Educating the data user:
the data archivist and bibliographic instruction

by Bliss B. Siman[1]
Associate Professor, Data Archivist
Baruch College, City University of New York

Information specialists have long been aware that graduates usually leave academia with only a rudimentary knowledge of information strategies and resources. The expansion of information access opportunities has not changed this situation radically. In the area of information about public data sources, the gap between what students know and what they could know, is tremendous. At Baruch, this problem is particularly important because there are few areas more dependent on information access and utilization, particularly in machine–readable formats, than business, and few academic disciplines, therefore, in which this "information gap" has greater significance. Since the Library at Baruch College provides access to information in almost all currently available formats: print, microform, audiodisc, microcomputer diskette, CD–ROM, and machine–readable data files, its instructional program has tried to include all of these information technologies in its workshops and courses. Sometimes this has come about quite by accident.

When the library began to collect machine–readable data files (MRDF), the activity was assigned to me based on my expressed interest. Since it was not a full–time assignment, I continued teaching in the bibliographic instruction program. In retrospect, it was a fortuitous combination of responsibilities. Working with data users, I became aware of the gaps in their information skills, the very skills I was teaching in my sections of the "Information Research in Business" course. Most data users were capable of using SAS or SPSS.x to analyze their data, and many could write elegant COBOL programs, or download data into LOTUS, but almost none had training in how to search for and identify quality data. Few graduate students or faculty members were sufficiently aware of the vast potential of public data for their research or teaching.

Surprisingly, many sophisticated faculty researchers continued to use datasets first introduced to them by their Ph.D. mentors simply because they were unaware of alternatives. Few were familiar with the varied storage media for data and how these could be effectively combined. For example, a faculty member working on a project using the Census of Population and Housing on magnetic tape might be totally unaware that portions of the work could be better accomplished using the same data in print. Clearly there were many exceptions to this bleak picture, but data users were not able to use the wide variety of sources available due to lack of information skills. Finding the means to overcome this deficiency became an important objective of a combined

[1] Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Washington, D.C., May 26–29, 1988

Data Archives/Bibliographic Instruction Program.

At the same time that the Data Library was being developed, other sections of the instructional program were beginning to include numeric data as part of the information process being taught. Online computerized retrieval services began to provide access to numeric data, and instruction in print sources of data was increasing in sophistication. Naturally, with all these programs being developed in the same division, there was a good deal of beneficial "cross-fertilization" in the planning and implementation of the instructional programs as well as the data service. Recognition of the competitive importance of numeric, quantitative information in business reinforced our desire to equip Baruch graduates, many of whom are the first generation of their families to go to college, with the data access skills they lacked. With respect to information literacy, business requirements were growing and we wanted to be sure that our students were prepared. Our method of attacking this problem also reflects a basic philosophy that information resources in general are underutilized due to the public's lack of training and education in research resources and research skills.

Consistent with this philosophy, the Data Resources Service immediately organized a seminar series to introduce users to numeric data sources. These seminars, which are still offered on a regular basis, were organized by subject field and publicized to faculty and graduate students. Unlike many similar seminars given in computer centers and data archives, these presentations included not only discussions of important datafiles, but also information on the same data in print, or the major print reference works in the same field. Often the data being discussed was also available through online vendors or on microcomputer diskette, and the criteria for using the various media were presented with practice problems to illustrate important points. The objective of the seminars was not to provide a "shopping list" of datafiles, but rather to equip the audience with the skills to find and evaluate the data needed for a particular project. Most importantly, no seminar failed to include information on important sources, whether data archives, government agencies or private vendors, of new data in the field, both general and specialized resources.

Models for locating data in a new field were discussed. These strategies paralleled those taught in the "Information Research in Business" course, in which students are equipped with the ability to identify sources in new fields as a basic part of modern information retrieval skills. The reception this information received underscored the need the modern data user has to identify quality data when beginning research in an unfamiliar area. For new graduates in entry-level positions, such skills can be invaluable. Many of these seminars included an online demonstration of some dataset using, where appropriate, SCSS (the interactive version of SPSS). The object of these seminars was to present MRDF in the context of other information sources, as well as to educate users in the available resources.

Initially, the target audiences were those who were already data users, individuals who had used data and were probably aware of the limitations of their knowledge of public data file availability. Because they were knowledgeable about data, they particularly appreciated training in strategies for finding data sources when the usual avenues were unproductive. These users were also receptive to information on selection of format of data, because they were also not aware of the many choices that could be made. Widening the scope of data knowledge and information retrieval skills among attendees was the primary objective of the first seminars. This objective remains central, although the seminars today are often presented to those who don't know very much about data. Of necessity, lists of data files are distributed when the seminar is attended by new data users. Although analytic issues are sometimes discussed, research methodology is never the focus. When Baruch became the coordinator of the ICPSR membership for the City University, these seminars were extended to the entire University with equal success.

Building on this basic format, we have experimented with workshops which include specific demonstrations of data sources that appear in online format, perhaps in print, and as machine-readable data files. For example, the Trinet database, which contains market share data for companies, is issued online and on magnetic tape. The Computer Search Services Librarian and I conducted this seminar to demonstrate the pros and cons of using numeric data online versus on tape. Baruch faculty and graduate students have become informed users of online information services, partly because these services have been free, and partly due to the excellent assistance that has been available to online users. But the consequence has been an expectation among the clients of the online services that all information, bibliographic and numeric, will be found neatly set up in database format, easily retrieved by a packaged query language. Seminars in which the nuances of using different formats has been presented, have increased the sophistication and information proficiency of faculty and students, especially the graduate students.

On the other hand, one can be too successful in encouraging data users to become aware of the wide variety of data sources available. The demand for data increases and the multiplicity of data requested creates budgetary problems. Since our graduate students are primarily in business fields, their need for expensive financial and economic data cannot always be met within the Data Resources Service budget. Finding ways in which to meet their needs is a constant challenge. Recently the number of very specific, very limited financial datasets being requested (example: 10 years of currency prices for five specific countries) went far beyond the capacity of the Baruch computerized information services. Methods of creating datasets, rather than purchasing them from private vendors were explored.

Using an online vendor, I. P. Sharp, we downloaded small amounts of data to meet specific needs. In order to make the data more widely available, they were also uploaded to magnetic tape on the mainframe and listed in the data archive holdings. As another alternative, microcomputer diskette data sets in Lotus 1-2-3 format were also created. We are hoping that knowledge of these data will encourage users to plan their graduate theses, etc., around data that are available rather than devising projects dependant on costly new data sets. However, philosophically, it was important not just to create these data sets ourselves, but to educate users to the benefits of this technique, a technique that could be very useful to faculty in their current work and to students in their future business roles. Consequently, the entire process was demonstrated at a data seminar. The demonstration included how the data were identified, accessed online, downloaded to diskette, uploaded to magnetic tape and then accessed using SAS. At the same time, the use of the data at each intermediary step was discussed, providing greater depth to the users' understanding of the pros and cons of each technique.

As the variety of formats increases, we see increased possibilities for this type of seminar. For example, seminars on the census, could present alternative methods of searching for data files, beginning with bibliographic searches online, as well as the choices to be made among formats: print, microfiche, Cendata, machine-readable data files, diskettes, CD-ROM, etc. Census data are very much underutilized at many colleges of the City University, and we believe such instruction will assist users in identifying valuable data not previously considered.

Most of this discussion has focused on the kind of training given sophisticated data users to improve their information skills. But the Library Instruction Division and the Data Resources Service have been equally engaged in a dialogue concerning the competence that undergraduate students should master in order to attain "information literacy" with respect to public data sources. Actually, this dialogue is part of a continuing discussion and re-evaluation of the entire bibliographic instruction program, necessary in a rapidly changing information environment. There is not yet consensus on what represents an adequate set of information skills.

Although the Library Instruction Division (LID) offers a basic course in information research in business (Library 1016, Information Sources in Business) the core of materials covered has changed over time. In fact, although all instructors teaching the course use the same text and give uniform exams, there is considerable flexibility in planning the curriculum. Instructors are encouraged to experiment with materials and share their successes or failures with colleagues. As the importance of public data sources became recognized within the LID, the department began discussing methods of including this information resource in the curriculum. Frankly, a complete answer has not been arrived at, although several configurations have been used.

Clearly, undergraduates do not need sophisticated knowledge of public data, but they do need some awareness of the role of raw data in the information process, of the availability of data for secondary analysis, and where data can be obtained. The inclusion of data sources in the curriculum reinforced some of the conceptual goals of the course as well. Mastery of the development of search strategies is an important objective of the course, and data files, due to the lack of bibliographic control, provide an excellent example of non-standard information search strategies. The process of searching for data is, of necessity, very different from the index searching model which undergraduates come to assume is relevant in all situations.

Awareness of raw data files and of what constitutes an authoritative source in this field reinforces another objective of the course, that of enhancing the students' ability to evaluate the information they consume on a daily basis. Polls and government reports based on data are constantly presented in newspapers, magazines, etc. Questioning the data on which these reports are based is an important attribute of the educated citizen in private or professional life. By including data as an information resource in the research course, we are equipping students the better to handle this important task.

Working with my colleagues, I developed several different instructional modules which presented numeric data and the different formats in which they come. The lectures contained explanations of machine-readable data files and the nature of secondary analysis. Basically, I wanted students to understand how data are used in business and academic research and the retrieval methods used to find data files. These lectures were tied to different units of the course, depending on the instructor: marketing in one case, social science research in another. In each case, the lectures built on what the students had already learned about information access and retrieval as well as emphasizing the evaluation of sources. The lectures always included a demonstration of data use in an area that would pique students' interest.

During these demonstrations, students were encouraged to participate in the development of a hypothesis and the testing of it with data at hand. Although this technique was borrowed from modules developed for sociology courses, the emphasis was placed on information dissemination and evaluation issues rather than on research methodologies. These demonstrations were very effective in making students aware of the whole process of research transmission, but it is not clear that the best combination of subject year and lecture format has yet been found.

Despite their success, these lectures have not yet become a standard part of the course for several reasons. As with every academic institution, we are understaffed. It is not always possible for me to do these lectures at the appropriate time in each instructor's syllabus. Nor do the instructors necessarily have the time in a single semester course to devote a full lecture to data files. The amount of material we would like to cover in a course is far greater than can be managed in a semester. Topics are constantly being juggled as we search for the optimum mix. However, the Department has a commitment to including the identification and retrieval of data files in the basic information course because we feel that use of these sources is an important information skill. Data users must first be able to find data, and the bibliographic instruction course can equip them with the skills to do the job efficiently and cost-effectively.□