Data Services Since the 1960s: Where are We Going?

by David Elesh[1]
Center for Public Policy,
Social Science Data Library,
Temple University,

Introduction

Twenty-five years ago, I wrote a proposal to the National Science Foundation for a Workshop on the Management of a Data and Program Library to promote the establishment of local university social science data archives. Although organized in only a few weeks, the Workshop attracted 100 people from 50 different institutions in the U.S. and Canada. We even had one person from Sweden, indicating the deep roots of Swedish Aarchival data activity. A few institutions already had data archives, and within a very short time, all did.

In preparation for my remarks here today, I went back to the Proceedings of the Workshop we published shortly afterward[2]. In the "Preface" to the Proceedings, I outlined what I thought were the benefits of establishing a data archive. For faculty, the benefits appeared to be greater productivity, the potential for investigating new kinds of problems, and the ability for scholars with limited funding and resources to access high quality data. Graduate students could be given a greater opportunity to gain research experience, and they shared with faculty the possibilities of examining new kinds of Problems and high quality data.

However, I reserved my greatest expectation of benefits for undergraduates. Because of the expense and time required for empirical social research, undergraduates in the sixties rarely experienced a genuine introduction to the social sciences as research disciplines. Instead they read brief and, often, highly simplified synopses of research that gave little indication of how quantitative social science is done. The ready availability of data and program libraries would, I thought, make possible realistic introductions to the social sciences.

Looking back now more than two decades, I think it fair to say that data archives have made a significant difference to faculty and graduate students and have not had much impact on undergraduates. For faculty and graduate students, data archives have vastly increased the amount of comparative and over-time research published. Better data also fostered greater sophistication in social scientific theories and analytical techniques. Twenty-five years ago, economists aside, most social scientists were content with theories based upon cross-sectional data, cross-tabulations, ordinary least squares, and path analytic techniques which required no more than ordinary least squares. We now find publications and dissertations dealing with dynamic theories based upon event history analysis, partial adjustment models, dynamic LISREL type models, and more[3]. Most of this work was made possible by the collections of data archives.

However, the impact of data archives on undergraduate education has been modest. I will acknowledge that data archives significantly altered patterns of instruction in undergraduate methods and statistics courses, but in most colleges and universities, these are "required" courses segregated from the substantive foci of their disciplines. In all too many cases, the links between these courses and substantive courses are left to the imagination of the students. The vast majority of undergraduate courses today are taught in a manner little different than decades ago. Faculty lecture about research that students find summarized in their texts, and the investigatory process that produced the results about which students read is about as much an enigma today as it was then. As a result, the skills taught in the methods and statistics quickly grow stale.

Ironically, the achievements of data archives in enriching faculty and graduate student work often has led to an institutional success which works against serving undergraduate education. Twenty years ago, data libraries had little or nothing to do with conventional libraries. They were creations of faculty members seeking to provide research and instructional resources for themselves, their colleagues, and their students, and they were housed in faculty offices or a few rooms down the hall. By and large, librarians did not understand computers and data and often were uninterested. Much has changed. At contemporary meetings of Inter-university Consortium for Political and Social Research Official representatives, there seem to be as many librarians as faculty—certainly they form a substantial fraction of those attending the meetings. Libraries are moving to accept data archives as important parts of their collections—a change that represents institutional commitments to data libraries as scholarly resources. What, then, is the irony? It is that even as libraries take on this new function, university and collegial support for libraries is either static or declining.

From 1975-76 to 1985-86, current fund expenditures for libraries in institutions of higher education fell from 3.1 percent of total expenditures to 2.6 percent[4]; and this occured during a period in which library costs for books and periodicals increased roughly 60 percentent faster than overall inflation.

Quite simply, the cost pressures universities have faced for more than a decade show no signs of improving soon, despite all the professed concern about the state of American education. This means that libraries are unlikely to have the kinds of staff resources required to help undergraduates use data resources. In fact, there is the real possibility that all users will suffer.

But the picture is not completely bleak. Much is changing in social scientific instruction, and an increasing number of undergraduates are gaining experience in doing quantitative social science. The introduction of microcomputers is slowly—very slowly—transforming undergraduate education. Texts increasingly come complete with analytical software and databases, and independent instructional packages and databases such as ShowCase have been adopted widely. Both types of innovations make it possible to introduce real, quantitative social scientific work to both lower and upper division undergraduates and usually find enthusiastic acceptance. But neither involves or leads to use of data archives. Why?

First, it is important to recognize that data archives were conceived in an era of mainframe computing and were meant to be analyzed by mainframe statistical packages. The fundamental meaning of this statement is that the knowledge base required to use these resources is simply much larger than for a PC. The architecture, organization, and funding of mainframe computing are designed for the researcher, not the instructor, and certainly not, excepting computer science students, the student.

Mainframe operating systems are far more sophisticated than those available on PCs. Mainframe statistical packages, while powerful, are typically intimidating in their complexity, and even those of us who routinely have required undergraduates to learn these packages sufficiently to get through our methods and statistics courses know that we must sacrifice some content to allow time for teaching basic computing skills. The learning curve for mainframe computing is a great deal steeper than for PCs.

In the past, we could defend the loss of statistical or methodological subject matter in the belief that knowledge of SPSS or SAS and the like formed part of the research skills we were trying to impart. Those of use who used the computer in our undergraduate instruction learned to create program and/or system files that shortcut many of the procedures we expected graduate students to learn. We used archived files because there were few alternatives, and setting a file up for a class was little different than setting one up for our own research use. But the skill level and time it requires to do these things are significant, and many social scientists simply did not and still do not have them.

Nor would they or their students find much help in the organization of computing. Because the machine or machines were located centrally, it was and is almost universally true that consultants were as well. One had to go to the Computing Center to use computers or seek assistance in using them. At the same time, the available consultants were and are typically programmers unfamiliar with statistical software and social science data. To a very large extent, users must learn the consultants' language in order to obtain help; they do not learn users' language. A few consultants might learn SPSS, SAS, BMDP, or the other statistical packages, but the demand for their services always exceeds the supply because social science users of the computer are greatly outnumbered by users in computer science, engineering, and the physical sciences, and the latter have the influence that attends greater external funding; thus central computing budgets favor the latter over the former.

Local data archives often tried to fill the void by offering assistance in use of the computer as well as of the data. Staff became expert in the use of tapes and the manipulation of large and complex files; in some institutions, they provide and have provided the primary consultative assistance in these areas.

Against this background, it is not surprising that analyses of archived data did not spread widely in undergraduate instruction.

While I think there is little doubt that the introduction of microcomputers can transform undergraduate instruction, I have some doubt that data libraries will be significant actors in the transformation. Micros eventually will succeed in changing social scientific instruction because they significantly lower the slope of the learning curve for computing. Students find PC operating systems easier to learn, and unlike the mainframe world, there are statistical packages specifically designed for instructional use which require far less faculty and student time to learn. However, generally these packages incorporate data which has been tailored for them, and the tools necessary to include other data sets are omitted. One can even find software that allows the student to place a disk in the lowliest PC, turn it on, and find him or herself in a menu driven analytical package capable of multivariate crosstabulations on a substantial number of variables

with adequate samples and with virtually instantaneous response.

Microcomputers also introduced a new market structure for computing and data. In the mainframe environments, computing is provided and funded centrally as a university or college function, and instructional costs are, at least partially, borne by tuition. Data files are also provided centrally and usually cost users nothing. However, with the introduction of microcomputers, the cost of the hardware, software, and data are increasingly being borne by the user as direct charges. Where universities or colleges supply microcomputing laboratories, the number of these institutions that have introduced "laboratory" fees to cover these costs grows with each passing year. And, as noted, software and data increasingly come either from text publishers or other third party vendors.

Clearly, publishers are seeking to make it significantly easier for students and faculty to analyze data. Clearly, too, if my history of the past quarter century or so is correct, greater ease-of-use is necessary if data analysis is to spread to substantive subjects. Although I have no hard evidence, I suspect that the effort to produce greater ease of use is producing instructional software that is increasingly valuable for research purposes—the development of analytical graphical displays is one example—which is an interesting reversal of direction for the traditional flow of technology.

It is possible for data libraries to participate in this transformation, but they will have to change their traditional modes of operation in several ways. First, they will have to work with faculty members to identify analytical software that is easy-to-use and capable of analyzing and presenting data in a way that the faculty member finds useful. Second, they will have to create files for that software. Typically, this will mean creating files on a mainframe, exporting them in ASCII, downloading them to a micro, and modifying them for the analytical program. The program may be a statistical package, a spreadsheet, a graphics program, a database program, or something else. The choices are larger in the micro world, and faculty demands are and can be expected to be diverse. Third, data libraries should attempt to develop expertise in exemplars of a number of software types—e.g., statistical packages, databases, spreadsheets, graphics—because it will be necessary if they are to be able to provide support for the files they create and because faculty are likely to ask for recommendations. Fourth, as networks expand and take on some of the functions of mainframes, data libraries will have to learn how to create and maintain data servers for users at all levels of sophistication. Fifth, data archives should look to the creation of display-formatted tables resident as files on disk as reference works for their most heavily utilized files.

While some tables on many subjects will be available on CD-ROMs from a number of vendors, it should be possible to create tables from archival holdings that serve the needs of particular programs at a cost significantly lower than would be required to manufacture a CD-ROM; software exists to compress such files and expand them as they are called by programs.

All of these possibilities for data archives require new investments—albeit at a relatively modest level—at a time when funding for new ventures is difficult. Given the cost pressures higher education now faces and will likely to face during the next decade, it is more likely that funds for these initiatives will come from a more efficient utilization of existing resources than from new ones.

One is supposed to close discussions of the future on an optimistic note, and I will try to do so. The transformation of computing offers substantial opportunities for using the data in our archives more broadly. We can move beyond our traditional support of faculty and graduate student research to make more of an impact on undergraduate instruction. But it will take initiative and a careful marshalling of resources. Otherwise, the past is, at best, all too likely to be prologue.

1.Paper presented at IASSIST 1990 in Poughkeepsie, New York.

2.Workshop on the Management of a Data and Program Library, Proceedings eds Magaret ONeil Adams, David Elesh, and Alice Robbins. Madison, WI, 1990.

3. I do not wish to re-open old, and typically, fruitless, debates about the relationship between theory and empirical research. I simply wish to note that neither theory nor research was much concerned with dynamic relationships in the 1960s.

4. U.S. Office of Education, Digest of Educational Statistics, Washington, D.C., Government Printing Office, 1990, p. 301.