterizes contemporary social research.

IS IT TIME FOR A CHANGE IN ORIENTATION?

The National Academy of Sciences committee surveying NSF's social science programs in 1976 never made crystal clear precisely what was being referred to by "*deficiencies* in the available base of social science data," and which of these were more or less responsible for "impeding the progress of research." It is pretty apparent from the committee's report, however, that data resource planning in social science was largely oriented toward filling topical gaps and producing lower levels of aggregation, larger-scale, longer-term data gathering efforts, and a more systematic approach. Though the importance of methodological accompaniments to assure good data quality was neglected neither in NSF's programs nor the committee's report, the effect of stressing data gaps and data shortfalls inevitably leads to more and more data getting collected and more and more data being retained--and that is exactly what has happened.

Now, one question that arises at this point is whether an orientation toward data amassment has had negative as well as positive consequences. The answer is "yes." If negative consequences is too strong a term, then we can at least speak of limiting effects. And if there have been negative consequences or limiting effects, then it <u>is</u> time for a change in orientation. But before proceeding to describe what the required change appears to be, it is crucial to make clear that such change in no way gainsays the compelling arguments put forth in many recent publications regarding the value of secondary analysis. Nor does it gainsay the need to have data available for reanalysis in order to test for bias in reported results, challenge data-driven theoretical assertions, and generally carry on the processes of scientific understanding in a field which is rarely able to conduct controlled experiments or reproduce the original conditions of an investigation. A change in orientation simply argues for diverting some amount of effort and devoting some portion of available resources to study the deeper aspects of the enterprise in which the field has become heavily engaged.

One negative consequence of the data gathering enterprise has been the pejoration of the term "data." This is no doubt connected with the fact that the enterprise is largely concerned with quantitative data in computer-manipulable form, but in any case the term data is now commonly used interchangeably with the terms observations, information and, worst of all, evidence. I daresay few really believe that data in and of themselves prove anything, but that's the way we have come to talk and, I fear, occasionally think. However, the more frequent tendency is to confuse data first with observations and then with information. I realize this gets pretty elementary, but contemporary social science data archives contain mostly recorded observations, not data. It sounds more imposing to speak of *data* archives, and it is certainly easier to raise money in the name of data than it is for just plain old observations, but the terminology is inaccurate. Observations become data only after they are placed in some analytical framework. As it is obvious from the general-purpose nature of the data archives, the same observations are destined to be interpreted as more than one kind of data.

A similar confusion prevails with respect to data and information. The two are not synonomous, though the exposition here is very difficult inasmuch as the relation is inferential and dependent upon the application of external structures. Simply state, it behooves us not to forget that any body of data is a mixture of information and noise, and that the proportions will vary according to the use to which the data are being put.

In the main, the signal to noise ratio in social science is typically much lower than in the physical sciences, which is a way of saying that the information content of a data base can be very meager, particularly when the data are employed to test hypotheses far afield from the hypotheses which motivated the data collection originally. It is thus ironic that the very success of large-scale, integrated data bases and the attendant data-processing technology often leads to a confusion of the technology with the natural semantics of information, which is heavily context-dependent. Thus the underlying assumptions appropriate to the context of one application may be totally inappropriate to the contexts of other applications. Moreover, the difficulty is compounded by the fact that in their research, social scientists are heavily dependent upon data files which were not generated for scientific purposes, such as census data, voting records, police and court records, governmental budgets, and so forth, and whose informational value relative to the kinds of scientific questions social scientists ask may be completely uncertain.

BRINGING INFORMATION SCIENCE INTO THE PICTURE

In the previous section of this paper, mention was made of the ancillary role played by the computer and information sciences in the building of NSF data programs. It was noted that those roles diminished after 1976, and that NSF's contributions to the data resource infrastructure of present-day social science has been carried on exclusively by the social science elements of NSF. This situation is changing. Given recent advances in information science, it seems particularly important to begin to apply newly-formulated principles of knowledge management to social science data resources precisely *because* their holdings--observations of social and behavioral phenomena in digital form--tend to be incomplete, imprecise, and error-prone due to the fuzzy nature of the phenomena being observed and the looseness of the data gathering process. Knowledge management facilitates the translation of user needs into expressions upon which a data base system can act. One example of possible applications to social data is the development of data base specification languages, that is, languages which would permit social science researchers to express their requirements in functional terms. These might then be translated into a database format, perhaps based on relational structures rather than representational ones, as is the present mode, which would help skirt the data dependence problem.

Other areas of potential application to social science data may come from information science's concern with descriptive classification, indexing, and the problems of relating variant terminology in a single retrieval system. The current work of Dolby is an example (3). Dolby argues that the correctness of data and data analysis involves correctness in meaning, and that correctness in meaning goes beyond matters of computer program correctness or the numerical accuracy of data. His approach concentrates on the use of classification structures to extend the formal treatment of meaning in computer-based data systems, and he has shown how such extensions can expose or reduce ambiguities and inconsistencies of meaning in such systems.

There are some other, more practical reasons to believe that the time has come to test out achievements in information science as they may be applied to stored social data. Urgencies created by current reductions in the quantity of social science-usable data generated by the federal statistical system is one reason; cutbacks in the funds available to support scientifically oriented social science data resources is another. It would help greatly if we could improve our ability to estimate the degree of redundancy (i.e., the amount of information overlap) among data collections, and if we could make progress in our ability to set data collection and maintenance priorities.

Considering the potential benefits of bringing information science into closer contact with social science data problems and opportunities, it has been decided to launch an initiative--still informal at this juncture--to make known our receptivity to proposals which combine or merge the subject matters normally covered by social science and information science independently. Such proposals will be handled jointly by NSF's Division of Social and Economic Science and its Division of Information Science and Technology (4). The Division of Social and Economic Science supports the establishment, evaluation, and improvement of social science data resources, research *on* social data, and the development of methods for analyzing such data. The Division of Information Science and Technology supports research to increase understanding of the properties of information transfer. We believe the future will show that this initiative was well advised.

Notes and References

(1) See, for example, Glaser, William A. Note on the work of the Council of Social Science Data Archives. Social Science Information, 1970 8(2):159-176. It may be of some historical interest to point out that the Council of Social Science Data Archives was the forerunner of IASSIST.

(2) Social and Behavioral Science Programs in the National Science Foundation. Final Report of the Committee on the Social Sciences in the National Science Foundation, National Research Council, National Academy of Sciences, Washington, D.C., 1976.

(3) Dolby, James L. Meaning from data: Implications for data analysis and database management systems. Paper presented at the meeting of the American Association for the Advancement of Science, Detroit, Michigan, 1983.

(4) Inquiries may be addressed to: Program Director for Measurement Methods and Data Resources, Div. of Social and Economic Science, NSF, Washington, D.C.