JUDITH S. ROWE and MARY RYAN

Library Service from Numerical Data Bases: The 1970 Census as a Paradigm

This article discusses some of the problems of introducing machine-readable data bases into the library service environment. The authors, a social scientist at a computer center, and a government documents librarian, describe the diverse approaches used in making tapes of the 1970 Census of Population and Housing *available to users through the library.*

LARGE RESEARCH LIBRARIES have traditionally been depositories for all of the maps and printed reports which are the products of each decennial census. Therefore it is a logical next step for them also to be the repository of these data in machine-readable form. First, this provides reference librarians with another resource for users whose needs are not satisfied by searching the printed materials, since the quantity of additional data which can be stored compactly on magnetic tape has made it possible for the Bureau of the Census to make available to the public at least ten times the amount of data available from any previous census. Second, research libraries are generally located at institutions which also have available large computers capable of selecting, digesting and analyzing these data and, if tapes are available, it becomes unnecessary for the user requiring a machine analysis to photocopy pages from reports and then keypunch the data. Rather, it becomes possible for a researcher to begin with the data already in machine-readable form and to proceed immediately to work with these data. Thus, we now have more data in a more usable format than ever before.

But why should the library be involved? Why not just store the tapes at the computer center and let the computer people worry about them? By doing this, the library would be abdicating part of its role as an information center. It would be denying users the opportunity of locating information at the one place we have trained them to look for it, the library. By the same token, since the acquisition of even a modest tape collection represents a substantial financial outlay, it is important to ensure that the responsibility for decisions on acquisition, on bibliographic documentation and control, and on reference service be allied with an organization with a commitment to public service and a continuity of collections and operation. Although many computer centers have a commitment to public service, computer centers are not libraries and their staffs do not have the library skills necessary for such a project. On the other hand, librarians do not generally have the computer expertise necessary to exploit to the fullest this new information resource.

What then is the solution? Actually

Judith Rowe is manager, Princeton-Rutgers Census Data Project, Princeton University Computer Center; Mary Ryan is head, Public Affairs Service, University of California, Los Angeles, library.

there is no one best solution. Different plans are now in operation and many more are possible. In this report we will describe in detail two very different approaches—at UCLA and at Princeton. As background, however, let us first present some of the questions which must be faced in arriving at a locally workable solution.

1. Who will decide which data should be purchased? To some extent this is a matter of money, but other considerations are also involved, most of them similar to those involved in the building of any library collection. How does one service all or most requests without the problems of supporting a huge amount of unused material?
2. What quantity and type of service will be provided? This depends on the availability of both funds and people. It also depends on the needs of users and on the nature of any other services of this sort already available in one area.
3. Where will the money come from for acquisitions, for computer time, for personnel, and for a myriad of miscellaneous expenses such as travel, duplicating, backup reels, telephones and clerical assistance? Many libraries are having great difficulties in maintaining even their traditional acquisitions and services. Assuming that one's total resources are not increased, does one reallocate resources and, if so, what criteria does one use, and just whose resources does one reallocate? It should be borne in mind that a complete collection of 1970 census summary tapes, purchased from the Bureau of the Census itself, would cost well over $100,000.
4. How is bibliographic control over machine-readable data bases to be exercised? Should records for these be interfiled in the card catalog? Given the complexity of these data bases and the volatile nature of machine storage, is the traditional catalog record sufficient? Are supplementary materials such as data description forms and content documentation codebooks necessary and, if so, how are the records to be integrated? No fixed solution yet exists, but an ALA subcommittee has been established to recommend rules for the cataloging of machine-readable data files.*
5. The whole question of staffing gives rise to many problems. What kind of staff and how much staff are needed and can be afforded? What background should they have? How are they to be trained? Under whose supervision should they be and how does one train the supervisor? Who will assume responsibility for those problems which are general to the nature of machine-readable data files and those which are specific to the census itself? These include such things as the vast amount of data involved, the relation of these data to the printed reports, the complexity of the tapes, questions of geographic coding, the necessity of maintaining and updating computer programs, errors and inconsistencies in the data, recalls of tapes, printed corrections to erroneous data on the tapes, problems in labeling and identifying data sets,

*John Byrum, head cataloger at Princeton University is chairman of the RTSD/CCS/DCC Subcommittee on rules for cataloging machine-readable data files. Other subcommittee members are: Lawrence W. S. Auld, Oregon State University, Henriette Avram, Library of Congress, Gerry Dobbin, University of British Columbia Library, Elizabeth Herman, U.R.L.—UCLA, Judith S. Rowe, Princeton University Computer Center.

and a great many problems in connection with documentation, ranging from its quantity, combined with inadequate indexing and correlation, right down to such minute problems as some of the items most needed by users being printed in colors that photograph poorly.

6. Where should the staff be housed? Where should the data tapes and the supporting documentation be stored? At the library? At the computer center? Somewhere on neutral territory? Or should the operation be separated, with reference service being provided at one location and data access and use at another? Here, as elsewhere, communication is a matter of paramount importance. Computer programmers, librarians, and social scientists often find it difficult to talk the same language. Is the answer an interdisciplinary specialist? There are also problems within the library itself. For every department in the library, from acquisitions to cataloging to reference, this new medium poses new problems and we can testify from experience that few precedents exist to aid in their solution.
7. What is the potential user community? Who will have direct or indirect access to the data? Is the service primarily designed for one's own campus users or for the outside public? Will there be preferential treatment, and, if so, on what basis will it be accorded? For example, would students, faculty, researchers with outside grants, and profit-making firms all have equal access and would they all pay the same rate for any charges involved?
8. Will there be user charges and, if so, on what basis will they be determined? If user charges are to be instituted, how much demand for service will there be, especially in a nonmonopoly situation, and how much will charges have to run? If demand is sporadic, for example, would charges have to be unrealistically high? Should they be the same as those of profit-making organizations providing similar services? What will one attempt to recoup with user charges and what will be the public relations effects? Probably few people would question charges for computer time or for staff time spent on special programming, but what about staff time spent in reference, orientation, and consultation?

At precisely what point, for example, would a university library halt the free reference and information service it offers its faculty on its census resources and put it on a for-sale basis? ("Well, Professor Jekyll, I can answer questions on Part I of the *Census User's Guide,* provided they don't touch on the summary tapes, but no questions about Part II," or: "You can look at the Census Bureau's fourth count documentation without charge, since it's depository, but it will cost you x dollars to use DUALabs' version.") At what point would conversations pass from the free orientation service stage and become priced consultations? Does one attempt to amortize the cost of the data base—and would there be public relations problems here? ("Well, Professor Jekyll, the reason you must pay to use $1,000 worth of the library's census data, while Professor Smith can use $1,000 worth of the library's Sanskrit manuscripts without charge, is that we amortize tape data but not manuscript data.") What about amortizing the cost of supporting tools, such as maps? ("Well, Professor Jekyll, you may look at the commercial map of Butte County free, since we

didn't buy it to support computerized census, but you must pay a fee to look at the census map for that county." "Professor Jekyll, why are you beginning to look like Mr. Hyde?")

Having surveyed some of the major problems, we will now look at two solutions. The University of California, Los Angeles, and Princeton approaches are very different in origin, scope, resources, and services, but they are similar in that in both instances libraries and librarians are fully involved.

The UCLA Program

At UCLA there was no central social science data library and no history of library involvement in machine-readable numerical data acquisitions. However there was a large computer available supporting several major statistical software packages, as well as some data processing personnel involved in other projects within the library, a large government documents collection, and a potential user community of several hundred.

Many commercial operations in the Los Angeles area had early announced their intention to serve as Census Summary Tape Processing Centers; in addition, a nonprofit, self-supporting Census Service Facility had been established at the University of California, Berkeley. This facility was operating under the joint auspices of Berkeley's Institute of Governmental Studies and its Survey Research Center, a group with a long history of service as a social science data archive. It had no connection with the library. It offered a wide range of services, emphasizing the production of standardized tabulations, as well as performing customized work tailored to individual requests. Although it would serve private organizations and individuals, its efforts were particularly directed towards academic and governmental users; for UCLA to have duplicated Berkeley's services would have been needless and wasteful.

However, it was obvious that students, faculty, and research personnel in many different fields would themselves need to be working directly with the machine-readable data on an individual basis. Unless a central source for the data and tools were available to them, various UCLA departments, schools, and institutes would each have to obtain the tapes independently and, while there would then be much wasteful duplication, there would still be no central source of information and no general availability of the data tapes.

Foreseeing this situation, the library undertook the responsibility for serving as a central campus resource for census tapes and for information about them, including appropriate cataloging. The library system was already involved with several machine-readable data bases, including MEDLARS and MARC. In addition, there was a Center for Information Services (CIS), funded by the National Science Foundation and then in the second of four phases, which has the specific purpose of giving the library the capability of acquiring, cataloging, and providing services for machine-readable data bases, whether bibliographical, numerical, or full-text. CIS has as its first priority the bibliographic files and its experimental services have included searching of CA Condensates, Compendex, CAIN, and the ERIC files. The census represented its first involvement with a numerical file.

Although the need for a census service was very apparent, the resources available to the library were extremely limited. Had an attempt been made to recover costs through user charges, potential revenue would have been minimal because of the existence of Berkeley's Census Service Facility, or would have been siphoned off from that facili-

ty. UCLA has no specific budget allocation for either census data acquisition, processing, or reference service. These are all paid for on an *ad hoc* basis out of the library budget, or by other parties, such as departments, willing to contribute.

Despite the administrative and financial difficulties, it was decided that the UCLA library would attempt to offer service, within the limits of its resources, to meet the most crucial campus needs. The course of action which seemed most appropriate, given the above framework of limited experience and stringent budgetary considerations, was for the library to join the START (Summary Tape Assistance, Research, and Training) Community organized by DUALabs (Data Use and Access Laboratories, Rosslyn, Virginia) under the sponsorship of the Center for Research Libraries, with aid from the Ford Foundation. Through this community, it would be possible to purchase tapes at a price substantially less than that of the Bureau of the Census and to take advantage of the programs already developed by DUALabs to avoid incurring the heavy cost of original programming.

Within the UCLA context, it was obvious that the logical library department to undertake the census tape service was the Public Affairs Service (PAS). Among PAS' key responsibilities is that of the library's government documents service. Thus, it receives the current printed reports from the Bureau of the Census, has a heavily-used reference service specializing in government documents, and has had long experience with the census printed reports. Even more important, PAS, which incorporated several older services such as government documents, had been created in 1968 to offer a coordinated information service to those working in the fields of government and public affairs, broadly interpreted. As a department of the research library, designed to supplement that library's more traditional resources, PAS was directed to place no limits on the kinds of material or forms of data that were acquired or used, so long as they were pertinent to the needs and interests of the clientele serviced. The census in machine-readable form is, in fact, a perfect example of the unconventional library resource which the department was created specifically to handle.

With this mandate, census tapes for California, plus the necessary tools—programs, the MEDLIST, etc.—and the needed documentation, were ordered. A specialized census reference service is now offered which includes extensive personalized orientation and a limited amount of consultation. General questions about the tapes are answered at PAS' regular reference desk, but this specialized service is a separate and distinct service within PAS. There were several reasons for this. First, the reference desk is an exceedingly busy place where questions must be answered expeditiously, or suggestions made to enable the user to start on his own search, so that the next reader waiting can be helped; since the typical initial census tape orientation takes at least one hour, it could not be handled as part of the regular service without causing that service to break down. Furthermore, there are in all nine librarians and seven others who are scheduled at the Public Affairs reference desk, and, given the time needed for someone to become trained in the census service and to keep up with the continuing flow of documentation, it was not economically feasible for all the staff to participate in the service.

This service is purely a reference and orientation service. It is not a production operation. Programs are available for the clientele to use, but the librarians do not themselves use them. Nor do

librarians offer such services as printing out data from the tapes for people, manipulating data for them, or doing any data processing. Instead, they make the data and the tools available to the user, so that he may do his own work. If he wants the processing done for him, he may obtain this service from Berkeley or another Summary Tape Processing Center at their stated rates.

The reference and orientation service is available to any inquirer and in fact the UCLA library serves as a User Contact Site of the Clearinghouse and Laboratory for Census Data, operated by DUALabs under contract with the Center for Research Libraries.

Copies of the tapes themselves and of the programs are kept at UCLA's Campus Computing Network (CCN) with access restricted to those who have received prior authorization from the library. Access is authorized by the library as a matter of course for anyone who has CCN computer time and he may then use the tapes and programs in accordance with CCN's standard procedures. This authorization system is designed to provide data on demand and usage to aid in planning. It enables the reference staff to ensure that potential users are acquainted with the documentation and are aware of the printed reports before they start their work with the tapes. It enables a degree of security to be maintained over the tapes, since it reduces the chance of damage caused by a completely uninitiated and untraceable user. In addition, insofar as users are willing—and this is done only with their consent—it permits the library to act as a clearinghouse to alert users to similar projects already underway on campus.

At the present time, no charge is made by the library for its census reference service or for access to its tapes and programs, though this may well change. CCN of course applies its standard procedures in relation to the computer time. If a user is unable to obtain CCN computer time, or prefers not to, CIS will sell him copies of any of UCLA's census tapes or programs, so that he may use them at his own computer facility. At one time, tape lending was considered, but the problems inherent in such a procedure made this impractical.

At present, therefore, the library's service is completely designed for "do-it-yourselfers." In some ways, it is analogous to typical academic library service on books in foreign languages. For example, bibliographies of German books, the German books themselves, dictionaries, indexes of translations, and directories of translators and translation centers are made available to the reader. He is offered reference service, helping him to identify these items—but not a translation service for German books.

There is a group of services ranging from reference through programming and keypunching, to which the user of census data, or of any data in machine-readable form, must have access. However, the decision as to which of these services will be supplied directly by the library and which will be handled elsewhere, with or without library involvement, is a matter of local option. These services are supplied somewhat differently at Princeton than at UCLA, largely because the origins of the Princeton Census project, and its financing, are different.

The Princeton Program

When the Princeton Library was approached by the Center for Research Libraries about acquiring the census tapes, two important precedents already existed: a tradition of cooperation with Rutgers for the purpose of avoiding duplication of special collections, and a prior acceptance of numerical ma-

chine-readable data files as a legitimate library resource. Since it was obviously less expensive to combine forces, an agreement was signed which created the Princeton-Rutgers Census Data Project. For the Princeton library this was a logical sequel to other similar steps. For example, Princeton is a member of the Inter-University Consortium for Political Research and the Consortium Membership Fee has for several years been part of the library budget, although related computer time and personnel services are funded by the Computer Center. When this decision was first made at Princeton, it was innovative. Today several other schools belonging to the Consortium have followed the Princeton pattern and more are considering doing so. However, it is not enough merely to foot the bill. Someone must undertake responsibility for the acquisition, the storage, and the use of the data. These responsibilities had been assumed for many years at Princeton by what is now known as the Social Science User Services (SSUS) section of the Computer Center and therefore it was logical for this group to perform the same functions in relation to the census tapes.

The advisory committee involved in the organization of the Census Project included, in addition to librarians, representatives of the Computer Center and of several departments in which there were potential census users. This committee confronted immediately the fact that funds would be needed for data acquisition and storage, computer use, and programming support. Contributions were therefore secured not only from the Princeton and Rutgers libraries but also from the budgets of several large research projects which already contained allocations for the acquisition of census data. The directors of these projects were all pleased with an arrangement which would afford them access to the data tapes without the burdens of acquisitions, bibliographic recording, physical control, or software development and at less cost than if individual purchases were to have been made.

Practically, how does the project function? The Princeton-Rutgers Census Data Project is recognized by the Bureau of the Census as a Summary Tape Processing Center and by the Clearinghouse and Laboratory for Census Data as a User Contact Site. This means that inquiries at the project office, which is located at the Princeton University Computer Center, frequently come from nonuniversity sources. However, on the four major campuses served, people are accustomed to looking for printed census data in their libraries and the project encourages them to continue doing so. With the aid of the Census Packet (a monthly acquisitions list covering both printed and machine-readable materials, which also includes lists of suggested readings, descriptions of maps and other geographic aids and miscellaneous training materials), Part II of the Census User's Guide and a one-day orientation, most of our reference librarians are aware of the potential of the machine-readable data. The librarians who are directly responsible for the Bureau's printed reports, particularly the Public Administration Librarian at Princeton and the Government Documents Librarian at Rutgers, can answer questions concerning the specific tables which are included in the machine-readable files and the geographic areas which are more adequately covered by these and, as a result, it is now quite common for the project to receive a call from a library user asking, for example, "Can I get Table 29 in the Second Count for all minor civil divisions in four New Jersey Counties?"

Reference librarians are still doing census reference work but with expanded resources. Those users who come directly to the project office, but without such a specific request, are first referred to the printed reports, samples of which are available for use there. Frequently, the next step is to send them to the nearest library census collection, but in those cases where it is evident that they will need machine-readable data, two options are open to them. They may have the necessary retrieval done for them, and for this there is a charge of $8.50 per hour for programming time with a $25.00 minimum per request or they may themselves use the available programs and access the necessary data tapes in which case they receive relevant orientation and documentation without charge. In either case they would pay for computer time but *not* for consultation or for use of the tapes.

In administering the project the Social Science User Services section of the Computer Center has provided a bridge between the library and the computer. The services performed by this group are similar to those provided by comparable centers throughout the country —sometimes at computer centers but more often attached to research institutes or social science departments. However, the Social Science User Services section is unique in that it maintains and has maintained for years an active relationship with both the reference and technical services staff of the Princeton library so that in many ways it functions as a special library. Its regular services include responsibility for acquisition and control of much of the social science and literary data in machine-readable form obtained by or generated at Princeton University, and then released for public use. This involves providing reference service not just for the data tapes but for all of the supporting documentation, including codebooks and records of physical and logical characteristics. SSUS also maintains three major statistical packages (OSIRIS, SPSS, and Data-Text) for analyzing data on the computer, and has access to a fourth, the Princeton-produced P-Stat. Consultation is available for any computer-oriented research project in the social sciences or the humanities; this process can cover all phases of methodology from questionnaire design to analysis. All of these services are provided without charge to members of the university community; users pay only for special programming and keypunching.

This then is how census tape service has developed at UCLA and at Princeton-Rutgers. Many other solutions are possible within a library framework, but whatever the approach finally adopted by a library, it is certain that there are many potential users not only of census data in machine-readable form but also of the vast array of other machine-readable data resources which are becoming publicly available. In spite of the obvious technical difficulties, these are clearly significant information resources and as such should not be ignored by libraries. Machine-readable information resources are now available to a greater or lesser extent at virtually every research-oriented college and university and in government agencies at all levels, but the number of instances in which the libraries at these institutions are involved or even aware of these resources is sadly small. No library, regardless of its lack of technical expertise, should completely surrender its responsibility as an information center to an academic department, a research group, or a computer center. It is not necessary that librarians hold these data physically in the library or that they process or even know how to process them, but it is necessary that reference librarians have enough knowl-

edge of these data to advise users of their existence, their general contents, and of the means by which they may be accessed. Failure to do so is likely to result in proliferation of competitive services in an area in which costs can become extremely high.

Financing

For those libraries which have the desire and the capability to become more heavily involved in providing data services, the question of financing may seem insoluble. Once a library has decided how far along the continuum of possible data services it wishes to move (and this decision will inevitably be colored by the nature of any existing service at one's location), the question of how to finance these new activities must be confronted. Assuming that a library's budget cannot be increased to meet this new demand, three alternatives are available, all of which may seem at odds with traditional library policy, but all of which have already been implemented by libraries. The first alternative is the re-allocation of existing library resources. The second is to institute user charges. These may be applied directly to each individual user or may be paid by departments, agencies, or research groups in a lump sum determined by usage. User charges might well be applied to all computer-related work but probably not to basic reference service in most libraries. It is certainly simpler and perhaps more acceptable to users to charge for keypunching, programming, and machine time than for orientation and basic consultation. The other alternative which could be implemented either separately or in combination with user charges is that of outside subsidy. Traditionally, libraries have provided out of their general fund for acquisitions requested by departments without expecting reimbursement from departments. Since, research budgets often contain provisions for machine-readable acquisitions, it seems entirely appropriate that when possible these funds be funneled through a central channel, the library, especially since a contract or grant may actually require that any such data become the property of the institution rather than of the individual. Although it must not be overlooked that many granting agencies specifically prohibit the purchasing of library resources on their funds (since the preexistence of adequate library facilities was a basic reason for awarding the grant to that institution in the first place) this seems an administrative problem capable of solution. All things considered, it seems logical that, before embarking on the major new activities that service from machine-readable data bases represents, a library might well solicit contributions from potential user groups, whether on-campus or off-campus, and employ any such contributions as seed money for initial acquisitions and processing.

To summarize, no library can completely abdicate its involvement in machine-readable data resources, unless it elects to abdicate part of its responsibility as an information center. However, since other nonlibrary centers may by default have assumed many of the functions involved, the degree of library activity must take account of the existing situation. At the very least, communication with the nonlibrary center would always be desirable, as would entries for machine-readable data files in the public catalog and basic reference service. This minimal activity must become an integral part of the service of every research library.